Regression in Action: Bias, squares, and measurement error

Econ 140, Section 3

Jonathan Old

Roadmap

1. Recap
2. Selection bias in action: Omitted Variable Bias
3. Quadratic Terms
4. Measurement Error

Any questions?

… Remember – Every question is useful!

Recap

Recap: Selection bias

We saw that whenever we do a difference-in-means comparison (or a regression), we get:

Estimating the effect of iPads on grades

Let us start with a difference-in-means comparison:

$$\Delta = E[\text{Grade}_i|\text{iPad}_i = 1] - E[\text{Grade}_i|\text{iPad}_i = 0]$$

Add and subtract $E[\text{Grade}_i(0)|\text{iPad}_i = 1]$:

$$= E[\text{Grade}_i(1)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 1] + E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

Use properties of expectations:

$$= E[\text{Grade}_i(1) - \text{Grade}_i(0)|\text{iPad}_i = 1] + E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

$$= \text{ATT} + \text{Selection bias}$$

Selection bias: Students with and without iPad have different potential grades: **even if they both *had* iPads, they would be different**.

6

Causal Effect

+

Selection Bias

- We cannot observe them - so we can never be sure!
- Econometrics is all about uncertainty. You can always state that there are different possibilities and you cannot know for sure

Recap: How to think about Selection Bias

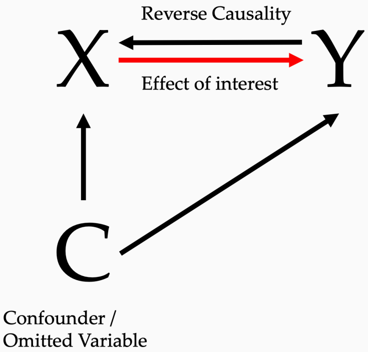

Figure 1: Selection bias

Selection bias in action: Omitted Variable Bias

Omitted Variable Bias (OVB)

Let's go through an example:

You are doing a summer internship with San-Francisco based tech company "Chirp", when your boss asks you to help evaluating the effect of "performance-based" or "incentive" pay on employees' job satisfaction. When you analyze the data, you realize that it's a mess. Until now, every manager could decide on the performance pay system in their team. Can you run a regression of job satisfaction on whether or not an individual works in a team that has a performance bonus in place?

You believe that:

$$\text{Satisfaction}_i = \beta_0 + \beta_1 \text{Performance pay}_i + \beta_2 \text{Manager quality}_i + u_i$$

$$\text{Manager quality}_i = \gamma_0 + \gamma_1 \text{Performance pay}_i + v_i$$

OVB Question

1. Run a "short" regression of satisfaction (S_i) on pay (P_i):

$$\mathrm{S}_i = \alpha_0 + \alpha_1 \mathrm{P}_i + e_i$$

OVB Question

1. Run a "short" regression of satisfaction (S_i) on pay (P_i):

$$S_i = \alpha_0 + \alpha_1 P_i + e_i$$

$$\alpha_1 = \frac{\text{Cov}(S_i, P_i)}{\text{Var}(P_i)}$$

2. Solve this expression with what we know

OVB Question

1. Run a "short" regression of satisfaction (S_i) on pay (P_i):

$$\mathrm{S}_i = \alpha_0 + \alpha_1 \mathrm{P}_i + e_i$$

$$\alpha_1 = \frac{\mathrm{Cov}(S_i, P_i)}{\mathrm{Var}(P_i)}$$

2. Solve this expression with what we know

$$\alpha_1 = \frac{\mathrm{Cov}(S_i, P_i)}{\mathrm{Var}(P_i)}$$

OVB Question

1. Run a "short" regression of satisfaction (S_i) on pay (P_i):

$$\mathsf{S}_i = \alpha_0 + \alpha_1 \mathsf{P}_i + e_i$$

$$\alpha_1 = \frac{\mathrm{Cov}(S_i, P_i)}{\mathrm{Var}(P_i)}$$

2. Solve this expression with what we know

$$\alpha_1 = \frac{\mathrm{Cov}(S_i, P_i)}{\mathrm{Var}(P_i)} = \frac{\mathrm{Cov}(\beta_0 + \beta_1 \mathsf{P}_i + \beta_2 \mathsf{M}_i + u_i, P_i)}{\mathrm{Var}(P_i)}$$

OVB Question

1. Run a "short" regression of satisfaction (S_i) on pay (P_i):

$$\text{S}_i = \alpha_0 + \alpha_1 \text{P}_i + e_i$$

$$\alpha_1 = \frac{\text{Cov}(S_i, P_i)}{\text{Var}(P_i)}$$

2. Solve this expression with what we know

$$\begin{aligned}\alpha_1 = \frac{\text{Cov}(S_i, P_i)}{\text{Var}(P_i)} &= \frac{\text{Cov}(\beta_0 + \beta_1 \text{P}_i + \beta_2 \text{M}_i + u_i, P_i)}{\text{Var}(P_i)} \\ &= \beta_1 \underbrace{\frac{\text{Cov}(P_i, P_i)}{\text{Var}(P_i)}}_{=1} + \beta_2 \underbrace{\frac{\text{Cov}(M_i, P_i)}{\text{Var}(P_i)}}_{=\gamma_1} + \underbrace{\frac{\text{Cov}(u_i, P_i)}{\text{Var}(P_i)}}_{=0, \text{ by assumption}}\end{aligned}$$

OVB Question

1. Run a "short" regression of satisfaction (S_i) on pay (P_i):

$$S_i = \alpha_0 + \alpha_1 P_i + e_i$$

$$\alpha_1 = \frac{Cov(S_i, P_i)}{Var(P_i)}$$

2. Solve this expression with what we know

$$\begin{aligned}
\alpha_1 = \frac{Cov(S_i, P_i)}{Var(P_i)} &= \frac{Cov(\beta_0 + \beta_1 P_i + \beta_2 M_i + u_i, P_i)}{Var(P_i)} \\
&= \beta_1 \underbrace{\frac{Cov(P_i, P_i)}{Var(P_i)}}_{=1} + \beta_2 \underbrace{\frac{Cov(M_i, P_i)}{Var(P_i)}}_{=\gamma_1} + \underbrace{\frac{Cov(u_i, P_i)}{Var(P_i)}}_{=0\text{, by assumption}} \\
&= \underbrace{\beta_1}_{\text{Actual effect}} + \underbrace{\beta_2 \gamma_1}_{\text{Omitted variable bias}}
\end{aligned}$$

The most important slide of this course (until the midterm)

We can summarize everything of OVB in three equations. Let Y_i be the outcome variable, X_i our regressor of interest, and Z_i the ”omitted” variable.

$$\begin{aligned}
&\text{[Long regression]} && Y_i = c_1 + \beta_L X_i + \delta Z_i + e_i \\
&\text{[Short regression]} && Y_i = c_2 + \beta_S X_i + u_i \\
&\text{[Auxiliary regression]} && Z_i = c_3 + \gamma X_i + v_i
\end{aligned}$$

Then, the **Omitted variable bias formula** states that:

$$\underbrace{\beta_S}_{\text{Short =}} = \underbrace{\beta_L}_{\text{Long +}} + \underbrace{\delta}_{\text{Omitted} \times} \cdot \underbrace{\gamma}_{\text{Included}}$$

We call $\delta\gamma$ the **omitted variable bias**. We can appraise the direction of the bias by multiplying our guesses for the signs of δ and γ.

Wait a minute!

We saw that the OLS coefficient in the short regression is:

$$\alpha_1 = \underbrace{\beta_1}_{\text{Actual effect}} + \underbrace{\beta_2\gamma_1}_{\text{Omitted variable bias}}$$

This looks awfully much like:

Estimating the effect of iPads on grades

Let us start with a difference-in-means comparison:

$$\Delta = E[\text{Grade}_i|\text{iPad}_i = 1] - E[\text{Grade}_i|\text{iPad}_i = 0]$$

Add and subtract $E[\text{Grade}_i(0)|\text{iPad}_i = 1]$:

$$= E[\text{Grade}_i(1)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 1] + E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

Use properties of expectations:

$$= E[\text{Grade}_i(1) - \text{Grade}_i(0)|\text{iPad}_i = 1] + E[\text{Grade}_i(0)|\text{iPad}_i = 1] - E[\text{Grade}_i(0)|\text{iPad}_i = 0]$$

$$= \text{ATT} + \text{Selection bias}$$

Selection bias: Students with and without iPad have different potential grades: **even if they both *had* iPads, they would be different**.

6

Control variables

Control variables are additional variables (or covariates) **included in a regression**. We do this for **various reasons** (in decreasing order of importance):

- To remove selection bias / omitted variable bias
- To increase precision of our estimates
- To know about the (conditional/partial) correlation of other variables
- To better predict the outcome

Let's see graphically how control variables work.

Any questions?

... Remember – Every question is useful!

Attendance and Evaluation

Please fill out a very short evaluation form here:
https://forms.gle/7neuAJbkAv17mSEp6

Quadratic Terms

Quadratic Terms

See in *RStudio*

Measurement Error

Measurement Error

See on Blackboard and in `RStudio`.

If time permits, see on chatGPT.