Regression Logic

EC 320: Introduction to Econometrics

Kyle Raze
Fall 2019

Prologue

Housekeeping

Problem Set 2

- Analytical problems due by Tuesday at 17:00 on Canvas
- Computational problems due by Friday at 17:00 on Canvas

Midterm 1 next week (Wednesday)

Midterm review on Monday

Last Time

1. Fundamental problem of econometrics
2. Selection bias
3. Randomized control trials

Regression Logic

Regression

Economists often rely on (linear) regression for statistical comparisons.

- *"Linear"* is more flexible than you think.

Regression analysis helps us make *other things equal* comparisons.

- We can model the effect of X on Y while **controlling** for potential confounders.
- Forces us to be explicit about the potential sources of selection bias.
- Failure to control for confounding variables leads to **omitted-variable bias**, a close cousin of selection bias

Returns to Private College

Research Question: Does going to a private college instead of a public college increase future earnings?

- **Outcome variable:** earnings
- **Treatment variable:** going to a private college (binary)

Q: How might a private school education increase earnings?

Q: Does a comparison of the average earnings of private college graduates with those of public school graduates isolate the economic returns to private college education? Why or why not?

Returns to Private College

How might we estimate the causal effect of private college on earnings?

Approach 1: Compare average earnings of private college graduates with those of public college graduates.

- Prone to selection bias.

Approach 2: Use a matching estimator that compares the earnings of individuals the same admissions profiles.

- Cleaner comparison than a simple difference-in-means.
- Somewhat difficult to implement.
- Throws away data (inefficient).

Approach 3: Estimate a regression that compares the earnings of individuals with the same admissions profiles.

The Regression Model

We can estimate the effect of X on Y by estimating a **regression model**:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Y_i is the outcome variable.
- X_i is the treatment variable (continuous).
- u_i is an error term that includes all other (omitted) factors affecting Y_i.
- β_0 is the **intercept** parameter.
- β_1 is the **slope** parameter.

Running Regressions

The intercept and slope are population parameters.

Using an estimator with data on X_i and Y_i, we can estimate a **fitted regression line**:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- $\hat{Y}_i$ is the **fitted value** of Y_i.
- $\hat{\beta}_0$ is the **estimated intercept**.
- $\hat{\beta}_1$ is the **estimated slope**.

The estimation procedure produces misses called **residuals**, defined as $Y_i - \hat{Y}_i$.

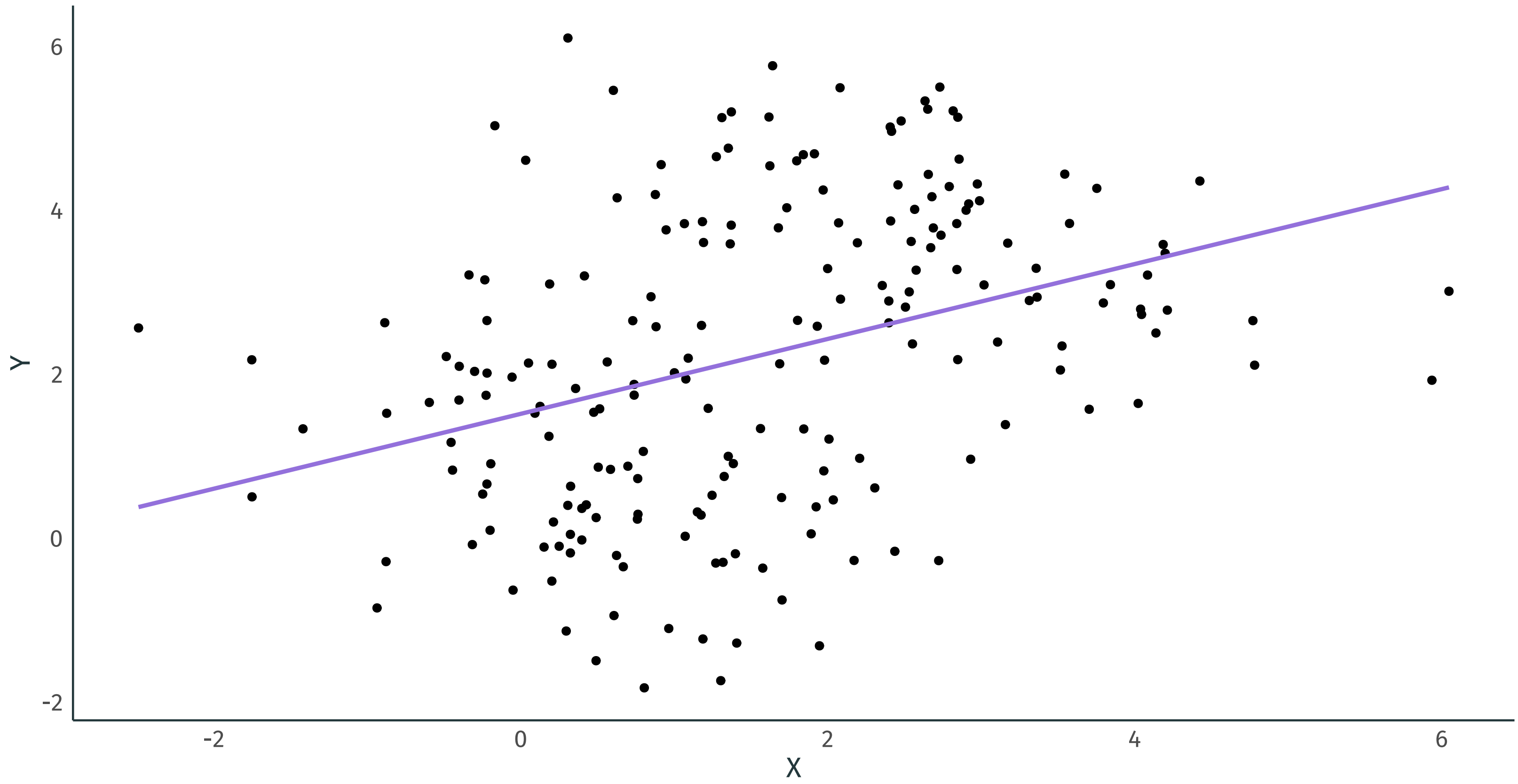
Running Regressions

In practice, we estimate the regression coefficients using an estimator called **Ordinary Least Squares** (OLS).

- Picks estimates that make $\hat{Y}_i$ as close as possible to Y_i given the information we have on X and Y.
- We will dive into the weeds after the midterm.

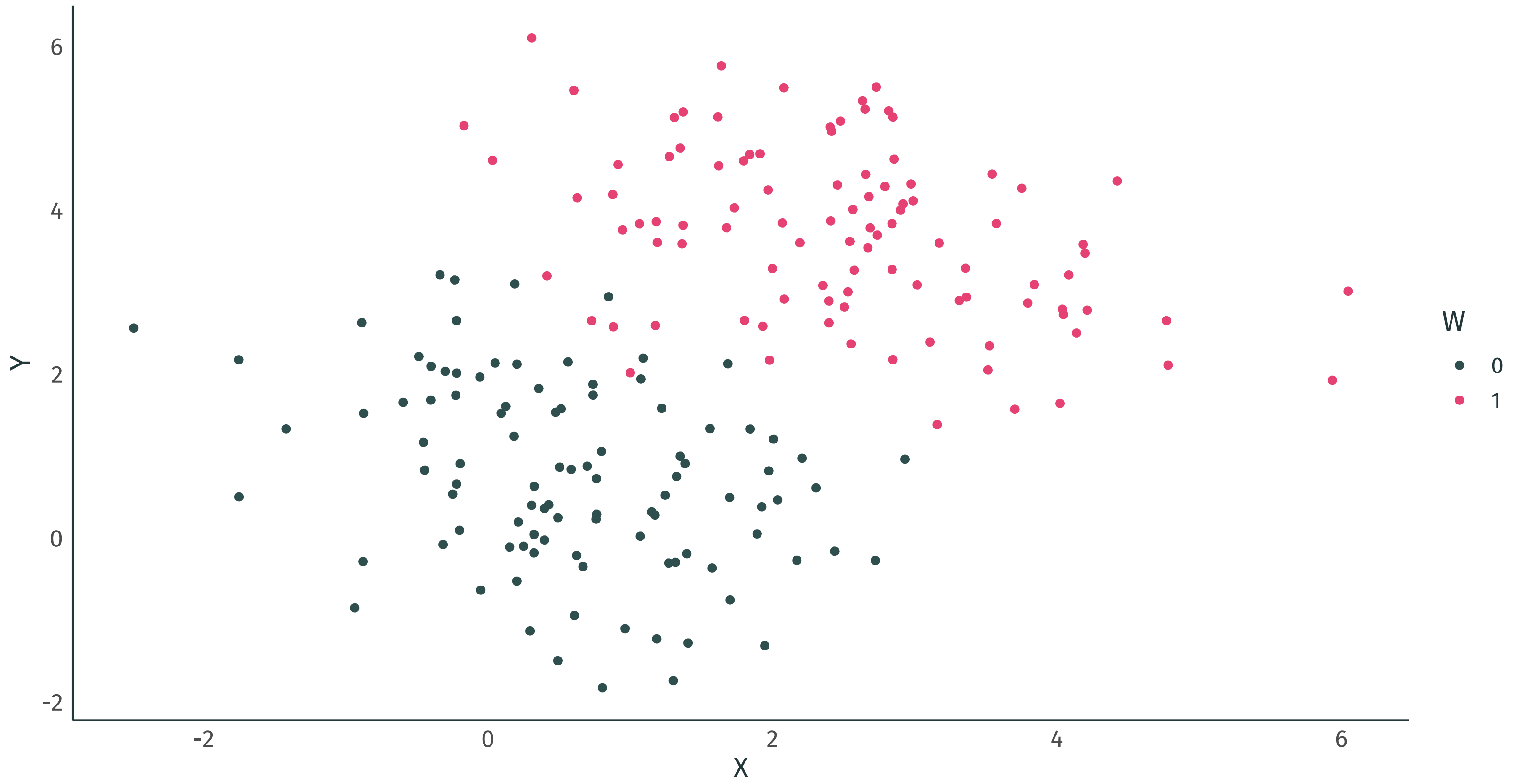
Running Regressions

OLS picks $\hat{\beta}_0$ and $\hat{\beta}_1$ that trace out the line of best fit. Ideally, we wound like to interpret the slope of the line as the causal effect of X on Y.



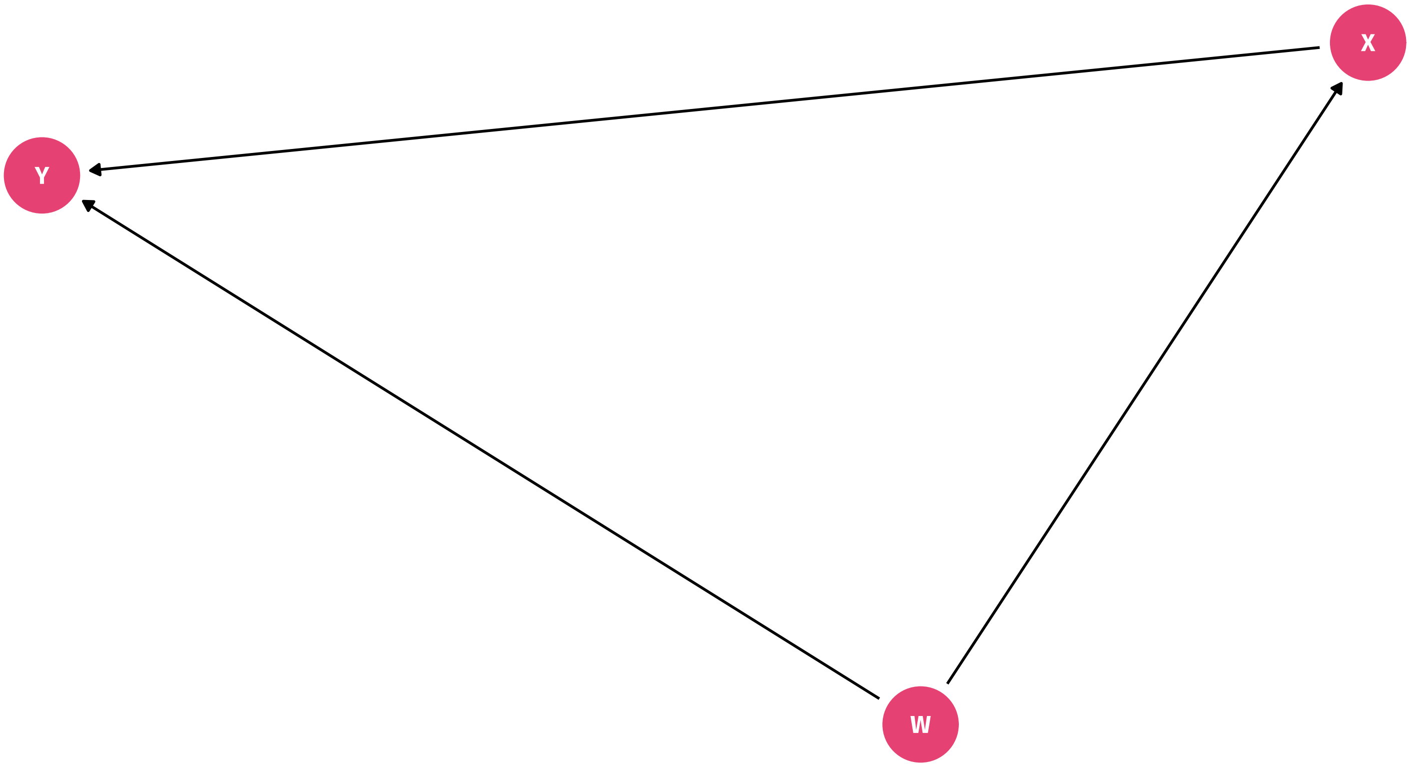
Confounders

However, the data are grouped by a third variable W. How would omitting W from the regression model affect the slope estimator?



Confounders

The problem with W is that it affects both Y and X. Without adjusting for W, we cannot isolate the causal effect of X on Y.



Controlling for Confounders

We can control for W by specifying it in the regression model:

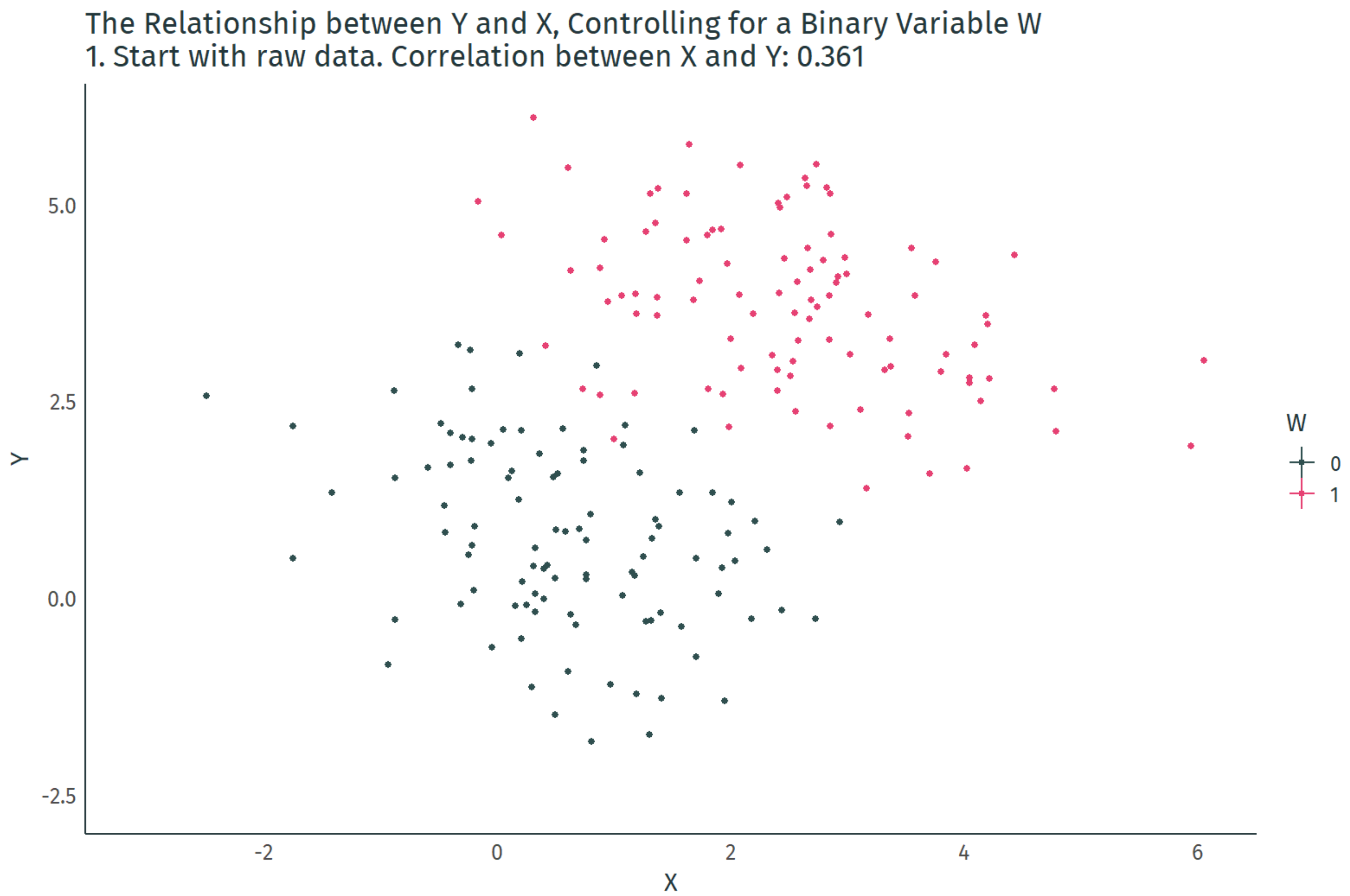
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

- W_i is a **control variable**.
- By including

  $$W_i$$

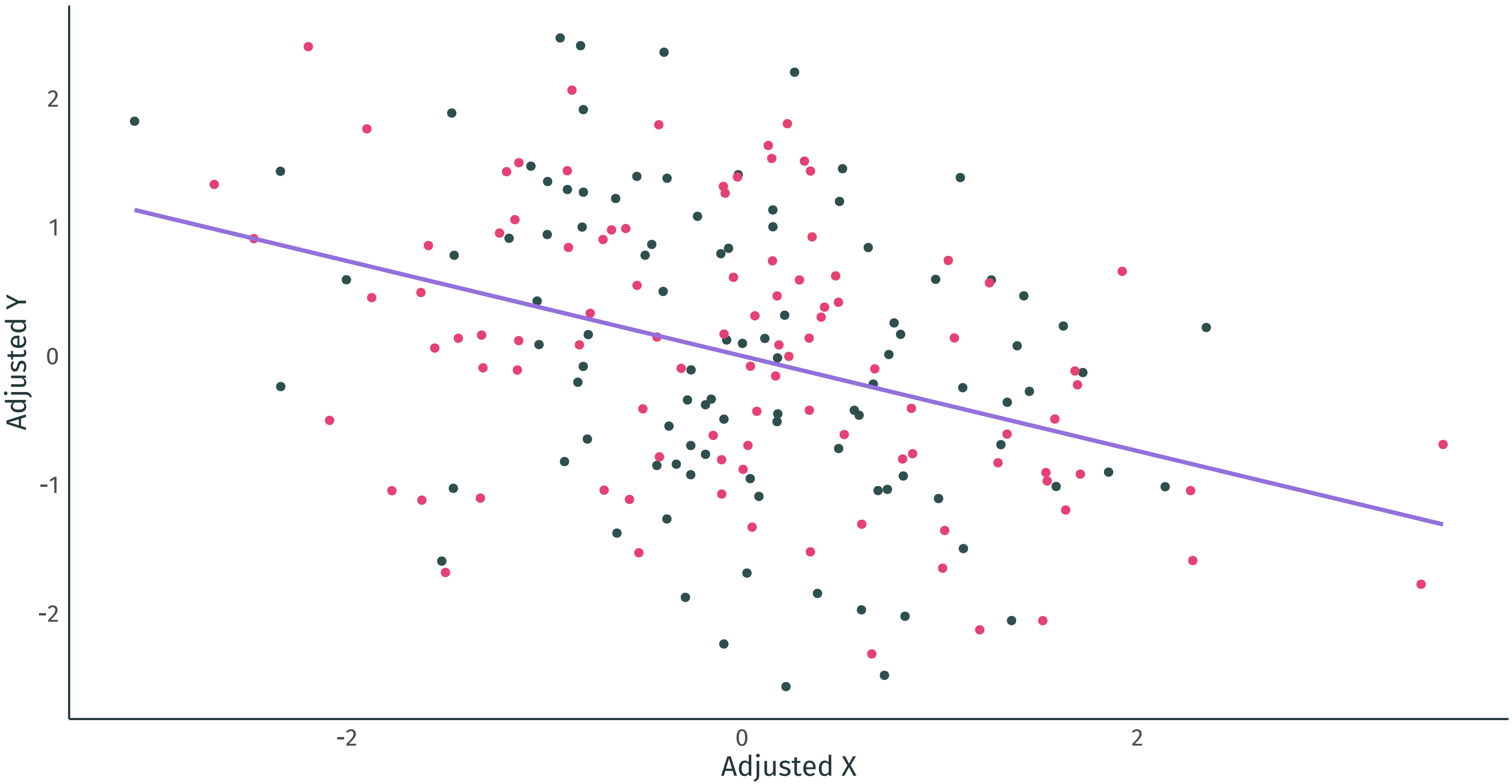
  in the regression, we can use OLS can difference out the confounding effect of W.
- **Note:** OLS doesn't care whether a right-hand side variable is a treatment or control variable, but we do.

Controlling for Confounders



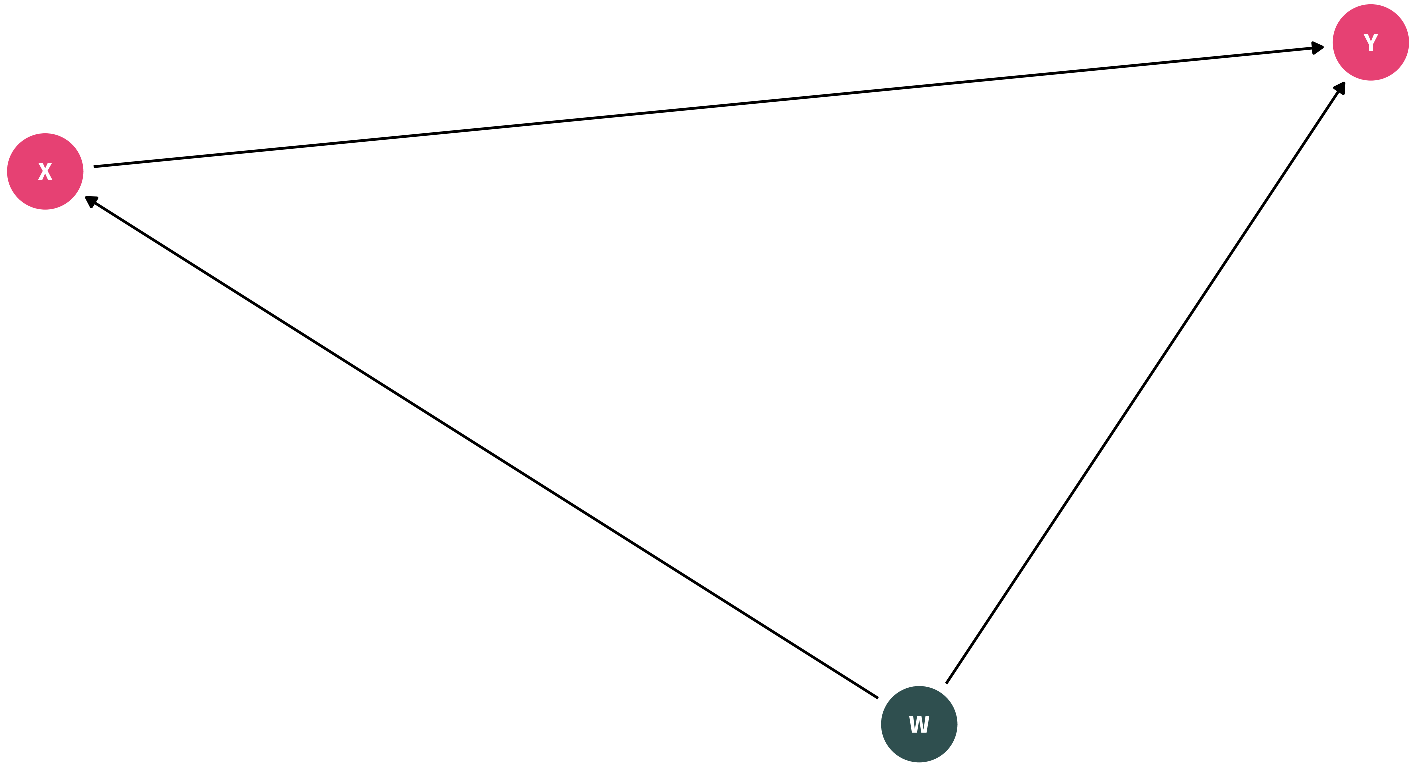
Controlling for Confounders

Controlling for W "adjusts" the data by **differencing out** the group-specific means of X and Y. **Slope of the estimated regression line changes!**



Controlling for Confounders

Can we interpret the estimated slope parameter as the causal effect of X on Y now that we've adjusted for W?



Controlling for Confounders

Example: Returns to schooling

Last class:

> **Q:** Could we simply compare the earnings those with more education to those with less?
> **A:** If we want to measure the causal effect, probably not.

What omitted variables should we worry about?

Example: Returns to schooling

Three regressions ***of*** wages ***on*** schooling.

Outcome variable: log(Wage)

Explanatory variable	**1**	**2**	**3**
Intercept	5.571	5.581	**5.695**
	(0.039)	(0.066)	**(0.068)**
Education	0.052	0.026	**0.027**
	(0.003)	(0.005)	**(0.005)**
IQ Score		0.004	**0.003**
		(0.001)	**(0.001)**
South			**-0.127**
			(0.019)

Omitted-Variable Bias

The presence of omitted-variable bias (OVB) precludes causal interpretation of our slope estimates.

We can back out the sign and magnitude of OVB by subtracting the slope estimate from a ***long*** regression from the slope estimate from a ***short*** regression:

$$\text{OVB} = \hat{\beta}_1^{\text{Short}} - \hat{\beta}_1^{\text{Long}}$$

Dealing with potential sources of OVB is one of the main objectives of econometric analysis!