

Regression Logic

EC 320: Introduction to Econometrics

Philip Economides

Winter 2022

Prologue

Housekeeping

Problem Set 2 due 01/24, which will address review content, fundamental thoughts and today's content.

First computational portion of problem sets, so make sure you've got **R** working **before Monday**.

So far we've identified the fundamental problem econometricians face. How do we proceed? **Regressions**

- Running models
- Confounders
- Omitted Variable Bias

Regression Logic

Regression

Modeling is about reducing something really complicated into something simple that still represents some part of the complicated reality.

- It's about telling stories that are easy to understand, and thus, easy to learn from

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

- "*Linear*" is more flexible than you think
- Describes the relationship between a dependent (endogenous) variable and one or more explanatory (exogenous) variable(s)

We will focus on the **simple univariate** case today.

Regression

Regression analysis helps us make *all else 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
- Why? The omitted variable, correlated with our covariate of interest, is sitting inside the error term causing chaos

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).
- β_0 is the **intercept** parameter. $\mathbb{E}[Y_i | X_i = 0] = \beta_0$
- β_1 is the **slope** parameter, which under the correct causal setting represents marginal change in X_i 's effect on Y_i . $\frac{\partial Y_i}{\partial X_i} = \beta_1$
- u_i is an error (disturbance) term that includes all other (omitted) factors affecting Y_i .

The Error term

u_i is quite special. If we consider the data generating process of variable Y_i , u_i captures all the unobserved variables that explain variation in Y_i .

- Always some error to our models, we just aim for it to be small relative to the challenge we face
- Some aspects of the observed data that was collected may also have been inputted correctly (measurement error)

The error term is the price we are willing to accept for a more simplified model.

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 $Y_i - \hat{Y}_i$.

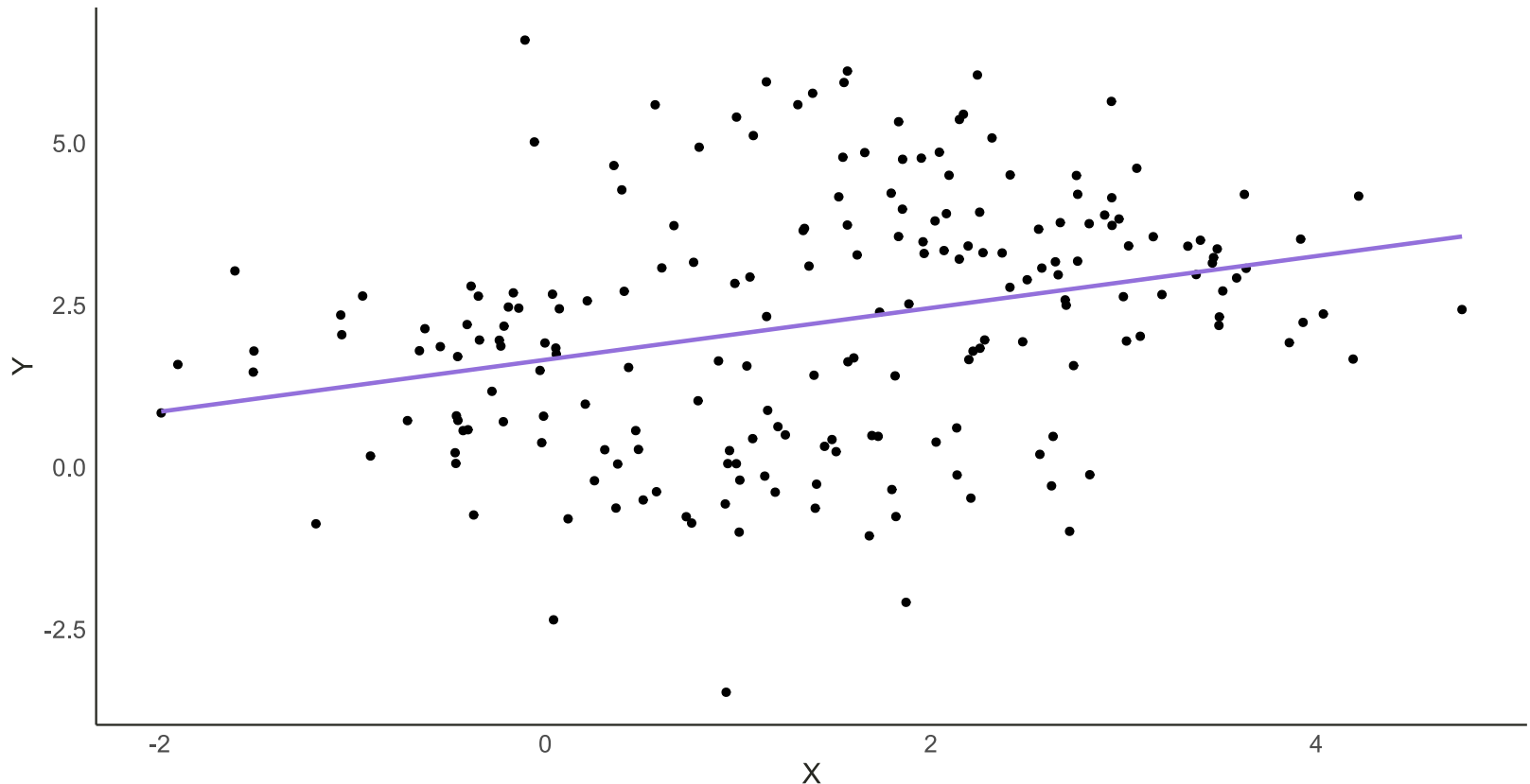
Running Regressions

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 .
- The sum of squared residuals, $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$, gives us an idea of how accurate our model is.
- **OLS** minimizes this sum.
- We will dive into the details next class.

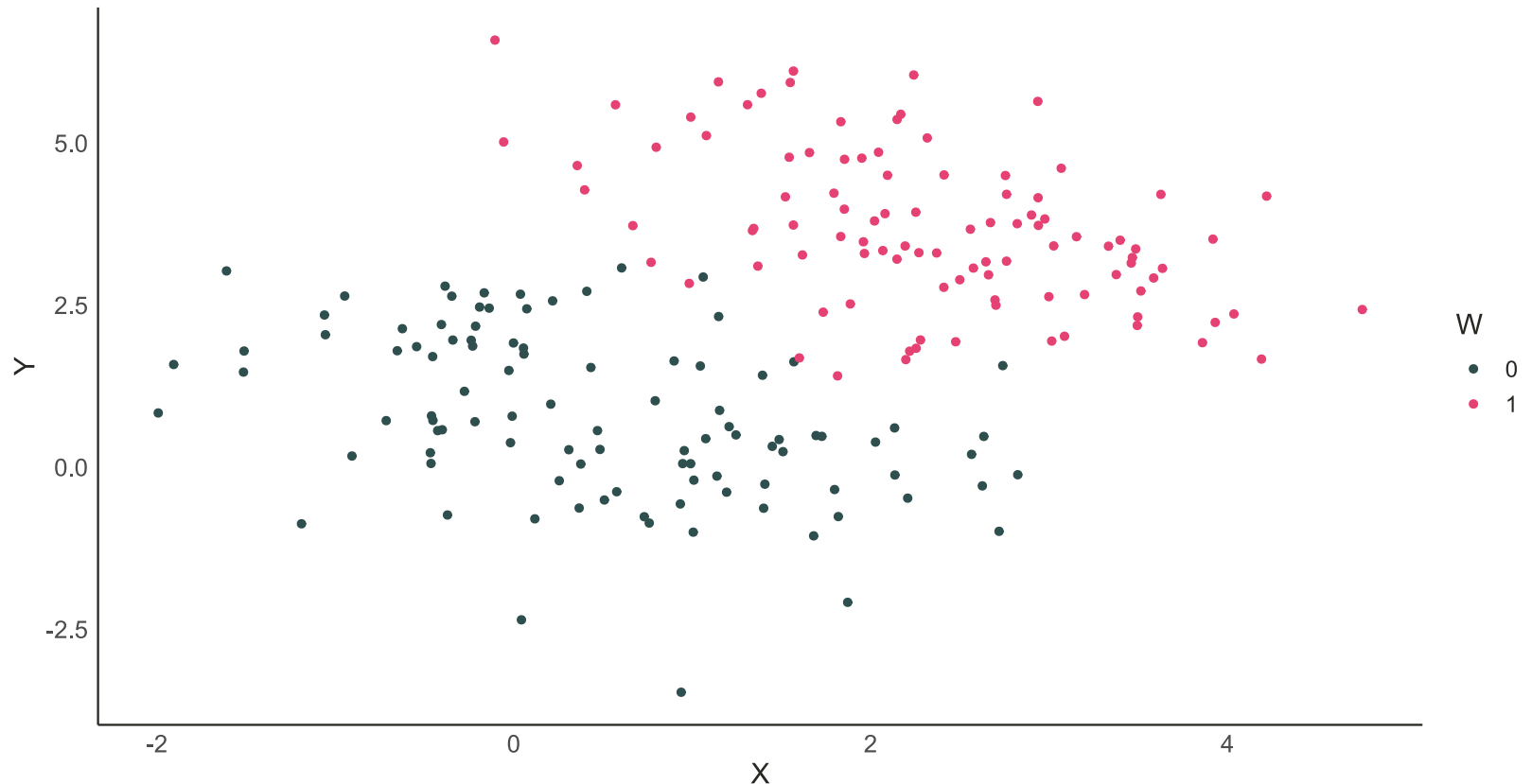
Running Regressions

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



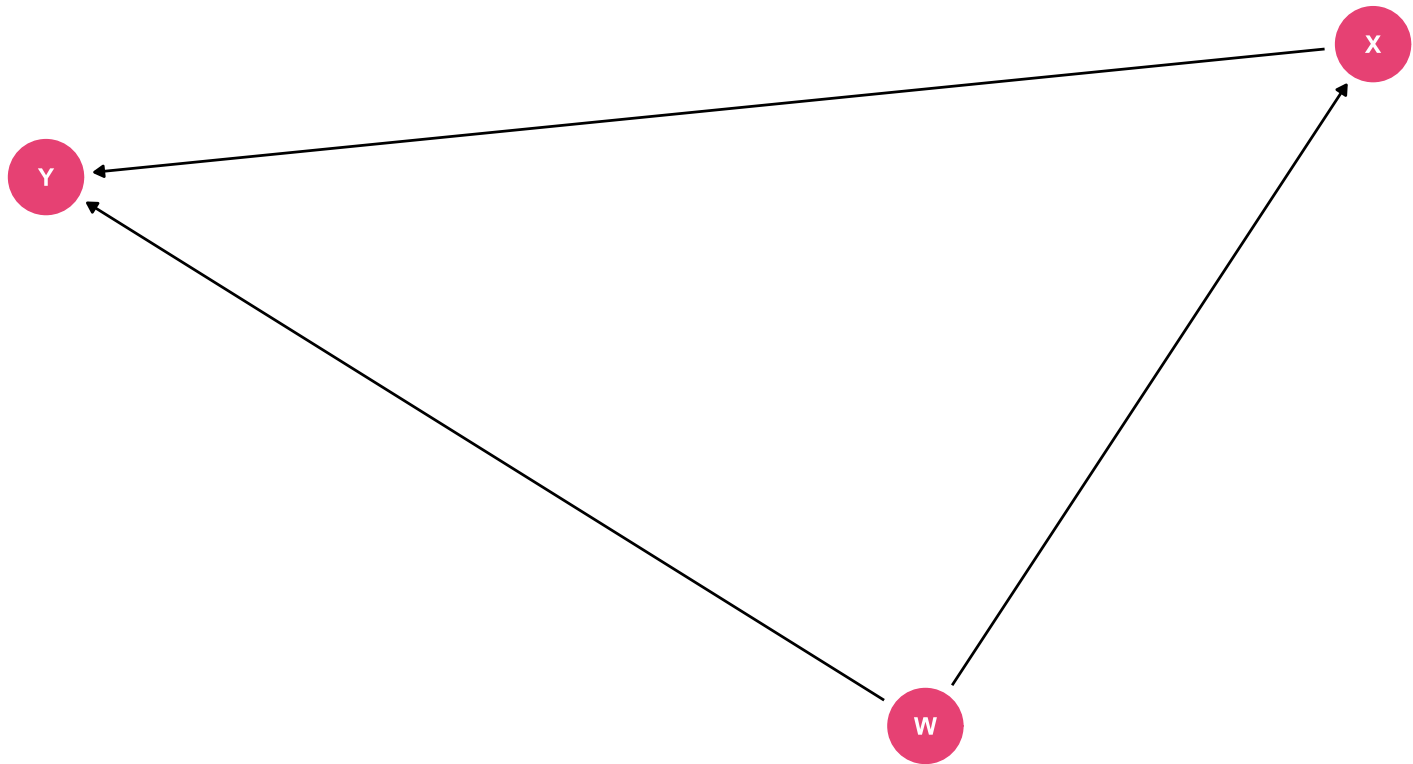
Confounders

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:

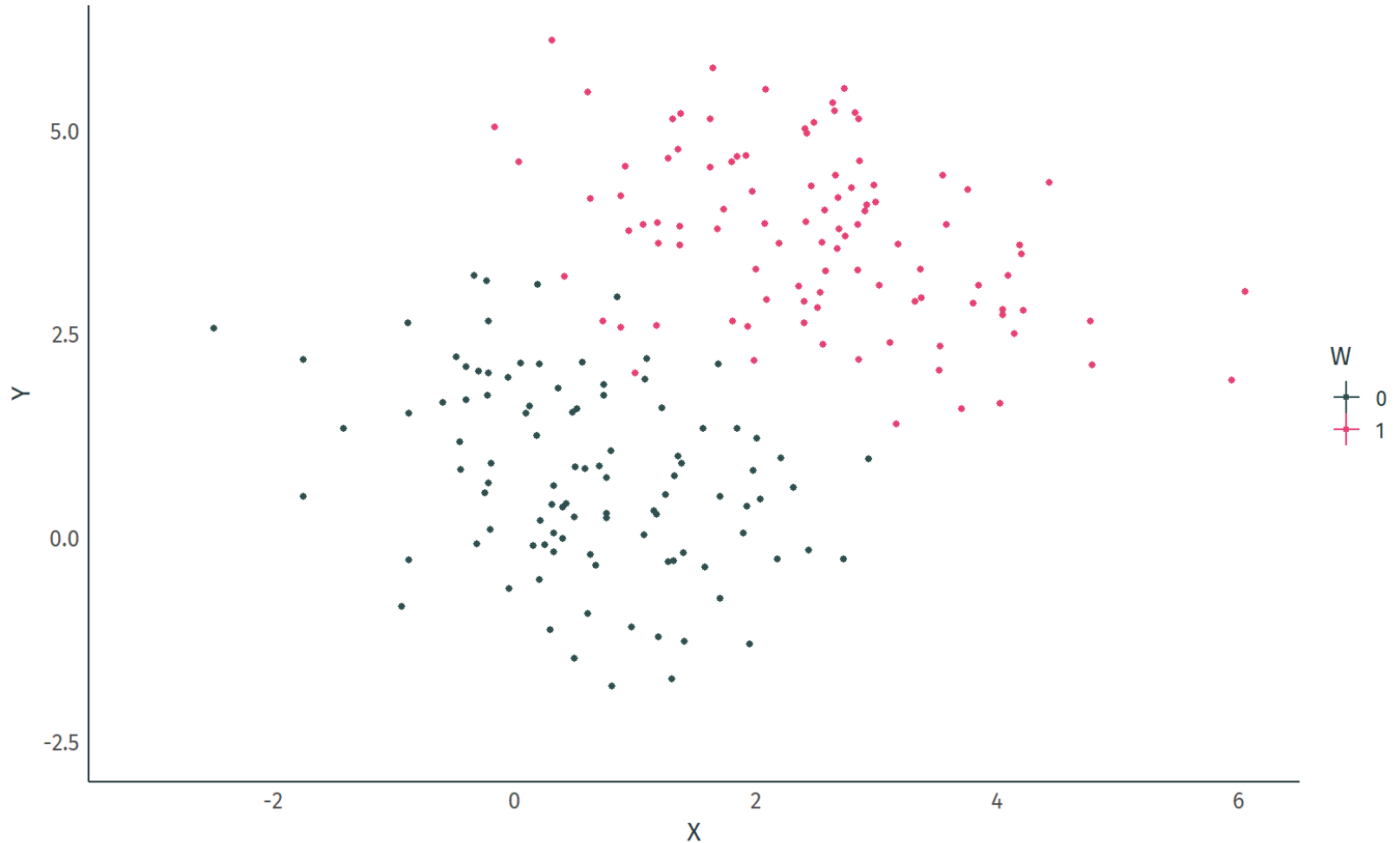
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

- 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

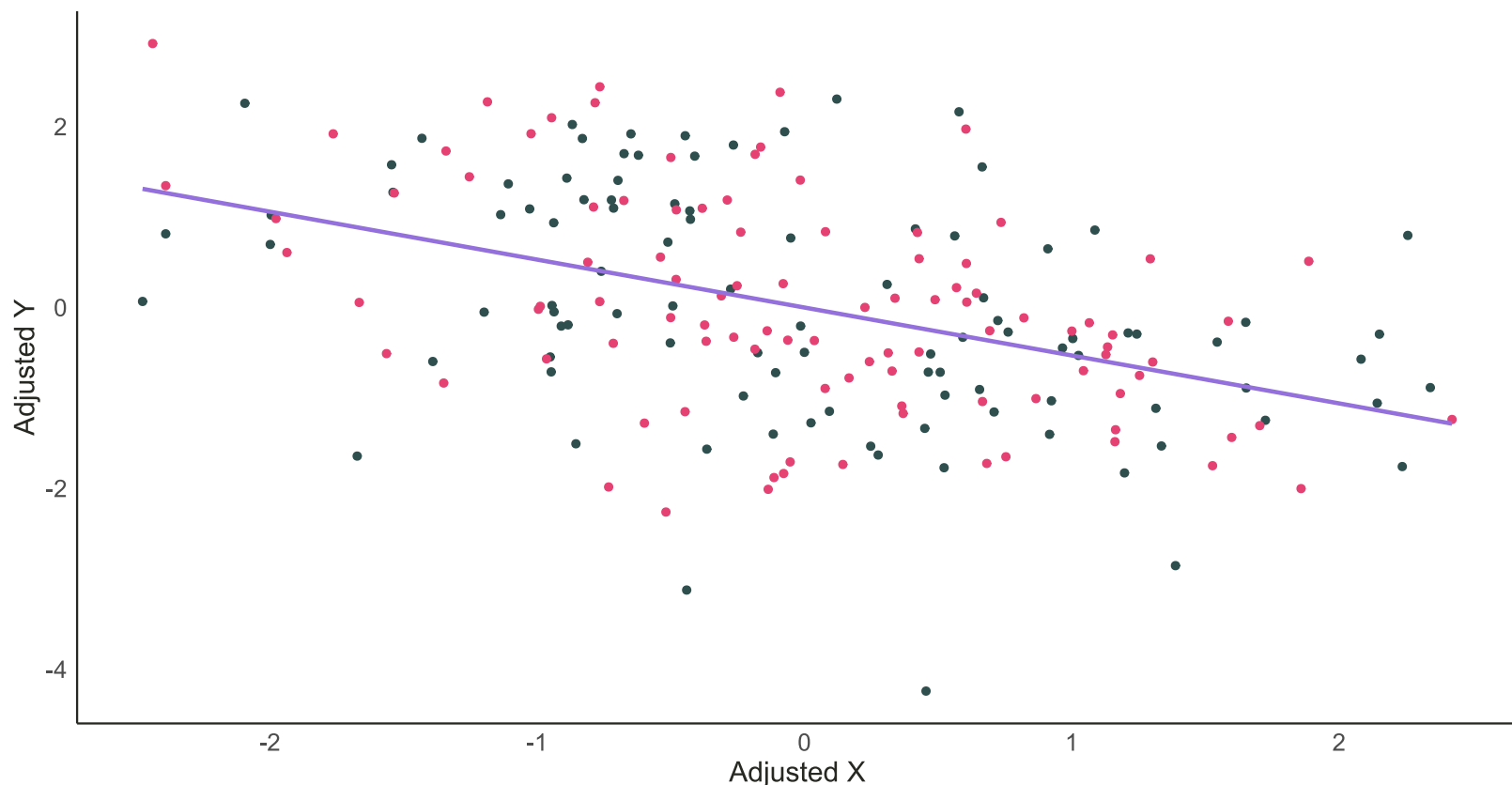
The Relationship between Y and X, Controlling for a Binary Variable W

1. Start with raw data. Correlation between X and Y: 0.361



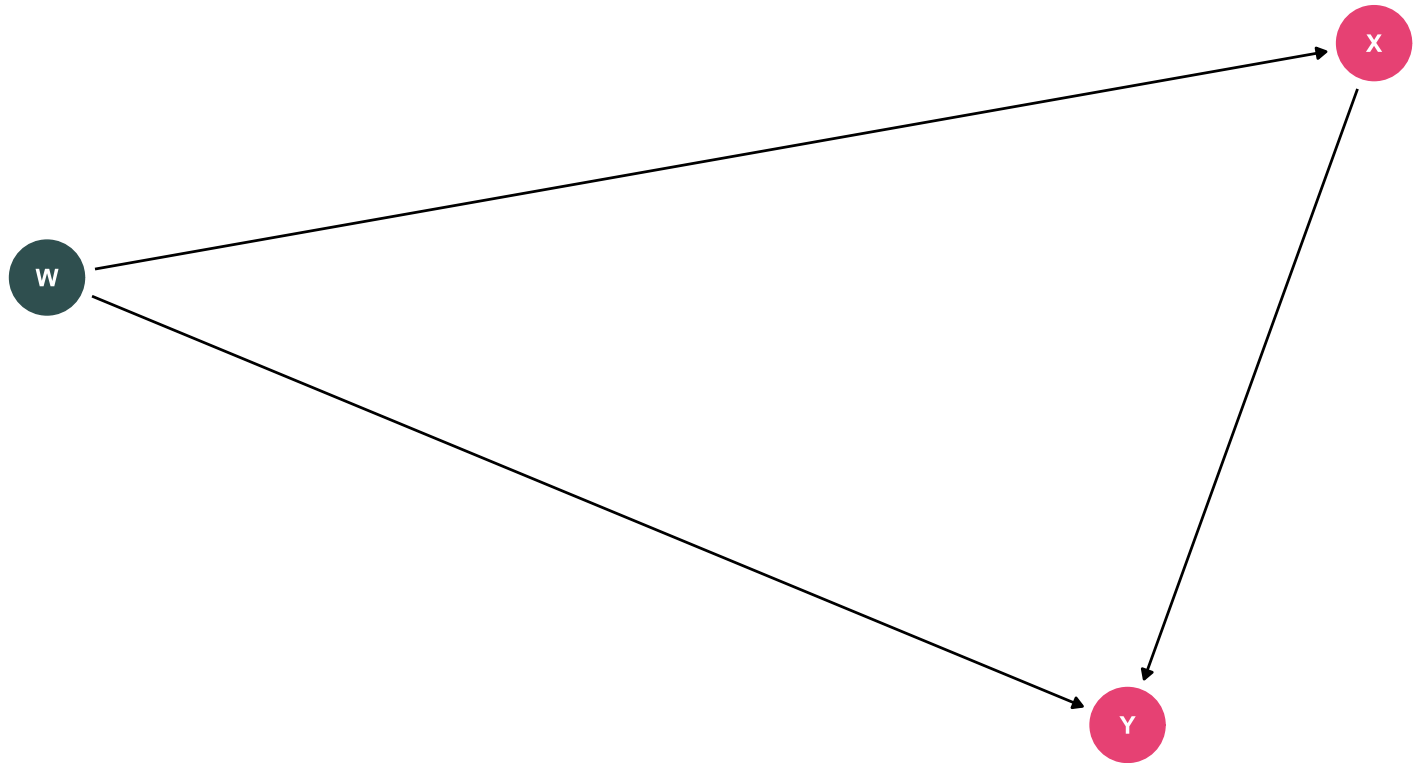
Controlling for Confounders

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?

Controlling for Confounders

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!

OVB vs. Irrelevant Variables

So if we risk bias as a result of excluding a variable, why not throw every possible variable and transformation of variables (log-linearized, squared, inverted) at the model?

- Time consuming
- Data not always available
- Irrelevant variables actually make matters **worse**

OVB vs. Irrelevant Variables

How can more variables cause trouble? **Loss of efficiency** in estimator while still unbiased.

- This is the classic **multicollinearity** problem
- If an irrelevant variable is highly correlated with your explanatory variable of interest, the standard error will increase
- Inference of the coefficient's significance becomes muddled by higher standard error term
- More details on what this looks like statistically next week

Summary

What to remember

- Regressions are models of how we imagine the data generating process plays out
- They are usually simplifications of real life observations
- A linear regression fits a line through the data to reveal the relationship between treatment and outcome
- Confounders, omitted variables and irrelevant variables all pose risks to the identification challenge involved in estimating a population parameter of interest in our regression model
- OLS is the most common algorithm for estimating regressions, and that is what our next lecture will focus on