GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Dr. Christian Czymara

FORSCHUNGSPRAKTIKUM I UND II: LÄNGSSCHNITTDATENANALYSE IN R

Directed acyclic graphs
session ii

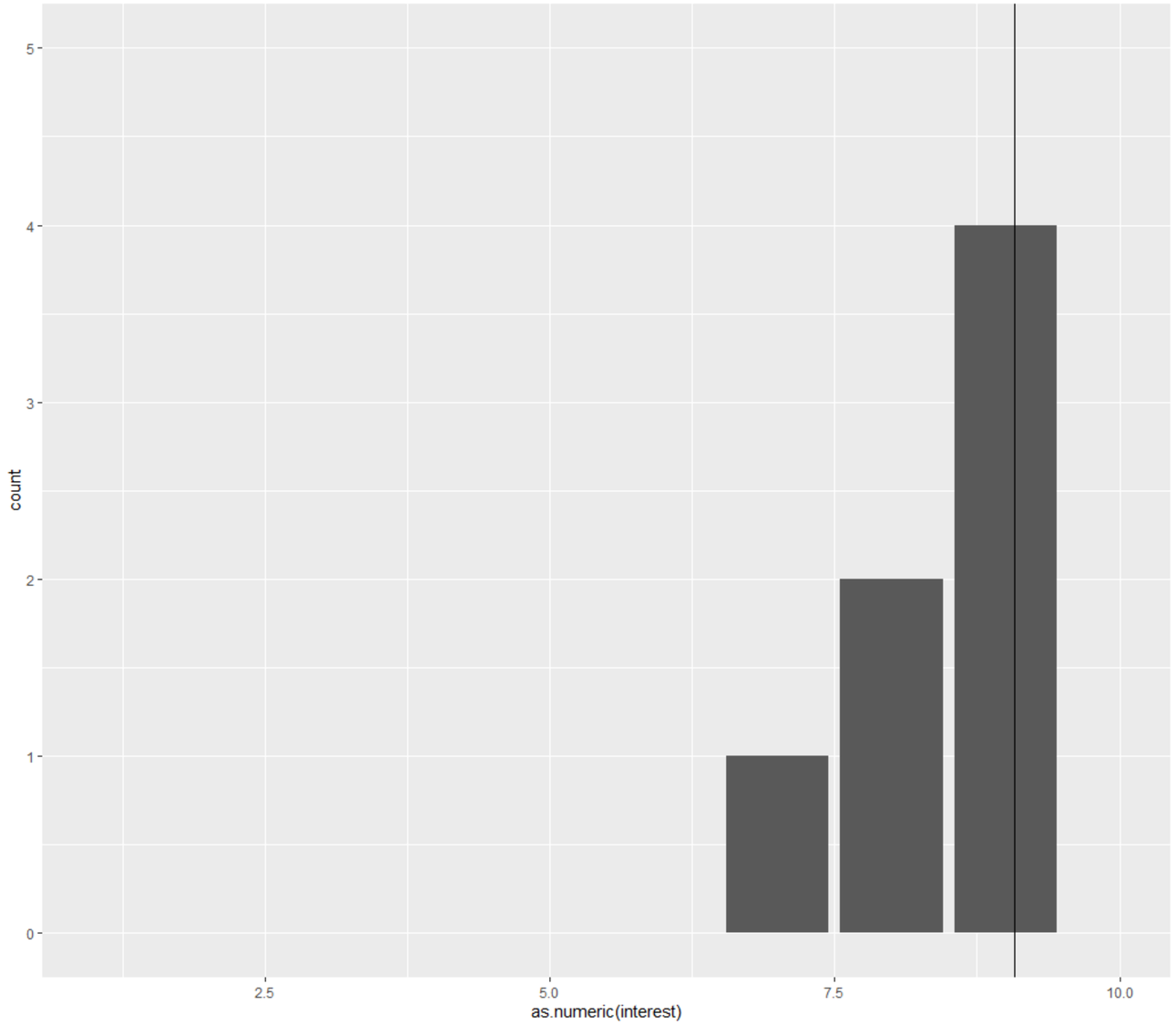
AGENDA

- Results of survey
- This week's focus will be on theory
- I.e., under which conditions can we justify to talk about a *causal* effect?
- Which implications does this have for our empirical model?
  - Which variables do we have to control?
  - Which variables must we *not* control?

SURVEY RESULTS

YOUR INTEREST

- Super motivated class: mean is 9/10

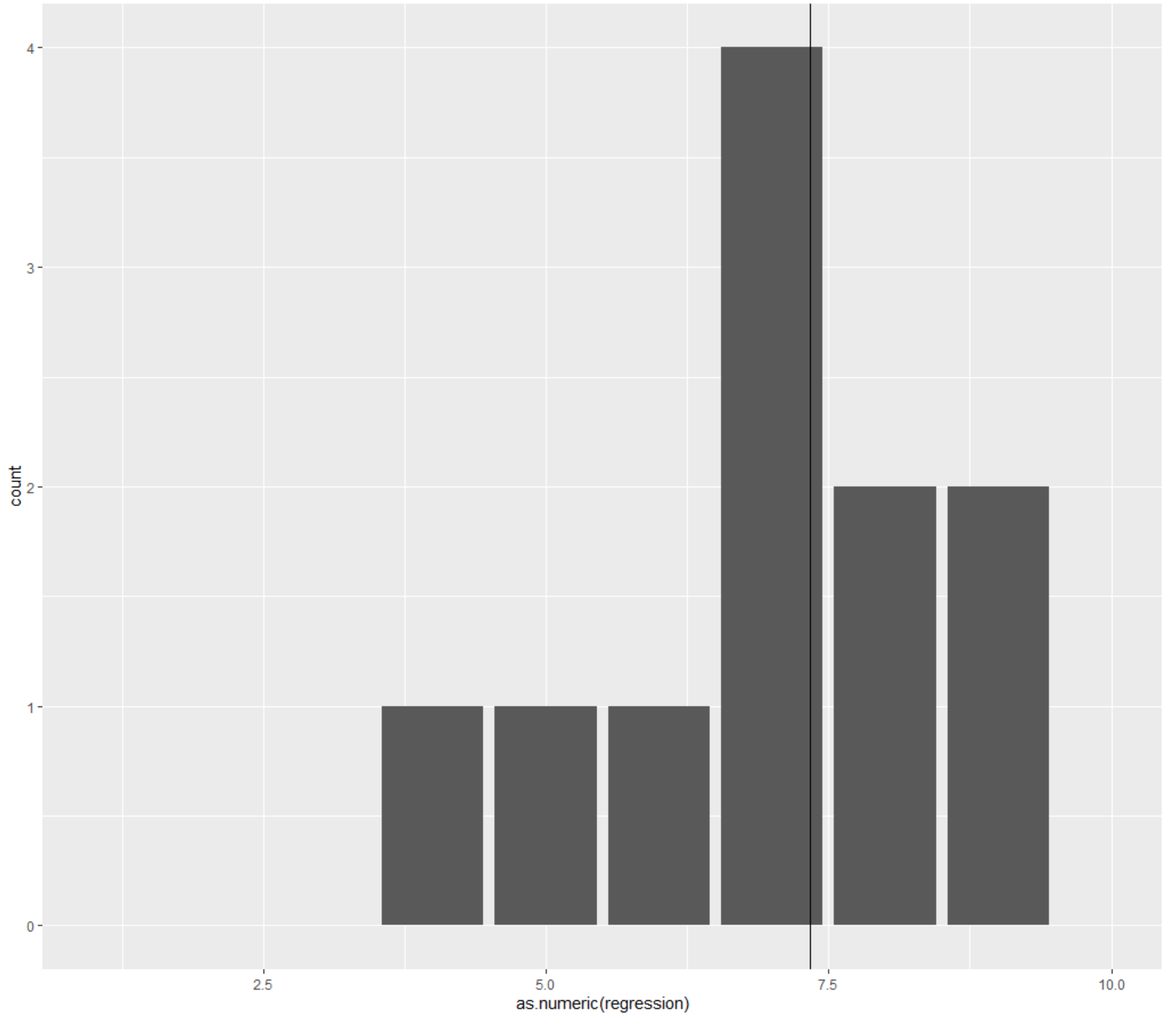


YOUR EXPECTATIONS

- *Auffrischung von Grundlagen*
- *Datenaufbereitung*
- *positiv- und negativ Beispielstudien*
- *missing values*
- *text data as time series / wiederholte Expertenumfragedaten zu pol. Parteien*
- *wärend des Kurses mit den eigen Daten arbeiten*
- *Unterstützung bei der Auswahl der Daten und der Fragestellung*
- *Understand how to write the analysis paper (what goes into the paper, what doesn't*

YOUR KNOWLEDGE

- Knowledge also rather high
- Two thirds are pretty familiar with regression models (>=7)
- A fourth has medium knowledge (between 4 and 6)
- No one is unfamiliar (below 4)



YOUR KNOWLEDGE

- Two thirds already read studies that analyze panel data
- … No one has never heard of it
- Over 80% have used linear regression before, over 40% logistic regression
- … but there are also students how don't know what linear or logistic regression is
- Almost 60% already work with R

ANNOUNCEMENTS

ADDITIONAL MASTER SEMINAR FOR THOSE INTERESTED IN ANALYZING PANEL DATA

- *Analyses of short-term changes over life events (seminar with hands-on panel data research)*
- https://qis.server.uni-frankfurt.de/qisserver/rds?state=verpublish&status=init&vmfile=no&moduleCall=webInfo&publishConfFile=webInfo&publishSubDir=veranstaltung&veranstaltung.veranstid=337704

PRAKTIKUM BEI DER STADT FRANKFURT

STADT FRANKFURT AM MAIN

DER MAGISTRAT

Bürgeramt, Statistik und Wahlen

Im Rahmen der Umfrage „Leben in Frankfurt 2022“ suchen wir ab August 2022

eine*n studentische*n Praktikant*in (m/w/d).

Die Umfrage „Leben in Frankfurt“ wird unter Bürgerinnen und Bürgern seit 1993 jährlich durch die Stadt Frankfurt am Main durchgeführt. Sie stößt bei der Bevölkerung auf eine hohe Akzeptanz und ist eine der etablierten städtischen Umfragen in Deutschland.
Im Rahmen der Umfrage mit mehr als 23.000 angeschriebenen Personen suchen wir, für konzeptionelle und redaktionelle Tätigkeiten, Studierende für ein Mehrwöchiges Praktikum im Zeitraum zwischen August und Oktober 2022.

Das erwartet Sie:

- Codierung von Befragungsergebnissen
- Verarbeitung und Aggregation umfangreicher Datenmatrizen
- Plausibilisierung von Rohdatensätzen
- Anwendung und Prüfung von Gewichtungsverfahren
- Berechnung von Auswertungsvariablen und Indikatoren
- Vorbereitung von Grafiken für Veröffentlichungen
- Redaktionsarbeiten im Rahmen der Veröffentlichung „Frankfurter Umfragen“
- ggf. Mitarbeit bei der Erstellung von Kurzpublikationen
- Prüfung von Umfrageergebnissen des Jahres 2022 und Konzeption sowie Vorbereitung der nächsten Welle im Jahr 2023

Das bringen Sie mit:

- Interesse an stadtgesellschaftlichen Entwicklungen
- ein Faible für empirische Sozialforschung
- methodisches Handwerkszeug bei der Verarbeitung von quantitativen Daten
- im Idealfall Kenntnisse in R

Das bieten wir:

- interessante Einblicke in eine der größten kommunalen Umfragen in Deutschland
- Mitarbeit in einem innovativen Team
- Anleitung mit methodischem Know-How für alle Arbeiten
- flexible Arbeitszeitgestaltung in Absprache mit der Projektleitung
- einen modernen Arbeitsplatz in der Innenstadt von Frankfurt am Main

Für weitere Auskünfte steht Ihnen Herr Stein unter der Rufnummer (069) 212-33422 oder per E-Mail unter umfragen@stadt-frankfurt.de zur Verfügung.

Wir freuen uns auf Ihre aussagekräftige Bewerbung. Bitte bewerben Sie sich bis zum 15.05.2022 per E-Mail an christian.stein@stadt-frankfurt.de.

CAUSAL GRAPHS

THE DIRTY C-WORD

- Nowadays, social scientists working with observational data are hesitant to make explicit causal statements
- Reviewer 2: "*Please remove causal language …*"
- However, these statements are then often made *implicit*
- "*Similar to when sex or drugs are made taboo, making explicit causal inference taboo does not stop people from doing it*" (Grosz et al. 2020: 11)
- This status quo is not helpful to anyone

DIRECTED ACYCLIC GRAPHS

- Instead of stopping to address causality altogether, be transparent about the assumptions that lie behind your causal effect
- Readers can decide whether or not they buy these assumptions
- One of the most accessible ways to make causal structure, and its underlying assumptions, transparent are directed acyclic graphs (DAGs)

TERMINOLOGY

ELEMENTS OF A DAG

1. Variables (nodes)
2. Arrows: Possible direct causal effects
3. Missing arrows: No direct causal effect (strong assumption!)

- DAGs are nonparametric: no statement about distribution of variables, or about functional form of effects or effect size

DIRECTED & ACYCLIC

- Directed: Arrows always point in (only) one direction, such as $x \rightarrow y$
- Acyclic: No cycles allowed (no $x \rightarrow y \rightarrow x$), *the future can't cause the past*

SIMPLE DAG EXAMPLE

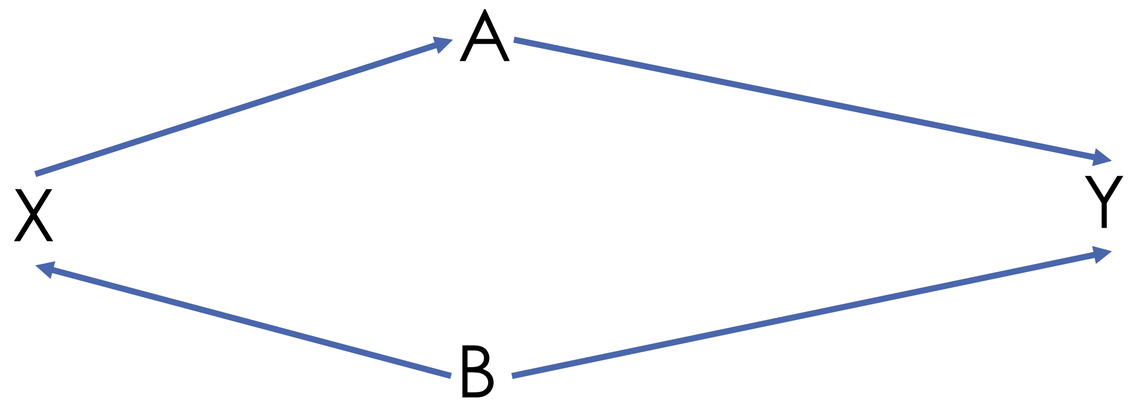
CONCEPTS

- Children: Variables directly caused by another variable (*A is a child of B*)
- Descendants: All variables caused by a variable (*A and Y are descendants of B*)
- Parents: Direct causes of variable (*B is a parent of A*)
- Ancestors: All direct and indirect causes of a variable (*A and B are ancestors of Y*)

CONCEPTS

- Paths: sequences of connected arrows (direction irrelevant, but only once through each variable)
- Cause paths: paths where all arrows point from x to y
- Non-causal paths: all paths that are not causal paths
- *Causal and non-causal paths are defined relative to a specific x and y* (see Keele et al. 2020)
- Aim: block all non-causal paths from x to y

SLIGHTLY MORE COMPLICATED DAG EXAMPLE



- $x \rightarrow a \rightarrow y$
  - → Causal path
- $x \leftarrow b \rightarrow y$
  - → Non-causal path

CONCEPTS

- Confounder: Variable that causes x and y
- Mediator: Variable that is caused by x and causes y
- Collider: Variable that is caused by x and y

ABSENCE OF EFFECTS YIELD REAL KNOWLEDGE

- Conventional social theory focusses on what relationships exist
- DAGs turn this logic around by focusing on which relationships do *not* exist
- *"Present arrows represent the analyst's ignorance. Missing arrows, by contrast, represent definitive claims of knowledge. It is the missing arrows […] that enable the identification of a causal effect. Adding arrows to an existing set of variables in a DAG […] never aids nonparametric identification."* (Elwert 2013: 248)

CONCEPTS

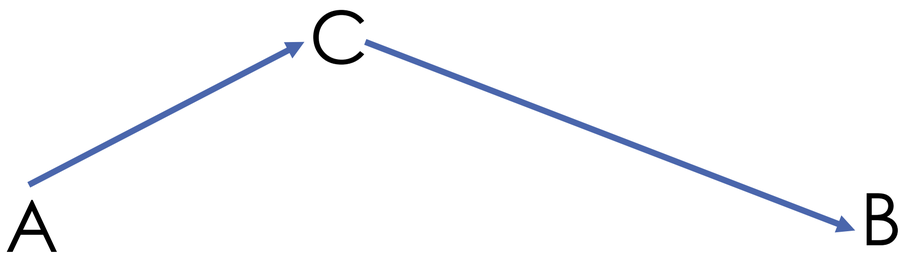
- D-separation: Path between two variables is *d-separated* (blocked or closed) when it…
  - Contains a non-collider that has been conditioned on
  - Contains a collider that has not been conditioned on (and no descendant of any other collider on the path has been conditioned on)
- Path is *d-connected* (unblocked or open) if it is not d-separated
- d: directional
- Conditioning means incorporating some information about that variable into the analysis (in our context, I will usually mean statistical controlling in regression models)

THREE SOURCES OF ASSOCIATION

CAUSAL STRUCTURES

1. Chains: A→C→B
2. Forks: A←C→B
3. Inverted forks: A→C←B

CAUSATION



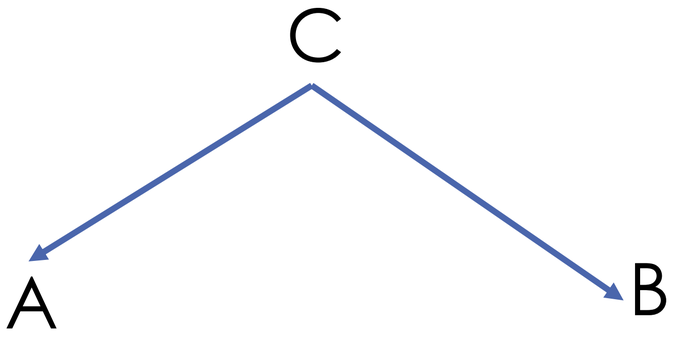
- Chains: A→C→B
- Two variables can be associated because one causes the other
- Here: A is associated with B because A indirectly causes B via mediator C
- Controlling for C would block the association from A to B

→*Overcontrol bias*

EXAMPLE

- Is there a race penalty in soccer (i.e., do referees give red cards more often to black than to white players)?
- Direct effect of race on red card
- Assume that black players more often play defense position
- … and defense players get more red cards
- Indirect effect of race through position
- Thus, controlling for position would remove part of the causal total effect of race

CONFOUNDING

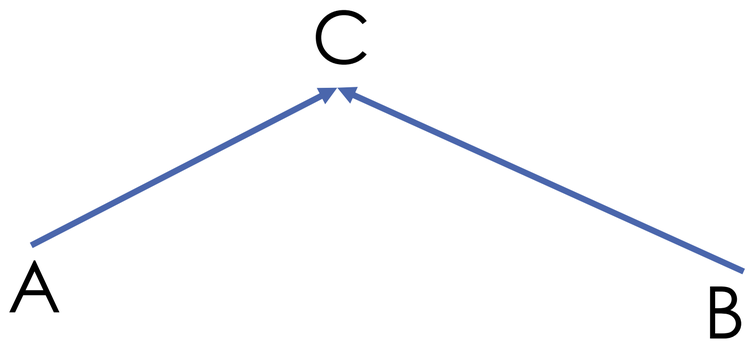


- Forks: A←C→B
- C is a "common cause" of A and B
- The association between A and B is spurious (biased) because it is induced only via C
- In this scenario, C is a "classical" control variable
- Controlling C would identify the causal (zero) effect of A on B

EXAMPLE

- Does watching soccer kill you?
- You might observe this association
- However, we know that men watch soccer more often
- … and men die earlier (higher risk behavior)
- Hence, gender is a confounder
- Accounting for gender closes the non-causal path from watching soccer to mortality (effect of watching soccer should become insignificant when gender is controlled)

ENDOGENOUS SELECTION

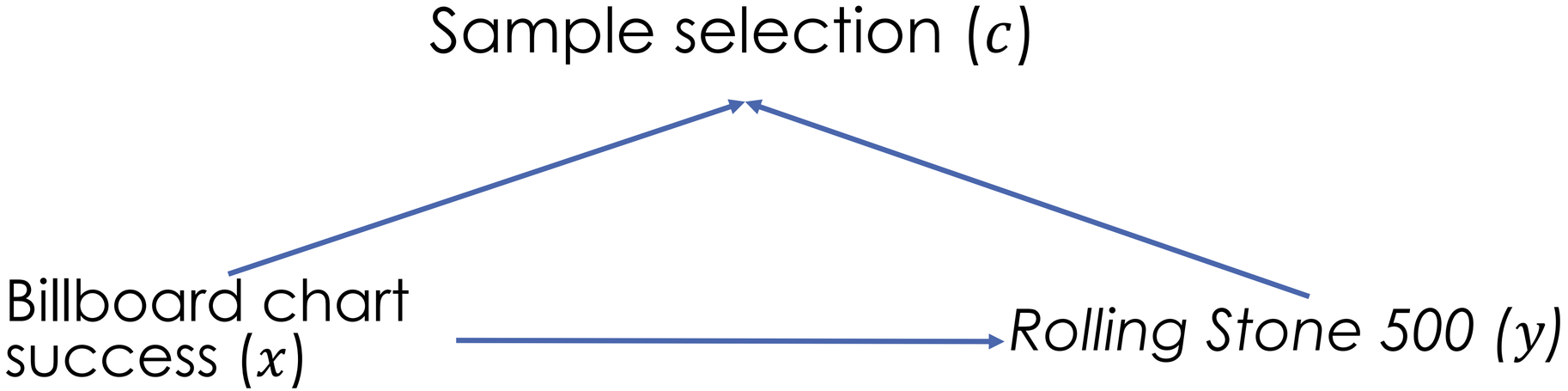


- Inverted forks: A→C←B
- C is a "common outcome" of A and B
- Not controlling C identifies the causal (zero) effect of A on B
- Controlling C would induce a spurious (non-zero) association between A and B
- *The association between A and B when controlling C is not causal!*

EXAMPLE

- Are commercially more successful music albums more likely to be included in *Rolling Stone* magazine's list of *500 Greatest Albums of All Time*? (Schmutz 2005)
- Data: *Rolling Stone 500* list plus 1,200 other successful albums
- Result: topping the Billboard charts has strong negative effect on being in the *Rolling Stone 500*
- What happened here?

EXAMPLE



- By sampling construction, x and y have strong positive effect on c
- Analyzing this sample implies conditioning on the collider *sample selection*
- Less successful albums severely undersampled
- See Elwert & Winship (2014)

SUMMING UP

HOW DO DAGS HELP US?

- DAGs are no panacea for causal inference
- However, DAGs force researchers to think about the causal structure of their theoretical model
- DAGs provide objective rules which variables should be controlled under which conditions
- Maybe most importantly, DAGs make theoretic assumptions underlying the empirical model very transparent

→Other can judge whether they "believe" the theoretical model – and thereby the empirical estimate – or not

TAKE HOME MESSAGES

- *"[…] it is not possible to decide on the proper set of control variables without an understanding of the underlying causal structure"* (Elwert 2013: 262)
- *"Carefully articulate an identification strategy. All causal analyses require one.*
- *Each treatment of interest requires a separate assessment of identification.*
- *Researchers should avoid providing any interpretation for estimates of control variables."* (Keele et al. 2020: 12)

NOT DISCUSSED HERE

- DAGs are often employed with more complex modelling procedures (e.g., matching, inverse probability weighting, nonparametric, or semiparametric regression methods)
- But the idea behind DAGs holds for simple OLS models as well
- Thinking about the causal structure of your theoretical model will *always* make a better study!

FREE COURSE FROM

- Causal Diagrams: Draw your assumptions before your conclusions
- Miguel Hernán (Kolokotrones Professor of Biostatistics and Epidemiology Harvard University)
- https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your

LITERATURE

- Elwert (2013). Graphical causal models. In: Handbook of causal analysis for social research. Springer Science & Business Media.
- Keele, Stevenson & Elwert (2020). The causal interpretation of estimated associations in regression models. Political Science Research and Methods, 8(1), 1-13.