ECON3389 Econometric Methods

Module 6
Instrumental Variable

Alberto Cappello

Department of Economics, Boston College

Spring 2025

Instrumental Variable Method

- The earliest application involved attempts to estimate demand and supply curves for a product.
- A simple but difficult question: How to find the supply or demand curves?
- Difficulty: We can only observe intersections of supply and demand, yielding pairs.
- Solution: Wright (1928) used variables that appear in one equation to shift this equation and trace out the other.
- The variables that do the shifting came to be known as Instrumental Variables method.
- IV can address the problems of omitted variable bias, measurement error, and reverse causality problems.

Instrumental Variable Method

- An endogenous variable is one that both we are interested in and is correlated with u.
- An exogenous variable is one that is uncorrelated with u.
- Historical note: "Endogenous" literally means "determined within the system," that is, a variable that is jointly determined with Y, that is, a variable subject to simultaneous causality.
- IV regression can be used to address OVB and errors-in-variable bias, not just to simultaneous causality bias.

Instrumental Variable Method

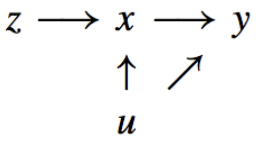
- Suppose a simple OLS regression:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Because $E[u_i|X_i] \neq 0$, we can use an instrumental variable Z_i to obtain a consistent estimate of the coefficient.
- Intuitively, we want to split X_i into two parts:
  1. One part that is correlated with the error term.
  2. One part that is uncorrelated with the error term.
- If we can isolate the variation in X_i that is uncorrelated with u_i, then we can use this part to obtain a consistent estimate of the causal effect of X_i on Y_i.

Instrumental Variable Method

- An instrumental variable Z_i must satisfy the following two properties:
  1. Instrumental relevance: Z_i should be correlated with the causal variable of interest X_i, thus
  $$\text{Cov}(X_i, Z_i) \neq 0.$$
  2. Instrumental exogeneity: Z_i is as good as randomly assigned and only affects Y_i through X_i, such that
  $$\text{Cov}(Z_i, u_i) = 0.$$
  or, equivalently, Exclusion restriction: IV affects the outcome only through the endogenous variable X_i, nothing else.



IV Estimator: Two Steps Least Squares (2SLS)

- Consider the covariance between Y_i and Z_i

$$\text{Cov}(Z_i, Y_i) = \text{Cov}[Z_i, (\beta_0 + \beta_1 X_i + u_i)] = \beta_1 \text{Cov}[Z_i, X_i]$$

thus if the instrument is valid

$$\beta_1 = \frac{\text{Cov}(Z_i, Y_i)}{\text{Cov}(Z_i, X_i)}$$

- We can estimate the causal effect of X_i on Y_i in two steps:
  1. First stage: Regress X_i on Z_i and obtain predicted values $\hat{X}_i$, if

$$\text{Cov}(Z_i, u_i) = 0, \text{ then } \hat{X}_i \text{ contains variation in } X_i \text{ that is uncorrelated with } u_i.$$

  2. Second stage: Regress Y_i on $\hat{X}_i$ to obtain the Two Stage Least Squares estimator $\hat{\beta}_{2SLS}$:

$$\hat{\beta}_{2SLS} = \frac{\sum(Y_i - \bar{Y})(\hat{X}_i - \bar{X})}{\sum(\hat{X}_i - \bar{X})^2}$$

IV Estimator: Two Steps Least Squares (2SLS)

- Because $\hat{X}_i = \pi_0 + \pi_1 Z_i$, then
$$\hat{X} = \pi_0 + \pi_1 Z_i$$
- We have
$$\hat{X}_i - \hat{X} = \pi_1 (Z_i - \bar{Z})$$
- Also, because π_1 is the estimating coefficient of Z_i on X_i, then based on the simple OLS formula:
$$\pi_1 = \frac{\sum (X_i - \bar{X})(Z_i - \bar{Z})}{\sum (Z_i - \bar{Z})^2}$$

IV Estimator: Two Steps Least Squares (2SLS)

- Substituting for $\hat{X}_i - \hat{X}$ in $\hat{\beta}_{2SLS}$ we obtain:

$$\hat{\beta}_{2SLS} = \frac{\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}$$

- This gives the 2SLS IV estimator:

$$\hat{\beta}_{2SLS} = \frac{\hat{Cov}_{ZY}}{\hat{Cov}_{ZX}} \quad \text{where } \hat{Cov} \text{is the empirical covariance}$$

- The 2SLS estimator of β_1 is the ratio of the sample covariance between Z and Y to the sample covariance between Z and X.
- If $Z_i = X_i$, then:

$$\hat{\beta}_{2SLS} = \hat{\beta}_{OLS}$$

Statistical Properties of IV Estimator

- Take the expected value of $\hat{\beta}_{2SLS}$ conditional on X, Z and replace $Yi = \beta_0 + \beta_1 X_i + u_i$
- Since $\mathbb{E}[u_i|Z] = 0$, it can be showed that $\hat{\beta}_{2SLS}$ is unbiased

$$\mathbb{E}[\hat{\beta}_{2SLS}|X, Z] = \beta_1$$

- $\hat{\beta}_{2SLS}$ is consistent
- The standard error $SE(\hat{\beta}_{2SLS})$ is:

$$SE(\hat{\beta}_{2SLS}) = \sqrt{\frac{1}{n}\sum(Z_i - \mu_Z)^2 \hat{u_i}^2 n} \left(\frac{1}{n}\sum(Z_i - \mu_Z)X_i\right)^2$$

- Because $\hat{\beta}_{2SLS}$ is normally distributed in large samples, hypothesis tests about β can be performed by computing the t-statistic, and a 95% large-sample confidence interval is given by:

$$\hat{\beta}_{2SLS} \pm 1.96 \times SE(\hat{\beta}_{2SLS})$$