Econometrics II: Econometric Modelling

Jürgen Meinecke

Research School of Economics, Australian National University

17 August, 2018

Roadmap

Instrumental Variables Estimation

- Definition and Role of Instrumental Variable
- The Two Stage Least Squares (TSLS) Estimator
- Properties of the TSLS Estimator
- Where Do Instruments Come From? (Part 1)

Recall from last lecture that there are five reasons why OLS may produce biased results:

1. Omitted variable bias
2. Wrong functional form
3. Errors-in-variables bias
4. Sample selection bias
5. Simultaneous causality bias

If any of these biases arise, we say that the explanatory variable is endogenous

Endogenous literally means "determined within the system"

We face the following generic estimation problem

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad \text{where} \quad \mathrm{E}[u_i | X_i] \neq \mathrm{E}[u_i].$$

After all we have learned during the past three weeks, we know that

- OLS Assumption 1 fails here
- the explanatory variable X_i is endogenous
- we cannot use EMET2007 results to estimate β_1 (i.e., we cannot use OLS estimation)

The core problem is that X_i is correlated with u_i

To make some progress, let's look at the relationship between X_i and u_i as follows: the regressor X_i can be thought of as consisting of two parts

- one part that may be correlated with u_i
- one part that is not

Enter the instrumental variable:
An IV essentially enables us to isolate these two parts of X_i

What exactly is an IV?

Definition (Instrumental Variable, Instrument)

An **instrumental variable (IV)** Z_i is a previously unused explanatory variable that is correlated with X_i but uncorrelated with u_i. Sometimes, an instrumental variable is simply referred to as **instrument**.

An IV needs to satisfy two technical conditions:

- Instrument relevance : $\rho_{Z_i X_i} \neq 0$
- Instrument exogeneity : $\rho_{Z_i u_i} = 0$

Definition (Instrument Valdidity)

An instrumental variable is called **valid** if it is relevant and exogenous.

(This definition implies that there exist *invalid* instruments. That seems a bit silly, but it is quite common in textbooks and the academic literature. Most of the time when I talk about an IV I presume it is actually a valid one.)

Where do IV come from?
(will see this later. . .)

If you have an IV, how can you it to estimate β_1?

Roadmap

Instrumental Variables Estimation

- Definition and Role of Instrumental Variable
- The Two Stage Least Squares (TSLS) Estimator
- Properties of the TSLS Estimator
- Where Do Instruments Come From? (Part 1)

As the name suggests, TSLS estimation has two stages

1. Run OLS of X_i on a constant and Z_i

   $$X_i = \pi_0 + \pi_1 Z_i + v_i$$

   (note: if IV is relevant then $\pi_1 \neq 0$)

   - Obtain OLS estimates $\hat{\pi}_0$ and $\hat{\pi}_1$
   - Because Z_i is uncorrelated with u_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with u_i
   - We don't know π_0 or π_1 but use estimates instead
   - Compute the predicted values of X_i: $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, for $i = 1, \ldots, n$
   - The predicted $\hat{X}_i$ can be viewed as the *exogenous* part of X_i
   - In the first stage, we therefore have used the instrument Z_i to isolate and extract the exogenous part of X_i

2. Run OLS of Y_i on a constant and $\hat{X}_i$ (instead of using X_i)
   - Obtain OLS estimates and denote them by $\hat{\beta}_0^{TSLS}$ and $\hat{\beta}_1^{TSLS}$
   - These are the IV estimates
   - This procedure is based on the following rewriting of the regression model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ &= \beta_0 + \beta_1 \hat{X}_i + \big(u_i + \beta_1 (X_i - \hat{X}_i)\big) \\ &=: \beta_0 + \beta_1 \hat{X}_i + w_i \end{aligned}$$

   - Is $\hat{X}_i$ exogenous here?

Let's set up the TSLS estimator more technically:

Definition (TSLS Estimator)

$$\hat{\beta}_0^{TSLS}, \hat{\beta}_1^{TSLS} := \operatorname*{argmin}_{b_0,b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 \hat{X}_i)^2,$$

where

$$\hat{X}_i := \hat{\pi}_0 + \hat{\pi}_1 Z_i$$

$$\hat{\pi}_0, \hat{\pi}_1 := \operatorname*{argmin}_{p_0,p_1} \sum_{i=1}^{n} (X_i - p_0 - p_1 Z_i)^2$$

If I asked you to solve the above minimization problem, what result would you get?

Turns out, coming up with the answer is easy

By analogy to the solution for the standard least squares problem, the TSLS estimator will be

$$\hat{\beta}_{1,TSLS} := \frac{\sum_{i=1}^{n}(\hat{X}_i - \bar{\hat{X}})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(\hat{X}_i - \bar{\hat{X}})(\hat{X}_i - \bar{\hat{X}})} = \frac{s_{\hat{X}Y}}{s^2_{\hat{X}}}$$

$$\hat{\beta}_{0,TSLS} := \bar{Y} - \hat{\beta}_{1,TSLS}\bar{\hat{X}}$$

This should not be surprising:
$\hat{\beta}_1^{TSLS}$ is the sample covariance between $\hat{X}_i$ and Y_i over the sample variance of $\hat{X}_i$

That is a straightforward application of the usual OLS formula

Roadmap

Instrumental Variables Estimation

- Definition and Role of Instrumental Variable
- The Two Stage Least Squares (TSLS) Estimator
- Properties of the TSLS Estimator
- Where Do Instruments Come From? (Part 1)

The benefit of the TSLS estimator is that it overcomes the endogeneity problem

Theorem

The TSLS estimator $\hat{\beta}_1^{TSLS}$ is consistent for β_1.

Proving this is a bit technical, but the intuition should be clear

By having an instrumental variable Z_i we are able to extract the exogenous variation out of X_i while making sure that any endogenous variation does not contaminate the estimates

Roadmap

Instrumental Variables Estimation

- Definition and Role of Instrumental Variable
- The Two Stage Least Squares (TSLS) Estimator
- Properties of the TSLS Estimator
- Where Do Instruments Come From? (Part 1)

Recap: what have we done on the previous slides?

We were sidestepping the endogeneity problem $\mathrm{E}[u_i|X_i] \neq \mathrm{E}[u_i]$

Key trick: introduce entirely new type of variable called IV

The beauty of the IV Z_i is that it is exogenous while, at the same time, it has some relationship with the endogenous variable X_i

We have (almost) done a *deus ex machina*:
“A device through which a seemingly unsolvable problem is suddenly and abruptly resolved by the contrived and unexpected intervention of some new event, character, ability or object“
(quoted with adaptations from Wikipedia)

Where should such a wonderful IV come from?

There is no golden rule for “deriving” IV

Instead, IV must be found on a case by case basis
(always depending on a particular application)

This turns out to be extremely tricky

Unfortunately, it’s more of an art than a science

During the past 20 years (or so) applied econometricians have written lots of research papers trying to solve endogeneity problems by finding interesting, exotic, controversial IV

Whenever a researcher suggests a particular variable to be an instrument, it is crucial to check whether the two conditions for valid instruments are satisfied: relevance and exogeneity

Example 1: Effect of Studying on Grades

Stinebrickner, R and Stinebrickner, T. R. (2008),
"The Causal Effect of Studying on Academic Performance,"
The B.E. Journal of Economic Analysis & Policy

Research question:
What is the effect on grades of studying for an additional hour per day?

- grades as measured by GPA
- study time (hours per day)

Data: 210 college freshmen, time use survey, random assignment of roommates

What is the endogeneity problem here?

What do authors suggest as IV:

$Z_i = 1$ if roommate brought video game at beginning of school year, zero otherwise

Justification:

- video games adversely affect studying
- freshmen whose roommates brought video games are likely to study less (all else equal)
- at the same time, the presence of video games should not have any other direct effect on grades
- (bear in mind that roommates were randomly assigned, so freshmen were not able to choose roommates based on their preferences for video gaming)

Do you think the IV is relevant and exogenous?

Finding: studying has large positive effect on grades (surprise!)

Example 2: Effect of economic growth on civil conflict

Miguel, E., Satyanath, S. and Sergenti, E. (2004), "*Economic Shocks and Civil Conflict: An Instrumental Variables Approach*", Journal of Political Economy

Research question:
Do economic conditions affect the likelihood of civil conflict?

What do they mean by

- economic conditions: real GDP growth
- civil conflict: use of armed forces between two parties with at least 25 deaths

Data: annual data on 41 African countries from 1981-1999

What is the endogeneity problem here?

What do authors suggest as IV: rainfall

Justification:

- rainfall adversely affects GDP “growth in economies that largely rely on rainfed agriculture“
- rainfall can be regarded as relatively random
- rainfall does not have a direct effect on civil conflict

Do you think the IV is relevant and exogenous?

Finding: growth is strongly negatively related to civil conflict

Problem Solving Exercises

1. In the model $Y_i = \beta_0 + \beta_1 X_i + u_i$, the dependent variable Y_i is measured with error. Instead of Y_i you observe $\tilde{Y}_i = Y_i + w_i$, where w_i is purely random.

   (i) Derive $\mathrm{E}\left[\hat{\beta}_1 | X_i\right]$.

   (ii) Derive $\mathrm{Var}\left[\hat{\beta}_1 | X_i\right]$.

   Modify the proofs of bias and variance from lecture 5 of last semester (given on the following two slides) to derive the results.

$$
\begin{aligned}
\mathrm{E}[\hat{\beta}_1|X_i] &= \mathrm{E}\left[\beta_1 + \frac{\sum_{i=1}^{n}(X_i-\bar{X})u_i}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\middle|X_i\right] \\
&= \beta_1 + \mathrm{E}\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})u_i}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\middle|X_i\right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\mathrm{E}\left[\sum_{i=1}^{n}(X_i-\bar{X})\cdot u_i\middle|X_i\right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sum_{i=1}^{n}\mathrm{E}\left[(X_i-\bar{X})\cdot u_i\middle|X_i\right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sum_{i=1}^{n}(X_i-\bar{X})\cdot\mathrm{E}\left[u_i\middle|X_i\right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sum_{i=1}^{n}(X_i-\bar{X})\cdot\mu_u \\
&= \beta_1 + \frac{\mu_u}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sum_{i=1}^{n}(X_i-\bar{X}) \\
&= \beta_1
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_1|X_i) &= \mathrm{Var}\left(\beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \middle| X_i\right) \\
&= \mathrm{Var}\left(\frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \middle| X_i\right) \\
&= \frac{1}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2} \mathrm{Var}\left(\sum_{i=1}^n (X_i - \bar{X})u_i \middle| X_i\right) \\
&= \frac{1}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2} \sum_{i=1}^n \mathrm{Var}\left((X_i - \bar{X})u_i \middle| X_i\right) \\
&= \frac{1}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2} \sum_{i=1}^n (X_i - \bar{X})^2 \mathrm{Var}(u_i|X_i) \\
&= \frac{1}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sigma_u^2 \\
&= \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}
\end{aligned}
$$