

Instrumental Variables

Erick F Molina

Summer, 2024

ITAM

This presentation is based on the references listed in the course syllabus and on notes by Arturo Aguilar, Mauricio Romero, Laura Schechter, Shaun McRae, Kirill Borusyak, among others. Any remaining errors are my own.

When do we use instrumental variables?

- When we face an endogeneity problem.
- This occurs when an explanatory variable is correlated with the error term in a regression model.
- This correlation violates one of the key OLS assumptions.
- As a result, OLS estimators are biased and inconsistent.

When do we use instrumental variables?

- In general, we use instrumental variables to mitigate the following issues:
 - Omitted variable bias (OVB).
 - Measurement error.
 - Simultaneity bias.
 - Reverse causality.
 - Imperfect compliance.

What conditions must an instrument satisfy?

- Consider a dependent variable Y , an explanatory variable X , and an instrument Z under the model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

where $\text{Cov}(X, \varepsilon) \neq 0$.

- The instrument Z must satisfy two conditions:
 - Relevance:** $\text{Cov}(X, Z) \neq 0$. The instrument is correlated with the endogenous regressor.
 - Exclusion:** $\text{Cov}(Z, \varepsilon) = 0$. The instrument affects the outcome only through the endogenous regressor.

How do we obtain IV estimators?

- Consider:

$$Y = \beta_0 + \beta_1 X + \underbrace{\gamma A + \nu}_{\varepsilon} \quad (2)$$

where Y is log wages, X is years of schooling, A is unobserved ability, and ν is an idiosyncratic shock.

- We can think of the observed error ε as being composed of A and ν .
- Problem: X is correlated with the error ε .
- Suppose there is a variable Z correlated with X , and uncorrelated with A and ν .

How do we obtain IV estimators?

$$\begin{aligned}\text{Cov}(Y, Z) &= \text{Cov}(\beta_0 + \beta_1 X + \gamma A + \nu, Z) \\ &= \beta_1 \text{Cov}(X, Z) + \gamma \text{Cov}(A, Z) + \text{Cov}(\nu, Z)\end{aligned}$$

If the instrument is relevant, $\text{Cov}(X, Z) \neq 0$:

$$\frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} = \beta_1 + \gamma \frac{\text{Cov}(A, Z)}{\text{Cov}(X, Z)} + \frac{\text{Cov}(\nu, Z)}{\text{Cov}(X, Z)}$$

If it also satisfies the exclusion restriction, $\text{Cov}(A, Z) = \text{Cov}(\nu, Z) = 0$:

$$\frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} = \beta_1$$

Two Stage Least Squares

- To estimate a model using instrumental variables, we typically use the Two-Stage Least Squares (2SLS) estimator. Suppose we have the following equations:
 - Structural (causal) model: $Y_i = \beta_0 + \beta_1 X_i + \eta_i$
 - First stage: $X_i = \gamma_0 + \gamma_1 Z_i + \zeta_i$
 - Second stage: $Y_i = \beta_0 + \beta_1 \hat{X}_i + \nu_i$
 - Reduced form: $Y_i = \psi_0 + \psi_1 Z_i + \varepsilon_i$
- To obtain the 2SLS estimator, we first estimate the first stage and compute $\hat{X}_i$. We then use this prediction in the second stage. This yields our 2SLS estimate.

2SLS estimator

- Recall that:

$$\hat{\beta} = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}$$

- If the instrument Z is relevant:

$$\beta_{IV} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)}$$

- Then we can write:

$$\beta_{IV} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} = \frac{\frac{\text{Cov}(Y, Z)}{\text{Var}(Z)}}{\frac{\text{Cov}(X, Z)}{\text{Var}(Z)}}$$

- Therefore...

$$\hat{\beta}_{2SLS} = \frac{\hat{\psi}_1}{\hat{\gamma}_1}$$

- Another way to see it (from the second stage):

$$\hat{\beta}_{2SLS} = \frac{\text{Cov}(\hat{X}, Y)}{\text{Var}(\hat{X})} = \frac{\text{Cov}(\gamma_0 + \gamma_1 Z, Y)}{\text{Var}(\gamma_0 + \gamma_1 Z)} = \frac{\gamma_1 \text{Cov}(Z, Y)}{\gamma_1^2 \text{Var}(Z)} = \frac{\text{Cov}(Z, Y)}{\gamma_1 \text{Var}(Z)}$$

- Rewriting γ_1 as:

$$\gamma_1 = \frac{\text{Cov}(Z, X)}{\text{Var}(Z)}$$

- Then:

$$\hat{\beta}_{2SLS} = \frac{\text{Cov}(Z, Y)}{\frac{\text{Cov}(Z, X)}{\text{Var}(Z)}} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, X)}$$

Example of an IV table

	OLS	First Stage	Second Stage
Years of Schooling	0.09*** (0.00)		0.26*** (0.03)
Near 4Y College		0.60*** (0.08)	
Observations	3,010	3,010	3,010
Dep. Var. Mean	6.262	13.263	6.262
Controls	✓	✓	✓

Some practical tips...

1. Look at the reduced form.
2. Look at the first stage: this is where we assess whether the instrument satisfies relevance.
3. Always report the simple OLS regression, even though we know X has an endogeneity problem.
4. Remember that **no** instrument is perfect.

Local Average Treatment Effects

- Concept introduced by Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996).
- Uses the potential outcomes framework.
- The endogenous variable X and the instrument Z are binary.

- Suppose we want to estimate the causal effect of years of schooling on an individual's wage: $C = Y(1) - Y(0)$.
- The average treatment effect (ATE) is:

$$ATE = \mathbb{E}[Y(1) - Y(0)]$$

- We know years of schooling may be endogenous.
- For this reason, Card (1995) used distance from home to college as an instrument. He constructed a dummy indicating whether an individual lived near a college.

LATE: Types of individuals

Recall that in practice, the relationship between distance and schooling is not the same for everyone, so we have:

	$X(0) = 0$	$X(0) = 1$
$X(1) = 0$	Never Takers	Defiers
$X(1) = 1$	Compliers	Always Takers

- Never Takers: never attend school.
- Always Takers: always attend school.
- Compliers: attend only if they live near a school.
- Defiers: the opposite of compliers (typically assumed not to exist).

Local Average Treatment Effect (LATE)

- LATE is the average causal effect for compliers.

- Formula:

$$\text{LATE} = \frac{\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]}{\mathbb{E}[X|Z = 1] - \mathbb{E}[X|Z = 0]}$$

- In our example, LATE is the causal effect of schooling on wages for those whose schooling is influenced by proximity.

LATE: Interpretation and limitations

- Standard IV identifies a LATE.
- LATE can vary depending on the instrument used.
- LATE applies only to the complier population.