# 第八讲 内生解释变量问题

模型中出现随机解释变量且与随机误差项相关时，OLS估计量是有偏的。

如果随机解释变量与随机误差项异期相关，则可以通过增大样本容量的办法来得到一致的估计量；

但如果是同期相关，即使增大样本容量也无济于事。这时，最常用的估计方法是工具变量法（Instrument variables）。

## 一、工具变量的选取

**工具变量**：在模型估计过程中被作为工具使用，以替代模型中与随机误差项相关的随机解释变量。

**选择为工具变量的变量必须满足以下条件：**

**（1）与所替代的随机解释变量高度相关；**

**（2）与随机误差项不相关；**

**（3）与模型中其它解释变量不相关，以避免出现多重共线性。**

# 二、工具变量的应用

**以一元回归模型的离差形式为例说明如下：**

$$y_i = \beta_1 x_i + \mu_i$$

用OLS估计模型，相当于用$x_i$去乘模型两边、对$i$求和、再略去$\sum x_i \mu_i$项后得到正规方程：

$$\sum x_i y_i = \beta_1 \sum x_i^2$$

解得

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \qquad (*)$$

由于$Cov(X_i, \mu_i)=E(X_i\mu_i)=0$，意味着大样本下

$$(\sum x_i\mu_i)/n\to 0$$

表明大样本下 $\hat{\beta}_1=\dfrac{\sum x_i y_i}{\sum x_i^2}$

成立，**即OLS估计量**具有一致性。

然而，如果$x_i$与$\mu_i$相关，即使在大样本下，也不存在 $(\sum x_i\mu_i)/n\to 0$，则

$$\hat{\beta}_1=\frac{\sum x_i y_i}{\sum x_i^2}$$

在大样本下也不成立，OLS估计量不具有一致性。

如果选择Z为X的工具变量，那么在上述估计过程可改为：

$$\sum z_i y_i = \beta_1 \sum z_i x_i + \sum z_i \mu_i$$

利用$E(z_i\mu_i)=0$，在大样本下可得到：

$$\tilde{\beta}_1 = \frac{\sum z_i y_i}{\sum z_i x_i}$$

关于$\beta_0$的估计，仍用$\tilde{\beta}_0 = \bar{Y} - \tilde{\beta}_1 \bar{X}$完成。

这种求模型参数估计量的方法称为**工具变量法(instrumental variable method)**，相应的估计量称为**工具变量法估计量**（**instrumental variable (IV) estimator**）。

对于矩阵形式：

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\mu}$$

采用工具变量法（假设$X_2$与随机项相关，用工具变量Z替代）得到的正规方程组为：

$$\mathbf{Z}'\mathbf{Y}=\mathbf{Z}'\mathbf{X}\boldsymbol{\beta}$$

参数估计量为：

$$\tilde{\boldsymbol{\beta}}=(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y}$$

其中 $\mathbf{Z}'=\begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{12} & \cdots & X_{1n} \\ Z_1 & Z_2 & \cdots & Z_n \\ \vdots & & & \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{bmatrix}$ 称为工具变量矩阵

# 三、工具变量法估计量是一致估计量

一元回归中，工具变量法估计量为

$$\tilde{\beta}_1 = \frac{\sum z_i(\beta_1 x_i + \mu_i)}{\sum z_i x_i} = \beta_1 + \frac{\sum z_i \mu_i}{\sum z_i x_i}$$

两边取概率极限得：

$$P\lim(\tilde{\beta}_1) = \beta_1 + \frac{P\lim \frac{1}{n}\sum z_i \mu_i}{P\lim \frac{1}{n}\sum z_i x_i}$$

如果工具变量Z选取恰当，即有

$$P\lim \frac{1}{n}\sum z_i \mu_i = \mathrm{cov}(Z_i, \mu_i) = 0 \qquad P\lim \frac{1}{n}\sum z_i x_i = \mathrm{cov}(Z_i, X_i) \neq 0$$

因此：

$$P\lim(\tilde{\beta}_1) = \beta_1$$

**注意：**

**1、在小样本下，工具变量法估计量仍是有偏的。**

$$E(\frac{1}{\sum z_i x_i}\sum z_i \mu_i) \neq E(\frac{1}{\sum z_i x_i})E(\sum z_i \mu_i) = 0$$

**2、工具变量并没有替代模型中的解释变量，只是在估计过程中作为“工具”被使用。**

上述工具变量法估计过程可等价地分解成下面的两步OLS回归：

第一步，用OLS法进行X关于工具变量Z的回归：

$$\hat{X}_i = \hat{\alpha}_0 + \hat{\alpha}_1 Z_i$$

第二步，以第一步得到的$\hat{X}$为解释变量，进行如下OLS回归：

$$\hat{Y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 \hat{X}_i$$

容易验证仍有：$\tilde{\beta}_1 = \dfrac{\sum z_i y_i}{\sum z_i x_i}$

因此，工具变量法仍是Y对X的回归，而不是对Z的回归。

**3、如果模型中有两个以上的随机解释变量与随机误差项相关，就必须找到两个以上的工具变量。但是，一旦工具变量选定，它们在估计过程被使用的次序不影响估计结果(Why？)。**

4、OLS可以看作工具变量法的一种特殊情况。

5、如果1个随机解释变量可以找到多个互相独立的工具变量，人们希望充分利用这些工具变量的信息，就形成了广义矩方法（**Generalized Method of Moments, GMM**）。

在**GMM**中，矩条件大于待估参数的数量，于是如何求解成为它的核心问题。

工具变量法是**GMM**的一个特例。

6、要找到与随机扰动项不相关而又与随机解释变量相关的工具变量并不是一件很容易的事

可以用$X_{t-1}$作为原解释变量$X_t$的工具变量。

# 四、案例分析

- 1.实验内容和数据来源
- 根据某统计资料，得到如下统计数据可用来估计教育投资的回报率。文件名：“grilic.dta”。
- 重要变量说明，lw80（80年工资对数），s80（80年时受教育年限），expr80（80年时的工龄），tenure80（80年时在现单位工作年限），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），mrt（婚姻虚拟变量，已婚=1），age（年龄）。
- 本实验使用该数据建立合理的模型对教育投资回报率进行探究，并对模型可能存在的内生性问题进行了检验和合理的处理。

- 实验操作指导

1、确定回归模型形式

根据研究问题的意义和数据，建立方程：

$lw80=\beta_1 s80 + \beta_2 expr80 + \beta_3 tenure80 + \varepsilon$

2、对模型进行回归

在 Stata 命令窗口中输入如下命令打开所用数据文件并对模型进行基本回归：

```
sysuse grilic.dta, clear
reg lw80 s80 expr80 tenure80
```

- 得到
- lw80=5.514+0.075s80+0.0202expr80+0.007tenure80(估计量均显著通过t检验)

- 2、对建立的模型进行分析
- 由于对于工资高低的解释中，个人能力是一个很重要的因素，然而这个因素却无法直接衡量，在实验数据中发现智商（iq）可以是能力的代理变量。
- 当然iq作为能力的代理变量是存在误差的，可能与扰动项有关，即我们认为可能iq可能存在内生性。而且若模型中加入iq作为解释变量，s80可能与扰动项中除去能力以外的扰动因素有关，即s80也可能具有内生性。因此建立的模型中iq,s80可能是模型中的内生解释变量。

- 3、对建立的模型进行分析
- 由于对于工资高低的解释中，个人能力是一个很重要的因素，然而这个因素却无法直接衡量，在实验数据中发现智商（iq）可以是能力的代理变量。
- 当然iq作为能力的代理变量是存在误差的，可能与扰动项有关，即我们认为可能iq可能存在内生性。而且若模型中加入iq作为解释变量，s80可能与扰动项中除去能力以外的扰动因素有关，即s80也可能具有内生性。因此建立的模型中iq,s80可能是模型中的内生解释变量。

# • 4、内生性处理

- 内生性处理方法中2SLS，GMM和迭代GMM方法，在Stata命令中格式是统一的：
- ivregress estimator y [varlist1] (varlist2 = varlist_iv) [if] [in] [weight] [, options]
- ivregress表示对模型进行内生性处理语句，其中estimator代指 2sls或者gmm两种方法，varlist1表示模型不存在内生性的解释变量，varlist2 = varlist_iv表示模型中存在内生性的变量和解释其的工具变量，if表示回归的条件，in表示回归的范围，weight表示回归中加入放入权重，options的内容如下表所示：

| options | 描述 |
| --- | --- |
| Model | |
| nonconstant | 不包括常数项 |
| hascons | 用户自己设定常数项 |
| GMM | |
| wmatrix(wmtype) | wmtype 可能是 robust，cluster clustvar，hac kernel, unadjusted |
| center | 权数矩阵采用中心距 |
| igmm | 采用迭代 gmm 估计法 |
| eps(#) | 指定参数的收敛标准。默认值为 eps(1e-6) |
| weps(#) | 权数矩阵的收敛标准。默认值为 wps(1e-6) |
| optimization options | 控制最优化的过程，很少用的 |
| SE/Robust | |
| vce(vcetype) | Vcetype 可能是 robust，cluster clustvar，hac kernel, unadjusted |
| Reporting | |
| level（#） | 设定置信区间 |
| First | 输出第一阶段的估计结果 |
| Small | 小样本下的自由度调整 |
| noheader | 仅显示估计系数表格 |
| depname | 显示替代变量的名称 |
| Eform(string) | 输出系数的指数形式并用 string 做其标签 |

- 下面介绍一种内生性处理方法：
- 2SLS法
- 此方法在Stata中命令语句格式如下，此命令就是将内生性处理命令语句基本格式estimator进行了具体化，所以基本结构和含义是相同的：
- ivregress 2sls y [varlist 1] (varlist2=instlist) [if] [in] [weight] [, options]
- 此命令语句2sls表示此种内生性处理语句是2sls方法，varlist1仍然表示不存在内生性的解释变量，varlist2 = varlist_iv表示模型中存在内生性的变量和解释其的工具变量，if表示回归的条件，in表示回归的范围，weight表示回归中加入放入权重，options内容与表8.9中的选项是一致的（除了GMM项）。
- 具体来说最常用的两个2SLS的命令语句：
- ivregress 2sls y x1 (x2 = z1 z2)
- 此命令表示2SLS法估计时使用的默认普通的标准差，且默认是结果只显示第二阶段的回归结果。
- ivregress 2sls y x1 (x2 = z1 z2), r first
- 此命令语句中其中 r表示使用稳健标准差，first表示在结果中显示第一阶段的回归。

- 本实验中，使用med,kww,mrt,age作为内生解释变量iq与s80的工具变量。使用2SLS法对模型进行估计时在Stata命令窗口中输入如下命令可以得到估计结果：
- **ivregress 2sls lw80 expr80 tenure80 (iq s80 = med kww mrt age)**
- 此命令表示使用2SLS法对模型进行估计，使用med,kww,mrt,age作为iq和s80的工具变量。

- 命令结果窗口中未显示第一阶段回归的结果，这里只是将最终结果列出来，读者可以自己试验看一下第一阶段结果。在最终的结果图中列示了instrumented(被使用工具变量解释的原解释变量)和instruments（所使用的工具变量）。注意：工具变量不会在最终的估计结果中出现。
- 从结果图中可以将模型具体化为如下的模型，且只有tenure的系数在10%的置信度下未通过显著性检验：lw80=4.089+0.0425s80+0.0175iq+0.0265expr80+0.00472tenure80，发现与原回归结果相比进行内生性处理后的估计值发生了很大变化，所以内生性处理是很有必要的。
- 另外，虽然知道2SLS法的两个步骤，但是若通过手动按步骤做出的结果是错的，因为手动方法计算时使用的残差序列是错的，所以此方法只能通过Stata来完成。

# • 5.内生性的检验：豪斯曼检验

- 前面已经解读了Stata中对内生性问题的处理，有时候仅从意义上观察不能确定内生性是否存在时，可以用以下方法检验模型是否存在内生性问题。
- 豪斯曼检验
- 在Stata中的命令语句为：
- hausman name-consistent [name-efficient] [,options]
- hausman语句表示豪斯曼检验，其中语句中 name-consistent是指一致估计量变量名，name-efficient是指有效估计量变量名；这两个变量的顺序是不能改变的，对于这两个估计量的估计在下面中会详细介绍。options内容如下表所示：

| Main | |
|---|---|
| constant | 计算检验统计量时加入常数项，默认值是排除常数项 |
| alleps | 使用所有方程进行检验，默认是只检验第一个方程 |
| skipeps(eqlist) | 检验时不包括 eqlist;此方程只能是方程名称不能是序号 |
| equations（matchlist） | 比较设定的方程 |
| Force | 即使假设条件不满足，仍进行检查 |
| df(#) | 使用#自由度。默认使用一致估计与有效估计的协方差矩阵的秩 |
| sigmamore | 协方差矩阵采用有效估计量的协方差矩阵 |
| sigmaless | 协方差矩阵采用一致估计量的协方差矩阵 |
| Advanced | |
| tconsistent(string) | 一致估计量栏的标题 |
| tefficient(string) | 有效估计量栏的标题 |

- 豪斯曼检验在检验一个模型是否存在内生性时的具体操作下面进行介绍。
- reg y x1 x2
- estimates store ols
- 这两个命令在对模型进行回归之后，存储OLS的估计结果为估计的有效估计量。
- ivregress 2sls y x1 (x2=z1 z2)
- 假设怀疑x2为内生解释变量，找到x2的工具变量进行2sls回归估计。
- estimates store iv
- 此命令存储2SLS估计的的结果为估计的一致估计量。
- hausman iv ols[,options]
- 此命令是根据以上的存储结果进行豪斯曼检验，然后根据得到的结果图进行判断。
- 常用检验语句是hausman iv ols，constant sigmamore

其中，“sigmamore”表示统一使用更有效的估计量（即 OLS）所对应的残差来计算$\hat{\sigma}^2$。这样有助于保证根据样本数据计算的 $\mathrm{Var}(\hat{\beta}_{IV})-\mathrm{Var}(\hat{\beta}_{OLS})$为正定矩阵。选择项“constant”表示 $\hat{\beta}_{IV}$ 与$\hat{\beta}_{OLS}$都包括常数项（默认值不包含常数项）。

- 本实验中，打开数据文件后，根据上面豪斯曼检验的操作步骤在命令窗口中输入如下命令：
- **reg lw80 s80 iq expr80 tenure80**
- **estimates store ols**
- **ivregress 2sls lw80 expr80 tenure80  s80 (iq=med kww mrt age)**
- **estimates store iv**
- **hausman iv ols,constant sigmamore**
- 从结果图可以看到豪斯曼检验的原假设是所有解释变量都是外生的，Stata检验结果显然表明模型以p=0的概率拒绝原假设，说明解释变量s80 iq为内生解释变量。