

# 搞定内生性，不可不知的工具变量法笔记

数量经济学 2022-08-31 22:21 发表于陕西

内生性(endogeneity)问题，是指由自变量与误差项相关所引发的估计偏倚及统计结果误导性等问题的总称，即违背了线性回归中的正交假定而产生的一系列问题。内生性问题看似简单，但目前已成为线性回归及其他回归模型中最为棘手的问题。

工具变量法是解决内生性问题的有效方法。

在工具变量估计中，第一，检验是否具有内生性，可以使用豪斯曼检验。第二，工具变量的正交性检验。（1）、强度条件，即工具变量应该与内生自变量具有较强的相关性，即该工具变量的应该能够代替或者表达原内生变量的信息，数学表达式为： $COV(Z, X) \neq 0$

（2）、排除限制条件，即工具变量应该与误差项不相关，也就是与因变量Y中不能被已有的自变量x所表达的部分无关（也是与误差项无关） $COV(Z, u) = 0$ 。

## 工具变量估计

二阶段最小二乘法的第一阶段就是利用原模型的内生解释变量对工具变量进行OLS，得到解释变量的拟合值；第二步，利用得到解释变量的拟合值对原模型进行最小二乘法，从而得到方程模型的估计值，这样就可以消除内生性的影响。

首先了解一下二阶段最小二乘法Stata中的命令为ivregress，语法格式为

```
1 ivregress estimator depvar [varlist1] (varlist2 = varlist_iv) [if] [in]
```

### 选项介绍

estimator分为2sls两阶段最小二乘、liml有限的信息最大似然(liml)、gmm广义矩方法(gmm)

depvar depvar 为被解释变量；

varlist1为外生解释变量；

varlist2 为所有的内生解释变量；

varlist\_iv为所有的工具变量；

在选项 options 中，

vce(robust)表示稳健型标准误

可使用 firstfirst 选项报告 2SLS 中第一阶段的回归结果

small表示小样本下的自由度调整

本文以伍德里奇第十五章数据mroz.dta为例，研究已婚妇女的教育回报，相关数据介绍如下：

```
1 use morz.dta
2 edit
3 desc
4 *被解释变量
5 label var lwage 已婚妇女工资的对数值
6 *解释变量
7 label var educ 受教育年数
8 label var exper 工作年限
9 label var expersq 工作年限平方
10
11 *工具变量
12 label var fatheduc 已婚妇女的父亲的受教育年数
13 label var motheduc 已婚妇女的母亲的受教育年限
```

```
. label var lwage 已婚妇女工资的对数值
.
. *解释变量
. label var educ 受教育年数
.
. label var exper 工作年限
.
. label var expersq 工作年限平方
.
.
. *工具变量
. label var fatheduc 已婚妇女的父亲的受教育年数
.
. label var motheduc 已婚妇女的母亲的受教育年数
.
. desc

Contains data from C:\Users\admin\Desktop\mroz.dta
  obs:      753
 vars:      22              10 Jan 2000 16:54
```

variable name	storage type	display format	value label	variable label
---------------	--------------	----------------	-------------	----------------

variable name	storage type	display format	value label	variable label
inlf	float	%9.0g		
hours	float	%9.0g		
kidslt6	float	%9.0g		
kidsge6	float	%9.0g		
age	float	%9.0g		
educ	float	%9.0g		受教育年数
wage	float	%9.0g		
repwage	float	%9.0g		
hushrs	float	%9.0g		
husage	float	%9.0g		
huseduc	float	%9.0g		
huswage	float	%9.0g		
faminc	float	%9.0g		
mtr	float	%9.0g		
motheduc	float	%9.0g		已婚妇女的母亲的受教育年数
fatheduc	float	%9.0g		已婚妇女的父亲的受教育年数
unem	float	%9.0g		
city	float	%9.0g		
exper	float	%9.0g		工作年限
nwifeinc	float	%9.0g		
lwage	float	%9.0g		已婚妇女工资的对数值
expersq	float	%9.0g		工作年限平方

计量经济学服务中心

其中研究问题为：

建立lnwage与educ、exper、expersq的方程，但是包括了影响已婚妇女工资的遗漏变量，可能存在内生性问题，其中能力会对工资产生影响，但是却与解释变量X中的educ相关，内生性存在。

因此需要寻找与能力相关，但是与误差项不相关的工具变量，认为已婚妇女的父亲和母亲的受教育年数跟已婚妇女的educ相关的，而这两个变量与已婚妇女的能力相关，可以替代原来内生变量的信息。因此，可以作为educ的工具变量。

相关操作代码为：

```

1 *OLS回归与2SLS对比
2
3 reg lwage educ exper expersq
4 est store OLS
5
6 ivregress 2sls lwage exper expersq (educ = motheduc fatheduc)
7 est store _2SLS
8

```


结果为：

```
. reg lwage educ exper expersq
```

Source	SS	df	MS	Number of obs	=	428
Model	35.0223023	3	11.6741008	F(3, 424)	=	26.29
Residual	188.305149	424	.444115917	Prob > F	=	0.0000
Total	223.327451	427	.523015108	R-squared	=	0.1568
				Adj R-squared	=	0.1509
				Root MSE	=	.66642

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.1074896	.0141465	7.60	0.000	.0796837	.1352956
exper	.0415665	.0131752	3.15	0.002	.0156697	.0674633
expersq	-.0008112	.0003932	-2.06	0.040	-.0015841	-.0000382
_cons	-.5220407	.1986321	-2.63	0.009	-.9124668	-.1316145

 计量经济学服务中心

```
. est store OLS
```

```
. ivregress 2sls lwage exper expersq (educ = motheduc fatheduc)
```

Instrumental variables (2SLS) regression	Number of obs	=	428
	Wald chi2(3)	=	24.65
	Prob > chi2	=	0.0000
	R-squared	=	0.1357
	Root MSE	=	.67155

lwage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
educ	.0613966	.0312895	1.96	0.050	.0000704	.1227228
exper	.0441704	.0133696	3.30	0.001	.0179665	.0703742
expersq	-.000899	.0003998	-2.25	0.025	-.0016826	-.0001154
_cons	.0481003	.398453	0.12	0.904	-.7328532	.8290538

```
Instrumented: educ
```

```
Instruments: exper expersq motheduc fatheduc
```

```
. est store _2SLS
```

 计量经济学服务中心

同时展现并对其进行对比，代码为：

```
1 esttab OLS _2SLS ,
2   title("已婚妇女教育投入回报影响研究") replace
3   mtitles("OLS回归" "2SLS回归结果" )
4       b(%6.3f) se
5       star( * 0.10 ** 0.05 *** 0.01 )
6       addnotes("*** 1% ** 5% * 10%") staraux r2 nogap compress
7
8
```

结果为：

### 已婚妇女教育投入回报影响研究

	(1) OLS回归	(2) 2SLS回~果
educ	0.107 (0.014)***	0.061 (0.031)**
exper	0.042 (0.013)***	0.044 (0.013)***
expersq	-0.001 (0.000)**	-0.001 (0.000)**
_cons	-0.522 (0.199)***	0.048 (0.398)

N	428	428
R-sq	0.157	0.136

Standard errors in parentheses

\*\*\* 1% \*\* 5% \* 10%

\* p<0.10, \*\* p<0.05, \*\*\* p<0.01

end of do-file

结果解释：

**EXAMPLE 15.5** Return to Education for Working Women

We estimate equation (15.40) using the data in MROZ. First, we test  $H_0: \pi_3 = 0, \pi_4 = 0$  in (15.41) using an  $F$  test. The result is  $F = 124.76$ , and  $p\text{-value} = .0000$ . As expected, *educ* is (partially) correlated with parents' education.

When we estimate (15.40) by 2SLS, we obtain, in equation form,

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

$$(.400) \quad (.031) \quad (.013) \quad (.0004)$$

$$n = 428, R^2 = .136.$$

The estimated return to education is about 6.1%, compared with an OLS estimate of about 10.8%. Because of its relatively large standard error, the 2SLS estimate is barely statistically significant at the 5% level against a two-sided alternative.

喜欢此内容的人还喜欢

[ArcPy百科]第三节：Geometry信息中的空间参考解析

虾神说D



统计上不显著的变量表明该变量对结果变量没有影响吗？

计量经济圈



常用: 主成分分析和因子分析的原理, 操作, 代码和案例讲解!

计量经济圈

