ECON3389 Econometric Methods

Module 1
Omitted Variables

Alberto Cappello

Department of Economics, Boston College

Spring 2025

Motivating Example

Research Question: What is the effect of a drug (X_1) on blood-pressure (Y)?
Challenge:

- Other variables like Diet and Stress (X_2) also affect blood-pressure .
- Omitting those variables from the regression model leads to biased estimates of the effect of the drug on blood-pressure.

True Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Omitted Model:

$$Y = \beta_0 + \beta_1 X_1 + u, \quad u = \beta_2 X_2 + \epsilon$$

What is Omitted Variable Bias?

Definition: Omitted Variable Bias (OVB) occurs when a relevant variable is left out of a regression model, leading to biased and inconsistent estimates of the regression coefficients.

Key Points:

- The omitted variable must be correlated with both the independent variable and the dependent variable.
- Failure to include such a variable distorts the estimated relationship between the included independent variables and the dependent variable.

Example:

- Estimating the effect of a drug on blood-pressure without accounting for other variables that might affect both blood-pressure and the decision to take the drug (omitted variable).

The OVB Problem

Suppose the true model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

If X_2 is omitted, the estimated model becomes:

$$Y = \beta_0 + \beta_1 X_1 + u$$

Where:

$$u = \beta_2 X_2 + \epsilon$$

Impact:

- The error term (u) is correlated with X_1, violating the assumption of zero correlation between regressors and errors.
- This leads to biased estimates of β_1.

Conditions for OVB

Omitted Variable Bias arises when:

1. The omitted variable (X_2) is correlated with the included independent variable (X_1):

$$\text{Cov}(X_1, X_2) \neq 0$$

2. The omitted variable (X_2) affects the dependent variable (Y):

$$\beta_2 \neq 0$$

If either condition is not met, OVB does not occur.

Derivation of Omitted Variable Bias

Suppose the true model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

If X_2 is omitted, the estimated model becomes:

$$Y = \beta_0 + \beta_1 X_1 + u, \quad \text{where } u = \beta_2 X_2 + \epsilon$$

OLS Estimate of β_1:

$$\hat{\beta}_1 = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)}$$

Substitute Y from the true model:

$$\text{Cov}(X_1, Y) = \text{Cov}(X_1, \beta_0 + \beta_1 X_1 + u) = \beta_1 \text{Var}(X_1) + \beta_2 \text{Cov}(X_1, u)$$

Substitute into $\hat{\beta}_1$:

$$\hat{\beta}_1 = \beta_1 + \underbrace{\frac{\text{Cov}(X_1, u)}{\text{Var}(X_1)}}_{\neq 0} \neq \beta_1$$

Instrumental Variables (IV): Motivation

Why IV?

- OVB makes ordinary least squares (OLS) estimates biased and inconsistent.
- Instrumental Variables (IV) provide a solution when an omitted variable is correlated with both the independent variable and the dependent variable.

Core Idea: Use an external variable (**instrument**) to isolate the variation in the independent variable that is not correlated with the omitted variable.

Example:

- Studying the effect of a drug on blood pressure. Use proximity to a pharmacy as an instrument for drug.

Instrumental Variables: Key Assumptions

For an instrument Z to be valid, it must satisfy:

1. **Relevance:** The instrument must be correlated with the independent variable.

$$\text{Cov}(Z, X) \neq 0$$

2. **Exogeneity:** The instrument must not be correlated with the error term.

$$\text{Cov}(Z, \epsilon) = 0$$

Implications:

- Relevance ensures the instrument has predictive power for the independent variable.
- Exogeneity ensures the instrument is not confounded by omitted variables.

Two-Stage Least Squares (2SLS)

How IV Works: Instrumental Variables regression is often implemented using **Two-Stage Least Squares (2SLS)**:

1. **First Stage:** Regress the endogenous variable (X) on the instrument (Z):

   $$X = \pi_0 + \pi_1 Z + v$$

   Obtain the predicted values $\hat{X}$.

2. **Second Stage:** Regress the dependent variable (Y) on $\hat{X}$:

   $$Y = \beta_0 + \beta_1 \hat{X} + \epsilon$$

Result: $\hat{\beta}_1$ is consistent, even when X is endogenous.

IV Example: Drug and Blood-Pressure

Scenario:

- Goal: Estimate the effect of a drug (X) on blood-pressure (Y).
- Problem: Stress and Diet (U) is omitted and correlated with both blood-pressure and drug.
- Instrument: Proximity to a pharmacy (Z).

Steps:

1. Regress drug (X) on proximity to a pharmacy (Z) to obtain $\hat{X}$.
2. Regress blood-pressure (Y) on $\hat{X}$ to estimate the causal effect.

Key: Proximity to a pharmacy affects blood-pressure only through its effect on drug.

Challenges with Instrumental Variables

1. **Weak Instruments:**
    - Instruments with low correlation with X lead to imprecise estimates.
    - Check using the F-statistic in the first stage ($F > 10$ is recommended).
2. **Validity of Exogeneity:**
    - Instruments may themselves be correlated with the omitted variable.
    - Requires careful reasoning and robustness checks.
3. **Overidentification:**
    - When multiple instruments are used, test overidentifying restrictions to ensure all instruments are valid.

Conclusion

Omitted Variable Bias:

- Occurs when a relevant variable is left out of a model.
- Leads to biased and inconsistent estimates.

Instrumental Variables:

- IV solves OVB by using instruments to isolate exogenous variation.
- Requires careful choice of instruments to ensure relevance and exogeneity.

Takeaway: Instrumental Variables are a powerful tool, but their validity depends on strong theoretical and empirical justification.