Week 1: Causal Inference and Directed Acyclic Graphs

Causal Inference & Structural Equation Modeling

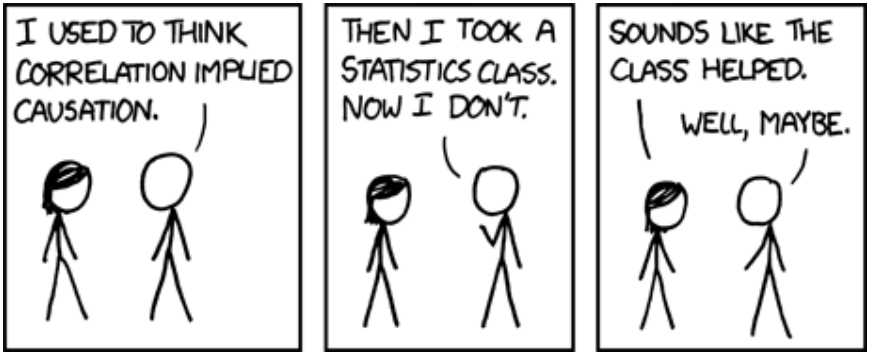
Noémi K. Schuurman
based on slides by Oisín Ryan

February 2022

Overview

- **Causal inference - intro**
- Causal Graphs, DAGs and SCMs
- Statistical dependencies implied by DAG structures
- Causal Discovery

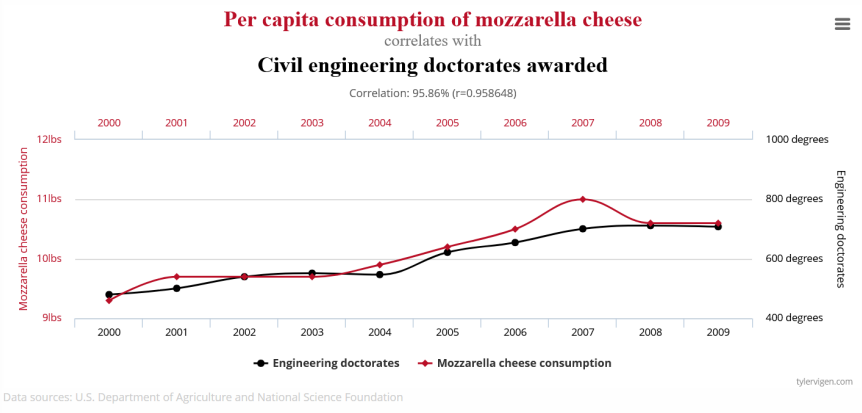
Correlation =/= Causation



https://imgs.xkcd.com/comics/correlation.png

Spurious Assocations

A **spurious association** is a non-causal association.



Check out: http://tylervigen.com/spurious-correlations

Correlation =/= Causation

bewonen. Het is dan ook steviger en duurzamer gebouwd, beter geïsoleerd, functioneler en heeft de uitstraling van een woonhuis – maar dan mini. Hoewel veel tiny houses op wielen staan, zijn het volgens de wet geen woonwagens. In een woonwagen mag je ook niet permanent wonen, terwijl dat bij een tiny house juist het doel is.

9 **Klein wonen (en ontspullen) maakt gelukkiger.** Wie klein gaat wonen, denkt ook vanzelf na over de vraag wat je nou echt nodig hebt en wat belangrijk voor je is. En dat brengt je dichter bij wie je bent en hoe je wilt leven. Je hoeft er niet meteen kleiner voor te gaan wonen, maar onderzoekers hebben wel een link ontdekt tussen mensen die veel spullen in huis hebben en het stresshormoon cortisol. Hoe meer spullen, hoe meer stress – zo bleek. En hoe paradoxaal het ook klinkt: minder woonoppervlakte levert volgens tiny-house-bewoners juist méér ruimte op. Namelijk: ruimte in je hoofd. Zonder de ballast van een groot huis, een hypotheek, al die bezittingen en verplichtingen, houd je meer tijd

Taken from magazine Flow "Het grote boek van minder"

Correlation =/= Causation

Living small scale (and getting rid of your "stuff") makes you happier. Who starts living on a small scale, automatically thinks about what they really need and what is important to them. And that brings you closer to who you are, and how you want to live. You don't have to immediately move to a smaller house, but researchers did find a link between people who have a lot of stuff in their house and the stress-hormone cortisol. The more stuff, the more stress - they found. And although it may sound paradoxical: according to tiny-house-residents less living space actually resulst in móre space. That is: space in your head. Without the burden of a large house, a morgage, all those possesions and obligations, you have more time and energy left to do what makes you happy. Reading, gardening, volunteer work, walking,

Translated from magazine Flow "Het grote boek van minder"

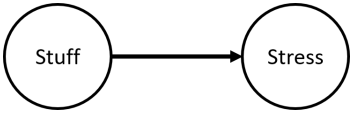
Causal interpretation by Flow:

Correlation =/= Causation

Living small scale (and getting rid of your "stuff") makes you happier. Who starts living on a small scale, automatically thinks about what they really need and what is important to them. And that brings you closer to who you are, and how you want to live. You don't have to immediately move to a smaller house, but researchers did find a link between people who have a lot of stuff in their house and the stress-hormone cortisol. The more stuff, the more stress - they found. And although it may sound paradoxical: according to tiny-house-residents less living space actually resulst in móre space. That is: space in your head. Without the burden of a large house, a morgage, all those possesions and obligations, you have more time and energy left to do what makes you happy. Reading, gardening, volunteer work, walking,

Translated from magazine Flow "Het grote boek van minder"

Causal interpretation by Flow:

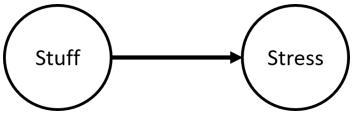


Correlation =/= Causation

Living small scale (and getting rid of your "stuff") makes you happier. Who starts living on a small scale, automatically thinks about what they really need and what is important to them. And that brings you closer to who you are, and how you want to live. You don't have to immediately move to a smaller house, but researchers did find a link between people who have a lot of stuff in their house and the stress-hormone cortisol. The more stuff, the more stress - they found. And although it may sound paradoxical: according to tiny-house-residents less living space actually resulst in móre space. That is: space in your head. Without the burden of a large house, a morgage, all those possesions and obligations, you have more time and energy left to do what makes you happy. Reading, gardening, volunteer work, walking,

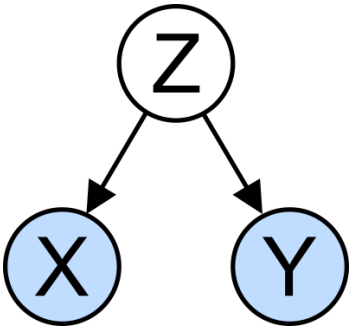
Translated from magazine Flow "Het grote boek van minder"

Causal interpretation by Flow:



Alternative Explanations?

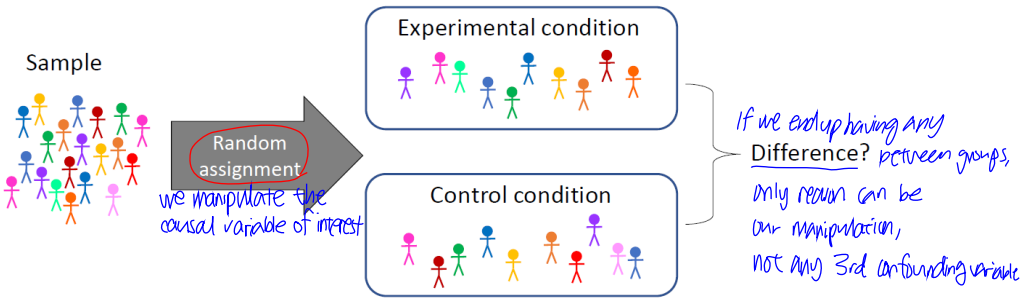
Confounding



Confounder: A variable (Z) that influences both the independent variable (X) and dependent variable (Y), causing a spurious association between them.

Solution a.: Experiments/ "Randomized Controlled Trials" → Gold standard of causal inference

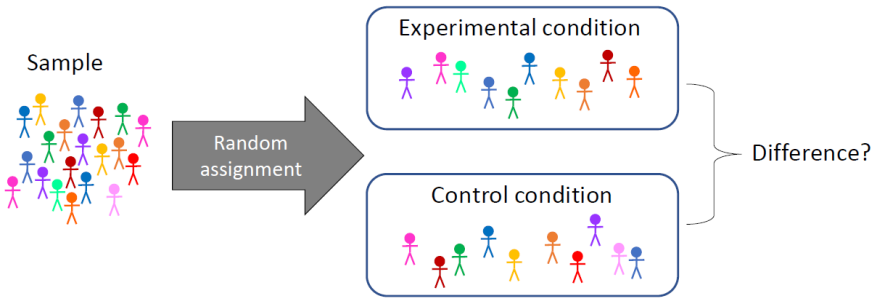
☆ Random assignment ensures that the independent variable is not affected by confounders.



Hence: Difference between groups = effect of the treatment.

Solution a.: Experiments/ "Randomized Controlled Trials"

Random assignment ensures that the independent variable is not affected by confounders.



Hence: Difference between groups = effect of the treatment.

Problems:

- Can go wrong: Drop-out, switching groups, contamination, etc.
- Often infeasible.

Solution b.: Avoid doing causality at all costs.

"I just care about description".

"I just care about prediction".

prediction doesn't tell us what's happening really...

Very often when you care about prediction,
you care about intervention.
& then you should probably care about causal mechanism.

Solution b.: Avoid doing causality at all costs. → It's not a real solution... ☹

"I just care about description".

"I just care about prediction".

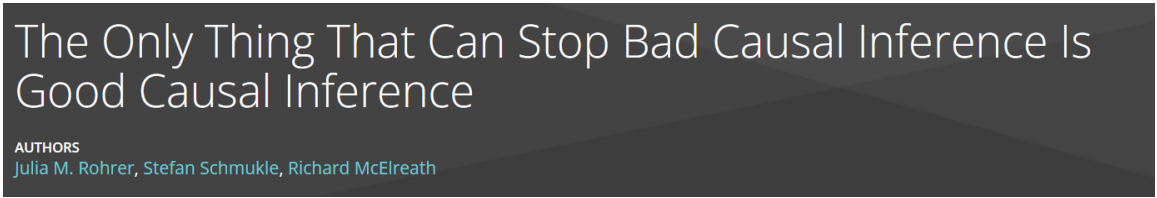
Solution c.: Secretly do causality, hope people won't notice.

Avoid explicitly talking about causality in your research.

- Hernán, M. A. (2018). The C-word: scientific euphemisms do not improve causal inference from observational data. *American journal of public health*.
- Hernán, M. (2018). The C-word: the more we discuss it, the less dirty it sounds.
- Haber, N., Smith, E. R., Moscoe, ... & CLAIMS research team. (2018). Causal language and strength of inference in academic and media articles shared in social media (CLAIMS): A systematic review. *PloS one*.
- Hamaker, E. L., Mulder, J. D., & van IJzendoorn, M. H. (2020). Description, prediction and causation: Methodological challenges of studying child and adolescent development. *Developmental cognitive neuroscience*.

the underlying dynamics between X and Y"
"X would engage increased Y"
"to what extent do genes drive an association between X and Y"
"can associations between X and Y be attributed to variations in"
"X would modulate Y"
"is X able to protect against Y"
"X is affected by Y"
"X may form a target of intervention"
"X can be primed, depending on Y"
"are differences in X accounted for by Y"
"does X elicit Y"
"the spillover from X to Y"
"(...) processes (...)"
"the impact of X on Y"
"the amount of X explained by Y"
"the interplay between X and Y"

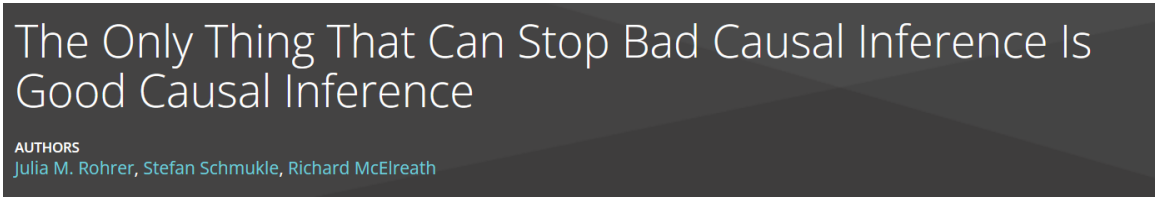
Solution d.: Do causal inference explicitly, in a principled way.!!



Explicitly:

- Be open about your causal interests, research questions.
- Specify your causal questions as clearly as possible: e.g. what kind of treatment/intervention are you interested in exactly.

Solution d.: Do causal inference explicitly, in a principled way.



Explicitly:

- Be open about your causal interests, research questions.
- Specify your causal questions as clearly as possible: e.g. what kind of treatment/intervention are you interested in exactly.

Principled way:

- Be clear about what assumptions you are making for your causal inference.
- Do you inference in a systematic way, for example using a causal inference framework.
- Make use of "triangulation": Do different studies that require different assumptions, to fill the gaps.

application & combination of several research methods in the study of same phenomenon

Causal Inference Frameworks & Techniques

- Structural Causal Models & Causal Graphs (Pearl) - lecture 1
- Potential Outcomes Framework (Rubin) - lecture 2

From wiki "Structural Causal Models - History":
"Sociologists originally called causal models **structural equation modeling**, but once it became a rote method, it lost its utility, leading some practitioners to reject any relationship to causality. Economists adopted the algebraic part of path analysis, calling it simultaneous equation modeling. However, economists still avoided attributing causal meaning to their equations (Pearl, Book of Why)."

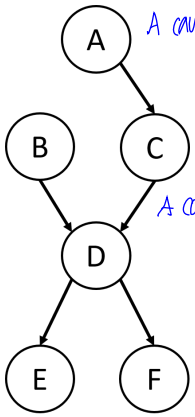
Overview

- Causal inference - intro
- Causal Graphs - DAGS
- Statistical dependencies implied by DAG structures
- Causal Discovery

Causal Graphs

All about

Example:



A causes C directly

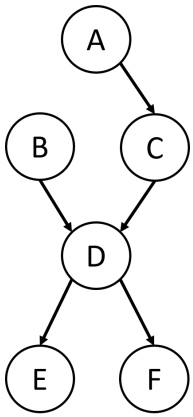
A causes D indirectly

- The circles, or 'nodes' represent variables.
- Arrows (edges) going directly from variable X to Y, represent that X directly causes Y. "Arrows are causal effects in causal graph."
- No arrow, means no causal effect.

$\Rightarrow$ Arrows are causal effect in causal graphs!

Causal Graphs

Example:



- The circles, or 'nodes' represent variables.
- Arrows going directly from variable X to Y, represent that X directly causes Y.
- No arrow, means no causal effect.
- 'Parents' directly cause 'Children'.
- 'Ancestors' directly or indirectly cause 'descendants'.

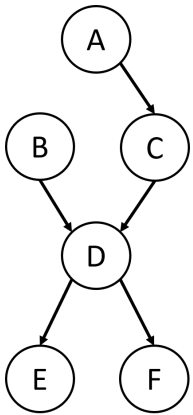
Ex) A is a parent of C / C is a child of A
D is a descendant of A & A is an ancestor of D

Take a "causal effect of X on Y" as: 'if we intervene and change the value of X, then as a result of this the value of Y will change'.

what causal effect actually means.

Directed Acyclical Graphs DAGs

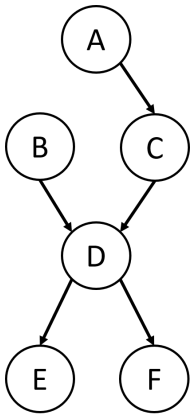
Example:



- Not necessarily causal graphs - but we will use them as causal graphs.

Directed Acyclical Graphs

Example:



The example in this figure is a DAG.

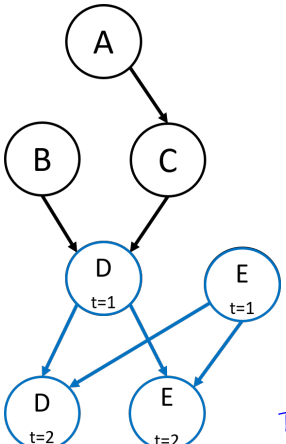
- Not necessarily causal graphs - but we will use them as causal graphs.
- Directed - Only directed edges/arrows allowed!
- Acyclic - no bidirectional effects or 'loops'

$\longleftrightarrow$ (X)

no cycles and such ...

Directed Acyclical Graphs

Example:



- Not necessarily causal graphs - but we will use them as causal graphs.
- Directed - Only directed edges/arrows allowed!
- Acyclic - no bidirectional effects or 'loops'

The trick is to involve time

In this figure you see the loophole to have 'loops' and 'bidirectional' effects in a DAG: Include time-specific variables!

we need to make connection

Tying Causal Directed Acyclical Graphs to Statistical Relationships

From a causal DAG, we can read of causal relationships, but also implied statistical relationships among the variables.

To do this, we use the following condition:

What kind of statistical dependencies does this DAG imply?

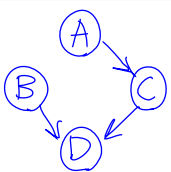
Global Markov Condition: ← we use

Every variable (node) is conditionally independent of its non-descendants, given its parents:

$X \perp\!\!\!\perp \text{non-descendants}(X) | \text{parents}(X)$.

$\perp\!\!\!\perp$: independent

$\not\perp\!\!\!\perp$: dependent



"D is ind. of A given C"

$D \perp\!\!\!\perp A \mid C$

Tying Causal Directed Acyclical Graphs to Statistical Relationships

From a causal DAG, we can read of causal relationships, but also implied statistical relationships among the variables.
To do this, we use the following condition:

Global Markov Condition:

Every variable (node) is conditionally independent of its non-descendants, given its parents:
X $\perp\!\!\!\perp$ non-descendants(X)|parents(X).

This means, to read of statistical dependencies from the graph, we can use markov factorization:

only their parents!

Markov Factorization: given by bunch of conditional dist. where the variables depend on

The joint density of the variables $P(X_1, \dots X_n)$ is given by $\prod_{i=1}^{n} P(X_i \mid Parents(X_i))$

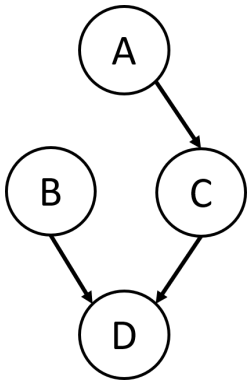
Note: This is based on the chain rule for random variables! Check out the wiki for "Chain rule probability"

Simpler than chain rule.

Markov Factorization Example

To read of statistical dependencies from the graph, we can use markov factorization to make things simpler:

Example:



Lets get the joint distribution of A, B, C, and D.

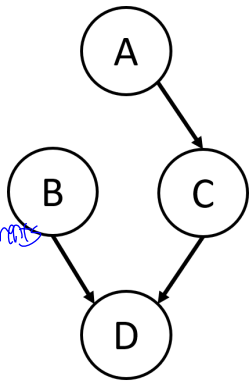
Markov Factorization Example

To read of statistical dependencies from the graph, we can use markov factorization to make things simpler:

Example:

Lets get the joint distribution of A, B, C, and D.

- Chain rule: $P(A, B, C, D) = P(D|A, B, C)P(C|A, B)P(B|A)P(A)$
- Markov Factorization: $\sim$ children being independent of their ancestors given their parents
  $P(A, B, C, D) = P(D|B, C)P(C|A)P(B)P(A)$ ← we can simplify this



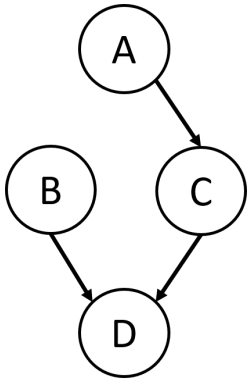
Markov Factorization Example

To read of statistical dependencies from the graph, we can use markov factorization to make things simpler:

Example:

Lets get the joint distribution of A, B, C, and D.

- Chain rule: $P(A, B, C, D) = P(D|A, B, C)P(C|A, B)P(B|A)P(A)$
- Markov Factorization: $P(A, B, C, D) = P(D|B, C)P(C|A)P(B)P(A)$



Markov Factorization based on the DAG makes things simpler.

Structural Causal Model ~ we tie equations to the graph.

Based on a DAG, we know exactly what conditional probability densities we need, to directly calculate the joint probability of two or more variables in the DAG.[1]

Of course, to actually calculate this, we also need to specify the functional forms of those densities in some way. We do this with a Structural Causal Model.

Then we tie equations to the graphs also.

<NOTE>
these all Markov-factorization tricks are basically what Bayesian samplers are using.

[1] This is also super handy in the context of Bayesian statistics...to figure what conditional distributions we need to obtain a joint posterior. In Bayesian stats context, DAGs are called 'Bayesian Networks'.

Structural Causal Models

A Structural Causal Model:
a set of equations describing causal relations between variables, and include noise terms ϵ that are independent of other noise terms and variables. → so those noise terms are not visible in DAGs. bcuz they're not related to anything

Notes:

- We consider variables that are stochastic/probabalistic.
- X causing Y means something like if I intervene and change the value of X, this changes the "probabilities" of the outcomes of Y. so it changes the prob instead of actual outcome of Y
- The noise terms are often not explicitly drawn in causal graphs (but in SEM path models they often are!). This is possible because they are (assumed) independent of all other variables.
- Specifying precise functional forms of the relationships, usually means introducing more assumptions (e.g., linearity)., normality , . . → soon as we go into SCMs, we need to specify those assumptions. (DAGs are just general)

Example: SCM with normal noise terms and linear conditional relationships

⟨SCM⟩

indicates that it's a causal relationship

$$X := \epsilon_X$$
$$Z := 2X + \epsilon_Z$$
$$Y := 1X + 2Z + \epsilon_Y$$

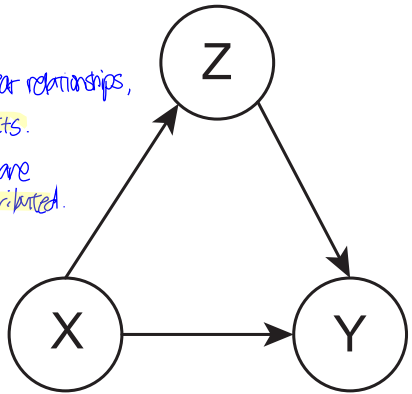
Here we specify linear relationships, linear causal effects. & the noise terms are normally distributed.

where

- $\epsilon_X, \epsilon_Z, \epsilon_Y$ are independently and identically distributed $\sim \mathcal{N}(0, 1)$

You can actually start calculating prob. distributions based on these.

⟨DAG⟩



Example: SCM with normal noise terms and linear conditional relationships

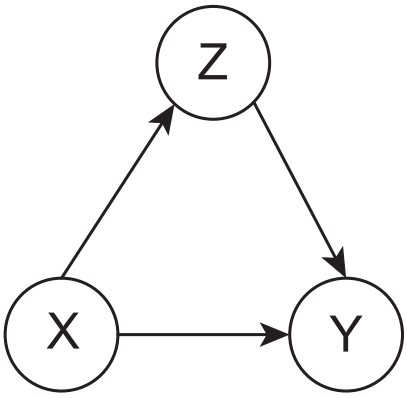
$$
\begin{aligned}
X &:= \epsilon_X \\
Z &:= 2X + \epsilon_Z \\
Y &:= 1X + 2Z + \epsilon_Y
\end{aligned}
$$

where $\epsilon_X, \epsilon_Z, \epsilon_Y$ are independently and identically distributed $\sim \mathcal{N}(0, 1)$

Conditional distributions:

- $X \sim \mathcal{N}(0, 1)$
- $Z|X \sim \mathcal{N}(2X, 1)$
- $Y|Z, X \sim \mathcal{N}(1X + 2Z, 1)$

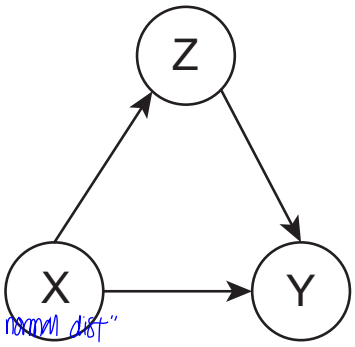
In Jags, you always specify cond. dist. like this.



Example: SCM with normal noise terms and linear conditional relationships

Conditional distributions:

- $X \sim \mathcal{N}(0, 1)$
- $Z|X \sim \mathcal{N}(2X, 1)$
- $Y|Z, X \sim \mathcal{N}(1X + 2Z, 1)$



Use Markov factorization: you get the joint dist by

$P(Y, X, Z) = P(Y|Z, X)P(Z|X)P(X)$

This results in the following multivariate normal:

" multivariate normal dist"

μ cov. matrix

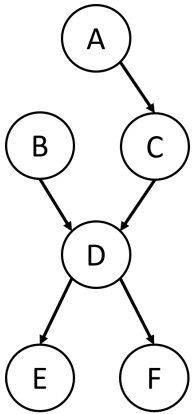
$$\begin{pmatrix} X \\ Z \\ Y \end{pmatrix} = \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 5 \\ 2 & 5 & 12 \\ 5 & 12 & 30 \end{pmatrix} \right]$$

Overview

Now, we are going to use DAGs to more intuitively read of statistical dependencies that are implied by causal structures.

- Causal inference - intro
- Causal Graphs, DAGs and SCMs
- Statistical dependencies implied by DAG structures
- Causal Discovery

Paths between variables in Graphs

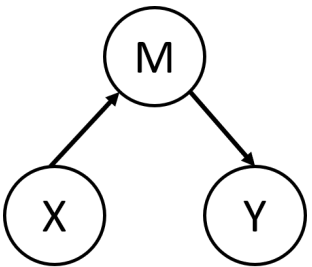


Arrows (any direction) connecting two variables?
→ path between those variables.

"Open paths": Imply statistical dependency between the variables.
"Blocked paths": Imply statistical independency between the variables.
No paths: Implies statistical independency between the variables.

Three Types of Paths in DAGs: 1) Chains

The Chain, aka Mediation.
M is the "mediator".

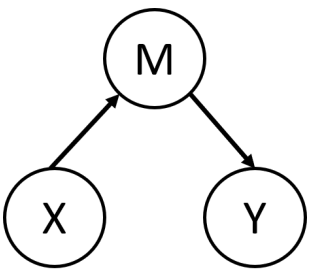


When we talk about causal inference,
typically we're interested in any kind of causal effects,
so it can be total, indirect, direct.

→ Here, X is indirectly causing Y via M (Mediation).

Three Types of Paths in DAGs: 1) Chains

The Chain, aka Mediation.
M is the "mediator".



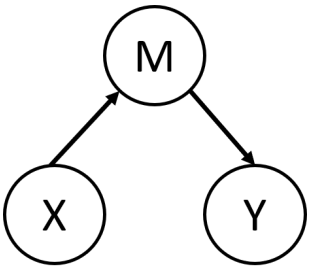
M transmits a causal association of X on Y.

- This results in a marginal association between X and Y: $X \not\perp\!\!\!\perp Y$ marginally X is dependent on Y
- Hence, a chain is an open path (between X and Y).

∵ bcuz there's an association between X & Y.

Three Types of Paths in DAGs: 1) Chains

The Chain, aka Mediation.
M is the "mediator".



M transmits a causal association of X on Y.

- This results in a marginal association between X and Y: $X \not\perp\!\!\!\perp Y$
- Hence, a chain is an open path (between X and Y).

Controlling for M blocks transmission of the causal effect:

- By conditioning on M, X and Y become *independent*: $X \perp\!\!\!\perp Y \mid M$ X&Y will be no longer dependent. no relationship
- We can block the open path between X and Y by conditioning on M.

∵ As soon as we know about M, we don't need X anymore to know about Y. all the info we need is in M.

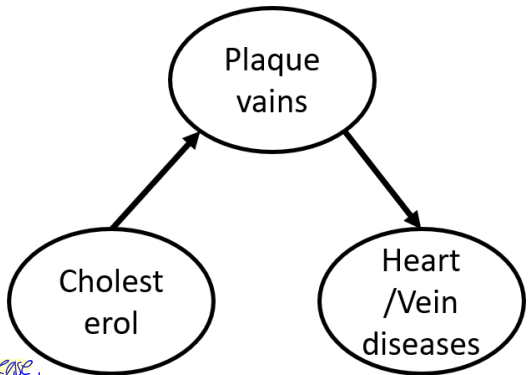
⇒ we have a marginal dependency between X & Y in a chain, but conditioning on M, we no longer have that dependency between X & Y.

3 Type of Paths in DAGs: 1) Chains - Example

The Chain, aka Mediation.
Plaque is the "mediator".

- Cholesterol $\not\perp\!\!\!\perp$ Heart-vein Diseases
- Cholesterol and Heart-vein Diseases are marginally dependent

- Cholesterol $\perp\!\!\!\perp$ Heart-vein Diseases | Plaque
- Cholesterol and Heart-vein Diseases are *in*dependent conditional on the amount of plaque in veins

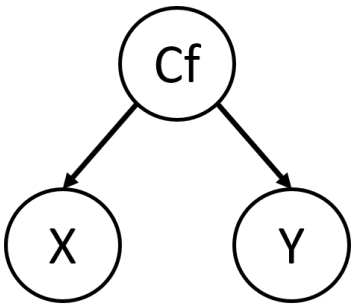


→ If we're interested in causal effect of cholestrol on heart disease, then we should not control for "plaque", bcuz there's an indirect causal effect via plaque!

Three Types of Paths in DAGs: 2) Forks

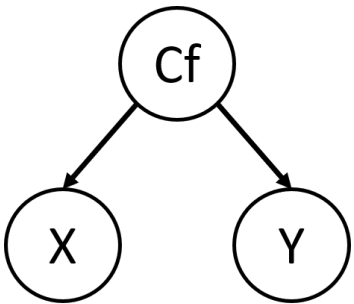
The Fork, aka Confounding.

Cf is the "confounder".



Three Types of Paths in DAGs: 2) Forks

The Fork, aka Confounding.
Cf is the "confounder".

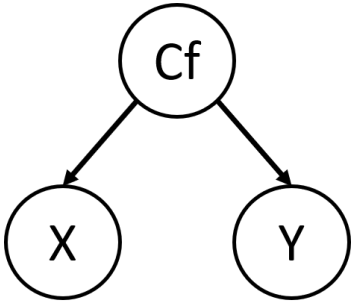


Cf transmits "NON-causal/spurious" association between X and Y:

- This results in a marginal association between X and Y: $X \not\perp\!\!\!\perp Y$
- Hence, a fork is an open path (between X and Y).

Three Types of Paths in DAGs: 2) Forks

The Fork, aka Confounding.
Cf is the "confounder".



Cf transmits NON-causal/spurious association between X and Y:

- This results in a marginal association between X and Y: $X \not\perp\!\!\!\perp Y$
- Hence, a fork is an open path (between X and Y).

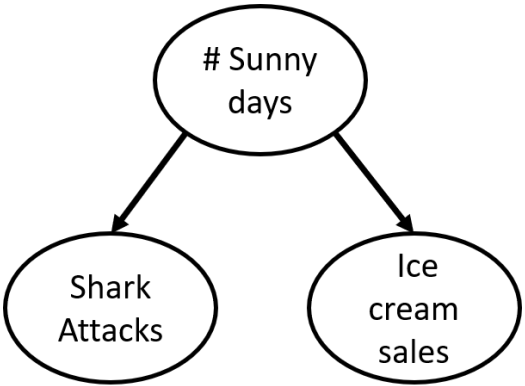
Controlling for Cf blocks transmission of the spurious effect:

- By conditioning on Cf, X and Y become *in*dependent: $X \perp\!\!\!\perp Y \mid Cf$
- We can block the open path between X and Y by conditioning on Cf.

→ In this case, you really want to control for the confounder, bcuz otherwise we get this association that is not causal!

3 Type of Paths in DAGs: 2) Forks - Example

The Fork, aka Confounding.
Number of Sunny Days is the "confounder".



- Shark Attacks $\not\perp\!\!\!\perp$ Ice Cream Sales
- Shark Attacks and Ice Cream Sales are marginally dependent

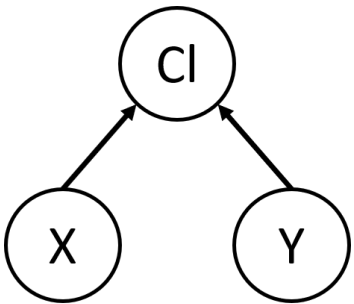
- Shark Attacks $\perp\!\!\!\perp$ Ice Cream Sales | Sunny Days
- The amount of shark attacks and ice cream sales are *in*dependent conditional on the number of sunny days

NOTE ★ Statistical patterns in marginal / conditional dependencies & independencies are the same in Chains and Forks! } 100 % ★★★

Three Types of Paths in DAGs: 3) Inverted Forks

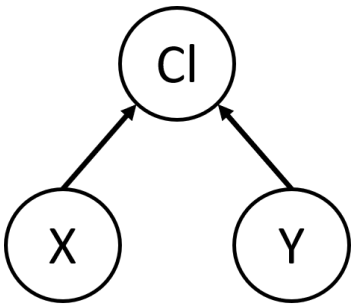
The Inverted Fork, aka Collider structure.

Cl is the "collider".



Three Types of Paths in DAGs: 3) Inverted Forks

The Inverted Fork, aka Collider structure.
CI is the "collider".

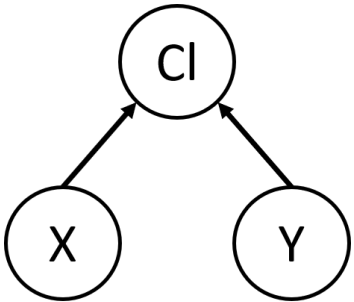


CI does NOT transmit association between X and Y:

- This results in marginal *in*dependence between X and Y: $X \perp\!\!\!\perp Y$
- The collider blocks the path between X and Y.

Three Types of Paths in DAGs: 3) Inverted Forks

The Inverted Fork, aka Collider structure.
Cl is the "collider".



Cl does NOT transmit association between X and Y:

- This results in marginal *in*dependence between X and Y: $X \perp\!\!\!\perp Y$
- The collider blocks the path between X and Y.

Controlling for Cl transmiths a spurious association between X and Y:

- By conditioning on Cl, X and Y become dependent: $X \not\perp\!\!\!\perp Y \mid Cl$
- Conditioning on the collider opens a path between X and Y.

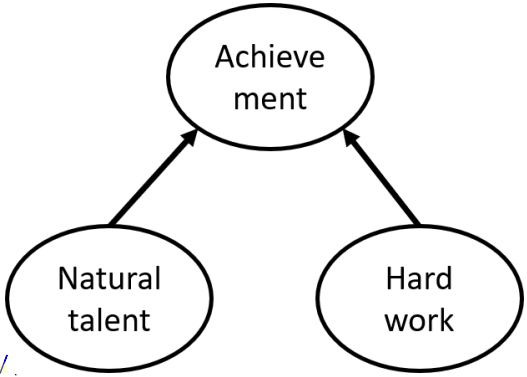
→ Accidentally conditioning on colliders, for ex. only sampling ppl w/ high Ach. can introduce selection bias.

Three Types of Paths in DAGs: 3) Inverted Forks

The Inverted Fork, aka Collider structure.
where Achievement is the "collider".

- Natural Talent $\perp\!\!\!\perp$ Working Hard
- Natural Talent and Working Hard are marginally *in*dependent

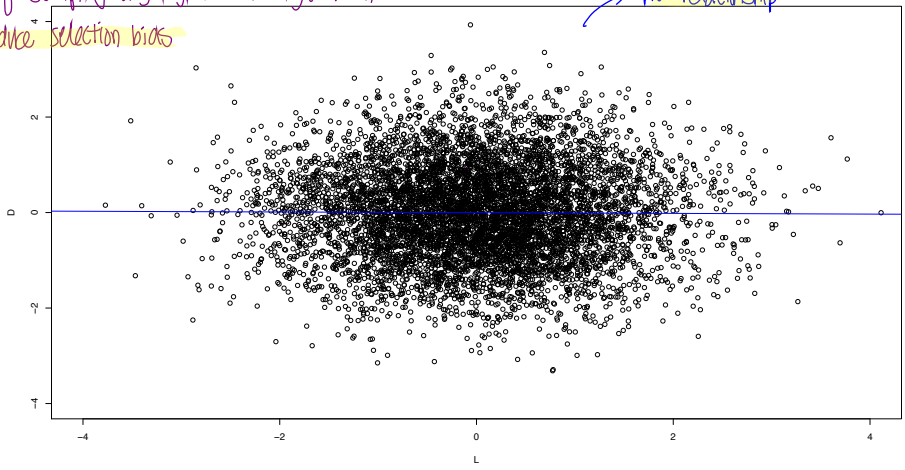
- Natural Talent $\not\perp\!\!\!\perp$ Working Hard | Achievement
- Natural Talent and Working Hard are dependent conditional on the level of Achievement.



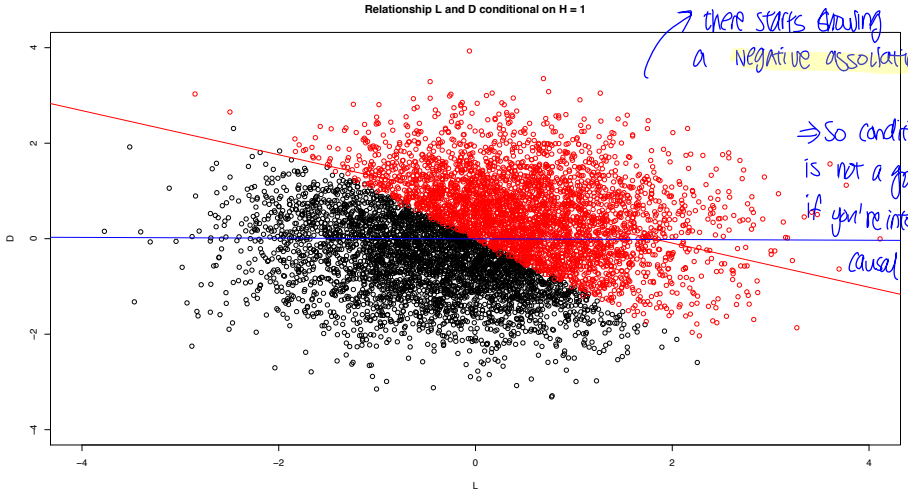
As soon as I know Achievement, it gives info. between N·T & H·W.
EX) If I know (Ach) is low & H·W is high, then I know N·T is low.

Collider Bias: Example

Accidentally conditioning on colliders, for ex. by sampling only high achievers by accident, can introduce selection bias



Collider Bias: Example



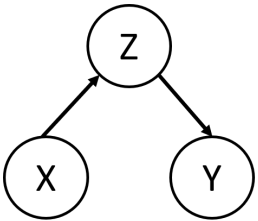
there starts showing a negative association.

⇒ So conditioning on colliders is not a good thing to do if you're interested in causal effects.

In red: only the high achievers. → neg. relationship pops up.

Three types of paths

Chain

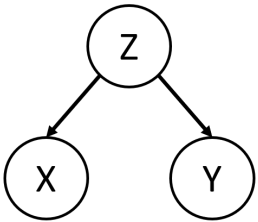


$X \not\perp\!\!\!\perp Y$

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

X: cholesterol
Z: plaque
Y: heart-vein disease

Fork

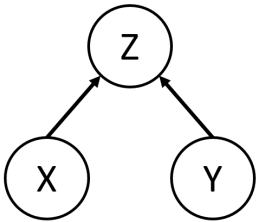


$X \not\perp\!\!\!\perp Y$

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

X: shark attacks
Z: sunny days
Y: ice cream sales

Collider



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

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

X: natural talent
Z: achievement
Y: working hard

* a.

When do causal (in)dependency and statistical (in)dependency align?

= Should we control for our 3rd variable if we are interested in causal effects either indirect / direct?

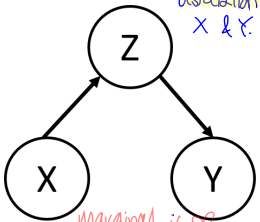
Three types of paths

∴ No, we should not control for Z, cuz otherwise we'll lose the association between X & Y.

∴ Yes, we should. If we do not, then we'll introduce a non-causal relationship between X & Y

∴ NO, we should not! If we do, collider will induce a relationship between X & Y which is NOT causal.

Chain



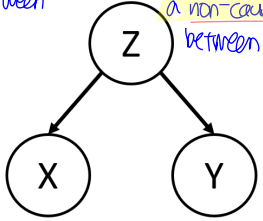
marginal is ♡

$X \not\perp\!\!\!\perp Y$ ♥

$X \perp\!\!\!\perp Y \mid Z$ ↯

X: cholesterol
Z: plaque
Y: heart-vein disease

Fork



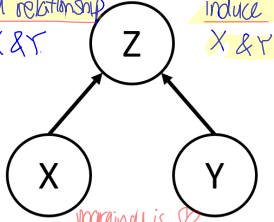
$X \not\perp\!\!\!\perp Y$ ↯

conditional is ♡

$X \perp\!\!\!\perp Y \mid Z$ ♥

X: shark attacks
Z: sunny days
Y: ice cream sales

Collider



marginal is ♡

$X \perp\!\!\!\perp Y$ ♥

$X \not\perp\!\!\!\perp Y \mid Z$ ↯

X: natural talent
Z: achievement
Y: working hard

→When we have chains, marginal relationship is a correct one, while for conditional relationship is the causal one, w/ colliders, marginal

When do causal (in)dependency and statistical (in)dependency align? relationship is the causal one! ☆

D-Separation Rules

Conditional (in)dependence, also for larger graphs, can be read off using *d-seperation rules*

Open path between variables: Variables are 'd-connected'.
No or Blocked path between variables: Variables are 'd-separated'.

D-Separation Rules

Conditional (in)dependence, also for larger graphs, can be read off using *d-seperation rules*

Open path between variables: Variables are ‘d-connected’.
No or Blocked path between variables: Variables are ‘d-separated’.

Marginal (in)dependencies

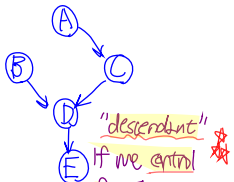
- Chains and Forks are *open paths* $\rightarrow$ marginal dependence $X \not\perp\!\!\!\perp Y$
- Collider structures are *blocked paths*. $\rightarrow$ marginal independence $X \perp\!\!\!\perp Y$

D-Separation Rules

Don't forget about their descendants!

Conditional (in)dependence, also for larger graphs, can be read off using *d-seperation rules*

Open path between variables: Variables are 'd-connected'.
No or Blocked path between variables: Variables are 'd-separated'.



Marginal (in)dependencies

- Chains and Forks are *open paths* $\rightarrow$ marginal dependence $X \not\perp\!\!\!\perp Y$
- Collider structures are *blocked paths.* $\rightarrow$ marginal independence $X \perp\!\!\!\perp Y$

Conditional (in)dependencies

- Conditioning on Mediators and Confounders *or their "descendants" block a path* $\rightarrow$ conditional *in*dependence $X \perp\!\!\!\perp Y|Z$
- Conditioning on Colliders *or their "descendants" open* a path. $\rightarrow$ conditional dependence $X \not\perp\!\!\!\perp Y|Z$

Also need to "consider descendants" of variables when you think about d-separation rules!

d-seperation: Exercise

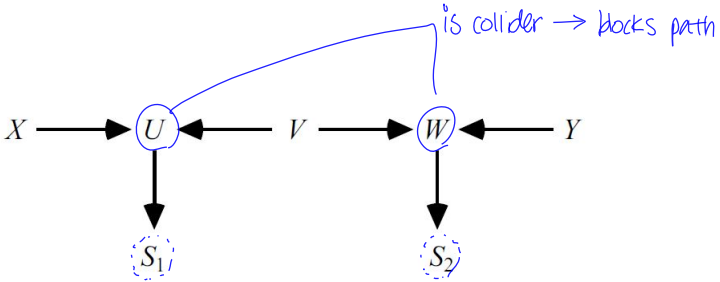


Figure 2.10

- Are X and Y marginally dependent? Nope.. Independent.
- Given which sets of variables are X and Y conditionally dependent?

d-seperation Exercise: Are X and Y marginally dependent?

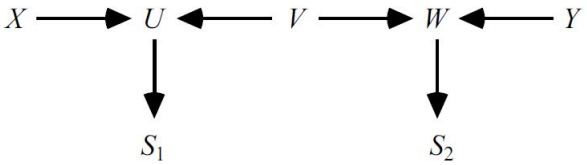


Figure 2.10

When not conditioning, both U and W block the path between X and Y. There are no open paths. → so they're independent.

X and Y are "d-seperated" "given the" "empty set" (given no variables)

means

This implies:

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

X and Y are *marginally* independent.

In this case, if we really wanna know the causal effect of X on Y, then we don't wanna condition on anything, cuz that will open spurious path.

d-seperation exercise: Given which sets of variables are X and Y conditionally dependent?

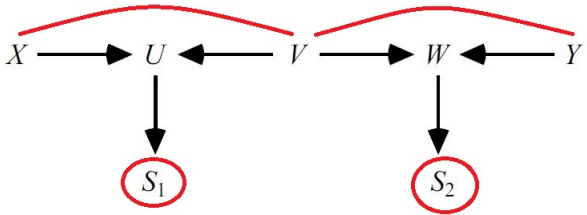


Figure 2.10

Conditioning on S_1 (see figure) or U creates an open path between X and V.
Conditioning on S_2 (see figure) or W creates an open path between V and Y.
So: X and Y are d-connected given a member of the set $\{U, S_1\}$ AND a member of the set $\{W, S_2\}$

For example: $X \not\perp\!\!\!\perp Y \mid \{S_1, S_2\}$; X and Y are dependent conditional on S_1 and S_2

d-seperation example: How can we close the path again?

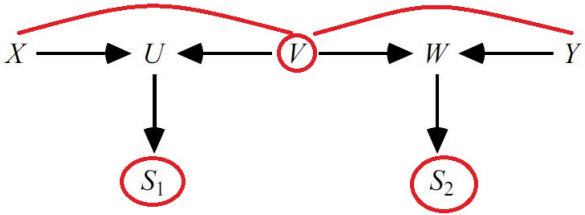


Figure 2.10

Conditioning on V in addition to S_1 or U and S_2 or W closes the path we opened by conditioning on colliders again.
X and Y are d-separated given a member of the set $\{U, S_1\}$ AND a member of the set $\{W, S_2\}$ AND V. ⇝ it becomes independent again

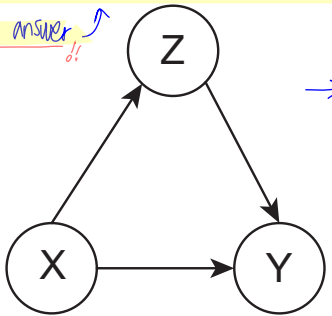
This for example implies: $X \perp\!\!\!\perp Y \mid \{S_1, S_2, V\}$
X and Y are independent conditional on S_1, S_2 and V

Using d-seperation to guide our causal analyses for observational data

I'm interested in the total causal effect of X on Y - both direct and indirect effects.
Interventions are not an option.

"What variables should and shouldn't I control for to obtain the causal effect?"

If I know the (TRUE) DAGs, then I can answer !!



→ In this case, we should NOT control for Z.

But if only interested in direct effect, then should control for Z.

(You need to figure out what effect you're interested in exactly!)

Using d-seperation to guide our causal analyses for observational data

I'm interested in the causal effect of X on Y. What variables should and shouldn't I control for?

basically confounding paths

1) Block all "*backdoor paths*": Paths from $X \to Y$ *that contain an arrow into* X $(...\to X)$
2) Don't open up any new spurious paths - don't condition on colliders *or on their descendants*
3) Leave all the directed paths you care about in-tact

don't control for mediators → over-control bias

Using d-seperation to guide our causal analyses for observational data

I'm interested in the causal effect of X on Y. What variables should and shouldn't I control for?

- Block all *backdoor paths*: Paths from $X \rightarrow Y$ *that contain an arrow into* X $(...\rightarrow X)$
- Don't open up any new spurious paths - don't condition on colliders *or on their descendants*
- Leave all the directed paths you care about in-tact

If controlling for Z gives us the causal effect., we have a set of confounders & if we control for 'em, we get the causal effect

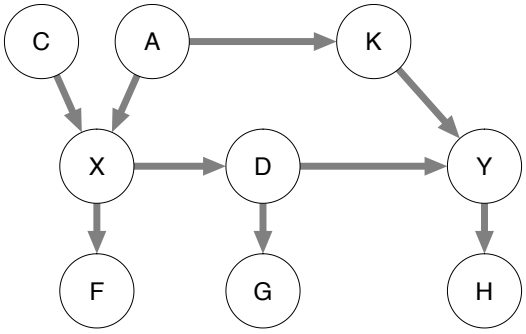
Backdoor criterion

If conditioning on a set of variables $\boldsymbol{Z}$ meets these goals, we say $\boldsymbol{Z}$ fulfills the 'backdoor criterion'. Adjusting for $\boldsymbol{Z}$ yields the causal effect of X on Y.

Valid Adjustment Set → fulfills the backdoor criterion..

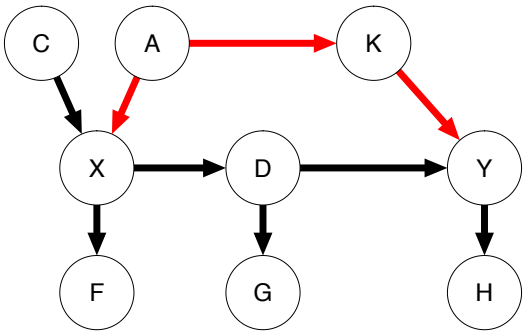
A set of variables $\boldsymbol{Z}$ that by conditioning on them allows us to correctly estimate the effect of X on Y , we call the 'Valid Adjustment Set' for the effect of X on Y.

Using d-seperation to guide our causal analyses for observational data



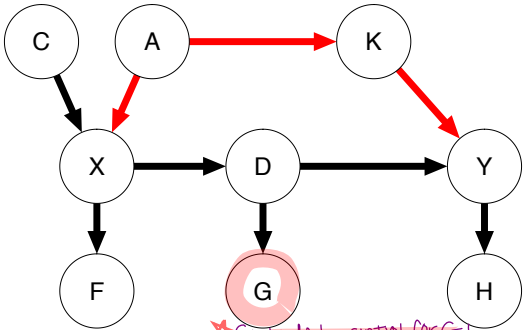
We want to estimate the causal effect of X on Y.

Using d-seperation to guide our causal analyses for observational data



$X \leftarrow A \rightarrow K \rightarrow Y$ is a "backdoor path" from X to Y

DAGs tell us how to estimate causal effects from observational data



Valid Adjustment Sets: $\{A\}$, $\{K\}$, $\{A, K\}$, $\{F, C, K\}$

Overview

But what if we do not know the DAG?!

- Causal inference - intro
- Causal Graphs, DAGs and SCMs
- Statistical dependencies implied by DAG structures
- **Causal Discovery** / causal learning

"Causal Discovery" or "Causal Learning"

Causal Discovery / Causal learning

Can we infer the causal structure from (observational) data?

Short answer:

- In general, **no**
- There is usually more than one SCM and DAG that can generate the same dataset.

Long answer:

- Yes, or at least, we can learn something about the causal structure.
- But only if we are willing to make certain assumptions about the causal system.
- Using "triangulation" - using different methods with different (causal) assumptions, we may learn even more.

Causal Discovery using DAGS and d-separation rules

<Causal Discovery using DAGs & d-separation rule>

Basic Idea:

1. Find all marginal and conditional independence relations present in the data
2. Draw the DAG in which all (and only) those independencies hold up based on the d-seperation rules.

Causal Discovery using DAGS and d-separation rules

Basic Idea:

1. Find all marginal and conditional independence relations present in the data
2. Draw the DAG in which all (and only) those independencies hold up based on the d-seperation rules.

1) causal system we wanna capture is actually a DAG.. If we have loops, & continuous time series, then it may not be able to be captured in DAG anymore,

100 %

Typical Assumptions:

1. *The causal system of interest can be captured in a DAG. → Then we can use d-separation rules
2. *No *unobserved* common causes. (**Sufficiency**) → No unobserved confounders
3. *No conditioning on *unobserved* colliders (no selection bias). we don't want an accidental selection bias..
4. *Faithfulness (tbd later) ~ related to Global Markov condition
5. various statistical assumptions for evaluating statistical dependencies
   e.g. linearity, normal dist. – etc.

Example 1: CI-based discovery

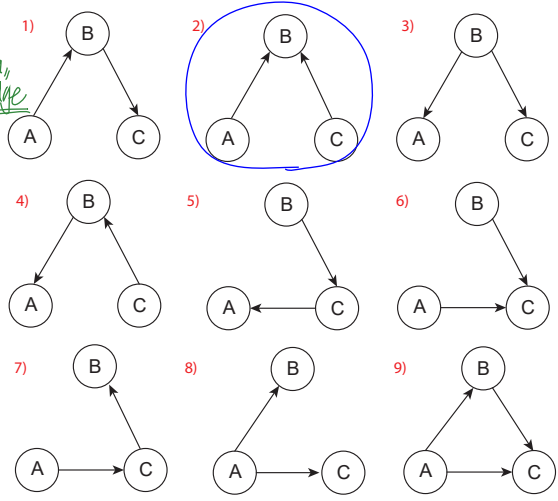
We have three variables in our dataset, A, B and C.

We know of the following (in)dependencies:

- $A \perp\!\!\!\perp C$
- $A \not\perp\!\!\!\perp C \mid B$

All other combinations of variables are dependent (e.g. $A \not\perp\!\!\!\perp B$ and $B \not\perp\!\!\!\perp C \mid A$)

What is the true (data-generating) DAG?



Example 1: CI-based discovery

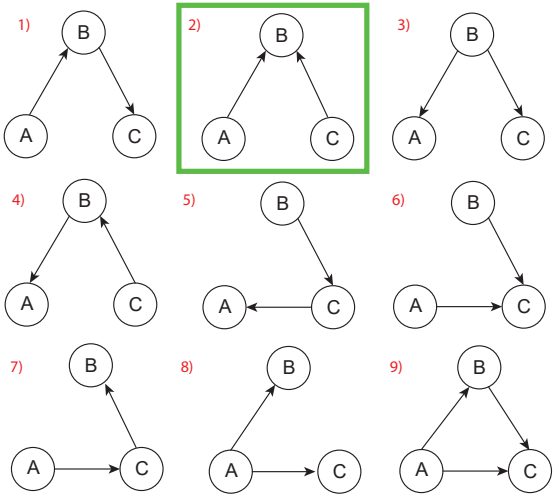
We have three variables in our dataset, A, B and C.

We know of the following (in)dependencies:

$$A \perp\!\!\!\perp C$$
$$A \not\perp\!\!\!\perp C \mid B$$

All other combinations of variables are dependent (e.g. $A \not\perp\!\!\!\perp B$ and $B \not\perp\!\!\!\perp C \mid A$)

What is the true (data-generating) DAG?



Causal Discovery with Conditional Independencies and DAGs

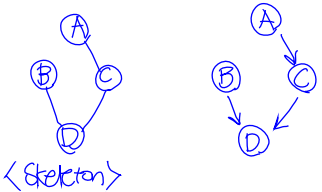
1) all the
Draw undirected edges between variables you are sure should be there.

Principle 1 if they're always always dependent, there has to be an edge,,

Two variables A and B are directly connected in the DAG (either A→B OR B→A) if, and only if, they are dependent conditional on every possible subset of the other variables including empty set

This includes marginal relationships, that is, when you" condition on no other variables".

Get all (in)dependencies and only directly link those variables that are always dependent: the result is the *skeleton* of the DAG.



<skeleton>

Causal Discovery with Conditional Indepencies and DAGs

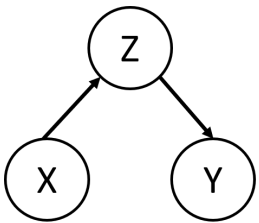
2)

Infer the direction of as many of the undirected edges as possible.

Reminder Intermezzo: Types of Paths

statistical dependency patterns are the same for these two! → cannot statistically distinguish them. as they are statistically equivalent!

Chain

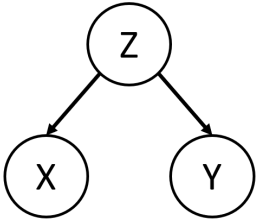


$X \not\perp\!\!\!\perp Y$

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

X: cholesterol
Z: plaque
Y: heart-vein disease

Fork

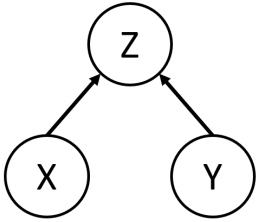


$X \not\perp\!\!\!\perp Y$

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

X: shark attacks
Z: sunny days
Y: ice cream sales

Collider



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

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

X: natural talent
Z: achievement
Y: working hard

* Which of these causal structures can be statistically distinguished from each other? Collider!

Types of Paths

This is what we'll use :)

Which of these causal structures can and cannot be statistically distinguished from each other?

Which have different conditional (in)dependence relations?

- Chains and forks are statistically equivalent!
- This applies to chains in either direction: $X \to Z \to Y$ is equivalent to $X \leftarrow Z \leftarrow Y$.
- Colliders are distinct from chains and forks.

Causal Discovery with Conditional Indepencies and DAGs

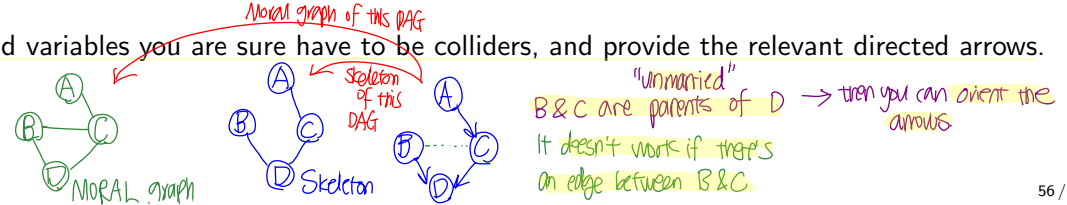
Infer the direction of as many of the undirected edges as possible.

Principle 2

If our skeleton contains a triplet X - Z - Y, where X and Y are marginally independent, we can orientate the arrows as $X \rightarrow Z \leftarrow Y$ if and only if X and Y are dependent conditional on every set of variables containing Z.

This works for 'immoral' colliders, where parents have a common child but are 'unmarried' - don't directly cause each other.

So, find variables you are sure have to be colliders, and provide the relevant directed arrows.



Example 2: CI-based discovery

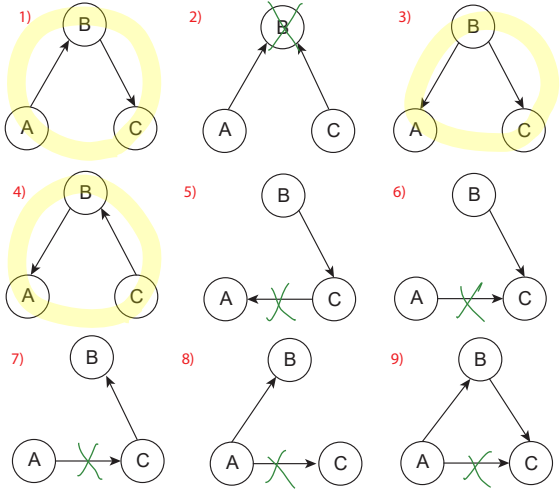
We have three variables in our dataset, A, B and C.

We know of the following (in)dependencies:

$$A \perp\!\!\!\perp C \mid B$$

All other combinations of variables are dependent

What is the data-generating DAG?



Example 2: CI-based discovery

We have three variables in our dataset, A, B and C.

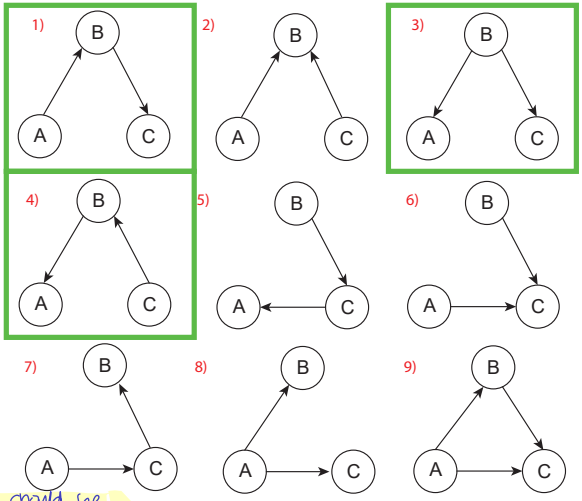
We know of the following (in)dependencies:

$$A \perp\!\!\!\perp C \mid B$$

All other combinations of variables are dependent

What is the data-generating DAG?

→ Now you already see there're 3 possible solutions
& Based on statistics alone, we cannot decide which one it should be.



Markov Equivalence

Typically CI-based methods find more than one DAGs that match the observational data.
The set of possible DAGs is called the *Markov Equivalence set*

more than one DAG that matches our observed dataset.

Markov Equivalence:

Two DAGs are *Markov Equivalent* if they satisfy the same d-separation statements, that is, the same set of (conditional) (in)dependence relations.

Markov Equivalence

Typically CI-based methods find more than one DAGs that match the observational data. The set of possible DAGs is called the *Markov Equivalence set*

Markov Equivalence:

Two DAGs are *Markov Equivalent* if they satisfy the same d-separation statements, that is, the same set of (conditional) (in)dependence relations.

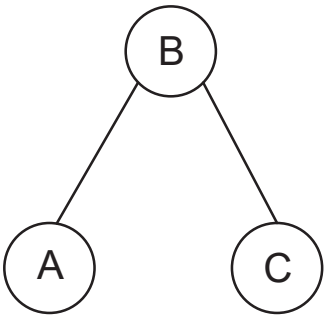
The Markov-Equivalance set is sometimes simply represented by a 'Completed Partially-Oriented Directed Acyclic Graph' **(CPDAG)**

a single graph

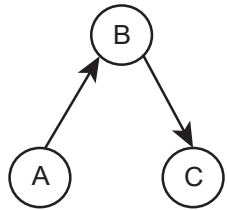
(by leaving out directed arrows)

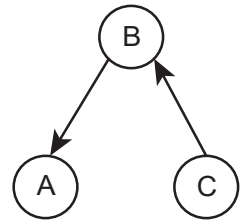
"colliders will have the arrows"

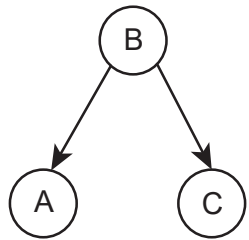
CPDAG



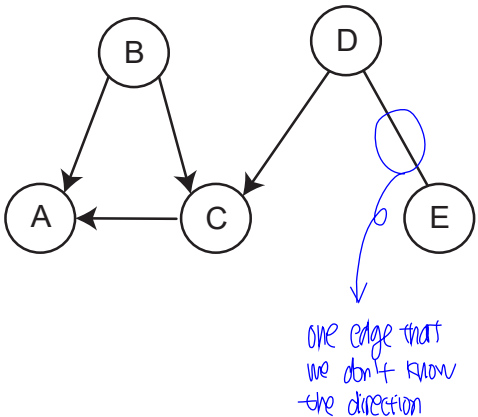
Marov-Equivalence Set



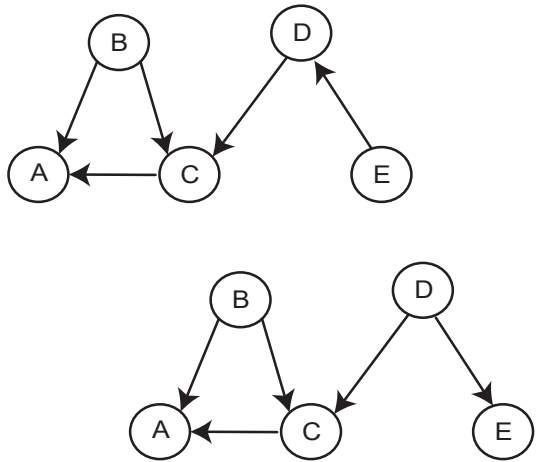




CPDAG



Marov-Equivalence Set



In Practice

PC algorithm; FCI algorithm (Spirtes et al. 2000)

- Do a quicker search without having to test all (in)dependencies
- Various extensions exist that deal with violations of *sufficiency*.
- Still rely on "faithfulness" assumption

all of them rely on ↑

Important Assumptions for CI-based discovery

For (sets of) variables X, Y, and Z:

Global Markov Condition: then statistical relationship follows. !!

If X and Y are d-separated by Z then $X \perp\!\!\!\perp Y \mid Z$: that's connection between the DAG & statistic

"The d-separation rules are appropriate" (which are the case if the causal structure is a DAG).

▷ d-separations in DAG also corresponds to conditional independence relations

Important Assumptions for CI-based discovery

For (sets of) variables X, Y, and Z:

Global Markov Condition:

If X and Y are d-separated by Z then $X \perp\!\!\!\perp Y \mid Z$

"The d-separation rules are appropriate" (which is the case if the causal structure is a DAG).

Faithfulness:

If $X \perp\!\!\!\perp Y \mid Z$ then X and Y are d-separated by Z

"The other way around works as well" - not always true even if the structure is a DAG.

Important Assumptions for CI-based discovery

For (sets of) variables X, Y, and Z:

Global Markov Condition:

If X and Y are d-separated by Z then $X \perp\!\!\!\perp Y \mid Z$

"The d-separation rules are appropriate" (which is the case if the causal structure is a DAG).

Faithfulness:

If $X \perp\!\!\!\perp Y \mid Z$ then X and Y are d-separated by Z

"The other way around works as well" - not always true even if the structure is a DAG.

Also: Various statistical assumptions, e.g., those we need to estimate and decide whether two variables can be considered (in)dependent.

Example of Violations of Faithfulness

$$
\begin{aligned}
A &:= \epsilon_A \\
B &:= .5A + \epsilon_B \\
C &:= -.25A + .5B + \epsilon_C
\end{aligned}
$$

A causes B
A & B cause C

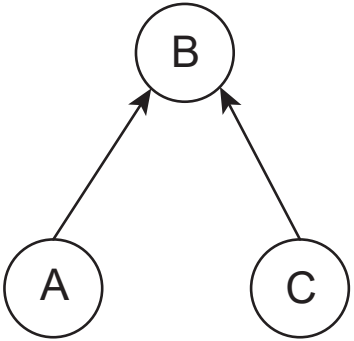
But if we look at the marginal relationship some of the edges exactly cancel out, which make it seem like they are not causally related. !!

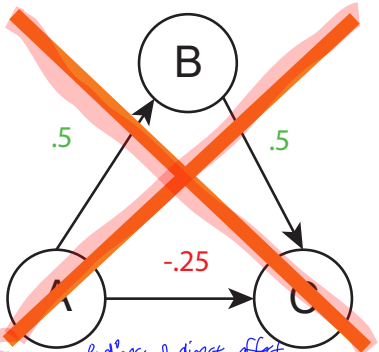
where

- $\epsilon_A, \epsilon_B, \epsilon_C$ are iid, $\sim \mathcal{N}(0, 1)$

$$
\begin{pmatrix} A \\ B \\ C \end{pmatrix} = \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & .5 & 0 \\ .5 & 1.25 & .5 \\ 0 & .5 & 1.25 \end{pmatrix} \right]
$$

Violations of Faithfulness





At least, in the population level, this assumption is fairly okay to make. Becuz the prob. that those values are exactly identical, is zero! it might be 0.001, but that's not the violation, since there's still dependency. However, this becomes problematic soon as we tie this to

$$A \perp\!\!\!\perp C$$
$$A \not\perp\!\!\!\perp C \mid B$$

Assume Faithfulness

indirect & direct effect summed together to 0zero!

↓

seem A & C are marginally independent, when we don't control for B when in fact, they're very causally dependent!!! & That is not sth.

Assumptions for CI-based discovery

Global Markov Condition:

$$X \text{ and } Y \text{ are d-seperated by } Z \implies X \perp\!\!\!\perp Y \mid Z$$

Faithfulness:

$$X \perp\!\!\!\perp Y \mid Z \implies X \text{ and } Y \text{ are d-seperated by } Z$$

Essentially: Paths never "perfectly cancel out"

Statistical (conditional) Independence $\underset{\text{follows}}{\implies}$ causal independence (d-seperation)

In Practice

PC algorithm; FCI algorithm (Spirtes et al. 2000)

- Do a quicker search without having to test all (in)dependencies
- Various extensions exist that deal with violations of *sufficiency*.
- Still rely on faithfulness.

Disadvantages:

- Population (in)dependencies are estimated: uses sample data + statistical tests
  - All of statistics is relevant here, e.g., sample size considerations
  - In a given sample : Type I and II errors
  - Faithfulness does NOT mean there are no false negatives!

all of the problems w/ estimation

In Practice

PC algorithm; FCI algorithm (Spirtes et al. 2000)

- Do a quicker search without having to test all (in)dependencies
- Various extensions exist that deal with violations of *sufficiency*.
- Still rely on faithfulness.

you have to have good conditional independence tests, for ex: t-test, chi-square test...

Disadvantages:

- Population (in)dependencies are estimated: uses sample data + statistical tests
  - All of statistics is relevant here, e.g., sample size considerations
  - In a given sample : Type I and II errors
  - Faithfulness does NOT mean there are no false negatives!
- CI testing easy if linear + Normal (partial correlation / regression) or discrete

  BUT Can be difficult in other cases (Shah & Peters, 2020) ex) non-linear relationship,
  - Non-parametric methods - difficult and requires large sample size

In Conclusion...

DAG/SCM causal modeling allows us *in theory* to make causal statements from observational data!

- Rests on our beliefs/assumptions the causal structure (e.g., the DAG being correct).
- And as soon as we estimate things (including DAGs), many more (statistical) assumptions pop up.
- Often we may not be able to evaluate these assumptions.
- BUT... by being explicit about our assumptions, we can openly debate and learn about them, and about our causal inferences of interest.

In Conclusion...

DAG/SCM causal modeling allows us *in theory* to make causal statements from observational data

- Rests on our beliefs/assumptions the causal structure (e.g., the DAG being correct).
- And as soon as we estimate things (including DAGs), many more (statistical) assumptions pop up.
- Often we may not be able to evaluate these assumptions.
- BUT... by being explicit about our assumptions, we can openly debate and learn about them, and about our causal inferences of interest.

- Next Week: Practicing all of the above in an R lab.
- Next Week: More info on the assignment(s).
- Next Lecture: Rubin's causal inference framework - very explicit about assumptions needed for estimating causal effects - and based on ideas related to missing data.

Week 3: Rubin's Causal Model & Controlling for Confounders

Causal Inference & Structural Equation Modeling

Noémi K. Schuurman
based on slides by Ellen Hamaker

February 2022

Overview

- **Potential Outcomes & Causal Effects**
- Estimating Causal Effects: Assumptions
- Estimating Causal Effects: Controlling for Confounders

Causal Inference Frameworks

In this course, we discuss two frameworks of causality:

- **Structural causal model & DAGS** by Pearl: Previous weeks
- **Potential outcomes framework** by Rubin (also Imbens): Now
- After: Apply both

his perspective: "causal inference is missing data problem"

Judea Pearl and Don Rubin

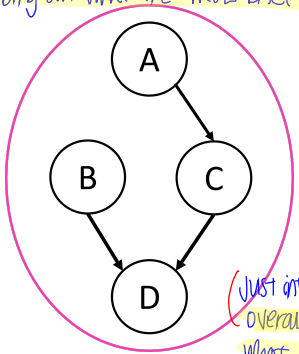
Causal Inference Goals

Today is more about CI, trying to identify specific causal effects

Causal Discovery or **Causal 'Learning'**

- Figuring out the causal structure among a bunch of variables - what are confounders, colliders, mediators?

Basically finding out what the TRUE DAG is.

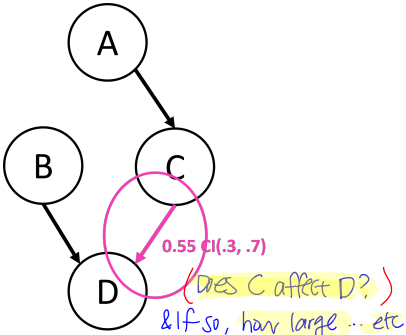


Just interested in the overall structure.

What directions of those arrows going in

Causal Identification

- Answering a specific, well-defined, causal question
- Estimating a specific, well-defined, causal effect



Does C affect D?

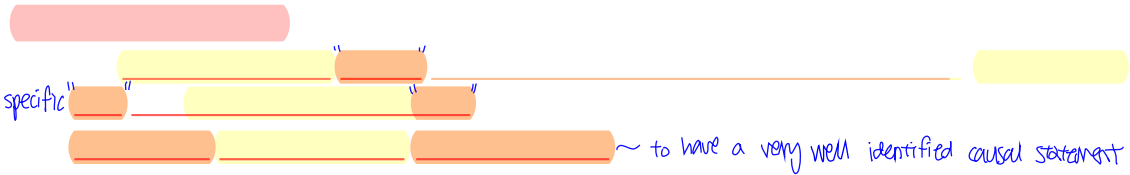
& if so, how large ... etc

Defining Causal Effects: Because Statements vs. Causal Statements

You need to be very careful w/ how you specify / define your causal questions

Examples of **because statements**: → considered as "incomplete" statement as a causal statement.

- My headache went away, because I took an aspirin.
- He is late, because he overslept.
- Hillary Clinton lost the election, because she is a woman.



↳ These are what's missing in "because ~ statements" above.

Defining Causal Effects: Because Statements vs. Causal Statements

"Causal intervention you are doing should be sth. that you can imagine if there no restrictions (e.g. ethical). Imagine a very specific experiment that you do, then you have a well-defined causal question."
(You should be able to imagine)

Examples of **because statements**:

- My headache went away, because I took an aspirin.
- He is late, because he overslept.
- Hillary Clinton lost the election, because she is a woman.

Causal statements should be based on:

- the outcome of some **action** (e.g.,intervention, manipulation, treatment), applied to a unit at a particular point in time
- **relative to** the outcome of **another action**

When the alternative action is not well described, the causal question is not well defined.

Defining Causal Effects: Potential Outcomes

Suppose Don has a headache. A well-defined causal question requires us to define:

- **treatment levels** (aka exposure): Aspirin ($X = 1$) and No Aspirin ($X = 0$) dichotomous causal variable, X
- **outcome**: Headache under both actions one hour later ($Y^{X=1}$ and $Y^{X=0}$)

2 potential outcomes

"specification of time" is impo!

Saying one week ≠ one hour would be very diff!
So it matters a lot potentially.

Defining Causal Effects: Potential Outcomes

Suppose Don has a headache. A well-defined causal question requires us to define:

- **treatment levels** (aka exposure): Aspirin ($X = 1$) and No Aspirin ($X = 0$)
- **outcome**: Headache under both actions one hour later ($Y^{X=1}$ and $Y^{X=0}$)

Suppose Don has these **two potential outcomes**:

- Potential outcome $Y^{X=1} = 0$ (i.e., no headache 1h after aspirin)
- Potential outcome $Y^{X=0} = 1$ (i.e., headache 1h after no aspirin)
- Causal effect: $Y^1 - Y^0 = 0 - 1$: reduction (i.e., improvement) due to Aspirin

in potential outcome framework: diff. between these two potential outcomes!

"Here you see a very clear definition of what a causal effect is"

problem is: we cannot observe both potential outcomes.

Defining Causal Effects: Potential Outcomes

Suppose Don has a headache. A well-defined causal question requires us to define:

- **treatment levels** (aka exposure): Aspirin ($X = 1$) and No Aspirin ($X = 0$)
- **outcome**: Headache under both actions one hour later ($Y^{X=1}$ and $Y^{X=0}$)

☆ NOTATION

Suppose Don has these **two potential outcomes**:

- Potential outcome $Y^{X=1} = 0$ (i.e., no headache 1h after aspirin)
- Potential outcome $Y^{X=0} = 1$ (i.e., headache 1h after no aspirin)
- Causal effect: $Y^1 - Y^0 = 0 - 1$: reduction (i.e., improvement) due to Aspirin

Either we give Don the aspirin, **or** we don't:

- the potential outcome we observe is the **fact**
- the potential outcome we do NOT observe is the **counterfact**

Defining Causal Effects: Potential Outcomes

Two key features to notice here are:

- the causal effect is defined at the **level of the unit** : level of the individual person (unit)
- we can only **observe one potential outcome per unit**

Defining Causal Effects: Potential Outcomes

Two key features to notice here are:

- the causal effect is defined at the **level of the unit**
- we can only **observe one potential outcome per unit**

The latter is known as: "**The fundamental problem of causal inference**" (p.947, Holland, 1986).

Defining Causal Effects: Potential Outcomes

Two key features to notice here are:

- the causal effect is defined at the **level of the unit**
- we can only **observe one potential outcome per unit**

The latter is known as: "**The fundamental problem of causal inference**" (p.947, Holland, 1986).

* Note: measuring the **same person at different occasions** (under different treatments), is NOT necessarily the solution; that **requires additional assumptions** (cf. Holland, 1986):

- temporal stability: effect of treatment does not depend on time : does not change over time
- causal transience: there is no lingering effect from the earlier treatment : earlier treatment does not change the effect of later treatments

Defining Causal Effects: Individual vs. Average Causal Effect

Causality is defined as the **difference in potential outcomes of an individual**:

Individual causal effect: In that unit, his/her (a particular person) causal effect ⇒ means that other units may have diff. causal effects.

$$ICE_i = Y_i^1 - Y_i^0$$

i: other unit may have different causal effect.

Defining Causal Effects: Individual vs. Average Causal Effect

Causality is defined as the **difference in potential outcomes of an individual**:

Individual causal effect:

$$ICE_i = Y_i^1 - Y_i^0$$

The ICE **may differ** for different individuals.

Defining Causal Effects: Individual vs. Average Causal Effect

Causality is defined as the **difference in potential outcomes of an individual**:

Individual causal effect:

$$ICE_i = Y_i^1 - Y_i^0$$

The ICE **may differ** for different individuals. We can't observe both potential outcomes, so we often focus on the **average causal effect** instead:

Average causal effect:

$$ACE = E[Y_i^1 - Y_i^0] = E[Y_i^1] - E[Y_i^0]$$

= difference between expected value of potential outcomes over diff ppl/units

* If you're interested in helping a specific individual, ACE may not be so useful, if that person is far from average.
* This is also what we did last week bb; estimating ACE via a regression model: "what happens to my expected value if change program"

Defining Causal Effects: Individual vs. Average Causal Effect

Causality is defined as the **difference in potential outcomes of an individual**:

Individual causal effect:

$$ICE_i = Y_i^1 - Y_i^0$$

The ICE **may differ** for different individuals. We can't observe both potential outcomes, so we often focus on the **average causal effect** instead:

Average causal effect:

$$ACE = E[Y_i^1 - Y_i^0] = E[Y_i^1] - E[Y_i^0]$$

(*Also referred to as individual treatment effect (ITE) and average treatment effect (ATE).)

Example: Individual vs. Average Causal Effect

diff-word for cause here

Treatment: aspirin ($X = 1$) or no aspirin ($X = 0$)
Outcome: headache ($Y = 1$) or no headache ($Y = 0$)

Example: Individual vs. Average Causal Effect

Treatment: aspirin ($X = 1$) or no aspirin ($X = 0$)
Outcome: headache ($Y = 1$) or no headache ($Y = 0$)

for each person, calculate

we have diff. units	Potential outcomes Y_i^1	Y_i^0	ICE $Y_i^1 - Y_i^0$
Charles	1	1	0
Susan	0	1	-1
Tracy	1	1	0
Ken	1	1	0
Pete	0	0	0
Helen	0	1	-1
Kate	0	0	0
George	0	1	-1

anyways.
For Charles, it doesn't matter. He'll get headache
For Susan, ICE = -1, indicating that taking aspirin helps.

we see that it differs for diff. persons!

Example: Individual vs. Average Causal Effect

Treatment: aspirin ($X = 1$) or no aspirin ($X = 0$)
Outcome: headache ($Y = 1$) or no headache ($Y = 0$)

	Potential outcomes		ICE
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$
Charles	1	1	0
Susan	0	1	-1
Tracy	1	1	0
Ken	1	1	0
Pete	0	0	0
Helen	0	1	-1
Kate	0	0	0
George	0	1	-1

Now, we want to calculate

ACE $= E[Y^1] - E[Y^0]$

when they were treated − when they weren't treated → take the difference between those expected values.

Example: Individual vs. Average Causal Effect

Treatment: aspirin ($X = 1$) or no aspirin ($X = 0$)
Outcome: headache ($Y = 1$) or no headache ($Y = 0$)

	Potential outcomes		ICE
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$
Charles	1	1	0
Susan	0	1	-1
Tracy	1	1	0
Ken	1	1	0
Pete	0	0	0
Helen	0	1	-1
Kate	0	0	0
George	0	1	-1

average

$\text{ACE} = E[Y^1] - E[Y^0] \quad = \frac{3}{8} - \frac{6}{8} = -0.375$ ⇒ taking aspirin helps.!!

Example: Naive Estimate Based on Observational Data

We only observe one of each individual's potential outcome!

Note: Observational, not experimental data... It's observational data → so we might have confounding going on. Keep in mind!

	Unobserved			Observed	
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i
Charles	**1**	1	0	1	1
Susan	**0**	1	-1	1	0
Tracy	**1**	1	0	1	1
Ken	**1**	1	-1	1	1
Pete	0	**0**	0	0	0
Helen	0	**1**	-1	0	1
Kate	0	**0**	0	0	0
George	0	**1**	-1	0	1

Example: Naive Estimate Based on Observational Data

We only observe one of each individual's potential outcome!

Note: Observational, not experimental data...

	Unobserved			Observed	
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i
Charles	**1**	1	0	1	1
Susan	**0**	1	1	1	0
Tracy	**1**	1	0	1	1
Ken	**1**	1	1	1	1
Pete	0	**0**	0	0	0
Helen	0	**1**	-1	0	1
Kate	0	**0**	0	0	0
George	0	**1**	-1	0	1

Probability of headache after aspirin: $E[Y|X = 1] = \frac{1+0+1+1}{4} = 0.75$

Example: Naive Estimate Based on Observational Data

We only observe one of each individual's potential outcome!
Note: Observational, not experimental data...

	Unobserved			Observed	
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i
Charles	**1**	1	0	1	1
Susan	**0**	1	-1	1	0
Tracy	**1**	1	0	1	1
Ken	**1**	1	-1	1	1
Pete	0	**0**	0	0	0
Helen	0	**1**	-1	0	1
Kate	0	**0**	0	0	0
George	0	**1**	-1	0	1

Probability of headache after aspirin: $E[Y|X=1] = \frac{1+0+1+1}{4} = 0.75$
Probability of headache after no aspirin: $E[Y|X=0] = \frac{0+1+0+1}{4} = 0.5$

Example: Naive Estimate Based on Observational Data

We only observe one of each individual's potential outcome!
Note: Observational, not experimental data...

	Unobserved			Observed	
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i
Charles	**1**	1	0	1	1
Susan	**0**	1	-1	1	0
Tracy	**1**	1	0	1	1
Ken	**1**	1	-1	1	1
Pete	0	**0**	0	0	0
Helen	0	**1**	-1	0	1
Kate	0	**0**	0	0	0
George	0	**1**	-1	0	1

Probability of headache after aspirin: $E[Y|X=1] = \frac{1+0+1+1}{4} = 0.75$
Probability of headache after no aspirin: $E[Y|X=0] = \frac{0+1+0+1}{4} = 0.5$

Hence: $E[Y|X=1] - E[Y|X=0] = 0.75 - 0.5 = 0.25$

We see sth. wierd happening here:
→ implies that taking aspirin makes it worse.
diff. is now positive
opposite conclusion!
If we draw a naive conclusion... solely based on

Example: Naive Estimate Based on Observational Data

the observed outcomes

We only observe one of each individual's potential outcome!
Note: Observational, not experimental data...

	Unobserved			Observed	
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i
Charles	**1**	1	0	1	1
Susan	**0**	1	-1	1	0
Tracy	**1**	1	0	1	1
Ken	**1**	1	-1	1	1
Pete	0	**0**	0	0	0
Helen	0	**1**	-1	0	1
Kate	0	**0**	0	0	0
George	0	**1**	-1	0	1

Probability of headache after aspirin: $E[Y|X=1] = \frac{1+0+1+1}{4} = 0.75$
Probability of headache after no aspirin: $E[Y|X=0] = \frac{0+1+0+1}{4} = 0.5$

only
if we focus on the observed data !!

Hence: $E[Y|X=1] - E[Y|X=0] = 0.75 - 0.5 = 0.25$

Naive conclusion: Aspirin increases one's chances of still having a headache 1 hour later.

Observing vs. Intervening

what we see here is again "correlation $\neq$ causation" is not

Observing $\neq$ intervening

"$E(Y|X=1) - E(Y|X=0)$ is **not the same** as $E(Y^1) - E(Y^0)$"

observing

intervening

: Taking observed data for the treated vs. not treated is very diff. from looking at potential outcomes if ppl were treated vs. not treated

Observing that $E(Y|X=1) \neq E(Y|X=0)$ does **not imply a causal effect** of X on Y. !!

Fancy way of saying correlation $=/=$ causation
there may be confounding, or colliders we conditioned on unknowingly!

Observing vs. Intervening

Observing $\neq$ intervening

$E(Y|X=1) - E(Y|X=0)$ is **not the same** as $E(Y^1) - E(Y^0)$

Observing that $E(Y|X=1) \neq E(Y|X=0)$
does **not imply a causal effect** of X on Y.

Fancy way of saying correlation $=/=$ causation
there may be confounding, or colliders we conditioned on unknowingly!

Pearl calls observing vs intervening the difference between **seeing** and **doing** (more next week).

In Pearl's context, instead of potential outcomes, "do-operator" w/ 2 diff. treatments

Recap

Causality is defined as the "**difference in potential outcomes of an individual**":

Individual causal effect:

$$ICE_i = Y_i^1 - Y_i^0$$

As we cannot observe both potential outcomes, we typically focus on the **average causal effect** instead:

Average causal effect:

$$ACE = E[Y_i^1 - Y_i^0] = E[Y_i^1] - E[Y_i^0]$$

There is a **difference between intervening and observing**:

Observing $\neq$ intervening

$E(Y|X=1) - E(Y|X=0)$ is **not necessarily the same** as $E(Y^1) - E(Y^0)$

key idea: "observing treated vs untreated $\neq$ comparing potential outcomes"

Causal Identification Steps

c.f., Goetghebeur et al., 2020:

1. Define **exposure**, and two levels of interest (e.g., aspirin vs. no aspirin)
   at least
2. Define **outcome variable** (e.g., headache 1 hour later)
   very precisely !!
3. Define **population** of interest (e.g., people who have a headache)
   whose causal effect, for which group? ~ very impo.

Causal Identification Steps

c.f., Goetghebeur et al., 2020:

1. Define **exposure**, and two levels of interest (e.g., aspirin vs. no aspirin)
2. Define **outcome variable** (e.g., headache 1 hour later)
3. Define **population** of interest (e.g., people who have a headache)
4. Formalize the **potential outcomes**, one for each level of treatment

have a very clearly-defined what problem you're trying to solve ~ then you

Causal Identification Steps

c.f., Goetghebeur et al., 2020:

1. Define **exposure**, and two levels of interest (e.g., aspirin vs. no aspirin)
2. Define **outcome variable** (e.g., headache 1 hour later)
3. Define **population** of interest (e.g., people who have a headache)
4. Formalize the **potential outcomes**, one for each level of treatment
5. Specify the causal effect in terms of a parameter to estimate: the **estimand** (e.g., the difference in means between potential outcome distributions)

"what is the thing we're trying to estimate exactly?"

Causal Identification Steps

c.f., Goetghebeur et al., 2020:

1. Define **exposure**, and two levels of interest (e.g., aspirin vs. no aspirin)
2. Define **outcome variable** (e.g., headache 1 hour later)
3. Define **population** of interest (e.g., people who have a headache)
4. Formalize the **potential outcomes**, one for each level of treatment
5. Specify the causal effect in terms of a parameter to estimate: the **estimand** (e.g., the difference in means between potential outcome distributions)
6. State the **assumptions** needed to estimate the causal effect (e.g., no unmeasured confounding)

Causal Identification Steps

c.f., Goetghebeur et al., 2020:

1. Define **exposure**, and two levels of interest (e.g., aspirin vs. no aspirin)
2. Define **outcome variable** (e.g., headache 1 hour later)
3. Define **population** of interest (e.g., people who have a headache)
4. Formalize the **potential outcomes**, one for each level of treatment
5. Specify the causal effect in terms of a parameter to estimate: the **estimand** (e.g., the difference in means between potential outcome distributions)
6. State the **assumptions** needed to estimate the causal effect (e.g., no unmeasured confounding)
7. **Estimate** the causal effect (i.e., choose a particular technique, such as regression, matching, weighting, etc.)

Is this linear effect, is there an interaction effect... etc. All those things you need to think about when you try to estimate specific causal effects.

This also relates to specifying the estimand:

Causal Identification Steps

c.f., Goetghebeur et al., 2020:

1. Define **exposure**, and two levels of interest (e.g., aspirin vs. no aspirin)
2. Define **outcome variable** (e.g., headache 1 hour later)
3. Define **population** of interest (e.g., people who have a headache)
4. Formalize the **potential outcomes**, one for each level of treatment
5. Specify the causal effect in terms of a parameter to estimate: the **estimand** (e.g., the difference in means between potential outcome distributions)
6. State the **assumptions** needed to estimate the causal effect (e.g., no unmeasured confounding)
7. **Estimate** the causal effect (i.e., choose a particular technique, such as regression, matching, weighting, etc.)
8. **Sensitivity analysis** (important, but not covered in this course)

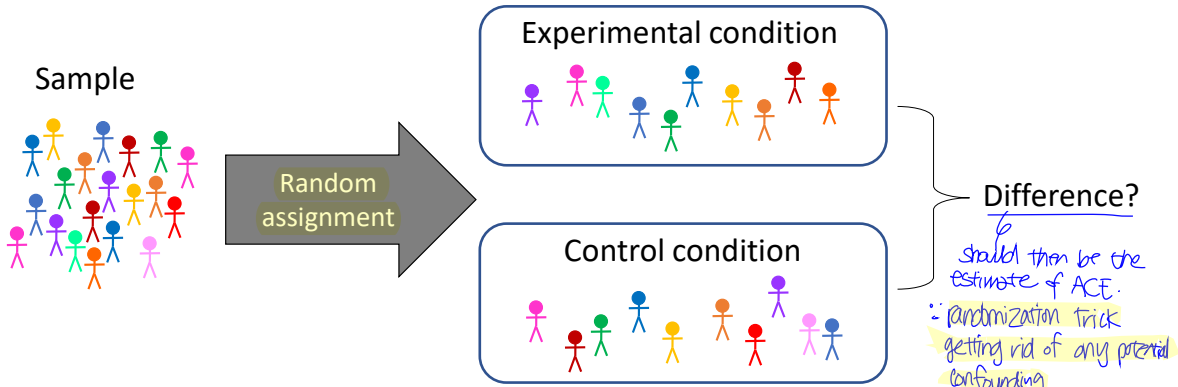
Overview

- Potential Outcomes & Causal Effects
- **Estimating Causal Effects: Assumptions**
- Estimating Causal Effects: Controlling for Confounders

Rubin's Potential Outcome Framework: Assumptions for identification of a (Average) causal effect based on observational data

"Assumptions for identification of an Average Causal Effect based on observational data"

RCTs: The gold standard for estimating ACEs



What we wanna have w/ observational data;

"randomized control trial"

Essentially, we want to mimic the situation we have in an RCT with random assignment e.g., every unit should be equally likely to be in the treatment or control group.

From Observation to Causation Assumption 1: Exchangeability

Exchangeability: in the context of potential outcome framework

For each unit Treatment is independent of their potential outcomes: $Y_i^1, Y_i^0 \perp\!\!\!\perp X_i$ → Your potential outcome should have nothing to do w/ in what group you end up.

a bit confusing ; talking about "potential" outcomes, not "observed outcomes" !

From Observation to Causation Assumption 1: Exchangeability

Exchangeability:

For each unit Treatment is independent of their potential outcomes: $Y_i^1, Y_i^0 \perp\!\!\!\perp X_i$

Violation of Exchangeability:

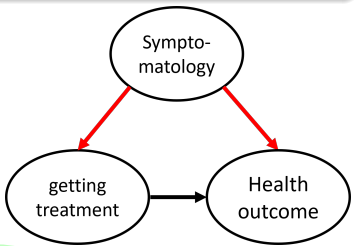
When there is a relation between people's assignment to treatment and their potential outcomes.

Essentially, your potential outcomes should have nothing to do with the treatment group.

You should have equal probability ending up in either the treatment or control group.

That's what $Y_i^1, Y_i^0 \perp\!\!\!\perp X_i$ this says!

Example:



From Observation to Causation Assumption 1: Exchangeability

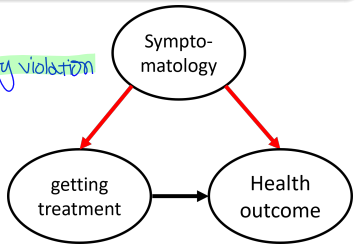
Essentially, whatever treatment you get, it should not depend on your potential result or the effects of treatment.
It's a very stringent assumption, & we often don't need this complete assumption

Exchangeability:

For each unit Treatment is independent of their potential outcomes: $Y_i^1, Y_i^0 \perp\!\!\!\perp X_i$

Violation of Exchangeability: → "confounding" is one ex. of exchangeability violation
When there is a relation between people's assignment to treatment and their potential outcomes.

Example:

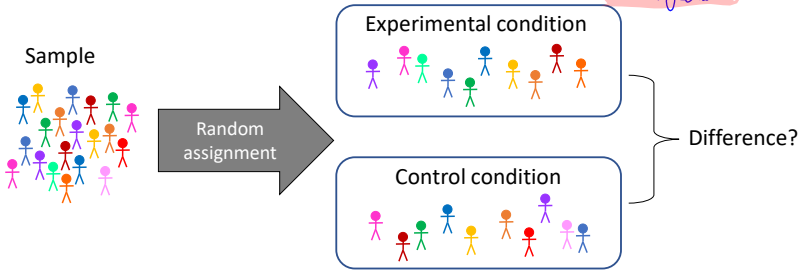


- The same idea of no backdoor paths between X and Y! in DAG framework
- SO, exchangeability = no confounding!

might be more general ∵ you don't necessarily need to think about 3rd variable.
It also means your outcome variable doesn't determine what treatments you get

Exchangeability related to RCTs

ppl in control & experimental condition should be "exchangeable."



Due to random assignment, in RCTs individuals receiving treatment ($X = 1$) are **exchangeable** with respect to their potential outcome with those who do not receive treatment ($X = 0$):

$$E[Y^1|X = 1] = E[Y^1|X = 0] \text{ and}$$
$$E[Y_i^0|X = 1] = E[Y^0|X = 0]$$

Exp. value of potential outcome from ppl who were treated = Exp. value of P.O. from ppl who weren't treated

Without random assignment (observational data), what treatment people get may depend on third variables that also relate to the potential outcomes...Confounding!

Exchangeability and Missing Data

) Here the link to missing data becomes very explicit!

Only part (at best half) of the potential outcomes is observed.
This makes causal inference a **missing data problem**. !!

Exchangeability and Missing Data

Only part (at best half) of the potential outcomes is observed.
This makes causal inference a **missing data problem**.

Exchangeability is about: What is the **missing data mechanism** for the potential outcomes?

Exchangeability and Missing Data

Only part (at best half) of the potential outcomes is observed.
This makes causal inference a **missing data problem**.

Exchangeability is about: What is the **missing data mechanism** for the potential outcomes?

- Missing Completely At Random (MCAR)?
- Missing At Random (MAR; random, once accounted for observed covariates)
- Missing Non At Random (MNAR; missing patterns depend on unobserved covariates or the outcome) ~ violation of exchangeability
- Check out: https://stefvanbuuren.name/fimd/sec-MCAR.html

full
Exchangeability implies a "MCAR" assumption for the missing potential outcomes, conditional on the Treatment variable.

confounder

treated ppl have higher scores on confounder.

↓

This is what can happen if you have observational data.

In this case, ppl select themselves into treatment based on their pre-symptoms (confounder)

Assumption: "Conditional Exchangeability"

Instead of assuming exchangeability, we may assume "**conditional exchangeability**": Conditional on a set of observed covariates, the potential outcomes are independent of treatment assignment.

Conditional Exchangeability:

$Y_i^1, Y_i^0 \perp\!\!\!\perp X_i | Z_i$; given a specific value of Z, the assignment is random.

↓

fix Z to a specific value

for ex) only looking at the group of ppl w/ really severe headache.

In that group, the assignment to treatment/control should be random. That's the assumption!

Assumption: Conditional Exchangeability

Instead of assuming exchangeability, we may assume **conditional exchangeability**: Conditional on a set of observed covariates, the potential outcomes are independent of treatment assignment.

Conditional Exchangeability:

$Y_i^1, Y_i^0 \perp\!\!\!\perp X_i | Z_i$; given a specific value of Z, the assignment is random.

That is, no **unobserved** confounding.

Assumption: Conditional Exchangeability

Instead of assuming exchangeability, we may assume **conditional exchangeability**: Conditional on a set of observed covariates, the potential outcomes are independent of treatment assignment.

Conditional Exchangeability:

$Y_i^1, Y_i^0 \perp\!\!\!\perp X_i | Z_i$; given a specific value of Z, the assignment is random.

That is, no **unobserved** confounding.

This implies *missing at random* (MAR) assumption!
(rather than MCAR, missing completely at random)

From Observation to Causation Assumption 2: Positivity

There must be exposed and unexposed participants at every combination of values of our observed confounders Z in the population under study.

= In every value of confounders, there must be treated & untreated ppl

↓

otherwise, we'll have (extrapolation) problem.

From Observation to Causation Assumption 2: Positivity

There must be exposed and unexposed participants at every combination of values of our observed confounders Z in the population under study.

In an RCT, positivity should be present by design. → ∵ bcuz it doesn't matter what value you have on confounding variables, you are just assigned to either control/treatment by chance!

So as long as there're enough ppl, it should all work in RCT.

From Observation to Causation Assumption 2: Positivity

There must be exposed and unexposed participants at every combination of values of our observed confounders Z in the population under study.

In an RCT, positivity should be present by design.

Violations can be spotted by:

- making tables of each categorical covariate and treatment (should be no empty cells)
- categorize a continuous covariate and make table (but this depends on number and width of categories)
- considering all combinations of covariates (becomes impossible, but then we use propensity scores, tbd later!)

as the number of covariates ↑

From Observation to Causation Assumption 2: Positivity

There must be exposed and unexposed participants at every combination of values of our observed confounders Z in the population under study.

In an RCT, positivity should be present by design.

Violations can be spotted by:

- making tables of each categorical covariate and treatment (should be no empty cells)
- categorize a continuous covariate and make table (but this depends on number and width of categories)
- considering all combinations of covariates (becomes impossible, but then we use propensity scores, tbd later!)

Positivity and exchangeability combined are also known as **strong ignorability**.

positivity + exchangeability = strong ignorability

From Observation to Causaion Assumption 3: Consistency

Consistency links observed outcomes Y_i to potential outcomes Y_i^X, through:

- $Y_i = Y_i^1$ for individuals with $X_i = 1$: ppl in treatment group. their observed outcome should be their potential outcome if they were treated.
- $Y_i = Y_i^0$ for individuals with $X_i = 0$: ppl in control group. their observed outcome should be their potential outcome for if they were not treated.

From Observation to Causaion Assumption 3: Consistency

Consistency links observed outcomes Y_i to potential outcomes Y_i^X, through:

- $Y_i = Y_i^1$ for individuals with $X_i = 1$
- $Y_i = Y_i^0$ for individuals with $X_i = 0$

Consistency:

$$Y_i = Y_i^x \text{ for } X_i = x_i$$

In words: the observed outcome Y_i equals the potential outcome for the treatment level that was observed.

From Observation to Causaion Assumption 3: Consistency

Consistency links observed outcomes Y_i to potential outcomes Y_i^X, through:

- $Y_i = Y_i^1$ for individuals with $X_i = 1$
- $Y_i = Y_i^0$ for individuals with $X_i = 0$

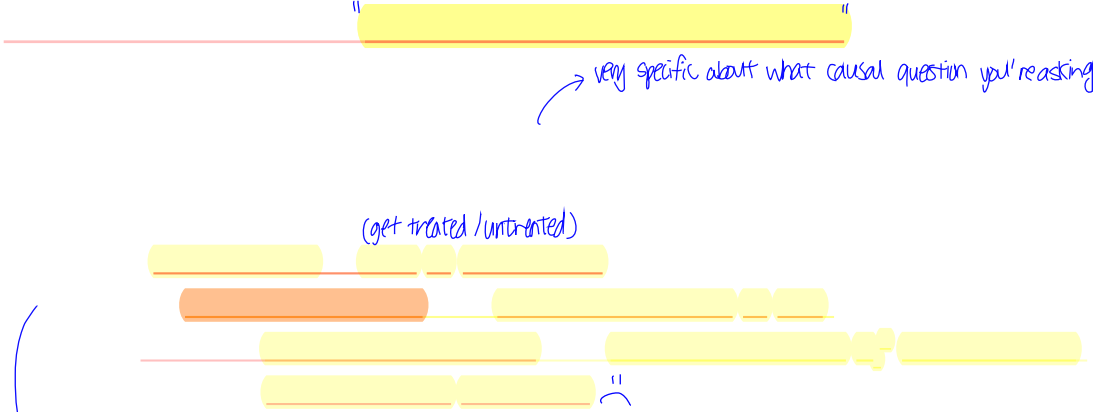
Consistency:

$$Y_i = Y_i^x \text{ for } X_i = x_i$$

In words: the observed outcome Y_i equals the potential outcome for the treatment level that was observed.

This assumption requires:

- well defined treatment and no-treatment conditions
- one treatment level implies one specific version of treatment : you shouldn't have diff. versions of treatment.
- no measurement error : If we wrote down treated, they were actually in the treated group & we wrote down their correct outcome value



Ex) diff. therapists giving the same treatment.
But some are really good, some aren't
If you have a bad one, it might be the
same as not being treated, but we wrote down
as "being treated". This might mess it up!

Consistency Violation: Different versions of treatment

Published in final edited form as:
Ann Epidemiol. 2016 October ; 26(10): 674–680. doi:10.1016/j.annepidem.2016.08.016.

Does water kill? A call for less casual causal inferences

Miguel A. Hernán[1,2]

If there are multiple ways to raise X from 0 to 1, this means:

- there are multiple treatments (i.e., multiple versions of $X = 1$)
- these may have different causal effects (i.e., multiple versions of Y_i^1 for one person)
- this renders the causal question **ill-defined**

E.g.: Changing BMI to affect weight - lower BMI achieved via Muscle loss vs Fat loss → You have to be more specific. Otherwise you're gonna jumble diff. causal effects.

Consistency Violation: Different versions of treatment

Published in final edited form as:
Ann Epidemiol. 2016 October ; 26(10): 674–680. doi:10.1016/j.annepidem.2016.08.016.

Does water kill? A call for less casual causal inferences

Miguel A. Hernán[1,2]

If there are multiple ways to raise X from 0 to 1, this means:

- there are multiple treatments (i.e., multiple versions of $X = 1$)
- these may have different causal effects (i.e., multiple versions of Y_i^1 for one person)
- this renders the causal question **ill-defined**

E.g.: Changing BMI to affect weight - lower BMI achieved via Muscle loss vs Fat loss

E.g.: Different therapists providing the treatment, some therapists are better than others.

Consistency Violation: Different versions of treatment

Published in final edited form as:
Ann Epidemiol. 2016 October ; 26(10): 674–680. doi:10.1016/j.annepidem.2016.08.016.

Does water kill? A call for less casual causal inferences

Miguel A. Hernán[1,2]

If there are multiple ways to raise X from 0 to 1, this means:

- there are multiple treatments (i.e., multiple versions of $X = 1$)
- these may have different causal effects (i.e., multiple versions of Y_i^1 for one person)
- this renders the causal question **ill-defined**

E.g.: Changing BMI to affect weight - lower BMI achieved via Muscle loss vs Fat loss

E.g.: Different therapists providing the treatment, some therapists are better than others.

Hernán: Specify a **target trial** with a causal question: "A detailed description of the RCT one *would* have done, had there been no ethical/practical limitations."

very explicit description what kind of RCT you have in mind if you're evaluating this causal effect

Podcast about target trial with Hernán: https://casualinfer.libsyn.com/casual-inference-talking-target-trials-with-miguel-hernan-episode-01

SUTVA ~ tied to the consistency idea..

(Stable Unit Treatment Value Assumption)

Others (including Rubin) use the "Stable Unit Treatment Assumption (SUTVA)";
it stems from going from the ICE to the ACE. contains consistency assumption: no diff. version of treatments

Stable unit treatment value assumption (SUTVA): potential observation on one unit should be unaffected by the particular assignment of treatments to the other units

For each unit:

- The potential outcomes do not vary with the treatments assigned to other units (i.e., RCT: "no interference")
- there are **no** different versions of each treatment level that lead to different potential outcomes (i.e., the consistency assumption).

"you have sb. in control group who lives together w/ sb in treatment group. & treated person influences the person in the control group in some ways" ~ This kind of interference shouldn't happen

Overview

- Potential Outcomes & Causal Effects
- Estimating Causal Effects: Assumptions
- **Estimating Causal Effects: Controlling for Confounders**

Assume we have only confounders & then we are gonna apply these techniques

And also assume together they are really good estimate of causal effect, assumed that they are all relevant confounders, of course.

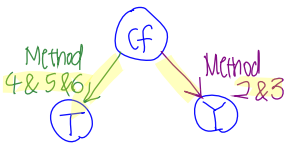
Controlling for Confounders

Many different techniques available.

We follow Schafer and Kang (2008) who discuss 9 different techniques, which can be divided into 4 main strategies:

- **No confounders** included: naive approach (Method 1)
- Model the relation between **confounders and Y**: account for confounding by including confounders as covariates (Methods 2 and 3)
- Model the relation between **confounders and X**: adjust the data to what would have been obtained in an RCT (Methods 4, 5, and 6)
- Model the relation between **confounders and X AND Y**: assumptions need to be correct for at least one of the two (Methods 7, 8 , and 9; not this course).

Note: In the lab, you will apply methods 1-6.



1. Naive Estimator: Prima Facie Effect

Simply the "observed difference" in means between treatment groups.

Prima facie effect

$$PFE = E[Y_i^1|X_i = 1] - E[Y_i^0|X_i = 0]$$

i.e., we assume "exchangeability": (first) we don't control for any confounders. So we need to assume that there's no confounders at all.

- $E[Y_i^1|X_i = 1] = E[Y_i^1|X_i = 0]$
- $E[Y_i^0|X_i = 0] = E[Y_i^0|X_i = 1]$

that is, there is no confounding.

This mean difference could, for example, be estimated using regression analysis with a dummy predictor.

→ we saw this earlier w/ aspirin example. & In many cases, this prolly is not a good way to go

Model the relation between confounders (Z_i) and outcome (Y_i)

Account for imbalance by including confounders as covariates in a model that relates the outcome and treatment.

→ Adding confounders as predictors in your regression model next to the treatment variables.

(Be sure that covariates are confounders!)

2a. ANCOVA (here, linear regression without interactions)

S&K refers to

Confounders are added as predictors ('control variables') to a linear regression of the outcome on the treatment variable.

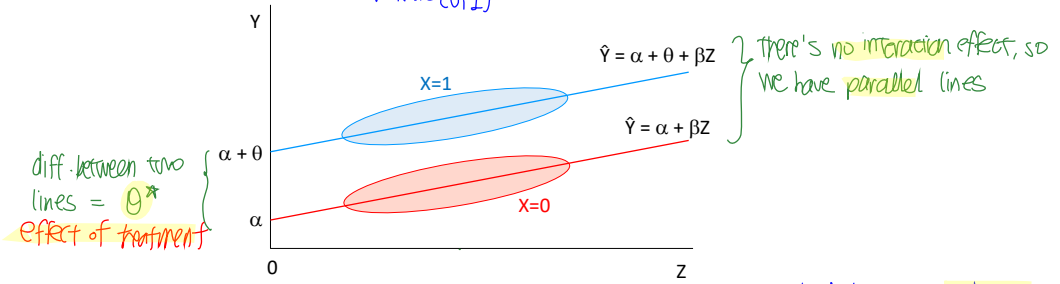
ANCOVA

estimate of causal effect

$$Y_i = \alpha + \theta X_i + \beta_1 Z_{i1} + \cdots + \beta_p Z_{ip} + e_i$$

treatment variable (0, 1)

effects of our confounding variables



Here θ is the estimator of the (causal) effect of X.

It is the effect of X when Z are equal to zero.

And since we don't have interaction effects, it's the same as keeping those confounders fixed at any other value!

2a. ANCOVA: Estimator Average Causal Effect

*Assumptions:

- conditional exchangeability - no unobserved confounders
- consistency
- correct model specification (e.g. linearity ...)
- (Also, are confounders and not colliders/mediators; no unobserved conditioning on colliders) are truly confounders

$$
\begin{aligned}
\text{ACE} &= E[Y_i^1 - Y_i^0] \\
&= E[Y_i^1] - E[Y_i^0] \\
&= E[Y_i^1 | X_i = 1, Z_i] - E[Y_i^0 | X_i = 0, Z_i] && \leftarrow \text{No unobserved confounding} \\
&= E[Y_i | X_i = 1, Z_i] - E[Y_i | X_i = 0, Z_i] && \leftarrow \text{Consistency (potential outcomes = observed outcome)} \\
&= \{\alpha + \theta + Z_i'\beta\} - \{\alpha + Z_i'\beta\} && \leftarrow \text{Linearity, no interactions} \\
&= \theta
\end{aligned}
$$

($X=1$) ($X=0 \rightarrow \theta$ drops out)

This part is questionable

θ = ACE!

BUT this is what we do here w/ ANCOVA.

2b. Regression analysis (linear regression more general than ANCOVA)

For example interactions between covariaties, treatment and covariates, non-linear effects, etc.

→ basically we add in more complicated predictor structures, which give us some additional thing to think about when we estimate our ACE.

ex) If we have an effect of (confounder)^2, still it's linear reg. but it's non-linear.

2b. Regression analysis (linear regression more general than ANCOVA)

For example interactions between covariaties, treatment and covariates, non-linear effects, etc.

Regression model with treatment*covariates interactions:

$$Y_i = \alpha + \theta X_i + \beta_1 Z_{i1} + \cdots + \beta_p Z_{ip} + \gamma_1 X_i \times Z_{i1} + \cdots + \gamma_p X_i \times Z_{ip} + e_i$$

interaction between treatment X + covariates Z

where

- β's are the slopes for outcome with covariates in the control group ($X_i = 0$)
- γ's are the interaction effects; the differences in slopes between the treatment group ($X_i = 1$) and the control group ($X_i = 0$).

⇒ As soon as you have interaction, the effects of treatments differ for different levels of covariates!

So we have different causal effects depending on the score you have on a particular covariate.

Also means that you don't necessarily have one causal effect, but you could consider multiple ones

The ACE then = effect of treatment for the "average" level of covariates → The effect depends on the level

Treatment*Covariate interactions

of covariates

With treatment*covariates interactions the treatment effect depends on the values of the covariates.

$$E\left[Y_i^1 - Y_i^0\right] = E\left[Y_i^1 | X_i = 1, Z_i\right] - E\left[Y_i^0 | X_i = 0, Z_i\right] = E\left[Y_i | X_i = 1, Z_i\right] - E\left[Y_i | X_i = 0, Z_i\right]$$
$$= \left\{\alpha + \theta + Z_i'(\beta + \gamma)\right\} - \left\{\alpha + Z_i'(\beta)\right\} = \theta + Z_i'\gamma$$

now the causal effect is θ + interaction effect × particular score on covariate Z_i

then you need

Average causal effect - for the average Z:

$$ACE = \theta + E\left[Z_i\right]'\gamma$$

you can estimate this using centering technique

Here you can really see the specific effect will depend on value of Z

Treatment*Covariate interactions

With treatment*covariates interactions the treatment effect depends on the values of the covariates.

$$E\left[Y_i^1 - Y_i^0\right] = E\left[Y_i^1 | X_i = 1, Z_i\right] - E\left[Y_i^0 | X_i = 0, Z_i\right] = E\left[Y_i | X_i = 1, Z_i\right] - E\left[Y_i | X_i = 0, Z_i\right]$$
$$= \left\{\alpha + \theta + Z_i'(\beta + \gamma)\right\} - \left\{\alpha + Z_i'(\beta)\right\} = \theta + Z_i'\gamma$$

Average causal effect - for the average Z:

$$ACE = \theta + E\left[Z_i\right]'\gamma$$

this two can be diff. bcuz we have the interaction effect

Average causal effect for the treated - for the average Z when X=1

$$ACE_1 = \theta + E\left[Z_i | X_i = 1\right]'\gamma$$

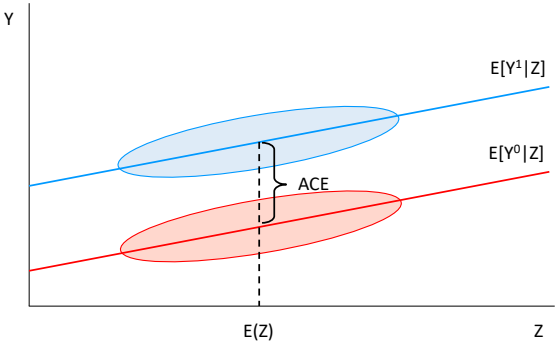
Average causal effect for the controls - for the average Z when X=0:

$$ACE_0 = \theta + E\left[Z_i | X_i = 0\right]'\gamma$$

* It's very impo. to realize these are different from ACE. (: Later we'll use specific techniques that actually only

Case 1: ACE in the ANCOVA model

estimate one of these, not ACE)

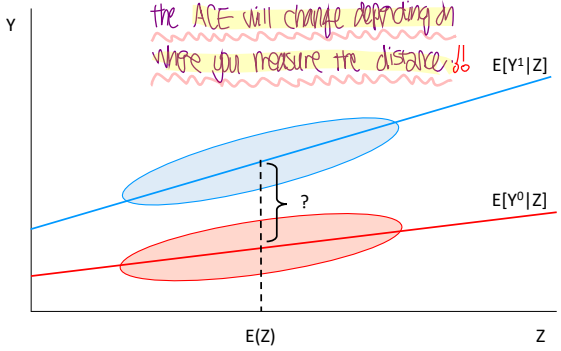


$$ACE = \theta + E\left[Z_i\right]\gamma = \theta$$

This is the classical ANCOVA scenario in which:

- there are no differences on the covariate Z between the treatment groups (akin to expectation under a RCT)
- there is no interaction between treatment X and covariate Z - regression lines are parallel.

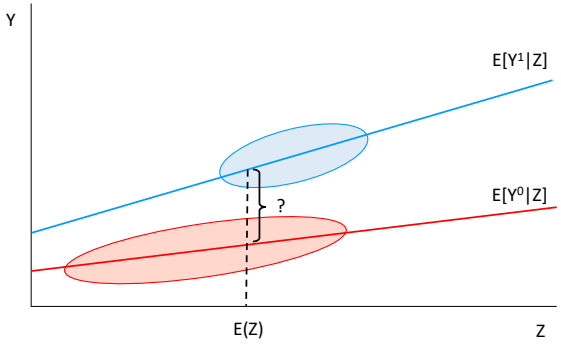
Case 2: What is it?



Explain:

- what kind of scenario this is
- what the quantity denoted by ? represents

Case 3: What is it?



Explain:

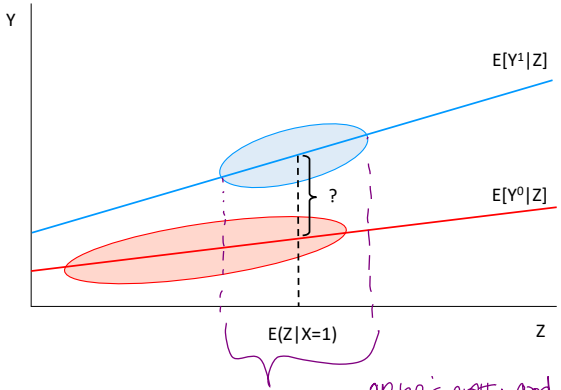
- what kind of scenario this is?
- what does the quantity denoted by ? represent?

ANOVA w/ interaction but positivity ass. violated

well the extrapolation is done...

estimate might not be good depending on how

Case 4: What is it?



Explain:

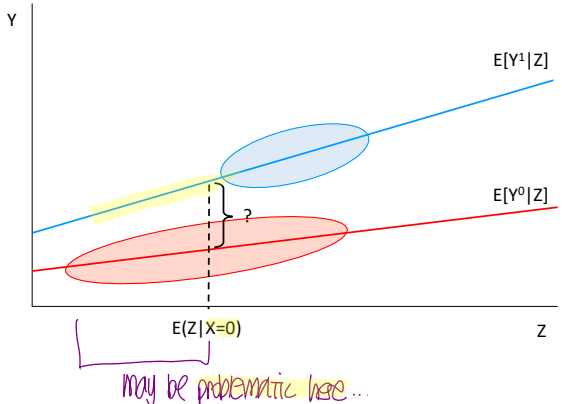
- what kind of scenario this is?
- what does the quantity denoted by ? represent?

ACE

overlap is pretty good for the treated group.

So this one probably is okay.

Case 5: What is it?



Explain:

- what kind of scenario this is?
- what does the quantity denoted by ? represent? ACE$_0$

Method 3. "Regression estimation"

estimate unobserved potential outcomes explicitly.
gonna basically impute those missing values based on those confounders

Idea: Estimate unobserved potential outcomes using the confounders

1. get parameter estimates for the confounders for units with $X_i = 1$: $\hat{\beta}^1$
2. Use this estimated model to predict for (the unobserved) potential outcomes when treated: $\hat{Y}_i^1 = Z_i' \hat{\beta}^1$
3. get parameter estimates for the confounders for units with $X_i = 0$: $\hat{\beta}^0$
4. Use this estimated model make prediction for (the unobserved) potential outcome when not treated: $\hat{Y}_i^0 = Z_i' \hat{\beta}^0$

Using these (estimates of) both potential outcomes for every person:

① Get the parameter estimate for the confounders; relationship between Y & confounders for the treated group → $\hat{\beta}^1$

② Use that model to predict the potential outcomes for the other group → $\hat{Y}_i^1 = Z_i' \hat{\beta}^1$

③ Then we do the same thing the other way around: get the par. estimate for the untreated group → $\hat{\beta}^0$

④ Use that to predict the potential outcomes for the treated group → $\hat{Y}_i^0 = Z_i \hat{\beta}^0$

⇒ We can just take the differences between those estimates of potential outcomes.

Method 3. "Regression estimation"

Idea: Estimate unobserved potential outcomes using the confounders

1. get parameter estimates for the confounders for units with $X_i = 1$: $\hat{\beta}^1$
2. Use this estimated model to predict for (the unobserved) potential outcomes when treated: $\hat{Y}_i^1 = Z_i'\hat{\beta}^1$
3. get parameter estimates for the confounders for units with $X_i = 0$: $\hat{\beta}^0$
4. Use this estimated model make prediction for (the unobserved) potential outcome when not treated: $\hat{Y}_i^0 = Z_i'\hat{\beta}^0$

Using these (estimates of) both potential outcomes for every person:

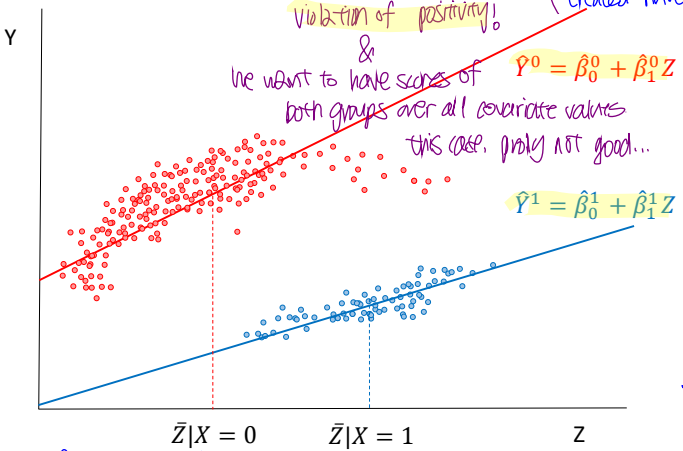
Regression estimate of ACE:

Average difference between "estimated" potential outcomes! (Notice the hat)

$$\hat{ACE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i^1 - \hat{Y}_i^0)$$

Note: We can either use predicted potential outcomes for only the unobserved potential outcomes, or for all -both observed an unobserved- potential outcomes.

Method 3. Regression Estimation



control : see they have relatively low value for the confounder.
treated have relatively high value for the confounder.

That's the violation of positivity!
&
We want to have scores of both groups over all covariate values
this case, prety not good...

If we have values all over the whole range of Z (confounders) then we could do much better job!

Important: Positivity Assumption!

For example: We need $\hat{\beta}^0$, based on observations around $\bar{Z}|X = 0$ to make predictions around $\bar{Z}|X = 1$.

If positivity is violated, this requires questionable extrapolation.

means
we have a lot of uncertainty about those estimated potential outcomes.

Model the relation between
confounders (Z_i) and treatment (X_i)

Propensity Scores

based on the relationship between confounders & treatment variable

All remaining techniques use **propensity scores**: the "probability of each unit of being treated".

Propensity scores (assuming conditional exchangeability - no unobserved confounding):

$$\pi_i = P[X_i = 1|Z_i]$$

They are easier to use instead of million different covariates...

Goal: Use propensity scores instead of confounder covariates in model in some way or form (depends on exact method).

For this to work, we want $Z_i \perp\!\!\!\perp X_i|\pi_i$.

Confounder & treatment (group) become independent of each other,

completely explain away the relationship w/ the propensity scores.

So you don't need confounders anymore, just use the propensity scores instead.

Bcuz they're based on all possible combinations of diff. confounders essentially

Propensity Scores

π_i are probabilities, hence often estimated with "logistic regression".

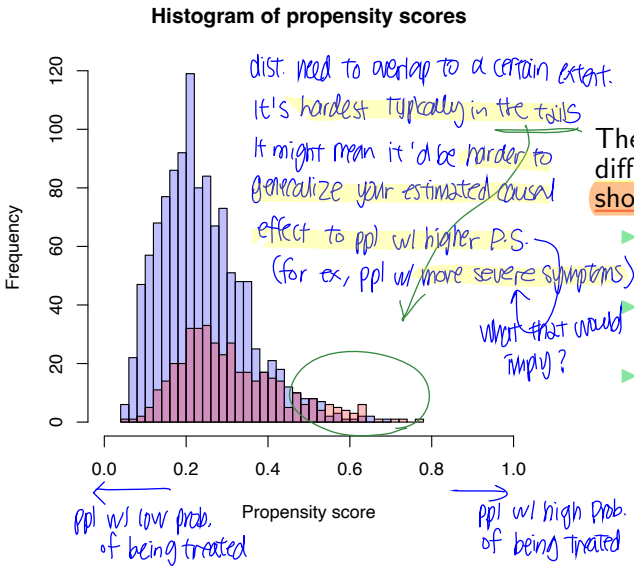
Propensity scores (assuming conditional exchangeability - no unobserved confounding):

$$\pi_i = P[X_i = 1|Z_i] = \frac{exp(Z_i'\phi)}{1+exp(Z_i'\phi)}$$

- Outcome variable: Treatment group (0,1)
- Predictors: Identified confounding variables
- Fit logistic regression model and save "predicted probabilities" per unit. → those are your estimated propensity scores

> In ideal situation,
everybody would have 50:50 probability of being treated.
but it's typically not the case in observational data.

Show overlap of propensity scores



The distributions of the propensity scores may differ for the treated and the untreated, but should overlap to serve causal inference.

- Non-overlapping areas imply a violation of positivity assumption.
- which can lead to problems for the methods using propensity scores
- e.g., issues with extrapolation.

the issue is still there, but now we have captured all diff. covariates into one variable.

Method 4. Matching

Matching implies you create pairs that consist of a treated and a non-treated unit, who have **identical covariate scores**.

Hard in case of many covariates: Propensity scores offer a solution.

Background: In an RCT we have: $P(Z|X=1) = P(Z|X=0)$

prob. of having a certain score on the covariate, given that you're in the treated group = should be the same = as in you're in the control group.

∵ Bcuz the covariates should not have anything to do w/ whether you end up in treatment / control groups.

Method 4. Matching

That's also sth. we wanna achieve w/ propensity scores

Matching implies you create pairs that consist of a treated and a non-treated unit, who have **identical covariate scores**.
Hard in case of many covariates: Propensity scores offer a solution.

Background: In an RCT we have: $P(Z|X=1) = P(Z|X=0)$

Balancing property balance out the covariates in each group via propensity score

$$P(Z|\pi = c, X = 1) = P(Z|\pi = c, X = 0)$$

If the propensity model is correct, then comparing treated and untreated individuals with the same π is a way of mimicking an RCT - you balance out the covariates among the (new) two groups.

Things to take care when matching:
- If we start w/ a treated group, we match ppl to the characteristic of treated group,
- And then we might leave out some ppl in the control group who are different from the treated group. And ACE1 can be different from the ACE0 or ACE.

Method 4. Matching

Matching implies you create pairs that consist of a treated and a non-treated unit, who have **identical covariate scores**.
Hard in case of many covariates: Propensity scores offer a solution.

Background: In an RCT we have: $P\big(Z|X=1\big) = P\big(Z|X=0\big)$

Balancing property:

$$P\big(Z|\pi=c, X=1\big) = P\big(Z|\pi=c, X=0\big)$$

If the propensity model is correct, then comparing treated and untreated individuals with the same π is a way of mimicking an RCT - you balance out the covariates among the (new) two groups.

!Note! We match units from the largest group to the characteristics of the units of the smallest group. !!! we are limited to the size of either control / treatment group, whichever group is the smallest.

Matching provides us not with the ACE, but with "**ACE1** or **ACE0**" ↙ As a result, what we get is not ACE, but ACE, of treated

Method 4. Matching

or ACE of control group, depending on what group we start with.

Many important matching techniques.

Important: Evaluating how successful matching was.

*Main critique of matching: There're so many researcher's degrees of freedom & modeling choices that might affect the end result. Matching is still very popular but it's still very heavily criticized

Check If the two matched groups have the similar values on covariates, & based on that make changes... (Researchers' df) again

When matching happens, the group sizes are equal!

Stratified by DIET

	0		1		SMD
n	1220		1220		
DISTR.1 (mean (SD))	0.71	(0.45)	0.71	(0.45)	0.007
BLACK (mean (SD))	0.18	(0.38)	0.17	(0.38)	0.004
NBHISP (mean (SD))	0.16	(0.36)	0.15	(0.36)	0.007
GRADE (mean (SD))	9.37	(1.35)	9.37	(1.34)	0.002
SLFHLTH (mean (SD))	2.36	(0.95)	2.35	(0.91)	0.011
SLFWGHT (mean (SD))	3.82	(0.69)	3.84	(0.70)	0.033
WORKHARD (mean (SD))	2.07	(0.86)	2.05	(0.85)	0.022
GOODQUAL (mean (SD))	1.81	(0.65)	1.84	(0.71)	0.049
PHYSFIT (mean (SD))	2.53	(0.97)	2.53	(0.93)	0.007
PROUD (mean (SD))	1.85	(0.77)	1.86	(0.79)	0.011
LIKESLF (mean (SD))	2.46	(1.05)	2.52	(1.06)	0.057
ACCEPTED (mean (SD))	2.33	(1.03)	2.35	(1.06)	0.023
FEELLOVD (mean (SD))	1.92	(0.87)	1.93	(0.90)	0.010

look at the differences in means between the matched groups. (there're rules of thumb → again Researchers' df)

Podcast about matching (and propensity scores, simulations, identifiability assumptions, etc.): https://seriousepi.blubrry.net/2021/03/01/1-16-finding-the-perfect-match-requires-common-support-matching-with-dr-anusha-vable/

Method 5. Inverse probability weighting

The **probability of received treatment** is:

- π_i for those who were **treated** $(X_i = 1)$
- $1 - \pi_i$ for those who were **NOT treated** $(X_i = 0)$

We'll use these as "weights"

Method 5. Inverse probability weighting

The **probability of received treatment** is:

- π_i for those who were **treated** $(X_i = 1)$
- $1 - \pi_i$ for those who were **NOT treated** $(X_i = 0)$

Among treated individuals $(X = 1)$, those with large π_i are *overrepresented* in comparison to those with small π_i.

low prob. of getting treated

Among untreated individuals $(X = 0)$, those with large $1 - \pi_i$ are *overrepresented* in control group comparison to those with small $1 - \pi_i$.

⇒ so we wanna correct for this w/ weights!

Method 5. Inverse probability weighting

The **probability of received treatment** is:

- π_i for those who were **treated** $(X_i = 1)$
- $1 - \pi_i$ for those who were **NOT treated** $(X_i = 0)$

Among treated individuals $(X = 1)$, those with large π_i are *overrepresented* in comparison to those with small π_i.

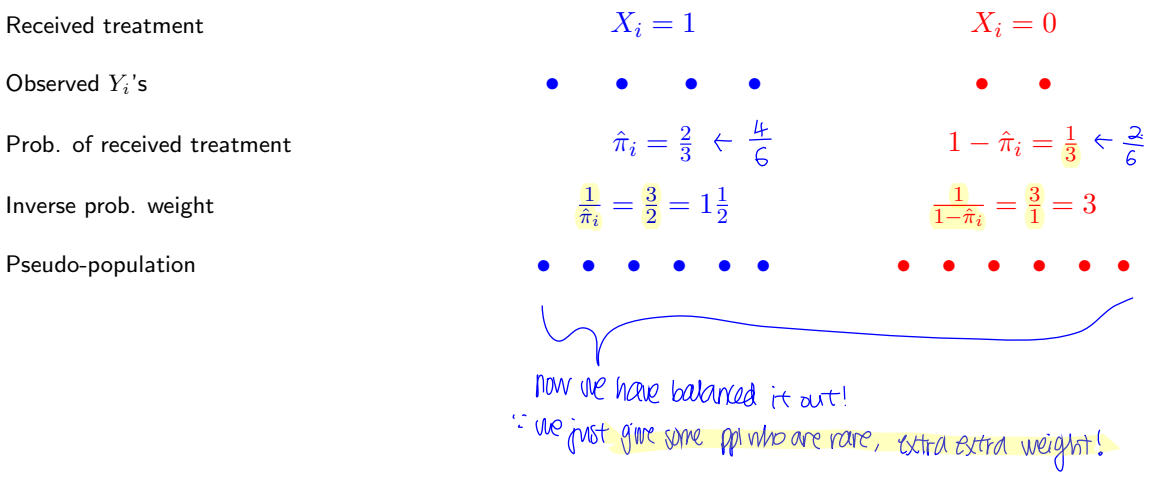
Among untreated individuals $(X = 1)$, those with large $1 - \pi_i$ are *overrepresented* in comparison to those with small $1 - \pi_i$.

To account for this inbalance, we:

- create a pseudo-population
- by weighing each unit by the inverse probability of their received treatment:
  - weight $\frac{1}{\hat{\pi}_i}$ for units with $X_i = 1$
  - weight $\frac{1}{1-\hat{\pi}_i}$ for units with $X_i = 0$

Method 5. Inverse probability weighting - Pseudo Population

Let's focus on the individuals in our sample who have $\hat{\pi}_i = \frac{2}{3}$.



Method 5. Inverse probability weighting - Pseudo Population

Let's focus on the individuals in our sample who have $\hat{\pi}_i = \frac{2}{3}$.

	$X_i = 1$	$X_i = 0$
Received treatment		
Observed Y_i's	● ● ● ●	● ●
Prob. of received treatment	$\hat{\pi}_i = \frac{2}{3}$	$1 - \hat{\pi}_i = \frac{1}{3}$
Inverse prob. weight	$\frac{1}{\hat{\pi}_i} = \frac{3}{2} = 1\frac{1}{2}$	$\frac{1}{1-\hat{\pi}_i} = \frac{3}{1} = 3$
Pseudo-population	● ● ● ● ● ●	● ● ● ● ● ●

Do this for each case (based on their own $1/\hat{\pi}_i$); the resulting pseudo-population:

- is (about) twice as large as the original sample !!
- makes $\pi_i = 0.5$ for every unit (as in an RCT) In this way, in this pseudo-population, the prob. of being treated & not treated should be 0.5, the same.
- and thus should have $Z_i \perp\!\!\!\perp X_i$ (as in an RCT)
- (should be checked if succesful)

Check balancing in pseudo-population

To check whether our pseudo-population **mimics an RCT**, we could check standardized mean differences on the covariates:

	Stratified by DIET 0	1	SMD
n	6014.23	6368.45	
DISTR.1 (mean (SD))	0.64 (0.43)	0.63 (0.43)	0.028
BLACK (mean (SD))	0.24 (0.43)	0.25 (0.44)	0.035
NBHISP (mean (SD))	0.15 (0.35)	0.13 (0.33)	0.061
GRADE (mean (SD))	9.20 (1.38)	9.26 (1.41)	0.037
SLFHLTH (mean (SD))	2.23 (0.94)	2.25 (0.91)	0.020
SLFWGHT (mean (SD))	3.33 (0.80)	3.15 (1.00)	0.196
WORKHARD (mean (SD))	2.12 (0.90)	2.14 (0.86)	0.017
GOODQUAL (mean (SD))	1.81 (0.67)	1.79 (0.68)	0.036
PHYSFIT (mean (SD))	2.30 (0.95)	2.31 (0.90)	0.014
PROUD (mean (SD))	1.78 (0.77)	1.77 (0.76)	0.018
LIKESLF (mean (SD))	2.18 (1.02)	2.14 (1.00)	0.045
ACCEPTED (mean (SD))	2.18 (1.01)	2.16 (1.02)	0.024
FEELLOVD (mean (SD))	1.81 (0.84)	1.79 (0.83)	0.028

ideally now the diff. between the groups on covariates should be small. What is small?.. ↓ Sth. to decide!

Note: The original sample was 6000 girls in total (i.e., 4780 with $X = 0$ and 1220 with $X = 1$).

now it's about double, about 6000 in both groups

IPW estimate of ACE

Computing the ACE using inverse probability weighting by hand:

For individuals with $X_i = 1$: weighted average for treated group

An estimate of $E[Y_i^1]$ is $\frac{\sum_i X_i Y_i / \hat{\pi}_i}{\sum_i X_i / \hat{\pi}_i}$

For individuals with $X_i = 0$: weighted average for control group

An estimate of $E[Y_i^0]$ is $\frac{\sum_i (1-X_i) Y_i / (1-\hat{\pi}_i)}{\sum_i (1-X_i) / (1-\hat{\pi}_i)}$

The IPW estimate of the ACE is now: calculated the diff. → average causal effect (ACE)

$$\hat{ACE} = \frac{\sum_i X_i Y_i / \hat{\pi}_i}{\sum_i X_i / \hat{\pi}_i} - \frac{\sum_i (1-X_i) Y_i / (1-\hat{\pi}_i)}{\sum_i (1-X_i) / (1-\hat{\pi}_i)}$$

} now you have point estimate of ACE but no SD yet. & they are a bit tricky to get.

In practice with R: Compute the weights, and use the package survey (in exercises). Sensitive to outliers. → If you have a person w/ really high/low prob. of being treated, & they prolly are really rare, then this person will get a very large weight. & this might be a problem... this also again relates to

6. Stratification / Subclassification / Blocking

the positivity assumption. You don't actually want these ppl to be bare in that sense.

Stratification is also referred to as blocking or subclassification. Propensity scores can also be used to:

1) create strata (e.g., 5 strata of 20% scores each) : divide ppl in diff groups based on their propensity scores.
2) estimate the ACE in each stratum separately (e.g., with mean difference or regression): $\hat{\theta}_s$
3) combine the stratum-specific ACEs in an overall ACE: $\hat{ACE} = \sum_i \frac{N_s}{N} \hat{\theta}_s$

Strata should be **narrow enough**: but not too narrow that you have infinite amount of strata.

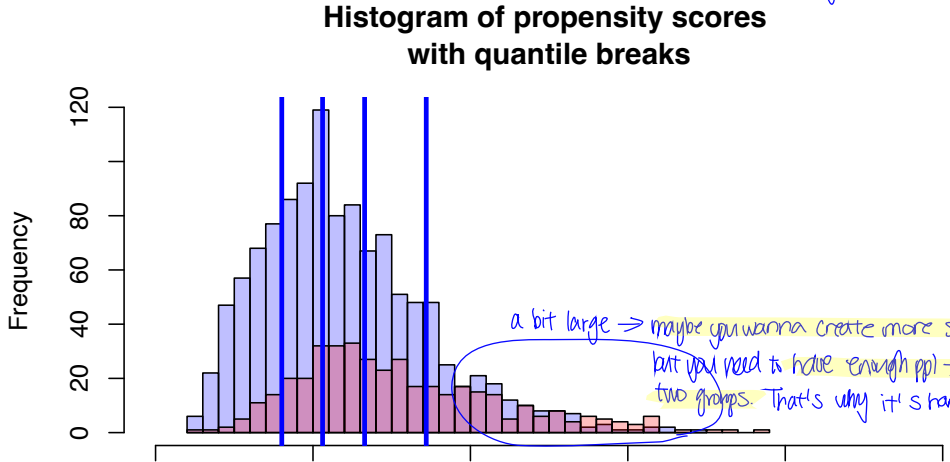
- within each stratum covariates do not make a difference; that is, it mimics an RCT
- Again, should be checked how succesful this was

Overlap

To allow for sensible comparisons within each stratum:

- there need to be treated and non-treated individuals in each stratum
- strata should not be too wide (otherwise make more strata)

Here again, there're lots of opportunities for Researchers' df. & capitalizing on chance on what you see in the data.



Controlling for Confounders: Assumptions

All the methods discussed require the identifiability assumptions: conditional exchangeability (no unobserved confounding), positivity, and consistency (and/or SUTVA).

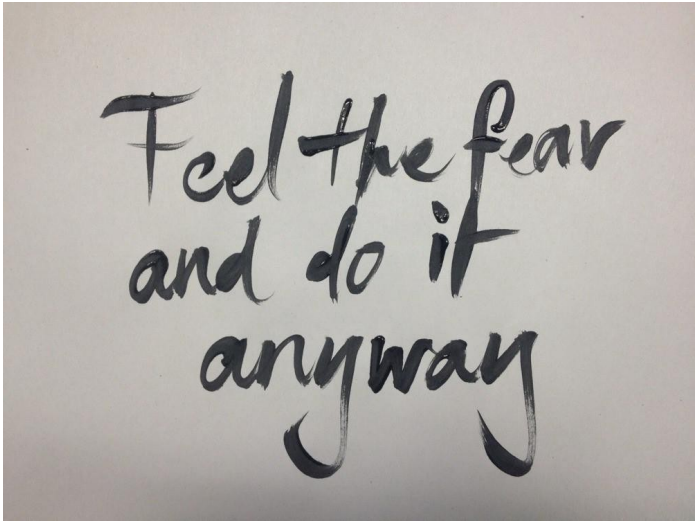
Additional method-specific assumptions:

- Prima facie effect (method 1) requires the assumption of no confounding at all - full exchangeability
- ANCOVA and regression estimation (methods 2 and 3) require correct specification of the outcome model
- matching, IPW, and stratification (methods 4, 5, and 6) require that (the use of) $\hat{\pi}_i$ balances the confounder distribution; it thus requires correct specification of the propensity score model
- dual-modeling strategies (methods 7, 8, and 9; not this course) require correct specification of either outcome model or propensity score model

Of course, it also requires to correctly identify confounders (rather than mediators, colliders).
Helpful if possible: Ensuring that the supposed confounders are measured prior to treatment.

so you are fairly sure that the arrows go from the confounders to treatments, not the other way around

Causal Inference: It's hard



Week 5: A Causal Inference Perspective on Methodological Issues

Causal Inference & Structural Equation Modeling

Noémi K. Schuurman
based on slides by Oisín Ryan and Ellen Hamaker

March 2022

Causal graphs and conceptual clarity

Causal models have very specific info. on what statistical result you might expect, but if you have certain statistical results, that does not contain all the info. we need on our causal operation... which means that we lose some info. if we only look at the statistics, e.g., means & cov. matrices... compared to if we knew the entire causal structure.

Causal models imply statistical models - causal models contain information not contained statistical models

- Many questions which are confusing, difficult or impossible to answer in purely statistical terms become clear when we take a causal inference perspective.
- I.e., Draw the relevant causal model!

Motto of Miguel Hernan: Draw your assumptions before your conclusions!

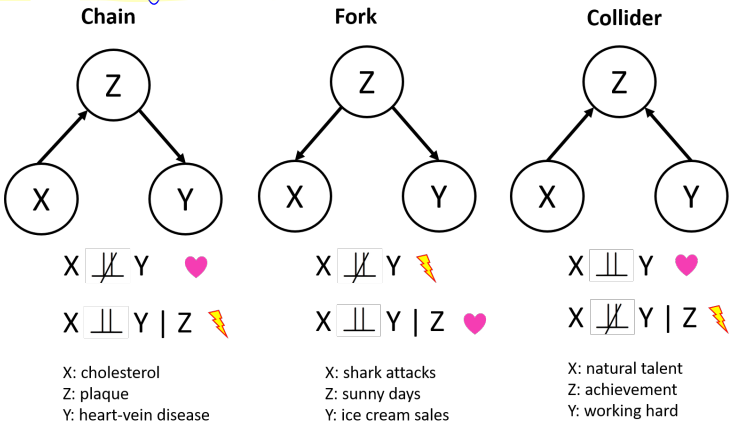
Overview

Interventions are more explicit in potential outcome framework.

- **DAGs & Interventions & the RCT**
- Selection Bias, Berksons Paradox, Simpsons Paradox ~ various paradoxes related to selection bias.
- Repeated Measures: Change Score vs Controlling for Pre-measure

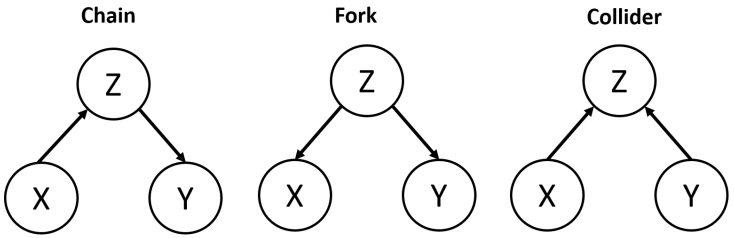
DAGs, SCMs, and interventions

Lecture 1: On which variables should we (not) condition to obtain the causal effect for observational data? ~w/o any interventions, what variable should we control for, based on a causal DAG?



DAGs, SCMs, and interventions

Lecture 1: On which variables should we (not) condition to obtain the causal effect for observational data?



Then how can we represent an intervention in a DAG?

- What if we intervened on the system represented by the DAG?

DAGs, SCMs, and interventions

Remember: Conditioning in observational data, and intervening are not the same actions.

another way of saying = statistical relationships are not the same as causal relationship!!

DAGs, SCMs, and interventions

Remember: Conditioning in observational data, and intervening are not the same actions.

- Conditioning on X, observational data: Let's look at people in our dataset that have X=1, and then at the people who have X=0.

DAGs, SCMs, and interventions

Remember: Conditioning in observational data, and intervening are not the same actions.

Here we don't know how they got that 1, there might have been some confounding involved...

- Conditioning on X, observational data: Let's look at people in our dataset that have X=1, and then at the people who have X=0.
- Intervening on X: Let's *set* these people's X to 1 (experimental group), and *set* these people's X to 0 (control group).

we have full control !! on which ppl have 1 & on which ppl have 0 → very impo difference!
that's why RCT is helpful bcuz it allows us to take control over values of a particular variable.

*Keep in mind they're diff! :) *

DAGs, SCMs, and interventions: The "do-operator"

To represent interventions in SCMs, we use the "do-operator":

do-operator: set X to a specific value

The do-operator $do(X = x)$ represents a "surgical intervention" to set the value of the variable X to a constant value x

DAGs, SCMs, and interventions: The "do-operator"

To represent interventions in SCMs, we use the "do-operator":

do-operator:

The do-operator $do(X = x)$ represents a "surgical intervention" to set the value of the variable X to a constant value x

"Surgical" interventions: Modularity assumption: Reminder of the DAG (causal structure) remains completely intact!

Assume that it is possible to intervene on a variable without fundamentally changing how it relates to other variables, e.g.:

- We can intervene on X without changing $p(Z \mid X)$
- We can intervene on one cause-effect mechanism without changing the others!

Read more about such assumptions by searching the jargon 'Modularity', 'Localized Interventions' and 'Fat Hand Interventions'. ⇝ affecting other variables by accident

Average Causal Effect - DAG edition

We can use the DAGs, SCM and the do-operator to define and estimate any causal effect based on an intervention we want:

Often we are interested in the effect of an intervention on the *mean* of our outcome variable - the effect of the intervention on average across different people.

Average causal effect of X (0 vs 1) on Y:

$$ACE = E[Y \mid do(X = x_1)] - E[Y \mid do(X = x_0)]$$

(expected value of Y given that we set X = x_1) − (exp. val. of Y given that we set X = x_0)

Average Causal Effect - DAG edition

We can use the DAGs, SCM and the do-operator to define and estimate any causal effect based on an intervention we want:

Often we are interested in the effect of an intervention on the *mean* of our outcome variable - the effect of the intervention on average across different people.

Average causal effect of X (0 vs 1) on Y:

$$ACE = E[Y \mid do(X = x_1)] - E[Y \mid do(X = x_0)]$$

Here it's more implicit in the do-operator... potential outcome framework, there it was very explicit what all

Note: $ACE = E[Y \mid do(X = x_1)] - E[Y \mid do(X = x_0)] = E[Y_i^1] - E[Y_i^0]$ of these assumptions are.

The idea is that the expression w/ "do-operator" is the same as the expression w/ the potential outcomes we had last week → ofc, that means that you have to adhere to "Essentially your intervention is like a successful RCT, that's the idea behind do-operator..." all of these assumptions we specified last week.

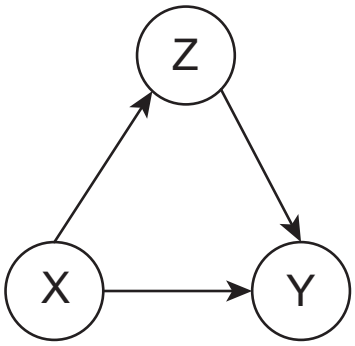
Interventions with DAGs and SCMs: Partial Mediation Example

In observational dataset, this case we don't wanna control for Z, cuz we're interested in a total effect of X on Y.

The observational DAG and SCM:

$$
\begin{aligned}
X &: \ bernoulli(0.5) \\
Z &:= 2X + \epsilon_Z \\
Y &:= 1X + 2Z + \epsilon_Y
\end{aligned}
$$

how they're generated



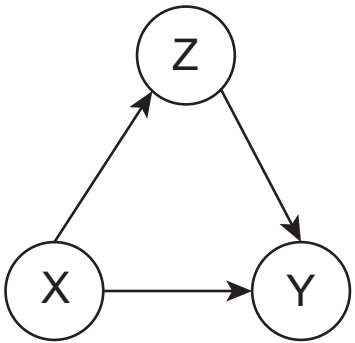
where

- X is bernoulli distributed (0 or 1) with probability 0.5, ϵ_Z, ϵ_Y are iid, $\sim \mathcal{N}(0, 1)$

Interventions with DAGs and SCMs: Partial Mediation Example

The observational DAG and SCM:

$$
\begin{aligned}
X &: \ bernoulli(0.5) \\
Z &:= 2X + \epsilon_Z \\
Y &:= 1X + 2Z + \epsilon_Y
\end{aligned}
$$



where

- X is bernoulli distributed (0 or 1) with probability 0.5, ϵ_Z, ϵ_Y are iid, $\sim \mathcal{N}(0,1)$

"Naive Causal Estimate"

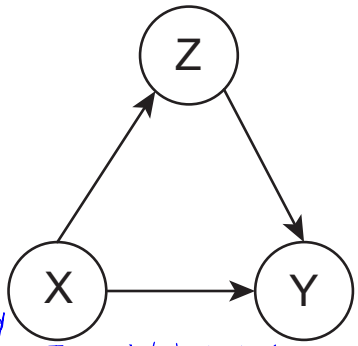
Prima Facie Effect: $E[Y \mid X = 1] - E[Y \mid X = 0] =$

In this case, you already know P.F effect is the same as causal effect bcuz there's no confounders or anything...

Interventions with DAGs and SCMs: Partial Mediation Example

The observational DAG and SCM:

$$X: \ bernoulli(0.5)$$
$$Z := 2X + \epsilon_Z$$
$$Y := 1X + 2Z + \epsilon_Y$$



where

- X is bernoulli distributed (0 or 1) with probability 0.5, ϵ_Z, ϵ_Y are iid, $\sim \mathcal{N}(0,1)$

For ppl w/ X=1, what expected value do they have for Y?

Then again look at ppl w/ X=0, and see the expected value of Y

Prima Facie Effect: $E[Y \mid X = 1] - E[Y \mid X = 0] =$

$5 - 0 = 5$ and this is combination of indirect path & direct path

Interventions with DAGs and SCMs: Partial Mediation Example

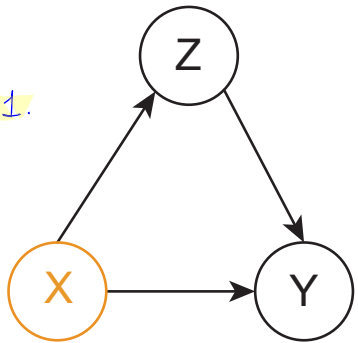
SCMs model with intervention "$do(X = 1)$": ~ meaning the SCM for X changes

$$X := 1$$
$$Z := 2X + \epsilon_Z$$
$$Y := 1X + 2Z + \epsilon_Y$$

No more probabilities involved any longer. We just fix it to be 1.

where

- ϵ_Z, ϵ_Y are iid, $\sim \mathcal{N}(0, 1)$



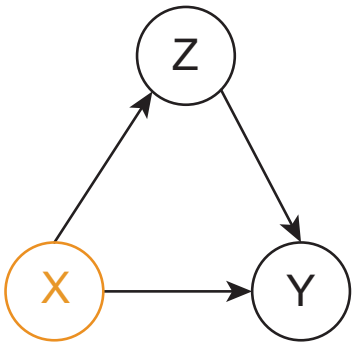
Interventions with DAGs and SCMs: Partial Mediation Example

SCMs model with intervention $do(X = 1)$.

$$
\begin{aligned}
X &:= 1 \\
Z &:= 2X + \epsilon_Z \\
Y &:= 1X + 2Z + \epsilon_Y
\end{aligned}
$$

where

- ϵ_Z, ϵ_Y are iid, $\sim \mathcal{N}(0, 1)$



Average Causal Effect: $E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] =$

Interventions with DAGs and SCMs: Partial Mediation Example

SCMs model with intervention $do(X = 1)$.

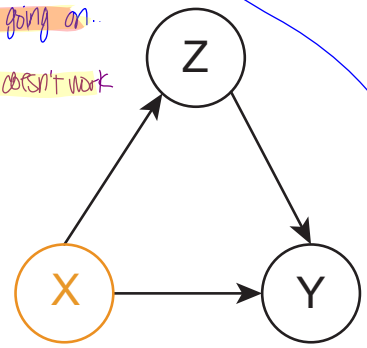
In this case that happens, because Primal Fact effect is a good estimator of true causal effect, because we don't have confounding going on.

$$X := 1$$
$$Z := 2X + \epsilon_Z$$
$$Y := 1X + 2Z + \epsilon_Y$$

It works here, but it doesn't work in general...

where

- ϵ_Z, ϵ_Y are iid, $\sim \mathcal{N}(0, 1)$



Average Causal Effect: $E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] =$
$5 - 0 = 5$ same thing as before.

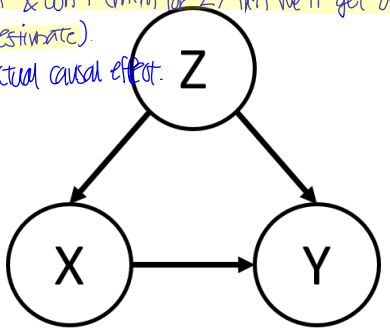
Interventions with DAGs and SCMs: "Partial Confounding" Example

Another observational DAG and SCM:

Now we have a situation w/ confounding! → If we just look at the effect of X on Y & don't control for Z, then we'll get a wrong estimate. (biased estimate)
You won't get an actual causal effect.

logistic reg.

$$logodds(X) := 2Z$$
$$Z := \epsilon_Z$$
$$Y := 1X + 2Z + \epsilon_Y$$



where

- X can be 0 or 1 and follows a logistic regression model, ϵ_Z, ϵ_Y are iid, $\sim \mathcal{N}(0, 1)$

Interventions with DAGs and SCMs: Partial Confounding Example

Another observational DAG and SCM:

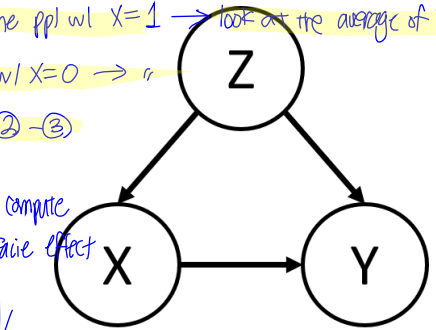
$$
\begin{aligned}
logodds(X) &:= 2Z \\
Z &:= \epsilon_Z \\
Y &:= 1X + 2Z + \epsilon_Y
\end{aligned}
$$

where

- X can be 0 or 1 and follows a logistic regression model, ϵ_Z, ϵ_Y are iid, $\sim \mathcal{N}(0,1)$

① simulate the data according to SCM
② look at the ppl w/ X=1 → look at the average of Y in that group
③ " w/ X=0 → "
④ subtract ② − ③

How to compute Prima Facie effect



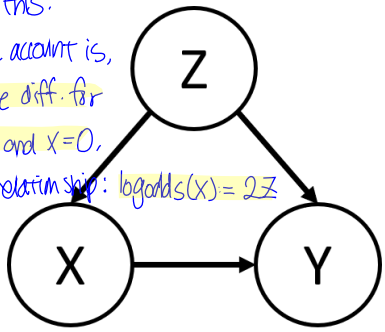
Prima Facie Effect: $E[Y \mid X = 1] - E[Y \mid X = 0] \sim$ appx. 4

Interventions with DAGs and SCMs: Partial Confounding Example

Another observational DAG and SCM:

$$
\begin{aligned}
logodds(X) &:= 2Z \\
Z &:= \epsilon_Z \\
Y &:= 1X + 2Z + \epsilon_Y
\end{aligned}
$$

you can also prove this:
what you need to take account is, the Z value would be diff. for ppl that have X=1 and X=0, becuz of the first relation ship: logodds(X) = 2Z



where

- X can be 0 or 1 and follows a logistic regression model, ϵ_Z, ϵ_Y are iid, $\sim \mathcal{N}(0,1)$

Prima Facie Effect: $E[Y \mid X = 1] - E[Y \mid X = 0] \sim$ → biased effect would be 4
$\sim 2.5 - (-1.5) = 4$

If someone proves this exactly I'll treat the class to boterkoek in our next meeting
(for the approximation via simulation see rcode on bb).

Interventions with DAGs and SCMs: Example 2

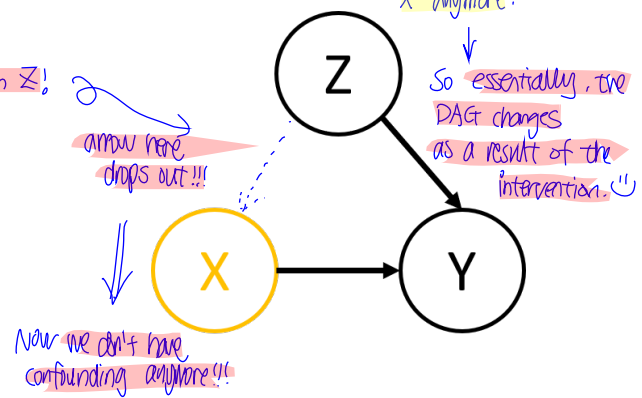
SCMs model the intervention $do(X = 1)$.

Note: X is now no longer affected by Z. We fully control the value of X.

$$X := 1$$
$$Z := \epsilon_Z$$
$$Y := 1X + 2Z + \epsilon_Y$$

where

- ϵ_Z, ϵ_Y are iid, $\sim \mathcal{N}(0, 1)$



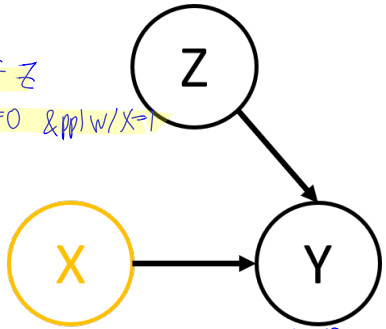
Interventions with DAGs and SCMs: Example 2

SCMs model the intervention $do(X = 1)$.

Note: X is now no longer affected by Z. We fully control the value of X.

$$X := 1$$
$$Z := \epsilon_Z$$
$$Y := 1X + 2Z + \epsilon_Y$$

In this situation, the expected value of Z does not differ between ppl w/ X=0 & ppl w/ X=1



where

- ϵ_Z, ϵ_Y are iid, $\sim \mathcal{N}(0, 1)$

$1 + E(2Z) - E(2Z) = 1$ actual causal effect = 1

Average Causal Effect: $E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] =$

in prima facie, it was 4.

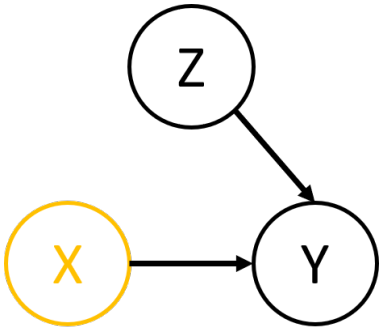
Interventions with DAGs and SCMs: Example 2

SCMs model the intervention $do(X = 1)$.
Note: X is now no longer affected by Z. We fully control the value of X.

$$
\begin{aligned}
X &:= 1 \\
Z &:= \epsilon_{\not Z} \\
Y &:= 1X + 2Z + \epsilon_Y
\end{aligned}
$$

where

- $\epsilon_{\not Z}, \epsilon_Y$ are iid, $\sim \mathcal{N}(0, 1)$



Average Causal Effect: $E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] =$
$1 - 0 = 1$, while in the observational case, we saw 4.

Interventions vs Conditioning

The point is again, observing is not the same as intervening!

Observing/Seeing $\neq$ Intervening/Doing:

$E[Y \mid A = a]$ is *not* necessarily the same as $E[Y \mid do(A = a)]$

If there's difference between the two, then there's confounding going on.

When statistical relationship $\neq$ causal effect, we say the former is *confounded*.

This is just another way of saying what confounding is!

Example 1 - partial mediation:

- $E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] = 5 - 0 = 5$
- $E[Y \mid X = 1] - E[Y \mid X = 0] = 2.5 - 0 = 5$

Interventions vs Conditioning

Observing/Seeing $\neq$ Intervening/Doing:

$E[Y \mid A = a]$ is *not* necessarily the same as $E[Y \mid do(A = a)]$

When statistical relationship $\neq$ causal effect, we say the former is *confounded*.

Example 1 - partial mediation:

- $E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] = 5 - 0 = 5$
- $E[Y \mid X = 1] - E[Y \mid X = 0] = 2.5 - 0 = 5$

) there was no diff.

Example 2 - partial confounding:

- $E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] = 1 - 0 = 1$
- $E[Y \mid X = 1] - E[Y \mid X = 0] \sim 2.5 - (-1.5) = 4$

) diff≠0 → there's confounding!

Interventions vs Conditioning

Observing/Seeing $\neq$ Intervening/Doing:

$E[Y \mid A = a]$ is *not* necessarily the same as $E[Y \mid do(A = a)]$

When statistical relationship $\neq$ causal effect, we say the former is *confounded*.

Example 1 - partial mediation:

- $E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] = 5 - 0 = 5$
- $E[Y \mid X = 1] - E[Y \mid X = 0] = 2.5 - 0 = 5$

Example 2 - partial confounding:

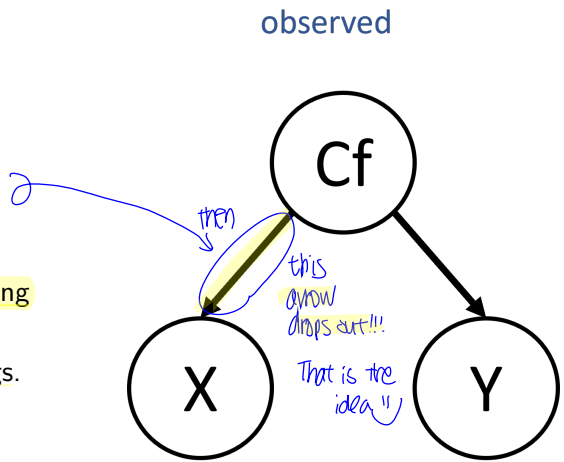
- $E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] = 1 - 0 = 1$
- $E[Y \mid X = 1] - E[Y \mid X = 0] \sim 2.5 - (-1.5) = 4$

Observationally, people with X=1 will have higher expected values for Z than people with X=0 ($\sim$ .8 vs -.8), so the mean of y will also be higher for the former!

Randomized Control Trials - Why They Work

RCTs are extremely powerful because randomization ensures no confounding.

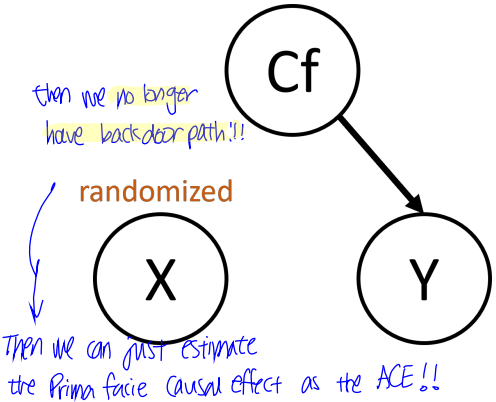
- Randomization (if succesfull) means having full control over the treatment variable.
- There can't be any backdoor paths if everyone has an equal probability of being treated or not
- But, RCTs not possible in many settings.



Randomized Control Trials - Why They Work

RCTs are extremely powerful because randomization ensures no confounding.

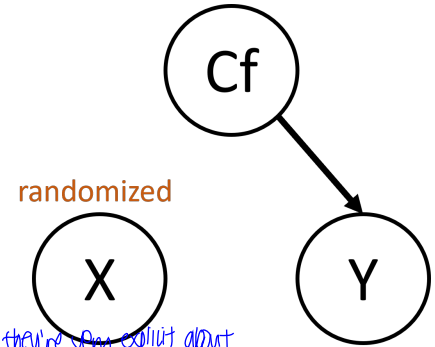
- Randomization (if succesfull) means having full control over the treatment variable.
- There can't be any backdoor paths if everyone has an equal probability of being treated or not
- But, RCTs not possible in many settings.



Randomized Control Trials - Why They Work

RCTs are extremely powerful because randomization ensures no confounding.

- Randomization (if succesfull) means having full control over the treatment variable.
- There can't be any backdoor paths if everyone has an equal probability of being treated or not
- But, RCTs not possible in many settings.



they're very explicit about what are! & what assumptions are required

Note: From the "Potential Outcomes perspective", (succesful) RCTs work because they by design adhere to all the assumptions (e.g., exchangeability, positivity, etc).

Simpsons Paradox

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Changes the sign or close to zero...
when we look at the specific part of population,
the association changes in some ways..

≈ "marginal relationships are not the same as conditional relationships!!"
And it also help thinking about what the underlying DAG is"

Simpsons Paradox

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Example (Pearl, Glymour & Jewell, 2016): no randomization

- 700 sick patients are given the choice to take a new drug: 350 "choose to" take it.
- We are interested in effects of a drug (D) on recovery (R). We also record the gender (G)
- Should we prescribe the drug?

Simpsons Paradox

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Example (Pearl, Glymour & Jewell, 2016):

- 700 sick patients are given the choice to take a new drug: 350 choose to take it.
- We are interested in effects of a drug (D) on recovery (R). We also record the gender (G)
- Should we prescribe the drug? <statistical result ↓>

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug		No drug
Men	81 out of 87 recovered (93%)	>	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	>	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	<	289 out of 350 recovered (83%)

Taking drug is a good idea!

Taking drug is NOT good?!

combined data shows the "opposite" result !!

Overview

- DAGs & Interventions & the RCT
- **Selection Bias, Berksons Paradox, Simpsons Paradox**
- Repeated Measures: Change Scores vs Controlling for Pre-measure

Simpsons Paradox

Counter-intuitive, but not really a paradox ~ it's completely in line w/ all the probability rules and maathrot.

- A marginal dependency ($P(R \mid D)$) is not necessarily the same as a conditional dependency ($P(R \mid D, G = 0)$)
- But which piece of information should we use to make treatment decisions?
- (Who) should we treat?!

But it still does not really help us to decide what should we do then.... which of these dependencies are relevant..

marginal dependency (where we don't control for Gender) is $\neq$ conditional dependency (where we do control for Gender)

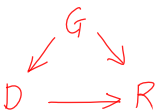
Simpsons Paradox

Counter-intuitive, but not really a paradox

- A marginal dependency ($P(R \mid D)$) is not necessarily the same as a conditional dependency ($P(R \mid D, G = 0)$)
- But which piece of information should we use to make treatment decisions?
- (Who) should we treat?!

Draw your DAG! based on 3 variables

Variables: Gender (sex; 'G'), Drug ('D') and Recovery ('R')



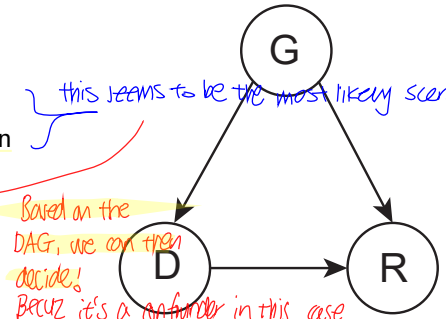
Simpsons Paradox

- Estrogen levels negatively affect recovery
- Women are more likely to take the drug than men
- We should condition on Gender - it blocks a backdoor path!

this seems to be the most likely scenario.

Yes,

Based on the DAG, we can then decide!
Because it's a confounder in this case



And then, the conclusion is: we do give drugs, because it helps recovering.

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Simpsons Paradox

Suppose that we measured post-treatment blood pressure (B) instead of gender, next to Drug taking (D) and Recovery (R).

Draw your DAG!

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	Drug		No drug
Low BP	81 out of 87 recovered (93%)	>	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	>	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	<	289 out of 350 recovered (83%)

Simpsons Paradox

Suppose that we measure post-treatment blood pressure (B) instead

- Statistical information is exactly the same!!
- B cannot cause drug taking
- The drug works in part by decreasing blood pressure
- We should **not** condition on blood pressure and in this case, that means we do NOT give the drugs!

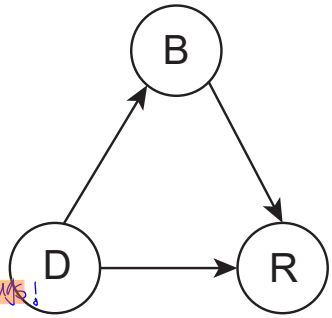
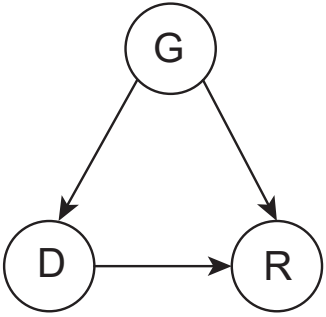


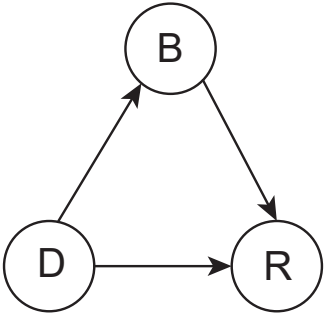
Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	Drug	No drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Simpsons Paradox

- Statistical information alone cannot provide the answer on when to treat (!)
- Two different DAGs can produce the exact same statistical dependencies in the observational setting
  - "Observationally equivalent"
- These DAGs imply different causal effects, and hence different models to estimate those effects from observational data.





Selection Bias

- when we condition on a specific sub-group, for ex. by doing some kinds of selection, for ex, we only look at student population... Usually it's a conditioning situation becuz of how you sampled. sampling only patients, students...etc → then we see the diff. between what we observe in the general population and what we had in a specific sub-population...

Berksons Paradox

Two phenomena which are statistically *independent* in the general population are statistically *dependent* in a sub-population that was selected.

Also know as: Selection Bias, Endogenous Selection Bias, Berskon's bias

Classic example: We are interested in the relationship between *Lung Cancer* (L) and *Diabetes* (D)

- General population, these two variables are independent.
- In a sample of *hospital patients*, there is a negative dependency - patients who don't have diabetes are *more likely* to have lung cancer.

Selection Bias

Berksons Paradox

Two phenomena which are statistically *independent* in the general population are statistically *dependent* in a sub-population that was selected.

Also know as: Selection Bias, Endogenous Selection Bias, Berskon's bias

Classic example: We are interested in the relationship between *Lung Cancer* (L) and *Diabetes* (D)

- General population, these two variables are independent.
- In a sample of *hospital patients*, there is a negative dependency - patients who don't have diabetes are *more likely* to have lung cancer.

Again, to figure out what's going on here:

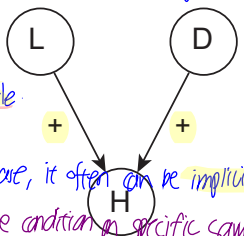
...Draw your DAG!

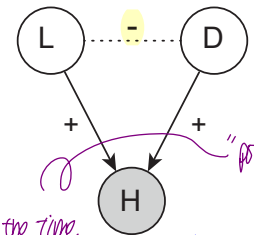
Selection Bias : we get the wrong result bcuz collecting patient samples, we condition on hospitalization.

ppl often don't think about it

So, in Simpson's paradox, it's very explicit. we just include a certain variable in the analysis or we don't.

But in Berksons paradox, we might have accidentally condition on some kind of variable & you might not be aware of this problem.... In Berksons case, it often can be implicit.





"potential Berksons paradox"

seems the most likely scenario

EX) very clear example where we condition on specific samples all the time = student population!!! we do so many experiments exclusively on students

- Lung cancer L and diabetes D cause hospitalization H
- By selecting participants from a hospital we *condition* on hospitalization ($H = 1$)
- If you are hospitalised, and you *don't* have diabetes, probably you do have lung cancer (Otherwise - why would you be in hospital?).
- $P(D|L = 1, H = 1) \neq P(D|L = 1) \neq P(D|do(L) = 1)$
- We have conditioned on a *collider*! ⇒ This would explain why we get a diff. result in the full population

$P(D|L=1, H=1) \neq P(D|L=1)$

$P(D|L=1, H=1) \neq P(D|do(L)=1)$

conditional dependencies do not have to be equal to marginal dependencies

Simpsons or Berksons or?

these & dependencies do not equal to causal dependencies

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Berksons Paradox ~typically mentioned in context of (pre-) selection idea

Two phenomena which are statistically *independent* in the general population are statistically *dependent* in a sub-population that was selected.

* What's the difference?

- In Simpsons, we find different relations when we control vs not control for a variable
- In Berksons, we find different relations as a result of 'accidental' selection via our sampling procedure.
- Either can be the result of collider bias or confounder bias or overcontrol bias (controlling for a mediator)

Berksons Simpsons

point is* ⇒ conditioning on a variable either by including in your analysis somehow, or by accidentally selecting a particular sample, might be problematic. It can be because of collider bias / confounder bias / overcontrol bias...

And to clear out what's going on → Draw A → G * Think about a causal model. (It's not so impo. whether

Simpsons or Berksons or?

you call it a Simpsons / Berksons paradox)

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Berksons Paradox

Two phenomena which are statistically *independent* in the general population are statistically *dependent* in a sub-population that was selected.

What's the difference?

- In Simpsons, we find different relations when we control vs not control for a variable
- In Berksons, we find different relations as a result of 'accidental' selection via our sampling procedure.
- Either can be the result of collider bias or confounder bias or overcontrol bias (controlling for a mediator).

Note 1. When we have a partial mediation where (for example) the direct effect is positive and the indirect effect is negative (or other way around), so the total effect is near zero - This is called a 'supression effect'.

↳ keep in mind: time-order of variables, & what the actual possible DAG might be.
cuz remember, "mediation & confounding" → we cannot tell them apart !!

Simpsons or Berksons or?

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Berksons Paradox

Two phenomena which are statistically *independent* in the general population are statistically *dependent* in a sub-population that was selected.

What's the difference?

- In Simpsons, we find different relations when we control vs not control for a variable
- In Berksons, we find different relations as a result of 'accidental' selection via our sampling procedure.
- Either can be the result of collider bias or confounder bias or overcontrol bias (controlling for a mediator).

Note 1. When we have a partial mediation where (for example) the direct effect is positive and the indirect effect is negative (or other way around), so the total effect is near zero - This is called a 'supression effect'.

Note 2. Some people relate Simpsons expressedly to confounding bias and Berksons to collider bias.

In any case...draw the causal model!

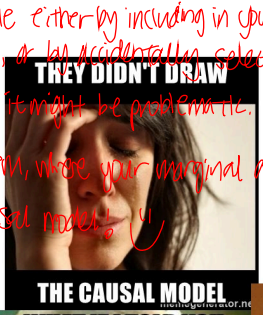
~Always a good idea!" Be very explicit about the mechanism that your summarizing equations or w/ drawing...

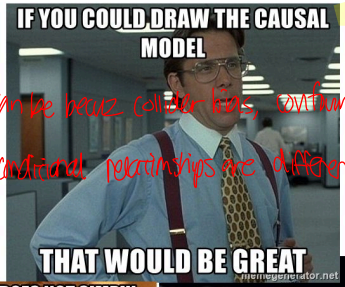
Conditioning on a variable either by including in your analysis in some ways, or by accidentally selecting a particular sample, it might be problematic. It can be because collider bias, confounder bias, or overcontrol bias.

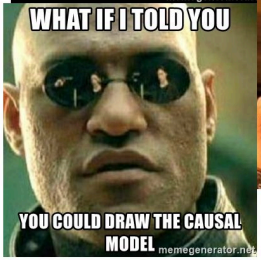
So

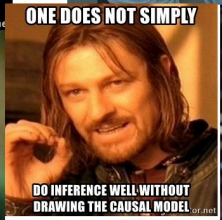
If you have this pattern, where your marginal and conditional relationships are different, Draw a DAG!

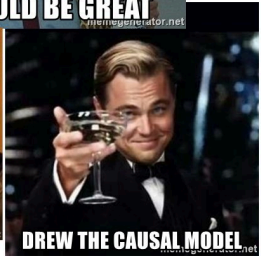
Think about the causal model! :)











Recap: Where are we?

We are interested in **the effect of treatment X on outcome Y**:

- What is the effect of dieting on psychological well-being (Schafer & Kang, 2008)?
- What is the effect of out-of-home-placement on children's well-being (Berger et al., 2009)?
- What is the effect of physical punishment of children's behavioral problems (Larzelere et al., 2010)?
- What is the effect of extra schooling on social economic status?

If we have only **observational data** for this, we should:

- be concerned about **confounding**, collider bias, overcontrol bias
- **including the right covariates** to account for this
- we can **use DAGs** to see what we should control for (i.e., condition on)

But a DAG is of course only as good as our theory is...

What covariates should be included?

Steiner, Cook, Shadish and Clark (2010) used a creative design:

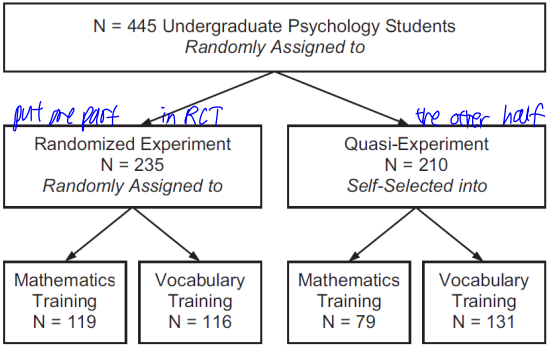


Figure 1. Overall design of the within-study comparison of the randomized experiment and the quasi-experiment.

This allows for a comparison of RCT results (true ACE) and observational results.

↳ How well the obs. study is recovering the true causal effect, assuming RCT was done properly

Results from Steiner er al. (2010)

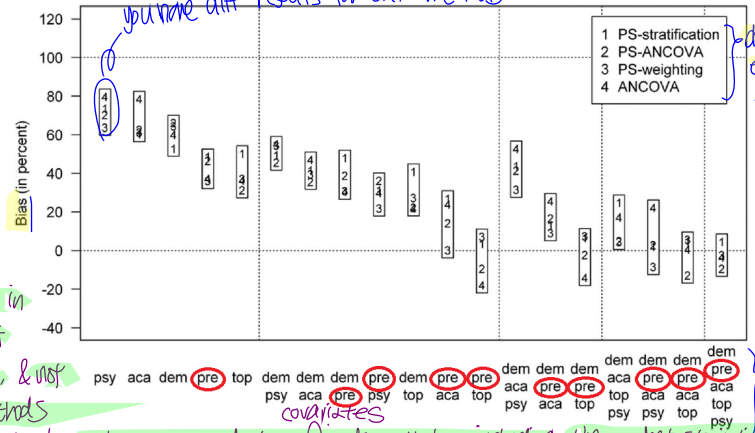


Figure 2. Remaining bias in vocabulary by construct set and analytic method (in the order of average bias). Construct sets are composed of demographics (dem), proxy-pretests (pre), prior academic achievement (aca), topic preference (top), and psychological predisposition (psy). PS = propensity score; ANCOVA = analysis of covariance.

- What they also found:

Overall, pre-test measure seems a valuable covariate to include.

Overview

ofc. you can already make a DAG in your mind & see when this might or might not apply! That's what we're gonna do now... :)

- DAGs & Interventions & the RCT
- Selection Bias, Berksons Paradox, Simpsons Paradox
- **Repeated Measures: Change Scores vs Controlling for Pre-measure**

Pre-Post Designs

Pre-post test designs: when the outcome is measured twice

- Lord's paradox
- ANCOVA vs. change score analysis
- Five scenarios
- How DAGs can help (Pearl, 2016)
- Unmeasured confounders (Kim & Steiner, 2019)

Lord's paradox

Comparing non-randomly assigned groups

Psychological Bulletin
1967, Vol. 68, No. 5, 304–305

A PARADOX IN THE INTERPRETATION OF GROUP COMPARISONS

FREDERIC M. LORD

Educational Testing Service

Attention is called to a basic source of confusion in the interpretation of certain types of group comparison data.

Comparing non-randomly assigned groups

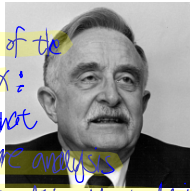
Psychological Bulletin
1967, Vol. 68, No. 5, 304–305

A PARADOX IN THE INTERPRETATION OF GROUP COMPARISONS

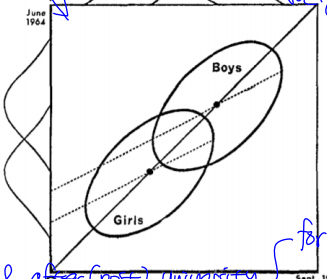
FREDERIC M. LORD

Educational Testing Service

Attention is called to a basic source of confusion in the interpretation of certain types of group comparison data.



visualization of the entire paradox: visualizing that the change score analysis is diff. from ANCOVA analysis where we control for pre-treatment weight.

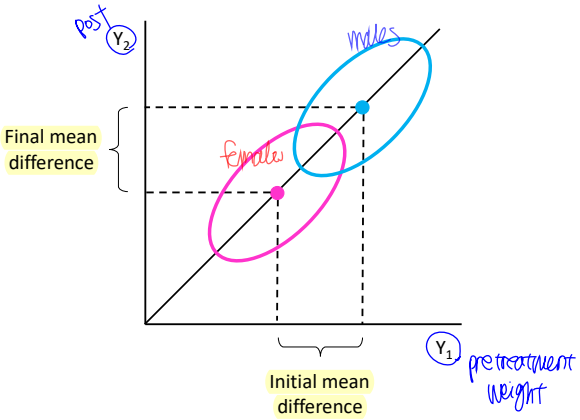


"A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex difference in these effects." (p.304)

treatment var. = diet

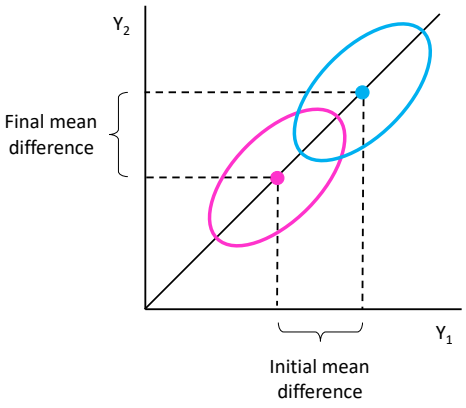
measure the weight before (pre) & after (post) university for each gender

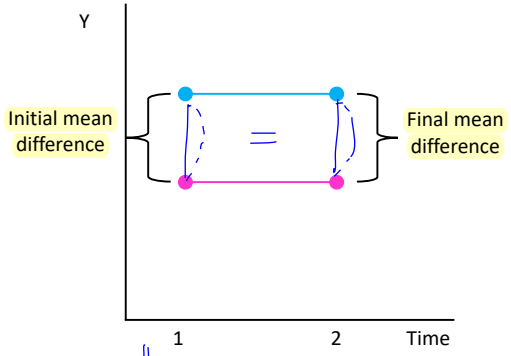
Statistician 1: Looks at the difference in the differences



$\Rightarrow$ Diff. in pre-weight between males & females is the same as in post-weight difference!

Statistician 1: Looks at the difference in the differences



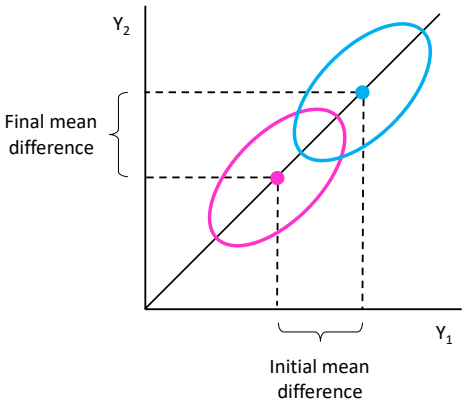


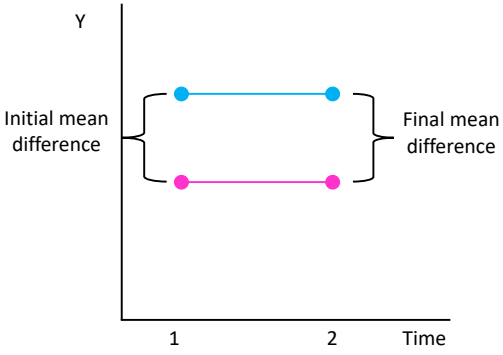
Diff. between boys & girls at Time 1

=

Time2

Statistician 1: Looks at the difference in the differences



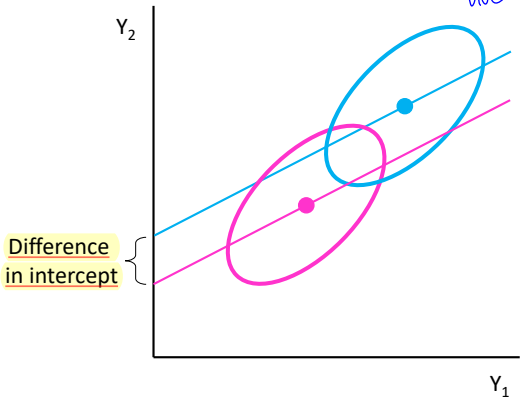


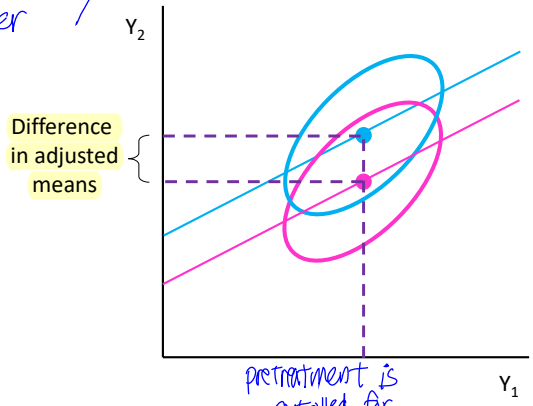
Conclusion: No change per group, so no difference in their change either. ⇒ University diet doesn't really matter...

Or **more generally**: There is no difference over time in the differences between the groups.

Statistician 2: Uses ANCOVA - Pre-weight as control variable

use & gender as covariates

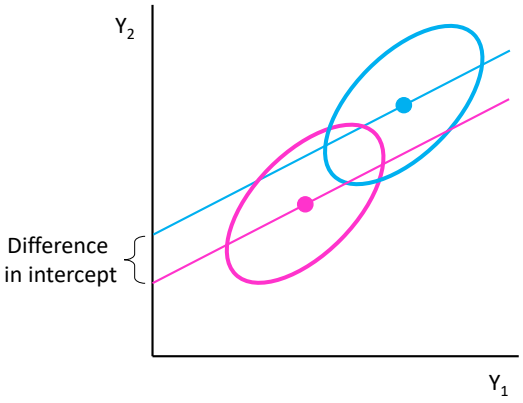


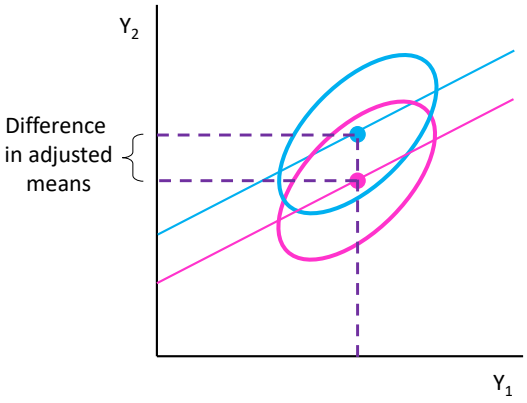


If we were to compare boys & girls that have the same pre-weights, then the adj mean in post-weights differ!!

now we DO see differences between them!

Statistician 2: Uses ANCOVA - Pre-weight as control variable





Conclusion: When comparing a boy and girl of **equal weight to begin with**, the boy **tends to weigh more afterwards** than the girl; hence, there is a difference, after correcting for initial differences.

Sex differences in effect of dining hall diet

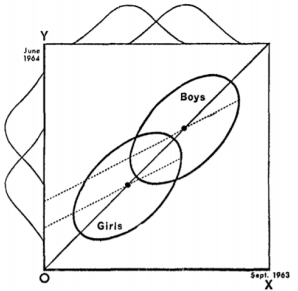


FIG. 1. Hypothetical scatterplots showing initial and final weight for boys and for girls.

Statistician 1: No difference Statistician 2: Boys gain more than girls

So in essence, he also is saying that statistics is not sufficient to decide what the correct approach is.

Lord's conclusion (p.305, 1967):

"The researcher wants to know how the groups would have compared if there had been no preexisting uncontrolled differences. The usual research study of this type is **attempting to answer a question that simply cannot be answered in any rigorous way on the basis of available data**."

Lord's paradox in empirical research

Larzelere et al. (2010) studied the effect of **corrective actions** on **problem behaviors** of 1,464 children aged 4 and 5.

Corrective action	ANCOVA result	Change score result
	Antisocial behavior	
Professional interventions		
Psychotherapy visits	.07**	.00
Ritalin	.07**	.04
Parental disciplinary actions		
Non-physical punishment	.03	−.08**
Physical punishment	.07**	−.05
Scolding/yelling	.06*	−.08**
"Hostile/ineffective" scale	.09**	−.15**
	Hyperactivity	
Professional interventions		
Psychotherapy visits	.03	−.02
Ritalin	.05*	−.00
Parental disciplinary actions		
Non-physical punishment	.07**	−.01
Physical punishment	.03	−.01
Scolding/yelling	.04*	−.05
"Hostile/ineffective" scale	.09**	−.08**

you see these arise very diff patterns.

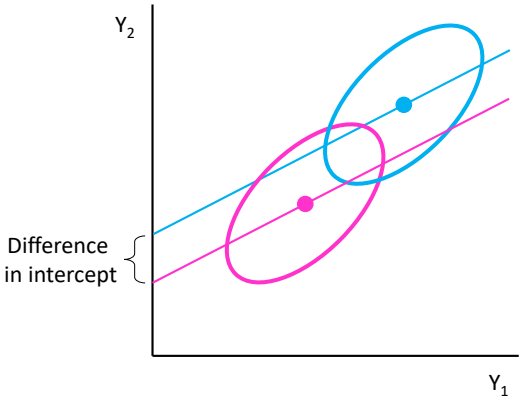
In ANCOVA context, "disciplinary action" seems to increase antisocial behavior.
In change score, it decreases antisocial behavior. similar w/hyperactivity.

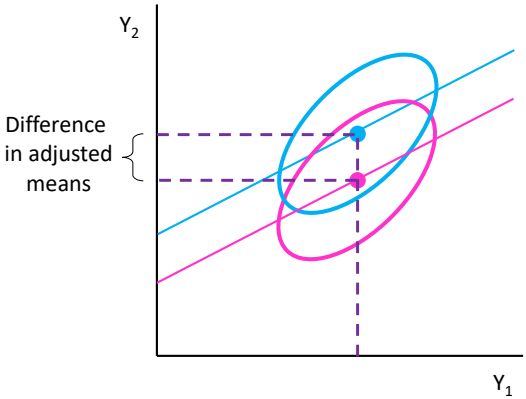
⇒ Not figuring out what result is correct, it'll imply very different practical advice to parents!!!

ANCOVA vs. Change score model

ANCOVA (popular in psychology)

Analysis of covariance (ANCOVA) in a pre-posttest design is based on including the pretest as covariate.





ACE based on ANCOVA model

The **causal effect** β_1 in the ANCOVA model is the **difference between treated and untreated** persons **with identical values on covariate** (i.e. the pretest). We can write this as the **expected difference in potential outcomes**:

what the ACE would be in ANCOVA context

$$
\begin{aligned}
ACE_{ANCOVA} &= E\left[Y_2^1\right] - E\left[Y_2^0\right] \\
&= E\left[Y_2^1 | X = 1, Y_1\right] - E\left[Y_2^0 | X = 0, Y_1\right] && \text{No unobserved confounding} \\
&= E\left[Y_2 | X = 1, Y_1\right] - E\left[Y_2 | X = 0, Y_1\right] && \text{Consistency} \\
&= \left(\beta_0 + \beta_1 + \beta_2 Y_1\right) - \left(\beta_0 + \beta_2 Y_1\right) && \text{Correct model} \\
&= \beta_1
\end{aligned}
$$

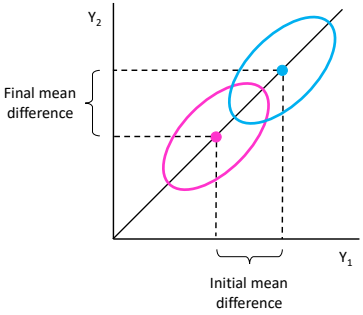
expected value of post. result controlling for pretreatment value for treated group − " for non-treated group

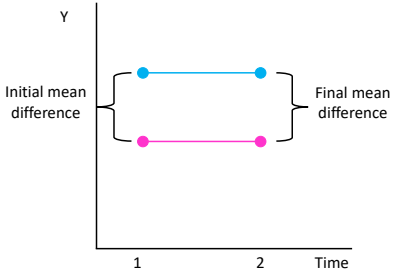
here $X = 1$ here $X = 0 \rightarrow \beta_1 \cdot 0$

β_1 ⇓ our estimate of effect of treatment X: just the regression coefficient of our treatment variable is our estimate of causal effect.

Change score model (popular in econometrics)

The change score model (or gain score model) is based on investigating difference-in-differences.





Change score model:

$$Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + \epsilon_{2i}$$

~ regress only on the treatment variable only control for X: the treatment of interest

where γ_1 is interpreted as the causal effect of X.

ACE based on Change score model

Let $G_i = Y_{2i} - Y_{1i}$ represent a person's **gain score** (aka change or difference score).

Then the **average causal effect** can be expressed as the expected difference in potential outcomes of G_i:

$$
\begin{aligned}
ACE_{CS} &= E\big[G^1\big] - E\big[G^0\big] \\
&= E\big[G^1|X=1\big] - E\big[G^0|X=0\big] && \text{No unobserved confounding} \\
&= E\big[G|X=1\big] - E\big[G|X=0\big] && \text{Consistency} \\
&= E\big[\gamma_0 + \gamma_1\big] - E\big[\gamma_0\big] && \text{Correct model} \\
&= \gamma_1
\end{aligned}
$$

Alternative expression of ACE_{CS}

Instead of expressing the **ACE** of the changes score model in terms of the regression parameters, we can also express it in terms of the **pre- and post-test means**, that is:

$$
\begin{aligned}
ACE_{CS} &= E\big[G^1\big] - E\big[G^0\big] \\
&= E\big[G^1|X=1\big] - E\big[G^0|X=0\big] && \text{No unobserved confounding} \\
&= E\big[G|X=1\big] - E\big[G|X=0\big] && \text{Consistency} \\
&= E\big[\{Y_2 - Y_1\}|X=1\big] - E\big[\{Y_2 - Y_1\}|X=0\big] \\
&= \big(E\big[Y_2|X=1\big] - E\big[Y_1|X=1\big]\big) - \big(E\big[Y_2|X=0\big] - E\big[Y_1|X=0\big]\big)
\end{aligned}
$$

ACE_{CS} as difference-in-differences

Thus we have

$$ACE_{CS} = \big(E[Y_2|X=1] - E[Y_1|X=1]\big) - \big(E[Y_2|X=0] - E[Y_1|X=0]\big)$$

diff. over time in group 1 — diff. over time in group 0

that is, the **ACE is equal to the difference between groups in their gain scores**.

Alternatively, we can write

$$ACE_{CS} = \big(E[Y_2|X=1] - (E[Y_2|X=0]\big) - \big(E[Y_1|X=1] - E[Y_1|X=0]\big)$$

group diff. at time 2 — group diff. at time 1

This is visually seen in change score model

that is, the ACE is equal to the difference over time in the difference between the groups (**difference-in-differences**).

Conclusion so far

With the **ANCOVA model** we answer the question: If the groups had been equal on the pre-test, would we observe a difference between them on the post-test?

$Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_{2i}$ If so (i.e., $\beta_1 \neq 0$), we conclude **treatment has an effect**.

With the **CS model**, we answer the question: Is the change over time different for the two groups?

$Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + \epsilon_i$ If so (i.e., $\gamma_1 \neq 0$), we conclude **treatment has an effect**.

It's impo. to realize that these are in essence diff. questions, which in result might have diff. answers.

Conclusion so far

With the **ANCOVA model** we answer the question: If the groups had been equal on the pre-test, would we observe a difference between them on the post-test?

$Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_{2i}$ If so (i.e., $\beta_1 \neq 0$), we conclude **treatment has an effect**.

With the **CS model**, we answer the question: Is the change over time different for the two groups?

$Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + \epsilon_i$ If so (i.e., $\gamma_1 \neq 0$), we conclude **treatment has an effect**.

Yet, these questions can lead to **different answers**.

Question: Would it help if we could decide whether we are interested in Y_{2i} or $G_i = Y_{2i} - Y_{1i}$ as the outcome? → They're diff. effects. okay.. But still then we want to know what should we do then?

DIY

We have **two models**:

- ANCOVA: $Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_{2i}$
- CSM: $Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + \epsilon_i$

1) **Rewrite** the ANCOVA model as a changes score model (i.e., with G_i as the outcome); what does this tell you?

2) **Rewrite** CSM as ANCOVA model (i.e., with Y_{2i} as the outcome); what does this tell you?

DIY

We have **two models**:

- ANCOVA: $Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_{2i}$
- CSM: $Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + \epsilon_i$

1) **Rewrite** the ANCOVA model as a changes score model (i.e., with G_i as the outcome); what does this tell you?

2) **Rewrite** CSM as ANCOVA model (i.e., with Y_{2i} as the outcome); what does this tell you?

1) ANCOVA as CSM: $Y_{2i} - Y_{1i} = \beta_0 + \beta_1 X_i + (\beta_2 - 1) Y_{1i} + e_{2i}$

DIY

We have **two models**:

- ANCOVA: $Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_{2i}$
- CSM: $Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + \epsilon_i$

1) **Rewrite** the ANCOVA model as a changes score model (i.e., with G_i as the outcome); what does this tell you?

2) **Rewrite** CSM as ANCOVA model (i.e., with Y_{2i} as the outcome); what does this tell you?

1) ANCOVA as CSM: $Y_{2i} - Y_{1i} = \beta_0 + \beta_1 X_i + (\beta_2 - 1) Y_{1i} + e_{2i}$

- It shows that doing an ANCOVA controlling for pre-test with either Y_{2i} or $G_i = Y_{2i} - Y_{1i}$ as the outcome leads to the same effect of X_i.

this will stay intact.

- reg. coef. for the pre-treatment variable, is $\beta_2 - 1$

DIY

We have **two models**:

- ANCOVA: $Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_{2i}$
- CSM: $Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + \epsilon_i$

1) **Rewrite** the ANCOVA model as a changes score model (i.e., with G_i as the outcome); what does this tell you?

2) **Rewrite** CSM as ANCOVA model (i.e., with Y_{2i} as the outcome); what does this tell you?

1) ANCOVA as CSM: $Y_{2i} - Y_{1i} = \beta_0 + \beta_1 X_i + (\beta_2 - 1) Y_{1i} + e_{2i}$

It shows that doing an ANCOVA controlling for pre-test with either Y_{2i} or $G_i = Y_{2i} - Y_{1i}$ as the outcome leads to the same effect of X_i.

2) CSM as ANCOVA: $Y_{2i} = \gamma_0 + \gamma_1 X_i + Y_{1i} + \epsilon_i$

DIY

We have **two models**:

- ANCOVA: $Y_{2i} = \beta_0 + \beta_1 X_i + \beta_2 Y_{1i} + e_{2i}$
- CSM: $Y_{2i} - Y_{1i} = \gamma_0 + \gamma_1 X_i + \epsilon_i$

1) **Rewrite** the ANCOVA model as a changes score model (i.e., with G_i as the outcome); what does this tell you?

2) **Rewrite** CSM as ANCOVA model (i.e., with Y_{2i} as the outcome); what does this tell you?

1) ANCOVA as CSM: $Y_{2i} - Y_{1i} = \beta_0 + \beta_1 X_i + (\beta_2 - 1) Y_{1i} + e_{2i}$

It shows that doing an ANCOVA controlling for pre-test with either Y_{2i} or $G_i = Y_{2i} - Y_{1i}$ as the outcome leads to the same effect of X_i.

2) CSM as ANCOVA: $Y_{2i} = \gamma_0 + \gamma_1 X_i + Y_{1i} + \epsilon_i$ This shows that the CSM can be considered a special case of ANCOVA (with $\beta_2 = 1$). In original ANCOVA situation, β_2 can take diff. values.

$\Rightarrow$ So, in a very specific situation, two models may give the same results!! for ex, when $\beta_2 = 1$, therefore

DIY: When are the ACEs the same?

also other scenarios...

When will the ANCOVA model and the change score model give the same ACE?

$$ACE_{CS} = \big(E\big[Y_2|X=1\big] - \big(E\big[Y_2|X=0\big]\big) - \big(E\big[Y_1|X=1\big] - E\big[Y_1|X=0\big]\big)$$

If we express the change score ACE in terms of β_1 & β_2, we get this

$$ACE_{ANCOVA} = \beta_1$$

$$ACE_{CS} = \beta_1 + (\beta_2 - 1)\big(E\big[Y_1|X=1\big] - E\big[Y_1|X=0\big]\big)$$

difference between pretreatment between the groups

ANSWER: These are **identical when** $(\beta_2 - 1)\big(E\big[Y_1|X=1\big] - E\big[Y_1|X=0\big]\big) = 0$

meaning that there's no diff. in pre-treatment between two groups

there're 2 scenarios!!

This is the case when either:

1) $\beta_2 = 1$: the **effect of the pretest on posttest within groups is 1** (as you had already found when rewriting a CSM as an ANCOVA model); or

2) $\big(E\big[Y_1|X=1\big] - E\big[Y_1|X=0\big]\big)$: there are **no initial group differences** (as in an RCT!)

$M_{Y1} = M_{Y0}$

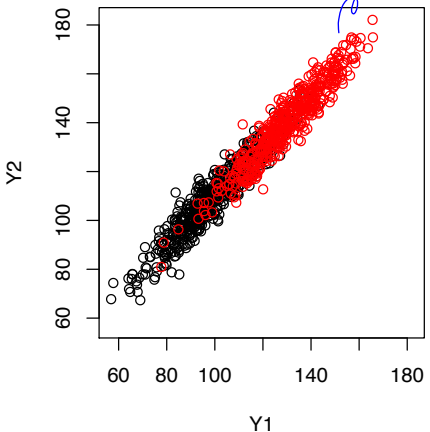
we'll also get the same result!! as w/ ANCOVA

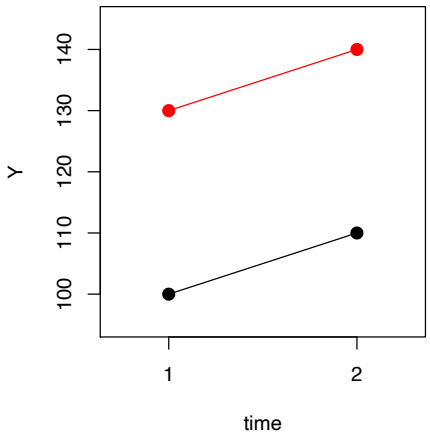
where there's no pre-treatment diff

Five scenarios

Scenario 1: No causal effect

shape here, because $\beta_2 = 1$: relationship between Y_1 & Y_2 is 1.

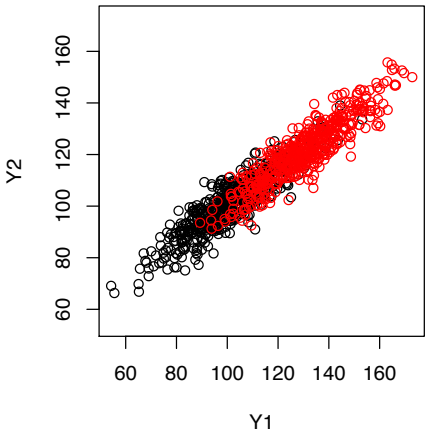


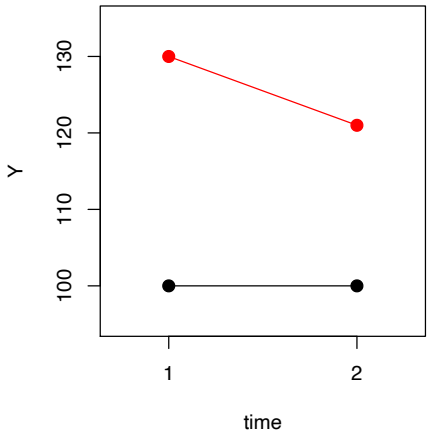


ANCOVA: $\beta_1 = 0$ so no causal effect

CSM: $\gamma_1 = 0$ so no causal effect

Scenario 2: CS model negative causal effect ~ They have diff. results. ? Becuz there're group differences in pre-treatment

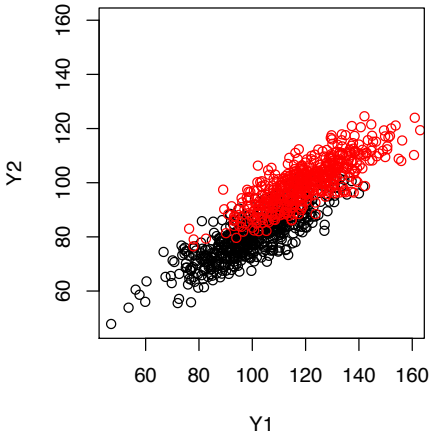


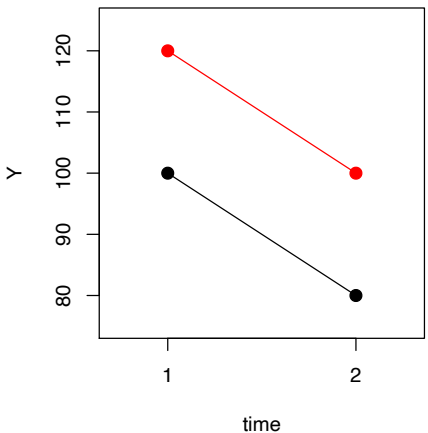


ANCOVA: $\beta_1 = 0$ so no causal effect

CSM: $\gamma_1 < 0$ so a (negative) causal effect

Scenario 3: ANCOVA model positive causal effect

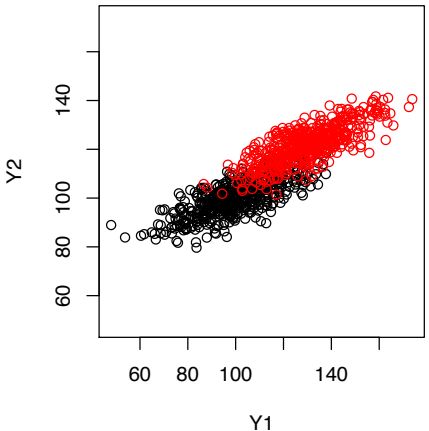


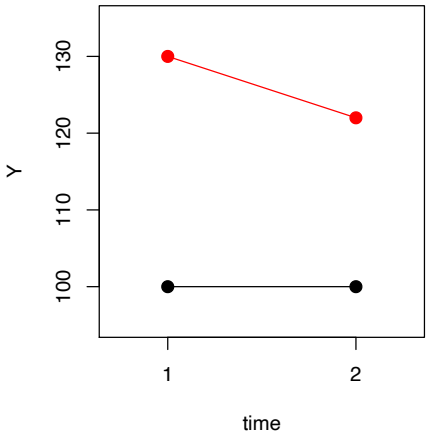


ANCOVA: $\beta_1 > 0$ so a (positive) causal effect

CSM: $\gamma_1 = 0$ so no causal effect

Scenario 4: Opposite conclusions regarding direction!

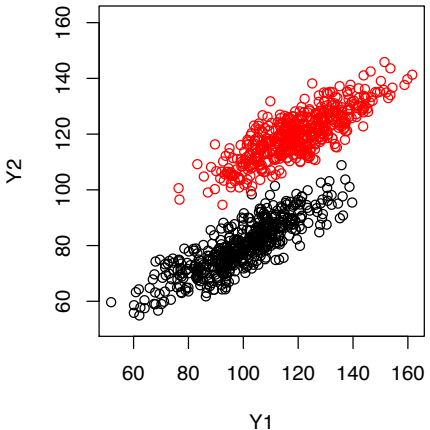


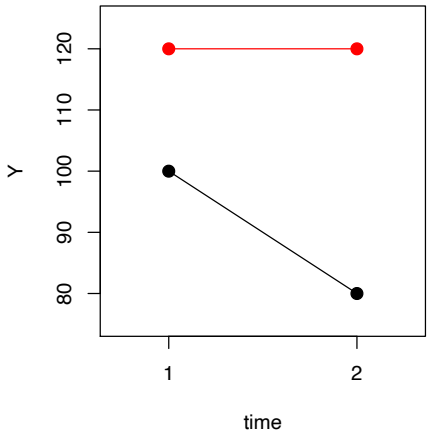


ANCOVA: $\beta_1 > 0$ so a (positive) causal effect

CSM: $\gamma_1 < 0$ so a (negative) causal effect

Scenario 5: Some agreement





ANCOVA: $\beta_1 > 0$ so a (positive) causal effect

CSM: $\gamma_1 > 0$ so a (positive) causal effect

So what now?

It can be shown that only when 1) $\beta_2 = 1$ and/or 2) $\mu_{1|1} = \mu_{1|0}$, are β_1 (ACE_{ANCOVA}) and γ_1 (ACE_{CSM}) identical.

Allison (p.109, 1990):

"It is unrealistic to expect either model to be best in all situations; [...] the choice will rarely be obvious, and there will almost always be some residual uncertainty. One should also consider the possibility that neither of these models is appropriate [...]."

Allison (p.100, 1990):

"A problem with much of the work comparing change score and regressor variable methods is that the conclusions are rarely based on an explicit model for generation of the data."

Draw your DAG! → to figure out what the right thing is.

How DAGs can help

One thing we saw in the previous result is that,
soon as we include the pre-test as a control variable,
also in the change-score model, then it becomes equivalent to ANCOVA.

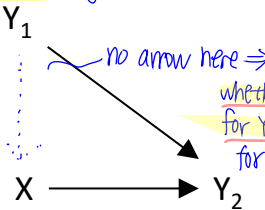
SO $\Rightarrow$ one impo. consideration is: (Do we need to control for pre-treatment or not?)

Should we control for pretest? - Y2 as outcome

Ofc. it depends on the actual DAG!

The key issue is whether **assignment is**:

Random assignment



no arrow here ⇒ doesn't matter whether we control for Y_1 or not for a causal effect

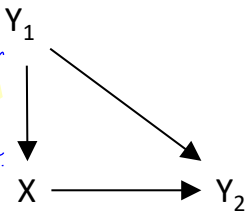
ANCOVA is preferred,

(statistical)

it has more power

But in terms of getting the right causal effect, it doesn't matter if we include it or not

Based on pretest



ANCOVA is correct;

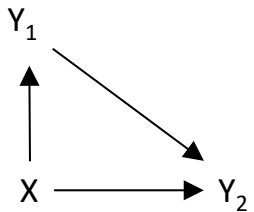
pretest is a **confounder**

↓

And in that case, we need to control for it to remove confounding

Treatment status determines the pre-treatment

Existing groups : mediating situation



We do not want to control unless you're only interested in

ANCOVA gives direct effect

pretest is a **mediator**

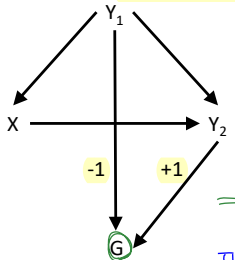
↓

we do not want to control for pre-treatment if we're interested in total effect

Controlling for Pre-Treatment - change score

Change/Gain score: $G_i = Y_{2i} - Y_{1i}$

Pretest as confounder



$\Rightarrow G := Y_2 - Y_1$

Need to block backdoor paths:

$X \leftarrow Y1 \rightarrow G$

$X \leftarrow Y1 \rightarrow Y2 \rightarrow G$

Then, again, the answer is the same as before: we should actually include pre-treatment in our change-score analysis to get the correct causal effect, the change-score.

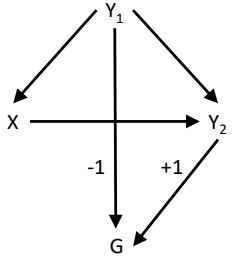
Control for pre-treatment (use ANCOVA) to get

$X \rightarrow Y2 \rightarrow G$

Controlling for Pre-Treatment - change score

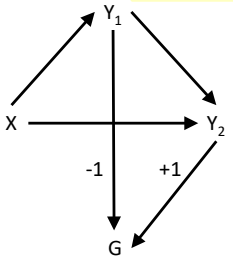
Change/Gain score: $G_i = Y_{2i} - Y_{1i}$

Pretest as confounder



Need to block backdoor paths:
$X \leftarrow Y1 \rightarrow G$
$X \leftarrow Y1 \rightarrow Y2 \rightarrow G$
Control for pre-treatment (use ANCOVA) to get
$X \rightarrow Y2 \rightarrow G$

Pretest as mediator

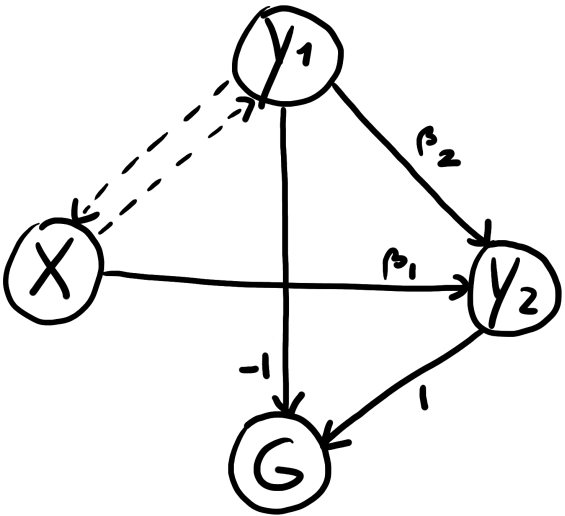


Total effect consists of:
$X \rightarrow Y2 \rightarrow G$
$X \rightarrow Y1 \rightarrow G$
$X \rightarrow Y1 \rightarrow Y2 \rightarrow G$

And then again, we do not want to control for pre-treatment.

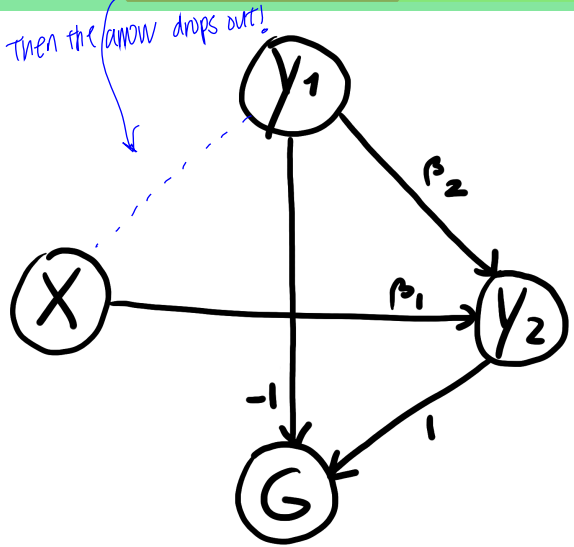
Do not control for pre-treatment (classical change score model) to get total effect; control for pre-treatment (ANCOVA) to get direct effect.

Revisit: When classical Change Score and ANCOVA model have the same results



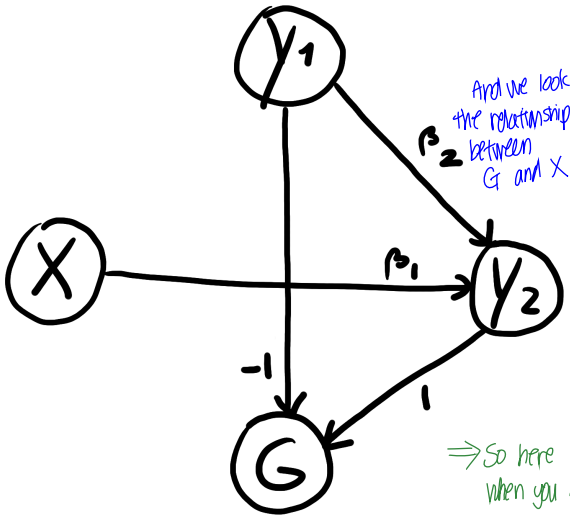
- When the expected values for Y1 are the same in each group (no group differences in Y1) $\mu_{111} = \mu_{110}$
- When the effect of Y1 on Y2 is equal to 1 $\beta_2 = 1$

Revisit: When classical Change Score and ANCOVA model have the same results: EVs Y1 the same in each group



Then what you'll see, the results become the same,

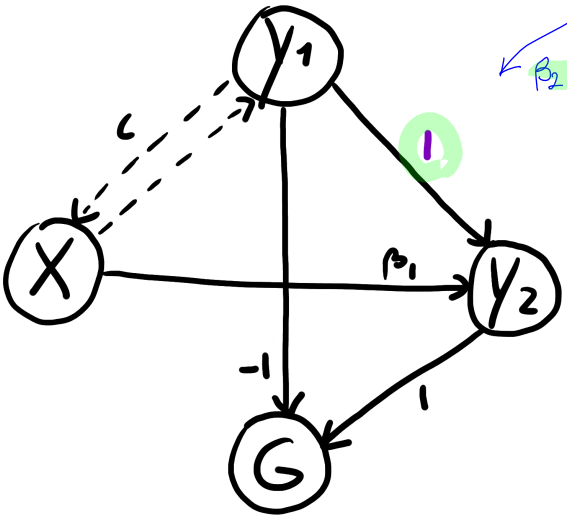
Revisit: When classical Change Score and ANCOVA model have the same results: EVs Y1 the same in each group



CS (do not control for y_1)	ANCOVA y_2 (control for y_1)
And we look at the relationship between G and X	So we have this model here
$G \sim X$	$y_2 \sim X + y_1$
paths X - G	paths X - y_2
$\beta_1 \times 1$	β_1
there's only one path left	

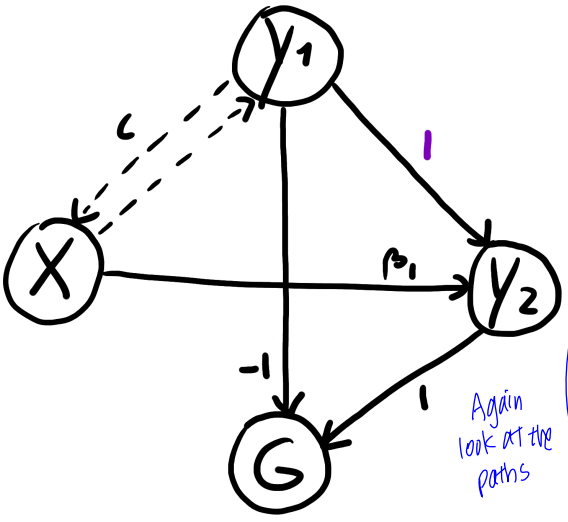
⇒ So here it's expressed in DAGs when you get the same result for both again.

Revisit: When classical Change Score and ANCOVA model have the same results: Effect of Y1 on Y2 equal to 1



β_2 is fixed to 1 now.

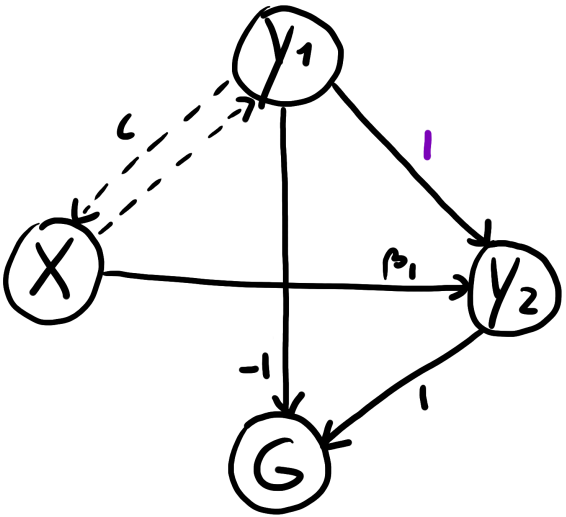
Revisit: When classical Change Score and ANCOVA model have the same results: Effect of Y1 on Y2 equal to 1



Again look at the paths

CS (do not control for y_1)	ANCOVA Y_2 (control for y_1)
$G \sim X$	$Y_2 \sim X + Y_1$
paths X - G	paths $x - y_2$
$X \leftrightarrow Y_1 \xrightarrow{1} Y_2 \xrightarrow{1} G$	$X \leftrightarrow Y_1 \xrightarrow{1} Y_2$
$X \leftrightarrow Y_1 \xrightarrow{-1} G$	$X \leftrightarrow Y_1 \xrightarrow{-1} G \leftarrow Y_2$
$X \xrightarrow{\beta_1} Y_2 \xrightarrow{1} G$	$X \xrightarrow{\beta_1} Y_2$
$1 + (-1) + \beta_1$	blocked ∵ we're controlling for Y_1
	blocked ∵ G is a collider

Revisit: When classical Change Score and ANCOVA model have the same results: Effect of Y1 on Y2 equal to 1



CS (do not control for y_1)	ANCOVA y_2 (control for y_1)
$G \sim X$	$y_2 \sim X + y_1$
paths $X - G$	paths $X - y_2$
$X \leftrightarrow y_1 \xrightarrow{1} y_2 \xrightarrow{1} G$	blocked: $X \leftarrow y_1 \xrightarrow{1} y_2$
$X \leftarrow y_1 \xrightarrow{1} G$	blocked: $X \leftarrow y_1 \rightarrow G \xrightarrow{1} y_2$
$X \xrightarrow{\beta_1} y_2 \xrightarrow{1} G$ =	$X \xrightarrow{\beta_1} y_2$

Timing is critical

We need to think about the relationship between the pre-treatment variable & the rest
↳ Do we need to control for it, YES/NO to get the correct causal effect. This stays the same whether you use change-score or ANCOVA...

The DAGs show that the **causal relation** between treatment X and pre-test Y1 is critical; it is about whether X or Y1 came first (i.e., their temporal order).

In Rubin's causal framework (Week 2), the timing of treatment, outcome and covariates is also considered critical.

If you know the timing of treatment & outcome variable, then you can make decision more easily, whether it is collider / mediator / confounder & accordingly whether you should control for it or not...

Holland (1986):

- exposure to a cause (i.e., treatment) occurs at a **specific time point or time interval**
- variables are thus divided into **pre-exposure** and **post-exposure**
- "The role of a response variable Y is to measure the effect of the cause, and thus **response variables** must fall into the **post-exposure class**." (p.946, Holland, 1986)
- **covariates** should come from the **pre-exposure phase**; then they cannot be affected by the treatment.

Critique of using change scores in DAGs

~ ppl do have opinions on whether you should use change score in the context of causal models...

Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Causal diagrams and change variables

should you look at change score at all?
or should you really focus on Y_2?

Eyal Shahar MD MPH[1] and Doron J. Shahar[2]

Abstract

Background The true change in the value of a variable between two time points is often assumed to be a cause or an effect of interest. To our knowledge, this assumption is based on intuition, rather than on any formal theoretical justification.

Methods We used causal directed acyclic graphs to explore the causal properties of a change variable, and critically examined competing structures.

Results Based on the proposed causal structure, a change variable (true change) is no more than a derived variable. It does not cause anything and is not of causal interest.

Conclusions A true change is not a variable in the physical world. Therefore, modelling the change between two time points is justified only in a few situations.

so ↓ The idea is that, there cannot really be arrows pointing outwards of G.
Would it be possible for G to be ever be a causal variable in that sense,
or can it only always be an outcome variable somewhere at the end..
& they're saying G is not so relevant, & not really interesting

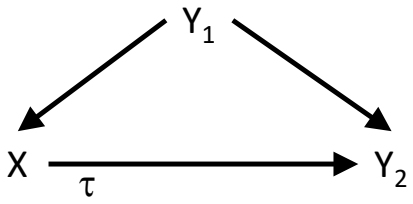
Unmeasured confounding

in the pre-post test design

When there's a specific kind of unmeasured confounding,
looking at change-score can be very useful!! ☺

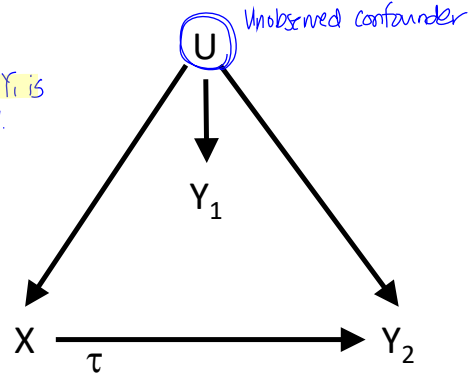
Pretest as a proxy for confounder

When Y_1 was **measured prior to treatment**, it could be a **confounder**; you need to control for it then (e.g., use ANCOVA model).

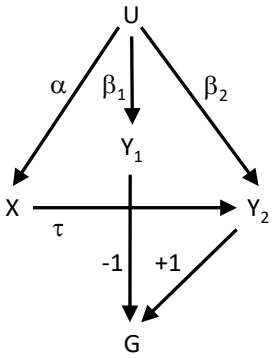


But Y_1 may also be a **proxy of an unobserved confounder**; controlling for Y_1 will only **partly remove bias** due to $X \leftarrow U \rightarrow Y_2$.

depending on how strongly Y_1 is related to U.



How can classical change score analysis help?



The interest is in $X \to G$; this is equal to $X \to Y2$

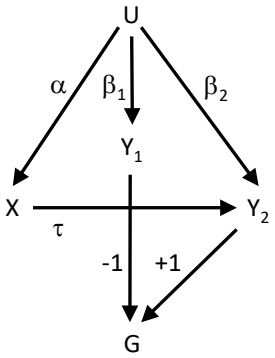
Other paths between X and G:

- $X \leftarrow U \rightarrow Y1 \rightarrow G$
- $X \leftarrow U \rightarrow Y2 \rightarrow G$
- $X \rightarrow Y2 \leftarrow U \rightarrow Y1 \rightarrow G$

When we have **linear relations** (and the variance of U is equal to 1), we get:

- First path: $-\alpha\beta_1$
- Second path: $\alpha\beta_2$
- Third path: 0 (contains the **collider** $Y2$)

How can classical change score analysis help?



The interest is in $X \rightarrow G$; this is equal to $X \rightarrow Y2$

Other paths between X and G:

- $X \leftarrow U \rightarrow Y1 \rightarrow G$
- $X \leftarrow U \rightarrow Y2 \rightarrow G$
- $X \rightarrow Y2 \leftarrow U \rightarrow Y1 \rightarrow G$

When we have **linear relations** (and the variance of U is equal to 1), we get:

- First path: $-\alpha\beta_1$
- Second path: $\alpha\beta_2$
- Third path: 0 (contains the **collider** $Y2$)

If $\beta_1 = \beta_2$, the first and second path **cancel each other out** (cf. Kim & Steiner, 2019)!

Important to realize (i.e., conclusion so far)

We have written the change score model as a **special case** of the ANCOVA model.

This may suggest we should just test **which model fits better**.

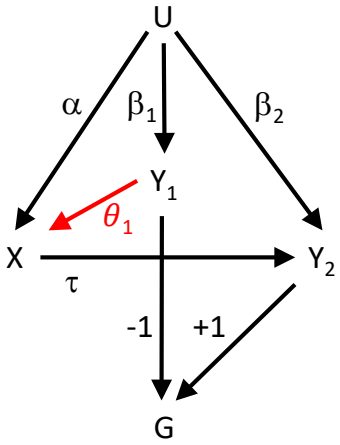
However, the point is **NOT** that we want to determine which of these two models generated the data!

The goal is to "**estimate the treatment effect without bias**" (true causal effect).

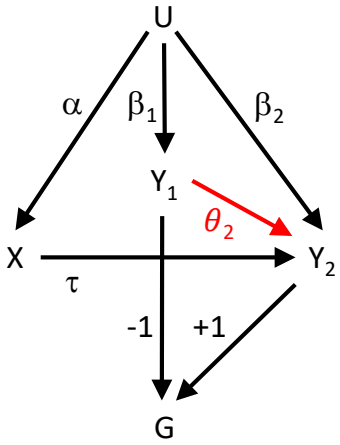
The **change score model** is:

- typically NOT considered a reasonable model as a data generating mechanism
- but a very useful model for estimating the causal effect (under very specific circumstances) ~ ex) $\beta_1 = \beta_2$

What if... we have slightly diff. situations, then it no longer works!!

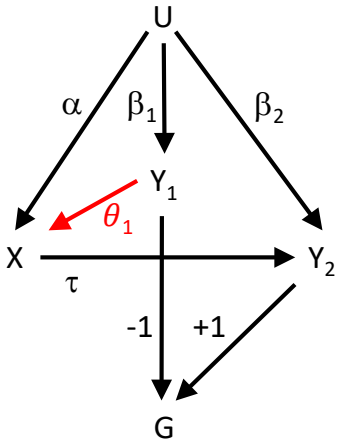


Pretest affects treatment

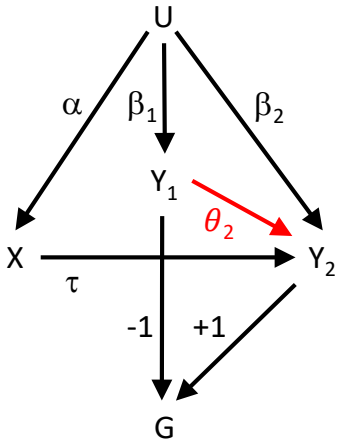


Pretest affects outcome

What if...



Pretest affects treatment

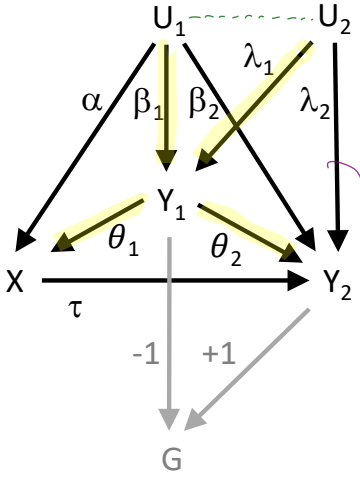


Pretest affects outcome

In these scenarios neither the Change score model nor the ANCOVA model give an unbiased estimate of the causal effect of X on Y2.

Also important to realize (wrt timing)

When pretest is from the **pretreatment phase**, it does **NOT** mean it can **only** be a confounder. Becuz there might be all kinds of unobserved variables that we didn't take into account.



we can have a variable that can be both confounder & collider, or both mediator & confounder.. specially when you have multiple variables that are affecting each other over time. Then, you have a problem...

The **pretest Y1** is:

- a **confounder**: $X \leftarrow Y1 \rightarrow Y2$
- a **collider**: $X \leftarrow U1 \rightarrow Y1 \leftarrow U2 \rightarrow Y2$

var. can be both confounder & collider..!!

So if we control for Y_1, we open a backdoor path by introducing a spurious relationship between U_1 & U_2 and introduce bias!

Summary

Use

ANCOVA (regress Y2 or G=Y2-Y1 on X and Y1):

- when Y1 is confounder of X and Y2
- or to get direct effect when Y1 is mediator

Use

Marginal model (regress Y2 on X):

- when Y1 is mediator and interest is in total effect of X on Y2

Use

Change score model (regress G=Y2-Y1 on X):

- when there is time-invariant unobserved confounding with stable effect : when we have this very specific kind of unobserved confounding
- (or when Y1 is mediator and the interest is in total effect of X on Y2-Y1)

Summary

ANCOVA (regress Y2 or G=Y2-Y1 on X and Y1):

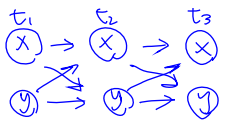
- when Y1 is confounder of X and Y2
- or to get direct effect when Y1 is mediator

Marginal model (regress Y2 on X):

- when Y1 is mediator and interest is in total effect of X on Y2

Change score model (regress G=Y2-Y1 on X):

- when there is time-invariant unobserved confounding with stable effect
- (or when Y1 is mediator and the interest is in total effect of X on Y2-Y1)



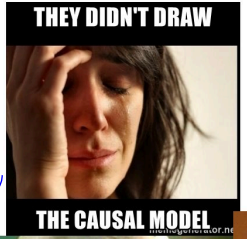
There is a lot more to say and study about causality and time

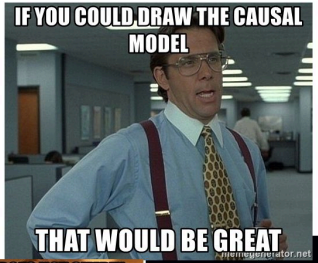
- Consider time-varying treatments, outcomes, and covariates
- Variables affecting themselves and each other continuously through time, effects may change over time, ...
- In (these) more complicated scenarios we may have variables that are simultaneously confounders/mediators/colliders.

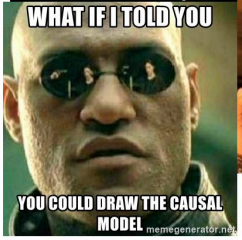
In any case...

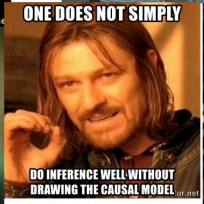
Do causal inference in a principled way!

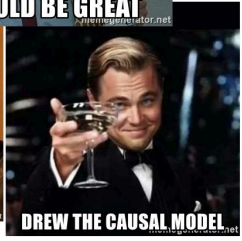
- Be explicit and clear about your causal interests/questions
- Specify your ideas (causal theory) in some causal graph and/or in equations (and SCM).. draw the DAG! :)
- make assumptions explicit
- Choose a causal analysis best tailored to your particular problem.
- Replicate, triangulate, critique, etc!











In any case...

Finally, remember these science key three...

- Measurement
- Theory formation
- Causal Inference