Studying Social Inequality with Data Science

INFO 3370 / 5371
Spring 2023

Causal inference: Connections to statistical modeling

Learning goals for today

By the end of class, you will be able to

- connect causal inference (a missing data problem)

  to statistical modeling (predicting missing data)

A running example

We should raise taxes on high earners to fund programs that seek to correct injustice

- 1 = Agree
- 0 = Disagree

A running example

We should raise taxes on high earners to fund programs that seek to correct injustice

- 1 = Agree
- 0 = Disagree

What is the average causal effect of taking this class on preferences for taxation to reduce injustice?

- why might it be big?
- why might it be small?
- why is it hard to know the answer?

Using potential outcomes

Each Row is a Student in This Class	Outcome under 3370	Outcome under no 3370
	$Y_1^{\text{Takes 3370}}$	$Y_1^{\text{No 3370}}$
	$Y_2^{\text{Takes 3370}}$	$Y_2^{\text{No 3370}}$
	$Y_3^{\text{Takes 3370}}$	$Y_3^{\text{No 3370}}$
	$Y_4^{\text{Takes 3370}}$	$Y_4^{\text{No 3370}}$
	$Y_5^{\text{Takes 3370}}$	$Y_5^{\text{No 3370}}$
	$Y_6^{\text{Takes 3370}}$	$Y_6^{\text{No 3370}}$

Y = We should raise taxes on high earners to fund programs that seek to correct injustice

Using potential outcomes

Y = We should raise taxes on high earners to fund programs that seek to correct injustice

Each Row is a Student in This Class

Outcome under 3370	Outcome under no 3370
$Y_1^{\text{Takes 3370}}$	?
$Y_2^{\text{Takes 3370}}$	?
$Y_3^{\text{Takes 3370}}$	?
$Y_4^{\text{Takes 3370}}$	?
$Y_5^{\text{Takes 3370}}$	?
$Y_6^{\text{Takes 3370}}$	?

Using potential outcomes

Each Row is a Student in This Class	Outcome under 3370	Outcome under no 3370
1	$Y_1^{\text{Takes 3370}}$	?
2	$Y_2^{\text{Takes 3370}}$	?
3	$Y_3^{\text{Takes 3370}}$	?
4	$Y_4^{\text{Takes 3370}}$	?
5	$Y_5^{\text{Takes 3370}}$	?
6	$Y_6^{\text{Takes 3370}}$	?

Y = We should raise taxes on high earners to fund programs that seek to correct injustice

How could we learn about the (?)

Strategy 1: People who nearly took the class

Strategy 1: People who nearly took the class

- Some of the class was on the waitlist
  - some got in
  - others didn't

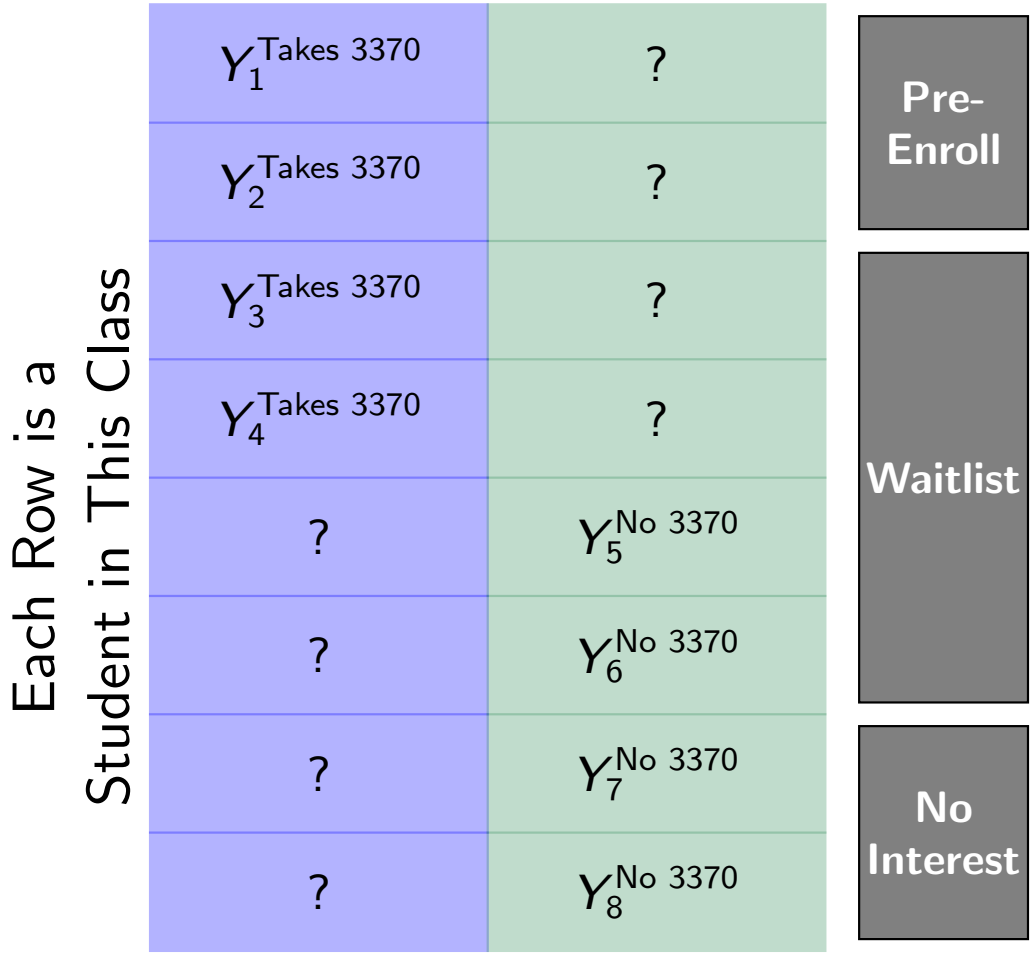
Strategy 1: People who nearly took the class

Y = We should raise taxes on high earners to fund programs that seek to correct injustice

Each Row is a Student in This Class

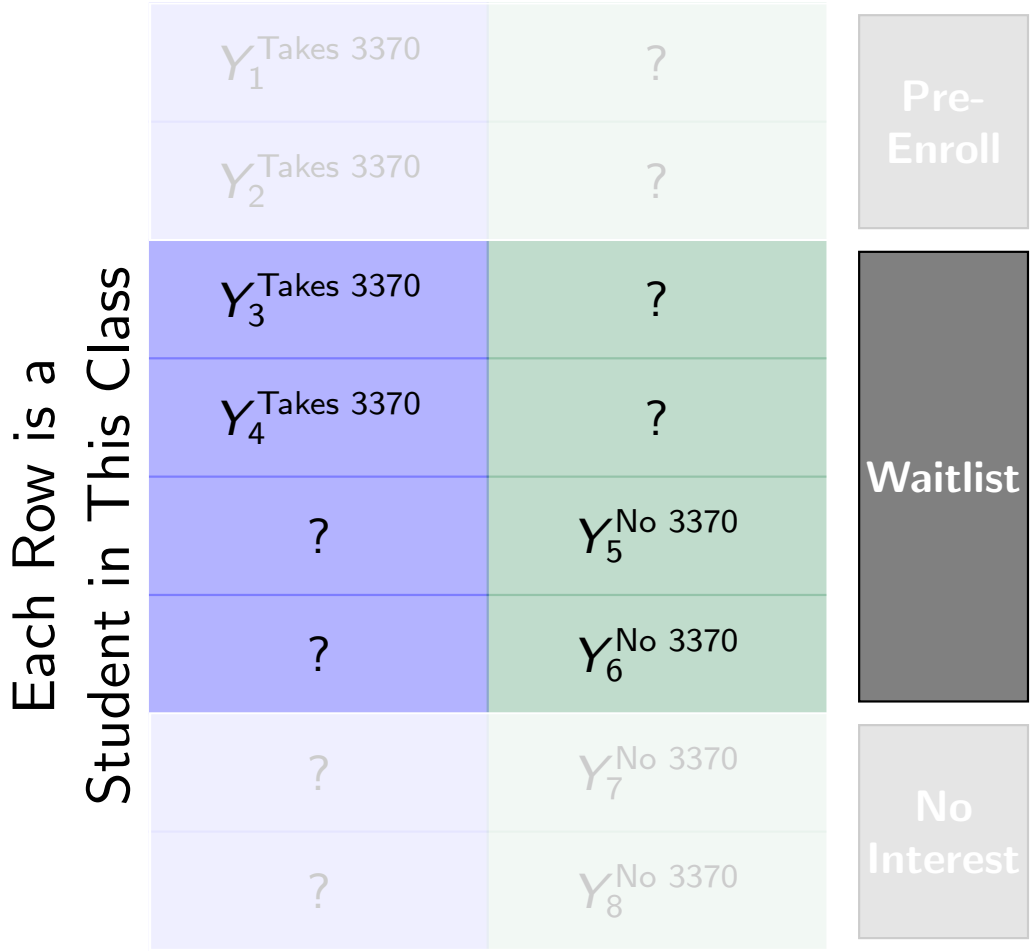
$Y_1^{\text{Takes 3370}}$	?
$Y_2^{\text{Takes 3370}}$	?
$Y_3^{\text{Takes 3370}}$	?
$Y_4^{\text{Takes 3370}}$	?
?	$Y_5^{\text{No 3370}}$
?	$Y_6^{\text{No 3370}}$
?	$Y_7^{\text{No 3370}}$
?	$Y_8^{\text{No 3370}}$

Strategy 1: People who nearly took the class



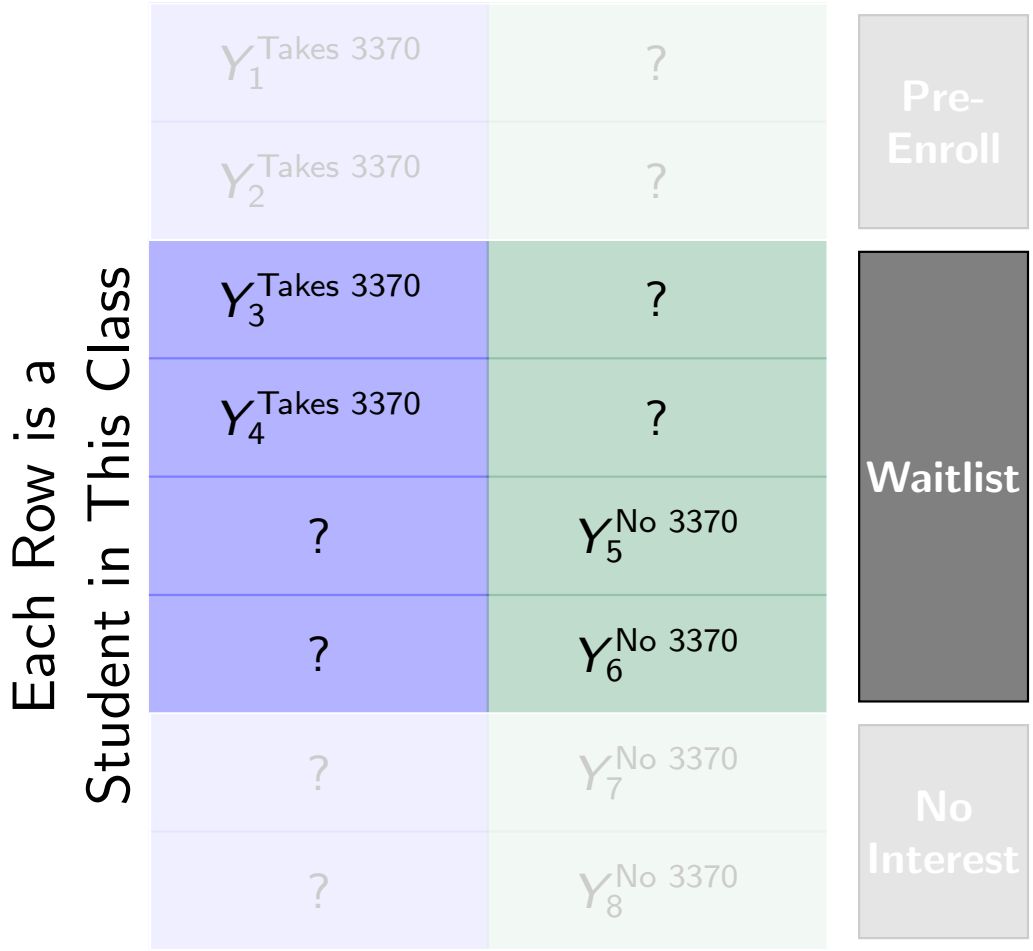
$Y =$ We should raise taxes on high earners to fund programs that seek to correct injustice

Strategy 1: People who nearly took the class



Y = We should raise taxes on high earners to fund programs that seek to correct injustice

Strategy 1: People who nearly took the class

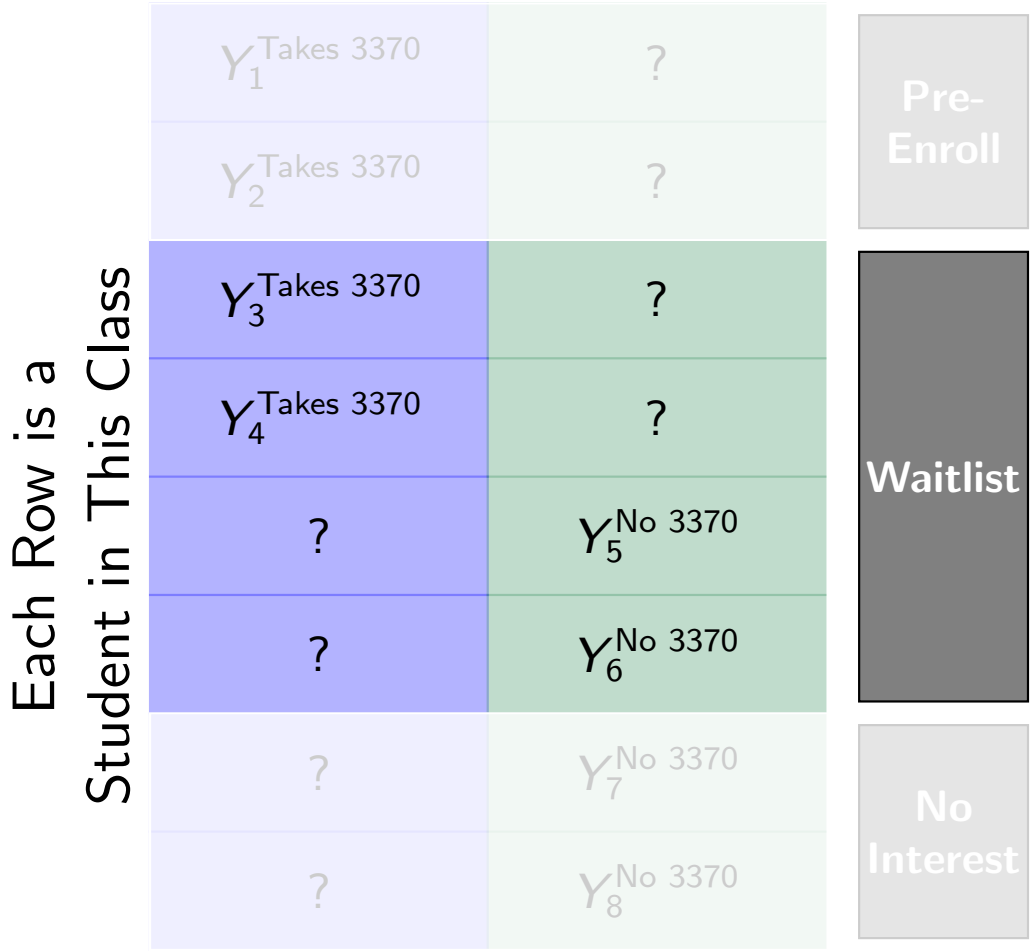


Y = We should raise taxes on high earners to fund programs that seek to correct injustice

Benefits of strategy

Drawbacks

Strategy 1: People who nearly took the class



Y = We should raise taxes on high earners to fund programs that seek to correct injustice

Benefits of strategy

Credible

Drawbacks

Limited target population

Strategy 1: People who nearly took the class

Generalizing: Causal strategies in this domain

Strategy 1: People who nearly took the class

Generalizing: Causal strategies in this domain

- instrumental variables

Strategy 1: People who nearly took the class

Generalizing: Causal strategies in this domain

- instrumental variables
- regression discontinuity

Strategy 1: People who nearly took the class

Generalizing: Causal strategies in this domain

- instrumental variables
- regression discontinuity
- interrupted time series

Strategy 1: People who nearly took the class

Generalizing: Causal strategies in this domain

- instrumental variables
- regression discontinuity
- interrupted time series

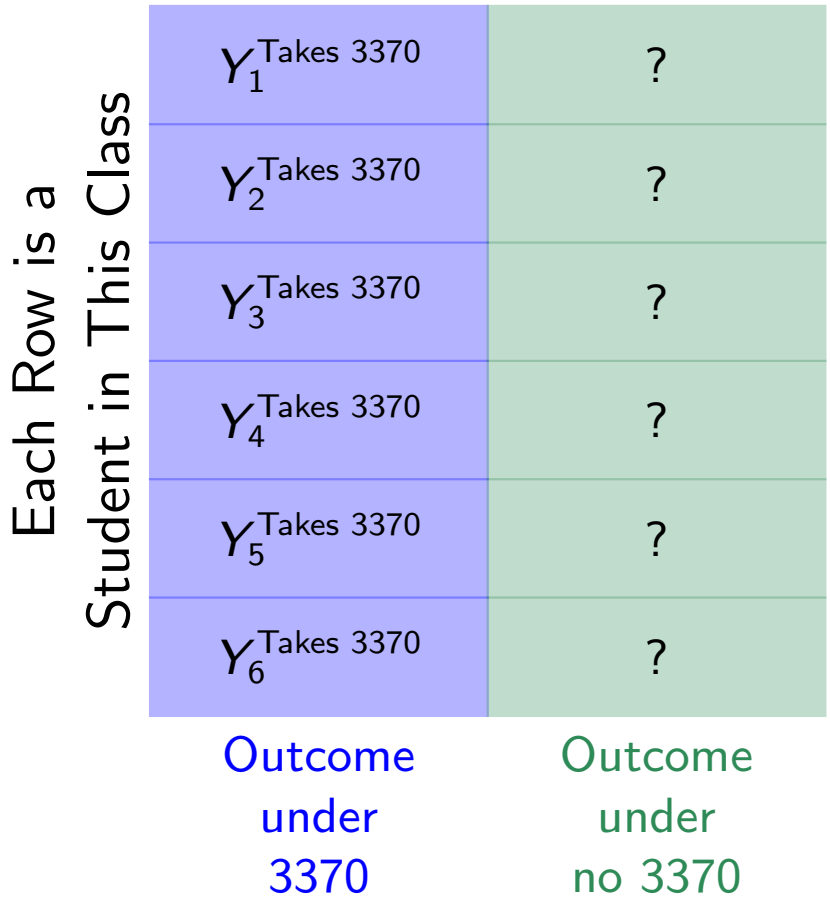
These strategies identify causal effects
by focusing on a feasible subpopulation
where treatment assignment is well-understood

Each Row is a Student in This Class

$Y_1^{\text{Takes 3370}}$	$Y_1^{\text{No 3370}}$
$Y_2^{\text{Takes 3370}}$	$Y_2^{\text{No 3370}}$
$Y_3^{\text{Takes 3370}}$	$Y_3^{\text{No 3370}}$
$Y_4^{\text{Takes 3370}}$	$Y_4^{\text{No 3370}}$
$Y_5^{\text{Takes 3370}}$	$Y_5^{\text{No 3370}}$
$Y_6^{\text{Takes 3370}}$	$Y_6^{\text{No 3370}}$
Outcome under 3370	Outcome under no 3370

Y = We should raise taxes on high earners to fund programs that seek to correct injustice

How could we learn about the (?)



Y = We should raise taxes on high earners to fund programs that seek to correct injustice

How could we learn about the (?)

Strategy 2: Find your look-alikes on relevant dimensions

Strategy 2: Find your look-alikes on relevant dimensions

For each of you, we could compare

1. your opinion after 3370
2. the average opinion of non-3370 students who look like you

Strategy 2: Find your look-alikes on relevant dimensions

For each of you, we could compare

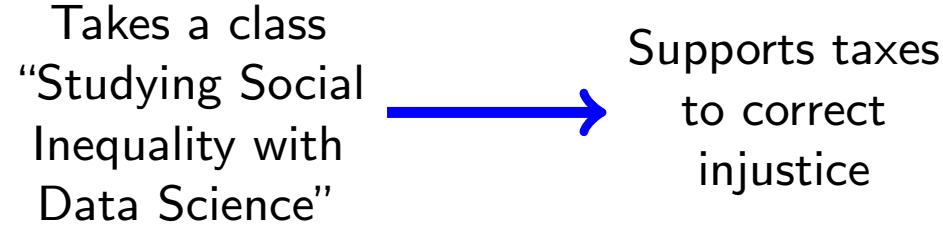
1. your opinion after 3370
2. the average opinion of non-3370 students who look like you

On what dimensions should they look like you?

Strategy 2: Find your look-alikes on relevant dimensions

Causal diagrams can help us reason about the adjustment set

- nodes are random variables
- edges are causal relationships

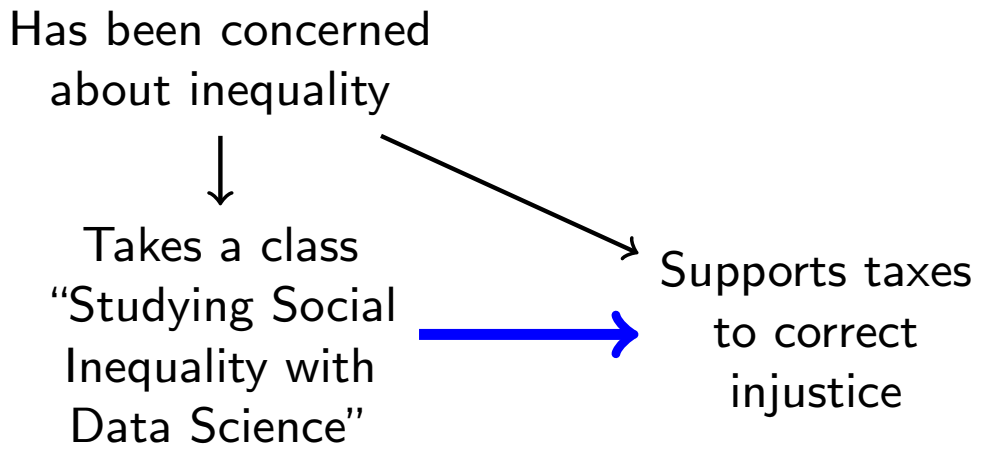


To learn more about causal graphs, see Pearl & Mackenzie 2018

Strategy 2: Find your look-alikes on relevant dimensions

Causal diagrams can help us reason about the adjustment set

- nodes are random variables
- edges are causal relationships

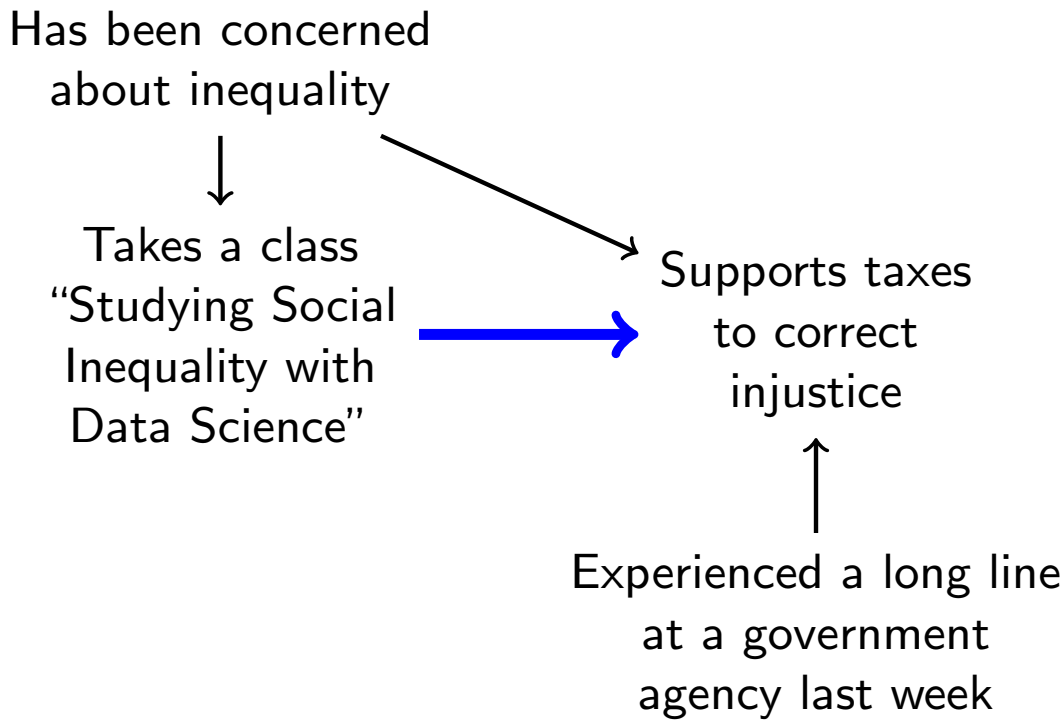


To learn more about causal graphs, see Pearl & Mackenzie 2018

Strategy 2: Find your look-alikes on relevant dimensions

Causal diagrams can help us reason about the adjustment set

- nodes are random variables
- edges are causal relationships

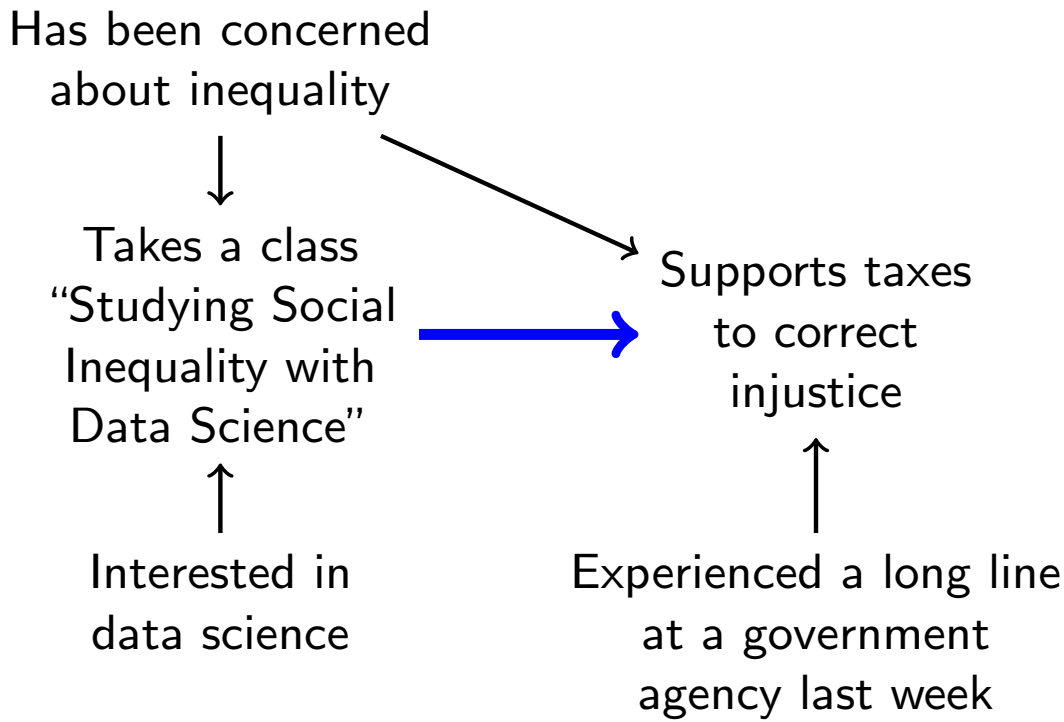


To learn more about causal graphs, see Pearl & Mackenzie 2018

Strategy 2: Find your look-alikes on relevant dimensions

Causal diagrams can help us reason about the adjustment set

- nodes are random variables
- edges are causal relationships



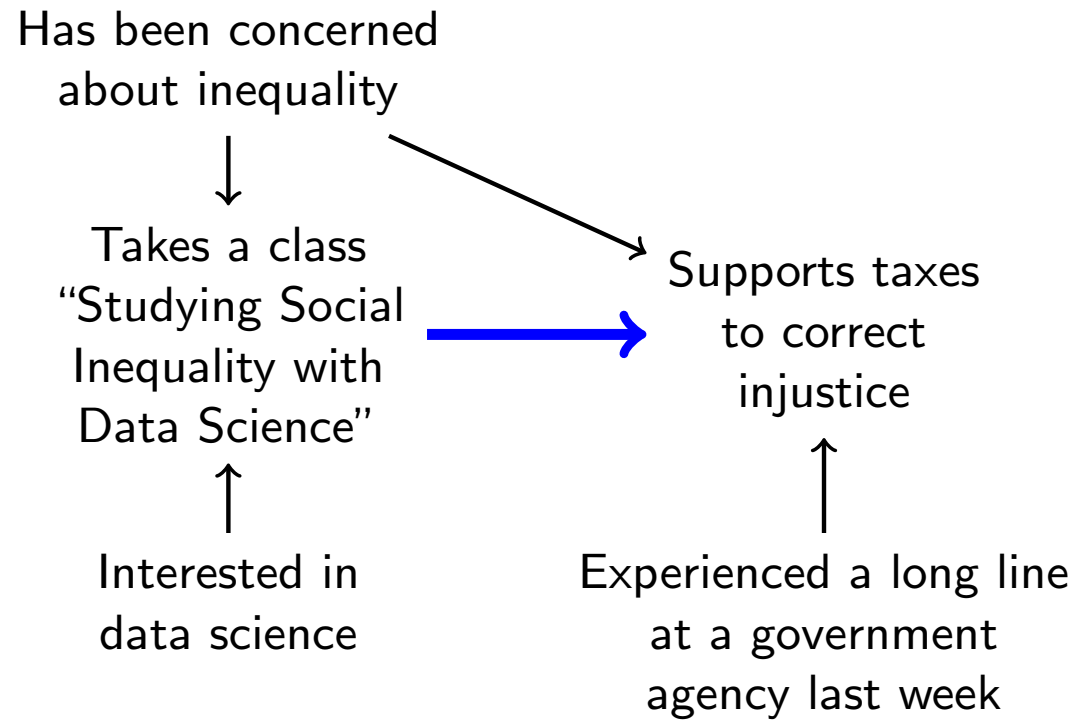
To learn more about causal graphs, see Pearl & Mackenzie 2018

Strategy 2: Find your look-alikes on relevant dimensions

Causal diagrams can help us reason about the adjustment set

- nodes are random variables
- edges are causal relationships

In this figure, what are the reasons taking 3370 is related to support for taxation?



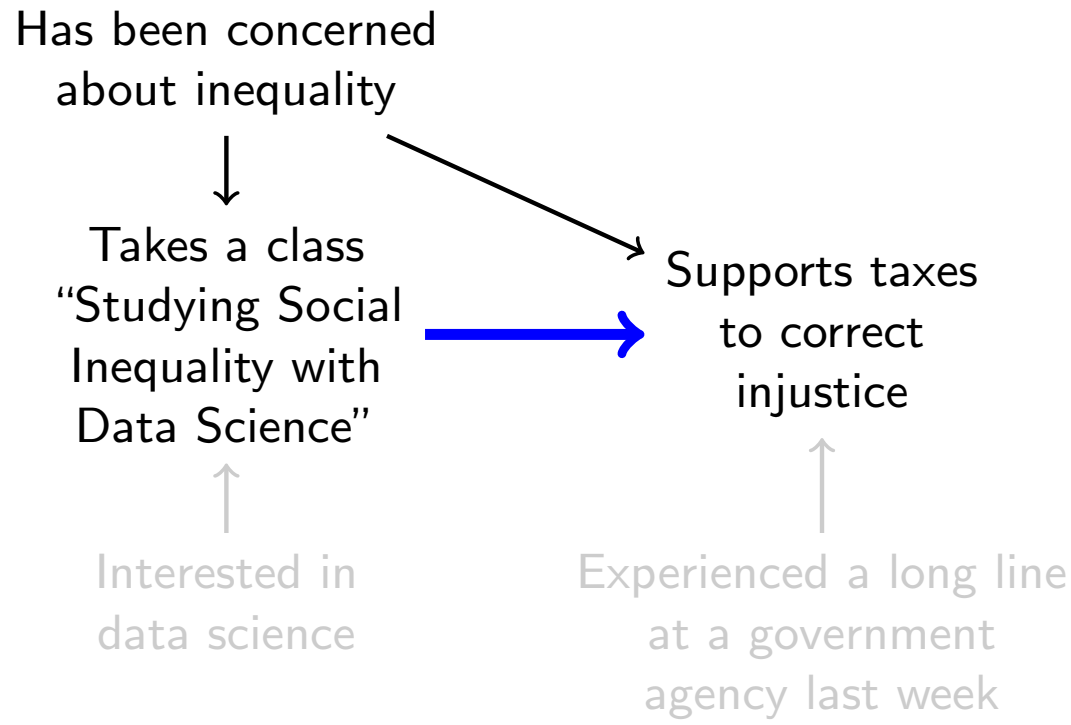
To learn more about causal graphs, see Pearl & Mackenzie 2018

Strategy 2: Find your look-alikes on relevant dimensions

Causal diagrams can help us reason about the adjustment set

- nodes are random variables
- edges are causal relationships

In this figure, what are the reasons taking 3370 is related to support for taxation?



To learn more about causal graphs, see Pearl & Mackenzie 2018

Strategy 2: Find your look-alikes on relevant dimensions

Benefits

- Full target population

Drawbacks

- May be less credible than approaches like the waitlist

Strategy 2: Generalizing to a model

Strategy 2: Generalizing to a model

Regression = Tool to predict data you don't see

Strategy 2: Generalizing to a model

Regression = Tool to predict data you don't see

- we don't see your outcome without 3370

Strategy 2: Generalizing to a model

Regression = Tool to predict data you don't see

- we don't see your outcome without 3370

Causal assumption: On average,

$$Y_{\text{You}}^{\text{No 3370}} \approx \text{E}(Y_{\text{Others}}^{\text{No 3370}} \mid \text{Look like you})$$

Strategy 2: Generalizing to a model

Regression = Tool to predict data you don't see

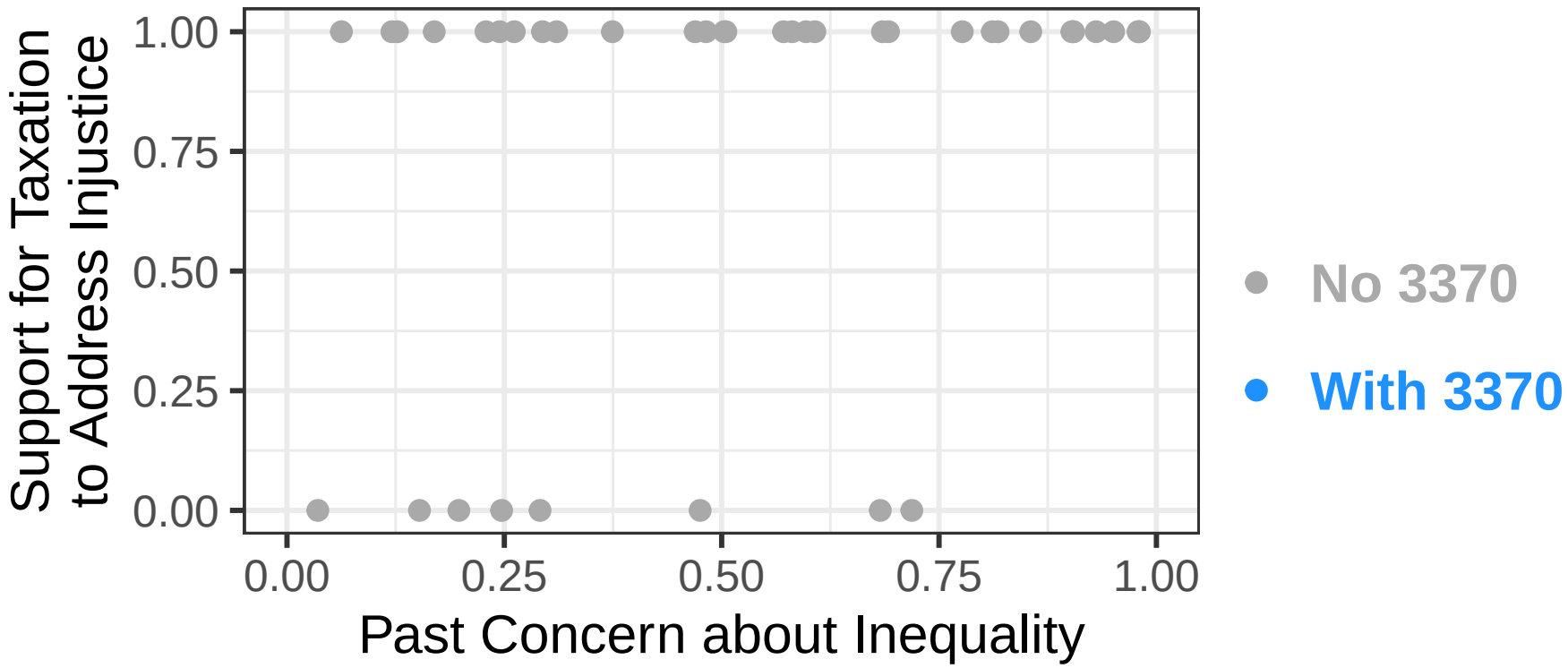
- we don't see your outcome without 3370

Causal assumption: On average,

$$Y_{\text{You}}^{\text{No 3370}} \approx \text{E}(Y_{\text{Others}}^{\text{No 3370}} \mid \text{Look like you})$$

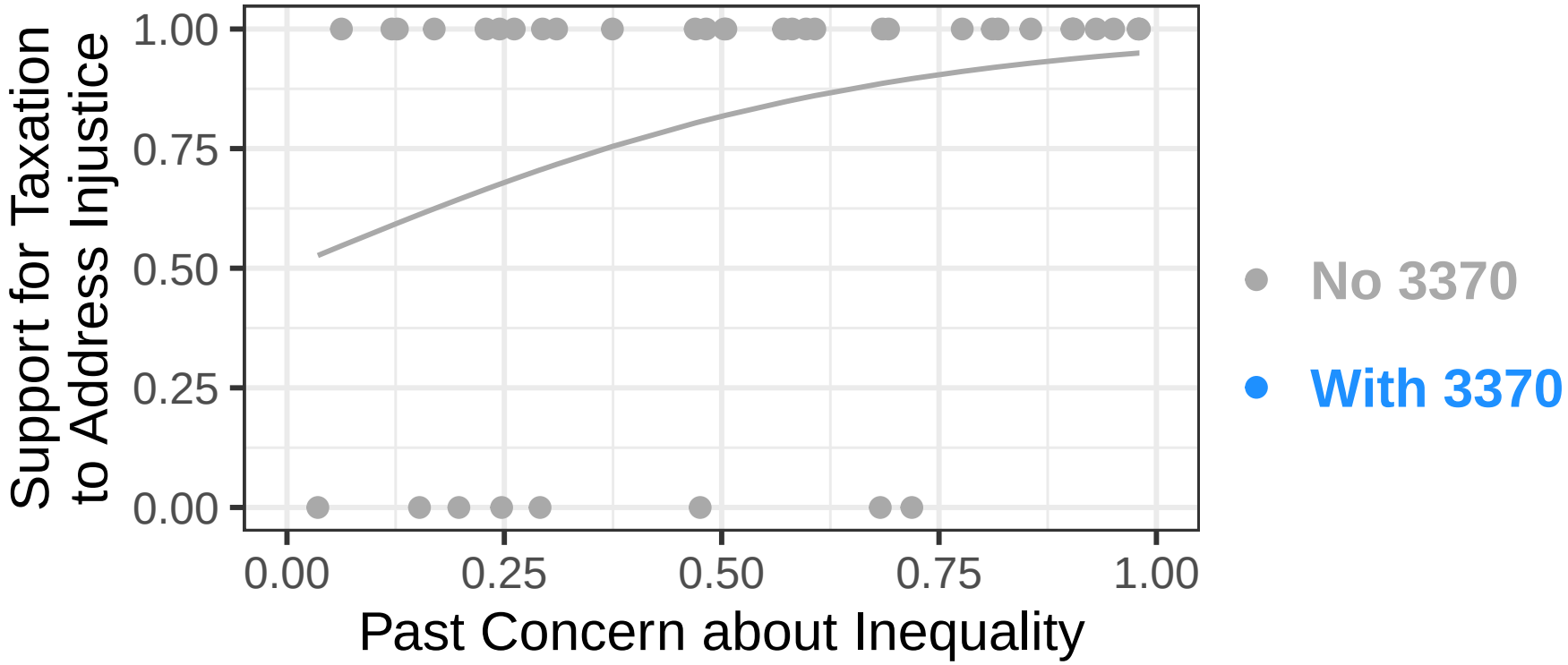
The right side can be modeled statistically

Strategy 2: Generalizing to a model



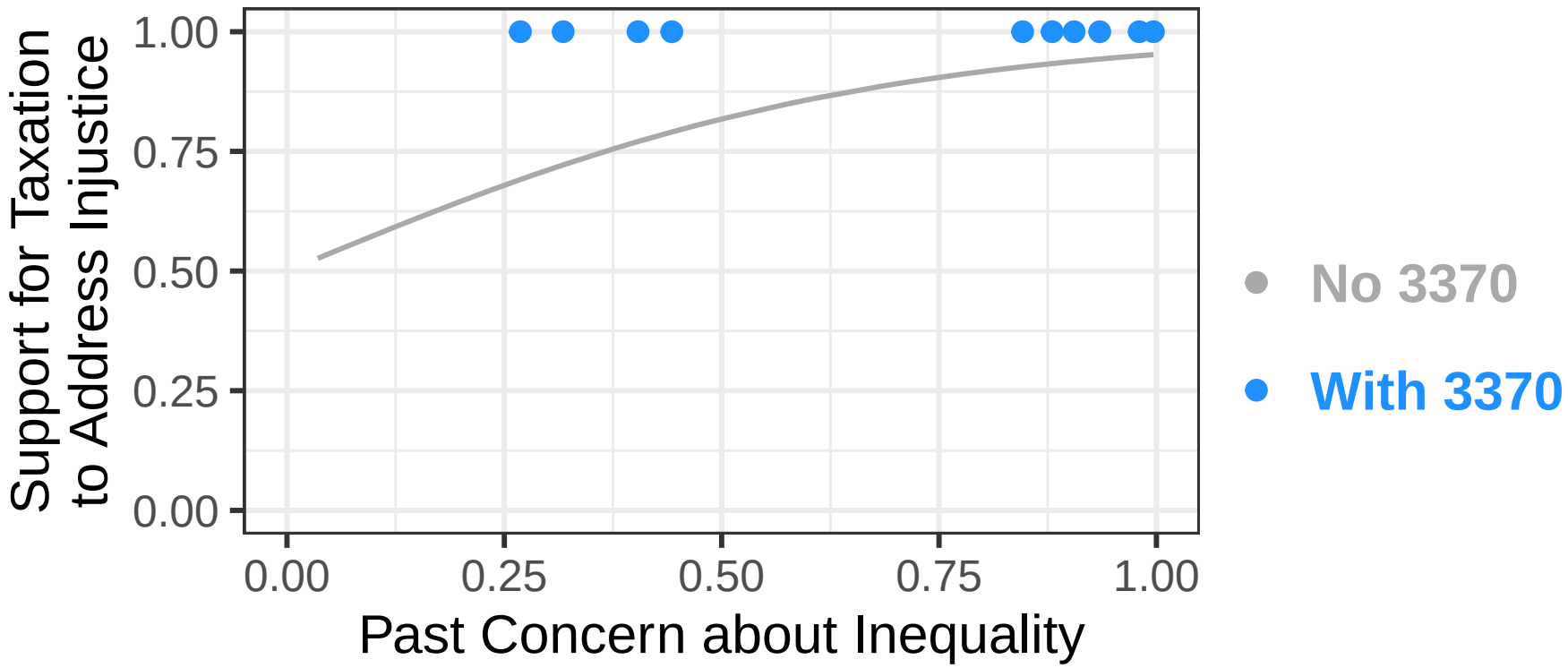
1) Find control units who didn't take this class

Strategy 2: Generalizing to a model



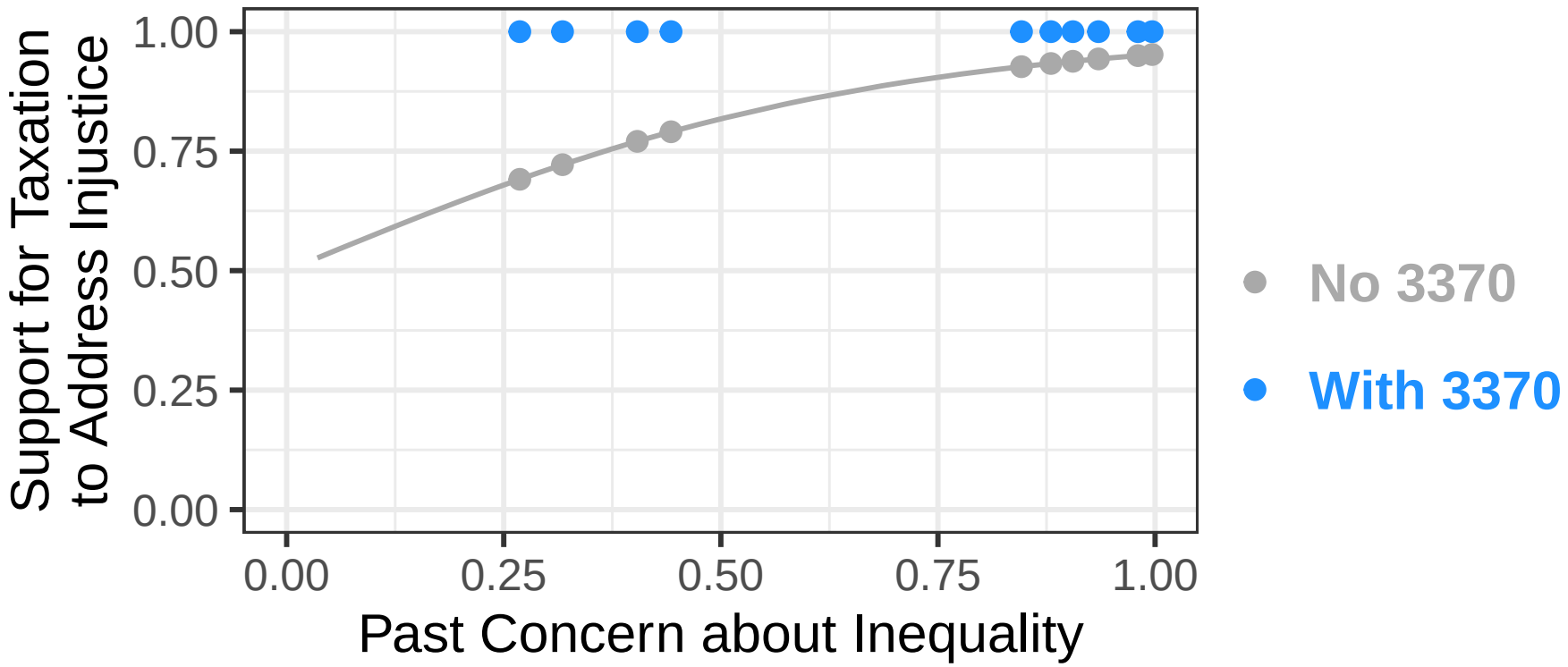
2) Model their outcomes given pre-treatment variables

Strategy 2: Generalizing to a model



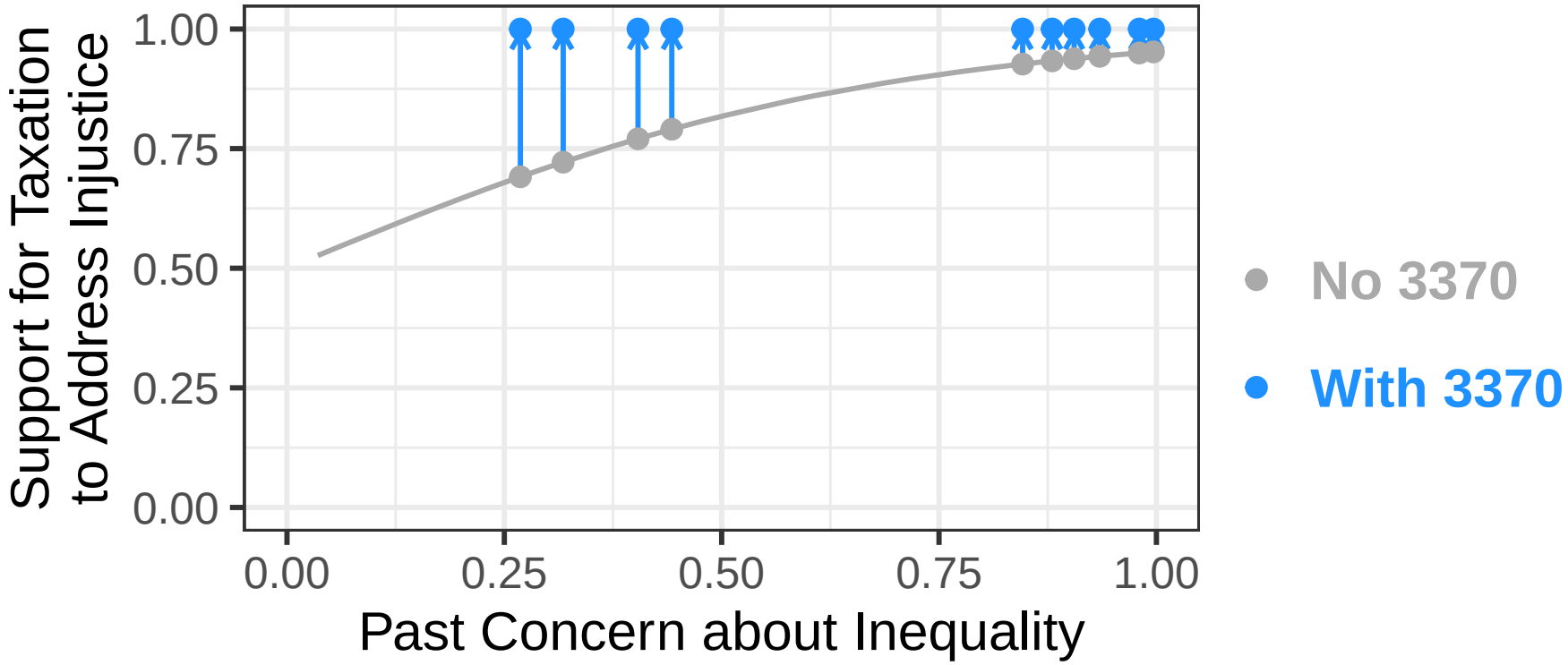
3) Find the treated units of interest

Strategy 2: Generalizing to a model



4) Predict their counterfactual outcomes

Strategy 2: Generalizing to a model



5) Infer causal effect for each person. Average over people

Strategy 2: Generalizing to a model

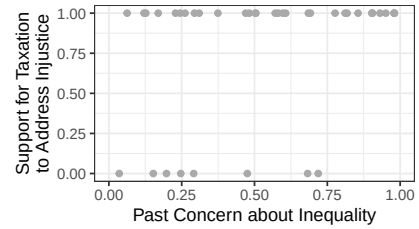
1) Find control units who didn't take this class

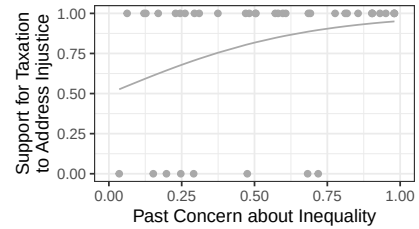
2) Model their outcomes given pre-treatment variables

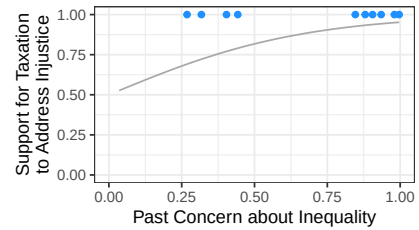
3) Find the treated units of interest

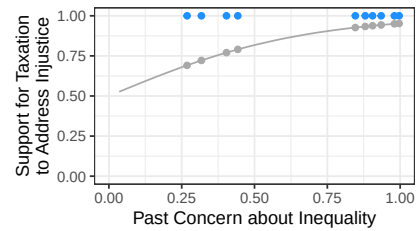
4) Predict their counterfactual outcomes

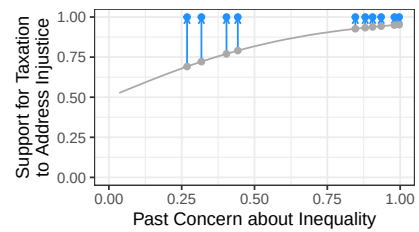
5) Infer causal effect for each person. Average over people









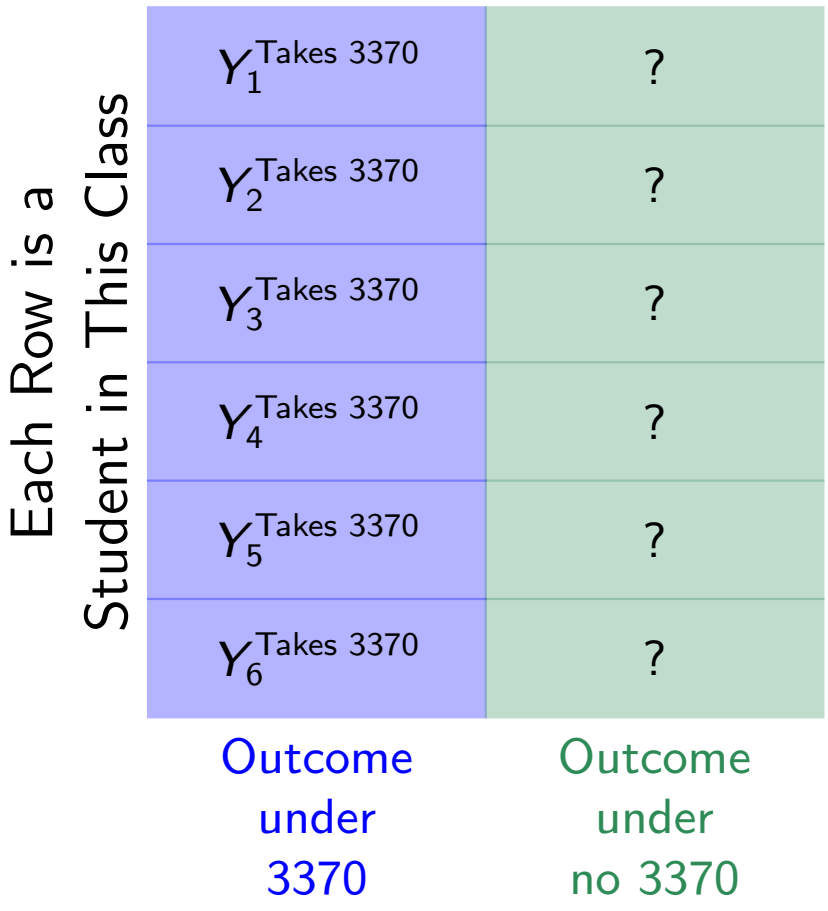


Summary: Causal inference is a missing data problem

Each Row is a Student in This Class

Outcome under 3370	Outcome under no 3370
$Y_1^{\text{Takes 3370}}$	$Y_1^{\text{No 3370}}$
$Y_2^{\text{Takes 3370}}$	$Y_2^{\text{No 3370}}$
$Y_3^{\text{Takes 3370}}$	$Y_3^{\text{No 3370}}$
$Y_4^{\text{Takes 3370}}$	$Y_4^{\text{No 3370}}$
$Y_5^{\text{Takes 3370}}$	$Y_5^{\text{No 3370}}$
$Y_6^{\text{Takes 3370}}$	$Y_6^{\text{No 3370}}$

Summary: Causal inference is a missing data problem



Summary: Causal inference is a missing data problem

Each Row is a Student in This Class

Outcome under 3370	Outcome under no 3370
$Y_1^{\text{Takes 3370}}$	$Y_1^{\text{No 3370}}$
$Y_2^{\text{Takes 3370}}$	$Y_2^{\text{No 3370}}$
$Y_3^{\text{Takes 3370}}$	$Y_3^{\text{No 3370}}$
$Y_4^{\text{Takes 3370}}$	$Y_4^{\text{No 3370}}$
$Y_5^{\text{Takes 3370}}$	$Y_5^{\text{No 3370}}$
$Y_6^{\text{Takes 3370}}$	$Y_6^{\text{No 3370}}$

General approach

Summary: Causal inference is a missing data problem

Each Row is a Student in This Class

Outcome under 3370	Outcome under no 3370
$Y_1^{\text{Takes 3370}}$	$Y_1^{\text{No 3370}}$
$Y_2^{\text{Takes 3370}}$	$Y_2^{\text{No 3370}}$
$Y_3^{\text{Takes 3370}}$	$Y_3^{\text{No 3370}}$
$Y_4^{\text{Takes 3370}}$	$Y_4^{\text{No 3370}}$
$Y_5^{\text{Takes 3370}}$	$Y_5^{\text{No 3370}}$
$Y_6^{\text{Takes 3370}}$	$Y_6^{\text{No 3370}}$

General approach

1) Define potential outcomes

Summary: Causal inference is a missing data problem

Each Row is a Student in This Class

Outcome under 3370	Outcome under no 3370
$Y_1^{\text{Takes 3370}}$	$Y_1^{\text{No 3370}}$
$Y_2^{\text{Takes 3370}}$	$Y_2^{\text{No 3370}}$
$Y_3^{\text{Takes 3370}}$	$Y_3^{\text{No 3370}}$
$Y_4^{\text{Takes 3370}}$	$Y_4^{\text{No 3370}}$
$Y_5^{\text{Takes 3370}}$	$Y_5^{\text{No 3370}}$
$Y_6^{\text{Takes 3370}}$	$Y_6^{\text{No 3370}}$

General approach

1) Define potential outcomes
2) Define target population

Summary: Causal inference is a missing data problem

Each Row is a Student in This Class

$Y_i^{\text{Takes 3370}}$	$Y_i^{\text{No 3370}}$
$Y_1^{\text{Takes 3370}}$	$Y_1^{\text{No 3370}}$
$Y_2^{\text{Takes 3370}}$	$Y_2^{\text{No 3370}}$
$Y_3^{\text{Takes 3370}}$	$Y_3^{\text{No 3370}}$
$Y_4^{\text{Takes 3370}}$	$Y_4^{\text{No 3370}}$
$Y_5^{\text{Takes 3370}}$	$Y_5^{\text{No 3370}}$
$Y_6^{\text{Takes 3370}}$	$Y_6^{\text{No 3370}}$
Outcome under 3370	Outcome under no 3370

General approach

1) Define potential outcomes
2) Define target population
3) Make causal assumptions

Summary: Causal inference is a missing data problem

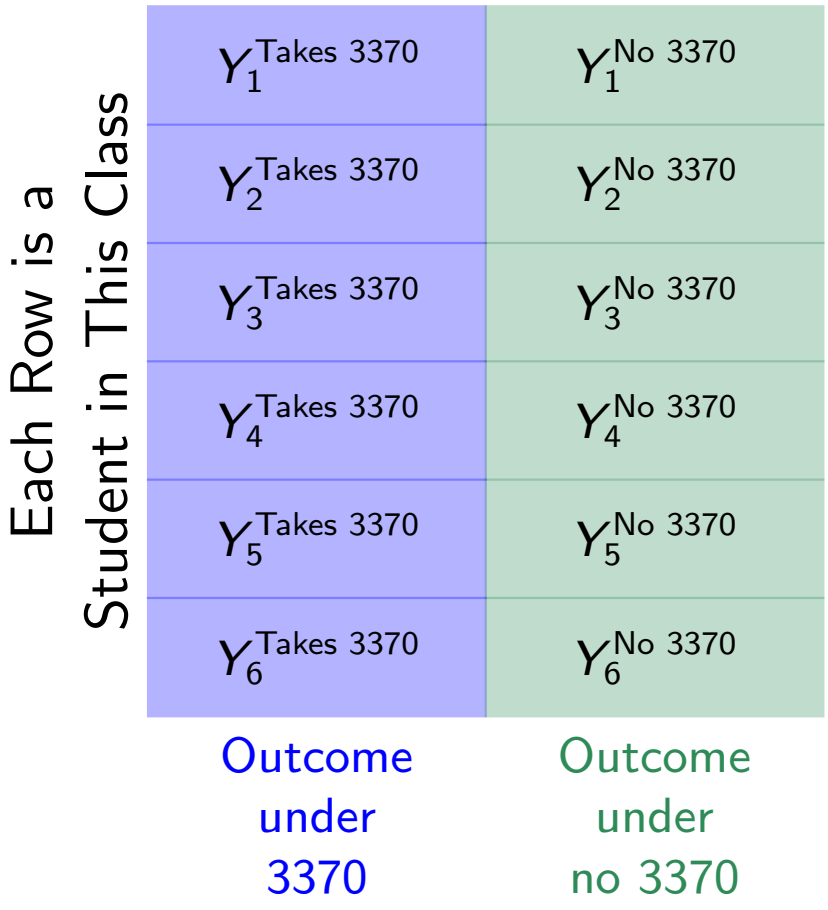
Each Row is a Student in This Class

Outcome under 3370	Outcome under no 3370
$Y_1^{\text{Takes 3370}}$	$Y_1^{\text{No 3370}}$
$Y_2^{\text{Takes 3370}}$	$Y_2^{\text{No 3370}}$
$Y_3^{\text{Takes 3370}}$	$Y_3^{\text{No 3370}}$
$Y_4^{\text{Takes 3370}}$	$Y_4^{\text{No 3370}}$
$Y_5^{\text{Takes 3370}}$	$Y_5^{\text{No 3370}}$
$Y_6^{\text{Takes 3370}}$	$Y_6^{\text{No 3370}}$

General approach

1) Define potential outcomes
2) Define target population
3) Make causal assumptions
4) Model unobserved outcomes

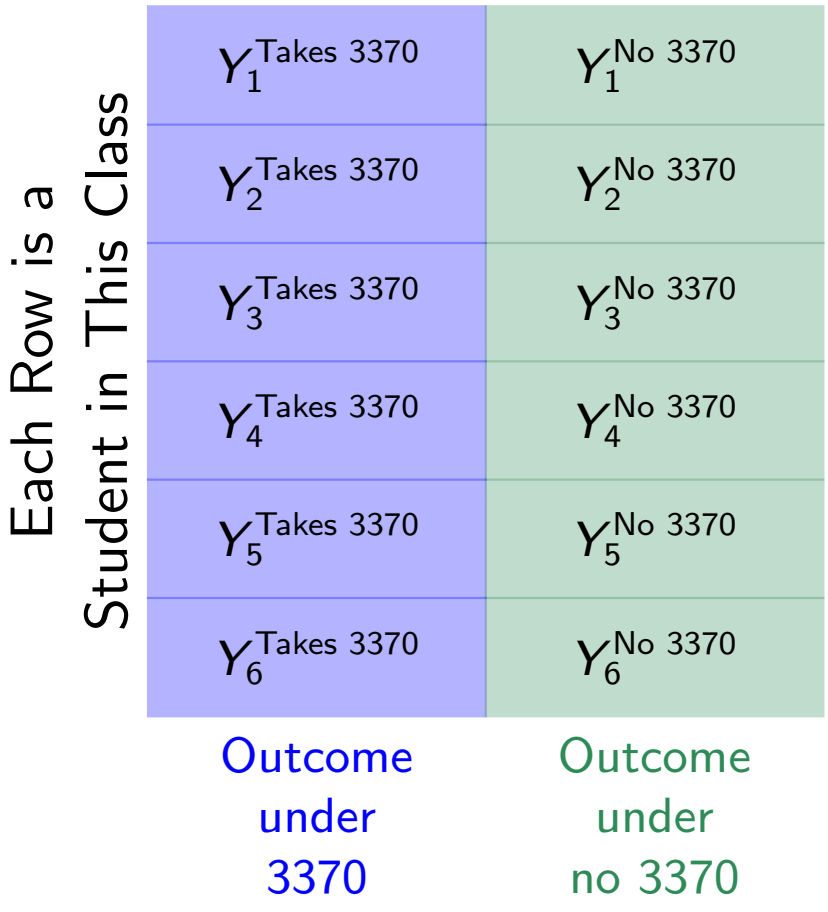
Summary: Causal inference is a missing data problem



General approach

1) Define potential outcomes
2) Define target population
3) Make causal assumptions
4) Model unobserved outcomes
5) Predict them

Summary: Causal inference is a missing data problem



General approach

1) Define potential outcomes
2) Define target population
3) Make causal assumptions
4) Model unobserved outcomes
5) Predict them
6) Report an average

In what settings

- is it important to ask a causal question about inequality?
- is it sufficient to ask a descriptive question?

Learning goals for today

By the end of class, you will be able to

- connect causal inference (a missing data problem)

  to statistical modeling (predicting missing data)