Eco 213: Basic Data Analysis and Econometrics
Lecture 9: Instrumental Variables
April 6, 2019

Dr. Garima Malik
Department of Economics
Shiv Nadar University

Outline

- Instrumental Variable
- 2SLS
- Weak Instruments

- In practice, there could be correlation between X and ε.
- Much of econometric work involves studying the process determining the explanators, to see how they might be correlated with ε.

- The ideal X variable has been randomly assigned.
- If X has been randomly assigned, then it contains no information about ε.
- However, true randomization is relatively uncommon.

- Often, an explanator is partially determined in a way that is random, or at least uncorrelated with ε.
- However, the explanator is also influenced by omitted variables, or determined endogenously, or is in some other way correlated with ε.

- Fortunately, econometricians have discovered a method for separating out the random elements of explanators from the elements that may be correlated with ε.
- Unfortunately, this method requires the data to include an **instrumental variable** with certain key properties.

Instrumental Variable

- An **Instrumental Variable** is a variable that is correlated with X but uncorrelated with ε.

- If Z_i is an instrumental variable:
  1. $E(Z_i X_i) \neq 0$
  2. $E(Z_i \varepsilon_i) = 0$

$$\hat{\beta}^{IV} = \frac{\Sigma z_i Y_i}{\Sigma z_i x_i}$$

- What is the probability limit of IV?

$$p\lim(\hat{\beta}_1^{IV}) = \beta_1 \frac{Cov(Z_i, X_i)}{Cov(Z_i, X_i)} + \frac{Cov(Z_i, \varepsilon_i)}{Cov(Z_i, X_i)} = \beta_1$$

If $Cov(Z_i, X_i) = 0$, the denominator equals 0, and the $p\lim$ does not exist.

If $Cov(Z_i, \varepsilon_i) \neq 0$, then $\hat{\beta}_1^{IV}$ is inconsistent.

Asymptotic Variance

- The asymptotic variance of β^{IV} is

$$\frac{1}{n}\sigma^2 \frac{p\lim \frac{1}{n}\Sigma z_i^2}{p\lim\left(\frac{1}{n}\Sigma z_i x_i\right)^2} \rightarrow \frac{1}{n}\sigma^2 \frac{Var(Z_i)}{Cov(Z_i, X_i)^2}$$

- The greater the covariance between X and Z, the lower the asymptotic variance.

- When we have just enough instruments for consistent estimation, we say the regression equation is **exactly identified.**
- When we have more than enough instruments, the regression equation is **over identified.**
- When we do not have enough instruments, the equation is **under identified** (and inconsistent).

- When the regression equation is over identified, we have more instruments than we need.
- We construct a new instrument that combines the original instruments.

IV estimation of the multiple regression model

Case 1: One endogenous variable, one instrument.

Case 2: One endogenous variable, more than one instruments. (Two stage least squares)

Case 3: More than one endogenous variables, more than one instruments. (Two stage least squares)

- Instrumental variables methods are much less efficient than OLS.
- The stronger the correlation between the instruments and the explanators, the more efficient IV is.
- If the correlation between Z and X is too low, then Z is a weak instrument, and 2SLS is not a helpful procedure.

Case 1: One endogenous variable, one instrument.

- Consider the following regression.

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \exp + u$$

- Suppose that educ is endogenous but exp is exogenous.

- To explain IV regression for multiple regression, it is often useful to use different notations for endogenous end exogenous variable.
- Let us use y for endogenous variable (i.e., correlated with u) and z for exogenous variables (i.e., uncorrelated with u).
- Then, we can write the model as:

$$y_1=\beta_0+\beta_1 y_2+\beta_2 z_1+u \text{ ....(1)}$$

y_1 is log(wage), y_2 is educ, and z_1 is exp.

- This model is called the structural equation to emphasize that this equation shows the causal relationship. Off course, OLS cannot be used to consistently estimate the parameters since y_2 is endogenous.
- If you have an instrument for y_2, you can consistently estimate the model. Let us call this instrument, z_2.

- As before, z_2 should satisfy (i) instrument exogeneity, and (ii) instrument relevance.

- For a multiple regression model, these conditions are written as:
  1. **The instrument exogeneity**

     $Cov(z_2, u)=0$ ……………………….(2)
  2. **The instrument relevance**

     $y_2=\pi_0+\pi_1 z_1+\pi_2 z_2+\text{error}$ ……………….(3)

     and $\pi_2 \neq 0$

- In addition, z_2 should **not** be a part of the structural equation (1). This is called the **exclusion restriction**.

All the exogenous variables included. This equation is often called the reduced form equation.

- Now, we have the following three conditions that can be used to obtain the IV estimators.

$$E(u)=0$$

$$Cov(z_1,u)=0$$

$$Cov(z_2,u)=0$$ (this is from the instrument exogeneity)

- Above method can be easily extended to the case where there are more explanatory variables (but only one endogenous variable).

- Consider the following model.

$$y_1=\beta_0+\beta_1 y_2+\beta_2 z_1+\beta_3 z_2+\beta_4 z_3+..+ \beta_k z_{k-1}+ u$$

- Suppose that z_k is the instrument for y_2. Then the IV estimators are the solution to the following equations.

$$\sum_{i=1}^{n} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1} - \ldots - \hat{\beta}_k z_{ik-1}) = 0$$

$$\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1} - \ldots - \hat{\beta}_k z_{ik-1}) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} z_{ik}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1} - \ldots - \hat{\beta}_k z_{ik-1}) = 0$$

Solution to the above equations are the IV estimators when there are many explanatory variables, but only one endogenous variable and one instrument.

Case 2: One endogenous variable, more than one instruments.

Two stage least squares

- Consider the following model with one endogenous variable.

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

- Now, suppose that you have two instruments for y_2 that satisfy the instrument conditions. Call them z_2 and z_3.

- You can apply IV method using either z_2 or z_3. But this produces two different estimators. Moreover, they are not efficient.
- There is a more efficient estimator.
- First, it is important to lay out the instrument conditions.

- For z_2 and z_3 to be valid instruments, they have to satisfy the following two conditions.

1. **Instrument exogeneity**

   $Cov(z_2, u)=0$ and $Cov(z_3, u)=0$

2. **Instrument relevance**

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

   and $\pi_2 \neq 0$ or $\pi_3 \neq 0$

Include all the exogenous variables

In addition, z_2 and z_3 should not be a part of the structural equation. These are called the **exclusion restrictions**.

- Instead of using only one instrument, we use a linear combination of z_2 and z_3 as the instrument.
- Since a linear combination of z_2 and z_3 also satisfies the instrument conditions, this is a valid method.
- The question is how to find the best linear combination of z_2 and z_3.

- It turns out that OLS regression of the following model provides the best linear combination.

$$y_2=\pi_0+\pi_1 z_1+ \pi_2 z_2+ \pi_3 z_3+\text{error}$$

- After you estimate this model, you get the predicted value of y_2.

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$$

- Since $\hat{y}_2$ is a combination of variables which are not correlated with u, $\hat{y}_2$ is not correlated with u as well. At the same time, $\hat{y}_2$ is correlate with y_2. Thus this is a valid instrument.

- Thus, we have the following three conditions that can be used to derive an IV estimator.

$E(u)=0$

$Cov(z_1,u)=0$

$Cov(\hat{y}_2,u)=0$

The sample counter part of the above equations are given by:

$$\sum_{i=1}^{n} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} \hat{y}_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

- This is a set of three equations with three unknowns $\hat{\beta}_0 \ \hat{\beta}_1 \ \hat{\beta}_2$
- Solution to these equations are special type of IV estimators called **the two stage least square estimators.**

The estimation procedures of the two stage least square (2SLS).

Stage 1. Estimate the following model using OLS and get the predicted value for y_2: $\hat{y}_2$.

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

Make sure to put all the exogenous variables

Stage 2. replace y_2 with $\hat{y}_2$, then estimate the following model using OLS.

$$y_1=\beta_0+\beta_1\hat{y}_2+\beta_2 z_1+error$$

OLS estimators of the coefficients are the two stage least square estimators (2SLS).

Case 3: More than one endogenous variables, more than one instruments

- Consider the following structural equation.

$$y_1=\beta_0+\beta_1 y_2+\beta_2 y_3+\beta_3 z_1+\beta_4 z_2+\beta_5 z_3+u_1$$

There are two endogenous variables, y_2 and y_3. Thus, OLS will be biased. In order to estimate this model with IV method, you need at least 2 instruments.

When you have multiple endogenous variables, **you need at least the same number of instruments as the endogenous variables.**

- Suppose you have 3 instruments: z_4 z_5 z_6. As usual, these instruments should satisfy 2 conditions. The first is that they should not be correlated with u_1 (Instrument exogeneity). The second is that they should be correlated with endogenous variable (instrument relevance).

The estimation procedure

- The 2SLS procedure when there are more than one endogenous variables is shown here.

- $y_1=\beta_0+\beta_1 y_2+\beta_2 y_3+\beta_3 z_1+\beta_4 z_2+\beta_5 z_3+u_1$

Suppose you have three Instruments : z_4 z_5 z_6.

- First stage: Estimate the following two reduced from regressions

$$y_2=\pi_{10}+\pi_{11}z_1+\pi_{12}z_2+\pi_{13}z_3+\pi_{14}z_4+\pi_{15}z_5+\pi_{16}z_6+\text{error}$$
$$y_3=\pi_{20}+\pi_{21}z_1+\pi_{22}z_2+\pi_{23}z_3+\pi_{24}z_4+\pi_{25}z_5+\pi_{26}z_6+\text{error}$$

Then obtain $\hat{y}_2$ and $\hat{y}_3$.

- The second stage: Estimate the following 'second stage regression'.

$$y_1=\beta_0+\beta_1\hat{y}_2+\beta_2\hat{y}_3+\beta_3z_1+\beta_4z_2+\beta_5z_3+u_1$$

The estimated coefficients are the 2SLS coefficients.

- The main trick to using instrumental variables is finding the instruments in the first place.
- When reading studies that employ instruments, be skeptical. Are the authors reasonably convincing that their proposed instruments are valid?

- Instrumental variables can be a powerful technique for drawing causal inferences from not-entirely-random processes.
- However, IV must be used with care.
- If instruments are weak, or correlated with ε, then IV will still be biased.