计量经济学

工具变量估计

张晨峰

2016年5月11日

华东理工大学商学院

4. 工具变量估计

主要内容

- 代理变量
- 回归模型的工具变量估计
- 两阶段最小二乘法（2SLS）
- 内生性检验和过度识别约束检验

4.1 代理变量

遗漏变量问题

- 代理变量
- 遗漏变量不随时间而变化，应用固定效应或差分
- 工具变量估计

4.1 代理变量

对无法观测的解释变量使用代理变量

方程中因遗漏变量而导致的偏误，解决的可能方式之一是找到遗漏变量的一个代理变量(proxy variable)。

工资方程

一个人的工资水平与他的可测教育水平及其他非观测因素的关系为

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + \mu$$

4.1 代理变量

遗漏变量的植入解

假设解释变量x_3^*观测不到，但我们有x_3^*的一个代理变量x_3。首先它要符合基本要求，即它应该和x_3^*有某种关系。

$$x_3^* = \delta_0 + \delta_1 x_3 + \nu_3$$

对u和ν_3的假定

- 误差u与x_1,x_2,x_3^*都不相关，且u与x_3也不相关。
- 误差ν_3与x_1,x_2,x_3都不相关，即$E(x_3^*|x_1, x_2, x_3) = E(x_3^*|x_3)$

于是，得到

$$y = (\beta 0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + \mu + \beta_3\nu_3$$

4.1 代理变量

TABLE 9.2 Dependent Variable: log(*wage*)

Independent Variables	(1)	(2)	(3)
educ	.065 (.006)	.054 (.007)	.018 (.041)
exper	.014 (.003)	.014 (.003)	.014 (.003)
tenure	.012 (.002)	.011 (.002)	.011 (.002)
married	.199 (.039)	.200 (.039)	.201 (.039)
south	−.091 (.026)	−.080 (.026)	−.080 (.026)
urban	.184 (.027)	.182 (.027)	.184 (.027)
black	−.188 (.038)	−.143 (.039)	−.147 (.040)
IQ	—	.0036 (.0010)	−.0009 (.0052)
educ·IQ	—	—	.00034 (.00038)
intercept	5.395 (.113)	5.176 (.128)	5.648 (.546)
Observations *R*-squared	935 .253	935 .263	935 .263

4.2 回归模型的工具变量估计

工具变量基本假定

假设有一个可观测的变量z，它满足两个假定

- 工具外生性：$Cov(z, \mu) = 0$
- 工具相关性：$Cov(z, x) \neq 0$

简单线性回归的工具变量估计量

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

当满足工具变量的两个假定时，β_1的IV估计量具有一致性

$$plim(\hat{\beta}_1) = \beta_1$$

用IV估计法做统计推断

在大样本的情况下，IV估计量近似服从正态分布。IV估计量和OLS估计量的渐近方差分别为

$$\hat{\beta}_{1,IV} = \frac{\sigma^2}{SST_x R^2_{x,z}}$$

$$\hat{\beta}_{1,OLS} = \frac{\sigma^2}{SST_x}$$

当x与μ不相关时进行IV估计的一个重要代价：IV估计量的渐近方差总是大于(有时远大于)OLS估计量的渐近方差。

低劣工具变量条件下IV的性质

z和x之间的弱相关可能产生严重的后果

$$plim\hat{\beta}_{1,IV} = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)}\frac{\sigma_u}{\sigma_x}$$

$$plim\hat{\beta}_{1,OLS} = \beta_1 + Corr(x,u)\frac{\sigma_u}{\sigma_x}$$

实践中特别有意思的所谓弱工具(weak instruments)，就被大致定义为z和x之间的相关度”很低“(但不为零)的问题。

多元回归模型的IV估计

考虑两个解释变量的标准线性模型

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \mu_1$$

我们称之为结构方程。如果用OLS估计，所有的估计量将是有偏且又不一致的。因此，我们寻找一个y_2的工具z_2，则关键假定为

- $E(\mu_1) = 0$
- $Cov(z_1, \mu_1) = E(z_1\mu_1) = 0$
- $Cov(z_2, \mu_1) = E(z_2\mu_1) = 0$

我们需要工具变量z_2和y_2偏相关，即回归模型

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \nu_2$$

此为约简型方程的一个例子。关键的识别条件是$\pi_2 \neq 0$

4.3 两阶段最小二乘法（2SLS）

单个内生解释变量

现在假定有两个被排斥在模型之外的外生变量：z_2和z_3。z_2和z_3不出现在模型中，且与误差项u_1不相关的假定被称为**排除性约束**(exclusion restrictions)。

为寻找最好的IV，我们选择与y_2最高度相关的线性组合。

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2$$

那么，y_2最好的IV是上式中z_j的线性组合，我们称之为y_2^*：

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$$

关键识别假定为：$\pi_2 \neq 0$或$\pi_3 \neq 0$

4.3 两阶段最小二乘法(2SLS)

单个内生解释变量

我们将y_2对z_1,z_2,z_3回归，获得拟合值$\hat{y}_2$。我们就可以用它作为y_2的IV。利用三个假定：

- $E(\mu_1) = 0$
- $Cov(z_1, \mu_1) = E(z_1\mu_1) = 0$
- $Cov(y_2^*, \mu_1) = E(y_2^*\mu_1) = 0$

求解三个正规方程，得到IV估计量。在多重工具条件下，IV估计量也叫作两阶段最小二乘(2SLS)估计量。原因很简单，当我们用$\hat{y}_2$作为y_2的IV时，IV估计值$\hat{\beta}_0$，$\hat{\beta}_1$和$\hat{\beta}_2$等同于从

$$y_1\text{对}\hat{y}_2\text{和}z_1$$

的回归中得到的OLS估计值。

多重共线性与2SLS

β_1的2SLS估计量的(渐近)方差近似地写为：

$$\frac{\sigma^2}{S\hat{S}T_2(1-\hat{R}_2^2)}$$

其中，$\sigma^2 = Var(\mu_1)$，$S\hat{S}T_2$是$\hat{y}_2$的总波动，$\hat{R}_2^2$是将$\hat{y}_2$对其他所有出现在结构方程中的外生变量做回归得到的R^2。2SLS的方差大于OLS的方差的原因有二。

- 根据构造，$\hat{y}_2$比y_2的波动性更小。
- $\hat{y}_2$与方程中外生变量之间的相关往往比y_2与这些变量之间的相关大得多。

4.3 两阶段最小二乘法(2SLS)

多个内生解释变量与方程识别阶条件

一般地，当我们在回归模型中有不止一个内生解释变量是，有可能遇到无法识别的问题。但是，我们可以容易地表述识别的一个必要条件，叫做阶条件。**方程识别的阶条件：**我们需要被排斥的外生变量至少与结构方程中包括的内生解释变量一样多。

4.4 内生性检验和过度识别约束检验

内生性检验

在解释变量外生时，2SLS估计量的有效性不如OLS。因此，有必要检验解释变量的内生性。假定我们仅有一个疑似内生变量

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \mu_1$$

豪斯曼(Hausman) 建议直接比较OLS和2SLS估计值，判断其差异是否在统计上显著。

为了判断两者是否显著不同，利用回归来检验似乎更加方便。此处为

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + \nu_2$$

因此，y_2与μ_1不相关的重要条件是ν_2与μ_1不相关。可以写成$\mu_1 = \delta_1 \nu_2 + e_1$。因此，我们用OLS估计

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \nu_2 + e_1$$

并用t统计量检验：H_0：$\delta_1 = 0$

4.4 内生性检验和过度识别约束检验

检验单个解释变量的内生性

1. 通过将y_2对所有外生变量回归，得到残差$\hat{v}_2$。
2. 在结构方程中添加$\hat{v}_2$，并用一个OLS回归检验$\hat{v}_2$的显著性。若其系数统计上显著异于零，我们便判断y_2确实是内生的。

过度识别约束检验

如果我们所拥有的工具多于得到一致估计结果所需要的工具数量，就能有效地检验它们中的一部分是否与误差不相关。其思想是，如果所有的工具都是外生的，那么，除了抽样误差外，**2SLS残差与工具应该不相关，且与这些工具的线性组合都不相关。**

过度识别约束检验

1. 用2SLS估计结构方程，获得2SLS残差$\hat{\mu}_1$。
2. 将$\hat{\mu}_1$对所有外生变量回归，得到R^2，即R_1^2。
3. 在所有的IV都与μ_1不相关的原假设下，$nR_1^2 \sim \chi_q^2$，其中q是模型之外的工具变量减去内生解释变量的总数目。

4.4 内生性检验和过度识别约束检验

EXAMPLE 15.8 **RETURN TO EDUCATION FOR WORKING WOMEN**

When we use *motheduc* and *fatheduc* as IVs for *educ* in (15.40), we have a single overidentifying restriction. Regressing the 2SLS residuals $\hat{u}_1$ on *exper*, $exper^2$, *motheduc*, and *fatheduc* produces $R_1^2 = .0009$. Therefore, $nR_1^2 = 428(.0009) = .3852$, which is a very small value in a χ_1^2 distribution (p-value = .535). Therefore, the parents' education variables pass the overidentification test. When we add husband's education to the IV list, we get two overidentifying restrictions, and $nR_1^2 = 1.11$ (p-value = .574). Subject to the preceding cautions, it seems reasonable to add *huseduc* to the IV list, as this reduces the standard error of the 2SLS estimate: the 2SLS estimate on *educ* using all three instruments is .080 (se = .022), so this makes *educ* much more significant than when *huseduc* is not used as an IV ($\hat{\beta}_{educ} = .061$, se = .031).