

Recitation 9: IV estimation

Seung-hun Lee

Columbia University
Undergraduate Introduction to Econometrics Recitation

November 17th, 2022

Instrumental variables

Why use IV? Deal with $E[u_i|X_i] \neq 0$

- Recall that OLS estimates can be biased when we have omitted variable bias, measurement error, and simultaneity bias
- Instrumental variable: Allows us to eliminate bias
 - ▶ When you have an independent variable X , there are parts of this variable that are correlated with u and the other parts that are independent of u .
 - ▶ If we are able to find Z that is correlated with X but not with u , using this Z variable would allow you to sort variable X into what is correlated with u and what is not.
 - ▶ Then, using the part that is not correlated with u , variable Z allows us to get unbiased estimates.
- Uses of IV
 - ▶ Endogenous variable: X determined as choice, not as given
 - ▶ Simultaneity bias: Y can lead to changes in X
 - ▶ Omitted variable bias: Unincorporated determinant of Y
 - ▶ Measurement error in X

Association vs causation

- Association refers to any type of statistical dependence
 - ▶ It mostly refers to correlation, which is

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}$$

- Running an OLS is practically equivalent to finding the correlation
 - ▶ From the population regression,

$$\beta_1 = \frac{E[X - E[X]]E[Y - E[Y]]}{(E[X - E[X]])^2} = \text{corr}(X, Y) \sqrt{\frac{\text{var}(Y)}{\text{var}(X)}}$$

implying that β_1 is obtained by multiplying $\text{corr}(X, Y)$ with some constant

- ▶ Same correlation direction (variances are nonnegative)
- ▶ Not good for finding causal effect of X on Y : Switching the two would give same results

What conditions should IV satisfy?

- Must satisfy
 - ▶ **Relevance:** Variable Z satisfies relevancy condition if $cov(X, Z) \neq 0$
 - ▶ **Exogeneity:** Variable Z satisfies exogeneity condition if $cov(Z, u) = 0$
- In words,
 - ▶ Variable Z should be somewhat correlated with the variable X
 - ▶ Variable Z should not be correlated with u
 - ▶ (For exogeneity): Variable Z should affect Y only through X , or when X is controlled for, Z alone should not affect Y (**exclusion**)
- More on exclusion (on model $Y = \beta_0 + \beta_1 X + u$)

$$\begin{aligned} cov(Z, u) &= cov(Z, Y - \beta_0 - \beta_1 X) = 0 \\ &= cov(Z, Y) - cov(Z, \beta_0) - cov(Z, \beta_1 X) = 0 \\ &\implies cov(Z, Y) = cov(Z, \beta_1 X) \end{aligned}$$

This condition means that Z is correlated with Y *only through* X

2SLS estimation

- **Regress with X as dependent, Z as independent variable.** Regress

$$X = \delta_0 + \delta_1 Z + v$$

From this regression, obtain the predicted values of X , denoted as $\hat{X} = \hat{\delta}_0 + \hat{\delta}_1 Z$. This \hat{X} is the part that is related with Z but is uncorrelated with u .

- **Regress with Y as dependent, \hat{X} as independent variable.** Regressing with this \hat{X} will satisfy the $E[u|\hat{X}] = 0$ condition, as the \hat{X} is uncorrelated with u . Thus, your regression equation looks like this:

$$Y = \beta_0 + \beta_1 \hat{X} + u$$

Then you run a OLS regression on the above equation and get the 2SLS estimator $\hat{\beta}_{\text{TSLs}}$.

Estimating using a covariance method

- Note that

$$\text{cov}(Z, Y) = \text{cov}(Z, \beta_1 X) \implies \text{cov}(Z, Y) = \beta_1 \text{cov}(Z, X)$$

- From this, we can get

$$\beta_1 = \frac{\text{cov}(Z, Y)}{\text{cov}(Z, X)}$$

where division is possible because we require a relevancy condition ($\text{cov}(Z, X) \neq 0$)

- Taking the sample counterparts, we get our IV estimator

$$\hat{\beta}_{TSLS} = \frac{s_{ZY}}{s_{ZX}}$$

Reduced form method

- Denote

$$X = \pi_0 + \pi_1 Z + v \text{ (where } \text{cov}(Z, v) = 0 \text{)}$$

$$Y = \gamma_0 + \gamma_1 Z + w \text{ (where } \text{cov}(Z, w) = 0 \text{)}$$

- Rewrite the first equation in terms of Z and get

$$Z = \frac{X}{\pi_1} - \frac{\pi_0}{\pi_1} - \frac{v}{\pi_1}$$

- Then plug this into the second equation. Reorganizing this equation, you should get

$$Y = \left(\gamma_0 - \frac{\pi_0 \gamma_1}{\pi_1} \right) + \left(\frac{\gamma_1}{\pi_1} \right) X + \left(w - \frac{\gamma_1}{\pi_1} v \right)$$

- As a result, β_1 from the equation with X as independent variable is $\beta_1 = \frac{\gamma_1}{\pi_1}$.

IV estimate is consistent!

- Consistency: 2SLS can be written as

$$\hat{\beta}_{1,\text{TSLS}} = \frac{s_{zy}}{s_{zx}} \simeq \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})}$$

As $n \rightarrow \infty$, we can show that $\hat{\beta}_{1,\text{TSLS}} \rightarrow \beta_1$

- ▶ $\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y}) \xrightarrow{p} \text{cov}(Z, Y)$
- ▶ $\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X}) \xrightarrow{p} \text{cov}(Z, X)$
- ▶ Note that $\text{cov}(Z, Y) = \text{cov}(Z, \beta_0 + \beta_1 X + u) = \beta_1 \text{cov}(Z, X) + \text{cov}(Z, u)$
- ▶ If Z is a valid IV, $\text{cov}(Z, u) = 0$
- ▶ So $\hat{\beta}_{1,\text{TSLS}} \rightarrow \frac{\beta_1 \text{cov}(Z, X)}{\text{cov}(Z, X)} = \beta_1$ QED

IV estimators are normally distributed (but only in large samples)

- Break down the 2SLS estimators into

$$\begin{aligned}\hat{\beta}_{\text{TSLs}} &= \frac{s_{zy}}{s_{zx}} = \frac{\sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})} = \frac{\sum_{i=1}^n (Z_i - \bar{Z})Y_i}{\sum_{i=1}^n (Z_i - \bar{Z})X_i} \\ &= \frac{\sum_{i=1}^n (Z_i - \bar{Z})(\beta_0 + \beta_1 X_i + u_i)}{\sum_{i=1}^n (Z_i - \bar{Z})X_i} = \beta_1 + \frac{\sum_{i=1}^n (Z_i - \bar{Z})u_i}{\sum_{i=1}^n (Z_i - \bar{Z})X_i} = \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})u_i}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})X_i}\end{aligned}$$

- After subtracting both sides by β_1 and multiplying both sides by \sqrt{n} , we get

$$\sqrt{n}(\hat{\beta}_{\text{TSLs}} - \beta_1) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n (Z_i - \bar{Z})u_i}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})X_i}$$

- Denominator $\xrightarrow{P} \text{cov}(Z, X)$ by (weak) law of large numbers
- Numerators $\sim N(0, \text{var}[(Z - \mu_Z)u])$ by central limit theorem.
- Thus, $\hat{\beta}_{\text{TSLs}}$ has a normal distribution.

Multivariate case and Identification issues

- Suppose that we have

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \delta_1 W_{1i} + \dots + \delta_l W_{li} + u_i$$

where X variables are endogenous and W variables are exogenous.

- Assume that we have found a total of m (not necessarily equal to k) variables that could qualify as IVs, all of them satisfying
 - ▶ **Just-identified:** When $m = k$. There are just enough instruments to identify k endogenous variables
 - ▶ **Overidentified:** When $m > k$. There are more than enough instruments.
 - ▶ **Underidentified:** When $m < k$. There are not enough instruments. The coefficients for X 's will not be identified
- Need at least as much instrumental variables as the number of endogenous regressors you have

IV assumptions

IV assumptions, multiple variable case

- IV1** $E[u_i | W_{1i}, \dots, W_{li}] = 0$ (At least for exogenous variables, this is satisfied)
- IV2** $(Y_i, X_{1i}, \dots, X_{ki}, W_{1i}, \dots, W_{li}, Z_{1i}, \dots, Z_{mi})$ are IID
- IV3** The Y, X, W, Z variables all have nonzero finite 4th moments
- IV4** The instruments are valid. That is $\text{cov}(Z_{ji}, u_i) = 0$ for all $j = 1, \dots, m$ and relevancy conditions are satisfied for all Z 's.

We can directly test relevance condition

- Assume we have

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

where only X_i is endogenous.

- Suppose we have m instruments. Then our first stage equation looks like

$$X_i = \pi_0 + \pi_1 Z_{1i} + \dots + \pi_m Z_{mi} + \pi_{(m+1)i} W_{1i} + \dots + \pi_{(m+r)i} W_{ri} + v_i$$

- We say our instrument is relevant if at least one of π_1, \dots, π_m is statistically nonzero.
- Run the F -test with these null and alternative hypothesis

$$H_0 : \pi_1 = \dots = \pi_m = 0 \text{ vs. } H_1 : \neg H_0$$

- By rule of thumb, we want F -statistics to be larger than 10. Otherwise, we have a weak instrument problem (and IV estimator may not be normally distributed even in large numbers)

We can get a taste of exogeneity condition with many instruments

- Consider the following case: We want to regress

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \delta_1 W_{1i} + \dots + \delta_l W_{li} + u_i$$

and we found m possible candidates for instrumental variables, Z_1, \dots, Z_m

- Overidentifying restriction test we conduct here is a J -test that uses TSLS estimators and not the hypothesized β values. The steps are as follows
 - 1 Regress the above equation using 2SLS. Obtain the residuals $\hat{u}_i = Y_i - \hat{Y}_i$
 - 2 Regress residuals onto instruments and other controls, namely

$$\hat{u}_i = \alpha_0 + \alpha_1 Z_{1i} + \dots + \alpha_m Z_{mi} + \alpha_{m+1} W_{1i} + \dots + \alpha_{m+l} W_{li} + v_i$$

- 3 Compute the F -statistics from $H_0 : \alpha_1 = \dots = \alpha_m = 0$ vs. $H_1 : \neg H_0$
- 4 Derive the J statistics as follows $J = m \times F$. J follows χ^2_{m-k} distribution
- 5 If you have an endogenous IV, you will end up rejecting the null
- 6 Then, you need to make a guess on which instrument is violating the exogeneity condition, drop those, and redo the above procedure. (Good luck!)

Overidentification test as a 'distance' between different IVs

- Assume a case with one endogenous variable X and two IVs (Z_1, Z_2)
- The estimates for the coefficient for X are

$$\hat{\beta}_{Z_1} = \frac{\text{cov}(Z_1, Y)}{\text{cov}(Z_1, X)}, \quad \hat{\beta}_{Z_2} = \frac{\text{cov}(Z_2, Y)}{\text{cov}(Z_2, X)}$$

The numbers look different, although both $\hat{\beta}$'s are suppose to be the same coefficient estimating impact of X_1 on Y .

- The overidentification test checks whether the differences between $\hat{\beta}_{Z_1}, \hat{\beta}_{Z_2}$ are large.
 - ▶ If they converge to same item, we do not have to worry. Otherwise, one or more IV might be faulty

What if we have just enough IVs?

- To be very honest, there is not much we can do in terms of rigorous testing approach. This is usually the area of judgement call.
- One way to do this is to use the approach of exclusion restriction - you argue that the instrumental variable Z affects Y only through X .
 - ▶ If Z affects Y directly without X , then the exclusion restriction fails.
- You try to persuade others based on logic, previous practices, or intuition (usually a combination of all three).