

# Omitted Variables Bias (OVV) and Causal Inference

---

ECON 490

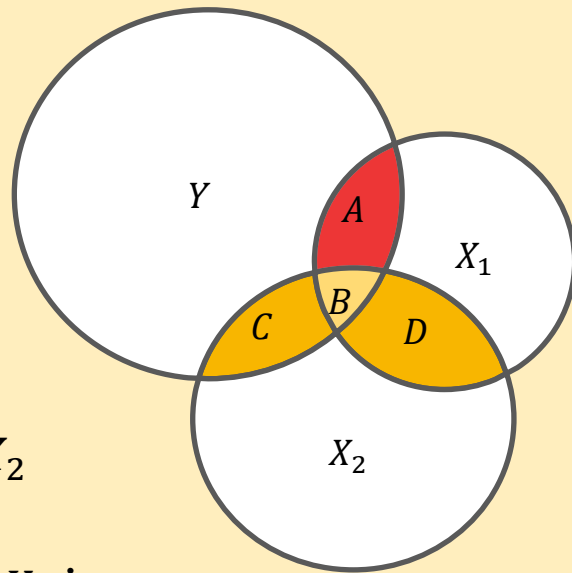
Taylor Mackay || Email: [tmackay@fullerton.edu](mailto:tmackay@fullerton.edu)

# Slides Overview

In these slides, we'll discuss:

- Defining omitted variables bias (OVB)
- Introducing causal inference
- Fixed effects and OVB

# Isolating Variation



$A$  represents covariance of  $Y$  and  $X_1$  that is unrelated to  $X_2$

In a regression of  $Y$  on  $X_1$  and  $X_2$ , our estimated effect of  $X_1$  is determined by the **unique** effect of  $X_1$  on  $Y$

The question for this week – what happens if we **don't** control for  $X_2$ ?

# Omitted Variables Bias (OVB)

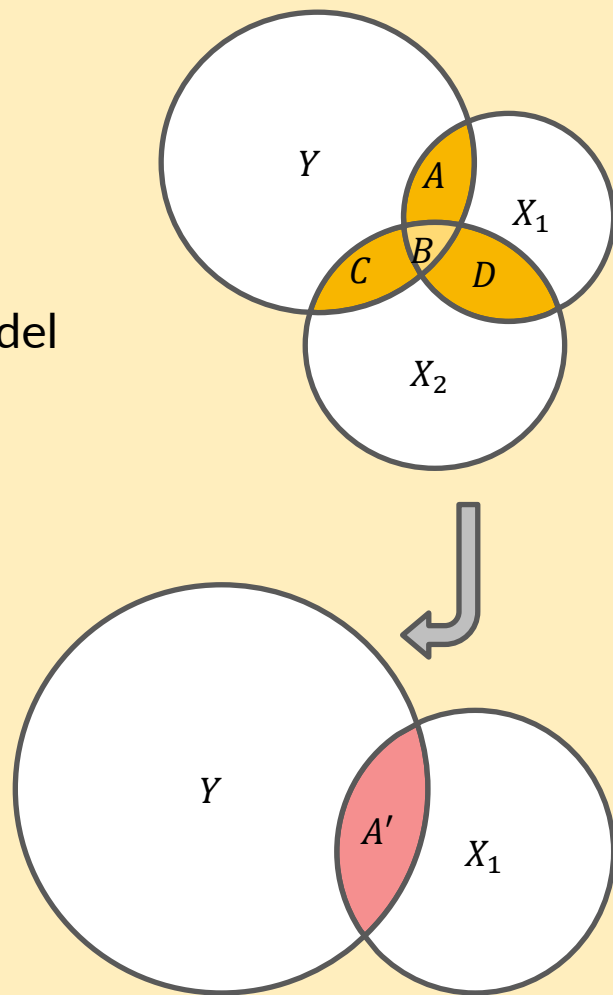
Let's suppose our initial Venn diagram is the “true” model

- What happens if we don't have data for  $X_2$ ?
- Then we're stuck with just observing  $X_1$

The “effect” of  $X_1$  in our **new** diagram is  $A'$

- But we know that's too big – it includes  $B$ !

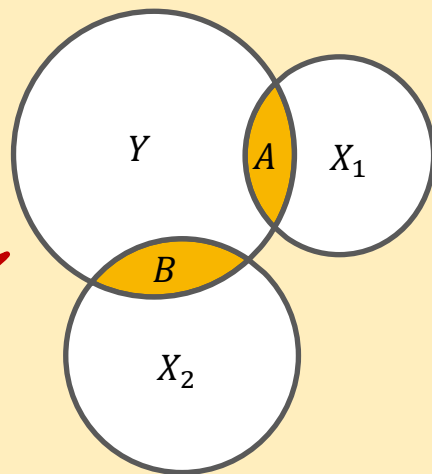
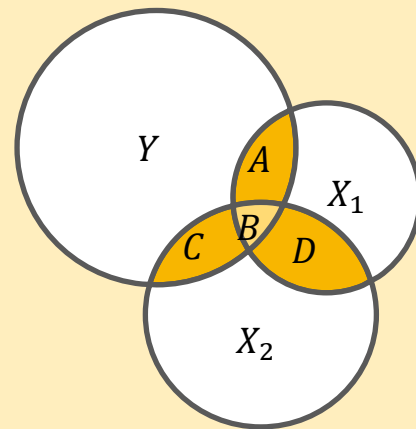
We **incorrectly** estimate the effect of  $X_1$  because we **couldn't** observe  $X_2$  – this is OVB



# Two Key Components of OVB

OVB requires **two** conditions:

- 1) Omitted  $X_2$  must covary with  $Y$
- 2) Omitted  $X_2$  must covary with  $X_1$



*If this is the true model,  
there's no OVB!*

# Direction of Bias

The **bias** in OVB means that our estimated OLS coefficient is wrong

- In other words, our estimated regression coefficient is too big or too small
- The direction of the bias depends on the underlying relationships

Suppose the true model is given by  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

We can refer to the following as the OVB model:  $Y = \alpha_0 + \alpha_1^{OVB} X_1 + e$

# Direction of Bias

Given the definitions from the last slide, we can compare  $\beta_1$  to  $\alpha_1^{OVB}$

- Table tells us, “What happens if we omit  $X_2$  from our regression?”

	$X_1$ and $X_2$ are <b>positively</b> correlated	$X_1$ and $X_2$ are <b>negatively</b> correlated
$Y$ and $X_2$ are <b>positively</b> correlated	Positive Bias ( $\beta_1 < \alpha_1^{OVB}$ )	Negative Bias ( $\beta_1 > \alpha_1^{OVB}$ )
$Y$ and $X_2$ are <b>negatively</b> correlated	Negative Bias ( $\beta_1 > \alpha_1^{OVB}$ )	Positive Bias ( $\beta_1 < \alpha_1^{OVB}$ )

# OVV and Descriptive vs. Causal Statements

Descriptive statement = “*The correlation between  $X_1$  and  $Y$  is  $A$* ”

- OVB provides helpful context – we know other factors probably matter!
- But claims are still informative – exploring data, relationships, patterns, etc.

Causal statement = “ $X_1$  **causes** *A change in  $Y$* ”

- Now OVB is a big deal – **can't** make (credible) causal claims if we have OVB
- Causal inference is broadly a set of tools for dealing with OVB



# OVB Example

What's the impact of a financial grant program on student GPAs?

- Give some students grants in Fall but not others, observe GPAs in Spring
- We want to estimate the following:

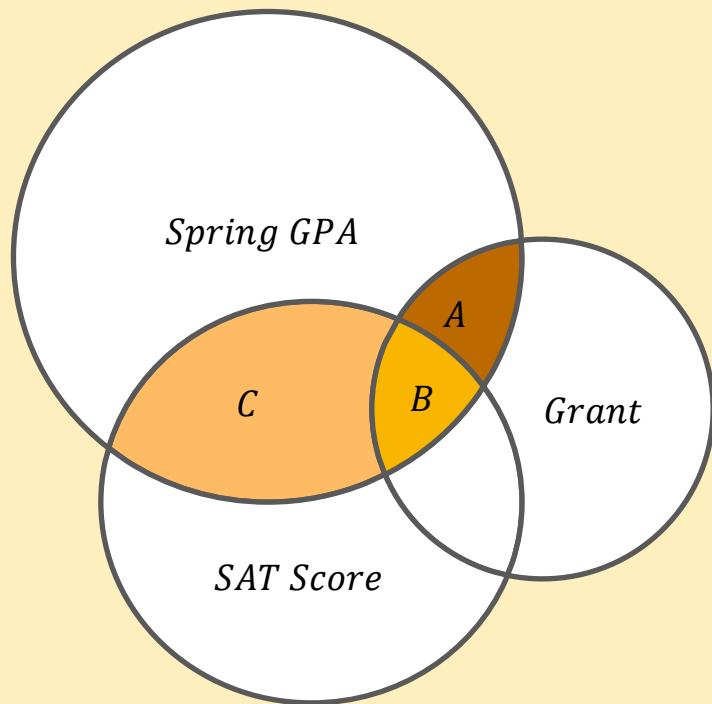
$$\text{Spring GPA} = \alpha_0 + \alpha_1 \text{Fall Grant} + u$$

Two possible ways of assigning grants:

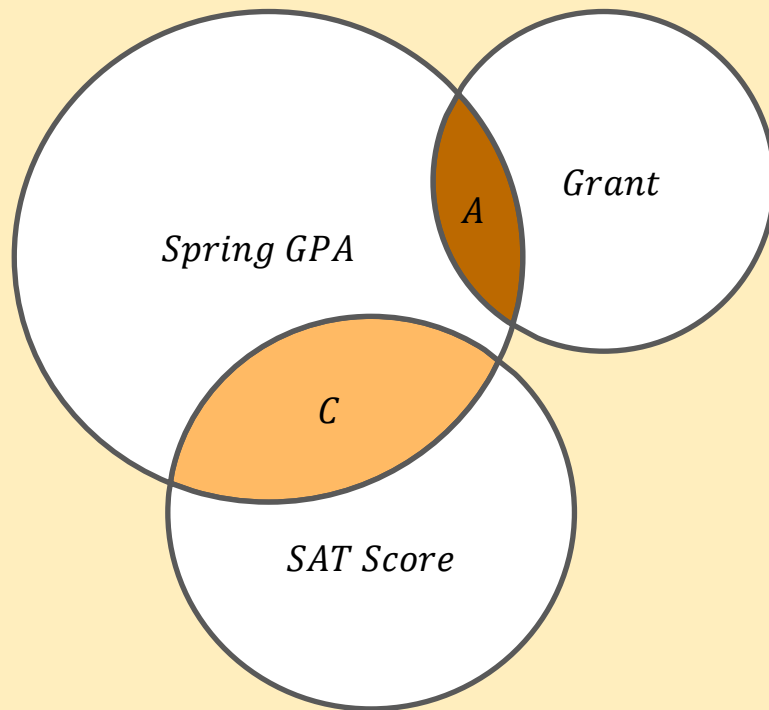
- Give to students with highest SAT scores
- Assign via coin flip

# OVB Example

Assigned to highest SAT scores



Assigned via coinflip



# Causal Inference without Randomization

Randomization solves the OVB problem by assigning values of  $X_1$

- Implemented via Randomized Control Trials (RCTs)
- Classic example = pharmaceutical trials

Economics often involves questions where randomization isn't feasible

Causal inference with observational data is the process of “finding” randomization

# Causal Inference with Observational Data

“Finding” randomization broadly means one of two things:

1. Identifying situations where treatment assignment is effectively random
2. Controlling for omitted variables to address OVB problem

We can broadly group causal inference models into the above categories

- (1) Includes instrumental variables (IV) & regression discontinuity (RD)
- (2) Includes fixed effects, event studies, diff-in-diff (DiD), etc.

# Group (1) – Finding Randomized Treatment

Instrumental Variables (IV) typically gets a lot of coverage in metrics classes

- Used to be more popular in applied research than it is now
- Why? Mix of new methods, potential for things to go wrong, etc.

Regression Discontinuity (RD) tries to “imitate” RCTs

- If we can't randomize treatment, find something that does it for us
- Laws, administrative rules, etc. that assign treatment based on cutoffs

# Regression Discontinuity (RD)

Let's return to our financial grants example

- Suppose everyone with  $SAT > 1200$  gets the grant
- OVB intuition – getting a grant is correlated with omitted factors

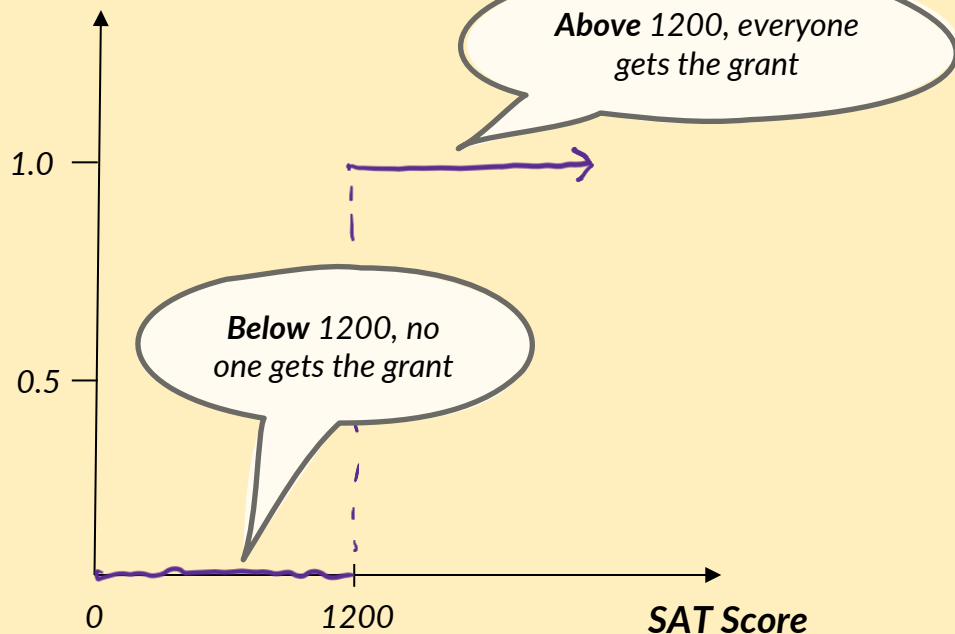
But what if we “zoom in” on SAT scores right around 1200?

- Difference between getting a 1190 vs. 1200 is mostly just luck
- Misread a single question, accidentally mark the wrong answer, etc.

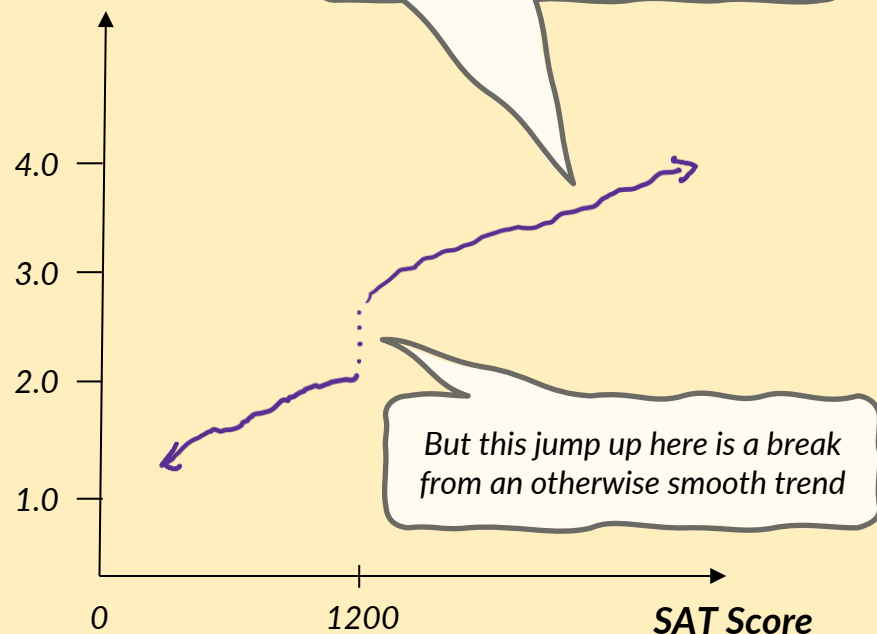
Grant assignment *effectively random* around 1200 threshold

# RD in Two Pictures

$Pr(\text{Getting Grant} \mid \text{SAT Score})$



Spring GPA



## Group (2) – Controlling for Omitted Variables

As a starting point, consider the following:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + u$$

Suppose we *knew* that this was the “true” regression equation

- Would we need to worry about OVB?
- No! Regression identifies *unique* effect of  $X_1$  *controlling* for  $X_2$



## Group (2) – Controlling for Omitted Variables

In practice, we almost never know the “true” model

- How do we know what to control for?
- Think about the potential “structure” of omitted variables

Do we think omitted variables...

- Differ across groups in a fixed or constant fashion? (*fixed effects*)
- Differ by group and across time? (*DiD, event studies, etc.*)

We can see how this works by considering an example using fixed effects

# Fixed Effects Example

What's the relationship between hours spent studying and test scores?

- Suppose we've got data for two students, A and B, for two tests
- Gives us (1) time studying and (2) test scores

Student	Test	Hours Studying	Test Score
A	1	2	90
A	2	3	96
B	1	5	78
B	2	7	83

# Fixed Effects Example

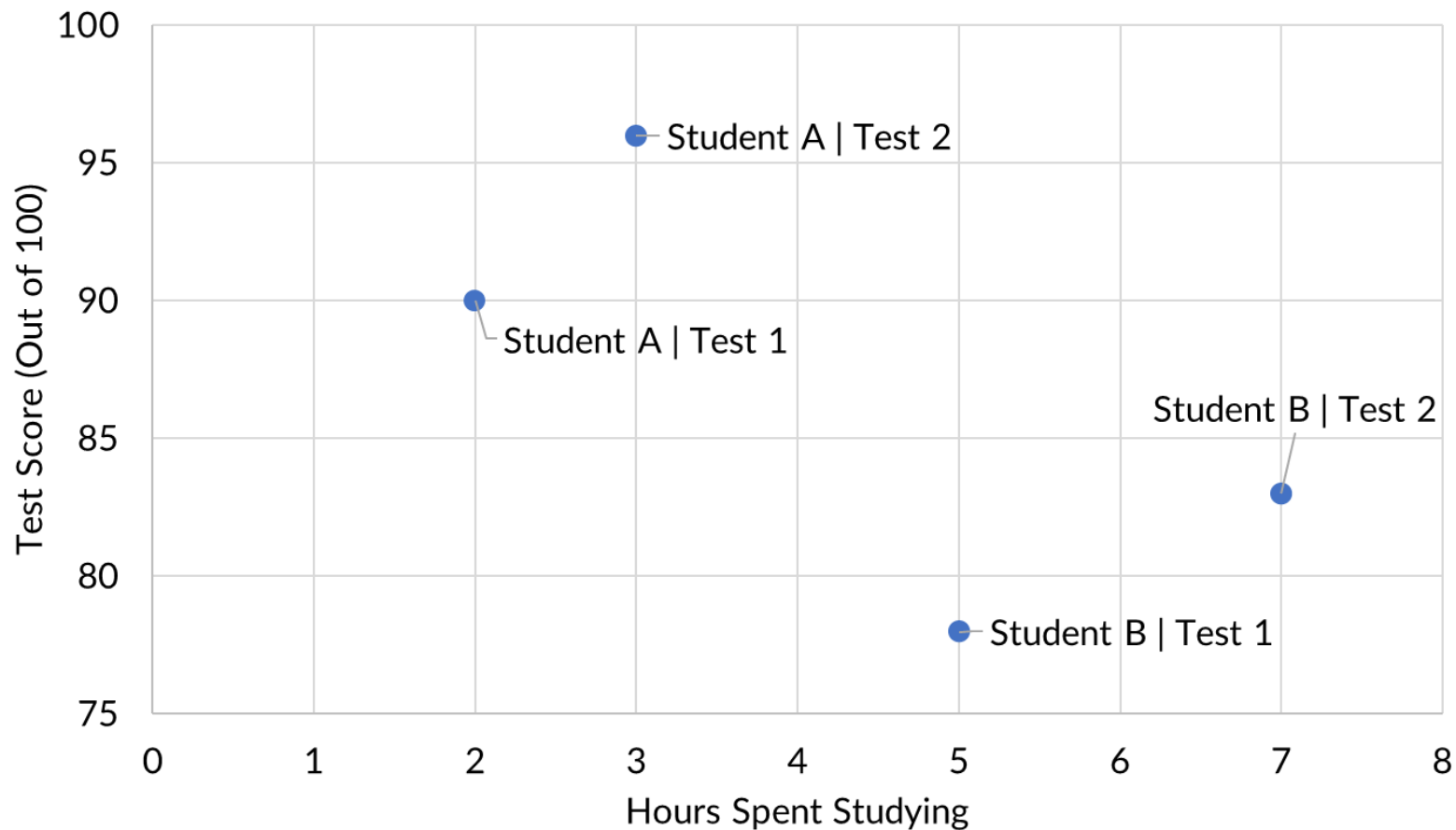
Given our data set, we can run the following regression:

$$Test\ Score_i = \beta_0 + \beta_1 Hours\ Studying_i + u_i$$

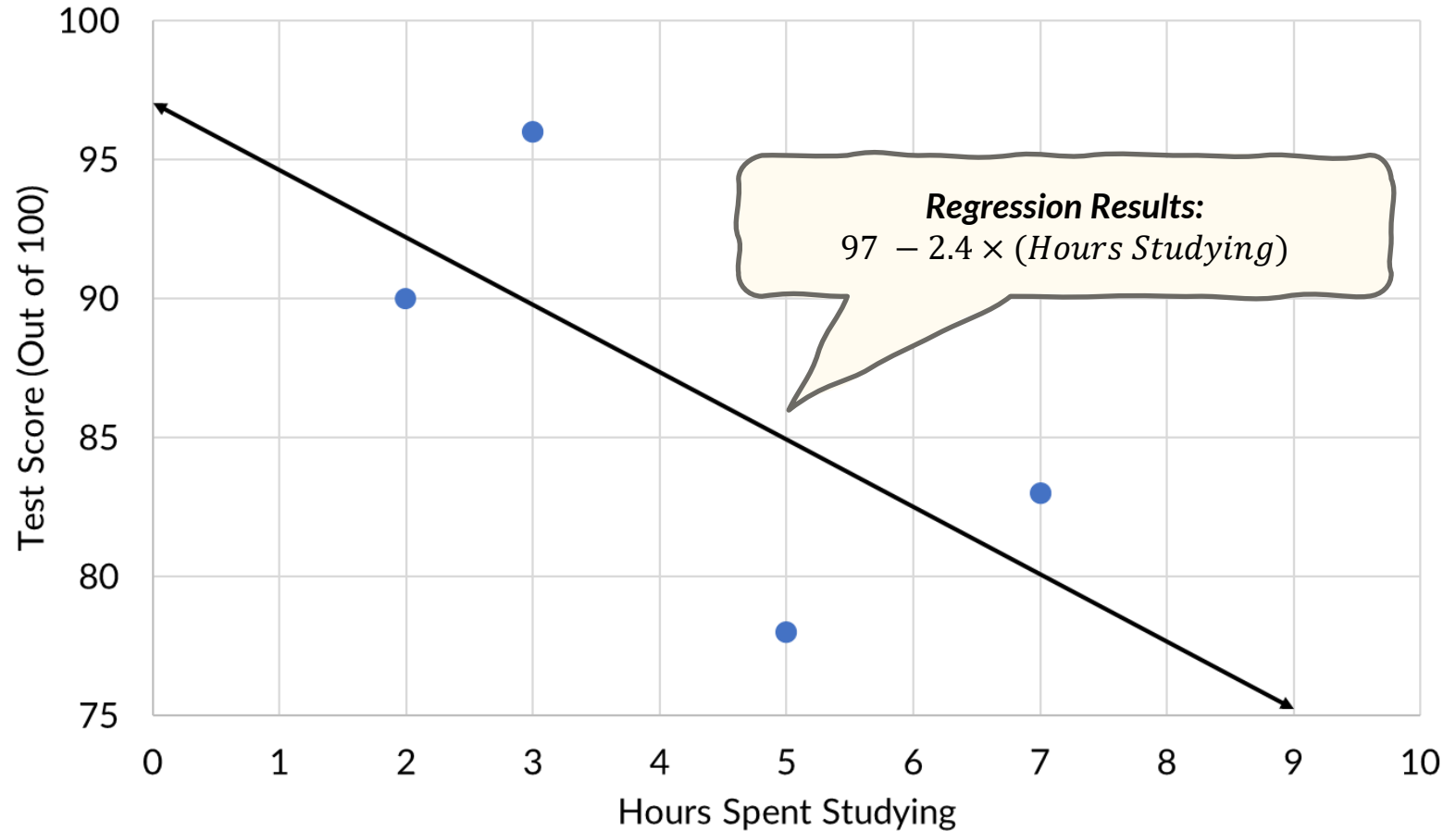
Two ways of interpreting our regression results:

- $\beta_1$  tells us the change in avg. scores associated with studying 1 more hour
- We can **predict** test scores given  $X$  hours studying using  $\hat{\beta}_0 + \hat{\beta}_1 X$

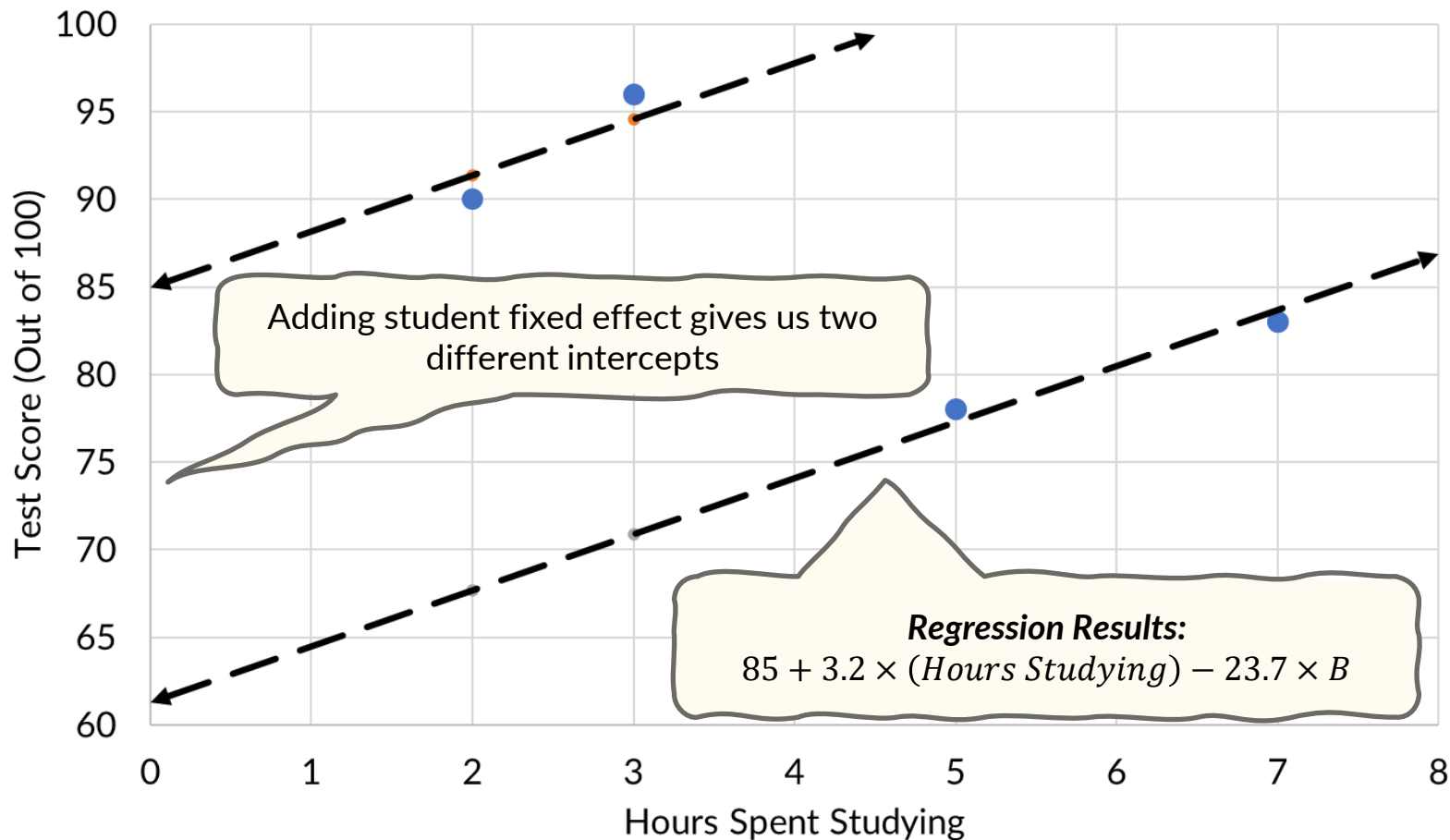
# Test Scores as a Function of Hours Studying



Test Scores as a Function of Hours Studying



## Test Scores as a Function of Hours Studying



# Interpreting Fixed Effects (FEs)

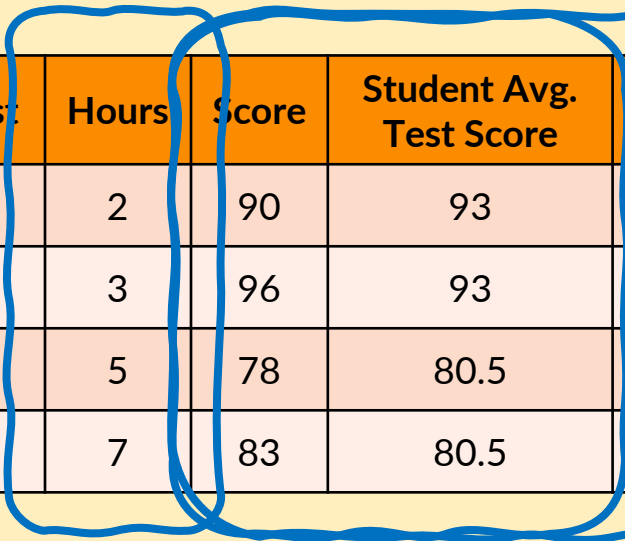
Student FE controlled for avg. differences between students

- In general, we say FEs ***absorb*** variation
- Here, our FE absorbed variation across students in avg. hours + scores

A key question if we're absorbing variation – what variation is left over?

- Refer to this as ***residual variation***
- We can see what this looks like in our example

# Residual Variation



Student	Test	Hours	Score	Student Avg. Test Score	Residual Score Variation	Student Avg. Hours	Residual Hours Variation
A	1	2	90	93	-3	2.5	-0.5
A	2	3	96	93	3	2.5	0.5
B	1	5	78	80.5	-2.5	6	-1
B	2	7	83	80.5	2.5	6	1



# Bias in Regression

In our example, students systematically differed in both:

- (1) avg. time spent studying and (2) avg. test performance
- This means “student” (broadly defined) was an omitted variable (OV)

Structure of OV in this example = fixed or constant difference in average Y and X

By creating a fixed effect for student, we solved the OVB problem

- Let's us *identify* the causal effect of time studying on test performance

# FEs and Causal Inference

FEs are the “building blocks” of research designs like DiD, event studies, etc.

In our example, student FE let us identify the causal effect of studying...

- So long as the **only** source of OVB was avg. differences between students
- This kind of reasoning underlies interpreting all group (2) methods

What might cause OVB? Do we have FEs that absorb that kind of variable?

- We can have fixed effects for time, place, people, etc.
- Adding them can solve OVB... but we still need residual variation!