Causal Inference
1 - Introduction to Research Design

Benjamin Elsner
benjamin.elsner@ucd.ie

Learning Outcomes

What I would like you to take away from this module:

- Know and understand the **state-of-the-art methods of applied econometrics**
- ...in particular methods of **causal inference** (used in 80+% of top publications)
- Critically **assess the empirical methods** used to answer **causal questions**
- **Develop research designs** for your work
- **Communicate research results** (written & in presentations)

Logistics

Credits: 10 ECTS
Office Hours: Wednesday, 9-10am, 12-1pm; book on Calendly
Contact email: benjamin.elsner@ucd.ie
Time: Wednesdays, 3-5pm on Zoom
Venue: D201

Blended Learning

I will **pre-record videos** that introduce the topic
$\Rightarrow$ you need to watch them BEFORE the lecture

Online lectures will be used for

- Presenting additional material
- Answering your questions
- Going through problem sets, etc...

Zoom Etiquette

You can make everyone's life so much better by following **three simple rules:**

1. Keep your **camera on**
2. Use your **name**
3. Mute your microphone unless you want to say something

It's a small course, so it's ok to

- unmute yourself and interrupt me
- use the chat function (will respond with a delay)
- use the "raise hand" tool

Communication outside the "Classroom"

Again, simple rules:

- The default is the Brightspace discussion forum
- I will also compile a list of FAQ on Brightspace (under "Module Tools")
- Email is only for communication with a personal/confidential content
- My email hours are: Monday 1-2pm, Friday 12-1pm

Materials

Reading lists of papers will be provided at the end of each set of lecture notes.

Textbooks:

MHE Angrist, Joshua and Jörn-Steffen Pischke. *Mostly Harmless Econometrics. An Empiricist's companion.* Princeton University Press, 2009.

CIM Cunningham, Scott. *Causal Inference: The Mixtape*, Yale University Press, 2021. Available for free on Cunningham's website

Another useful more **general econometrics textbook** is:
Wooldridge, Jeffrey. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, M.A.: MIT Press 2010 (2nd edition).

For an **introduction to causal inference**, I recommend:
Angrist, Joshua and Jörn-Steffen Pischke. *Mastering 'Metrics: The Path from Cause to Effect* Princeton University Press, 2014.

Statistical Software

The most commonly used programmes among applied economists are **Stata and R**

I highly recommend using R. R is

- free
- highly versatile
- thanks to RStudio much more user-friendly than it used to be

But you can use any software package you like, as long as it fits the task

Assessment

The assessment will be **based on five components**:

- Final exam: 50%
- Replication and practice exercise: 20%
- 5 problem sets: 25%
- Presentation: 5%

You need to pass each component to pass the course. We use the alternative linear grade scale.

Assessment

Final exam

- 2-hour exam; type, duration and date tba

Replication and practice exercise

- this will teach you how to run an analysis and write it up
- Work in randomly assigned groups
- You will receive a paper to replicate
- You have to add at least one extension to the analysis (instructions given)
- I will provide materials in early February
- You have about six weeks to complete the exercise
- Presentations in the last two weeks

Presentation

I will randomly assign you to 5-6 presentation groups

Each group has to **present a recent paper that uses a given method**

Give a 15min presentation in the live lecture. Among others, comment on:

- What is the research question the paper wants to answer?
- How does the identification strategy work? What is the identifying assumption?
- How do the authors justify the identifying assumption?

Assessment

Problem sets

- There will be **five problem sets**; best four will be graded
- You typically have **one week** to solve them
- You can work in **groups 3-5 people**
- Submit one joint solution per group

Rules for submission

- Submit through Brightspace
- All in one pdf file
- Code should be in the appendix
- Scans of handwritten derivations are ok
- Screenshots of statistical software are not ok!
- ...results need to be presented in tables or graphically
- Show proof that you use version control (more on that later)

Managing Expecations

This is a PhD course

- it requires **a lot of work**
- students have to **learn to solve problems**
- ...this is what research is all about
- problem sets are not always in sync with the lectures

The course focuses on the **most relevant methods in applied econometrics**

- this inevitably leaves out other important methods
- examples: time series, advanced panel methods, survival analysis...

Managing Expecations

This is a PhD course

What I will NOT teach you

- how to **use statistical software**
- how to **handle and clean datasets**

But don't panic, you will **learn this "by doing"**

Your peers are your lab

Course outline

1. Introduction to Research Design
2. Research Practice
3. Instrumental Variables and Marginal Treatment Effects
4. Regression Discontinuity and Kink Designs
5. Difference-in-Differences (advanced)
6. Synthetic Controls
7. Bounding
8. Advanced topics (time permitting)
   - Fixed effect estimation — new developments
   - Shift-share instruments
   - Mediation analysis

1) Causality reloaded

Social Norms in Econometrics

An important element of this course is to teach the **social norms about "how to do research"**

These norms are important to understand **what distinguishes a good (read: successful) paper** from a **not-so-good one**

As with all social norms, **they constantly change**. Some older economists call the methods I teach here a **fad**. But these methods have dominated top journals for about 20 years now...

Cookbook approach vs. Traditional econometrics teaching

In a **traditional M.Sc/PhD econometrics course** (and in the major textbooks), **all methods are created equal**

But **not all methods are equally important** in current research in applied microeconom(etr)ics

Challenge: **try to publish a paper** that uses one of the following **methods in a top journal**

- Heckman two-step selection model
- Propensity score matching
- Oaxaca-Blinder wage decompositions
- Random effects models
- ...

Cookbook approach vs. Traditional econometrics course

Have these **methods become obsolete**? **Absolutely not!**

- Many state-of-the-art methods build upon them
- Social norms change: new varieties of these models may re-appear
- Example: logit models; heavily used since economists discovered machine learning

Bottom line: **"traditional" and "cookbook"** knowledge are **complements**, not substitutes

Causality

In Econometrics 1, **causality** was introduced via **potential outcomes**

In this lecture, we discuss a different approach: **DAGs**

We learn:

- how to think about causal questions in **causal diagrams** (DAGs)
- to develop **research designs based on DAGs**
- to detect **common pitfalls in empirical analyses**

This lecture is based on

- MHE, Ch. 2, 3.2
- CIM, Ch. 4, 5

Causality

Oxford dictionary: **the relationship between cause and effect**

Causality is a theoretical concept. It **cannot be** (directly) **tested with data**

$\Rightarrow$ to make a causal statement, one needs a **clear theory**

The **methods of causal inference** are "rhetorical devices"

- they allow us to establish causality **under certain assumptions**
- since we want to **identify a causal effect**, these are called **identifying assumptions**

Causality

Formally, in econometrics (and beyond), causality involves two random variables: a **treatment** D and an **outcome** Y

$$D \rightarrow Y$$

The **treatment** can either be **binary,** $D \in \{0, 1\}$ or **continuous** $D \in \mathbb{R}$

We speak of a **causal effect of D on Y** if a **change in D triggers a change in Y**

Causal Diagrams

Causal diagrams (also called "**directed acyclical graphs**", or DAGs) are a powerful tool to understand:

- how **causal effects** can be identified from **observational data**
- which **variables** we should or should not **condition on**

DAGs are common in **computer science** and are slowly making their way into econometrics

Here we will briefly introduce DAGs.

Book recommendation:

- *The Book of Why* (Pearl & Mackenzie, 2018)
- For a more profound treatise, see Pearl (2009)

Causal Diagrams

Ingredients

- **nodes**: random variables
- **arrows**: causal relationships
- missing arrows indicate the absence of a causal relationship

Direct causal effect of the **treatment** D on the **outcome** Y

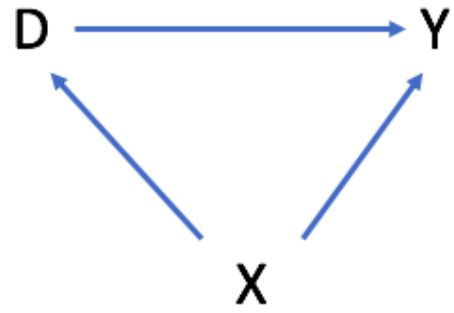
$$D \to Y$$

Indirect causal effect: D affects Y through a **mediator X**

$$D \to X \to Y$$

Causal Diagrams - Confounders

A common challenge in applied econometrics is to **separate a causal effect** from the **influence of confounders**



Here we have two paths:

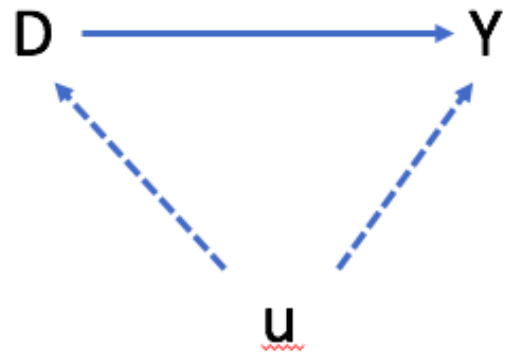
- The **direct path**: $D \rightarrow Y$
- A **backdoor path**: $D \leftarrow X \rightarrow Y$

As long as there is no collider (introduced in a few slides), we speak of **backdoor path with a condfounder** as being ***open***

We can only **identify the causal effect** $D \rightarrow Y$ if we **condition on/adjust for X**

Causal Diagrams - Confounders

Problem: **often we don't observe a confounder**



u lies on the **backdoor path** from *D* to *Y* but is **unobservable** ($\Rightarrow$ dashed line)

- open backdoor $\Rightarrow$ *u* is a confounder

Problem: selection into treatment. In microeconomics we learn

- **people** make **rational choices**...
- ...as do **firms**
- ...as do **governments**

Causal Diagrams - Confounders

Examples for **selection into treatment**:

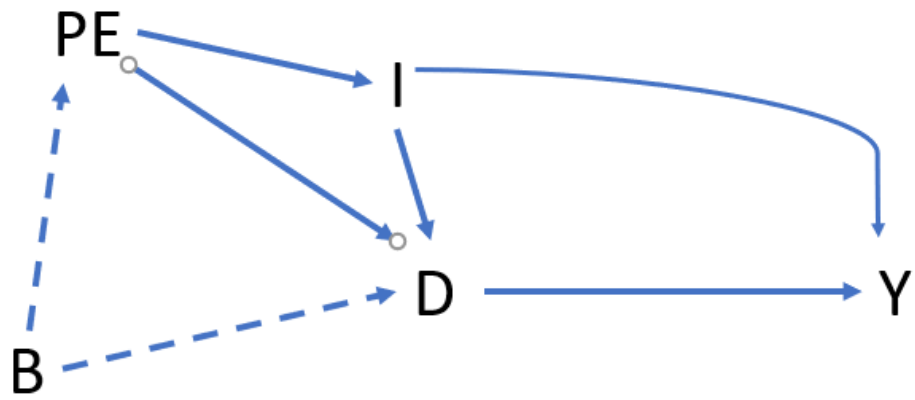
Going to the gym makes you healthier

- good reason to believe so
- but people who go to the gym are different from those who don't
- observed correlation ≠ causation

Exporting boosts firm profitability

- good reason to believe so
- but exporters are different in many ways from non-exporters
- observed correlation ≠ causation

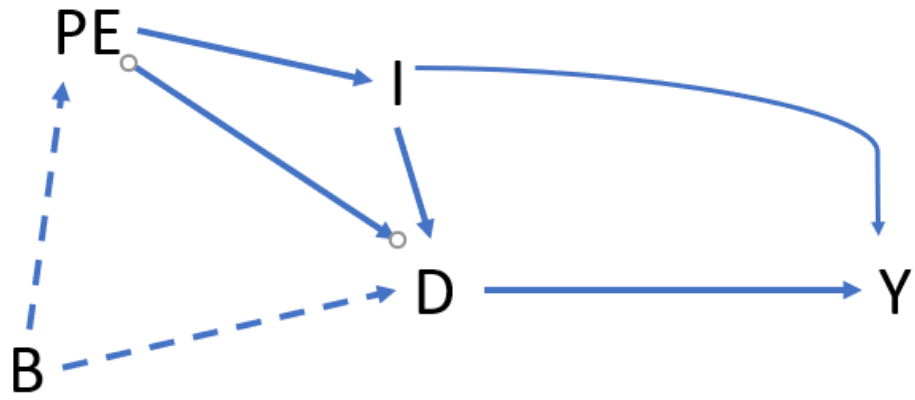
Causal Diagrams - Confounders



We are interested in the **effect of education D on earnings Y**, but also need to think about parental education (PE), family income (I) and unobserved family background (B)

- **Causal effect**: $D \rightarrow Y$
- **Backdoor path 1**: $D \leftarrow I \rightarrow Y$
- **Backdoor path 2**: $D \leftarrow PE \rightarrow I \rightarrow Y$
- **Backdoor path 3**: $D \leftarrow B \rightarrow PE \rightarrow I \rightarrow Y$

Causal Diagrams - Confounders

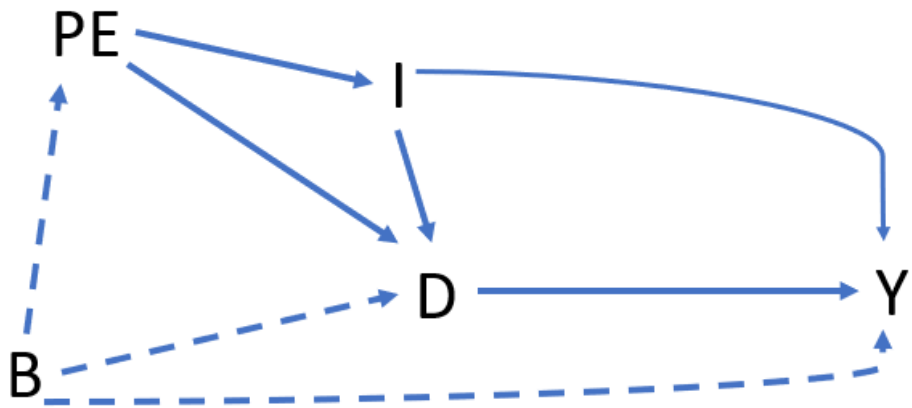


To identify the causal effect, we need to **shut the backdoor paths** 1-3

- we can do so by **conditioning on *I***
- i.e. we control for *I* in a regression
- we could also control for *PE*, but this wouldn't help with identification

Causal Diagrams - Confounders

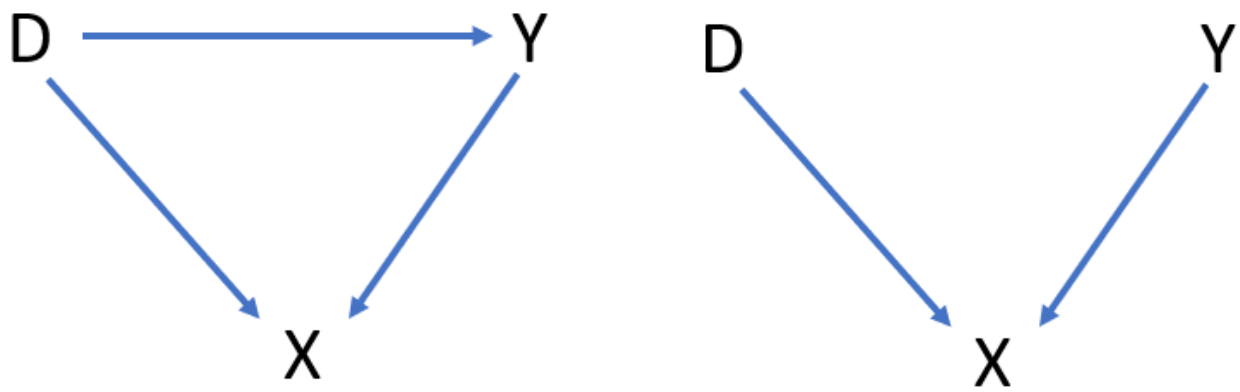
Note that this reasoning **depends on the DAG being the correct one**



- If $B \rightarrow Y$, we would have an **additional open backdoor path**
- In that case, **controlling for** I would **not be sufficient**
- If we cannot observe B, we know that our estimate is most likely biased

Causal Diagrams - Colliders

Unlike confounders, **colliders** are a little known **source of bias**

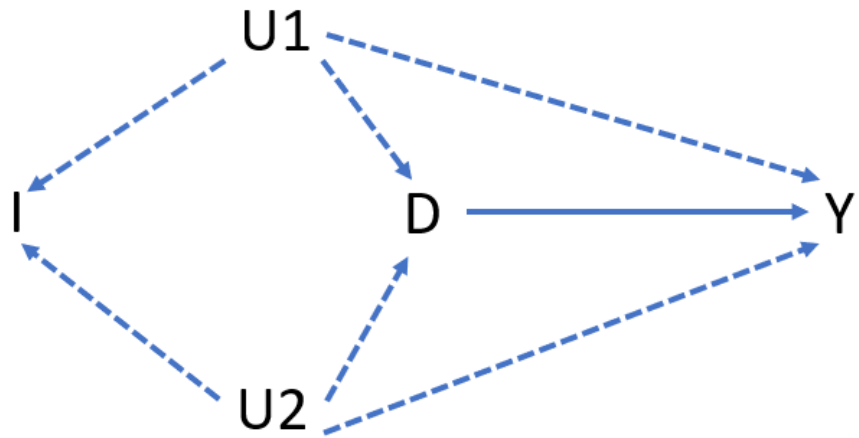


In both examples the **backdoor path** $D \rightarrow X \leftarrow Y$ is **closed**

Conditioning on a collider can open a backdoor path and **lead to bias**

- In particular, it can induce a **spurious correlation** (between D and Y)

Causal Diagrams - Colliders



To deconfound $D \rightarrow Y$, we would need to **control for** $U1$ **and** $U2$

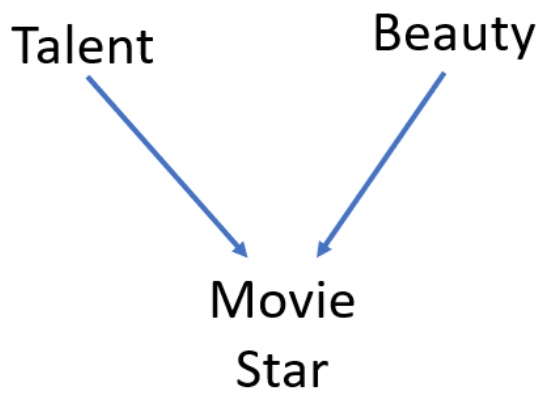
But what if we **controlled for an observable variable** I **instead?**

- $D \leftarrow U1 \rightarrow I \leftarrow U2 \rightarrow Y$
- $D \leftarrow U2 \rightarrow I \leftarrow U1 \rightarrow Y$

Controlling for I makes the situation worse because it opens both backdoor paths

Colliders - Example from Cunningham (2020)

...among **movie stars**, we can observe a **negative correlation between talent and beauty**



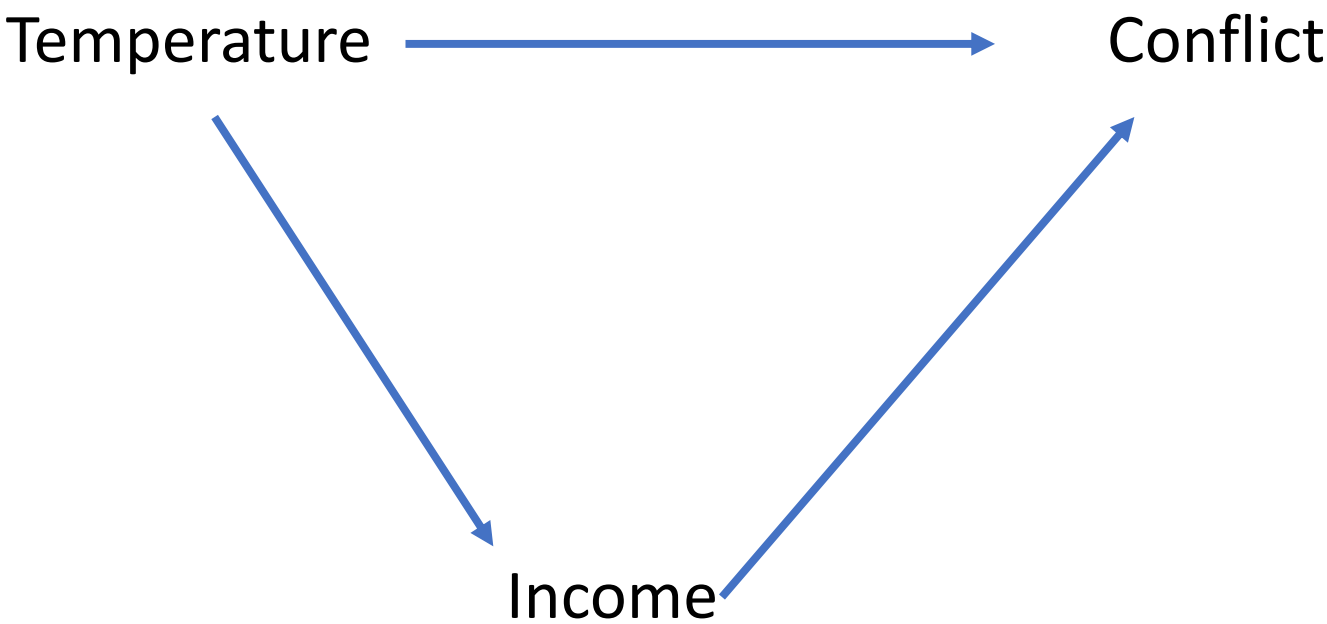
If talent and beauty are unrelated in the population,

- then the observed correlation may reflect **collider bias**
- due to **non-random sample selection**

Colliders - Example from Cunningham (2020)

Suppose movie stars are those in the top 15% of *score=beauty+talent*

The Bad Control Problem: Condition on a Mediator



"*We estimate the effect of temperature on conflict irrespective of income*"

Credit: Marshall Burke's Blog (G-FEED)

The Bad Control Problem

Conditioning on a mediator introduces **selection bias**

Income is not as good as randomly assigned.
$\Rightarrow$ it is a function of temperature

Conditioning on income will lead to a downward bias

- The direct effect is probably negative
- Temperature reduces income
- Lower income $\rightarrow$ more conflict

The Bad Control Problem

Simulation results (true effect in Column 1):

	(1) conflict	(2) conflict
temperature	0.0540***	0.0402***
	(80.43)	(30.77)
income		-0.00277***
		(-12.30)
_cons	-0.557***	-0.558***
	(-52.61)	(-53.15)
N	10000	10000

The Bad Control Problem

In many cases, bad control problems can be easily detected

- If a variable is on the **causal path, don't control for it**

But sometimes **bad controls** are the result of **sample selection**.

Example: racial bias in policing

Racial Bias in Police Use of Force (Fryer, 2019)

Administrative data from NYC, Texas, Florida, LA County

Observes all stops of the police

- race of person stopped
- use of force by the police
- contextual variables (place, time, ...)

Findings:

- Disproportionate use of force against Blacks and Hispanics
- This is true even when controlling for context

Racial Bias in Police Use of Force (Fryer, 2019)

Fryer acknowledges several **potential problems**:

- Mis-reporting of the use of force
- Probability of interacting with the police is higher for Blacks
- Whites and Blacks stopped by the police may differ on average

Critique by Knox *et al.* (2020): bias "goes deeper"

Bad Controls: Endogenous Sample Selection

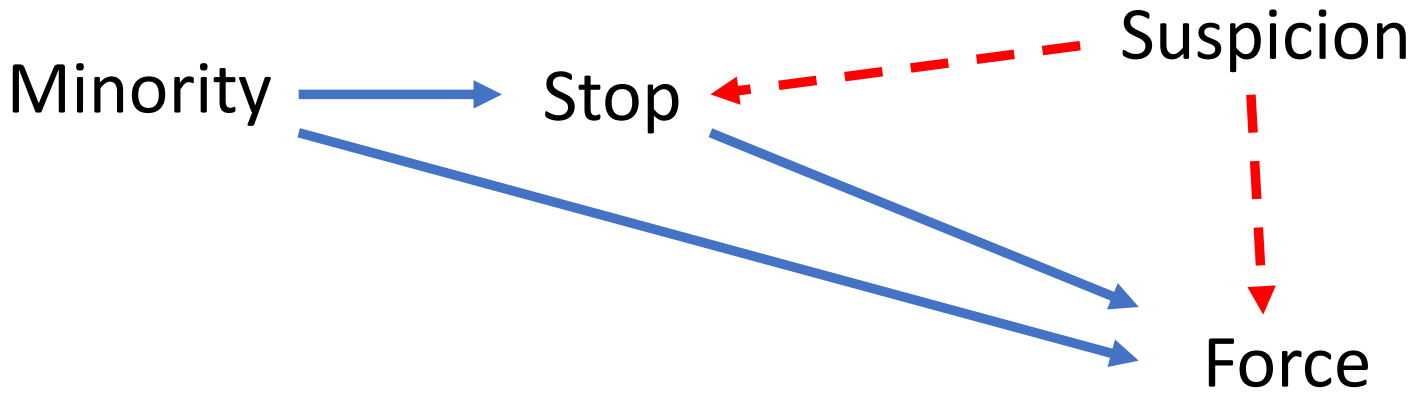
Problem: it is **not random who is stopped by the police**

- Officer behavior is unobservable
- No information on people who are observed but not investigated

Knox *et al.* (2020): this is equivalent to

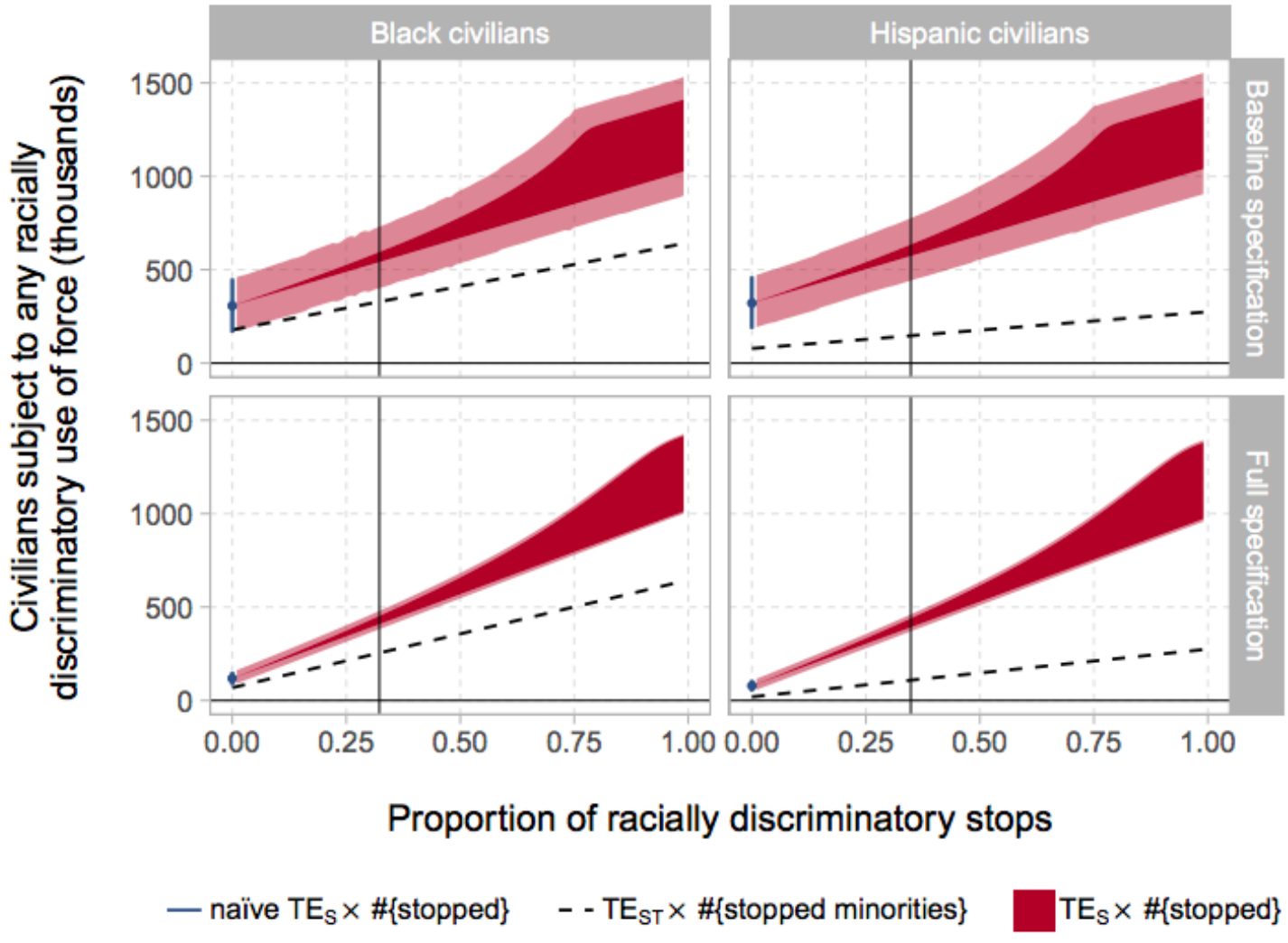
- **conditioning on a mediator**
- while not accounting for a confounder

Bad Controls: Endogenous Sample Selection



Studies only use observations with *Stop* $= 1$

Bounding exercise in Knox *et al.* (2020)



⇒ Ignoring the probability of stopping leads to a **severe underestimation** of the racial gap in use of force

Controlling for Variables in a Regression

The **main takeaway** from studying **causal diagrams**

- they clarify **which variables** we should (and should not) **control for**

Control for **confounders** (use the backdoor criterion)

Do not control for **colliders**

Do not control for **mediators** (“bad controls”)

Controlling for Variables in a Regression

Causal diagrams are rarely shown in papers, but they are a very **useful first step** when thinking about **causality**

A researcher has to **take a stand on causal relationships** between variables

- what is a confounder, mediator, collider?
- this requires some theoretical reasoning
- and cannot be answered just by looking at data

Further Readings

Imbens (2020): PO vs DAGs

- Self-recommending

Montgomery *et al.* (2018): bad control problem in experiments

- Insightful description based on potential outcomes and DAGs

Schneider (2020): collider bias in economic history research

- How to detect and overcome collider bias (applications)

References I

Cunningham, Scott. 2020. *Causal Inference: The Mixtape*. Yale University Press.

Fryer, Roland G. 2019. An Empirical Analysis of Racial Differences in Police Use of Force. *Journal of Political Economy*, **127**(3), 1210–1261.

Imbens, Guido W. 2020. Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature*, **58**(4), 1129–79.

Knox, Dean, Lowe, Will, & Mummolo, Jonathan. 2020. Administrative Records Mask Racially Biased Policing. *American Political Science Review*, **114**(3), 619–637.

Montgomery, Jacob M., Nyhan, Brendan, & Torres, Michelle. 2018. How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It. *American Journal of Political Science*, **62**(3), 760–775.

Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*. 2nd edn. New York, NY, USA: Cambridge University Press.

Pearl, Judea, & Mackenzie, Dana. 2018. *The Book of Why: The New Science of Cause and Effect*. 1st edn. New York, NY, USA: Basic Books, Inc.

Schneider, Eric B. 2020. Collider bias in economic history research. *Explorations in Economic History*, **78**, 101356.