Econometrics and Applications

Kirby CHEN

Academic Year 2024-2025

Contents

1 Lecture 3: Endogeneity and Instrumental Variables

1.1 Motivation and Overlook

Example:

- Omitted variables bias
- Measurement error
- Simultaneous equations bias (reverse causality)

Our Goal

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The endogenous variable x has a real impact on Y, and we aim to find the true value of β_1.

1. **Using an Instrumental Variable to Derive the Model's Covariance**

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Taking the covariance of both sides with the instrumental variable z:

$$\mathrm{cov}(Y, z) = \mathrm{cov}(\beta_0 + \beta_1 X + \varepsilon, z)$$

Expanding the covariance expression:

$$\mathrm{cov}(Y, z) = \mathrm{cov}(\beta_0, z) + \beta_1 \times \mathrm{cov}(X, z) + \mathrm{cov}(\varepsilon, z)$$

Since the instrumental variable z is uncorrelated with both β_0 and the error term ε, these covariance terms disappear:

$$\text{cov}(Y, z) = \beta_1 \times \text{cov}(X, z)$$

Solving for β_1:

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

Instrumental Variables (IV) estimator of β_1, β_{IV}.

2. **Reduced-form Equation: Indirect Least Square, ILS**

$$x = \delta_0 + \delta_1 \times z + u$$

$$Y = \pi_0 + \pi_1 \times z + v$$

Reduced-form equation: Writing an endogenous variable in terms of exogenous variables.

$$x = \delta_0 + \delta_1 \times z + u$$
$$Y = \pi_0 + \pi_1 \times z + v$$

$$\delta_1 = \frac{\text{cov}(x, z)}{\text{var}(z)}$$
$$\pi_1 = \frac{\text{cov}(Y, z)}{\text{var}(z)}$$

We know:

$$Y = \beta_0 + \beta_1 \times x + \varepsilon$$

Regression coefficient:

$$\beta_1 = \frac{\text{cov}(Y, x)}{\text{var}(x)}$$

Using the instrumental variable:

$$\frac{\pi_1}{\delta_1} = \frac{\frac{\text{cov}(Y,z)}{\text{var}(z)}}{\frac{\text{cov}(x,z)}{\text{var}(z)}} = \frac{\text{cov}(Y, z)}{\text{cov}(x, z)} = \beta_{IV} = \beta_1$$

$$x = \delta_0 + \delta_1 \times z + u$$

$$Y = \pi_0 + \pi_1 \times z + v$$

$$\delta_1 = \frac{\text{cov}(x, z)}{\text{var}(z)}$$

$$\pi_1 = \frac{\text{cov}(Y, z)}{\text{var}(z)}$$

*****Reduced-form Equation

$$x = \delta_0 + \delta_1 \times z + u$$

$$Y = \pi_0 + \pi_1 \times z + v$$

$$\delta_1 = \frac{\text{cov}(x, z)}{\text{var}(z)}$$

$$\pi_1 = \frac{\text{cov}(Y, z)}{\text{var}(z)}$$

*****Indirect Least Squares (ILS) Method

$$Y = \beta_0 + \beta_1 \times x + \varepsilon$$

$$= \beta_0 + \beta_1 \times (\delta_0 + \delta_1 \times z + u) + \varepsilon$$

$$= \beta_0 + \beta_1 \times \delta_0 + \beta_1 \times \delta_1 \times z + \beta_1 \times u + \varepsilon$$

$$= (\beta_0 + \beta_1 \times \delta_0) + \beta_1 \times \delta_1 \times z + (\beta_1 \times u + \varepsilon)$$

$$\pi_0 = \beta_0 + \beta_1 \times \delta_0, \quad \pi_1 = \beta_1 \times \delta_1, \quad v = \beta_1 \times u + \varepsilon$$

Question: when IVs more than endogenous variables, the above two method fails.

3. **Two Stage Least Squares (2SLS/TSLS)**

*First Stage

$$x = \delta_0 + \delta_1 \times z + u$$
$$x = \hat{\delta_0} + \hat{\delta_1} \times z + \hat{u}$$
$$\hat{x} = \delta_0 + \delta_1 \times z$$

*Second Stage

$$Y = \beta_{0,2SLS} + \beta_{1,2SLS} \times \hat{x} + \varepsilon_{2SLS}$$

*Does the Model Have Endogeneity?

$$Y = \beta_0 + \beta_1 \times x + \varepsilon$$

$$= \beta_0 + \beta_1 \times (\hat{x} + \hat{u}) + \varepsilon$$
$$= \beta_0 + \beta_1 \times \hat{x} + \beta_1 \times \hat{u} + \varepsilon$$

$$\text{cov}(\hat{x}, \varepsilon_{2SLS}) = \text{cov}(\hat{x}, \beta_1 \times \hat{u} + \varepsilon)$$

$$= \beta_1 \times \text{cov}(\hat{x}, \hat{u}) + \text{cov}(\hat{x}, \varepsilon) = 0$$

When there exists many IVs:

*First Stage

$$x = \delta_0 + \delta_1 \times z_1 + \delta_2 \times z_2 + u$$
$$\hat{x} = \hat{\delta_0} + \hat{\delta_1} \times z_1 + \hat{\delta_2} \times z_2$$

*Second Stage

$$Y = \beta_{0,2SLS} + \beta_{1,2SLS} \times \hat{x} + \varepsilon_{2SLS}$$

1.2 Math Section

Assumption:

1. **Linearity**: $Y = X\beta + \epsilon$.
2. **Full rank**: $\text{rank}(X) = k$.
3. **Exogeneity**: $\mathbb{E}[\epsilon|X] = 0$.

Law of iterated expectations: $$\mathbb{E}[\epsilon] = \mathbb{E}[\mathbb{E}[\epsilon|X]] = \mathbb{E}[0] = 0.$$

4. **Homoscedasticity and nonautocorrelation**:

$$\text{Var}(\epsilon_i|X) = \sigma^2, \quad i = 1, 2, \ldots, n.$$

$$\text{Var}(\epsilon_i, \epsilon_j|X) = 0, \quad i \neq j, \quad \text{Var}(\epsilon_i \epsilon) = \sigma^2 I.$$

5. X may be fixed and random.

We assume that there is an additional vector of variables z_i, with $L \geq k$.

(1) **Exogeneity**: z_i is uncorrelated with disturbance ϵ_i.

(2) **Relevance**: z_i is correlated with explanatory variable x_i.

(3) **Homoscedasticity**: $\mathbb{E}[\epsilon_i^2|z_i] = \sigma^2$.

(4) **Random Sampling** $(x_i, z_i, \epsilon_i) \overset{iid}{\sim}$.

(5) **Moments of x_i and z_i**:

$$\mathbb{E}[x_i x_i'] = Q_{XX} < \infty, \quad \text{rank}(Q_{XX}) = k.$$

$$\mathbb{E}[z_i z_i'] = Q_{ZZ} < \infty, \quad \text{rank}(Q_{ZZ}) = L.$$

$$\mathbb{E}[z_i x_i'] = Q_{ZX} < \infty, \quad \text{rank}(Q_{ZX}) = k.$$

$$(L \times k) \quad (\text{since } L \geq k).$$

(6) **Exogeneity of Instruments**:

$$\mathbb{E}[\epsilon_i|b_i] = 0.$$

1. **OLS is biased**.

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon.$$

$$\mathbb{E}[\hat{\beta}|X] = \beta + \mathbb{E}[(X'X)^{-1}X'\epsilon|X].$$

$$= \beta + (X'X)^{-1}X'\mathbb{E}[\epsilon|X].$$

$$= \beta + (X'X)^{-1}X'\eta \neq \beta$$

(biased).

2. **OLS is inconsistent in big sample**.

Recall: $\mathbb{E}[\epsilon|X] = 0, \quad \mathbb{E}[\epsilon_i x_i]$

$$= \mathbb{E}\left[\mathbb{E}[\epsilon_i x_i|X]\right] = \mathbb{E}\left[x_i \mathbb{E}[\epsilon_i|X]\right] = 0.$$

---

2. **OLS is inconsistent**.

$$\mathbb{E}[x_i\epsilon_i] = \mathbb{E}[x_i\eta] \neq 0.$$

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\epsilon_i\right).$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{p} Q_{XX}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i\epsilon_i \xrightarrow{p} \eta \neq 0.$$

$$\Rightarrow \hat{\beta} \xrightarrow{p} \neq \beta.$$

moment non.

$$\mathbb{E}[x_i\epsilon_i] = \mathbb{E}[x_i(y_i - x_i'\beta)] = 0.$$

OLS? .

3. **A method of moment estimator $\beta_{\textbf{mom}}$ sets the sample analogue to 0**:

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - x_i'\beta_{\text{mom}}) = 0.$$

$$\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i x_i'\right)\beta_{\text{mom}} = 0.$$

.

$$\left(\sum_{i=1}^{n} x_i x_i'\right)\beta_{\text{mom}} = \sum_{i=1}^{n} x_i y_i.$$

$$\beta_{\text{mom}} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\sum_{i=1}^{n} x_i y_i\right).$$

$$= (X'X)^{-1}X'y = \beta_{\text{ols}}.$$

IV Model Assumptions

- (1), (2), (3) were replaced with (7).

$$\mathbb{E}[x_i|z_i] = 0.$$

$$\mathbb{E}[z_i\epsilon_i] = \mathbb{E}[\mathbb{E}[z_i\epsilon_i|z_i]] = \mathbb{E}[z_i\mathbb{E}[\epsilon_i]] = 0.$$

$$\mathbb{E}[z_i(y_i - x_i'\beta)] = 0.$$

(In sample),

$$\frac{1}{n}\sum_{i=1}^{n} z_i'(y_i - x_i'\beta_{IV}) = 0.$$

$$\sum_{i=1}^{n} z_i y_i - \left(\sum_{i=1}^{n} z_i x_i'\right)\beta_{IV} = 0.$$

$$\left[\sum_{i=1}^{n} z_i x_i'\right]\beta_{IV} = \sum_{i=1}^{n} z_i y_i.$$

If $L = k$, then

$$\beta_{IV} = \left(\sum_{i=1}^{n} z_i x_i'\right)^{-1}\left(\sum_{i=1}^{n} z_i y_i\right).$$

$$\beta_{IV} = (Z'X)^{-1}Z'y.$$

$$\beta_{OLS} = (X'X)^{-1}X'y.$$

WTS: Consistency
When $L = k$, $\mathbb{E}[z_i x_i'] = Q_{ZX}$, and:

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'y.$$

$$= (Z'X)^{-1}Z'(X\beta + \epsilon).$$

$$= \beta + (Z'X)^{-1}Z'\epsilon.$$

$$= \beta + \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i \epsilon_i\right).$$

$$\xrightarrow{p} \mathbb{E}[z_i x_i'] = Q_{ZX}, \quad \text{for using WLLN}.$$

$$\Rightarrow \hat{\beta}_{IV} \xrightarrow{p} \beta + (\mathbb{E}[z_i x_i'])^{-1}\mathbb{E}[z_i\epsilon_i].$$

$$\mathbb{E}[z_i\epsilon_i] = \mathbb{E}[\mathbb{E}[z_i\epsilon_i|z_i]] = \mathbb{E}[z_i\mathbb{E}[\epsilon_i|z_i]].$$

$$= \mathbb{E}[z_i \cdot 0] = 0.$$

$$\Rightarrow \hat{\beta}_{IV} \xrightarrow{p} \beta.$$

IV estimator is consistent.
WTS: Asymptotic normality proof

$$\hat{\beta}_{IV} - \beta = \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i \epsilon_i\right).$$

By CLT,

$$\sqrt{n}(\hat{\beta}_{IV} - \beta) = \left[\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right]^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i \epsilon_i\right).$$

$$\xrightarrow{p} Q_{ZX}$$

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} z_i \epsilon_i\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} z_i \epsilon_i - \mathbb{E}[z_i \epsilon_i]\right).$$

$$\xrightarrow{d} N(0, \sigma^2 Q_{ZZ}).$$

$$\mathrm{Var}(z_i \epsilon_i) = \mathbb{E}[z_i \epsilon_i - 0](z_i \epsilon_i - 0)'.$$

$$= \mathbb{E}[z_i \epsilon_i \epsilon_i' z_i'] = \mathbb{E}[\epsilon_i^2 z_i z_i'].$$

$$= \mathbb{E}[\mathbb{E}[\epsilon_i^2 | z_i] z_i z_i'].$$

$$= \sigma^2 \mathbb{E}[z_i z_i'] = \sigma^2 Q_{ZZ}.$$

By Slutsky's theorem,

$$\sqrt{n}(\hat{\beta}_{IV} - \beta) \to dN(0, \sigma^2 Q_{ZX}^{-1} Q_{ZZ} Q_{ZX}^{-1}).$$

Consistency.
But IV is biased:

$$\hat{\beta}_{IV} = \beta + (Z'X)^{-1} Z' \epsilon.$$

$$\mathbb{E}[\hat{\beta}_{IV} | X, Z] = \beta + (Z'X)^{-1} Z' \mathbb{E}[\epsilon | X, Z] \neq \beta.$$

$$\hat{\beta}_{IV} = (Z'X)^{-1} Z' y.$$

Matrix dimensions:

$$Z : n \times L, \quad Z' : L \times n, \quad X : n \times k.$$

$$L > k.$$

When $L > k$:

$$X \to Z \quad \text{projection.}$$

$$P_Z = Z(Z'Z)^{-1} Z'.$$

$$= ZCZ'Z'.$$

L L Z Z . .

$$\hat{X} = P_Z X.$$

$$\hat{X} = Z(Z'Z)^{-1}Z'X.$$

$$L \times L, \quad L \times n, \quad L \times k.$$

$$\hat{\beta}_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y.$$

$$= (X'P_Z X)^{-1}X'P_Z y.$$

Replaced the Z.

Question: Does the instrumental variable z need to be uncorrelated with the dependent variable y?

No!

- The instrumental variable z affects the dependent variable y through the endogenous variable x:
$$z \to x \to y$$
- The instrumental variable z does not directly affect the dependent variable y:
$$\text{cov}(z, y|x) = 0$$
- The instrumental variable z **can and must** influence the dependent variable y **only through** the endogenous variable x.

Suppose that there is a set of instrumental variables $Z = (Z_0 \quad Z_1 \quad \ldots Z_K)$ that meet the following condition:

1. plim $n^{-1}Z'X = Q_{ZX}$ (non-singular)
2. plim $n^{-1}Z'Z = Q_{ZZ}$ (positive definite)
3. plim $n^{-1}Z'u = 0$

$$Y = X\beta + u \Rightarrow Z'Y = Z'X\beta + Z'u$$

Let $\tilde{\beta}$ be an estimator of β. Then we have:

$$Z'Y = Z'X\tilde{\beta} + Z'\tilde{u} \Rightarrow Z\tilde{U} =$$

$$Z'(Y - X\tilde{\beta}) \Rightarrow \tilde{u} = Y - X\tilde{\beta}$$

$$(Z'\tilde{u})(Z'\tilde{u}) = (Z'Y - Z'X\tilde{\beta})'(Z'Y - Z'X\tilde{\beta})$$

$$= Y'Z'ZY - 2\tilde{\beta}'X'Z'ZY + \tilde{\beta}'X'Z'Z'X\tilde{\beta}$$

$$\frac{\partial(Z'\tilde{u})(Z'\tilde{u})}{\partial\tilde{\beta}} = -2X'Z'ZY + 2X'Z'Z'X\tilde{\beta} = 0$$

hence $X'Z'ZY = X'Z'Z'X\tilde{\beta}$. Then premultiplying by $(X'Z)^{-1}$ leads to

$$\tilde{\beta}^{IV} = (Z'X)^{-1}Z'Y$$

We further have:

$$\tilde{\beta}^{IV} = (Z'X)^{-1}Z'(X\beta + u)$$

$$= \beta + (Z'X)^{-1}Z'u$$

$$\text{plim } \tilde{\beta}^{IV} = \beta + \left[\text{plim}\left(\frac{Z'X}{n}\right)\right]^{-1} \cdot \text{plim}\frac{Z'u}{n}$$

$$= \beta + Q_{ZX}^{-1} \cdot 0 = \beta$$

Therefore $\tilde{\beta}^{IV}$ is consistent.

1.3 Two-Stage Least Squares (2SLS)