Recitation 9

Seung-hun Lee

Columbia University

Instrumental variables

Motivation

- 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 correlate 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

Conditions

- 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*

Instrumental variables

Estimation: 2SLS

- **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 $\hat{\beta}_{TSLS}$.

Instrumental variables

Estimation: Covariance method

- Note that

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

- From this, we can get

$$\hat{\beta}_1 = \frac{cov(Z, Y)}{cov(Z, X)}$$

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

Instrumental variables

Estimation: Reduced form method

- Denote

$$X = \pi_0 + \pi_1 Z + v \text{ (where } cov(Z, v) = 0)$$
$$Y = \gamma_0 + \gamma_1 Z + w \text{ (where } cov(Z, w) = 0)$$

- 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

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

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

Instrumental variables

Properties

- Consistency: 2SLS can be written as

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

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

Instrumental variables

Extending to Multivariate case

- Suppose that we have

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

  where X variables are endogenous and W variables are exogenous.
- Assume that we have found a total of m (not necessarily, equal) variables that could qualify as IVs, all of them satisfying
  - **IV1** $E[u_i|W_{1i}, ..., W_{li}] = 0$ (At least for exogenous variables, this is satisfied)
  - **IV2** $(Y_i, X_{1i}, .., X_{ki}, W_{1i}, .., W_{li}, Z_{1i}, ..., Z_{mi})$ are IID
  - **IV3** The Y, X, W, Z variables all have nonzero finite 4th moments
  - **IV4** The instruments are valid. That is $cov(Z_{ji}, u_i) = 0$ for all $j = 1, ..., m$ and relevancy conditions are satisfied for all Z's.

Instrumental variables

Extending to Multivariate case: Identification

- A parameter is **identified** if different values of the parameter produce different distributions of the data.
- In other words, there is a one-to-one matching of the parameters and the distributions.
- If it is the case that the same distribution can be obtained from different parameter values, we say that the parameters are not identified
- If we have k endogenous regressors and m IV's
  - **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

Instrumental variables

Extending to Multivariate case: Overidentification tests

- **Overidentification test** applies to the case where $m > k$ and aims to test whether the instrument variables we found are valid.
- Assume a case with one endogenous variable X and two IVs (Z_1, Z_2)
- The estimates for the coefficient for X are

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

  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