Controls

EC 607, Set 06

Edward Rubin
Spring 2020

Prologue

Schedule

Last time

The conditional independence assumption: $\{\mathbf{Y}_{0i}, \mathbf{Y}_{1i}\} \perp\!\!\!\perp \mathbf{D}_i \big| \mathbf{X}_i$

I.e., conditional on some controls $(\mathbf{X}_i)$, treatment is as-good-as random.

Today

- Omitted variable bias
- Good *vs.* bad controls

Upcoming

- Topics: Matching estimators
- Admin: Assignment and midterm

Omitted-variable bias

Omitted-variable bias

Revisiting an old friend

Let's start where we left off: Returns to schooling.

We have two linear, population models

$$\text{Y}_i = \alpha + \rho \text{s}_i + \eta_i \tag{1}$$
$$\text{Y}_i = \alpha + \rho \text{s}_i + \text{X}_i' \gamma + \nu_i \tag{2}$$

We should not interpret $\hat{\rho}$ causally in model (1) (for fear of selection bias).

For model (2), we can interpret $\hat{\rho}$ causally ***if*** $\text{Y}_{\text{s}i} \perp\!\!\!\perp \text{s}_i | \text{X}_i$ (CIA).

In other words, the CIA says that our **observable vector X_i must explain all of correlation between s_i and η_i**.

Omitted-variable bias

The OVB formula

We can use the omitted-variable bias (OVB) formula to compare regression estimates from **models with different sets of control variables**.

We're concerned about selection and want to use a set of control variables to account for ability (A_i)—family background, motivation, intelligence.

$$\mathrm{Y}_i = \alpha + \beta \mathrm{s}_i + v_i \tag{1}$$
$$\mathrm{Y}_i = \pi + \rho \mathrm{s}_i + \mathrm{A}_i' \gamma + e_i \tag{2}$$

What happens if we can't get data on A_i and opt for (1)?

$$\frac{\mathrm{Cov}(\mathrm{Y}_i,\, \mathrm{s}_i)}{\mathrm{Var}(\mathrm{s}_i)} = \rho + \gamma' \delta_{As}$$

where δ_{As} are coefficients from regressing A_i on s_i.

Omitted-variable bias

Interpretation

Our two regressions

$$\mathrm{Y}_i = \alpha + \beta \mathrm{s}_i + v_i \tag{1}$$
$$\mathrm{Y}_i = \pi + \rho \mathrm{s}_i + \mathrm{A}_i' \gamma + e_i \tag{2}$$

will yield the same estimates for the returns to schooling

$$\frac{\mathrm{Cov}(\mathrm{Y}_i,\, \mathrm{s}_i)}{\mathrm{Var}(\mathrm{s}_i)} = \rho + \gamma' \delta_{As}$$

if (**a**) schooling is uncorrelated with ability ($\delta_{As} = 0$) *or* (**b**) ability is uncorrelated with earnings, conditional on schooling ($\gamma = 0$).

Omitted-variable bias

Example

Table 3.2.1, The returns to schooling

	1	2	3	4
Schooling	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
Controls	None	Age Dum.	2 + Add'l	3 + AFQT

Here we have four specifications of controls for a regression of log wages on years of schooling (from the NLSY).

Omitted-variable bias

Example

Table 3.2.1, The returns to schooling

	1	2	3	4
Schooling	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
Controls	None	Age Dum.	2 + Add'l	3 + AFQT

Column 1 (no control variables) suggests a 13.2% increase in wages for an additional year of schooling.

Omitted-variable bias

Example

Table 3.2.1, The returns to schooling

	1	2	3	4
Schooling	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
Controls	None	Age Dum.	2 + Add'l	3 + AFQT

Column 2 (age dummies) suggests a 13.1% increase in wages for an additional year of schooling.

Omitted-variable bias

Example

Table 3.2.1, The returns to schooling

	1	2	3	4
Schooling	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
Controls	None	Age Dum.	2 + Add'l	3 + AFQT

Column 3 (column 2 controls plus parents' ed. and self demographics) suggests a 11.4% increase in wages for an additional year of schooling.

Omitted-variable bias

Example

Table 3.2.1, The returns to schooling

	1	2	3	4
Schooling	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
Controls	None	Age Dum.	2 + Add'l	3 + AFQT

Column 4 (column 3 controls plus AFQT† score) suggests a 8.7% increase in wages for an additional year of schooling.

† *AFQT* is *Armed Forces Qualification Test*.

Omitted-variable bias

Example

Table 3.2.1, The returns to schooling

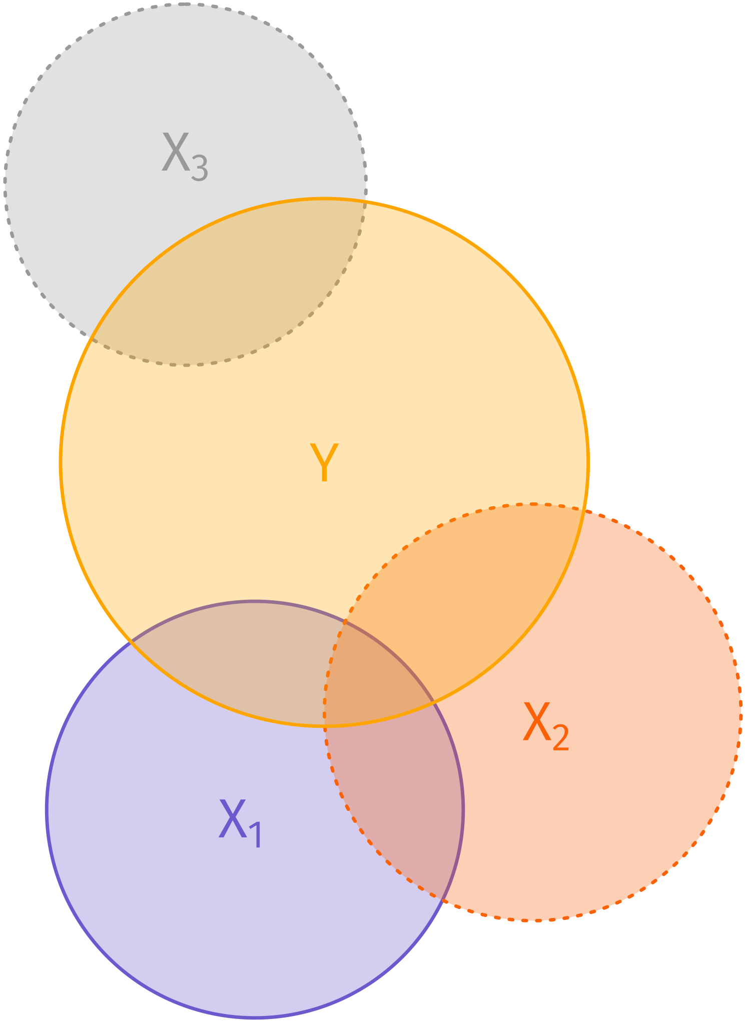
	1	2	3	4
Schooling	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
Controls	None	Age Dum.	2 + Add'l	3 + AFQT

As we ratchet up controls, the estimated returns to schooling drop by 4.5 percentage points (34% drop in the coefficient) from **Column 1** to **Column 4**.

$$\frac{\mathrm{Cov}(\mathrm{Y}_i,\, \mathrm{s}_i)}{\mathrm{Var}(\mathrm{s}_i)} = \rho + \gamma' \delta_{As}$$

If we think **ability positively affects wages**, then it looks like we also have **positive selection into schooling**.

Omitted: X_2 and X_3



Omitted-variable bias

Note

This OVB formula **does not** require either of the models to be causal.

The formula compares the regression coefficient in a **short model** to the regression coefficient on the same variable in a **long model**.†

† Here, ***long model*** refers to a model with more controls than the ***short model***.

Omitted-variable bias

The OVB formula and the CIA[†]

In addition to helping us think through and sign OVB, the formula

$$\frac{\text{Cov}(\text{Y}_i,\, \text{s}_i)}{\text{Var}(\text{s}_i)} = \rho + \gamma' \delta_{As}$$

drives home the point that we're leaning *very* hard on the conditional independence assumption to be able to interpret our coefficients as causal.

Q When is the CIA plausible?

A Two potential answers

1. Randomized experiments
2. Programs with arbitrary cutoffs/lotteries

† The title for my first spy novel.

Control variables play an enormous role in our quest for causality (the CIA).

Q Are "more controls" always better (or at least never worse)?

A No. There are such things as...

Bad controls

Bad controls

Defined

Q What's a *bad* control—when can a control make a bad situation worse?

A *Bad controls* are variables that are (also) affected by treatment.

Q Okay, so why is it bad to control using a variable affected by treatment?

Hint It's a flavor of selection bias.

Let's consider an example...

Bad controls

Example

Suppose we want to know the **effect of college graduation on wages**.

1. There are only two types of jobs: blue collar and white collar.
2. White-collar jobs, on averge, pay more than blue-collar jobs.
3. Graduating college increases the likelihood of a white-collar job.

Q Should we control for occupation type when considering the effect of college graduation on wages? (Will occupation be an omitted variable?)

A No. Imagine college degrees are randomly assigned. When we condition on occupation, we compare degree-earners who chose blue-collar jobs to non-degree-earners who chose blue-collar jobs. Our assumption of random degrees says **nothing** about random job selection.

Bad controls

Formal-ish derivation

More formally, let

- W_i be a dummy for whether i has a white-collar job
- Y_i denote i's earnings
- C_i refer to i's **randomly assigned** college-graduation status

$$\text{Y}_i = \text{C}_i \text{Y}_{1i} + (1 - \text{C}_i)\, \text{Y}_{0i}$$
$$\text{W}_i = \text{C}_i \text{W}_{1i} + (1 - \text{C}_i)\, \text{W}_{0i}$$

Becuase we've assumed C_i is randomly assigned, differences in means yield causal estimates, *i.e.*,

$$E[\text{Y}_i \mid \text{C}_i = 1] - E[\text{Y}_i \mid \text{C}_i = 0] = E[\text{Y}_{1i} - \text{Y}_{0i}]$$
$$E[\text{W}_i \mid \text{C}_i = 1] - E[\text{W}_i \mid \text{C}_i = 0] = E[\text{W}_{1i} - \text{W}_{0i}]$$

Bad controls

Formal-ish derivation, continued

Let's see what happens when we throw in some controls—*e.g.*, focusing on the the wage-effect of college graduation for white-collar jobs.

$$E[\mathrm{Y}_i \mid \mathrm{W}_i = 1,\ \mathrm{C}_i = 1] - E[\mathrm{Y}_i \mid \mathrm{W}_i = 1,\ \mathrm{C}_i = 0]$$

$$= E[\mathrm{Y}_{1i} \mid \mathrm{W}_{1i} = 1,\ \mathrm{C}_i = 1] - E[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1,\ \mathrm{C}_i = 0]$$

$$= E[\mathrm{Y}_{1i} \mid \mathrm{W}_{1i} = 1] - E[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1]$$

$$\begin{aligned} &= E[\mathrm{Y}_{1i} \mid \mathrm{W}_{1i} = 1] - E[\mathrm{Y}_{0i} \mid \mathrm{W}_{1i} = 1] \\ &\quad + E[\mathrm{Y}_{0i} \mid \mathrm{W}_{1i} = 1] - E[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1] \end{aligned}$$

$$= \underbrace{E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} \mid \mathrm{W}_{1i} = 1]}_{\text{Causal effect on white-collar workers}} + \underbrace{E[\mathrm{Y}_{0i} \mid \mathrm{W}_{1i} = 1] - E[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1]}_{\text{Selection bias}}$$

Formal-ish derivation, continued

By introducing a bad control, we introduced selection bias into a setting that did not have selection bias without controls.

Specifically, the selection bias term

$$E[\mathrm{Y}_{0i} \mid \mathrm{W}_{1i} = 1] - E[\mathrm{Y}_{0i} \mid \mathrm{W}_{0i} = 1]$$

describes how college graduation changes the composition of the pool of white-class workers.

Note Even if the causal effect is zero, this selection bias need not be zero.

Bad controls

A trickier example

A timely/trickier example: Wage gaps (*e.g.*, female-male or black-white).

Q Should we control for occupation when we consider wage gaps?

- What are we trying to capture?
- If we're concerned about discrimination, it seems likely that discrimination also affects occupational choice and hiring outcomes.
- Some motivate occuption controls with groups' differential preferences.

What's the answer?

Bad controls

Proxy variables

Angrist and Pischke bring up an interesting scenario that intersects omitted-variable bias and bad controls.

- We want to estimate the returns to education.
- Ability is omitted.
- We have a proxy for ability—a test taken after schooling finishes.

We're a bit stuck.

1. If we omit the test altogether, we've got omitted-variable bias.
2. If we include our proxy, we've got a back control.

With some math/luck, we can bound the true effect with these estimates.

Bad controls

Example

Returning to our OVB-motivated example, we control for occupation.

Table 3.2.1, The returns to schooling

	1	2	3	4	5
Schooling	0.132	0.131	0.114	0.087	0.066
	(0.007)	(0.007)	(0.007)	(0.009)	(0.010)
Controls	None	Age Dum.	2 + Add'l	3 + AFQT	4 + Occupation

Schooling likely affects occupation; how do we interpret the new results?

Bad controls

Conclusion

Timing matters.

The right controls can help tremendously, but bad controls hurt.

Table of contents

Admin

Controls