ECON42720 Causal Inference and Policy Evaluation

1 Foundations of Causality

Ben Elsner (UCD)

About this Lecture

In this lecture, we will learn about **causal inference** based on **causal diagrams** (also called **directed acyclical graphs**, or 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**

Resources

This lecture is based on

- Cunningham (2020), Mixtape: Chapter 3
- Huntington-Klein (2022), The Effect: Chapters 6-8

The Effect is very accessible and goes step-by-step through all things DAGs. *The Mixtape* chapter is more concise. Both are excellent resources.

Part of the material of this lecture is also covered in the first two lectures of my **YouTube playlist** on Causal Inference: LINK

Find more about the course on the **course page**

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**

$$D \to X \to Y$$

How to Construct and Use a DAG

Step 1: Construct a DAG

1. **Identify the causal question** you want to answer
2. **Identify the variables** that are relevant for the causal question
3. **Draw a DAG** that represents the causal relationships between the variables

Challenges:

- Which DAG is the right one?
- Every arrow and the absence of an arrow is an assumption
- Is the DAG too simplistic or too complex?

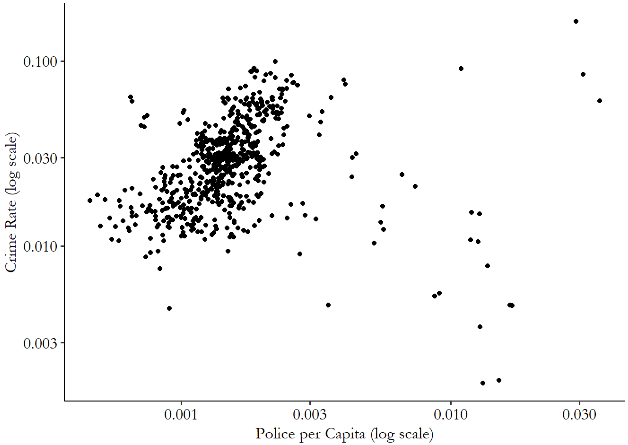
How to Construct and Use a DAG

Step 2: Causal Identification: use a DAG to identify the causal effect of interest

- Determine which causal forces you need to eliminate to answer the causal question
- This gets done through **closing back door paths** (more on that soon)
- Once that's done we can use **standard econometric methods** to estimate the causal effect of interest
- But that's also what's really hard in practice

Example: Police Presence and Crime

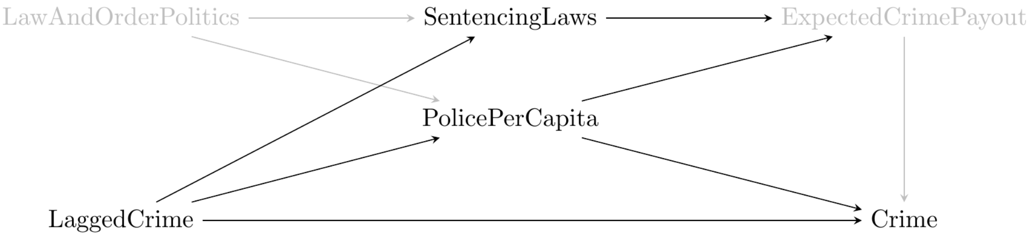
Classic case: the correlation is positive. Why?



Source: The Effect, Figure 6.5

Example: Police Presence and Crime

A DAG can help. . .



Source: The Effect, Figure 6.6

Example: Police Presence and Crime

Assumptions in the DAG:

1. LaggedCrime doesn't cause LawAndOrderPolitics
2. PovertyRate isn't a part of the data generating process
3. LaggedPolicePerCapita doesn't cause PolicePerCapita (or anything else for that matter)
4. RecentPopularCrimeMovie doesn't cause Crime

Trade-off

- **Omit too many variables**: DAG is too simplistic and we may omit variables that are very important
- **Omit too few variables**: if a DAG becomes too complex, it is very difficult to identify the causal effect of interest

How to Draw Causal Diagrams

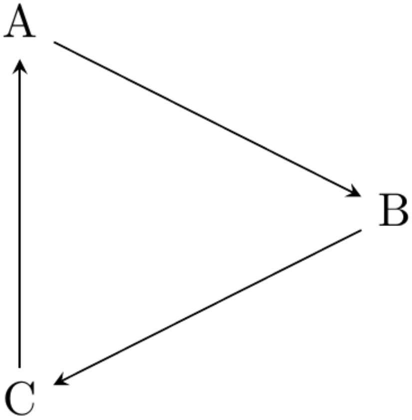
Huntington-Klein (2022), Ch. 7, offers many useful tips on how to draw causal diagrams

- **Thinking about the "data-generating process"**, i.e. all the relevant variables and their causal relationships
- **Simplifying DAGs** by getting rid of redundant or unimportant variables
- **Avoiding cycles** (i.e. loops) in DAGs

Cycles in DAGs

Cycles are a big **no-no in DAGs**

- They imply that a variable causes itself
- Challenge: teach a cycle to a computer. . .



How to Avoid Cycles?

$$\text{IPunchYou} \longleftrightarrow \text{YouPunchMe}$$

We can **add a time dimension**. . .

$$\begin{tikzcd}
\text{IPunchYou}_t \arrow[dr] & \text{IPunchYou}_{t+1} \\
\text{YouPunchMe}_t \arrow[ur] & \text{YouPunchMe}_{t+1}
\end{tikzcd}$$

How to Avoid Cycles?

Or we can add a **variable that only affects one of the variables in the cycle**

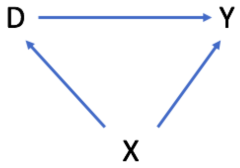
Example: I flip a coin, and if it's head I'll. . .

$$\text{CoinFlip} \longrightarrow \text{IPunchYou} \longrightarrow \text{YouPunchMe}$$

Now there is no cycle because the coin flip only affects my decision, but not yours.
You just react to my decision.

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:

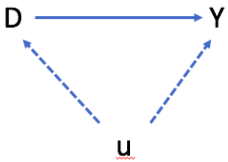
- The **direct path**: $D \rightarrow$ Y
- A **backdoor path**: $D \leftarrow X \rightarrow Y$

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**:

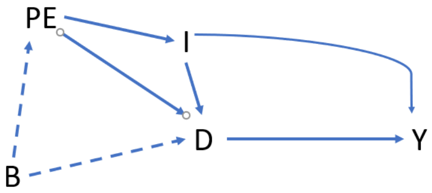
Going to the gym makes you healthier

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

Exporting boosts firm profitability

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

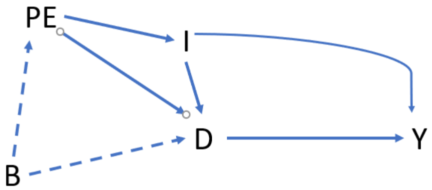
Causal Diagrams - Confounders



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

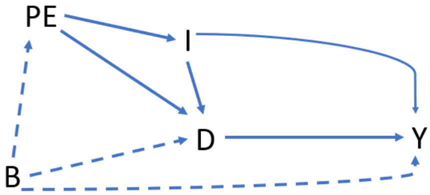


To identify the causal effect, we need to **shut the backdoor paths** 1-3

- 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

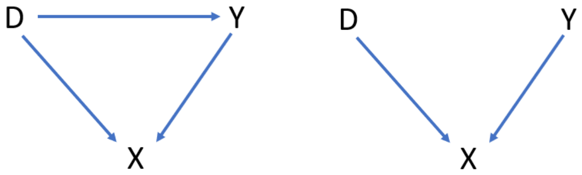
Note that this reasoning **depends on the DAG being the correct one**



- 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**

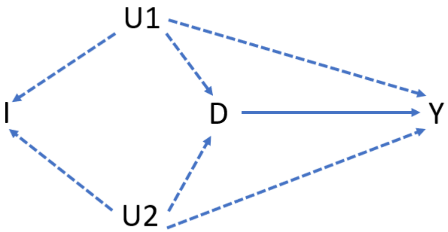


In both examples the **backdoor path** $D \rightarrow X \leftarrow Y$ is **closed**

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** $U2$

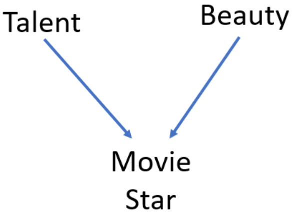
But what if we **controlled for an observable variable** I **instead?**

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

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**



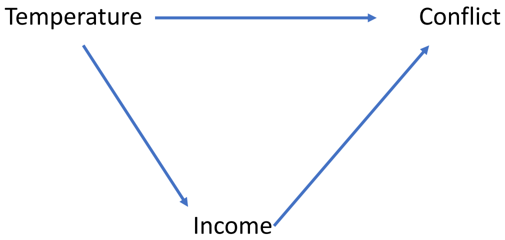
If talent and beauty are unrelated in the population,

- 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**

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

Conditioning on income will lead to a downward bias.

- The direct effect is probably positive
- High 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

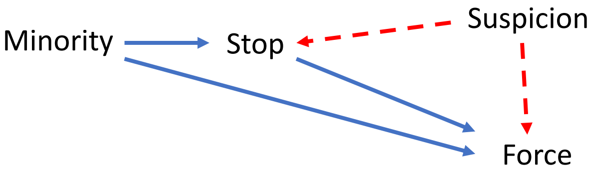
Problem: it is **not random who is stopped by the police**.

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

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

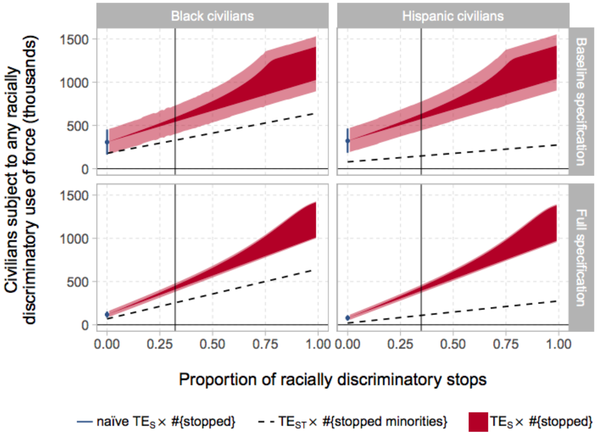
- **conditioning on a mediator**
- while not accounting for a confounder

Bad Controls: Endogenous Sample Selection



Studies only use observations with $Stop = 1$

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



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

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)

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”)

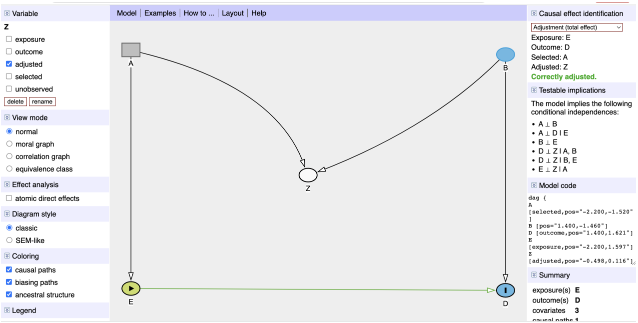
Controlling for Variables in a Regression

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

Drawing DAGs: Dagitty



Link to Dagitty Browser

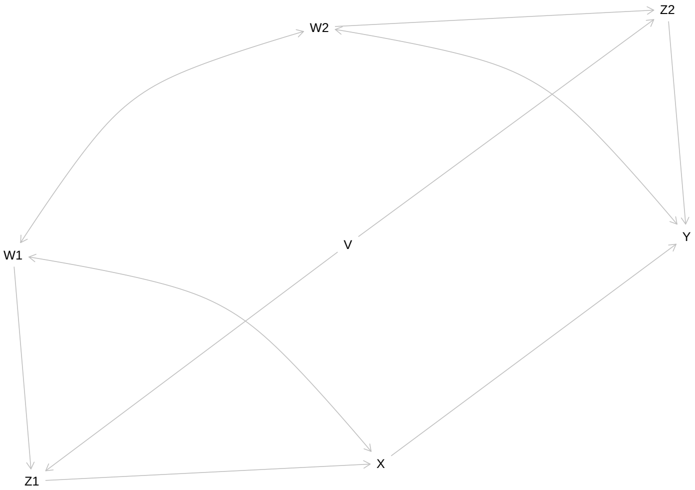
Drawing DAGs with R and Dagitty

```
library(dagitty)

g1 <- dagitty( "dag {
    W1 -> Z1 -> X -> Y
    Z1 <- V -> Z2
    W2 -> Z2 -> Y
    X <-> W1 <-> W2 <-> Y
}")

plot(graphLayout(g1))
```

Drawing DAGs with R and Dagitty



References

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.

Huntington-Klein, Nick. 2022. *The Effect: An Introduction to Research Design and Causality*. Chapman and Hall/CRC.

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.

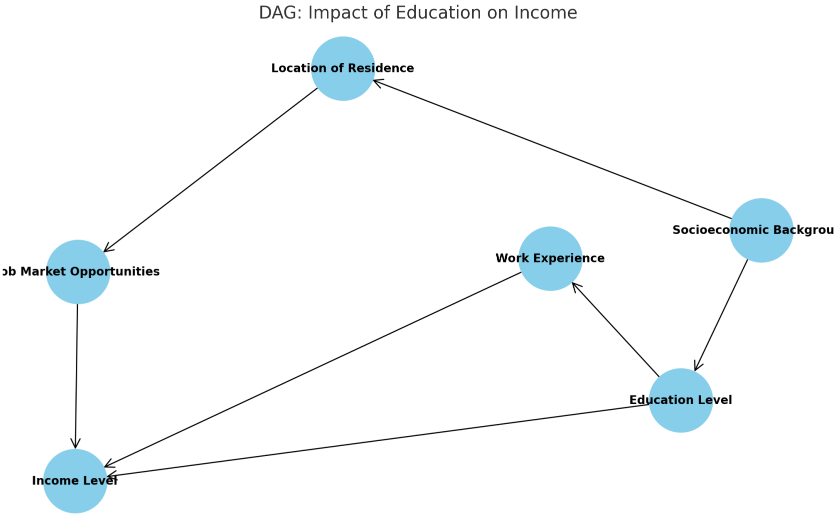
Group work 1

Draw a DAG for the following **causal relationships**:

- Exporting (by a firm) $- >$ Firm profitability
- Participation in a job training programme $- >$ likelihood of re-employment
- Exposure to an earthquake *in-utero* $- >$ health at age 50
- Attendance of a mixed-sex school $- >$ gender attitudes later in life
- Experience of conflict early in life $- >$ voting later in life

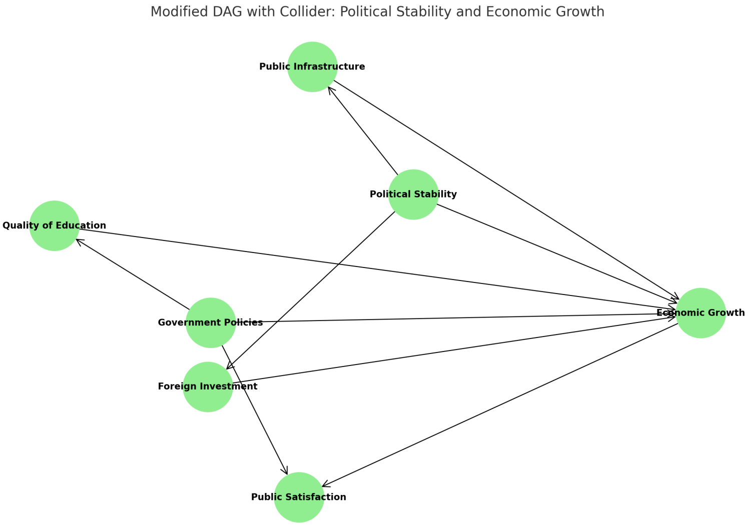
Group work 2

Deconfound the following DAG of the effect of education on income:



Group work 3

Deconfound the following DAG of the effect of education on income:



benjamin.elsner@ucd.ie

www.benjaminelsner.com

Sign up for office hours

YouTube Channel

@ben_elsner

LinkedIn

Contact

Prof. Benjamin Elsner
University College Dublin
School of Economics
Newman Building, Office G206
benjamin.elsner@ucd.ie

Office hours: book on Calendly