第七讲 工具变量法

许文立*[1,2]

[1]安徽大学经济学院
[2]安徽生态与经济发展研究中心

January 12, 2018

正如第五讲所述，一项经验研究可能存在的主要问题：遗漏变量偏误、变量测量误差、反向因果、模型设定错误、样本选择偏误、异方差和序列相关。第四讲和第六讲分别呈现了消除某种遗漏变量偏误的方法－－多元回归和面板数据模型。多元回归应对遗漏变量数据可用情形；面板数据模型引入个体固定效应和时间固定效应来消除遗漏变量数据不可用，且截面或时间单一维度变化时的遗漏变量偏误。

一来，上述两种解决方法均相当于在回归模型中增加核心解释变量(X_{it}) 以外的自变量($Z_{it}, \alpha_i, \gamma_t$)；二来，变量测量误差和反向因果所引起的问题，并不能由多元回归和面板数据模型直接解决。那么，除此之外，还有没有其他方法来解决遗漏变量偏误、变量测量误差、反向因果问题呢？

工具变量(Instrumental variables,IV)回归就是获得X与u相关的总体回归函数未知系数一致估计量的一种常用方法。**IV的核心思想：把核心解释变量X 的变动分解成两个部分：一个部分与误差项u相关，另一个部分与误差项u不相关。如果我们有资料、数据、信息来分离出第二部分，那么，我们**

*E-mail：xuweny87@163.com。

就可以只关注于误差项无关的第二部分，丢弃引起OLS估计偏误的第一部分。这些表征X变动，且与u不相关的数据信息可能来源于一个或多个其他变量，这些变量就称为**工具变量(IV)**。如字面意思，这些变量被作为工具来分离出X中与u无关的部分，从而确保回归系数估计量具有一致性。

下面，详细阐述工具变量法的作用原理及其应用。

1 一元回归与单工具变量

如第五讲所述，引起X与u相关的原因可能是遗漏变量、变量误差、反向因果。无论什么原因，如果我们有一个有效的工具变量I，那么，我们就可以利用工具变量估计量来估计出X对Y的效应。总回归模型为

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{1}$$

如果$corr(X_i, u_i) \neq 0$，OLS估计量就是非一致的。我们就可以用工具变量I_i来分离出X_i中与u_i无关的部分。

依据是否与误差项u相关，可以把变量划分成**内生变量**和**外生变量**，前者与误差项相关，后者与误差项无关。这两个名字可以追溯至多方程模型，即“内生变量”由模型内决定，“外生变量”由模型外决定。这两个变量在后面的DSGE模型中还会见到。

有效的工具变量必须同时满足两个条件：

（1）工具变量相关性条件：$corr(I_i, X_i) \neq 0$

（2）工具变量外生性条件：$corr(I_i, u_i) = 0$

工具变量相关性表明一个工具变量的变动与X的变动相关。工具变量外生性

表明工具变量抓住了X中外生变化的部分。这两个条件对于工具变量回归非常重要。

1.1 两阶段最小二乘TSLS

两阶段最小二乘(TSLS)是用来估计IV估计量的方法。正如该方法的名字所述，IV估计量是通过两个阶段计算出来的。

第一阶段，把X分解成两个部分：可能与误差项相关的部分和与误差项无关的第二部分。

具体来说，第一阶段用X对工具变量I回归：

$$X_i = \pi_0 + \pi_1 I_i + v_i \tag{2}$$

公式（2）就把X分解成：v_i和$\pi_0 + \pi_1 I_i$。由于I_i是外生的，因此，$\pi_0 + \pi_1 I_i$与u_i无关，而剩下的v_i与u_i相关。据此，我们可以用样本数据估计出$\hat{\pi}_0, \hat{\pi}_1$，然后从OLS回归中得到X的预测值$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 I_i$。

第二阶段，利用与误差项无关的第二部分估计β_1。也就是说，Y_i对$\hat{X}_i$回归，利用OLS估计系数，进而得到TSLS估计量，$\hat{\beta}_0^{TSLS}$!$\hat{\beta}_1^{TSLS}$

I. 香烟需求

下面，我们使用美国48个州1985-1995年的香烟销售相关数据。我们用这些数据来估计香烟的需求弹性。工具变量——销售税：来自于一般销售税中对烟草征收的税。香烟消费量，Q_i^{cig}，是第i 个州人均消费香烟包数。价格，P_i^{cig}，含税实际平均价格。

在进行TSLS之前，大家首先要关注工具变量是否有效，也就说，我们选择

的工具变量是否满足前面的两个条件——相关性和外生性。后面，我们会详细给出如何通过统计工具来检验工具变量的有效性。在此之前，我们先来看看这销售税是否能作为一个有效的工具变量。

首先，工具变量相关性。销售税越高，香烟的税后价格也会越高，那么，销售税与香烟价格具有相关性；

其次，工具变量外生性。一般来说，销售税在各州之间是不同的，但是这种差异并不主要由香烟需求决定，而是由于政治考虑。因此，我们也可以认为销售税是外生的。

根据上述TSLS的两个阶段，用1995年的48个州的数据，我们先来看看第一阶段回归：

$$ln(\tilde{P}_i^{cig}) = 4.62 + 0.31SalesTax_i \tag{3}$$

回归结果均在1%下显著，而且如经济理论预测的，销售税越高，税后价格就越高。回归方程的$R^2 = 0.47$，也就睡说，销售税变化解释了47%的香烟价格变动。

在第二阶段中，用$ln(Q_i^{cig})$对$ln(\tilde{P}_i^{cig})$来进行回归。回归结果是

$$ln(\tilde{Q}_i^{cig}) = 9.72 - 1.08ln(\tilde{P}_i^{cig}) \tag{4}$$

也就是说，第一阶段的预测值$ln(\tilde{P}_i^{cig})$，被用作第二阶段的回归量。但是，软件中输出的结果是$ln(P_i^{cig})$，而不是$ln(\tilde{P}_i^{cig})$。因此，TSLS估计为

$$ln(\tilde{Q}_i^{cig}) = 9.72 - 1.08ln(P_i^{cig}) \tag{5}$$

TSLS的估计结果显示，香烟的需求弹性是富有弹性的：价格提高1%，香烟需求量下降1.08%。

从第四、五、六讲我们可以知道，上述估计结果可能存在遗漏变量偏误。

2 IV回归

在一般化的IV回归模型中，主要包括四种类型的变量：被解释变量Y；内生解释变量X；外生解释变量W；工具变量I。一般来说，可能存在多个内生解释变量，多个外生解释变量和多个工具变量。

回归系数是**恰好识别**，如果工具变量与内生解释变量一样多；

回归系数是**过度识别**，如果工具变量比内生解释变量还多；

回归系数是**识别不足**，如果工具变量比内生解释变量还少；

需要注意的是：**IV回归中，至少要有与内生回归因子（内生解释变量）一样多的工具变量。**

在IV中包含外生变量或控制变量是为了确保工具变量与误差项不相关。

【小贴士】（以下内容来源于“SociologyOfDrink”微信公众号2018-01-05期张友浪”控制变量是否越多越好）控制变量一般没有因果解释，但控制变量又很重要，那么，怎么选择控制变量呢？一般来说我们只需控制能够同时影响解释变量和被解释变量的变量（confounder）。但是，我们在投稿时经常会收到审稿人的意见，说这个没有控制那个也没有控制。或者自己有意无意的在回归中包含了过多的控制变量。而控制变量过多往往会造成模型损失自由度，模型不够简洁，模型过度拟合，甚至会让我们得出错误的结论。这些问题在小样本回归中尤为严重。因此，我们在回归时，真正要考虑的是，什么样的变量

不需要被控制？

根据某一备选变量在因果关系中的位置，可以分以下几种情况讨论：

（1）既不影响解释变量，也不影响被解释变量。很多社会经济变量不仅仅因为常见，而被人们要求加入模型。但如果没有可信的理论来支撑对解释变量和被解释变量的影响的话，不应将其纳入模型；

（2）只影响解释变量的变量。这类变量与关键解释变量没有直接影响，不影响我们对解释变量影响的估计，当然无需纳入模型中控制；

（3）只影响被解释变量的变量。这类变量与关键解释变量不存在理论上的相关性，不会造成遗漏变量偏误，无需控制；

（4）被解释变量影响，又影响解释变量的变量，即中介变量（mediator）。考虑到我们的关键解释变量对被解释变量的影响往往是通过一个或多个渠道（因果链条可无限细分），这时就要分两种情况做决定：如果该中介变量处在我么假设的因果链条中，那就应该将其去掉，因为加入这个变量会让解释变量的影响从全部影响减弱为直接影响，而间接影响则被中介变量吸收，从而削弱了我们对解释变量整体效应的估计；如果该中介变量并不处在我们假设的因果链条中，那就应该保留，这时对自变量影响的估计就会自动排除竞争性解释的影响，有助于提高估计结果的可信度；

（5）当然，有时候，会遇到审稿人搞不清因果关系，坚持要求加入某个变量来控制。考虑到硬怼审稿人没什么好下场，因此建议，审稿人说什么就是什么，听从他们的意见，控制审稿意见要求的变量。

2.1 TSLS

当回归方程中只有一个内生解释变量X和多个外生变量时，回归方程为

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i \tag{6}$$

如前所述，X可能与误差项相关，W则不与误差项相关。

根据前文TSLS的两个阶段。在**第一阶段**中，用X与所有的外生变量——外生解释变量W和工具变量I进行回归。

$$X_i = \pi_0 + \pi_1 I_{1i} + \cdots + \pi_m I_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i \tag{7}$$

在TSLS的**第二阶段**，先用（7）式估计的$\hat{X}_i$来替代（6）式中的X，然后进行OLS估计。由此，得到的$\beta_0, \beta_1, \cdots$就是TSLS的估计量。

【小贴士】

1、在进行第一阶段回归时，除了工具变量外，所有的外生变量（或控制变量）都需要包含在其中。而第二阶段回归也需要包含这些外生变量。

2、当有多个内生解释变量时，第一阶段的工具变量回归是每个内生解释变量分别与它对应的工具变量进行回归，然后再将所有的估计值放入第二阶段回归中得到TSLS估计量。

2.2 例子：香烟需求

在第一节的香烟需求例子中，我们只估计了单变量。但是这可能会存在遗漏变量偏误，例如，州的收入水平可能既影响香烟需求，也影响销售税。那么，这

还会使得工具变量外生性条件不被满足。因此，下面，我们在回归中包含州的收入水平。TSLS的估计结果为

$$ln(\tilde{Q_i^{cig}}) = 9.43 - 1.14ln(P_i^{cig}) + 0.21ln(Inc_i) \tag{8}$$

上面的回归系数的标准误为0.37，是恰好识别，也就是说单一内生解释变量对应单一工具变量。除了销售税作为工具变量之外，还可以使用州的烟草特种税。因此，这种税是第二种可能的工具变量。烟草特种税（CigTax）提高会增加香烟的价格，因此满足相关性条件，如果它与误差项无关，那么它也满足外生性条件。下面，我们用两个工具变量来重新进行TSLS 估计，估计结果如下：

$$ln(\tilde{Q_i^{cig}}) = 9.89 - 1.28ln(P_i^{cig}) + 0.28ln(Inc_i) \tag{9}$$

上述两个IV的估计系数标准误为0.25，比较（9）式和（8）式的标准误，我们可以看出，（9）的标准误比（8）下降了三分之一左右。这是因为（9）式利用了更多的信息，两个IV解释了更大的价格变动。

那么，上述IV估计结果可信吗？很遗憾，我们不能立马回答上述问题，因为可信度依赖于IV是否有效。因此，IV有效的两个条件——相关性和外生性就必须要被检验。

3 如何检验IV的有效性

我们仍然用美国48个州的1985-1995年的香烟销售数据。我们来估计长期价格弹性，因此用十年的的数据来进行回归，例如，我们用香烟销售量的

对数之差$ln(Q^{cig}_{i,1995}) - ln(Q^{cig}_{i,1985})$,价格的对数之差$ln(P^{cig}_{i,1995}) - ln(P^{cig}_{i,1985})$和收入的对数之差$ln(Inc^{cig}_{i,1995}) - ln(Inc^{cig}_{i,1985})$。两个工具变量是$SalsTax_{i,1995} - SalsTax_{i,1985}$，$CigTax_{i,1995} - CigTax_{i,1985}$。

结果呈现在图1中，每一列都是不同的回归，都是用TSLS估计量，唯一的差别是工具变量不同。第一列是只包含销售税这个工具变量；第二列是只包含烟草税这个工具变量；第三列是包含两个工具变量。从图1中的结果可以看出，三列结果的第一行均在5%水平下显著为负。但这个结果可信吗？这依赖于我们使用的工具变量是否有效。

Dependent variable: $\ln(Q^{cigarettes}_{i,1995}) - \ln(Q^{cigarettes}_{i,1985})$

Regressor	(1)	(2)	(3)
$\ln(P^{cigarettes}_{i,1995}) - \ln(P^{cigarettes}_{i,1985})$	−0.94** (0.21)	−1.34** (0.23)	−1.20** (0.20)
$\ln(Inc_{i,1995}) - \ln(Inc_{i,1985})$	0.53 (0.34)	0.43 (0.30)	0.46 (0.31)
Intercept	−0.12 (0.07)	−0.02 (0.07)	−0.05 (0.06)
Instrumental variable(s)	Sales tax	Cigarette-specific tax	Both sales tax and cigarette-specific tax
First-stage *F*-statistic	33.70	107.20	88.60
Overidentifying restrictions *J*-test and *p*-value	—	—	4.93 (0.026)

These regressions were estimated using data for 48 U.S. states (48 observations on the 10-year differences). The data are described in Appendix 12.1. The *J*-test of overidentifying restrictions is described in Key Concept 12.6 (its *p*-value is given in parentheses), and the first-stage *F*-statistic is described in Key Concept 12.5. Individual coefficients are statistically significant at the *5% significance level or **1% significance level.

Figure 1: 香烟需求

（I）我们首先来看看，**工具变量的相关性**。

工具变量相关性的作用类似于“样本规模的作用”。因为工具变量与内生解释变量越相关，说明IV回归中包含X的信息越多，TSLS估计量越准确，这就类似与样本量越大，估计结果越准确。

工具变量解释X变动的部分较少时，这种工具变量成为**弱工具变量**。在上面

的例子中，如果选取香烟生产企业到州的距离作为工具变量，这可能就是一个弱工具变量。尽管这个距离越远，香烟的销售价格越高，但是香烟较轻，运输成本可能在其价格中并不是主要组成部分，因此，距离的变动很可能只能解释价格变动中的一小部分。因此，生产距离可能就是一个弱工具变量。那么，如何检验一个工具变量是否是弱工具变量？如果是弱工具变量，我们应该怎么处理？

上面已经说过，工具变量的作用类似于大样本的作用。如果存在弱工具变量问题，正态分布就不是TSLS估计量抽样分布的良好近似，那么，TSLS估计量就不再可信。那么，工具变量相关性程度有多大才是一个良好的分布近似呢？这个答案很复杂，但是幸运的是，我们在实践中有一些经验规则可用：

检验弱工具变量的经验规则：当只有一个内生解释变量时，检验弱工具变量的方法就是计算TSLS第一阶段的F统计量。**第一阶段F统计量**为包含在工具变量中的信息提供了一个不错的测量指标：包含的信息越多，F统计量越大。**经验规则是；如果第一阶段F统计量大于10，不存在弱工具变量；如果小于10，可能就是弱工具变量。**

我们从图1中可以看到，三个TSLS的回归结果中，一阶段F统计量分别为33.7,107.2和88.6，因此，我们选择的工具变量不是弱工具变量。

那如果上面的一阶段F统计量小于10，**也就是说存在弱工具变量，我们该怎么办呢？**

如果存在弱工具变量，且有一些工具变量比另一些更弱。那么，就应该舍弃那些最弱的工具变量。当我们放弃一些弱工具变量时，TSLS估计量的标准误可能会变大，但是**请记住“原始标准误没有任何意义”！**

但是，如果系数恰好识别，也就是一个内生解释变量，只有一个工具变量，

且是弱工具变量时，我们就不能舍弃这个弱工具变量了。即使在过度识别时，没有足够的强工具变量来取得识别效果，舍弃弱工具变量也不会有任何帮助。这种情况下，我们可以干两件事：

（1）去寻找其他的更强的工具变量。说起来容易，做起来难！这需要我们对所研究的问题有足够的认识，并且能重写设计和收集相关数据。

（2）仍然使用弱工具变量，但是估计方法不用TSLS，而用其他估计方法，例如有限信息极大似然（LIML）估计量。

（II）**工具变量外生性**

如果有一个内生解释变量，多个工具变量，那么，我们可以计算出多个TSLS估计量（每个工具变量计算一个）。假设有两个工具变量，那么，我们计算的两个TSLS估计量不同。但是如果两个工具变量都是外生的，那么，它们会十分接近。如果我们估计的两个TSLS估计量差异非常大，那么，我们就要非常警觉：要么其中一个工具变量不是外生的，要么两个都不是外生的。在过度识别情形下，**过度识别限制检验（J统计量）**就是在对多个工具变量TSLS估计量进行比较。

总之，在过度识别情形下，我们能计算出多个TSLS估计量，然后比较它们是否接近，即通过stata计算出**J统计量**。如果是精确识别情形，我们就不能比较，实际上，这个时候的J统计量为0。

从图1中可以看出，（1）列和（2）列只有一个工具变量，因此，不存在J统计量。而在第（3）列中，有两个工具变量，是过度识别情形，因此，可以计算J统计量，其结果为4.93，它服从卡方分布，5%的临界值为3.84，因此，它在5%的水平下拒绝两个工具变量都是外生的假设。这是因为两个工具变量的TSLS估计量差异很大。J统计量拒绝原假设意味着第（3）列的估计是基于无

效的IV估计，因为IV外生性条件不满足。那么，这是否就意味着我们估计的三个结果都不行呢？J统计量拒绝原假设只意味着两个工具变量中至少有一个是内生的，那么，我们可以推断：第一，销售税是外生的，烟草税不是，那么，（1）中的结果就是可靠的；第二，烟草税是外生的，销售税不是，那么,（2）中的结果是可靠的；第三，两个都不是外生的，那么，三个结果在统计意义上都可靠。

千万要记住：统计证据并不能告诉我们哪一种可能是正确的，因此这就需要我们根据我们的经验以及经济理论去argue。

4 哪里去寻找有效的工具变量呢？

在实践中，IV回归最难的就是找到有效的工具变量。虽然如此，但是还是有两种指导性的方法：

（I）遵循经济理论来找工具变量。例如，IV回归的发明者P. Wright（1928）通过他对农业市场的理解，使他认识到所寻找的IV不是使需求曲线移动，而是使供给曲线移动，因此，他就想到用农业地区的天气条件作为有效IV。经济理论法最成功的领域就是金融经济学。在这个领域，有一些投资者行为的经济模型通常是非线性的，此时，不能使用IV估计。因此，将IV方法扩展到非线性模型时，这种扩展方法就是广义矩估计（GMM）。但是经济理论太抽象，并不总是能找到一个有效的IV。

（II）构造工具变量。从这种视角出发，我们要去寻找那些引起内生解释变量变化的随机事件，从这些随机事件中剥离出X变动的外生冲击。

5 Stata命令

**********; 上述回归

数据和stata命令可以给我发邮件所要！