

Economics 326
Methods of Empirical Research in Economics
Lecture 20: Instrumental variable estimation

Vadim Marmer
University of British Columbia

May 5, 2010

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:

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

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 under estimated depending on the sign of $\text{Cov}(X_i, U_i)$.

- ▶ Tests and confidence intervals are invalid.

Sources of endogeneity

There are 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

$$\ln Wage_i = \beta_0 + \beta_1 Education_i + \beta_2 Ability_i + V_i,$$

where V_i is uncorrelated with *Education* and *Ability*.

- ▶ Since *Ability* is unobservable, the econometrician regresses $\ln Wage$ against *Education*, and $\beta_2 Ability$ "goes" into the error part:

$$\begin{aligned}\ln Wage_i &= \beta_0 + \beta_1 Education_i + U_i, \\ U_i &= \beta_2 Ability_i + V_i.\end{aligned}$$

- ▶ *Education* is correlated with *Ability*: we can expect that $Cov(Education_i, Ability_i) > 0$, $\beta_2 > 0$, and therefore

$$Cov(Education_i, U_i) > 0.$$

Thus, OLS will overestimate the return to education.

Simultaneity

- ▶ Consider the following 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:

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

- ▶ Note that Q^d and Q^s are not observed separately, we observe only the equilibrium values Q .

Simultaneity

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

- ▶ Solving for P , we obtain

$$0 = (\beta_0^d - \beta_0^s) + (\beta_1^d - \beta_1^s) P + (U^d - U^s),$$

or

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

- ▶ 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 the following labour supply model for married women:

$$Hours_i = \beta_0 + \beta_1 Children_i + \text{Other Factors} + U_i,$$

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:

$$Children_i = \gamma_0 + \gamma_1 Hours_i + \text{Other Factors} + V_i,$$

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

- ▶ As a result, $Cov(Children_i, U_i) \neq 0$, and the effect of family size cannot be estimated by OLS.

Errors in variables

- ▶ Consider the following 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

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

- ▶ 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

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

$$U_i = V_i - \beta_1 \varepsilon_i,$$

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

- ▶ 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

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

- ▶ Suppose that in addition, the econometrician observes another variable Z_i , called the instrumental variable, that satisfies the following conditions:
 1. The IV is exogenous: $\text{Cov}(Z_i, U_i) = 0$.
 2. The IV determines the endogenous regressor:
 $\text{Cov}(Z_i, X_i) \neq 0$.
- ▶ When an IV variable satisfying those conditions is available, it allows us to estimate the effect of X on Y consistently.

IV regression

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

- ▶ Consider the following IV estimator of β_1 :

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

- ▶ Write

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

Consistency of the IV estimator

$$\text{Cov}(Z_i, U_i) = 0, \quad (1)$$

$$\text{Cov}(Z_i, X_i) \neq 0. \quad (2)$$

- ▶ Using the LLN (and under some additional technical conditions), (1) implies that

$$\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}_n) U_i \rightarrow_p \text{Cov}(Z_i, U_i),$$

and (2) implies that

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

- ▶ The IV estimator is consistent if (1) and (2) are true:

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

Examples: 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 the 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 a 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 and return to education

- ▶ Angrist and Krueger, 1991, *QJE*, suggested using school start age policy to estimate β_1 in
$$\ln Wage_i = \beta_0 + \beta_1 Education_i + \beta_2 Ability_i + V_i.$$
- ▶ We need to find an IV variable Z such that
$$Cov(Ability_i, Z_i) = 0 \text{ and } Cov(Education_i, Z_i) \neq 0.$$
- ▶ They argue that due to compulsory schooling laws, the season of birth variable satisfies the IV conditions:
 - ▶ A child has to attend the school until he reaches a certain drop-out age.
 - ▶ Students born in the first quarter of the year, reach the legal drop-out age before their classmates who were born later in the year.
 - ▶ The quarter of birth dummy variable is correlated with education.
 - ▶ The quarter of birth is uncorrelated with ability.

Example: Sibling-sex composition and labor supply

- ▶ Angrist and Evans, 1998, *AER*, argue that the parents' preferences for a mixed sibling-sex composition can be used to estimate β_1 in $Hours_i = \beta_0 + \beta_1 Children_i + \dots + U_i$.
- ▶ We need to find an IV Z such that $Cov(U_i, Z_i) = 0$ and $Cov(Children_i, Z_i) \neq 0$.
- ▶ Consider a dummy variable that takes on the value one if the sex of the second child matches the sex of the first child.
- ▶ If the 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.

The asymptotic distribution of the IV estimator

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

► Write

$$\sqrt{n}(\hat{\beta}_{1,n}^{IV} - \beta_1) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n (Z_i - \bar{Z}_n) U_i}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z}_n) X_i} \rightarrow_d \frac{N\left(0, E\left((Z_i - EZ_i)^2 U_i^2\right)\right)}{\text{Cov}(Z_i, X_i)}.$$

► Thus,

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

Variance estimation

$$\sqrt{n} \left(\hat{\beta}_{1,n}^{IV} - \beta_1 \right) \rightarrow_d N \left(0, V^{IV} \right), \text{ where } V^{IV} = \frac{E \left((Z_i - EZ_i)^2 U_i^2 \right)}{\left(\text{Cov} (Z_i, X_i) \right)^2}.$$

- ▶ 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}

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

- ▶ In finite samples, we use the following approximation:

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