

Instrumental Variables Estimation and Two-Stage Least Squares

Lecture 3/1 – 2/2014

Chayanee Chawanote

Outline

- Problem of endogenous variables
- Basic IV estimation
- The method of 2-stage least squares
- Testing for endogeneity and testing for overidentifying restrictions

Why use instrumental variables (IV)?

- Endogenous: one or more regressors are correlated with the error term, $\text{Cov}(x_{ik}, u_i) \neq 0$. Then, OLS cannot consistently estimate.
- Instrumental variables: exogenous variables z_{ij} which are correlated with the endogenous regressor, but not correlated with the error term, $\text{Cov}(z_{ij}, u_i) = 0$.
- 3 main problems need IV:
 - Omitted variables
 - Simultaneity and reverse causality (from structural equation)
 - Measurement errors (Errors-in-variables)

Omitted variables

- If we have omitted variables bias (unobserved heterogeneity), how we deal with it?
 - 1) Ignore, but get biased and inconsistent estimators
 - 2) Try to find a suitable proxy variable for the unobserved variable
 - 3) Assume the omitted variable does not change over time. Then, use fixed effects or first-differencing methods
- Now, use IV

What should be an instrumental variable?

- In order for a variable, z , to serve as a valid instrument for x , the following must be true
 - The instrument must be exogenous: $\text{Cov}(z,u) = 0$
 - The instrument must be correlated with the endogenous variable x : $\text{Cov}(z,x) \neq 0$
- We have to use common sense and economic theory to decide if it makes sense to assume $\text{Cov}(z,u) = 0$
- We can test if $\text{Cov}(z,x) \neq 0$
 - Just testing $H_0: \pi_1 = 0$ in $x = \pi_0 + \pi_1 z + v$
 - Sometimes refer to this regression as the first-stage regression

IV Estimation in the Simple Regression Case

- For $y = \beta_0 + \beta_1 x + u$, and $\text{Cov}(z, u) = 0$, $\text{Cov}(z, x) \neq 0$, write β_1 in terms of population covariances:

$$\text{Cov}(z, y) = \beta_1 \text{Cov}(z, x) + \text{Cov}(z, u)$$

- So, $\beta_1 = \text{Cov}(z, y) / \text{Cov}(z, x)$
- Then the IV estimator for β_1 (sample analogs) is

$$\hat{\beta}_1 = \frac{\sum (z_i - \bar{z})(y_i - \bar{y})}{\sum (z_i - \bar{z})(x_i - \bar{x})}$$

- and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- $\text{plim}(\hat{\beta}_1) = \beta_1$ provided that assumption $\text{Cov}(z, u) = 0$, $\text{Cov}(z, x) \neq 0$ are satisfied; otherwise, IV is inconsistent

IV Estimation in the Simple Regression Case

- Impose homoskedasticity assumption on IV estimator

$$E(u^2 | z) = \sigma^2 = \text{Var}(u)$$

- Asymptotic variance of $\hat{\beta}_1$ is $\frac{\sigma^2}{n\sigma_x^2\rho_{xz}^2}$

- Hence, to estimate s.e. for IV

– σ_x^2 is computed from sample variance of x

– ρ_{xz}^2 is from $R^2_{x,z}$

– σ^2 is computed from IV residuals

Estimated asymptotic standard error = $\left[\frac{\hat{\sigma}^2}{SST_x R^2_{x,z}} \right]^{1/2}$

IV vs. OLS estimation

- Variance in IV case differs from OLS only in the R^2 from regressing x on z
- Since $R^2 < 1$, IV variance is always larger
- However, IV is consistent, while OLS is inconsistent when $\text{Cov}(x, u) \neq 0$
- If $R^2_{x,z}$ is small (no strong relationship between x and z), then the IV variance can be much larger than the OLS variance
- The stronger the correlation between z and x , the smaller the IV variance