Introduction: Causation

William Lowe

Hertie School of Governance

6th September 2020

Causation

Causal inference is the main form of explanation in empirical social science

> *Q: Why?*
> *A: Because.*

Causation

Causal inference is the main form of explanation in empirical social science

> *Q: Why?*
> *A: Because.*

Note: Lots of apparently *non-causal* explanation depends on it too. For example, you may invoke my *principles* or *the law* to explain my action, but if those are not also *causes* of my action, there's no explanation yet.

For our purposes, we'll care about causation when we care about

- → understanding how institutions work
- → evaluating policy impact
- → fairness and discrimination

Causation

One important feature of causal explanation is that it is *contrastive* (Sober, 2020)

> *Reporter: Why do you rob banks?*
> *Willy Sutton: Because that's where the money is.*

Causation

One important feature of causal explanation is that it is *contrastive* (Sober, 2020)

> *Reporter: Why do you rob banks?*
> *Willy Sutton: Because that's where the money is.*

There are at least three possible causal questions here:

- → Why banks rather than post offices?
- → Why robbing rather than working?
- → Why do it yourself rather than hiring a gang?

Each one invokes a different counterfactual and implies a different *causal estimand*.

Causal estimand

The counterfactual contrast you want to estimate

How to think formally about causation

We're going to use three complementary frameworks for thinking systematically about causation

1. Structural equations
2. Graphs
3. Potential outcomes

These correspond to different focuses

1. Nature: The mechanisms a.k.a. 'the Science'
2. Nature's joints: How *variables* relate to one another in these mechanisms
3. Nature's creatures: How *cases* relate to one another

How to think formally about causation

What we get to work with is…*data*: columns of numbers organised into cases and variables, and their associations

- → The joint probability distribution of variables
- → All the conditional distributions
- → All the independence relationships

All of that stuff is either

1. generated by nature according to some causal structure that we'd love to know about, or
2. generated by nature according to some *other* structures that look like noise from this one
3. Random noise

Hint: Unless you are doing quantum physics, you can assume 3 is just 2.

Samples and populations

Usually the data we have is

- → a sample from a population
- → a population, which could have been different
- → a population, which we can only measure imperfectly

(For many statistical purposes these are treated the same, or at least very similarly)

Consequently many of those observed associations will be noisy version of the true ones

But we have causal purposes, which are different, and we will assume we have the population

- → things are plenty hard enough even then

'The Science'

What do we need to assume about the science?

We will assume

- → we can express it all in equations that relate a variable on the left hand side (a 'dependent' variable) to one or more variables on the right hand side (the 'explanatory' variables)
- → that each equation represents a distinct *mechanism* (modularity)
- → that the equations are *complete* up to random noise

We can be fairly vague here because causal inference doesn't care about your subject matter (much)

Representational constraints

But there are some constraints (for this course)

- → We cannot separately things that are logically connected (duh)
- → We cannot model *instantaneous* feedback

So what makes these equations structural?

- → Their coefficients are real causal effects

Graphs

Not every detail of the structural equation is needed to learn about causal relations, e.g. if

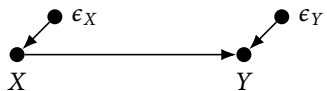
$$Y = \beta_0 + X\beta_X + X^2\beta_X^2 + \epsilon_Y$$

then we can abstract away the functional form and write it as

$$X = \epsilon_X \qquad\qquad Y = f(X, \epsilon_Y)$$

(remember that we can *always* divide Y into $\mathrm{E}[Y \mid X]$ and orthogonal noise ϵ_Y)

This is a recipe to generate $P(X, Y) = P(Y \mid X)P(X)$ that we can draw like this:



Outside factors

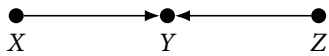
Usually we'll assume there's *always* some ϵ or other representing other external causes not systematically related to the one's we are interested in.

→ So we'll mostly suppress them unless we need it for something.

It's only important that nothing represented ϵ is a common cause of variables we have drawn

Graphs

If $Y = f(X, Z)$ for some complicated f then we'll draw some variant of



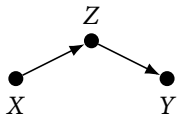
to represent the connection between variables.

→ *Don't* read this as saying X and Z act separately on Y. They may interact in any way.

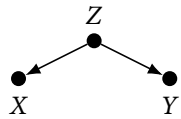
→ *Do* read it as claiming we can imagine intervening on X or Z or Y separately.

Graph elements

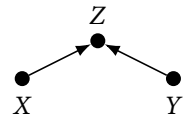
We'll think of graphs as compositions of the following three types of structures



mediator



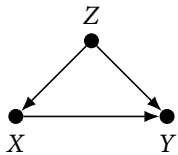
fork, common cause



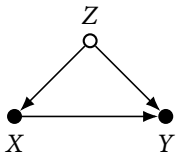
collider, common effect

Graph elements

Some common situations with extra notation:



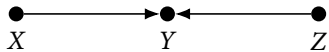
The effect of X on Y is confounded by Z



The effect of X on Y is confounded by Z but we can't measure it

The effect of X on Y is confounded by... something

Implications of graphs



Even this very small graph has observable implications:

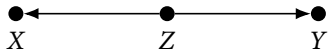
$$\begin{aligned} P(Y, X, Z) &= P(Y \mid X, Z)P(X \mid Z)P(Z) \\ &= P(Y \mid X, Z)P(X)P(Z) \end{aligned}$$

The second line says

- → In the data: Z is independent of X i.e. $P(X \mid Z) = P(X)$ or equivalently $P(X, Z) = P(Z)P(X)$
- → But this is *only* true because of the graph structure

Implications of graphs

For example, in this graph (a fork)



X is *not* independent of Y because they share a common cause Z

The decomposition is now

$$P(Y, X, Z) = P(X \mid Z)P(Y \mid Z)P(Z)$$

Nevertheless, X and Y are *conditionally* independent, given Z:

$$P(X, Y) \neq P(X)P(Y)$$
$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

Notation

This $P(X, Y, Z)$ notation can get cumbersome when all we really want to know about is independencies

Often we'll write

$$X \perp\!\!\!\perp Y$$

when X is independent of Y and

$$X \perp\!\!\!\perp Y \mid Z$$

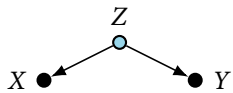
when X is conditionally independent of Y given Z

Conditioning

Counterintuitively, conditioning can both

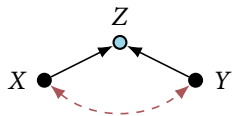
- → make dependent variables independent (mediators and forks)
- → make independent variable dependent (colliders and their children)

Let's condition on Z



$X \not\perp\!\!\!\perp Y$

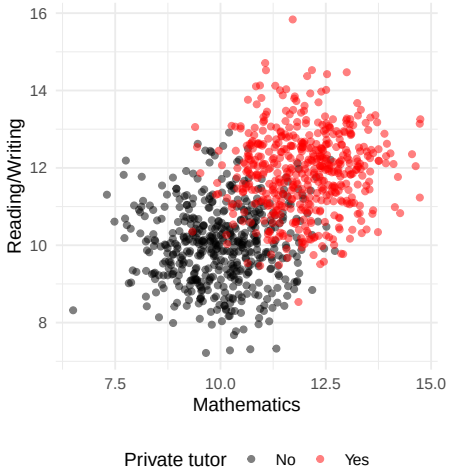
$X \perp\!\!\!\perp Y \mid Z$

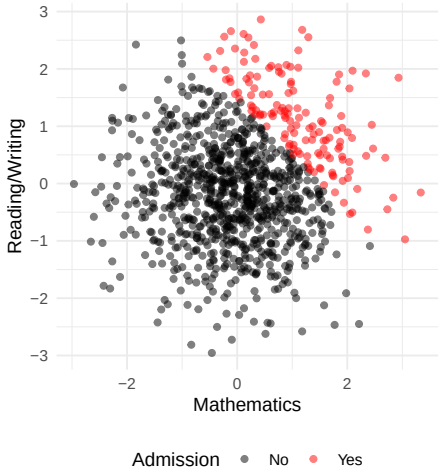


$X \perp\!\!\!\perp Y$

$X \not\perp\!\!\!\perp Y \mid Z$

Conditioning





Using implications about conditioning

Knowing the graph of the equations, not even the equations themselves, provides a guide to independencies we should see in the data

This is important because it means we can work the other way…

Using implications about conditioning

Knowing the graph of the equations, not even the equations themselves, provides a guide to independencies we should see in the data

This is important because it means we can work the other way…

If Y and X are not independent, but are conditionally independent given Z

- → our first graph $Y \longrightarrow Z \longleftarrow X$ cannot be right
- → which means X and Z do not jointly cause Y

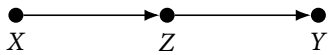
Cool.

- → Causal inference from observational data (plus expert qualitative knowledge)

Limitations: Observational equivalence

Unfortunately there are limits to these sorts of inferences

If X and Y are not independent, but are conditionally independent given Z, this graph (a mediation) is still possible



So we haven't identified the graph (and therefore the causal relationships) purely from data

But we have done so up to *observationally equivalent* graph structures

Limitations: Observational equivalence

Graphs that imply all the same conditional independencies are called Markov equivalent

Markov equivalence

Two graphs are Markov equivalent when they they have the same skeleton (same variables and links) and the same collider structures (Pearl & Verma, 1991).

How to *find* the Markov equivalent set?

Algorithms (Glymour et al., 2019)

- → If everything is observed: Spirtes-Glymour-Scheine (SGS) or the PC algorithm
- → If there are latent variables causal induction (CI) (and FCI, if functional information is available)

A good intuitive description of SGS is ch.22 of Shalizi (2019).

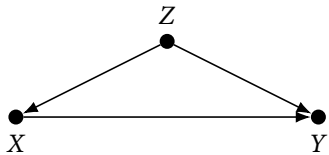
Limitations: Observational equivalence

How to distinguish between the remaining candidates?

1. Domain expertise
2. Experiments! (i.e. interventions)
3. Functional form assumptions, e.g. monotonicity
4. Distributional form assumptions, e.g. non-Normality

We won't get too much into the last two.

Limitations: Faithfulness



could represent these equations (still suppressing ϵ)

$$X = \gamma_0 + Z\gamma_Z$$
$$Y = \beta_0 + X\beta_X + Z\beta_Z$$

Now let

$$\beta_X = -\gamma_Z\beta_Z$$

For all parameter combination like *that*

$$X \perp\!\!\!\perp Z$$

Limitations: Faithfulness

This is an example of *unfaithfulness*.

Unfaithfulness

When there are (conditional) independencies in the data that are not implied by the graph structure

How often might we expect this to happen?

→ In theory: never. These events have probability zero in continuous parameter spaces

→ In finite samples: sometimes, but only by accident

How often might we expect this to *nearly* happen

→ Err, maybe quite a lot of the parameter space is *nearly* unfaithful (Uhler et al., 2013)

Limitations: Faithfulness

All the discovery algorithms require faithfulness to work.

→ It's hard to imagine how they'd work otherwise

Nearly all our social science is going to assume it too.

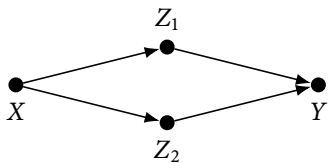
Note: sometimes we create unfaithfulness ourselves, for good reasons, e.g. in matching

Making changes

Usually we are going to be interested in the effect of changing some variable X on another one Y.

In simple cases it's going correspond to just one arrow, e.g. the effect of $X \longrightarrow Y$, or several 'hops', e.g. the effect of X on Y via Z

Or *all* the paths from X to Z if there is more than one way that X affects Z, what is often called the 'total effect', e.g.



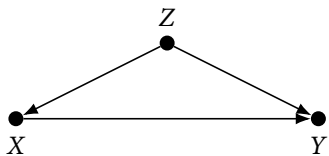
We might even ask about just *one* of the paths, say via Z_1 'holding Z_2 the others constant'

Simple effects

So what actually is the effect of X on Y? Let's take a familiar situation:

$$X_i = \alpha_X + \gamma_Z Z_i$$
$$Y_i = \alpha_Y + \beta_X X_i + \beta_Z Z_i$$

where X is 0 or 1, and we ignore any independent noise, e.g. affecting Y.



In this case the effect of X on Y is β_X. Obviously. But why?

Simple effects

We've just drawn a graph where Z *confounds* the X to Y relationship

It's pretty clear that a simple difference of Y means for $X = 1$ and $X = 0$ (or a regression without Z) would *not* return β_X

→ What it *would* return is given by the *omitted variable formula* you learned about in that one statistics class

What if *we* stepped into this little linear world and adjusted X ourselves, then looked at Y?

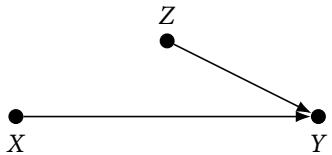
Simple effects

We've just drawn a graph where Z *confounds* the X to Y relationship

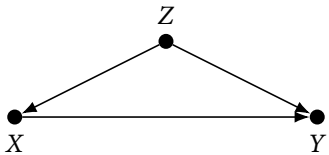
It's pretty clear that a simple difference of Y means for $X = 1$ and $X = 0$ (or a regression without Z) would *not* return β_X

→ What it *would* return is given by the *omitted variable formula* you learned about in that one statistics class

What if *we* stepped into this little linear world and adjusted X ourselves, then looked at Y? It would look like this:

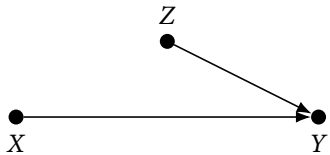


P(Y | X) IN TWO WORLDS



Pre-intervention

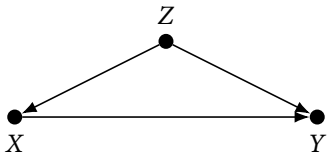
$$\sum_z^Z P(Y \mid X, Z = z)P(X \mid Z = z)P(Z = z)$$



Post-intervention

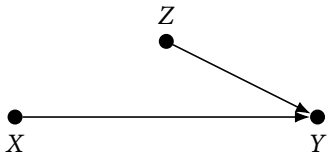
$$\sum_z^Z P(Y \mid X, Z = z)P(Z = z)$$

P(Y | X) IN TWO WORLDS



Pre-intervention

$\sum_z^Z P(Y \mid X, Z = z)P(X \mid Z = z)P(Z = z)$



Post-intervention

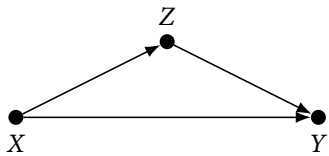
$\sum_z^Z P(Y \mid X, Z = z)P(Z = z)$

The adjustment formula

$P(Y \mid \text{do}(X = x)) = \sum_z^Z P(Y \mid X, Z = z)P(Z = z)$ where Z are parents of X

Identification

But what if the graph were (changing one arrow)



then the post intervention world doesn't look any different.

Since Z doesn't affect X it wouldn't be affected by our intervention. In the this case the adjustment formula says:

$$P(Y \mid \text{do}(X = x)) = P(Y \mid X)$$

No need to do anything with Z

Identification strategies for graphs

Informally, our strategy for identifying a causal effect $X \longrightarrow Y$ is to

- → block (by conditioning) all the paths that add non-causal association
- → not 'open' (by conditioning) any associational 'paths' that are not causal

Formally we should condition on the minimal set of variables that *d-separate* X and Y

d-separation

A set of nodes Z d-separated X from Y if for a node from Z is present on no paths between them that contain colliders or the children of colliders and all paths between them that do.

- → This generalizes the adjustment criterion
- → You can calculate d-separation by hand, or just use DAGitty.

Potential outcomes

Let's take a look at the alternative way to think about causal effects, in terms of potential outcomes

First, consider the causal effect of X on Y for *me*.

The difference between my Y if, say, $X = 1$ and my Y had my X (counterfactually) been 0 instead.

$$\Delta = Y^{X=1} - Y^{X=0}$$

Let's give these potential outcomes (one of which will eventually be the actual outcome while the other remains the counterfactual outcome) some names.

(Annoyingly the naming scheme differs and we can't do much about that.)

Let's try $Y^{X=0}$ for the outcome if my $X = 0$ and $Y^{X=1}$ if it's 1.

Potential outcomes

Remember that $Y^{X=0}$ and $Y^{X=1}$ are 'all the ways things could go with respect to Y', or the 'distribution of possible Ys'

They are related to Y by *consistency*, which here requires that

$$Y = XY^{X=1} + (1 - X)Y^{X=0}$$

More generally

Consistency

if $X = x$ then $Y = Y^{X=x}$

Potential outcomes match outcomes where they contact reality

→ It would be pretty weird if it weren't true

→ but doesn't say anything about what the other outcome should be

Potential outcomes

You might imagine that the equations (and the graph) anchor all the potential outcomes.

That's true, but for some reason many economists like to treat potential outcomes as primitives rather than extensions of the mechanism defined by a graph.

A note on notation: when it's obvious we're thinking of X then we often write the potential outcomes as just Y^1 and Y^0

Or more generally Y^x

Potential outcomes

The tricky thing about potential outcomes is that we can and do treat things like Y^x like regular variables.

For example, we can ask whether

$$Y^0 \perp\!\!\!\perp Z$$

which is the same as asking

> *Would knowing the value of Z help you predict what Y would be if X were set to 0?*

If they are independent, clearly it would not.

Potential outcomes

In practice we nearly always end up wondering about the *whole schedule* of potential outcomes in relation to X, e.g. whether

$$Y^1, Y^0 \perp X$$

so if X were some kind of treatment

> *Would knowing whether you were assigned to treatment or control condition predict how well you would respond to it?*

Why as that interesting?

One very useful implication of this kind of independence is that if you and I have different treatment assignments, we can act as each other's counterfactual.

Which implies that comparing us helps identifies Δ

Potential outcomes

As an aside, it feels a bit weird to say that to say that Y^0 and Y^1 independent of (unpredictable from) X

If X affects Y how could they be?

So it helps to remember that this independence only means that we wouldn't predict *the distribution of your possible responses* to X to be predictable from X.

Causal effects and independence

Individual treatment effects are permanently unobservable

→ Not quite true, but pretty unlikely

but an *average treatment effect* is within reach

$$\begin{aligned} \text{ATE} &= \mathrm{E}[Y^1 - Y^0] \\ &= \mathrm{E}[Y^1] \; - \; \mathrm{E}[Y^0] \end{aligned}$$

provided we can arrange that $Y^1, Y^0 \perp\!\!\!\perp X$.

Reminder: that is because then the $X = 0$ group act as the counterfactual for the $X = 1$ group, and vice versa.

This type of independence is rather rare

→ sometimes we can make it true with randomized experiments

Conditional independence

But we can loosen the requirements to say that

$$Y^1, Y^0 \perp\!\!\!\perp X \mid Z$$

for some Z variables that we think make X=0 and X=1 not comparable

Here we are admitting that sometimes the range of possibilities for Y is different depending on whether your X value *but* that is because of Z.

How?

In the sense that Z both predicts your X value and also what Y you will realize, but that once we *know* what Z you have $Y^1, Y^0 \perp\!\!\!\perp X$ again.

Causal effects

Notice that the effect of X on Y could be quite different for different values of Z, opposite even

So if we want to get the average causal effect of X on Y we

- → Split the data into levels of Z
- → Compute the proportion of cases at that level
- → Compute the causal effect in that level
- → Average the effects, weighted by the level proportions

In practice we usually do with with a regression model and call this 'controlling for Z'.

Conditional effects

Sometimes we also want specific effects, e.g. for a specific subgroup

→ For just cases where $Z = 0$

in which case we can just throw away all the cases with other Z values and examine what's left.

This is an example of an easily forgotten concept:

> *Systematically keeping some observations and ignoring others is a form of conditioning*

This becomes important when we think of missing data, sample selection, censoring, and many unexpectedly related problems.

→ In these cases we will think of there being a selection variable S, such that only if $S_i = 1$ does a case i appear in our data

ATT AND ALL THAT

One particularly policy relevant subgroup effect is the effect of treatment on *the cases that were actually treated*.

That is

$$\begin{aligned} \text{ATT} &= \mathrm{E}[Y^1 - Y^0 \mid X = 1] \\ &= \mathrm{E}[Y^1 \mid X = 1] \; - \; \mathrm{E}[Y^0 \mid X = 1] \end{aligned}$$

This is rather hard to represent distinctly in a graph

→ Graphs represent relationships between variables, but we're talking about the *values* of a single variable

ATT AND ALL THAT

One considerable advantage of potential outcomes is the ability to define very fine-grained causal estimands, e.g.

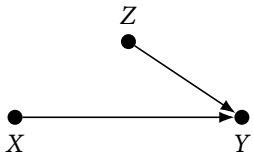
- → The average treatment effect
- → The conditional average treatment effect on $Z = 1$
- → The average treatment effect on the treated
- → The local average treatment effect (in RDD, instrumental variable analysis, surveys with dropout)
- → The direct or indirect effect of X (in mediation problems)
- → The direct effect of X on the treated

and we are just warming up…

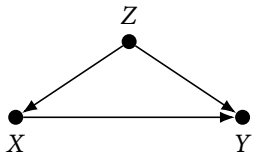
Independence is not enough

We have done the standard exposition of potential outcomes, but there is a problem:

→ How to choose Z?

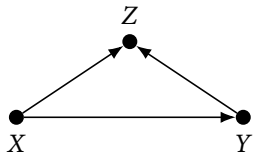


$Y^1, Y^0 \perp\!\!\!\perp X$



$Y^1, Y^0 \not\perp\!\!\!\perp X$

$Y^1, Y^0 \perp\!\!\!\perp X \mid Z$



$Y^1, Y^0 \perp\!\!\!\perp X$

$Y^1, Y^0 \not\perp\!\!\!\perp X \mid Z$

Oops…

Independence is not enough

Turns out, old advice about controlling for confounders is no good either

Old definition: Confounding happens when you fail to condition on variables 'correlated with' (i.e. not independent of) both X and Y

Right? Nope.

Independence is not enough

Turns out, old advice about controlling for confounders is no good either

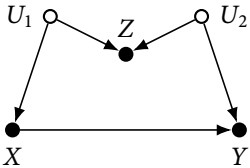
Old definition: Confounding happens when you fail to condition on variables 'correlated with' (i.e. not independent of) both X and Y

Right? Nope.

New definition: Confounding happens when you fail to condition on variables that d-separate X from Y.

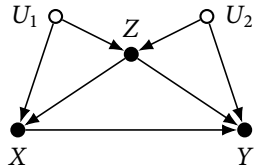
And for that, we've got to have opinions about the graph structure

Example



$$Y^1, Y^0 \perp\!\!\!\perp X$$
$$Y^1, Y^0 \not\perp\!\!\!\perp X \mid Z$$

Conditioning on Z is not even necesssary, and is harmful



$$Y^1, Y^0 \perp\!\!\!\perp X$$
$$Y^1, Y^0 \not\perp\!\!\!\perp X \mid Z$$

Conditioning on Z is necessary, but adding U_1 and/or U_2 will recover independence (see also Ding & Miratrix, 2015; Greenland, 2003)

Summing up

Despite this rather partisan presentation, we'll be using the both algebraic, graph, and potential outcome reasoning in the course

- → Equations when we have stronger theory about detailed mechanism
- → Graphs when we have (or only want to assume) weaker theory about detailed mechanisms
- → Potential outcomes when we want to specify causal estimands precisely
- → or see the implications of assumptions at the level of cases
- → or read the economics literature…

In case you were curious, potential outcome and graph manipulation are formally equivalent.

- → just very different to work with

Just like other languages, sometimes it takes a long time to say what you want, and sometimes you find there's a word for your sentence

- → Looking at you, German.

History: Graphs and equations

Biology, Statistics

→ Wright (1934) introduced path diagrams for genetics, and 'Wright's Rules'

Economics

→ Started early on graphs (Haavelmo, 1943; Strotz & Wold, 1960)

→ had a 'credibility revolution' (Angrist & Pischke, 2010; Leamer, 1983)

→ now leans strongly towards potential outcomes (Imbens, 2019)

Psychology, Sociology

→ Structural Equation Modeling (SEM) built after Cowles Commission (1954)

→ In the 70s by Jöreskog at ETS and Wold at Uppsala University.

→ Still kind of *tense* about causal inference…

History: Graphs and equations

Computer Science

- → In Bayesian expert systems research via probability on graphs, but with little causal focus: (Neapolitan, 1990; Pearl, 1989)
- → In artificial intelligence: Pearl (2000)

Political science

- → A stronghold of the Neyman-Rubin causal model but increasingly graphical: Sekhon, Imai, Green.

Epidemiology

- → Graphs and potential outcomes in equal measure, pioneered by Hernan and Robins.
- → That's why we're reading their book.

References

Angrist, J. D. & Pischke, J.-S. (2010). 'The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics'. *Journal of Economic Perspectives*, *24*(2), 3–30.

Ding, P. & Miratrix, L. W. (2015). 'To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias'. *Journal of Causal Inference*, *3*(1), 41–57.

Glymour, C., Zhang, K. & Spirtes, P. (2019). 'Review of Causal Discovery Methods Based on Graphical Models'. *Frontiers in Genetics*, *10*.

Greenland, S. (2003). 'Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias'. *Epidemiology*, *14*(3), 300–306.

Haavelmo, T. (1943). 'The Statistical Implications of a System of Simultaneous Equations'. *Econometrica*, *11*(1), 1.

References

Imbens, G. (2019, July). *Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics* (w26104). National Bureau of Economic Research. Cambridge, MA.

Leamer, E. E. (1983). 'Let's take the con out of econometrics'. *The American Economic Review*, *73*(1), 31–43.

Neapolitan, R. E. (1990). 'Probabilistic reasoning in expert systems: Theory and algorithms'. Wiley.

Pearl, J. (1989). 'Probabilistic reasoning in intelligent systems: Networks of plausible inference'. Kaufmann
OCLC: 553972645.

Pearl, J. (2000). 'Causality: Models, reasoning, and inference'. Cambridge University Press.

References

Pearl, J. & Verma, T. (1991). Equivalence and synthesis of causal models. In B. D'Ambrosio & P. Smets (Eds.), *UAI '91: Proceedings of the seventh annual conference on uncertainty in artificial intelligence*. Morgan Kaufmann.

Shalizi, C. R. (2019). *Advanced Data Analysis from an Elementary Point of View*.

Sober, E. (2020). 'A theory of contrastive causal explanation and its implications concerning the explanatoriness of deterministic and probabilistic hypotheses'. *European Journal for Philosophy of Science*, *10*(3), 34.

Strotz, R. H. & Wold, H. O. A. (1960). 'Recursive vs. Nonrecursive Systems: An Attempt at Synthesis (Part I of a Triptych on Causal Chain Systems)'. *Econometrica*, *28*(2), 417.

Uhler, C., Raskutti, G., Bühlmann, P. & Yu, B. (2013). 'Geometry of the faithfulness assumption in causal inference'. *The Annals of Statistics*, *41*(2), 436–463.

Wright, S. (1934). 'The method of path coefficients'. *The Annals of Mathematical Statistics*, *5*(3), 161–215.