工具变量

汇报人：李竞开

时间：2020/12/10

汇报内容

- 工具变量运用的检验
- 工具变量使用步骤
- 工具变量运用举例
- 工具变量运用常见问题

工具变量运用的检验

- 是否需要使用工具变量：内生性检验
- 工具变量是否满足相关性：弱工具变量检验
- 工具变量是否是外生的：过度识别检验

工具变量运用的检验

- 是否需要使用工具变量：内生性检验
  - Durbin-Wu-Hausman χ^2_J 检验
  - 回归形式的Wu-Hausman F检验

工具变量运用的检验

- Durbin-Wu-Hausman χ_J^2 检验

  如果可能的内生变量 D_i 有一个合理的工具变量 Z_i：

  - 构造检验统计量如下：

    $$H=\left(\widehat{\beta_1^{2SLS}}-\widehat{\beta_1^{OLS}}\right)'\left[Avar\left(\widehat{\beta_1^{2SLS}}\right)-Avar\left(\widehat{\beta_1^{OLS}}\right)\right]^{-1}\left(\widehat{\beta_1^{2SLS}}-\widehat{\beta_1^{OLS}}\right)\sim\chi_J^2$$

  - 原假设H_0：D_i是外生的
  - 在原假设下，OLS估计量和工具变量估计得到的参数估计是一致的；备择假设下，OLS估计量是有偏的，所以二者差别 $\widehat{\beta_1^{2SLS}}-\widehat{\beta_1^{OLS}}$ 应该较大，H值远远不等于零。故H值较大时，拒绝原假设，D_i是内生的。

工具变量运用的检验

- 回归形式的Wu-Hausman F检验

  考虑简单模型：

$$Y_i = \alpha + \beta_1 D_i + e_i$$
$$D_i = \gamma_0 + \gamma_1 Z_i + u_i$$
$$cov(D_i, e_i) = cov(u_i, e_i)$$

  - 原理：检验D_i的外生性就是检验$cov(u_i, e_i)$是否为零。
  - 操作：把干扰项的线性关系表示为 $e_i = \rho u_i + \tau_i$，代入原估计方程$Y_i = \alpha + \beta_1 D_i + \rho u_i + \tau_i$，使用$\hat{u}_i = D_i - \widehat{\gamma_0} - \widehat{\gamma_1} Z_i$代入估计检验原假设$H_0: \rho = 0$，如果$\rho$不显著等于0，则拒绝$D_i$是外生的假设。

工具变量运用的检验

- 工具变量是否满足相关性：弱工具变量检验
  - 弱工具变量在有限样本和大样本下都对估计方差有很大负面影响，要避免使用弱工具变量。
  - 只有一个内生变量：观察2SLS中第一阶段关于所有工具变量系数同时为0的F检验，如果F值非常低，存在弱工具变量问题。（临界值由一些学者给出）
  - 存在多个内生变量：Stock and Yogo (2005) 提供了一个检验方法。计算一个被称作Minimum eigenvalue的统计量，如果该统计量高于Stock and Yogo (2005) 给出的对应关键值，则不存在弱工具变量的问题。

工具变量运用的检验

- 工具变量是否是外生的：过度识别检验

  本质而言，我们不可能检验这个条件，因为干扰项无法被观测。但可以一定程度上对外生条件进行检验。

  - 恰当识别：无法检验
  - 要得到一致的估计量$\widehat{\beta_1}$的前提假设是工具变量外生。

    $cov(Z_i, \hat{e_i}) = cov\left(Z_i, Y_i - \widehat{\beta_1} D_i\right) = cov(Z_i, Y_i) - \widehat{\beta_1} cov(Z_i, D_i) = cov(Z_i, Y_i) - \frac{cov(Z_i, Y_i)}{cov(Z_i, D_i)} cov(Z_i, D_i) = 0$

    通过$cov(Z_i, \hat{e_i}) = 0$检验工具变量外生性没有意义 。

工具变量运用的检验

- 工具变量是否是外生的：过度识别检验
  - 过度识别：假设存在两个工具变量，先假设第一个变量Z_1是外生的，并且只用Z_1作为工具变量得到系数$\widehat{\beta_1^{Z_1}}$,那么残差$\mathrm{e}_i^{\widehat{Z_1}} = \mathrm{Y_i} - \widehat{\beta_1^{Z_1}} D_i$是一致的。接着使用残差对第二个工具变量检验$cov\left(\mathrm{e}_i^{\widehat{Z_1}}, Z_2\right) = 0$即可检验第二个工具变量的外生性。
  - 必须先假设一个工具变量的外生性再去检验另一个的外生性，对结果的解释也必须谨慎：Z_2不通过外生性检验有可能是因为对Z_1的外生性假设不正确。因此，两个中至少有一个是内生的。Z_2通过外生性检验也可能是因为对Z_1的外生性假设不正确。因此，通过检验时也不能得出所有工具变量都是外生的结论。

工具变量运用的检验

- 工具变量是否是外生的：过度识别检验
  - 实际运用中的操作：假设检验
  - 原假设H_0:所有工具变量都是外生的。
  - 检验统计量：使用所有工具变量用2SLS进行回归得到残差项。将残差项作为被解释变量，所有工具变量作为解释变量OLS回归，得到R^2。检验统计量$NR^2 \sim x_q^2$，如果NR^2大于相关x_q^2关键值，得出结论：不是所用工具变量都是外生的，如果NR^2小于相关x_q^2关键值，通过了过度识别检验（但仍不能确定所有工具变量都是外生的）。
  - x_q^2的取值：Sargan和Basmann的χ_q^2统计值；Wooldridge的稳健得分过度识别检验；Hansen的J统计值。

工具变量的使用步骤

- 清晰地定义研究问题，描述经济机制，设置基本模型，对基本模型进行OLS回归，理解并描述模型可能存在的内生性问题的原因（反向因果，测量误差，遗漏变量等）。
- 根据经济机制和理论基础选择有效的工具变量，并解释工具变量的外生性和相关性。
  - 相关性一般比较容易解释。
  - 最具挑战性、最关键之处：使用描述性语言和经济原理说明工具变量的外生性。
- 使用工具变量估计法对模型进行估计，同时进行必要的统计检验，并谨慎的对结果进行解释。
  - 检验变量的内生性：Hausman检验。
  - 检验工具变量的相关性：报告第一阶段的F统计量，检验是否存在弱工具变量问题。
  - 检验工具变量的外生性：过度识别检验。

工具变量的使用步骤

- 将工具变量估计结果和OLS结果进行比较，理解为何结果存在差异。
- Ps:第三步中的三个检验只是辅助性检验，它们得到的结论是有限的，不能够代替前两步中对处置变量可能内生的原因和工具变量有效性的讨论。

工具变量运用举例

- The Colonial Origins of Comparative Development : An Empirical Investigation
- 研究问题：好的社会制度对经济发展是否有促进作用
- 经济机制：良好的社会制度意味着更好的产权保护和更少的扭曲资源配置的政策，会促进资产和人力资源的投入，并更有效率的产出。

工具变量运用举例

- 基本模型

$$\log \text{ppp}\, GDP = \alpha + \beta_1 avexpr + \beta_2 lat\_abst + e$$

- $\log ppp\ GDP$：按购买力平价计算出的GDP取对数
- $avexpr$：1985~1995年企业免受政府盘剥的指数平均值，该值越大证明制度越好（国家制度）
- lat_abst：首都的纬度（地理位置）

工具变量运用举例

数据展示

```
. use "C:\Users\lenovo.LAPTOP-A3K6LB3T\Desktop\工具变量\maketable3.dta"

. des

Contains data from C:\Users\lenovo.LAPTOP-A3K6LB3T\Desktop\工具变量\maketable3.dta
  obs:           376
 vars:            11                          18 Jan 2010 22:27
                                              (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
lat_abst        float   %9.0g                 Abs(latitude of capital)/90
euro1900        float   %9.0g                 European settlers 1900, AJR
excolony        float   %9.0g                 =1 if was colony FLOPS definiti
avexpr          float   %9.0g                 average protection against expropriation risk
logpgp95        float   %9.0g                 log PPP GDP pc in 1995, World Bank
cons1           float   %9.0g                 cons on exec in 1st year indep
indtime         float   %9.0g                 years independent: 1995 minus firstyr
democ00a        float   %9.0g                 democracy in 1900
cons00a         float   %9.0g                 constraint on executive in 1900
extmort4        float   %9.0g                 corrected mort.
logem4          float   %9.0g                 log settler mortality
-------------------------------------------------------------------------------
Sorted by:
```

工具变量运用举例

OLS回归结果

```
. regress logpgp95 avexpr lat_abst,robust

Linear regression                               Number of obs     =        111
                                                F(2, 108)         =     183.95
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6225
                                                Root MSE          =      .7108
```

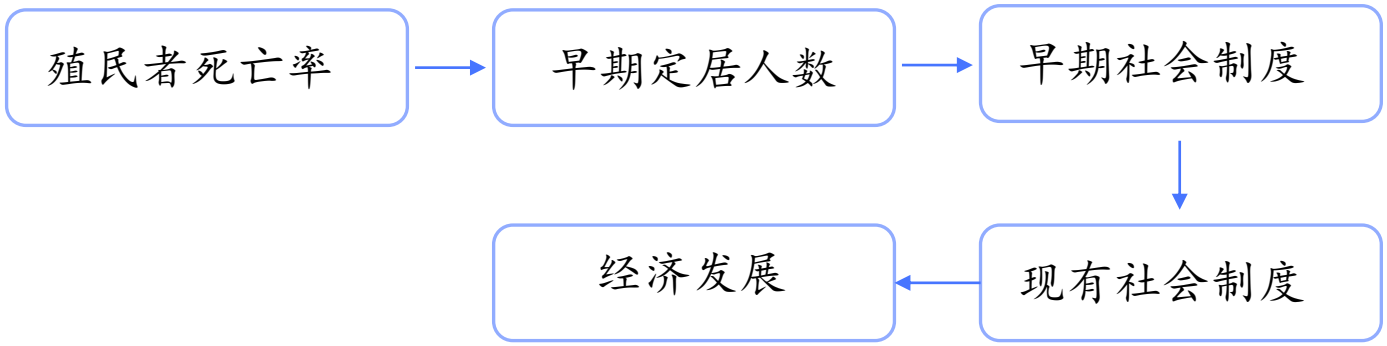
logpgp95	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
avexpr	.4634816	.0521728	8.88	0.000	.360066	.5668972
lat_abst	.8721613	.4993736	1.75	0.084	-.1176837	1.862006
_cons	4.872922	.2807513	17.36	0.000	4.316424	5.42942

工具变量运用举例

- OLS可能存在的问题
  - 发展较好的经济体更有可能建立良好的制度（反向因果）
  - 文化差异等其他变量同时影响国家经济状况和制度（混淆路径）
  - 社会制度不易准确衡量，测量可能存在较大偏差（测量误差）
  - 存在内生性问题：$\mathrm{cov}(avexpr, e) \neq 0$

工具变量运用举例

- 寻找有效的工具变量
  - 相关性：国家现在的社会制度一定程度上是过去制度的延续，早期社会制度又与欧洲殖民者的殖民政策有关。



  - 外生性：殖民者死亡率的外生性最强。

工具变量运用举例

- 使用工具变量法对模型进行估计

```
. ivregress 2sls logpgp lat_abst (avexpr = logem4),first

First-stage regressions
_____________________
```

Number of obs	= 70
F(2, 67)	= 19.53
Prob > F	= 0.0000
R-squared	= 0.3682
Adj R-squared	= 0.3494
Root MSE	= 1.2523

avexpr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lat_abst	3.125466	1.203964	2.60	0.012	.7223438	5.528588
logem4	-.4537058	.1304823	-3.48	0.001	-.7141496	-.1932619
_cons	8.094943	.7590112	10.67	0.000	6.57995	9.609936

工具变量运用举例

- 使用工具变量法对模型进行估计

```
Instrumental variables (2SLS) regression          Number of obs   =         70
                                                  Wald chi2(2)    =      39.18
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.0670
                                                  Root MSE        =     1.0159
------------------------------------------------------------------------------
    logpgp95 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avexpr |    1.029084   .2332977     4.41   0.000      .5718288    1.486339
    lat_abst |   -1.784366   1.494275    -1.19   0.232     -4.713092    1.144359
       _cons |     1.65175   1.322986     1.25   0.212     -.9412546    4.244754
------------------------------------------------------------------------------
Instrumented:  avexpr
Instruments:   lat_abst logem4
```

工具变量运用举例

- 对工具变量进行检验
  - Hausman检验解释变量是否外生

```
. estat endogenous

  Tests of endogeneity
  Ho: variables are exogenous

  Durbin (score) chi2(1)          =  16.4466  (p = 0.0001)
  Wu-Hausman F(1,66)              =  20.2691  (p = 0.0000)

.
```

工具变量运用举例

- 对工具变量进行检验
  - 检验工具变量是否为弱工具变量

```
. estat firststage
```

First-stage regression summary statistics

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	F(1,67)	Prob > F
avexpr	0.3682	0.3494	0.1529	12.0905	0.0009

Minimum eigenvalue statistic = 12.0905

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

	5%	10%	20%	30%
2SLS relative bias	(not available)			

	10%	15%	20%	25%
2SLS Size of nominal 5% Wald test	16.38	8.96	6.66	5.53
LIML Size of nominal 5% Wald test	16.38	8.96	6.66	5.53

工具变量运用举例

- 对工具变量进行检验
  - 过度识别情况下，检验工具变量是否是外生的
  - 检验工具变量有效性的另一个条件是外生性，此时需要过度识别，采用殖民者死亡率和定居人数同时作为工具变量进行估计。

```
. ivregress 2sls logpgp lat_abst (avexpr = logem4 euro1900),first

First-stage regressions
```

```
                                  Number of obs     =         69
                                  F(   3,     65)   =      15.07
                                  Prob > F          =     0.0000
                                  R-squared         =     0.4103
                                  Adj R-squared     =     0.3830
                                  Root MSE          =     1.2271
```

avexpr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lat_abst	1.845078	1.355143	1.36	0.178	-.8613306	4.551486
logem4	-.3197937	.1453475	-2.20	0.031	-.6100726	-.0295148
euro1900	.0160626	.0077541	2.07	0.042	.0005765	.0315487
_cons	7.458333	.8117933	9.19	0.000	5.83707	9.079597

工具变量运用举例

- 对工具变量进行检验
  - 过度识别情况下，检验工具变量是否是外生的

```
Instrumental variables (2SLS) regression          Number of obs   =         69
                                                  Wald chi2(2)    =      47.23
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.1599
                                                  Root MSE        =     .96035
```

logpgp95	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
avexpr	.9735225	.1886416	5.16	0.000	.6037918	1.343253
lat_abst	-1.629427	1.291443	-1.26	0.207	-4.160609	.9017557
_cons	1.975545	1.073346	1.84	0.066	-.1281749	4.079265

```
Instrumented:  avexpr
Instruments:   lat_abst logem4 euro1900
```

工具变量运用举例

- 对工具变量进行检验
  - 过度识别检验

```
. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  .104916  (p = 0.7460)
  Basmann chi2(1)        =  .098985  (p = 0.7531)
```

工具变量运用举例

- 解释工具变量的经济显著性，并将估计结果和OLS结果进行对比，理解结果为何有差异
  - 经济显著性：制度对经济发展水平有较大实质性影响
  - 和OLS结果相比：2SLS的回归系数比OLS的回归系数高。同时，我们知道，前面提到的测量误差会导致向下偏差，逆向关系和缺少变量会导致向上偏差。因此，测量误差可能是造成系数差别的主要原因。

工具变量运用常见问题

- 用计量软件估计工具变量模型，不要自己手动进行两步回归。
- 第一阶段回归的解释变量应包含所有的外生变量。
- 避免使用组均值作为工具变量。
- 避免使用内生变量的滞后项作为工具变量。
- 模型含有二次项的工具变量的用法。
- 模型存在交叉项时工具变量的用法。
- 工具变量估计结果只是局部平均处置效应。
- 工具变量越多越好吗？
- 工具变量是解决内生性的万灵药吗？

工具变量运用常见问题

- 用计量软件估计工具变量模型，不要自己手动进行两步回归

$$Avar\left(\widehat{\beta^{2SLS}}\right) = \widehat{\sigma_e^2}\left(\widehat{\boldsymbol{D}}'\widehat{\boldsymbol{D}}\right)^{-1}$$

$$\widehat{\sigma_e^2} = \frac{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}}{N}$$

$$\hat{\boldsymbol{e}} = \mathbf{Y} - \mathbf{D}'\widehat{\beta^{2SLS}}$$

注意此时计算残差时用到的时内生变量的原值而不是预测值，手动计算容易出$\hat{\boldsymbol{e}} = \mathbf{Y} - \widehat{\boldsymbol{D}}'\widehat{\beta^{2SLS}}$的错误。

工具变量运用常见问题

- 第一阶段回归的解释变量应包含所有的外生变量。
  - 工具变量使用的一个常见的错误是，只使用工具变量作为第一阶段回归的解释变量。
  - 正确的做法是将工具变量和所有的外生的解释变量都包含在第一步回归中（如果解释变量包含个体固定效应和时间固定效应，那么也应该加入第一步回归）
  - 不这么做的后果：第二步的参数估计不具备一致性。

工具变量运用常见问题

- 避免使用组均值作为工具变量
  - 使用组内均值的理由：组内个体特征与组内均值有关（相关性）；组内其他个体的平均或加总特征不直接影响个体（外生性）。
  - 实际上，组内均值作为工具变量，不满足外生性的要求。

工具变量运用常见问题

- 避免使用内生变量的滞后项作为工具变量
  - 使用内生变量的滞后项作为工具变量的要求

$$Y_{i,t} = \alpha + \beta X_{i,t} + e_{i,t}$$

$$cov(X_{i,t}, X_{i,t-1}) \neq 0; cov(X_{i,t-1}, e_{i,t}) = 0$$

  - 以上一方面要求内生变量是序列相关的，另一方面要求干扰项不存在序列相关，存在矛盾。

工具变量运用常见问题

- 模型含有二次项的工具变量的用法
  - 考虑模型$Y = \alpha + \beta_1 X + \beta_2 X^2 + e$，$X$有工具变量$Z$
  - 正确做法是：分别使用Z和Z^2作为工具变量，使用计量软件估计模型。

工具变量运用常见问题

- 模型存在交叉项时工具变量的用法
  - 考虑模型$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e$，$X_1$有工具变量$Z_1$，$X_2$外生
  - 正确做法：Z_1作为X_1的工具变量，$Z_1 X_2$作为$X_1 X_2$的工具变量。

工具变量运用常见问题

- 工具变量估计结果只是局部平均处置效应
  - OLS参数估计值使用了解释变量所有的信息去估计被解释变量的变化，因此OLS估计得到的系数是平均处置效应：解释变量变动一个单位，引起的被解释变量均值的变化。
  - 工具变量估计系数是使用了与工具变量相关的信息去估计被解释变量的变化，得到的系数描述的是对于那些其特征X会收到工具变量Z影响的个体，X变化一个单位引起的他们的Y均值的变化，即局部平均处置效应，局部指的是一部分个体。
  - 工具变量的估计结果是局部平均处置效应也意味着选取不同的工具变量会得到不同的估计结果。

工具变量运用常见问题

- 工具变量越多越好吗？
  - 找到更多的好的工具变量可以提高有效性，使用差的工具变量会放大估计偏差。
  - 使用多工具变量权衡有效性和偏差性。
  - 避免添加弱工具变量和同质工具变量。
  - 比较不同工具变量的估计结果来决定是否添加某个工具变量。

工具变量运用常见问题

- 工具变量是解决内生性的万灵药吗？
  - 好的工具变量可以有效地解决内生性问题。
  - 外生性无法通过统计检验确认。
  - 实际运用中外生性和相关性很难兼得：一个较为明确的外生工具变量通常相关性很低，一个相关性较高的工具变量又不满足外生性。