

Lecture 19: Instrumental Variables

Economics 326 — Introduction to Econometrics II

Vadim Marmer, UBC

Endogeneity

- In the linear regression model,

$$Y_i = \beta_0 + \beta_1 X_i + U_i,$$

the condition for consistent estimation of β_1 by OLS is that X is **exogenous**:

$$\text{Cov}(X_i, U_i) = 0.$$

- When $\text{Cov}(X_i, U_i) \neq 0$, we say that the regressor X is **endogenous**.
- When the regressor is **endogenous**, the OLS estimator is **inconsistent**:

$$\begin{aligned}\hat{\beta}_{1,n} - \beta_1 &= \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n) U_i}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2} \\ &\rightarrow_p \frac{\text{Cov}(X_i, U_i)}{\text{Var}(X_i)} \neq 0.\end{aligned}$$

Consequences of endogeneity

- The causal effect of X on Y is not estimated consistently:

$$\hat{\beta}_{1,n} \rightarrow_p \beta_1 + \frac{\text{Cov}(X_i, U_i)}{\text{Var}(X_i)}.$$

The effect can be over- or underestimated depending on the sign of $\text{Cov}(X_i, U_i)$.

- Tests and confidence intervals are invalid.

Sources of endogeneity

- Several possible sources of endogeneity:
 1. Omitted explanatory variables.
 2. Simultaneity.
 3. Errors in variables.
- All result in regressors correlated with the errors.

Omitted explanatory variables

- Suppose that the true model is

$$\ln \text{Wage}_i = \beta_0 + \beta_1 \text{Education}_i + \beta_2 \text{Ability}_i + V_i,$$

where V_i is uncorrelated with Education and Ability.

- Since Ability is unobservable, the econometrician regresses ln Wage against Education, and β_2 Ability goes into the error:

$$\begin{aligned}\ln \text{Wage}_i &= \beta_0 + \beta_1 \text{Education}_i + U_i, \\ U_i &= \beta_2 \text{Ability}_i + V_i.\end{aligned}$$

- Education is correlated with Ability: we can expect that $\text{Cov}(\text{Education}_i, \text{Ability}_i) > 0$, $\beta_2 > 0$, and therefore

$$\text{Cov}(\text{Education}_i, U_i) > 0.$$

Thus, OLS will overestimate the return to education.

Simultaneity

- Consider a **demand-supply** system:

$$\begin{aligned}\text{Demand: } Q^d &= \beta_0^d + \beta_1^d P + U^d, \\ \text{Supply: } Q^s &= \beta_0^s + \beta_1^s P + U^s,\end{aligned}$$

where Q^d = quantity demanded, Q^s = quantity supplied, P = price.

- The quantity and price are determined **simultaneously** in the equilibrium:

$$Q^d = Q^s = Q.$$

- Q^d and Q^s are not observed separately; we observe only the equilibrium values Q .

Simultaneity

- Solving for P :

$$\begin{aligned}Q^d &= Q^s \\ \beta_0^d + \beta_1^d P + U^d &= \beta_0^s + \beta_1^s P + U^s,\end{aligned}$$

so

$$P = -\frac{\beta_0^d - \beta_0^s}{\beta_1^d - \beta_1^s} - \frac{U^d - U^s}{\beta_1^d - \beta_1^s}.$$

- Thus,

$$\text{Cov}(P, U^d) \neq 0 \text{ and } \text{Cov}(P, U^s) \neq 0.$$

The demand-supply equations cannot be estimated by OLS.

Simultaneity

- Consider a labour supply model for married women:

$$\text{Hours}_i = \beta_0 + \beta_1 \text{Children}_i + \text{Other Factors} + U_i,$$

where Hours = hours of work, Children = number of children.

- It is reasonable to assume that women decide **simultaneously** how much time to devote to career and family.
- Thus, while we may be mainly interested in the effect of family size on labour supply, there is another equation:

$$\text{Children}_i = \gamma_0 + \gamma_1 \text{Hours}_i + \text{Other Factors} + V_i,$$

and Children and Hours are determined **simultaneously** in an equilibrium.

- As a result, $\text{Cov}(\text{Children}_i, U_i) \neq 0$, and the effect of family size cannot be estimated by OLS.

Errors in variables

- Consider a model:

$$Y_i = \beta_0 + \beta_1 X_i^* + V_i,$$

where X_i^* is the “true” regressor.

- Suppose that X_i^* is not directly observable. Instead, we observe X_i that measures X_i^* with an error ε_i :

$$X_i = X_i^* + \varepsilon_i.$$

- Since X_i^* is unobservable, the econometrician has to regress Y_i against X_i .

Errors in variables

- The model for Y_i as a function of X_i can be written as

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 (X_i - \varepsilon_i) + V_i \\ &= \beta_0 + \beta_1 X_i + V_i - \beta_1 \varepsilon_i, \end{aligned}$$

or

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + U_i, \\ U_i &= V_i - \beta_1 \varepsilon_i. \end{aligned}$$

Errors in variables

- We can assume that

$$\text{Cov}(X_i^*, V_i) = \text{Cov}(X_i^*, \varepsilon_i) = \text{Cov}(\varepsilon_i, V_i) = 0.$$

- However,

$$\begin{aligned} \text{Cov}(X_i, U_i) &= \text{Cov}(X_i^* + \varepsilon_i, V_i - \beta_1 \varepsilon_i) \\ &= \text{Cov}(X_i^*, V_i) - \beta_1 \text{Cov}(X_i^*, \varepsilon_i) \\ &\quad + \text{Cov}(\varepsilon_i, V_i) - \beta_1 \text{Cov}(\varepsilon_i, \varepsilon_i) \\ &= -\beta_1 \text{Var}(\varepsilon_i) \neq 0. \end{aligned}$$

- Thus, X_i is **endogenous** and β_1 cannot be estimated by OLS.

Instrumental variable (IV)

- Consider

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + U_i, \\ \text{Cov}(X_i, U_i) &\neq 0. \end{aligned}$$

- Suppose that the econometrician observes another variable Z_i , called the **instrumental variable**, that satisfies the following conditions:

- The IV is **exogenous**: $\text{Cov}(Z_i, U_i) = 0$.
- The IV **determines** the endogenous regressor: $\text{Cov}(Z_i, X_i) \neq 0$.

- When an IV satisfying those conditions is available, it allows us to estimate the effect of X on Y consistently.

IV regression

- Consider the **IV estimator** of β_1 :

$$\hat{\beta}_{1,n}^{IV} = \frac{\sum_{i=1}^n (Z_i - \bar{Z}_n) Y_i}{\sum_{i=1}^n (Z_i - \bar{Z}_n) X_i}.$$

- Substituting $Y_i = \beta_0 + \beta_1 X_i + U_i$:

$$\begin{aligned}\hat{\beta}_{1,n}^{IV} &= \frac{\sum_{i=1}^n (Z_i - \bar{Z}_n) (\beta_0 + \beta_1 X_i + U_i)}{\sum_{i=1}^n (Z_i - \bar{Z}_n) X_i} \\ &= \beta_1 + \frac{\sum_{i=1}^n (Z_i - \bar{Z}_n) U_i}{\sum_{i=1}^n (Z_i - \bar{Z}_n) X_i}.\end{aligned}$$

Consistency of the IV estimator

- The IV conditions:
 - Exogeneity:** $\text{Cov}(Z_i, U_i) = 0$.
 - Relevance:** $\text{Cov}(Z_i, X_i) \neq 0$.

- By the LLN:

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}_n) U_i &\rightarrow_p \text{Cov}(Z_i, U_i), \\ \frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}_n) X_i &\rightarrow_p \text{Cov}(Z_i, X_i).\end{aligned}$$

- The IV estimator is consistent:

$$\begin{aligned}\hat{\beta}_{1,n}^{IV} &= \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}_n) U_i}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}_n) X_i} \\ &\rightarrow_p \beta_1 + \frac{0}{\text{Cov}(Z_i, X_i)} = \beta_1.\end{aligned}$$

Natural experiments

- Theoretically, the causal effect can be estimated from **controlled experiments**:
 - To estimate the return to education, select a random sample of children, **randomly** assign how many years of education they should have, and measure their income several years after graduation.
 - To estimate the effect of family size on labor supply, select a random sample of parents and **randomly** assign how many children they should have, and measure their labor market outcomes.
- Such an approach is infeasible due to high cost and/or ethical reasons.
- Natural experiments:** use the random variation in the variable of interest to estimate the causal effect.

Example: Compulsory schooling laws

- Angrist and Krueger (1991, *QJE*) suggested using school start age policy to estimate β_1 in

$$\ln \text{Wage}_i = \beta_0 + \beta_1 \text{Education}_i + \beta_2 \text{Ability}_i + V_i.$$

- We need an IV Z such that $\text{Cov}(\text{Ability}_i, Z_i) = 0$ and $\text{Cov}(\text{Education}_i, Z_i) \neq 0$.
- They argue that due to compulsory schooling laws, the **season of birth** satisfies the IV conditions:

- A child must attend school until reaching a certain drop-out age.
- Students born in the first quarter reach the legal drop-out age before classmates born later in the year.
- The quarter-of-birth dummy is correlated with education.
- The quarter of birth is uncorrelated with ability.

Example: Sibling-sex composition

- Angrist and Evans (1998, *AER*) argue that parents' preferences for a mixed sibling-sex composition can be used to estimate β_1 in

$$\text{Hours}_i = \beta_0 + \beta_1 \text{Children}_i + \dots + U_i.$$

- We need an IV Z such that $\text{Cov}(U_i, Z_i) = 0$ and $\text{Cov}(\text{Children}_i, Z_i) \neq 0$.
- Consider a dummy variable equal to one if the sex of the second child matches the sex of the first child.
- If parents prefer a mixed sibling-sex composition, they are more likely to have another child if their first two children are of the same sex.
- The same-sex dummy is correlated with the number of children.
- Since sex mix is randomly determined, the same-sex dummy is exogenous.

Asymptotic distribution of the IV estimator

- Write

$$\begin{aligned} \sqrt{n}(\hat{\beta}_{1,n}^{IV} - \beta_1) &= \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n (Z_i - \bar{Z}_n) U_i}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}_n) X_i} \\ &\rightarrow_d \frac{N(0, \text{E}[(Z_i - \text{E}[Z_i])^2 U_i^2])}{\text{Cov}(Z_i, X_i)}. \end{aligned}$$

- Thus,

$$\begin{aligned} \sqrt{n}(\hat{\beta}_{1,n}^{IV} - \beta_1) &\rightarrow_d N(0, V^{IV}), \text{ where} \\ V^{IV} &= \frac{\text{E}[(Z_i - \text{E}[Z_i])^2 U_i^2]}{(\text{Cov}(Z_i, X_i))^2}. \end{aligned}$$

Variance estimation

- Let $\hat{\beta}_{0,n}^{IV} = \bar{Y}_n - \hat{\beta}_{1,n}^{IV} \cdot \bar{X}_n$.
- Let $\hat{U}_i = Y_i - \hat{\beta}_{0,n}^{IV} - \hat{\beta}_{1,n}^{IV} X_i$.
- Estimate V^{IV} by

$$\hat{V}_n^{IV} = \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 \hat{U}_i^2}{\left(\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}_n) X_i\right)^2}.$$

- In finite samples, we use the approximation:

$$\hat{\beta}_{1,n}^{IV} \overset{a}{\sim} N\left(\beta_1, \frac{\hat{V}_n^{IV}}{n}\right).$$