Omitted Variable Bias

Gov 2001: Quantitative Social Science Methods I

Week 10, Lecture 20

Spring 2026

For Today

Required Reading

- Angrist & Pischke, §3.2.2 (pp. 59–68)
- Blackwell, Chapter 6 (omitted variables section)

Today: What happens when you leave out an important variable?

Roadmap

1. The omitted variable problem
2. The OVB formula
3. Direction of bias
4. Examples
5. Bad controls

Part I: The Omitted Variable Problem

The Setup

True model ("long regression"):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

What we estimate ("short regression"):

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + u_i$$

Question: What is the relationship between $\hat{\alpha}_1$ and β_1?
If X_2 affects Y and is correlated with X_1, we have **omitted variable bias**.

Why This Matters

We can't include every possible variable.

- Some variables are unobserved (ability, motivation)
- Some variables are hard to measure
- We might not know what variables matter

Key question: Does omitting a variable *bias* our estimate of β_1?
And if so, in which direction?

Part II: The OVB Formula

Deriving the OVB Formula

Long regression (true):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Short regression (estimated):

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + u_i$$

Auxiliary regression (X_2 on X_1):

$$X_{2i} = \delta_0 + \delta_1 X_{1i} + v_i$$

Here $\delta_1 = \text{Cov}(X_1, X_2)/\text{Var}(X_1)$.

Deriving the Formula

Substitute the auxiliary regression into the long regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2(\delta_0 + \delta_1 X_{1i} + \nu_i) + \varepsilon_i$$

Collect terms:

$$Y_i = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)X_{1i} + (\beta_2\nu_i + \varepsilon_i)$$

This looks like the short regression with:

$$\alpha_0 = \beta_0 + \beta_2\delta_0$$
$$\alpha_1 = \beta_1 + \beta_2\delta_1$$

The OVB Formula

Omitted Variable Bias Formula

$$\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \hat{\delta}_1$$

or equivalently:

$$\text{Short} = \text{Long} + (\text{Effect of omitted}) \times (\text{Relationship with included})$$

The bias:

$$\text{Bias} = \hat{\alpha}_1 - \hat{\beta}_1 = \hat{\beta}_2 \cdot \hat{\delta}_1$$

The Two Components of Bias

$$\text{Bias} = \underbrace{\beta_2}_{\text{Effect of } X_2 \text{ on } Y} \times \underbrace{\delta_1}_{\text{Relationship between } X_1 \text{ and } X_2}$$

For there to be bias, BOTH must be non-zero:

1. $\beta_2 \neq 0$: The omitted variable affects Y
2. $\delta_1 \neq 0$: The omitted variable is correlated with X_1

If either is zero, there is no bias!

When There's No Bias

Case 1: $\beta_2 = 0$
The omitted variable doesn't affect Y. No problem omitting it.

Case 2: $\delta_1 = 0$
The omitted variable is uncorrelated with X_1.
Even if it affects Y, it doesn't bias our estimate of β_1.

Omitting a variable is only a problem if it affects Y AND is correlated with X_1.

Part III: Direction of Bias

Determining the Direction of Bias

$$\text{Bias} = \beta_2 \times \delta_1$$

	$\delta_1 > 0$ (X_1, X_2 positively correlated)	$\delta_1 < 0$ (X_1, X_2 negatively correlated)
$\beta_2 > 0$ (Omitted raises Y)	Positive bias $\hat{\alpha}_1 > \beta_1$	Negative bias $\hat{\alpha}_1 < \beta_1$
$\beta_2 < 0$ (Omitted lowers Y)	Negative bias $\hat{\alpha}_1 < \beta_1$	Positive bias $\hat{\alpha}_1 > \beta_1$

Positive Bias: An Example

Research question: Effect of education on wages

Omitted variable: Ability

- $\beta_2 > 0$: Higher ability $\Rightarrow$ higher wages
- $\delta_1 > 0$: Higher ability $\Rightarrow$ more education (correlation)

Result: Positive bias.
Simple regression of wages on education **overstates** the true effect.
Part of what we attribute to education is actually ability.

Ability is a **confounder**—it affects both X_1 and Y.
Omitting it biases the estimated effect of education.

Negative Bias: Another Example

Research question: Effect of class size on test scores

Omitted variable: School resources (per-pupil spending)

- $\beta_2 > 0$: More resources $\Rightarrow$ higher test scores
- $\delta_1 < 0$: More resources $\Rightarrow$ smaller classes (negative correlation)

Result: $(+) \times (-) =$ Negative bias.
Simple regression might show small classes hurt scores—but this is confounded by resources.

Part IV: Using the OVB Formula

Quantifying the Bias

If you run both regressions, you can check:

$$\hat{\alpha}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \hat{\delta}_1$$

Example:

- Short regression: $\widehat{\text{Wage}} = 5.2 + 0.12 \cdot \text{Educ}$
- Long regression: $\widehat{\text{Wage}} = 4.8 + 0.08 \cdot \text{Educ} + 0.15 \cdot \text{Ability}$
- Auxiliary: $\widehat{\text{Ability}} = 2 + 0.27 \cdot \text{Educ}$

Check: $0.12 = 0.08 + 0.15 \times 0.27 = 0.08 + 0.04 = 0.12$ ✓

Reasoning About Omitted Variables

Often we can't observe X_2. But we can still reason about bias.

Ask yourself:

1. What's an important omitted variable?
2. Does it affect Y? (What sign is β_2?)
3. Is it correlated with X_1? (What sign is δ_1?)
4. What's the direction of bias?

This helps you interpret results even when you can't control for everything.

Bounding the True Effect

If you know the direction of bias:

- Positive bias: $\hat{\alpha}_1 > \beta_1$
  The true effect is **smaller** than the estimate.
- Negative bias: $\hat{\alpha}_1 < \beta_1$
  The true effect is **larger** than the estimate.

This gives you a **bound** on the true effect—useful even without observing the omitted variable.

Part V: Bad Controls

The “Bad Controls” Problem

Adding controls is not always good!

Rule: Don’t control for variables that are:

1. Affected by X_1 (post-treatment)
2. On the causal path from X_1 to Y (mediators)

Controlling for these can **introduce bias** where there was none, or **change the estimand** to something you don’t want.

Example: Controlling for a Mediator

Question: Effect of education on wages

$$
\begin{tikzcd}
\text{Education} \arrow[r] \arrow[rr, bend left] & \text{Occupation} \arrow[r] & \text{Wages}
\end{tikzcd}
$$

If you control for occupation:
You block part of the effect of education (the part working through occupation).
You'd only get the "direct" effect, not the total effect.

Example: Controlling for Post-Treatment Variable

Question: Effect of job training on wages

$$\text{Training} \longrightarrow \text{Employment} \longrightarrow \text{Wages}$$

If you control for employment status:
Training might increase wages *by getting people employed.*
Controlling for employment removes this effect.
Among the employed, training might show little effect—but that's misleading!

Example: Collider Bias

Question: Is fundraising ability correlated with candidate quality?

$$\begin{tikzcd}
\text{Candidate Quality} \arrow[dr] & & \text{Fundraising Ability} \arrow[dl] \\
& \text{Win Election} &
\end{tikzcd}$$

Quality and fundraising might be uncorrelated among all candidates.
But among winners, they might appear negatively correlated.
Lower-quality candidates who won must have compensated with more money.
Conditioning on a **collider** induces spurious correlation.

Guidance on Control Variables

Good controls:

- Variables that cause both X and Y (confounders)
- Pre-treatment covariates

Bad controls:

- Variables caused by X (post-treatment)
- Variables on the causal path from X to Y (mediators)
- Colliders (caused by both X and Y)

Think carefully about causal structure before adding controls!

Summary

Key results:

1. OVB formula: Short = Long + $\beta_2 \times \delta_1$
2. Bias requires **both**:
   - Omitted variable affects Y ($\beta_2 \neq 0$)
   - Omitted variable correlates with X_1 ($\delta_1 \neq 0$)
3. Direction of bias = $\text{sign}(\beta_2) \times \text{sign}(\delta_1)$
4. **Bad controls**: Don't control for mediators or post-treatment variables

Looking Ahead

Next week: Regression Extensions

- Interaction terms
- Nonlinear transformations (logs, polynomials)
- F-tests for joint hypotheses
- Matrix notation

OVB reasoning will continue to be important throughout the course.

Omitting a variable biases your estimate if
the variable affects Y AND correlates with X.

$$\text{Bias} = \beta_2 \times \delta_1$$

But be careful: controlling for the wrong variables
can also introduce bias.