

# Controls

EC 425/525, Set 6

Edward Rubin  
28 April 2019

# Prologue

# Schedule

## Last time

The conditional independence assumption:  $\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i | X_i$   
I.e., conditional on some controls ( $X_i$ ), treatment is as-good-as random.

## Today

- Omitted variable bias
- Good vs. bad controls

## Upcoming

Topics: Matching estimators

# Follow up

Example from last class.

# Advice

Atom...

# 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

$$Y_i = \alpha + \rho s_i + \eta_i \quad (1)$$

$$Y_i = \alpha + \rho s_i + X'_i \gamma + \nu_i \quad (2)$$

# Omitted-variable bias

## Revisiting an old friend

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

We have two linear, population models

$$Y_i = \alpha + \rho s_i + \eta_i \quad (1)$$

$$Y_i = \alpha + \rho s_i + X'_i \gamma + \nu_i \quad (2)$$

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

# Omitted-variable bias

## Revisiting an old friend

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

We have two linear, population models

$$Y_i = \alpha + \rho s_i + \eta_i \quad (1)$$

$$Y_i = \alpha + \rho s_i + X'_i \gamma + \nu_i \quad (2)$$

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**  $Y_{si} \perp\!\!\!\perp s_i | X_i$  (CIA).

# Omitted-variable bias

## Revisiting an old friend

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

We have two linear, population models

$$Y_i = \alpha + \rho s_i + \eta_i \quad (1)$$

$$Y_i = \alpha + \rho s_i + X'_i \gamma + \nu_i \quad (2)$$

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*  $Y_{si} \perp\!\!\!\perp s_i | X_i$  (CIA).

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

# 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**.

# 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.

$$Y_i = \alpha + \beta s_i + v_i \quad (1)$$

$$Y_i = \pi + \rho s_i + A'_i \gamma + e_i \quad (2)$$

# 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 ( $\mathbf{A}_i$ )—family background, motivation, intelligence.

$$Y_i = \alpha + \beta s_i + v_i \tag{1}$$

$$Y_i = \pi + \rho s_i + \mathbf{A}'_i \gamma + e_i \tag{2}$$

What happens if we can't get data on  $\mathbf{A}_i$  and opt for (1)?

# 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.

$$Y_i = \alpha + \beta s_i + v_i \quad (1)$$

$$Y_i = \pi + \rho s_i + A'_i \gamma + e_i \quad (2)$$

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

$$\frac{\text{Cov}(Y_i, s_i)}{\text{Var}(s_i)} = \rho + \gamma' \delta_{As}$$

where  $\delta_{As}$  are coefficients from regressing  $A_i$  on  $s_i$ .

# Omitted-variable bias

## Interpretation

Our two regressions

$$Y_i = \alpha + \beta s_i + v_i \tag{1}$$

$$Y_i = \pi + \rho s_i + A'_i \gamma + e_i \tag{2}$$

will yield the same estimates for the returns to schooling

$$\frac{\text{Cov}(Y_i, s_i)}{\text{Var}(s_i)} = \rho + \gamma' \delta_{As}$$

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
<b>Schooling</b>	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
<b>Controls</b>	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
<b>Schooling</b>	0.132 (0.007)	0.131 (0.007)	0.114 (0.007)	0.087 (0.009)
<b>Controls</b>	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
<b>Schooling</b>	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
<b>Controls</b>	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
<b>Schooling</b>	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
<b>Controls</b>	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
<b>Schooling</b>	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
<b>Controls</b>	None	Age Dum.	2 + Add'l	3 + AFQT

**Column 4** (column 3 controls plus AFQT<sup>†</sup> 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
<b>Schooling</b>	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)
<b>Controls</b>	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**.

# Omitted-variable bias

## Example

Table 3.2.1, The returns to schooling

	1	2	3	4
<b>Schooling</b>	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)

<b>Controls</b>	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**.

$$\frac{\text{Cov}(Y_i, s_i)}{\text{Var}(s_i)} = \rho + \gamma' \delta_{As}$$

# Omitted-variable bias

## Example

Table 3.2.1, The returns to schooling

	1	2	3	4
<b>Schooling</b>	0.132	0.131	0.114	0.087
	(0.007)	(0.007)	(0.007)	(0.009)

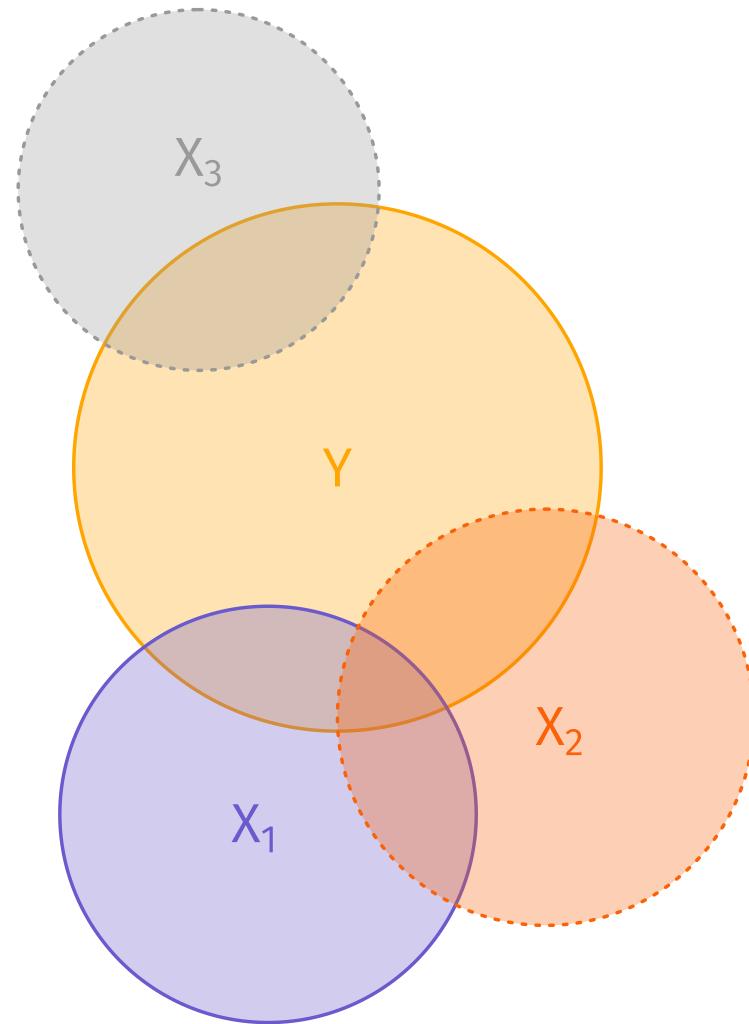
<b>Controls</b>	None	Age Dum.	2 + Add'l	3 + AFQT
-----------------	------	----------	-----------	----------

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**.

$$\frac{\text{Cov}(Y_i, s_i)}{\text{Var}(s_i)} = \rho + \gamma' \delta_{As}$$

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**.<sup>†</sup>

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

# Omitted-variable bias

## The OVB formula and the CIA<sup>†</sup>

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

$$\frac{\text{Cov}(Y_i, s_i)}{\text{Var}(s_i)} = \rho + \gamma' \delta_{As}$$

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

<sup>†</sup> The title for my first spy novel.

# Omitted-variable bias

## The OVB formula and the CIA<sup>†</sup>

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

$$\frac{\text{Cov}(Y_i, s_i)}{\text{Var}(s_i)} = \rho + \gamma' \delta_{As}$$

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?

<sup>†</sup> The title for my first spy novel.

# Omitted-variable bias

## The OVB formula and the CIA<sup>†</sup>

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

$$\frac{\text{Cov}(Y_i, s_i)}{\text{Var}(s_i)} = \rho + \gamma' \delta_{As}$$

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

<sup>†</sup> 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?

# 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.

# 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?

# 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.

# 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 average, pay more than blue-collar jobs.
3. Graduating college increases the likelihood of a white-collar job.

# 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 average, 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?)

# 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 average, 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.

# 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 average, 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.

# 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 average, 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,

# 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 average, 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.

# 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 average, 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

# 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

$$Y_i = C_i Y_{1i} + (1 - C_i) Y_{0i}$$
$$W_i = C_i W_{1i} + (1 - C_i) W_{0i}$$

# 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

$$Y_i = C_i Y_{1i} + (1 - C_i) Y_{0i}$$
$$W_i = C_i W_{1i} + (1 - C_i) W_{0i}$$

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

$$E[Y_i | C_i = 1] - E[Y_i | C_i = 0] = E[Y_{1i} - Y_{0i}]$$

$$E[W_i | C_i = 1] - E[W_i | C_i = 0] = E[W_{1i} - W_{0i}]$$

# Bad controls

## Formal-ish derivation, continued

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

# Bad controls

## Formal-ish derivation, continued

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

$$E[Y_i | W_i = 1, \text{C}_i = 1] - E[Y_i | W_i = 1, \text{C}_i = 0]$$

# Bad controls

## Formal-ish derivation, continued

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

$$E[Y_i \mid W_i = 1, C_i = 1] - E[Y_i \mid W_i = 1, C_i = 0]$$

$$= E[Y_{1i} \mid W_{1i} = 1, C_i = 1] - E[Y_{0i} \mid W_{0i} = 1, C_i = 0]$$

# Bad controls

## Formal-ish derivation, continued

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

$$\begin{aligned} & E[Y_i \mid W_i = 1, C_i = 1] - E[Y_i \mid W_i = 1, C_i = 0] \\ &= E[Y_{1i} \mid W_{1i} = 1, C_i = 1] - E[Y_{0i} \mid W_{0i} = 1, C_i = 0] \\ &= E[Y_{1i} \mid W_{1i} = 1] - E[Y_{0i} \mid W_{0i} = 1] \end{aligned}$$

# Bad controls

## Formal-ish derivation, continued

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

$$E[Y_i \mid W_i = 1, C_i = 1] - E[Y_i \mid W_i = 1, C_i = 0]$$

$$= E[Y_{1i} \mid W_{1i} = 1, C_i = 1] - E[Y_{0i} \mid W_{0i} = 1, C_i = 0]$$

$$= E[Y_{1i} \mid W_{1i} = 1] - E[Y_{0i} \mid W_{0i} = 1]$$

$$\begin{aligned} &= E[Y_{1i} \mid W_{1i} = 1] - E[Y_{0i} \mid W_{1i} = 1] \\ &\quad + E[Y_{0i} \mid W_{1i} = 1] - E[Y_{0i} \mid W_{0i} = 1] \end{aligned}$$

# Bad controls

## Formal-ish derivation, continued

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

$$\begin{aligned} & E[Y_i \mid W_i = 1, C_i = 1] - E[Y_i \mid W_i = 1, C_i = 0] \\ &= E[Y_{1i} \mid W_{1i} = 1, C_i = 1] - E[Y_{0i} \mid W_{0i} = 1, C_i = 0] \\ &= E[Y_{1i} \mid W_{1i} = 1] - E[Y_{0i} \mid W_{0i} = 1] \\ &= E[Y_{1i} \mid W_{1i} = 1] - E[Y_{0i} \mid W_{1i} = 1] \\ &\quad + E[Y_{0i} \mid W_{1i} = 1] - E[Y_{0i} \mid W_{0i} = 1] \\ &= \underbrace{E[Y_{1i} - Y_{0i} \mid W_{1i} = 1]}_{\text{Causal effect on white-collar workers}} + \underbrace{E[Y_{0i} \mid W_{1i} = 1] - E[Y_{0i} \mid W_{0i} = 1]}_{\text{Selection bias}} \end{aligned}$$

# Bad controls

## Formal-ish derivation, continued

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

# Bad controls

## 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

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

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

# Bad controls

## 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

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

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).

# 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?

# 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 occupation controls with groups' differential preferences.

# 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 occupation 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.

# 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.

# 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
<b>Schooling</b>	0.132	0.131	0.114	0.087	0.066
	(0.007)	(0.007)	(0.007)	(0.009)	(0.010)
<b>Controls</b>	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

- 1. Schedule
- 2. Follow up
- 3. Advice

## Controls

- 1. Omitted-variable bias
  - The formula
  - Example
  - OVB Venn
  - OVB and the CIA
- 2. Bad controls
  - Defined
  - Example
  - Formalization(ish)
  - Trickier example
  - Bad proxy conundrum
  - Empirical example