小白学统计|面板数据分析与Stata应用笔记（六）

本期内容： 内生性与工具变量法

> 面板数据分析与Stata应用笔记整理自慕课上浙江大学方红生教授的面板数据分析与Stata应用课程，笔记中部分图片来自课程截图。
>
> 笔记内容还参考了陈强教授的《高级计量经济学及Stata应用（第二版）》

一篇高质量的经验研究论文需要高度重视内生性问题的处理

内生性问题及解决方法

短面板和长面板数据的分析技术都只考虑了一种内生性问题，即不可观测的个体效应与解释变量相关，对于这一问题，我们只需要对个体效应进行控制即可解决。

而一种普遍的内生性问题——解释变量与误差项存在相关性普遍存在于计量方程中，如果这个问题不加以重视和处理那么回归结果就变的不可相信。

内生性问题的来源

内生性问题主要来自于三个方面，分别为：遗漏变量、联立性以及度量误差

1、遗漏变量

遗漏变量是指可能与解释变量相关的变量，本来应该加以控制，但却没有控制的变量。这些变量最后进入了误差项，从而导致误差项与解释变量相关，进而导致了内生性问题。

2、联立性

联立性是指一个计量方程中的核心解释变量A对被解释变量B产生影响，反过来，被解释变量B又对A产生影响。

如果B对A有正向的影响，正向冲击就会导致A增加，从而导致核心解释变量A与误差项正相关。

如果B对A有负向的影响，正向冲击就会导致A降低，从而导致核心解释变量A与误差项负相关。

3、度量误差

度量误差可以分为解释变量的度量误差和被解释变量的度量误差。

（1）解释变量存在度量误差

假设真实的模型为：

$$y = \alpha + \beta x^* + \varepsilon \quad （1）$$

其中，$Cov(x^*, \varepsilon) = 0$，而 x^* 无法精确观测，我们只能观测到x，二者之间的关系是：

$$x = x^* + \mu \quad （2）$$

且满足 $Cov(x^*, \mu) = 0$， $Cov(\mu, \varepsilon) = 0$。

将上述（1）（2）两式合并可以得到：

$$y = \alpha + \beta x + (\varepsilon - \beta\mu) \quad （3）$$

此时，因为 μ 与x相关，所以能够观测到的解释变量x与新的误差项 $(\varepsilon - \beta\mu)$ 存在相关关系，从而产生内生性，在这种情况下，估计得到的系数绝对值会偏小。

（2）被解释变量存在度量误差

假设真实的模型为：

$$y^* = \alpha + \beta x + \varepsilon \quad （4）$$

其中，$Cov(x, \varepsilon) = 0$，而 y^* 无法精确观测，只能观测到y，二者之间的关系是：

$$y = y^* + v \quad （5）$$

其中，v 为测量误差，两式合并可以得到：

$$y = \alpha + \beta x + (\varepsilon + v) \quad （6）$$

此时，如果 $Cov(x, v) = 0$，则OLS估计量仍是一致的，但估计结果可能会增大扰动项的方差，若 $Cov(x, v) \neq 0$，就会产生内生性问题。

内生性会带来的问题

在存在内生性解释变量的情况下，OLS估计量有偏且不一致。

只要任何一个解释变量与随机扰动项相关，全部解释变量的系数都会有偏、不一致。

解决内生性问题的方法

通常有两种方法解决内生性问题即使用内生变量的滞后一期和工具变量法。

1、使用内生变量的滞后一期

一般来说，内生变量的上一期与当期误差项并不存在相关关系，所以可以考虑使用内生变量的滞后一期替代当期的内生变量。这种方法较为简单，并且在直觉上可行，但这种方法的缺点是：不能够回答当期的内生变量对当期的被解释变量的影响程度；而且，上一期的内生变量也可能因为遗漏变量而具有内生性。

2、工具变量法

工具变量是指某一个变量与模型中解释变量高度相关，但却不与误差项相关，估计过程中被作为工具使用，以替代模型中与误差项相关的解释变量的变量。

工具变量法则是使用工具变量进行估计的方法。

工具变量法最常用的估计方法为：两阶段最小二乘法（TSLS）。

*** 两阶段最小二乘法

两阶段最小二乘法，顾名思义，是指分两阶段进行最小二乘估计。

> 第一阶段：将内生性变量作为被解释变量，工具变量和方程中的外生变量作为解释变量，来进行最小二乘估计；
> 第二阶段：用第一阶段估计得到的内生变量的预测值替换内生变量，再进行最小二乘估计。

*** 两阶段最小二乘法的原理

> 第一阶段：消除了潜在内生解释变量的内生性，通过外生变量的预测回归，得到这些变量的外生性部分。
> 第二阶段：利用第一阶段得到外生的预测回归的拟合值进行回归，进而消除偏误。

需要注意的是：

工具变量法估计的标准误始终大于OLS估计的标准误，工具变量与内生性解释变量的相关性越强，工具变量法估计的标准误就越低，估计精度就越高。

工具变量法对工具变量的选择有着非常高的要求，好的工具变量会使得结果更加精确，而不当的工具变量会使得工具变量法的估计结果比最小二乘法的估计结果更加糟糕。如果实在难以找到一个好的工具变量，那么选择OLS估计也是一个不错的结果，但是需要在文章解释OLS估计是高估还是低估了结果，这有利于对问题的分析。如果内生变量与误差项是正相关则是高估了结果，如果是负相关，则是低估了结果。

工具变量法的检验

使用工具变量法进行估计时，我们需要对工具变量进行三项检验，分别为：内生性检验、相关性检验、外生性检验。

1、内生性检验

内生性检验即检验核心变量是否具有内生性。如果我们关心的核心解释变量不具有内生性，我们就没有必要使用工具变量法进行估计，而如果我们使用了工具变量法虽然得到了一致估计量，但并不是有效估计量。

2、相关性检验

相关性检验是检验工具变量是否与内生变量之间存在强相关关系。如果使用的工具变量是弱工具变量，则会导致内生变量估计的标准系数偏大。

3、外生性检验

外生性检验是检验工具变量是否与误差项不相关。如果工具变量与误差项相关，则不满足外生性条件，那么使用工具变量法估计很可能会比OLS估计的结果更糟糕。

工具变量法的检验方法

> ***
> 对工具变量的三大检验，一般来说，我们应该先做相关性检验，因为，如果存在弱工具变量，则两阶段最小二乘法的估计结果会比OLS的估计结果更加糟糕。此外，弱工具变量会使内生性检验的Hausman test和外生性检验的Hansen's J的结果产生偏差。
> ***

1、相关性检验方法

相关性检验是通过构造辅助回归来对工具变量与内生变量之间的相关性进行检验。构造的辅助回归即为两阶段最小二乘法的第一阶段，用所有的外生解释变量（包括工具变量）对潜在的内生解释变量做OLS回归。

对辅助回归的结果，我们首先观察工具变量的系数符号是否符合理论预期，其次观察F值是否大于10，如果大于10则表明工具变量与内生解释变量存在强相关

以Acemoglu等人2001年的论文举例说明，文章中的变量如下所示。

举例:Acemoglu(2001) "The Colonial Origins of Comparative Development: An Empirical Investigation"

模型：$logpgp95 = \alpha\, avexpr + \beta\, lat\_abst + u$

被解释变量	logpgp95	log PPP GDP pc in 1995, World Bank
内生解释变量	avexpr	average protection against expropriation risk
外生解释变量	lat_abst	Abs(latitude of capital)/90
工具变量	logem4	log settler mortality

由于制度与人均GDP之间存在着双向因果关系，所以制度是内生变量。为了解决制度变量的内生性问题，Acemoglu等人选择的工具变量是殖民地的死亡率。这是因为，殖民地的死亡率越高，殖民者就越不倾向于移植好的制度，而更倾向于采取掠夺性的制度安排，而那时的制度对现在仍然有影响。

使用Acemoglu等人的截面数据做第一阶段OLS回归，结果如下所示。

```
reg avexpr logem4 lat_abst
```

```
. regress  avexpr lat_abst logem4
```

F>10?

Source	SS	df	MS		
Model	40.2217245	2	20.1108623	Number of obs =	64
Residual	95.6645164	61	1.56827076	F(2, 61) =	12.82
Total	135.886241	63	2.15692446	Prob > F =	0.0000
				R-squared =	0.2960
				Adj R-squared =	0.2729
				Root MSE =	1.2523

avexpr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lat_abst	2.001775	1.337176	1.50	0.140	-.6720747	4.675624
logem4	-.5102681	.1410186	-3.62	0.001	-.7922521	-.228284
_cons	8.529432	.8123128	10.50	0.000	6.905113	10.15375

由检验结果可知，F值等于12.82，大于10，表明工具变量与内生性变量之间存在着强相关关系。此外，工具变量logem4对内生性变量avexpr产生了显著的负向影响，这符合工具变量背后的逻辑。

** 我们也可以使用如下命令进行相关性检验。

```
ivreg logpgp95 last_abst(avexpr=logem4),first
```

ivreg命令是在括号中"="的左边放入内生性解释变量"="的右边放入工具变量，选项[first]表示既报告第一阶段的回归结果也报告第二阶段的回归结果。

上面的两种检验工具变量与内生变量相关性的方法适合于方程中存在一个内生性解释变量。那么，如果方程中有多个内生性解释变量，我们又该如何做先关性检验呢。

Stock/Yogo给出了检验规则：

如果弱识别检验的最小特征值统计量大于Stock/Yogo的15% maximal IV size所对应的临界值，我们就可以认为工具变量不存在弱相关问题。

对多个内生性变量进行相关性检验的命令为[ivreg2]或者[xtivreg2]

仍以上文的例子为例，对内生性解释变量使用命令[ivreg2]进行检验，结果如下所示。

```
ivreg2 logpgp95 last_abst(avexpr=logem4)
```

```
Weak identification test (Cragg-Donald Wald F statistic):                13.093
Stock-Yogo weak ID test critical values: 10% maximal IV size             16.38
                                         15% maximal IV size              8.96
                                         20% maximal IV size              6.66
                                         25% maximal IV size              5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
```

由检验结果可知，弱识别检验的最下特征值统计量为13.09，大于所对应的临界值8.96，所以，我们可以认为工具变量不存在弱相关。

如果检验结果显示为弱工具变量，我们又该如何解决呢？

一般有三种办法来解决弱工具变量问题。

- 选择更好的工具变量
- 做冗余检验将弱相关的工具变量剔除掉，冗余检验的原假设是指定的工具变量是多余的，ivreg/xtivreg2提供了选项redundant(varlist)
- 利用有限信息最大似然法（LIML）对弱工具变量不敏感。在大样本下，LIML与两阶段最小二乘估计是渐进等价的，当存在弱工具变量时，LIML的小样本性质可能优于两阶段最小二乘法。具体实现命令为：（ivreg …，liml）或（ivreg2 …，liml）

2、内生性检验方法

内生性检验首先假定模型中存在内生性解释变量进行两阶段最小二乘回归；然后再假定不存在内生性变量进行普通回归；最后使用Hausman检验对内生性问题进行检验。

如果检验结果的P值小于0.1，表明两个回归系数存在显著的系统性差异，意味着关注的核心变量具有内生性。如果P值大于0.1，表明两个回归系数不存在系统性差异，意味着关注的核心变量不存在内生性问题。

继续以Acemoglu等人的数据为例进行内生性检验，Hausman检验程序为：

```
ivreg logpgp95 lat_abst(avexpr=logem4)
est store iv
reg logpgp95 lat_abst avexpr
hausman iv
```

```
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       iv            .       Difference          S.E.
-------------+------------------------------------------------------------
     avexpr  |     .995704     .4678871        .527817        .2121926
   lat_abst  |    -.6472071     1.576884       -2.224091        1.130514
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from ivreg
        B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        6.19
                Prob>chi2 =      0.0453
```

检验结果显示，P值为0.0453，小于0.1，拒绝原假设，表明两个模型的回归系数在10%的显著性水平下存在显著性的差异，认为检验的变量是内生性变量。

内生性检验有一个非常便捷的命令，即[ivreg2,endog]或[xtivreg,endog()]，我们只要在上述命令中将内生变量放入endog选项中即可。我们也可以直接输入命令

```
ivreg2 logpgp95 lat_abst(avexpr=logem4),endog(avexpr)
```

```
Sargan statistic (overidentification test of all instruments):         0.000
                                                (equation exactly identified)
-endog- option:
Endogeneity test of endogenous regressors:                             15.239
                                                  Chi-sq(1) P-val =    0.0001
Regressors tested:   avexpr
```

由检验结果我们可以看到，P值为0.0001，同样表明制度变量是内生性变量。

3、外生性检验

如果我们选择的工具变量的个数恰好等于内生变量的个数，这时是恰好识别的，这种情况公认是无法进行外生性检验的即检验所选择的工具变量是否和误差项不相关，我们只能定性讨论或依赖于专家的意见。

如果是过度识别的情况即工具变量的个数大于内生变量的个数，我们就可以检验所选择的工具变量是否与误差项不相关。对工具变量外生性的检验我们可以使用命令ivreg2或者xtivreg2。检验的原假设是工具变量与误差项不相关。当P值小于0.1时拒绝原假设，说明工具变量与误差项相关，工具变量不具有外生性；当P值大于0.1时，接受原假设，说明工具变量与误差项不相关，工具变量具有外生性。

值得注意的是，ivreg2 和 xtivreg2可以同时做上述三大检验。

此外，在上述的分析中，我们没有考虑模型误差项的异方差和自相关问题，如果误差项存在异方差或自相关时，两阶段最小二乘虽然是一致估计量但却并不是有效估计量，而更为有效的方法是“广义矩估计（GMM）”，GMM方法使用的前提条件是工具变量数大于内生变量数，两阶段最小二乘估计存在异方差或自相关。

GMM的实现命令为：直接在ivreg2或xtivreg2命令之后添加gmm选项即可。