All About Regression

EC 350: Labor Economics

Kyle Raze

Winter 2022

All About Regression

Econometrics

The objective? Identify the effect of a treatment variable D on an outcome variable Y.[†]

- **How?** Find a way to shut down **selection bias**.

Regression analysis

> A set of statistical processes for quantifying the relationship between a dependent variable (*e.g.*, an outcome) and one or more independent variables (*e.g.*, a treatment or a control variable).

A bundle of useful tools for doing econometrics!

[†] The other objective? Forecast future values of key outcome variables, such as unemployment, GDP, customer retention, *etc*. But that's a different subject for a different course.

All About Regression

Regression analysis

Economists often rely on regression analysis to make various statistical comparisons.

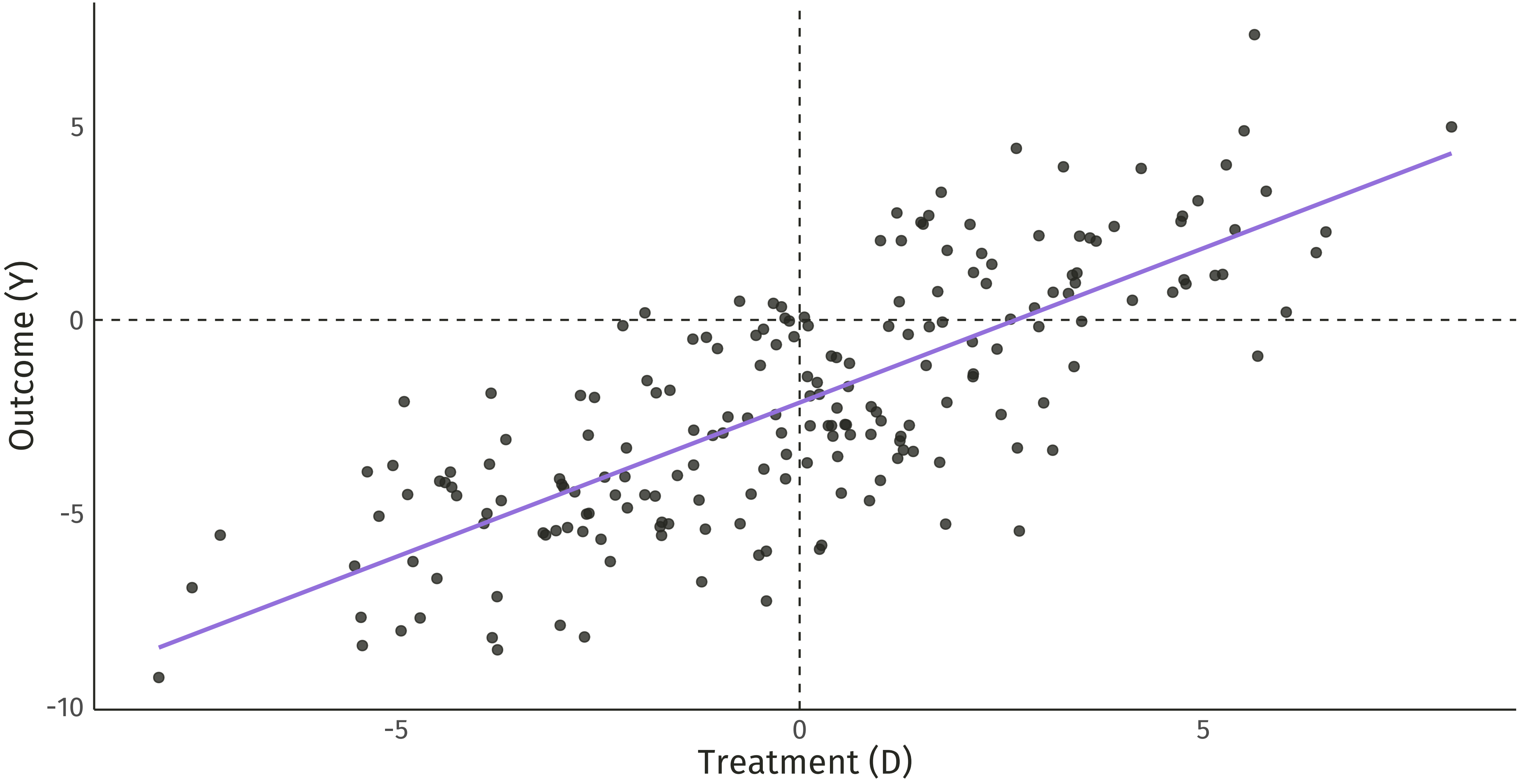
- Can facilitate *other things equal* comparisons.
- Can shut down selection bias by explicitly **controlling for confounding variables**.
- Failure to control for confounding variables? $\longrightarrow$ **omitted-variable bias**.

Our objective? Learn how to interpret the results of a regression analysis.

1. **Literal interpretation**
   - Interpret the size and statistical significance of regression coefficient estimates.
   - Know your way around a regression table.
2. **Big-picture interpretation**
   - What do the estimates imply about the effects of a treatment?
   - Should we trust the estimates? Do they reflect a causal relationship?

Simple linear regression

Simple linear regression



Simple linear regression

Model

We can express the relationship between the **outcome variable** and the **treatment variable** as linear:

$$Y_i = \alpha + \beta\, D_i + \varepsilon_i$$

- i indexes an individual.
- α = the **intercept** or constant.
- β = the **slope coefficient**.
  - Imagine for now that D_i can take on many different values (*e.g.*, more than just 0 or 1).
- ε_i = the **error term**.

Simple = Only one independent variable.

Simple linear regression

Model

The **intercept** tells us the expected value of Y_i when $D_i = 0$.

$$Y_i = \alpha + \beta \, D_i + \varepsilon_i$$

Part of the regression line, but almost never the focus of an analysis.

- In practice, omitting the intercept would bias estimates of the slope coefficient—the object we really care about.

Model

The **slope coefficient** tells us the expected change in Y_i when D_i increases by one.

$$Y_i = \alpha + \beta \, D_i + \varepsilon_i$$

"A one-unit increase in D_i *is associated with* a β-unit increase in Y_i."

Under certain (strong) assumptions about the error term (*e.g.*, no selection bias), β represents the causal effect of D_i on Y_i.

- "A one-unit increase in D_i *leads to* a β-unit increase in Y_i."
- Otherwise, it's just the *association of* D_i *with* Y_i, representing a non-causal correlation.

Model

The **error term** reminds us that D_i isn't the only variable that affects Y_i.

$$Y_i = \alpha + \beta \, D_i + \varepsilon_i$$

The error term represents all other factors that explain Y_i.

- **So what?** If some of those factors influence D_i, then omitted-variable bias will contaminate estimates of the slope coefficient.

Simple linear regression

Example

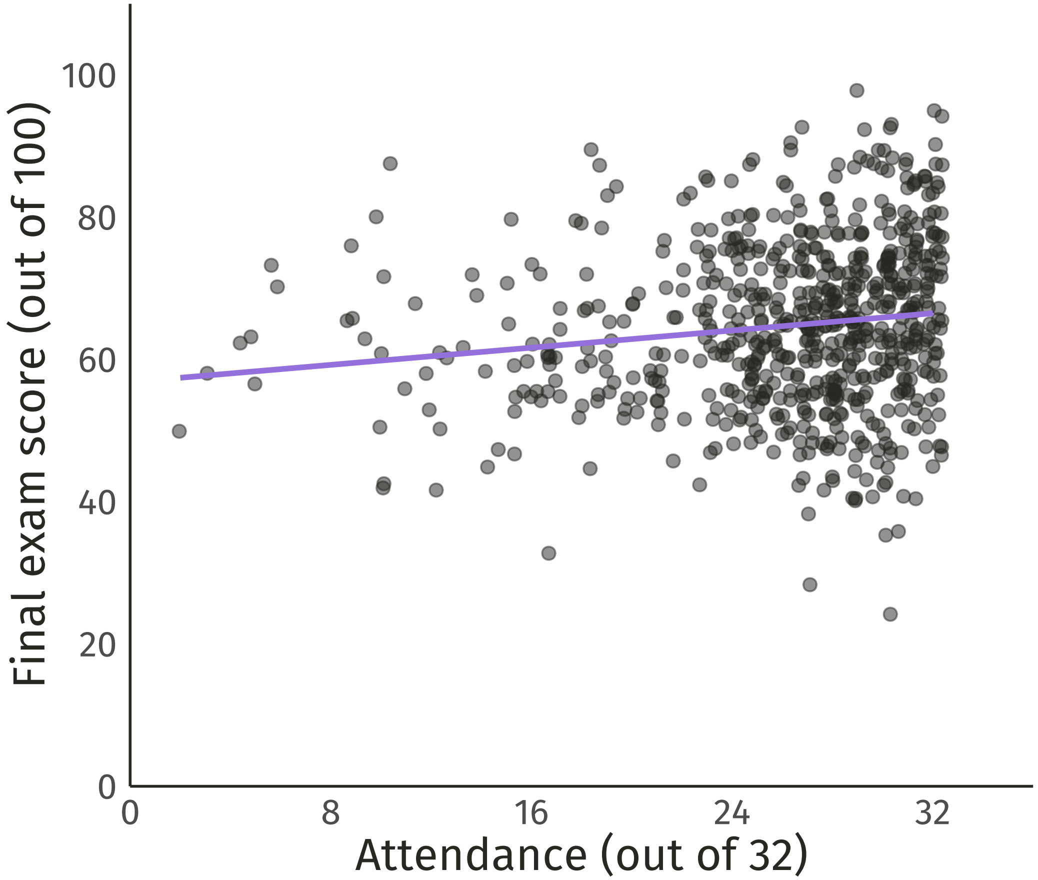
Q: How does attendance affect performance?

As a first attempt at an answer, we can estimate a regression of final exam scores on attendance:

$$\text{Final}_i = \alpha + \beta\,\text{Attend}_i + \varepsilon_i$$

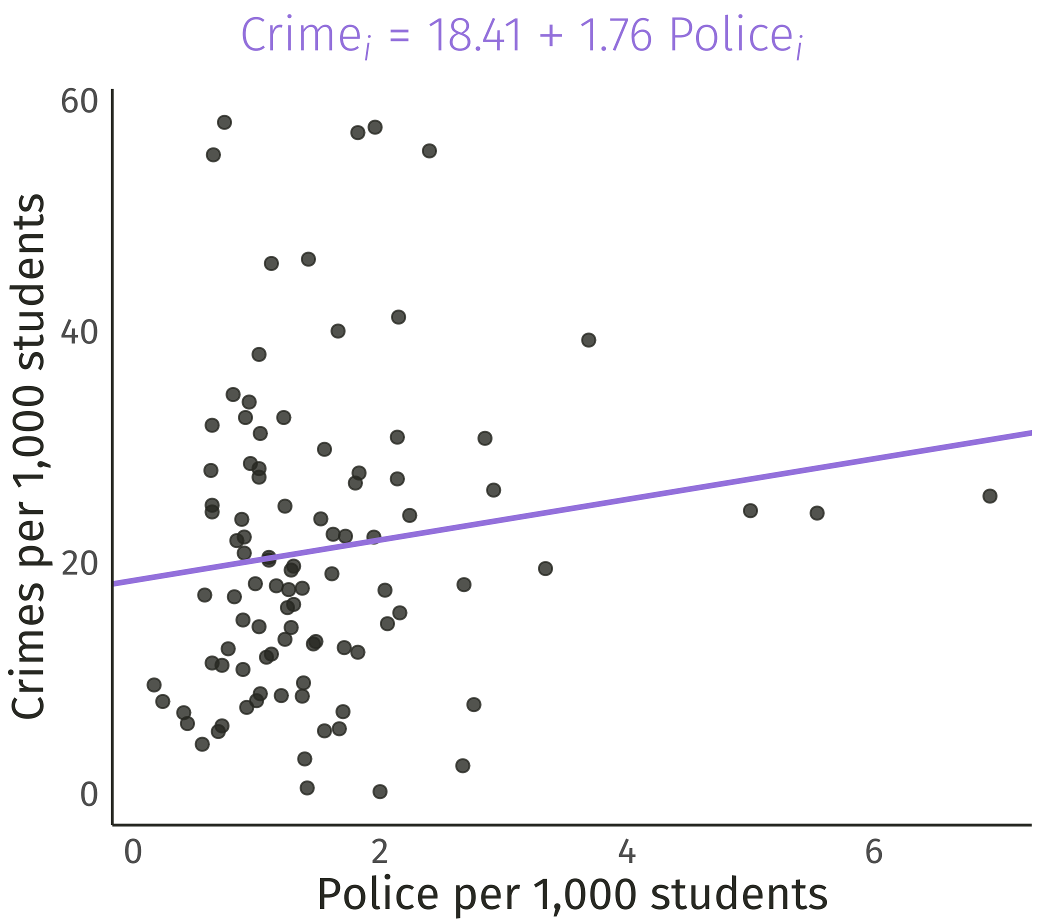
Parameter	**(1)**
Intercept	**56.82**
	(2.19)
Attendance	**0.3**
	(0.08)

Standard errors in parentheses.



Simple linear regression

Example



Q: Do police on college campuses reduce crime?

- What does the slope coefficient tell us?

Q: Does this mean that police *cause* crime!?

- Why or why not?

For an interesting discussion of the causal effects of police staffing on crime and arrests—and how those effects vary by race—check out episode 55 of the *Probable Causation* podcast.

Simple linear regression

Estimation

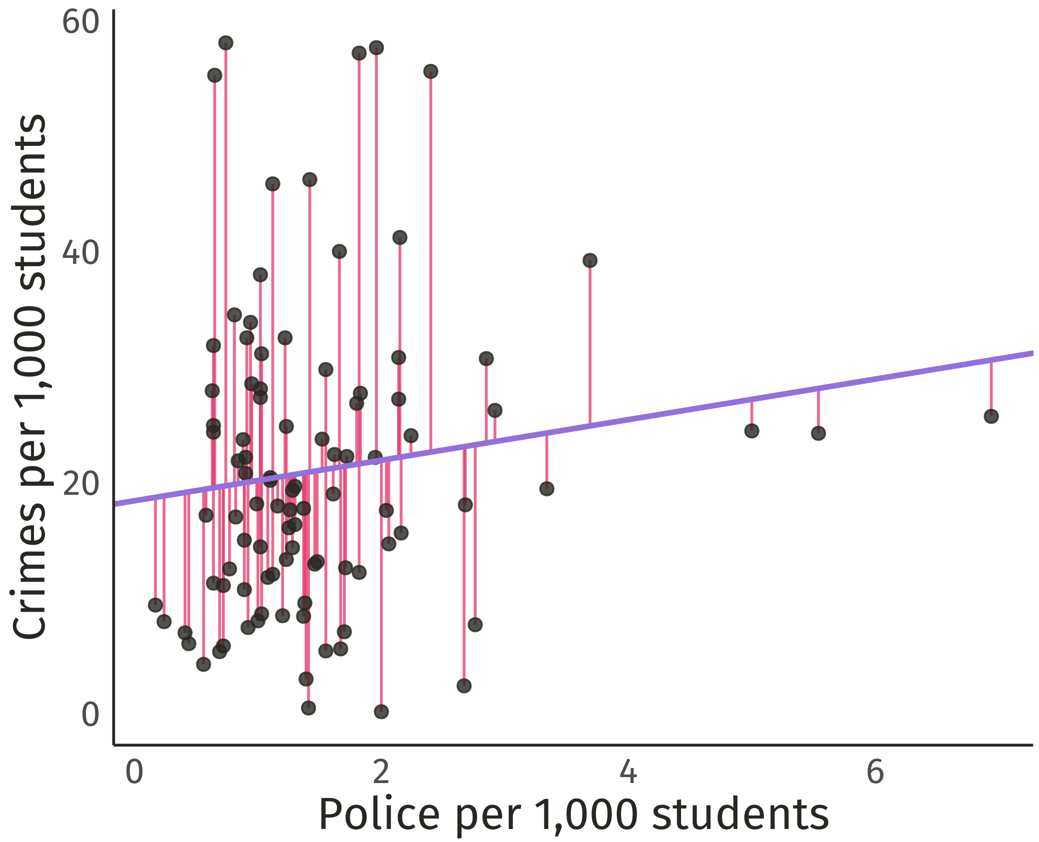
Q: Where does the regression line come from?

A: A routine called **ordinary least squares (OLS)**.

How does OLS work?

- Every "fitted line" produces **residuals**.
- Residual = actual − **predicted**

$\text{Crime}_i = 18.41 + 1.76 \text{ Police}_i$



Simple linear regression

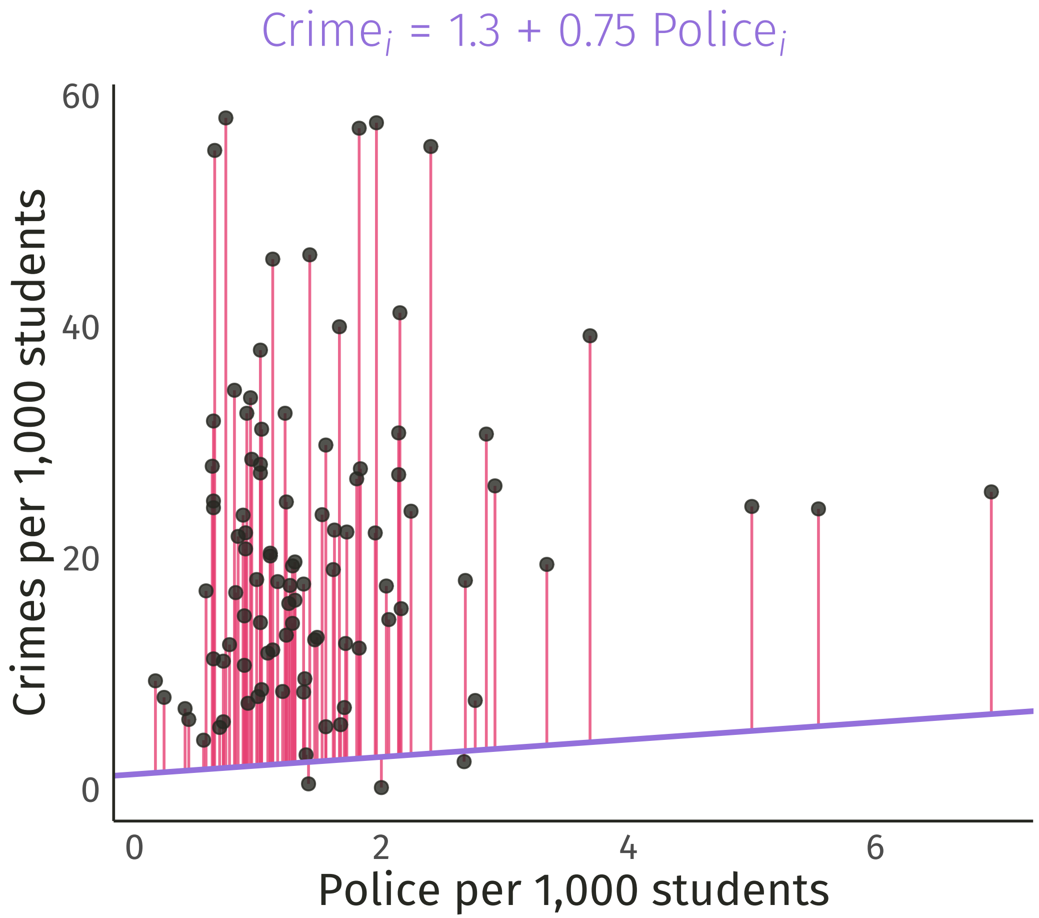
Estimation

Q: Where does the regression line come from?

A: A routine called **ordinary least squares (OLS)**.

How does OLS work?

- Some fitted lines generate bigger residuals than others.



Simple linear regression

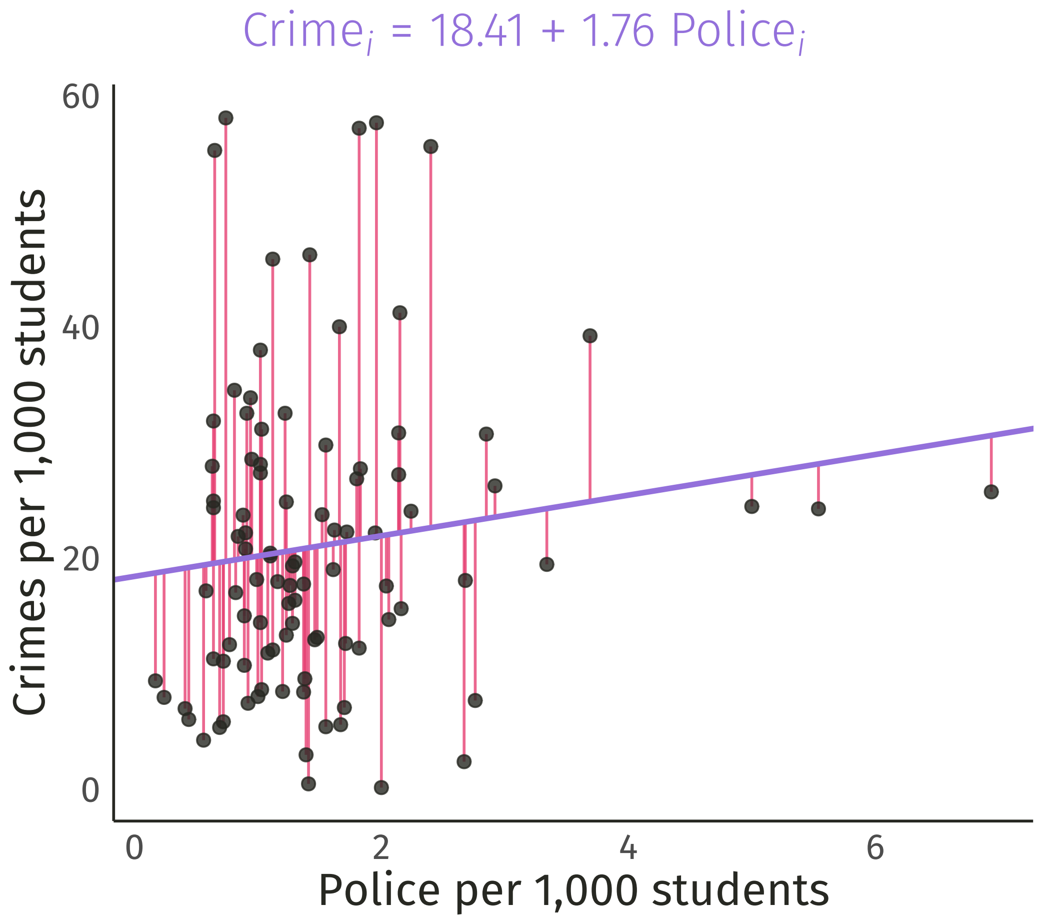
Estimation

Q: Where does the regression line come from?

A: A routine called **ordinary least squares (OLS)**.

How does OLS work?

- The "line of best fit" is the line that **minimizes** the **sum of squared residuals**.
- **Q:** Why squared?
- Using math you'll see in EC 320 or matrix algebra, OLS does this without the guesswork.



Simple linear regression

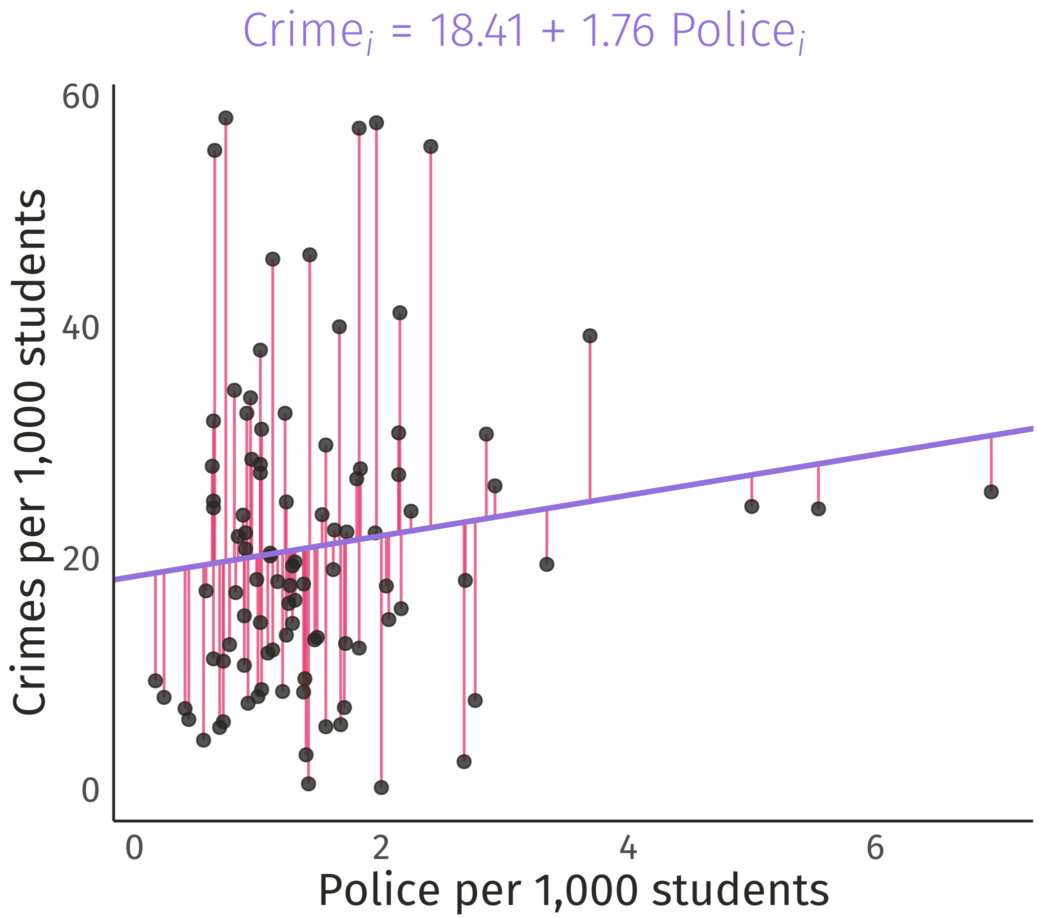
Estimation

Q: Where does the regression line come from?

A: A routine called **ordinary least squares (OLS)**.

How does OLS work?

- **"Squares?"** Sum of squared residuals.
- **"Least?"** Minimize that sum.
- **"Ordinary?"** Oldest, most common way of estimating a regression.



Simple linear regression

Example: Returns to education

The optimal investment in education by students, parents, and legislators depends in part on the monetary *return to education*.

Thought experiment:

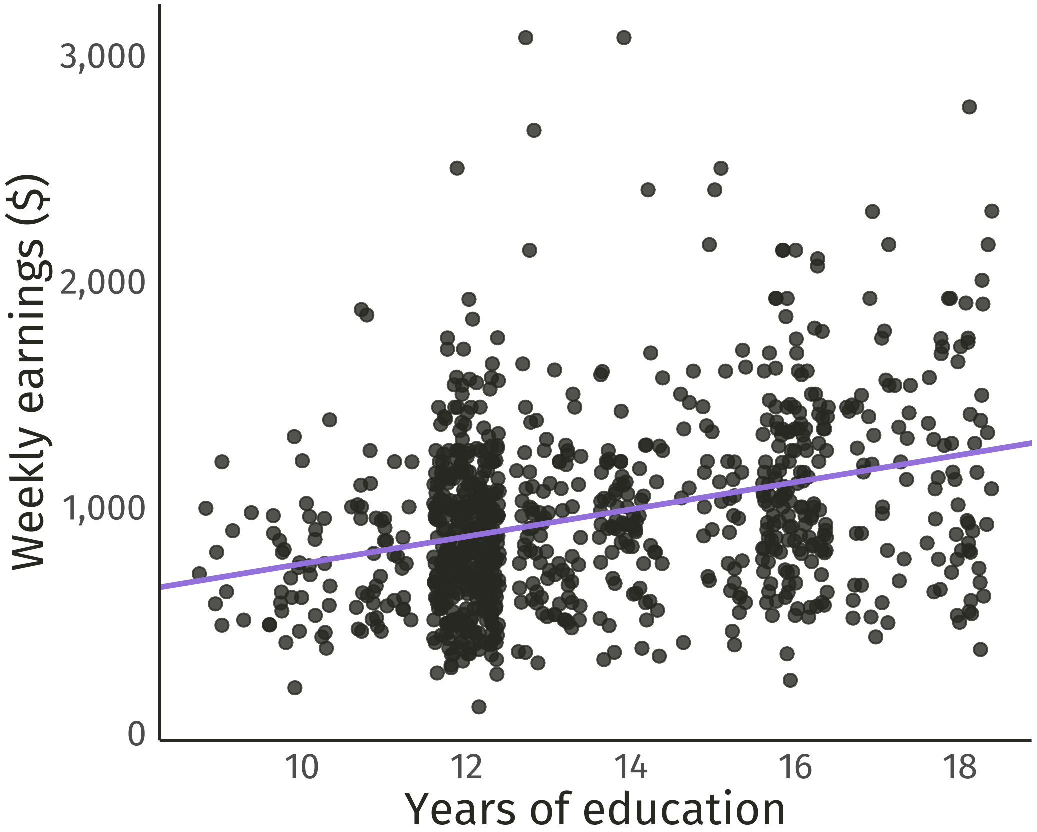
- Randomly select an individual.
- Give her an additional year of education.
- How much do her earnings increase?

The change in her earnings describes the **causal effect** of education on earnings.

Simple linear regression

Example: Returns to education

$\text{Earnings}_i = 146.95 + 60.21\ \text{Schooling}_i$



Q: How much extra money can a worker in this sample expect from an additional year of education?

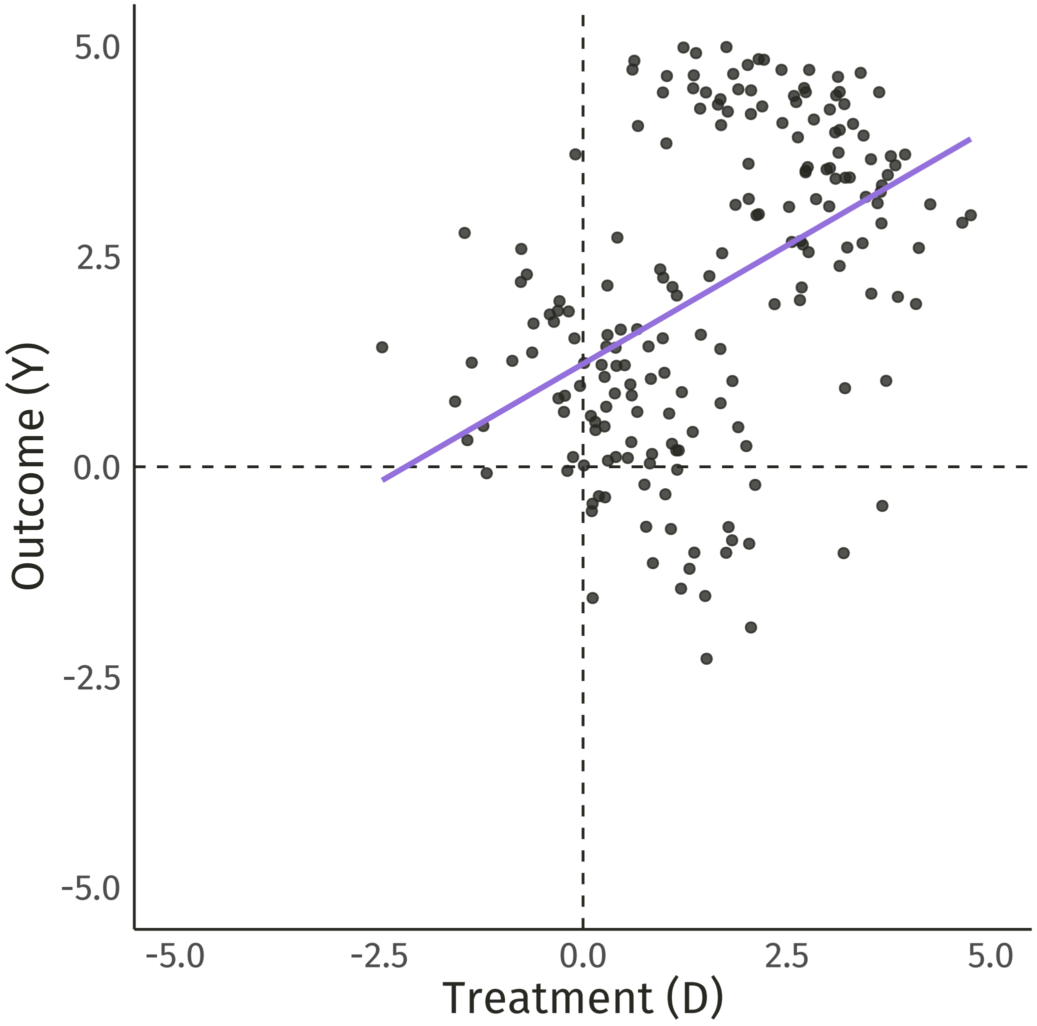
- How do you know?

Q: Does this number represent the causal return to an additional year of education?

- What other variables could be driving the relationship?

Making adjustments

Making adjustments



We can produce a fitted line by estimating a regression of an outcome on a treatment:

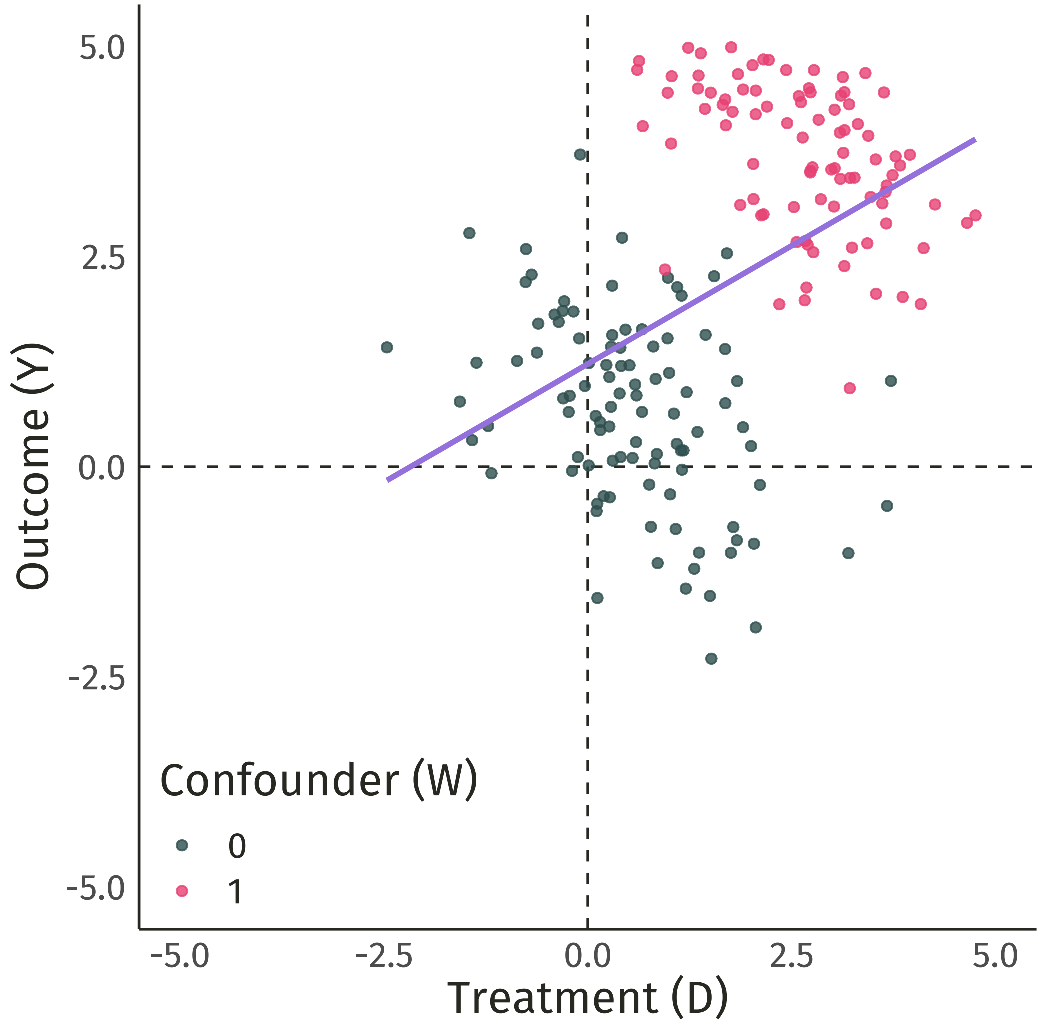
$$Y_i = \alpha + \beta\, D_i + \varepsilon_i$$

β describes how the outcome changes, *on average*, when treatment changes.

Parameter	**(1)**
Intercept	**1.22**
	(0.18)
Treatment	**0.56**
	(0.08)

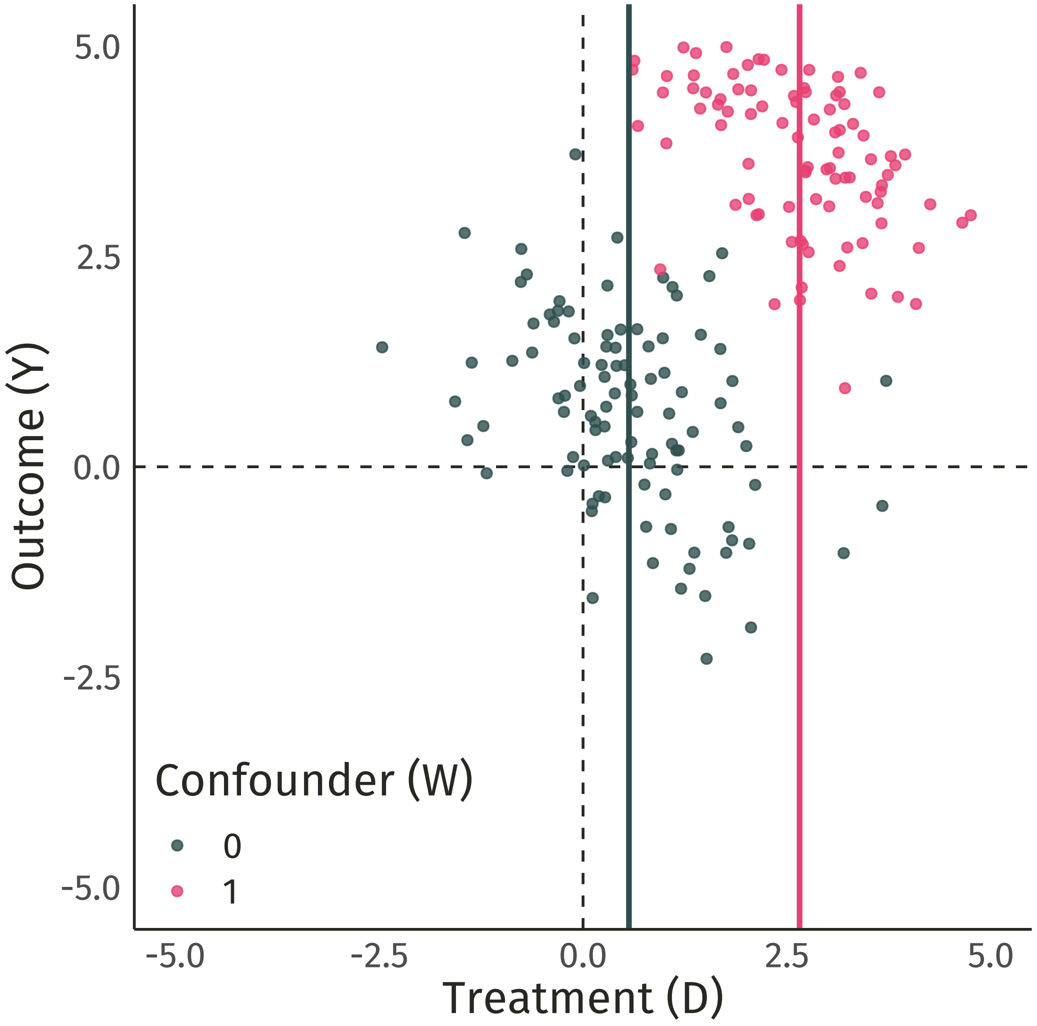
Standard errors in parentheses.

Making adjustments



However, we might worry that a third variable W_i confounds our estimate of the effect of the treatment on the outcome.

Making adjustments



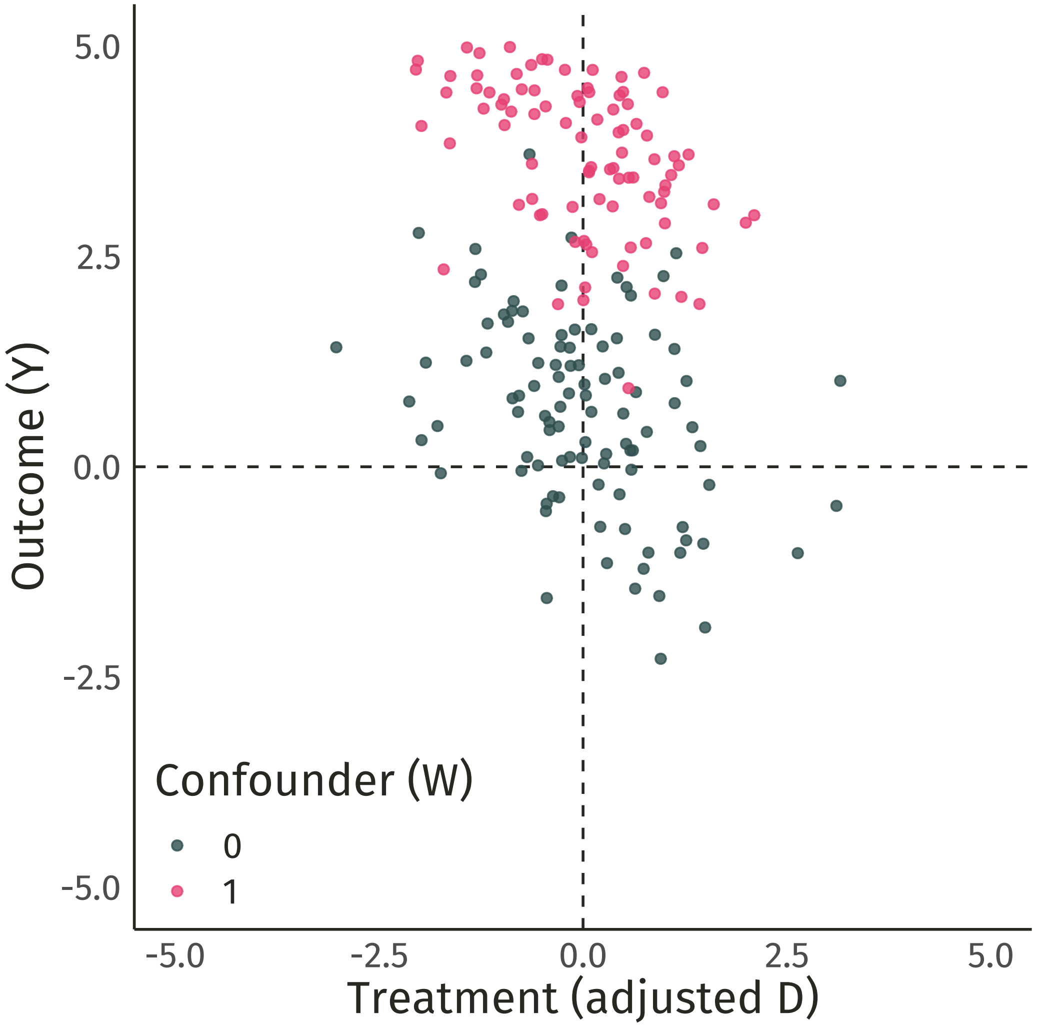
If data on the confounder exists, it can be added to the regression model:

$$Y_i = \alpha + \beta\, D_i + \gamma\, W_i + \varepsilon_i$$

Q: How does OLS "adjust" for the confounder?

- **Step 1:** Figure out what differences in D are explained by W.

Making adjustments



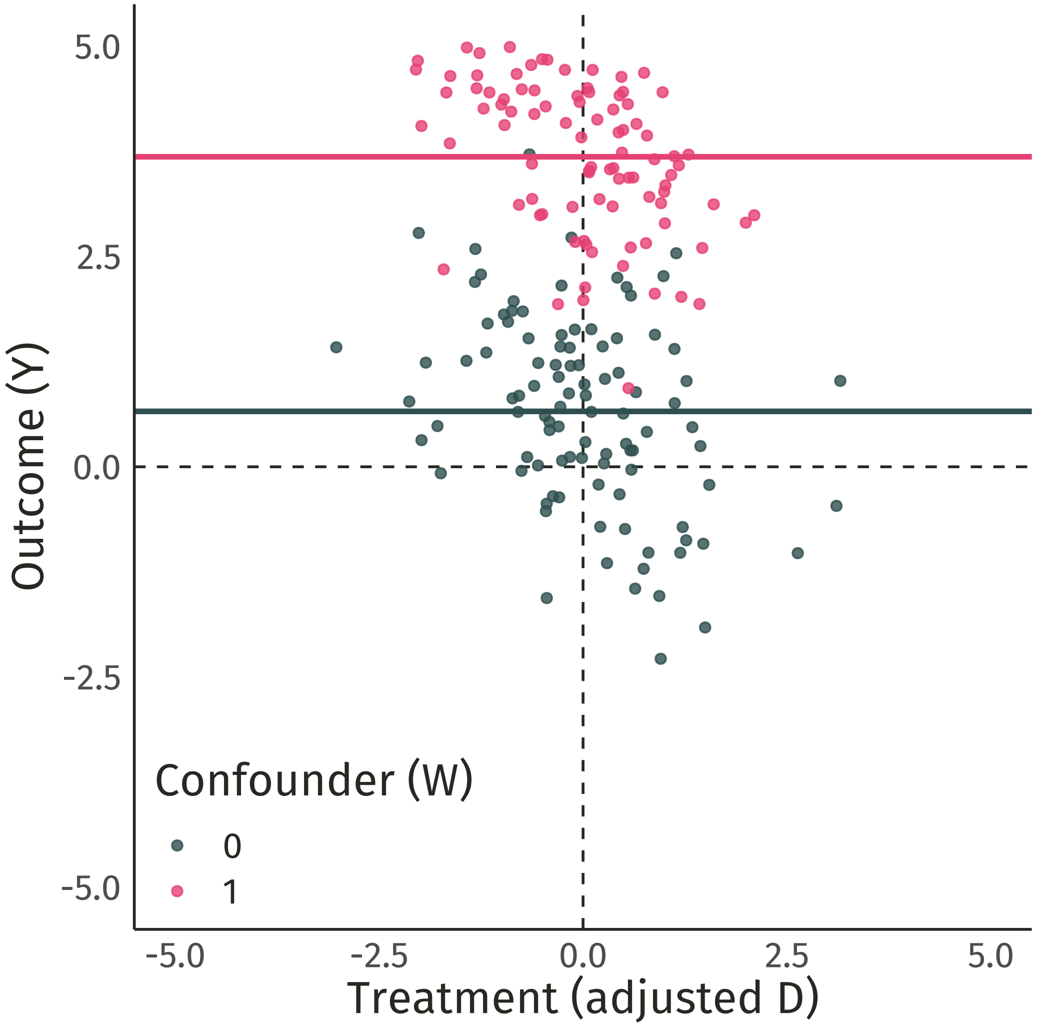
If data on the confounder exists, it can be added to the regression model:

$$Y_i = \alpha + \beta\, D_i + \gamma\, W_i + \varepsilon_i$$

Q: How does OLS "adjust" for the confounder?

- **Step 2:** Remove differences in D explained by W.

Making adjustments



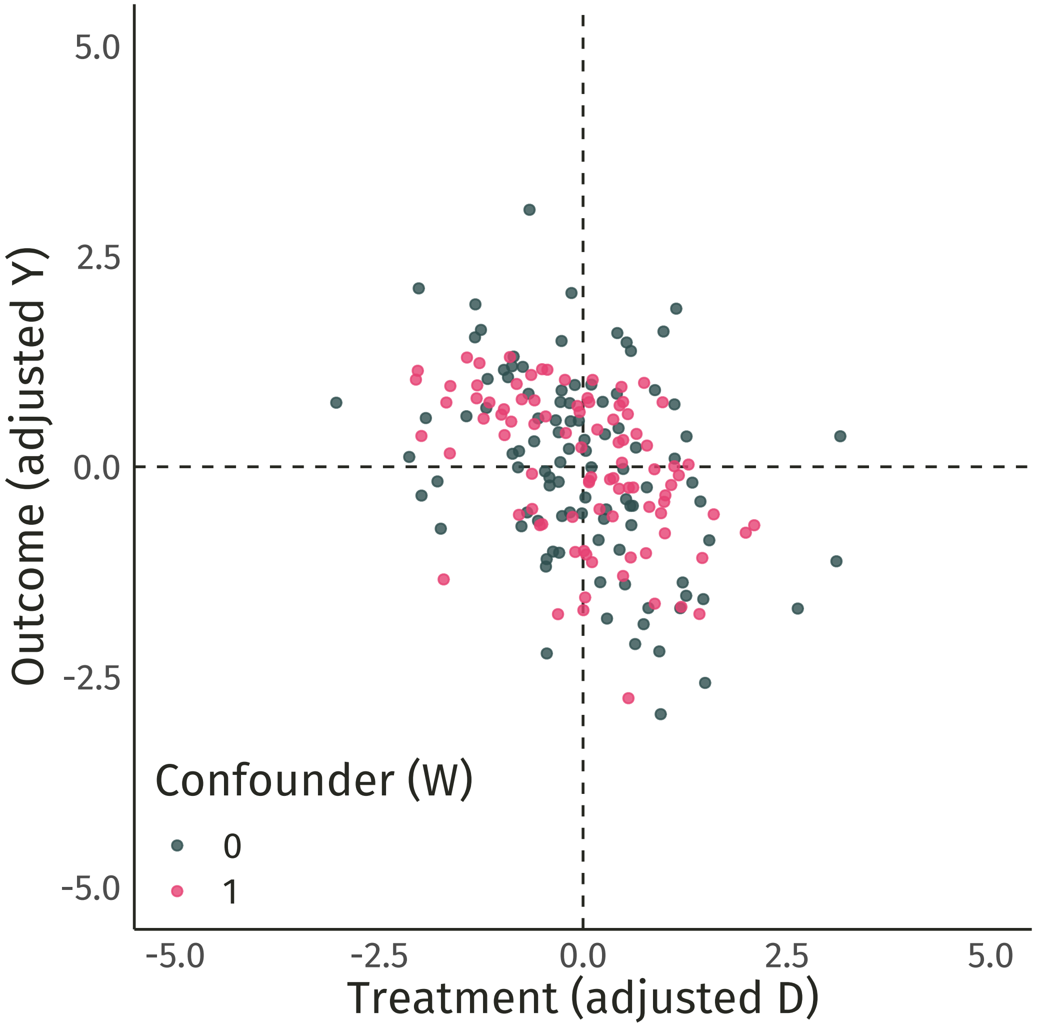
If data on the confounder exists, it can be added to the regression model:

$$Y_i = \alpha + \beta\, D_i + \gamma\, W_i + \varepsilon_i$$

Q: How does OLS "adjust" for the confounder?

- **Step 3:** Figure out what differences in Y are explained by W.

Making adjustments



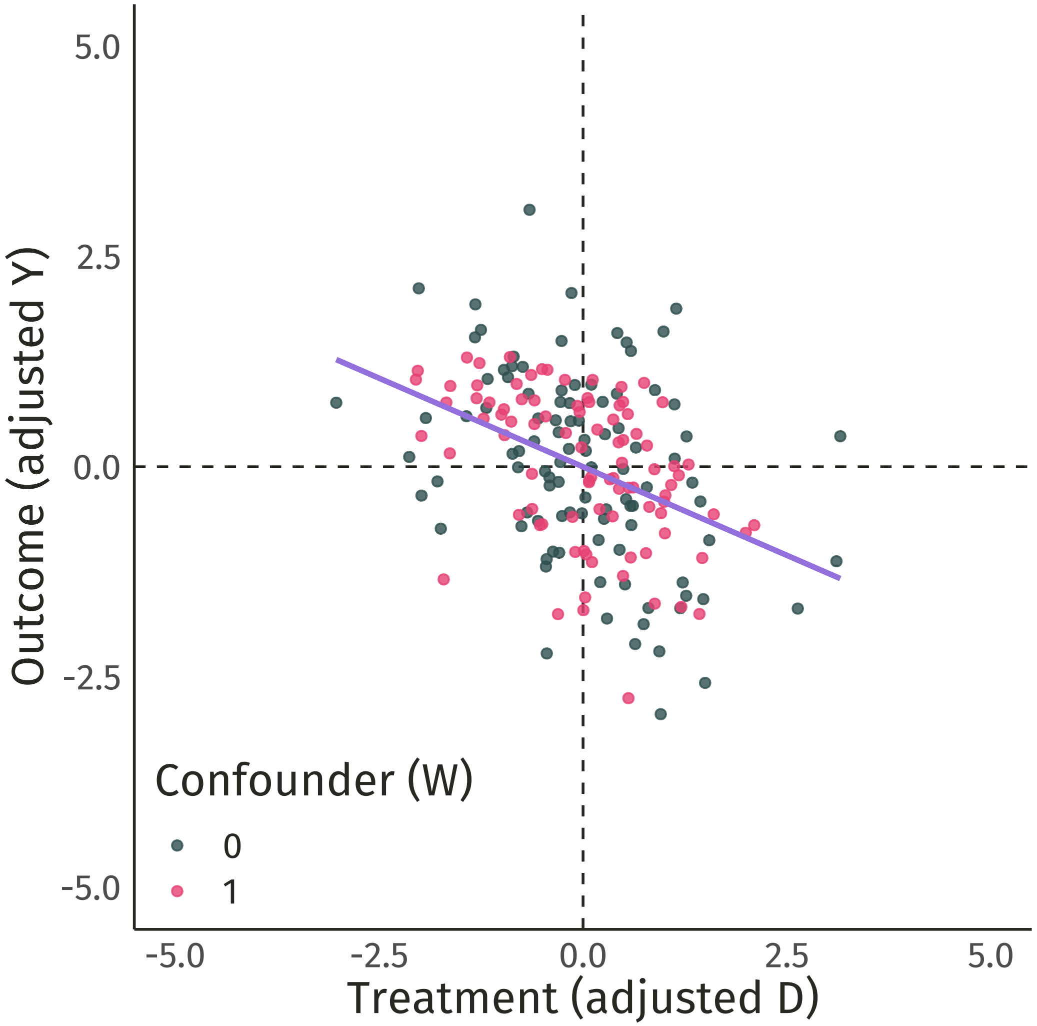
If data on the confounder exists, it can be added to the regression model:

$$Y_i = \alpha + \beta\, D_i + \gamma\, W_i + \varepsilon_i$$

Q: How does OLS "adjust" for the confounder?

- **Step 4:** Remove differences in Y explained by W.

Making adjustments



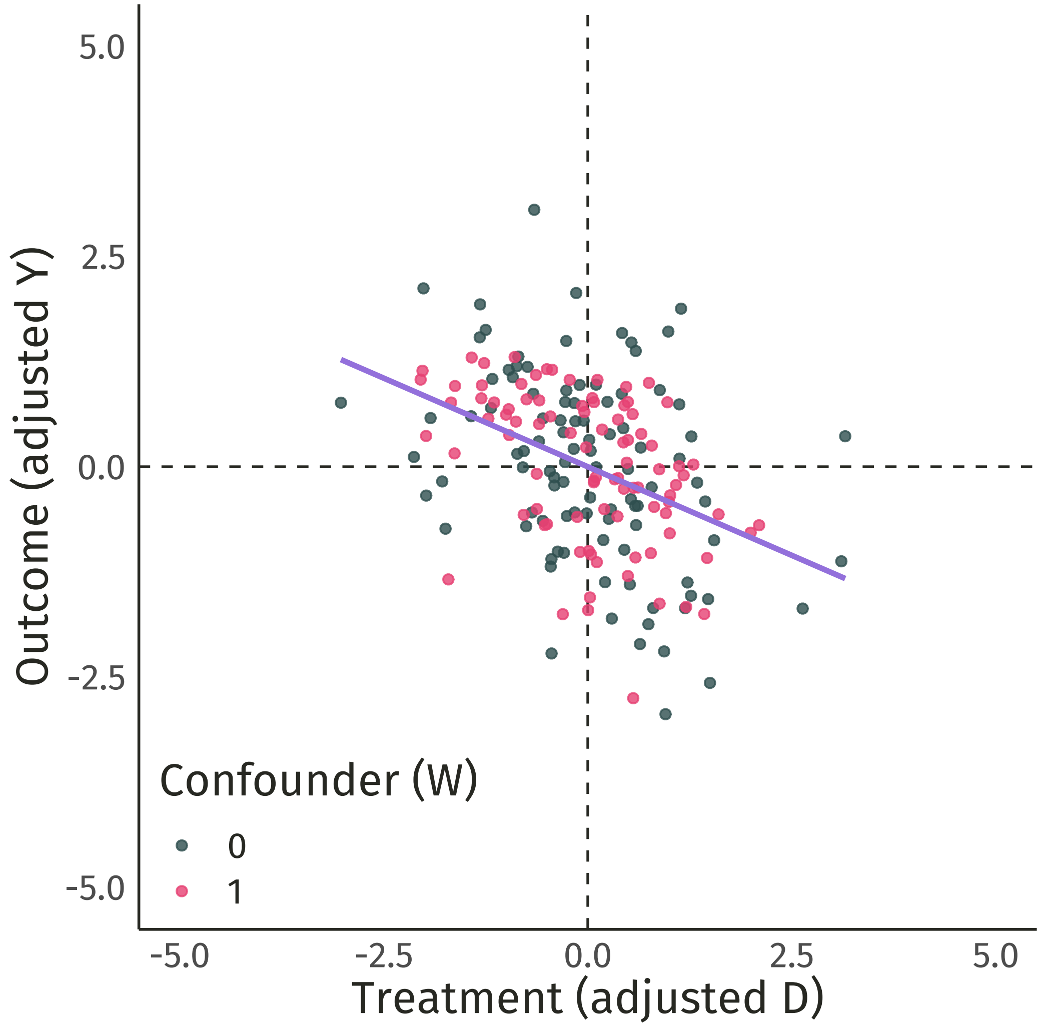
If data on the confounder exists, it can be added to the regression model:

$$Y_i = \alpha + \beta\, D_i + \gamma\, W_i + \varepsilon_i$$

Q: How does OLS "adjust" for the confounder?

- **Step 5:** Fit a regression through the adjusted data.

Making adjustments



If data on the confounder exists, it can be added to the regression model:

$$Y_i = \alpha + \beta\, D_i + \gamma\, W_i + \varepsilon_i$$

Parameter	**(1)**	**(2)**
Intercept	1.22	**0.9**
	(0.18)	**(0.1)**
Treatment	0.56	**-0.42**
	(0.08)	**(0.07)**
Confounder		**3.91**
		(0.2)

Standard errors in parentheses.

Omitted-variable bias

Example: Returns to education

Outcome: Weekly Earnings

Parameter	**1**	**2**
Intercept	146.95	**-128.89**
	(77.72)	**(92.18)**
Schooling (Years)	60.21	**42.06**
	(5.70)	**(6.55)**
IQ Score (Points)		**5.14**
		(0.96)

Standard errors in parentheses.

Bias from omitting IQ score

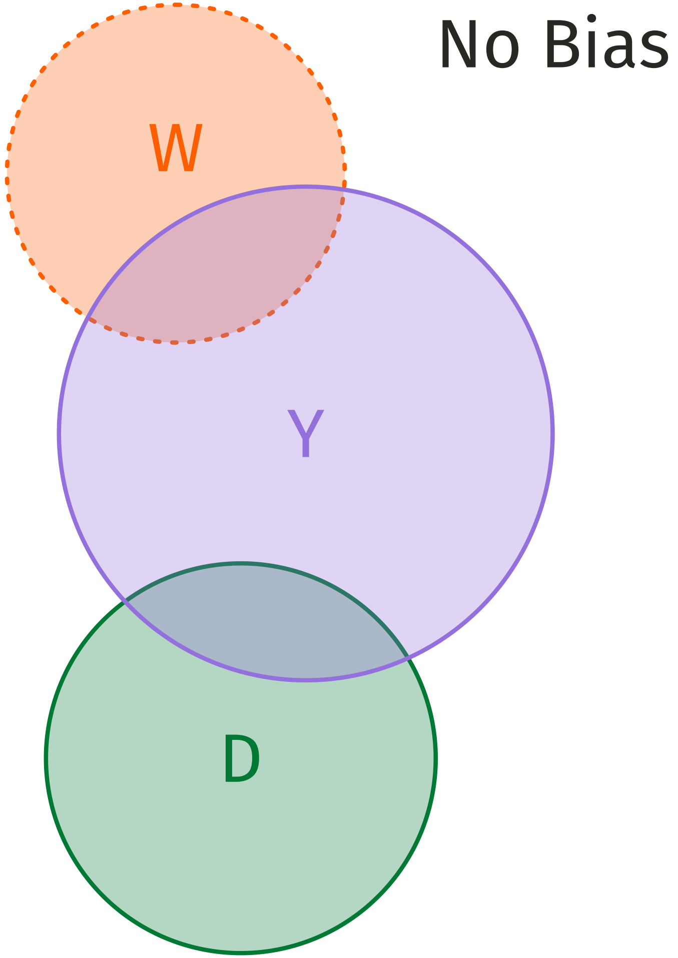
= "short" − "long"

= 60.21 − 42.06

= 18.15

The first regression mistakenly attributes some of the influence of intelligence to education.

Omitted-variable bias



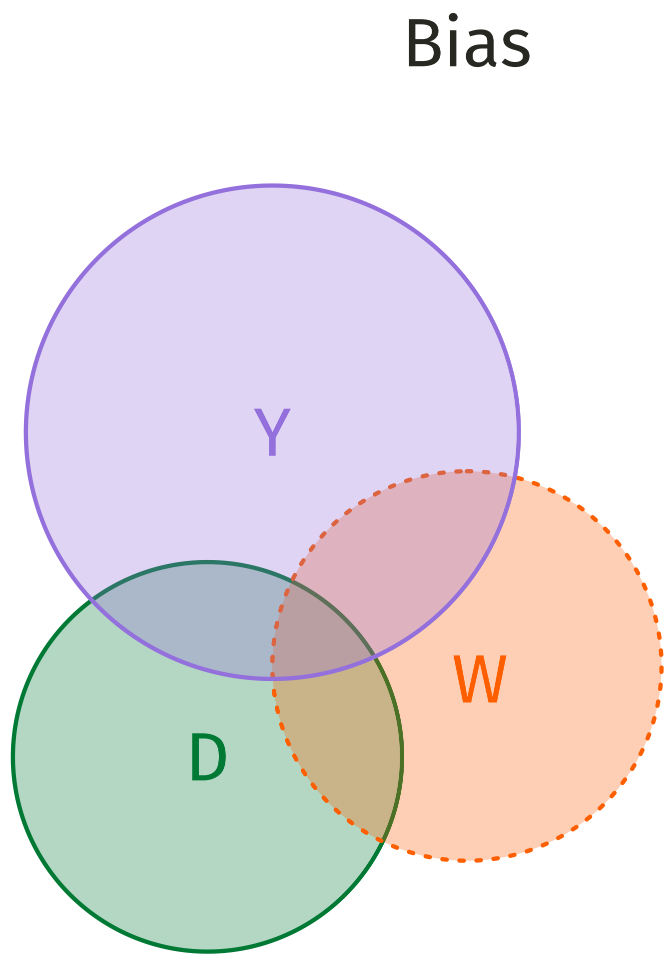
Y = Outcome

D = Treatment

W = Omitted variable

If **W** is correlated with both **D** and **Y** ⟶ omitted variable bias ⟶ regression fails to isolate the causal effect of **D** on **Y**.

Omitted-variable bias



Y = Outcome

D = Treatment

W = Omitted variable

If **W** is correlated with both **D** and **Y** ⟶ omitted variable bias ⟶ regression fails to isolate the causal effect of **D** on **Y**.

Housekeeping

MLK Jr. Day: No class or office hours on Monday the 17th.

Pre-recorded lecture for Wednesday the 19th.

- I will try to post it sometime next week.
- In the meantime, enjoy your weekend!

Assigned reading for next week: Snapping back: Food stamp bans and criminal recidivism by Cody Tuttle (2019).

- Best to read it *after* you watch next week's lecture.
- Reading Quiz 3 due the following week (Monday the 24th).

Problem Set 1 due on Friday the 21st by 11:59pm.

- Covers everything though next Wednesday.