Week 5: Frameworks for Causal Inference

Brandon Stewart[1]

Princeton

October 1–3, 2024

[1]These slides are heavily influenced by Matt Blackwell, Justin Grimmer, Jens Hainmueller, Erin Hartman, Kosuke Imai, Ian Lundberg, and Matt Salganik.

Where We've Been and Where We're Going...

- Last Week
  - hypothesis testing
  - core ideas in causal inference
  - potential outcomes
- This Week
  - nonparametric structural equation models and DAGs
  - experiments
  - selection on observables
- Next Week
  - estimating conditional expectation functions
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Nonparametric Structural Equation Models

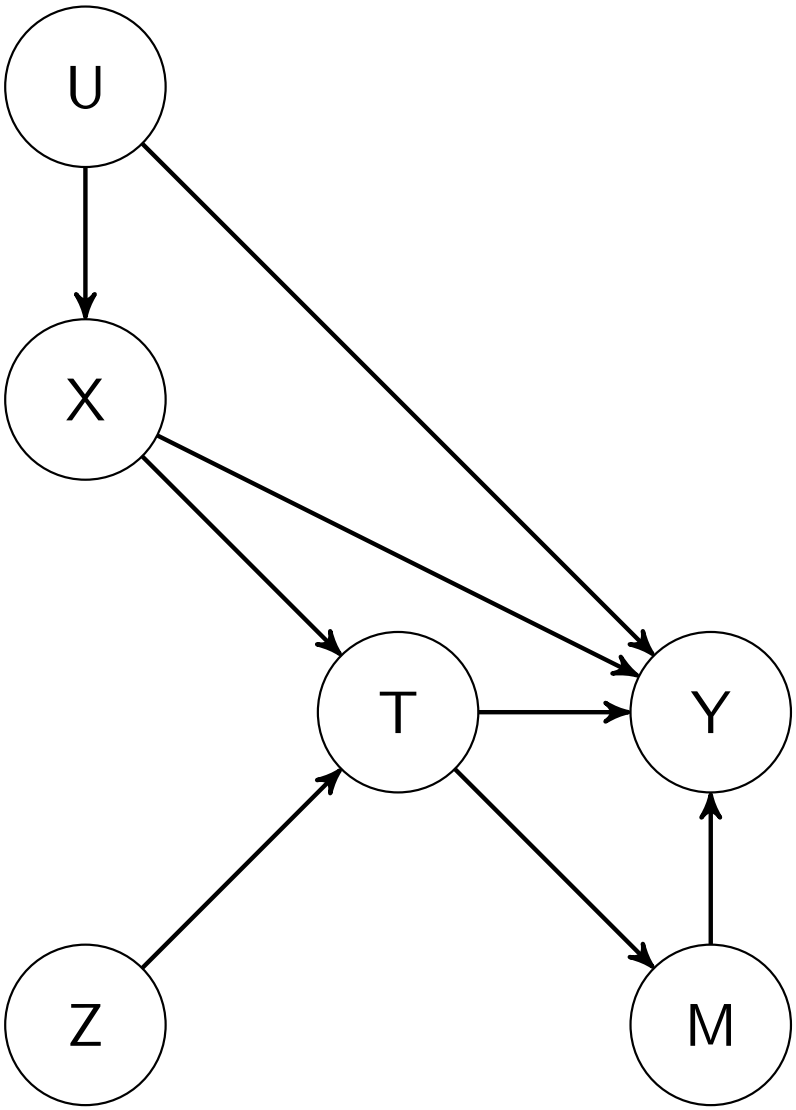
- The work we discuss here comes out of developments by Judea Pearl and others
- Idea of nonparametric structural equations (NPSEMs), is that we will write each variable as a general function of the variables that it "listens to" (i.e. could affect it directly if intervened on).
  e.g. $T = f(X, U_T)$ for some arbitrary function $f()$
- From this we can derive a set of rules (called the do-calculus) which allow us to relate causal quantities and non-causal quantities.
- The `do()` operator denotes the act of intervening
  e.g. $p(Y|T = 1)$ is the distribution of the outcome for those who are observed to receive treatment but $p(Y|do(T = 1))$ is the distribution among those when we intervene to treat.
- Identification involves using the do-calculus to write the causal quantity (which involves do()) in terms of an empirical quantity (which involves only observed data).

Graphical Models

- A general framework for representing causal relationships ('what listens to what' in the NPSEM) based on directed acyclic graphs (DAG)
- A way to encode assumptions about the causal structure precisely and separately from assumptions about estimation.
- The do-calculus and a DAG are all you need to determine whether causal quantities are identified.
- Nice software that takes the graph and returns an identification strategy: DAGitty at `http://dagitty.net` (also accessible directly in R).

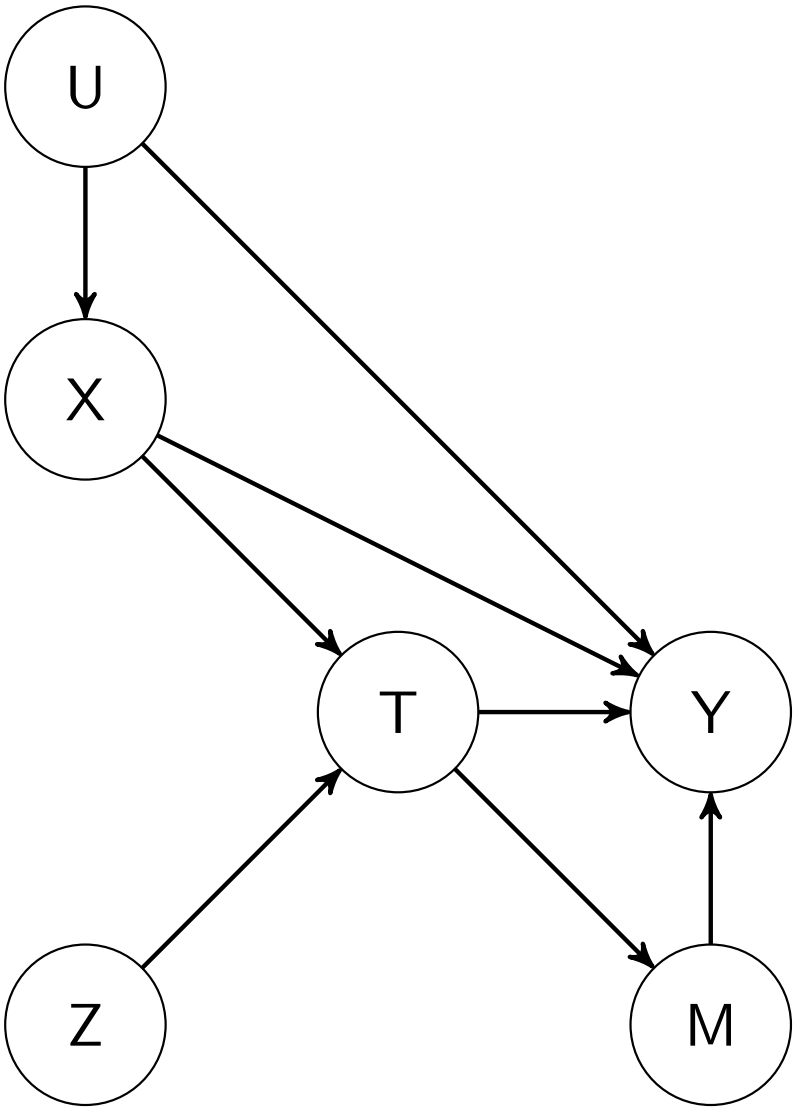
Code Idea: causal reasoning before statistical reasoning

Components of a DAG



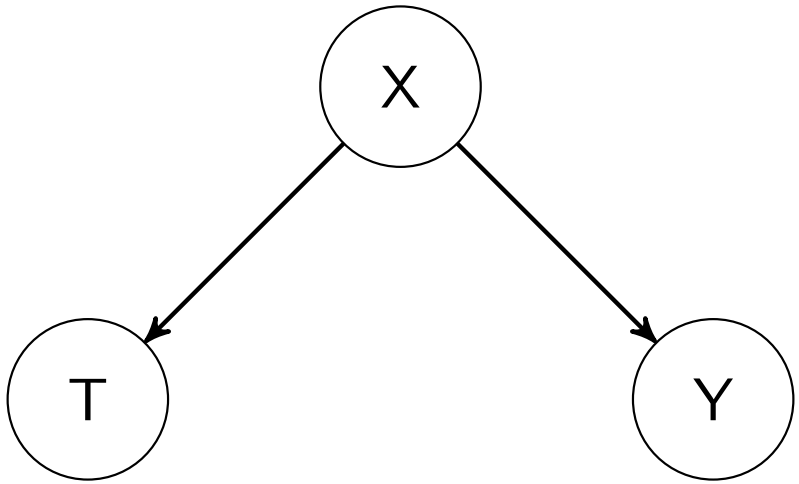
- nodes represent variables (unobserved typically called U or V)
- (directed) arrows represent causal effects
- absence of nodes represents no common causes of any pair of variables
- absence of arrows represents no causal effect
- positioning conveys no mathematical meaning but often is oriented left-to-right with causal ordering for readability.
- dashed lines are used in context dependent ways
- all relationships are non-parametric

Relationships in a DAG



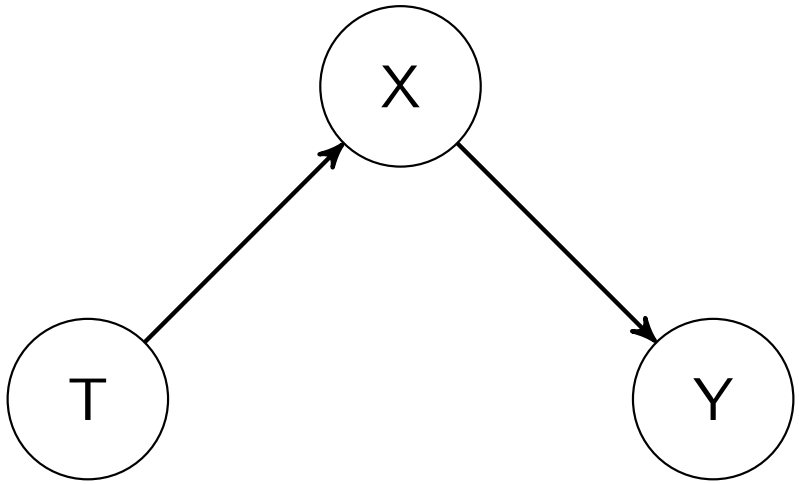
- Parents (Children): directly causing (caused by) a node
- Ancestors (Descendants): directly or indirectly causing (caused by) a node
- Path: a route that connects the variables (path is causal when all arrows point the same way)
- Acyclic implies that there are no cycles and a variable can't cause itself
- Causal Markov assumption: condition on its direct causes, a variable is independent of its non-descendents.
- We will talk in depth about two types of relationships: confounders and colliders

Forks (Confounders)



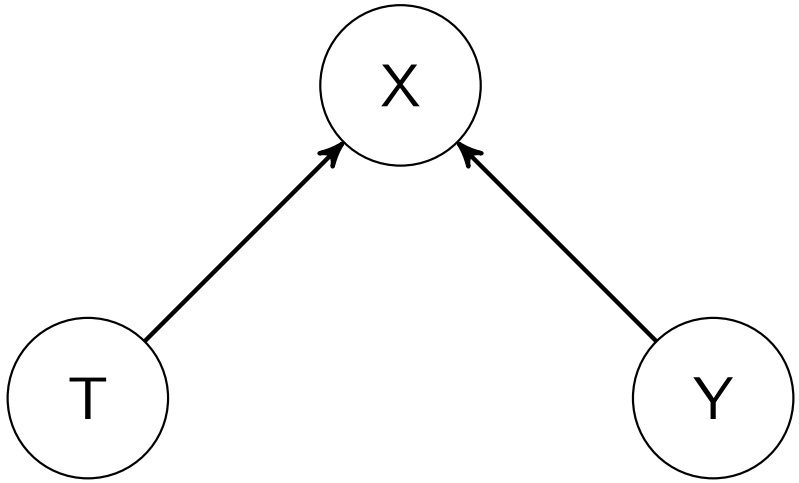
- This is a fork structure.
- When T is a treatment and Y is an outcome, we call X a confounder (or common cause).
- Even without a causal effect or directed edge between T and Y they will have a marginal associational relationship.
- Conditional on X, T and Y are unrelated in this graph.
- We can think of conditioning on the middle of a fork as blocking the flow of association.

Chains (Mediators)



- This is a chain structure.
- When T is a treatment and Y is an outcome, we call X a mediator on the path from T to Y.
- Even without a causal effect or directed edge between T and Y they will have a marginal associational relationship.
- Conditional on X, T and Y are unrelated in this graph even though T has a causal effect on Y.
- We generally don't want to condition on mediators because they include part of the effect we want to capture.

Colliders



- X is now a collider because two arrows point into it
- In this scenario T and Y are not marginally associated
- If we control for X, we create association between T and Y even if one was not there before.

Colliders are scary because you can induce dependence

Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable

Felix Elwert[1] and Christopher Winship[2]

[1]Department of Sociology, University of Wisconsin, Madison, Wisconsin 53706; email: elwert@ssc.wisc.edu

[2]Department of Sociology, Harvard University, Cambridge, Massachusetts 02138; email: cwinship@wjh.harvard.edu

Annu. Rev. Sociol. 2014. 40:31–53

First published online as a Review in Advance on June 2, 2014

The Annual Review of Sociology is online at soc.annualreviews.org

This article's doi: 10.1146/annurev-soc-071913-043455

Keywords

causality, directed acyclic graphs, identification, confounding, selection

Abstract

Endogenous selection bias is a central problem for causal inference. Recognizing the problem, however, can be difficult in practice. This article introduces a purely graphical way of characterizing endogenous selection bias and of understanding its consequences (Hernán et al. 2004). We use causal graphs (directed acyclic graphs, or DAGs) to highlight that endogenous selection bias stems from conditioning (e.g., controlling, stratifying, or selecting) on a so-called collider variable, i.e., a variable that is itself caused by two other variables, one that is (or is associated with) the treatment and another that is (or is associated with) the outcome. Endogenous selection bias can result from direct conditioning on the outcome variable, a post-outcome variable, a post-treatment variable, and even a pre-treatment variable. We highlight the difference between endogenous selection bias, common-cause confounding, and overcontrol bias and discuss numerous examples from social stratification, cultural sociology, social network analysis, political sociology, social demography, and the sociology of education.

Colliders: When drawing a DAG helps

Example extended from Elwert & Winship 2014

Hypothetical substantive question:

Does acting ability causally affect the probability of marriage?

Hypothetical approach: Estimate on a sample of Hollywood actors and actresses.

We want to estimate:

Acting ability $\longrightarrow$ Marriage

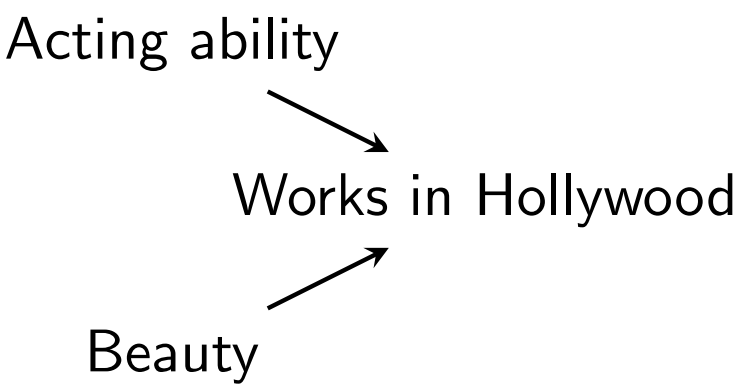
Should we worry about this design? It depends on our theory about how these variables are related. We can argue about identification with a DAG.

Colliders: When drawing a DAG helps

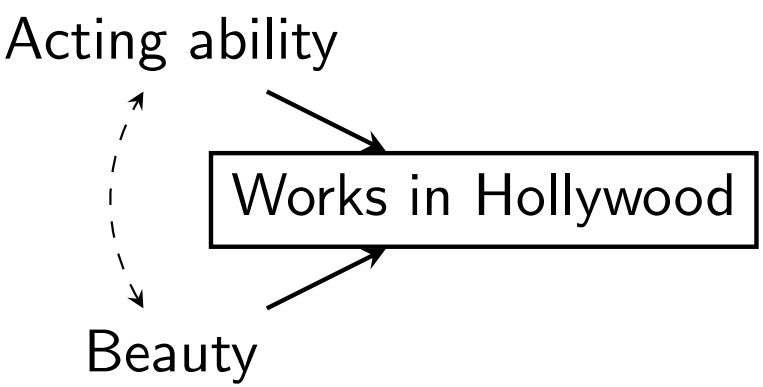
Example extended from Elwert & Winship 2014

Suppose working in Hollywood is a function of two factors: acting ability and beauty. In the general population, these two are uncorrelated. However, among those who work in Hollywood, those who are bad at acting must be beautiful.

True DAG



Conditional on Hollywood



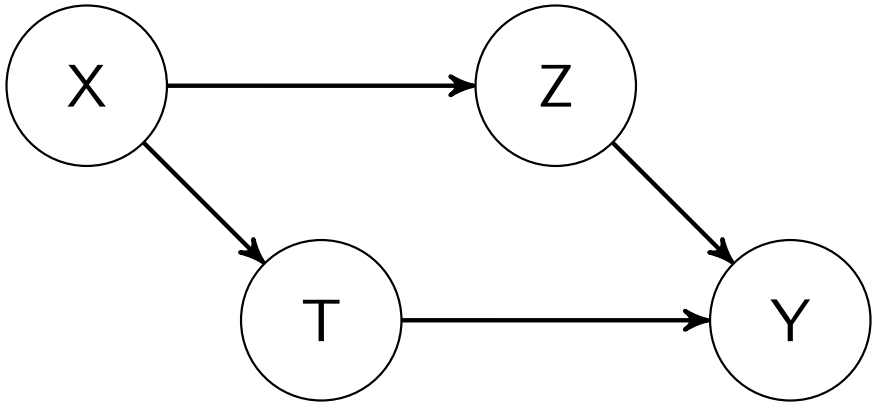
Colliders: When drawing a DAG helps

Example extended from Elwert & Winship 2014

This is an example of conditioning on a collider! We induce a negative association between acting ability and beauty.

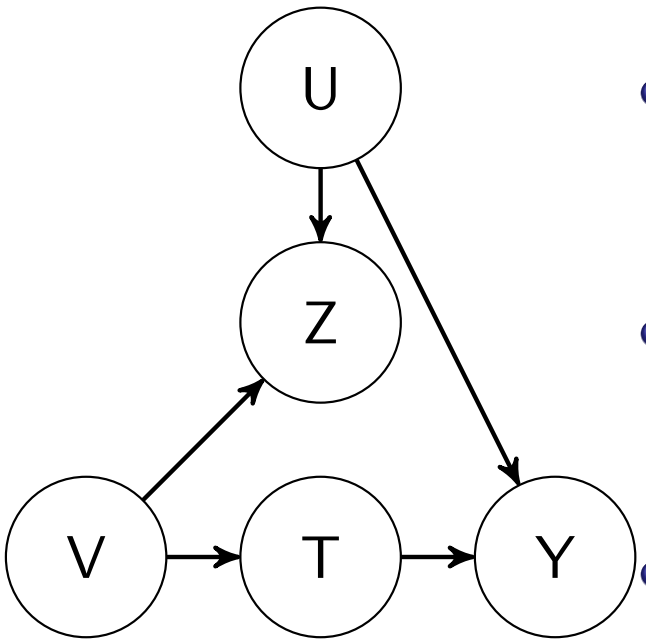
Under the assumptions above, our results are driven by collider conditioning!

From Confounders to Back-Door Paths



- Identify causal effect of T on Y by conditioning on X, Z or X and Z
- We can formalize this logic with the idea of a back-door path
- A back-door path is "a path between any causally ordered sequence of two variables that begins with a directed edge that points to the first variable." (Morgan and Winship 2013)
- Two paths from T to Y here:
  1. $T \rightarrow Y$ (directed or causal path)
  2. $T \leftarrow X \rightarrow Z \rightarrow Y$ (back-door path)
- Observed marginal association between T and Y is a composite of these two paths and thus does not identify the causal effect of T on Y
- We want to block the back-door path to leave only the causal effect

Colliders and Back-Door Paths



- Z is a collider and it lies along a back-door path from T to Y
- Conditioning on a collider on a back-door path does not help and in fact causes new associations
- Here we are fine unless we condition on Z which opens a path $T \leftarrow V \leftrightarrow U \rightarrow Y$ (this particular case is called M-bias)
- So how do we know which back-door paths to block?

Example: COVID

ARTICLE

https://doi.org/10.1038/s41467-020-19478-2 OPEN

Collider bias undermines our understanding of COVID-19 disease risk and severity

Gareth J. Griffith[1,2,4], Tim T. Morris[1,2,4], Matthew J. Tudball[1,2,4], Annie Herbert[1,2,4], Giulia Mancano[1,2,4], Lindsey Pike[1,2], Gemma C. Sharp[1,2], Jonathan Sterne[2], Tom M. Palmer[1,2], George Davey Smith[1,2], Kate Tilling[1,2], Luisa Zuccolo[1,2], Neil M. Davies[1,2,3] & Gibran Hemani[1,2,4]✉

Numerous observational studies have attempted to identify risk factors for infection with SARS-CoV-2 and COVID-19 disease outcomes. Studies have used datasets sampled from patients admitted to hospital, people tested for active infection, or people who volunteered to participate. Here, we highlight the challenge of interpreting observational evidence from such non-representative samples. Collider bias can induce associations between two or more variables which affect the likelihood of an individual being sampled, distorting associations between these variables in the sample. Analysing UK Biobank data, compared to the wider cohort the participants tested for COVID-19 were highly selected for a range of genetic, behavioural, cardiovascular, demographic, and anthropometric traits. We discuss the mechanisms inducing these problems, and approaches that could help mitigate them. While collider bias should be explored in existing studies, the optimal way to mitigate the problem is to use appropriate sampling strategies at the study design stage.

Received: 28 May 2020; Accepted: 8 October 2020;
Published online: 12 November 2020

Example: COVID

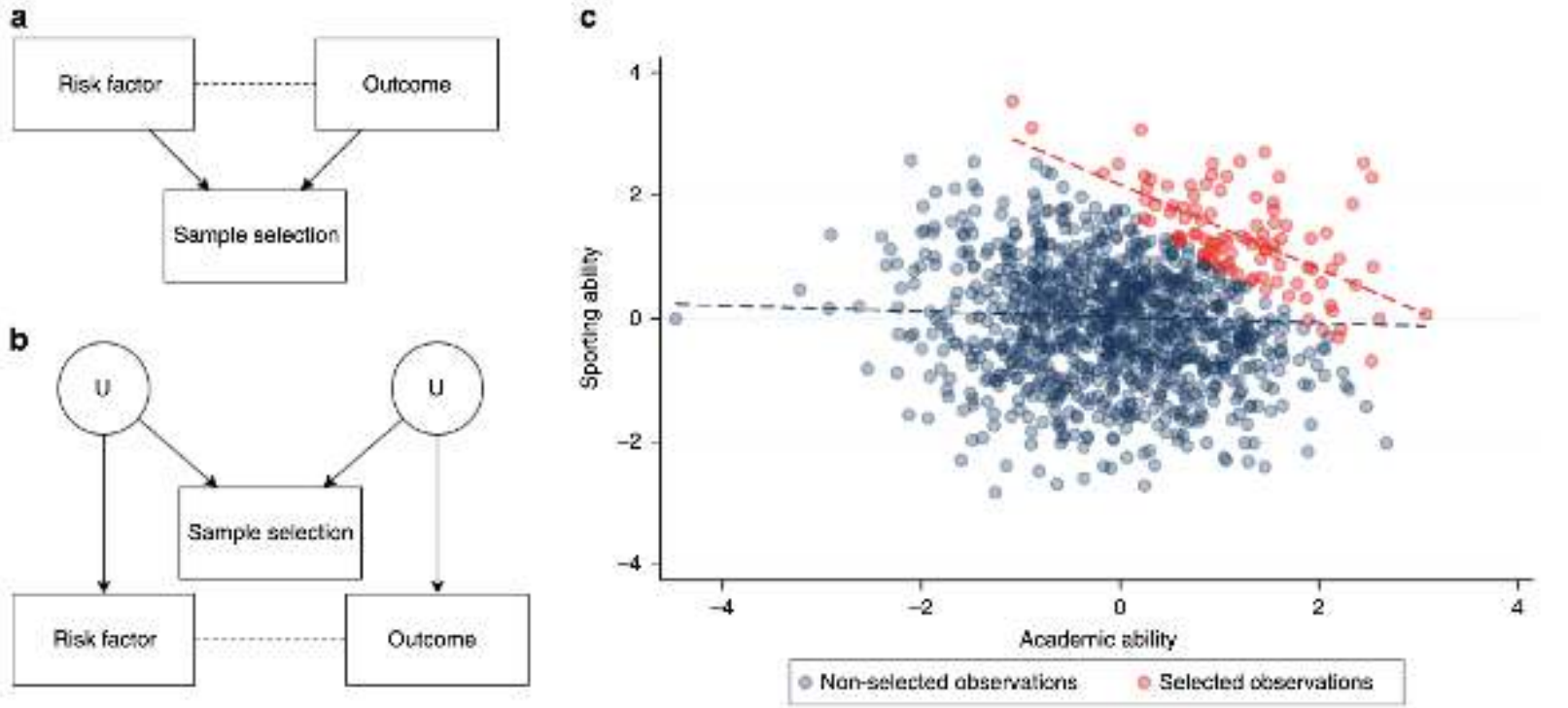


Fig. 1 Illustrative example of collider bias. a A directed acyclic graph (DAG) illustrating a scenario in which collider bias would distort the estimate of the causal effect of the risk factor on the outcome. Directed arrows indicate causal effects and dotted lines indicate induced associations. Note that the risk factor and the outcome can be associated with sample selection indirectly (e.g. through unmeasured confounding variables), as shown in **b**. The type of collider bias induced in graph (**b**) is sometimes referred to as M-bias. To illustrate the example in **a**, consider academic ability and sporting ability to each influence selection into a prestigious school. As shown in **c**, these traits are negligibly correlated in the general population (blue dotted line), but because they are selected for enrolment they become strongly correlated when analysing only the selected individuals (red dotted line).

Example: COVID

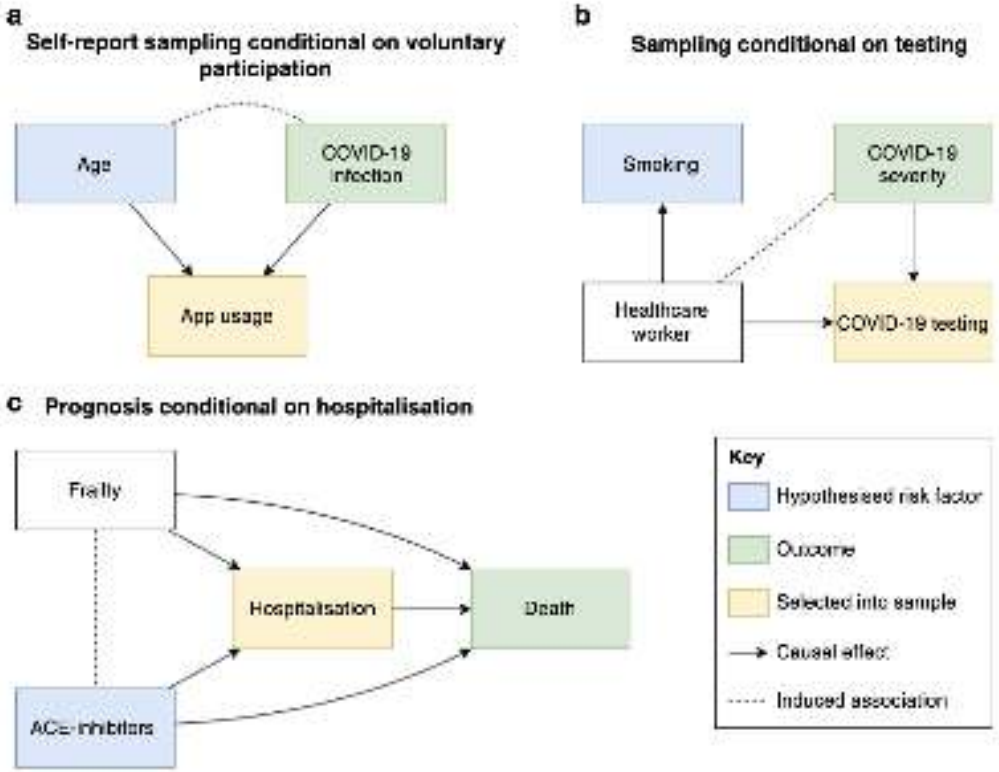


Fig. 2 Collider bias induced by conditioning on a collider in three scenarios relating to COVID-19 analysis. These are simplified Directed Acyclic Diagrams where only the main variables of interest have been represented for sake of illustrating collider bias scenarios. All assume no unspecified confounding or other biases. Rectangles represent observed variables and solid directed arrows represent causal effects. The dashed line represents an induced association when conditioning on the collider, which in these scenarios are variables that indicate whether an individual is selected into the sample **a** When some hypothesised risk factor (e.g. age) and outcome (e.g. COVID-19 infection) each associate with sample selection (e.g. voluntary data collection via mobile phone apps), the hypothesised risk factor and outcome will be associated within the sample. The presence and direction of these biases are model dependent; where causes are supra-multiplicative they will be positively associated in the sample; where they are sub-multiplicative they will be negatively correlated, and where they are exactly multiplicative they will remain unassociated. We extend this scenario in **b** where the association between the hypothesised risk factor and the collider does not need to be causal. **c** When inferring the influence of some hypothesised risk factor on mortality, in an unselected sample the risk factor for infection is a causal factor for death (mediated by COVID-19 infection). However, if analysed only amongst individuals who are known to have COVID-19 (i.e. we condition on the COVID-19 infection variable) then the risk factor for infection will appear to be associated with any other variable that influences both infection and progression. In many circumstances, this can lead to a risk factor for disease onset that appears to be protective for disease progression. Each of these scenarios represents those described in the main text.

D-Separation

- Graphs provide us a way to think about conditional independence statements. Consider disjoint subsets of the vertices A, B and C
- A is D-separated from B by C if and only if C blocks every path from a vertex in A to a vertex in B
- A path p is said to be blocked by a set of vertices C if and only if at least one of the following conditions holds:
  1. p contains a chain structure $a \to c \to b$ or a fork structure $a \leftarrow c \to b$ where the node c is in the set C
  2. p contains a collider structure $a \to y \leftarrow b$ where neither y nor its descendents are in C
- If A is not D-separated from B by C we say that A is D-connected to B by C

Three Rules of Do-Calculus

Define: $G_{\overline{T}}$ as the subgraph of G with all edges incoming to T deleted, and $G_{\underline{T}}$ the subgraph of G with all outgoing edges from T deleted.

Rule 1 (Insertion/Deletion of Observations):

$$\text{If } Y \perp\!\!\!\perp Z \mid T, X \text{ in } G_{\overline{T}}, \text{then}$$
$$P(y \mid \text{do}(t), z, x) = P(y \mid \text{do}(t), x).$$

Rule 2 (Action/Observation Exchange):

$$\text{If } Y \perp\!\!\!\perp T \mid X, Z \text{ in } G_{\underline{T},\overline{Z}},$$
$$\text{then } P(y \mid \text{do}(t), x, \text{do}(z)) = P(y \mid t, x, \text{do}(z)).$$

Rule 3 (Insertion/Deletion of Actions):

$$\text{If } Y \perp\!\!\!\perp T \mid X, Z \text{ in } G_{\overline{T(X)Z}},$$
$$\text{then } P(y \mid \text{do}(t), \text{do}(z), x) = P(y \mid \text{do}(z), x).$$

Backdoor Criterion

- Generally we want to know if we can nonparametrically identify the average effect of T on Y given a set of possible conditioning variables X
- Backdoor Criterion for X
  1. No node in X is a descendent of T (i.e. don't condition on post-treatment variables!)
  2. X D-separates every path between T and Y that has an incoming arrow into T (backdoor path)
- In essence, we are trying to block all non-causal paths, so we can estimate the causal path.
- Backdoor criterion is just one way to identify the effect: but its the most popular approach in the social sciences and what we are trying to do 99% of the time.
- We will see some other approaches late in the semester.

Derivation of Backdoor Adjustment Using Do-Calculus

We aim to express the causal effect $P(y \mid \mathrm{do}(t))$ in terms of observational quantities by adjusting for confounders X.

Step 1: Introduce X by conditioning:

$$P(y \mid \mathrm{do}(t)) = \sum_x P(y, x \mid \mathrm{do}(t)) = \sum_x P(y \mid x, \mathrm{do}(t)) P(x \mid \mathrm{do}(t))$$

Step 2: Use Rule 3 to remove an intervention: $P(x \mid \mathrm{do}(t)) = P(x)$

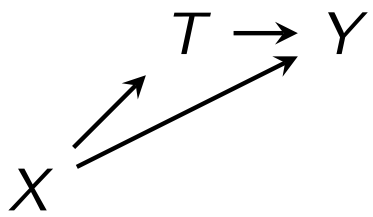
$$P(y \mid \mathrm{do}(t)) = \sum_x P(y \mid x, \mathrm{do}(t)) P(x)$$

Step 3: Use Rule 2 to treat an intervention as an observation.

$$P(y \mid \mathrm{do}(t)) = \sum_x P(y \mid x, t) P(x)$$

Blocking backdoor paths: College and earnings

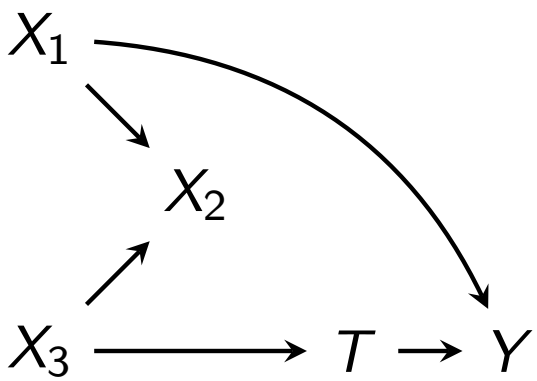
What do we need to include to block all backdoor paths between college and earnings?



Ability, parents' income, parents' education, extended family who pay for college and help you find a job, neighborhood characteristics that affect high school quality and also the availability of local jobs, ... lots of things!

Non-causal paths: Part 2

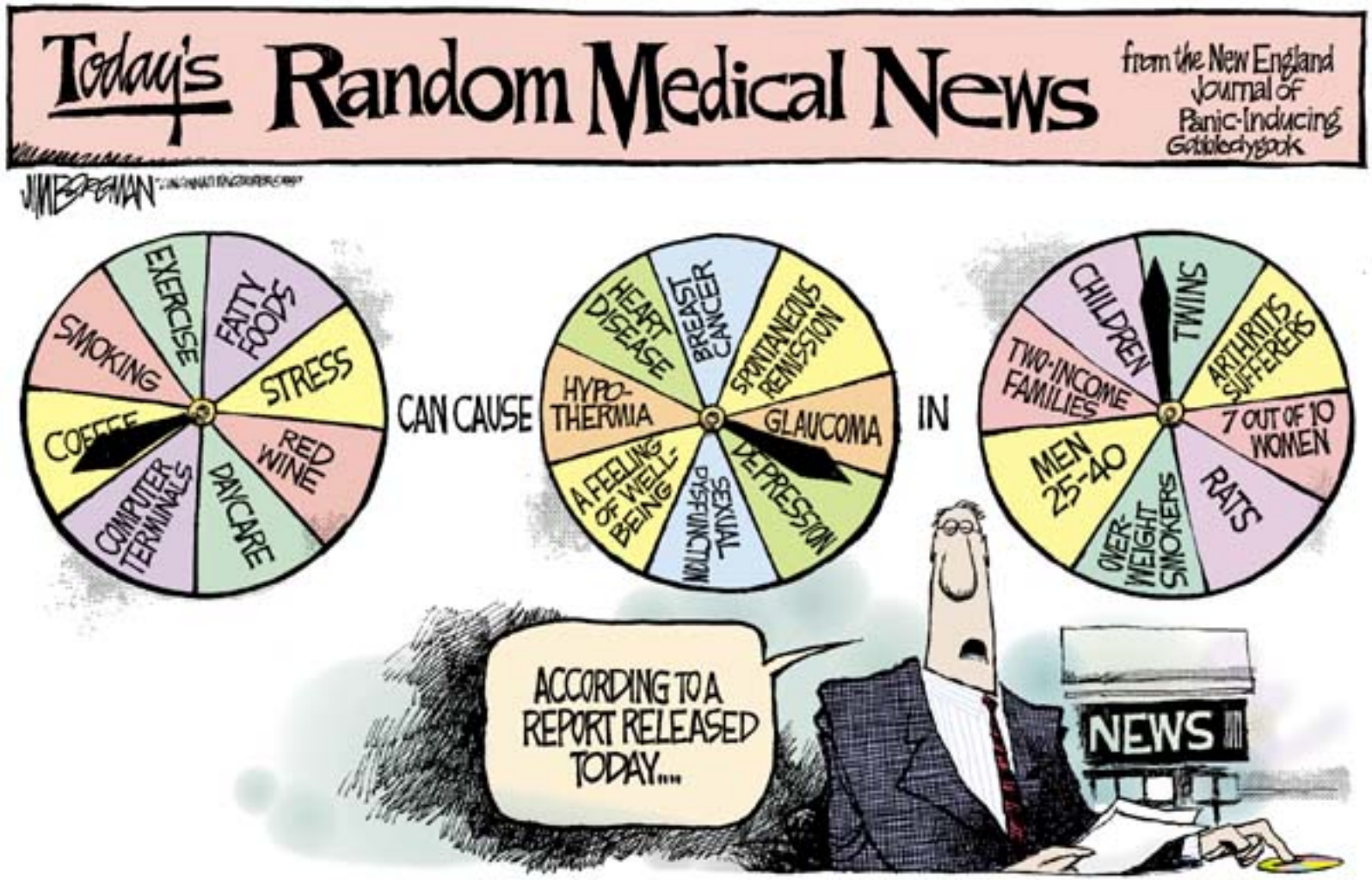
Now consider this graph. Is there an unblocked backdoor path from T to Y?



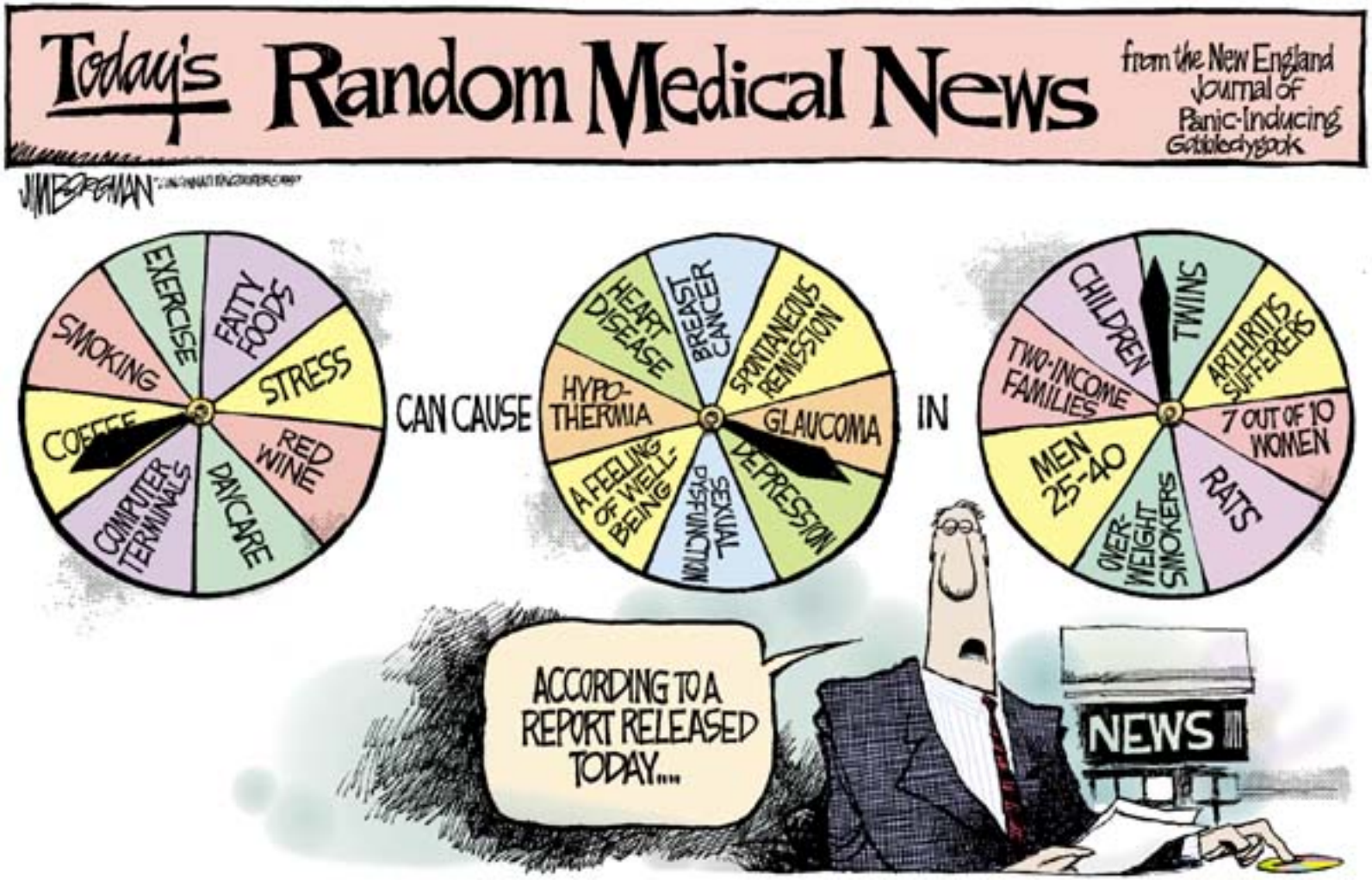
No need to condition! X_2 already blocks this path. it is a collider.

Thoughts on DAGs and Potential Outcomes

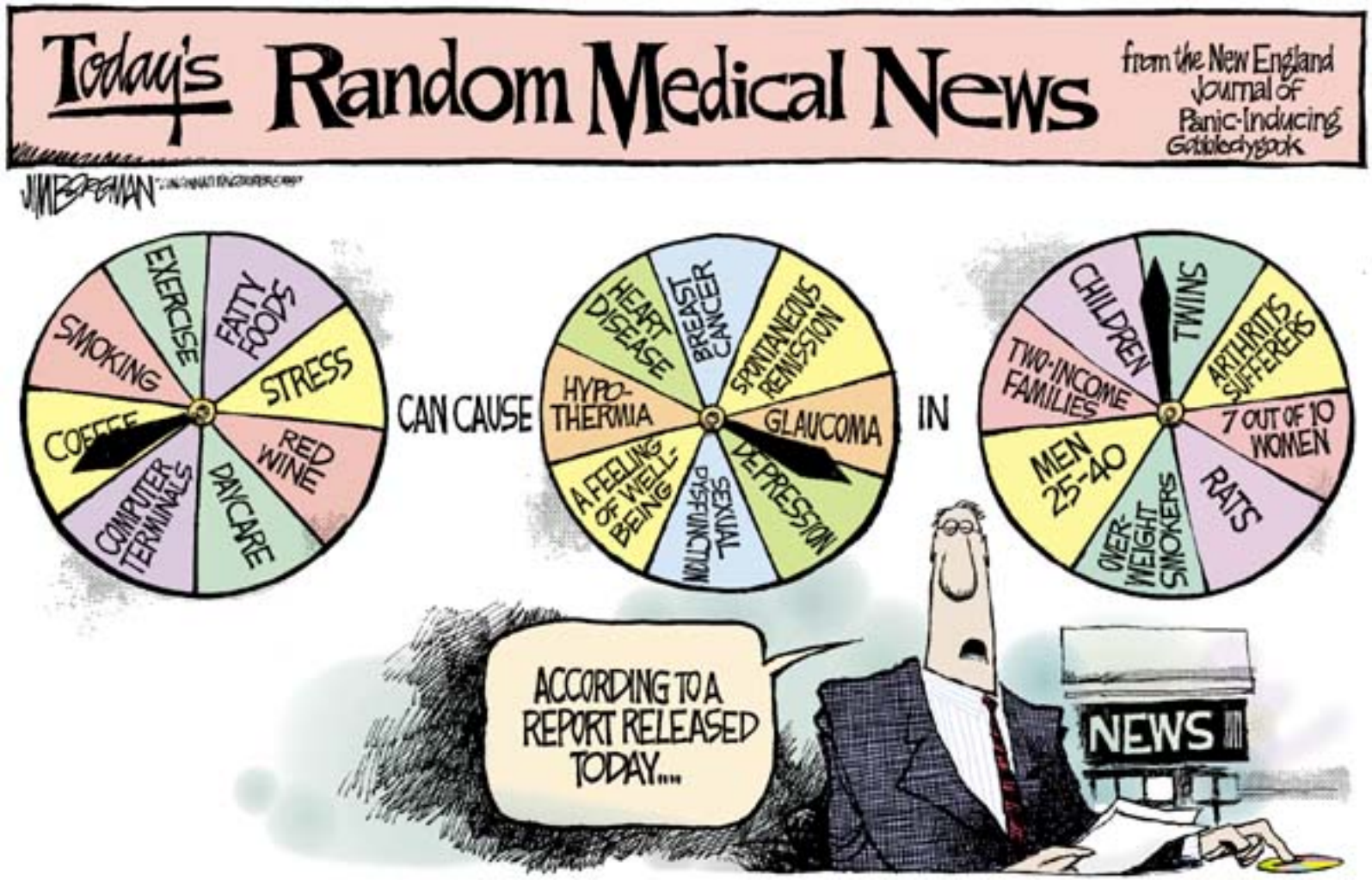
- Two very different languages for talking about and thinking about causal inferences.
- Potential outcomes is very focused on thinking about the treatment assignment mechanism and helpful for heterogeneity of treatment effects.
- Potential outcomes is also less of a "foreign language" for most statisticians, but in my experience lumps together a lot of identification assumptions in opaque ignorability conditions.
- Graphical Models with DAGs are very visually appealing but the operations on the graph can be challenging
- DAGs very helpful for thinking through identification and the entire causal process
- Note that both are about non-parametric identification and not estimation.
  - Good: provides a very general framework that applies in non-linear scenarios and interactions
  - Tricky: identification results for identification only holds when variable is completely controlled for (which may be difficult!)



Lancet 2001: negative correlation between coronary heart disease mortality and level of vitamin C in bloodstream (controlling for age, gender, blood pressure, diabetes, and smoking)



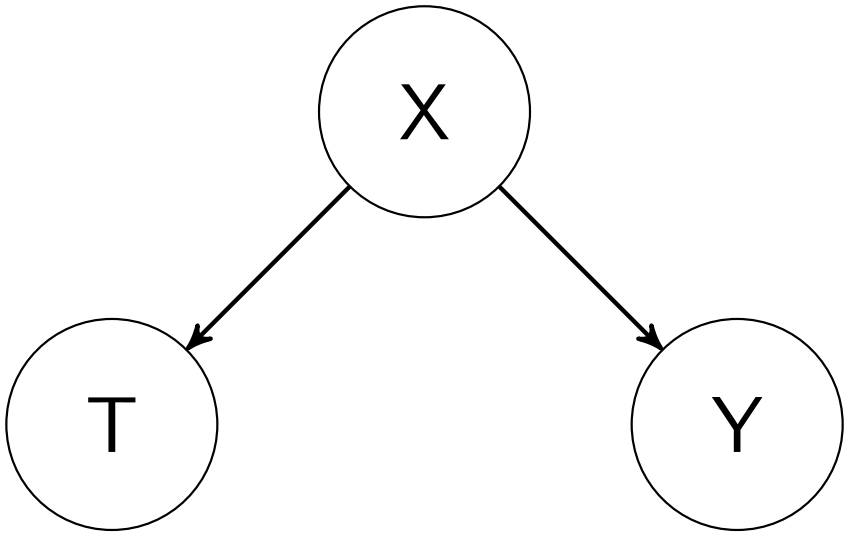
Lancet 2002: no effect of vitamin C on mortality in controlled placebo trial (controlling for nothing)



Lancet 2003: comparing among individuals with the same age, gender, blood pressure, diabetes, and smoking, those with higher vitamin C levels have lower levels of obesity, lower levels of alcohol consumption, are less likely to grow up in working class, etc.

Why So Much Variation?

Confounders



Observational Studies and Experimental Ideal

- Randomization forms gold standard for causal inference, because it balances observed and unobserved confounders
- Cannot always randomize so we do observational studies, where we adjust for the observed covariates and hope that unobservables are balanced
- Better than hoping: design observational study to approximate an experiment
  - "The planner of an observational study should always ask himself [sic]: How would the study be conducted if it were possible to do it by controlled experimentation" (Cochran 1965)

Experiment review

- An experiment is a study where assignment to treatment is controlled by the researcher.
  - $P(T_i = 1)$ is controlled and known by researcher.
- In the ideal randomized experiment, two assumptions hold by design:
  1. Positivity: assignment is probabilistic: $0 < P(T_i = 1) < 1$
     - ★ No deterministic assignment.
  2. Ignorability: $P(T_i = 1 | \mathbf{Y}(1), \mathbf{Y}(0)) = P(T_i = 1)$
     - ★ Treatment assignment does not depend on any potential outcomes.
     - ★ Sometimes written as $T_i \perp\!\!\!\perp (\mathbf{Y}(1), \mathbf{Y}(0))$

Why do Experiments Help?

Remember selection bias?

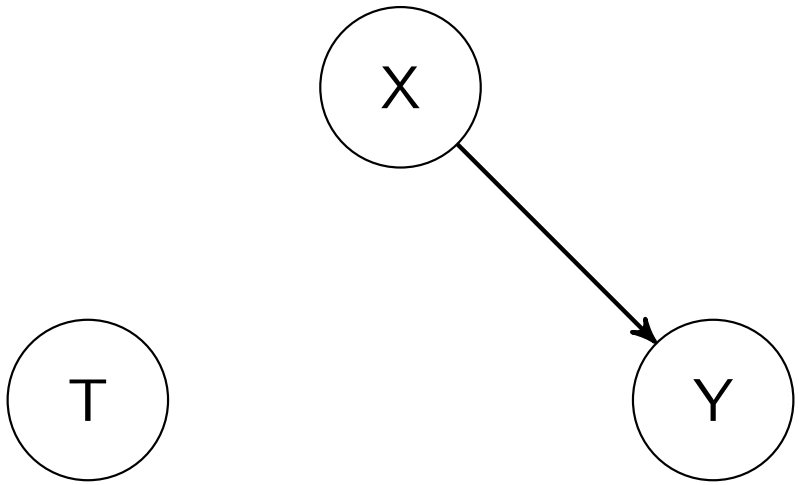
$$
\begin{aligned}
&E[Y_i|T_i=1] - E[Y_i|T_i=0] \\
&\quad = E[Y_i(1)|T_i=1] - E[Y_i(0)|T_i=0] \\
&\quad = E[Y_i(1)|T_i=1] - E[Y_i(0)|T_i=1] + E[Y_i(0)|T_i=1] - E[Y_i(0)|T_i=0] \\
&\quad = \underbrace{E[Y_i(1) - Y_i(0)|T_i=1]}_{\text{Average Treatment Effect on Treated}} + \underbrace{E[Y_i(0)|T_i=1] - E[Y_i(0)|T_i=0]}_{\text{selection bias}}
\end{aligned}
$$

In an experiment we know that treatment is randomly assigned. Thus we can do the following:

$$E[Y_i(1)|T_i=1] - E[Y_i(0)|T_i=0] = E[Y_i(1)] - E[Y_i(0)]$$

When all goes well, an experiment eliminates selection bias.

Randomization Removes the Arrows into the Treatment



This ensures any observed or unobserved pretreatment covariates have the same distribution in the treatment and the control group. That is, they are balanced.

Stratified Designs

- Stratified randomized experiment seek to remove bad randomizations where covariates are unbalanced by chance.
- Core Procedure:
  - form J blocks, b_j, $j = 1, \ldots, J$ based on the covariates
  - completely randomized assignment within each block.
  - randomization probability depends on the block variable, B_i
  - conditional ignorability: $T_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | B_i$.
- Generally, stratified designs mean that the probability of treatment depends on a covariate, X_i: $P(T_i = 1 | X_i = x)$.
- Conditional randomization assumptions:
  - Positivity: $0 < P(T_i = 1 | X_i = x) < 1$ for all i and x.
  - Unconfoundedness: $P(T_i = 1 | \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)) = P(T_i = 1 | X_i)$

Identification for Stratified Random Experiments

Can we identify the ATE under these stratified designs? Yes!

$$\begin{aligned}
\mathbb{E}[Y_i(1) - Y_i(0)] &= \mathbb{E}_X\Big\{\mathbb{E}[Y_i(1) - Y_i(0)|X_i]\Big\} && \text{(iterated expectations)} \\
&= \mathbb{E}_X\Big\{\mathbb{E}[Y_i(1)|X_i] - \mathbb{E}[Y_i(0)|X_i]\Big\} \\
&= \mathbb{E}_X\Big\{\mathbb{E}[Y_i(1)|T_i = 1, X_i] - \mathbb{E}[Y_i(0)|T_i = 0, X_i]\Big\} && \text{(ignorability)} \\
&= \mathbb{E}_X\Big\{\mathbb{E}[Y_i|T_i = 1, X_i] - \mathbb{E}[Y_i|T_i = 0, X_i]\Big\} && \text{(SUTVA)}
\end{aligned}$$

- ATE is just the weighted average of the within-strata differences in means.
- Identified because the last line is a function of observables.
- The averaging is over the distribution of the strata $\rightsquigarrow$ size of the blocks.

Experiments

- The benefit of experiments is that key decisions hold by design and that you have balance along observed AND unobserved covariates.
- We aren't really covering experiments in this class as a broader discussion would require discussion of topics like randomization schemes, blocking, power calculations, internal/external validity, pre-registration and ethics.
- We cover a bit because experiments are a dominant metaphor. Much of the work on observational studies is trying to find a circumstance where we can pretend we have a stratified experiment.
- Of course, in real experiments sometimes people don't always do what we say (compliance problems).

Avoiding Common Areas of Confusion

- contribution not attribution: we care about a difference which doesn't make it the main reason, nor does it imply a morality claim, it doesn't make T the reason it happened, it doesn't mean that T is "responsible" for Y
- T can 'cause' Y if it is neither necessary nor sufficient
- If you know that on average A causes B and B causes C this doesn't mean you know that A causes C (example A$\rightarrow$B for one subgroup, B$\rightarrow$C for second subgroup, still no A$\rightarrow$C)
- estimation of causal effects does not require identical treatment and control groups
- you need a clear counterfactual to have a well-defined causal effect. For example of 'the recession was caused by Wall Street' may make intuitive sense but is it well-defined?

http://egap.org/methods-guides/10-things-you-need-know-about-causal-inference

Designing Observational Studies

- Rubin (2008) argues that we should still "design" our observational studies:
  - Pick the ideal experiment to this observational study.
  - Hide the outcome data (minimize snooping!).
  - Try to estimate the randomization procedure.
  - Analyze this as a stratified randomized experiment with this estimated procedure.
- This is the perspective on observational design that comes out of the potential outcomes framework.
- Core idea: imagine that your data was generated by a stratified randomized experiment.

Identification Under Selection on Observables

Identification Assumption

1. $(Y(1), Y(0)) \perp\!\!\!\perp T|X$ *(selection on observables)*
2. $0 < P(T = 1|X) < 1$ *with probability one (common support)*

Identification Result

Given selection on observables we have

$$\begin{aligned} \mathbb{E}[Y(1) - Y(0)|X] &= \mathbb{E}[Y(1) - Y(0)|X, T = 1] \\ &= \mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0] \end{aligned}$$

Therefore, under the positivity assumption:

$$\begin{aligned} \tau_{ATE} &= \mathbb{E}[Y(1) - Y(0)] = \int \mathbb{E}[Y(1) - Y(0)|X]\, dP(X) \\ &= \int \left(\mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0]\right) dP(X) \end{aligned}$$

Identification Under Selection on Observables

Identification Assumption

1. $(Y(1), Y(0)) \perp\!\!\!\perp T|X$ *(selection on observables)*
2. $0 < P(T = 1|X) < 1$ *with probability one (common support)*

Identification Result

Similarly, for the Average Treatment Effect on the Treated (ATT),

$$\begin{aligned} \tau_{ATT} &= \mathbb{E}[Y(1) - Y(0)|T = 1] \\ &= \int \left(\mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0]\right) dP(X|T = 1) \end{aligned}$$

To identify τ_{ATT} the selection on observables and common support conditions can be relaxed to:

- $Y(0) \perp\!\!\!\perp T|X$ *(Selection on Observables for Controls)*
- $P(T = 1|X) < 1$ *(Weak Positivity)*

Identification Under Selection on Observables

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	$Y_i(1)$	$Y_i(0)$	T_i	X_i
1	$\mathbb{E}[Y(1)\|X=0,T=1]$	$\mathbb{E}[Y(0)\|X=0,T=1]$	1	0
2			1	0
3	$\mathbb{E}[Y(1)\|X=0,T=0]$	$\mathbb{E}[Y(0)\|X=0,T=0]$	0	0
4			0	0
5	$\mathbb{E}[Y(1)\|X=1,T=1]$	$\mathbb{E}[Y(0)\|X=1,T=1]$	1	1
6			1	1
7	$\mathbb{E}[Y(1)\|X=1,T=0]$	$\mathbb{E}[Y(0)\|X=1,T=0]$	0	1
8			0	1

Identification Under Selection on Observables

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	$Y_i(1)$	$Y_i(0)$	T_i	X_i
1	$\mathbb{E}[Y(1)\|X=0,T=1]$	$\mathbb{E}[Y(0)\|X=0,T=1]=$	1	0
2		$\mathbb{E}[Y(0)\|X=0,T=0]$	1	0
3	$\mathbb{E}[Y(1)\|X=0,T=0]$	$\mathbb{E}[Y(0)\|X=0,T=0]$	0	0
4			0	0
5	$\mathbb{E}[Y(1)\|X=1,T=1]$	$\mathbb{E}[Y(0)\|X=1,T=1]=$	1	1
6		$\mathbb{E}[Y(0)\|X=1,T=0]$	1	1
7	$\mathbb{E}[Y(1)\|X=1,T=0]$	$\mathbb{E}[Y(0)\|X=1,T=0]$	0	1
8			0	1

$(Y(1), Y(0)) \perp\!\!\!\perp T|X$ implies that we conditioned on all confounders. The treatment is randomly assigned within each stratum of X:

$$\mathbb{E}[Y(0)|X=0,T=1] = \mathbb{E}[Y(0)|X=0,T=0] \text{ and}$$
$$\mathbb{E}[Y(0)|X=1,T=1] = \mathbb{E}[Y(0)|X=1,T=0]$$

Identification Under Selection on Observables

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	$Y_i(1)$	$Y_i(0)$	T_i	X_i
1	$\mathbb{E}[Y(1)\|X=0,T=1]$	$\mathbb{E}[Y(0)\|X=0,T=1]=$	1	0
2		$\mathbb{E}[Y(0)\|X=0,T=0]$	1	0
3	$\mathbb{E}[Y(1)\|X=0,T=0]=$	$\mathbb{E}[Y(0)\|X=0,T=0]$	0	0
4	$\mathbb{E}[Y(1)\|X=0,T=1]$		0	0
5	$\mathbb{E}[Y(1)\|X=1,T=1]$	$\mathbb{E}[Y(0)\|X=1,T=1]=$	1	1
6		$\mathbb{E}[Y(0)\|X=1,T=0]$	1	1
7	$\mathbb{E}[Y(1)\|X=1,T=0]=$	$\mathbb{E}[Y(0)\|X=1,T=0]$	0	1
8	$\mathbb{E}[Y(1)\|X=1,T=1]$		0	1

$(Y(1), Y(0)) \perp\!\!\!\perp T|X$ also implies

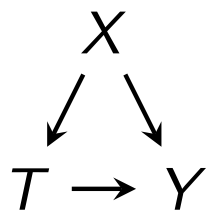
$$\begin{aligned} \mathbb{E}[Y(1)|X=0,T=1] &= \mathbb{E}[Y(1)|X=0,T=0] \text{ and} \\ \mathbb{E}[Y(1)|X=1,T=1] &= \mathbb{E}[Y(1)|X=1,T=0] \end{aligned}$$

Big problem

- How can we determine if no unmeasured confounding holds if we didn't assign the treatment?
- Put differently:
  - What covariates do we need to condition on?
- One way, from the assumption itself:
  - $P[T_i = 1|\mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = P[T_i = 1|\mathbf{X}]$
  - Include covariates such that, conditional on them, the treatment assignment does not depend on the potential outcomes.
- Another way: use DAGs and look at back-door paths.

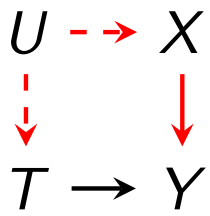
Backdoor paths and blocking paths

- Backdoor path: is a non-causal path from T to Y.
  - Would remain if we removed any arrows pointing out of T.
- Backdoor paths between T and $Y \rightsquigarrow$ common causes of T and Y:



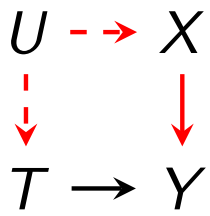
- Here there is a backdoor path $T \leftarrow X \rightarrow Y$, where X is a common cause for the treatment and the outcome.

Other types of confounding



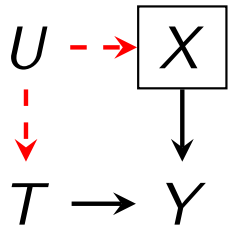
- T is enrolling in a job training program.
- Y is getting a job.
- U is being motivated
- X is number of job applications sent out.
- Big assumption here: no arrow from U to Y.

Other types of confounding



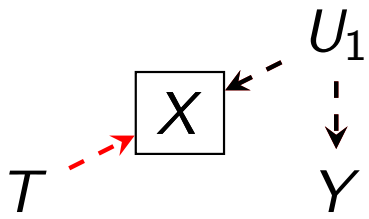
- T is exercise.
- Y is having a disease.
- U is lifestyle.
- X is smoking
- Big assumption here: no arrow from U to Y.

What's the problem with backdoor paths?



- A path is blocked if:
  1. we control for or stratify a non-collider on that path OR
  2. we do not control for a collider.
- Unblocked backdoor paths $\rightsquigarrow$ confounding.
- In the DAG here, if we condition on X, then the backdoor path is blocked.

Not all backdoor paths



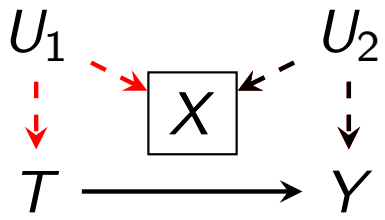
- Conditioning on the posttreatment covariates opens the non-causal path.
  - $\rightsquigarrow$ selection bias.

Don't condition on post-treatment variables

Every time you do, a puppy cries.

(just kidding. but seriously, this is one of the easiest ways to mess up your analysis if you don't know what you are doing.)

M-bias



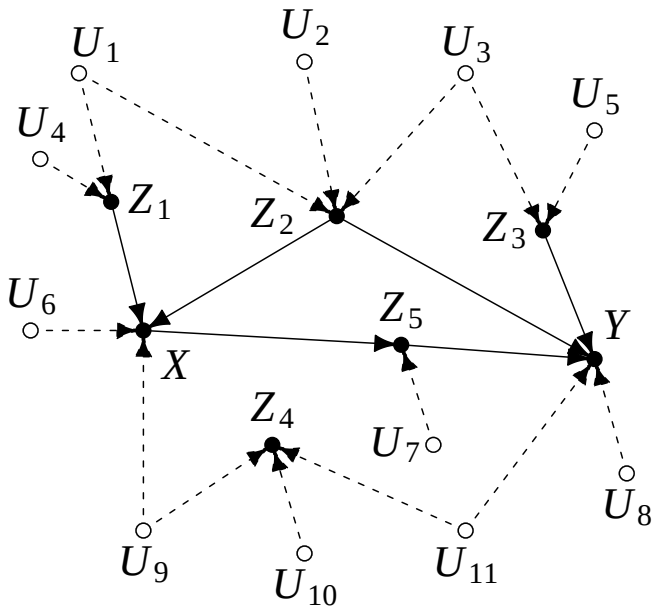
- Not all backdoor paths induce confounding.
- This backdoor path is blocked by the collider X that we don't control for.
- If we control for $X \rightsquigarrow$ opens the path and induces confounding.
  - Sometimes called M-bias.
- Controversial because of differing views on what to control for:
  - Rubin thinks that M-bias is a "mathematical curiosity" and we should control for all pretreatment variables
  - Pearl and others think M-bias is a real threat.
  - See the Elwert and Winship piece for more!

Backdoor criterion

- Can we use a DAG to argue for no unmeasured confounders?
- Pearl answered yes, with the backdoor criterion, which states that the effect of T on Y is identified if:
  1. No backdoor paths from T to Y OR
  2. Measured covariates are sufficient to block all backdoor paths from T to Y.
- First is really only valid for randomized experiments.
- The backdoor criterion is fairly powerful. Tells us:
  - if there is confounding given this DAG,
  - if it is possible to remove the confounding, and
  - what variables to condition on to eliminate the confounding.

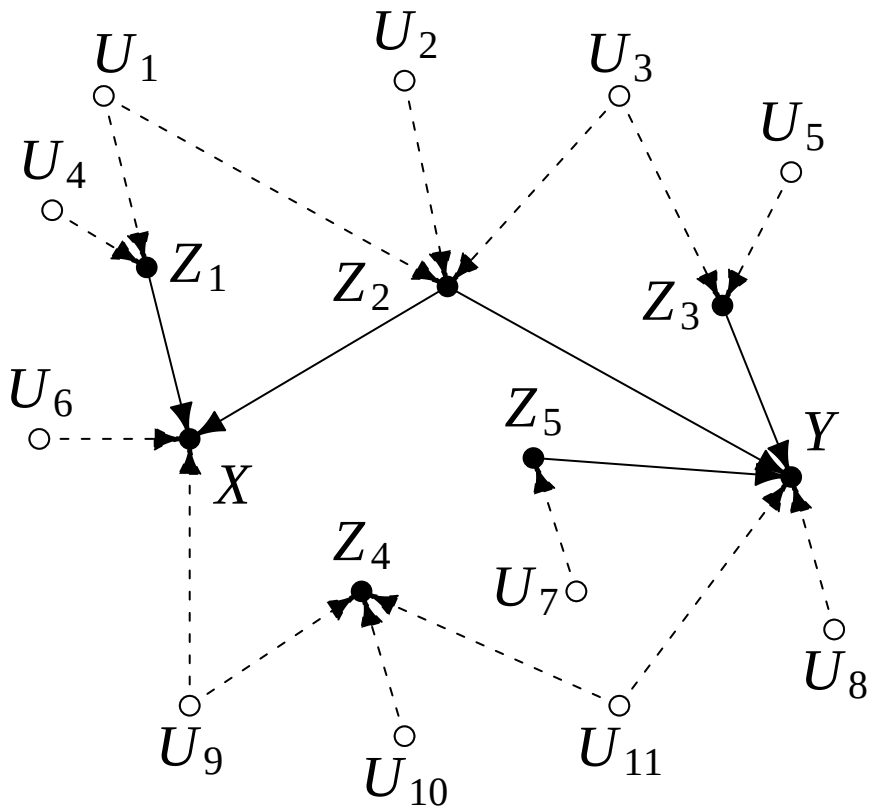
Example: Sufficient Conditioning Sets

We want to estimate the effect of X on Y for this DAG.



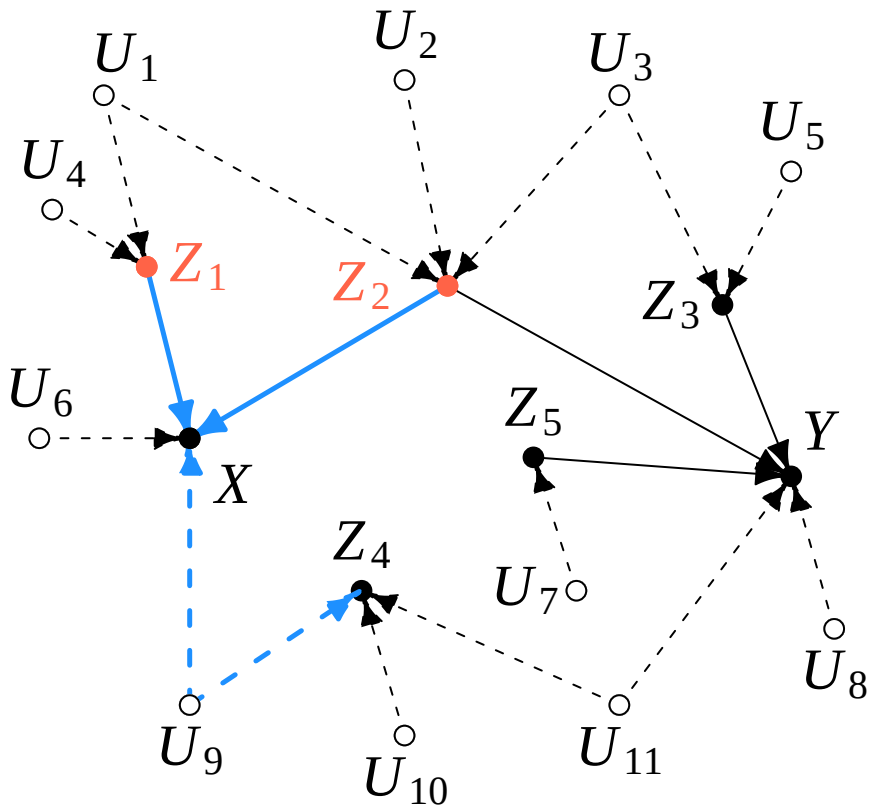
Remove arrows out of X.

Example: Sufficient Conditioning Sets



Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

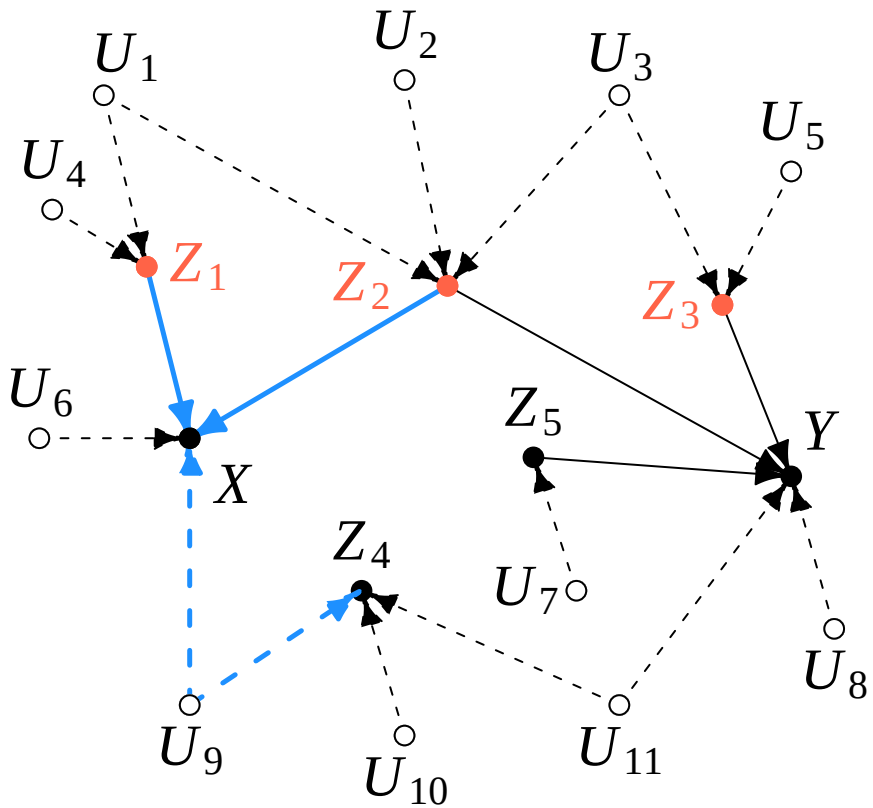
Example: Sufficient Conditioning Sets



No unblocked backdoor paths if we condition on Z_1 and Z_2

Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

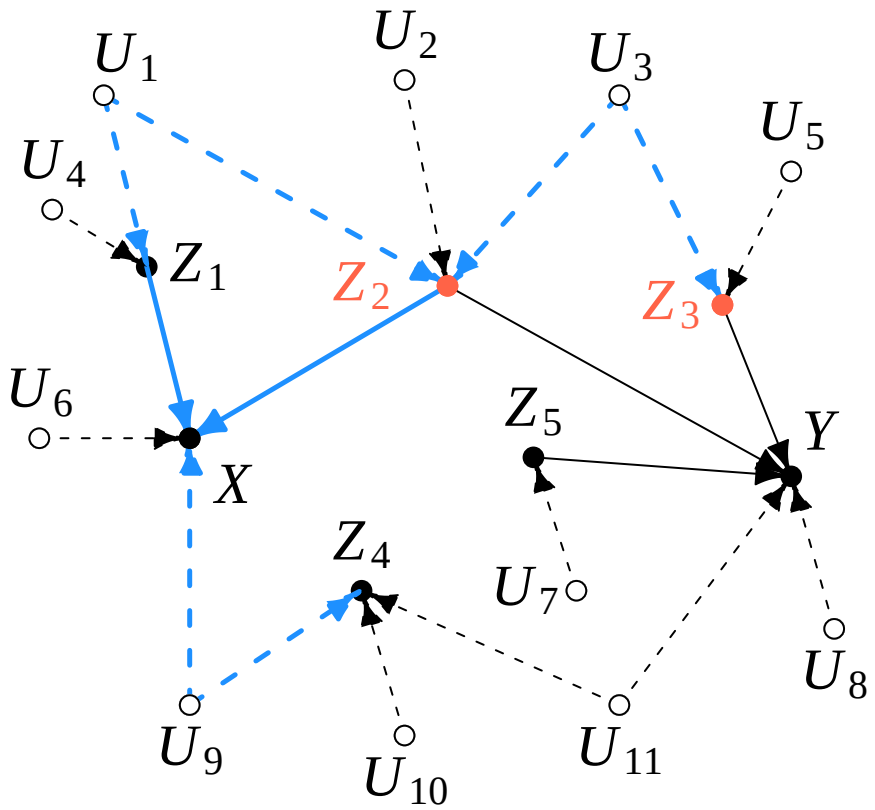
Example: Sufficient Conditioning Sets



No unblocked backdoor paths if we condition on Z_1, Z_2, and Z_3

Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

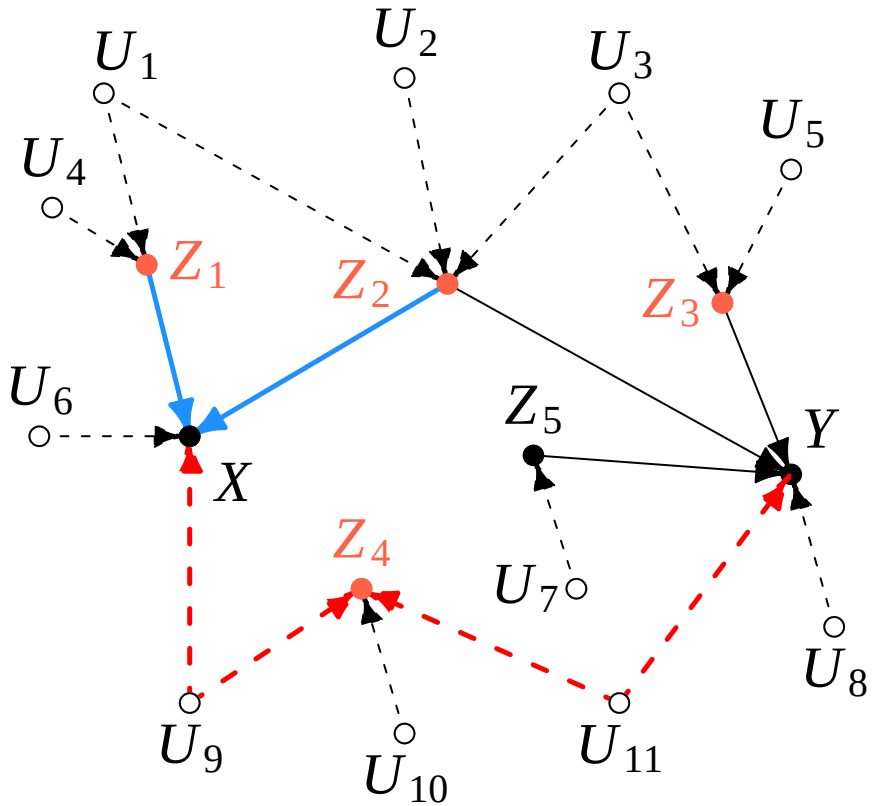
Example: Sufficient Conditioning Sets



No unblocked backdoor paths if we condition on Z_2 and Z_3

Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

Example: Non-sufficient Conditioning Sets

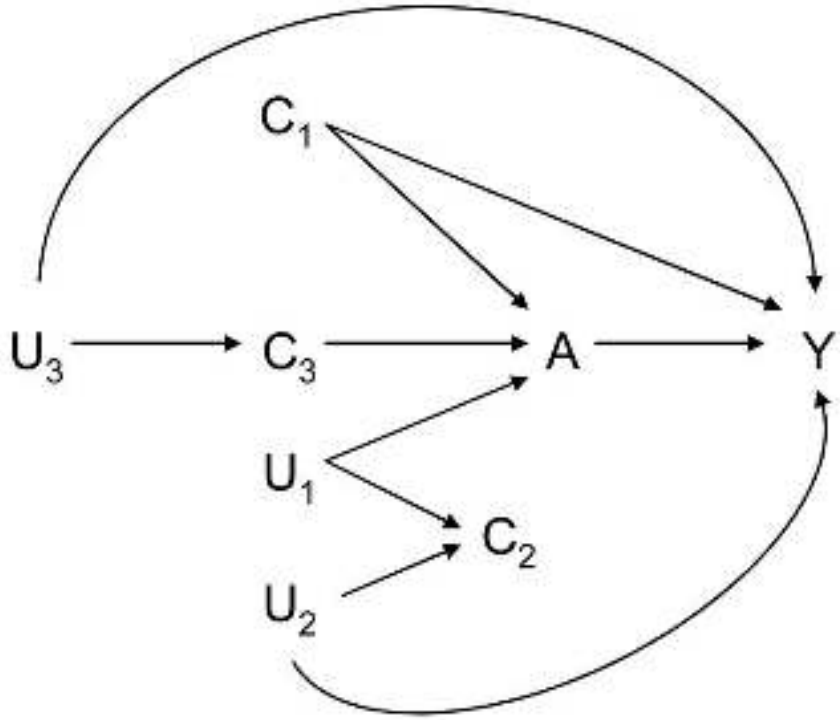


There are unblocked paths if we condition on Z_1, Z_2, Z_3, Z_4

Recall that paths are blocked by "unconditioned colliders" or conditioned non-colliders

More examples in Morgan and Winship

Implications (via Vanderweele and Shpitser 2011)



Two common criteria fail here:

1. Choose all pre-treatment covariates
   (would condition on C_2 inducing M-bias)
2. Choose all covariates which directly cause the treatment and the outcome
   (would leave open a backdoor path $A \leftarrow C_3 \leftarrow U_3 \rightarrow Y$.)

No unmeasured confounders is not testable

- No unmeasured confounding places no restrictions on the observed data.

$$\underbrace{\left(Y_i(0) \middle| T_i = 1, X_i\right)}_{\text{unobserved}} \overset{d}{=} \underbrace{\left(Y_i(0) \middle| T_i = 0, X_i\right)}_{\text{observed}}$$

- Recall, $\overset{d}{=}$ means equal in distribution.
- No way to directly test this assumption without the counterfactual data, which is missing by definition!
- With backdoor criterion, you must have the correct DAG.

Assessing no unmeasured confounders

TABLE VI
THE FOX NEWS EFFECT: INTERACTIONS AND PLACEBO SPECIFICATIONS

	Interactions		Placebo specifications		
	Presid. Rep. vote share 2000–1996		Presidential Republican vote share		
			2000–1996	1996–1992	1992–1988
Dep. var.	(1)	(2)	(3)	(4)	(5)
Availability of Fox News via cable in 2000	0.0109 (0.0042)***	0.0105 (0.0039)***	0.0036 (0.0016)**	−0.0024 (0.0031)	0.0026 (0.0026)
Availability of Fox News via cable in 2003			−0.0001 (0.0012)		

- Can do "placebo" tests, where T_i cannot have an effect (lagged outcomes, etc)
- Della Vigna and Kaplan (2007, QJE): effect of Fox News availability on Republican vote share
  - Availability in 2000/2003 can't affect past vote shares.
- Unconfoundedness could still be violated even if you pass this test!

So Where Does That Leave Us?

- If we can identify the right covariates $\mathbf{X}$ to get conditional ignorability we can do selection on observables.
- No unmeasured confounders $\approx$ randomized experiment.
- These variables are those that block backdoor paths and we want to be *really* sure we know what we are doing if we condition on a post-treatment variable.
- This still leaves us with a tricky estimation problem as we now need to work with distributions conditional on $\mathbf{X}$.

Discrete covariates

- Suppose that we knew that T_i was unconfounded within levels of a binary X_i.
- Then we could always estimate the causal effect using iterated expectations as in a stratified randomized experiment:

$$\begin{aligned}
&\mathbb{E}_X\Big\{\mathbb{E}[Y_i|\,T_i=1,X_i]-\mathbb{E}[Y_i|\,T_i=0,X_i]\Big\}\\
&=\underbrace{\Big(\mathbb{E}[Y_i|\,T_i=1,X_i=1]-\mathbb{E}[Y_i|\,T_i=0,X_i=1]\Big)}_{\text{diff-in-means for } X_i=1}\ \underbrace{\mathbb{P}[X_i=1]}_{\text{share of } X_i=1}\\
&\quad+\underbrace{\Big(\mathbb{E}[Y_i|\,T_i=1,X_i=0]-\mathbb{E}[Y_i|\,T_i=0,X_i=0]\Big)}_{\text{diff-in-means for } X_i=0}\ \underbrace{\mathbb{P}[X_i=0]}_{\text{share of } X_i=0}
\end{aligned}$$

- Stratification is great because it makes no assumptions about parametric form.
- More generally stratification just involves a weighted average of subgroup means.

Stratification Example: Smoking and Mortality (Cochran, 1968)

TABLE 1
DEATH RATES PER 1,000 PERSON-YEARS

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

Stratification Example: Smoking and Mortality (Cochran, 1968)

TABLE 2
MEAN AGES, YEARS

Smoking group	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Stratification

To control for differences in age, we would like to compare different smoking-habit groups with the same age distribution

One possibility is to use stratification:

- for each country, divide each group into different age subgroups
- calculate death rates within age subgroups
- average within age subgroup death rates using fixed weights (e.g. number of cigarette smokers)

Stratification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers?

$15 \cdot (11/40) + 35 \cdot (13/40) + 50 \cdot (16/40) = 35.5$

Stratification: Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for Pipe Smokers if they had same age distribution as Non-Smokers?

$15 \cdot (29/40) + 35 \cdot (9/40) + 50 \cdot (2/40) = 21.2$

Smoking and Mortality (Cochran, 1968)

TABLE 3

ADJUSTED DEATH RATES USING 3 AGE GROUPS

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	28.3	12.8	17.7
Cigars/pipes	21.2	12.0	14.2

Limits to Stratification

- So, great, we can stratify. Why not do this all the time?
- There are three core problems:
  1. Empty cells (combinations of covariates X_i which have either no treatment or control)
  2. Continuous covariates (empty cells are inevitable!)
  3. Estimation variance (even if we didn't have empty cells, we might have very few observations to estimate the mean making it noisy).
- One option is to discretize into bins or regularize our estimates in various ways, but for many settings we may want another approach.

Preview: From Stratification to Regression

- Nonparametric identification yielded this estimand:

$$\mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}\Big[\mathbb{E}[Y_i \mid T_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i \mid T_i = 0, \mathbf{X}_i]\Big]$$

- What if we could estimate the inner conditional expectations? This is what we will do with regression models for the remainder of the semester.
- We will sometimes write the empirical quantity of interest as

$$\mu_t(\mathbf{x}) = E[Y_i(t) \mid T = t, \mathbf{X}_i = \mathbf{x}]$$

- Ignorability tell us that the counterfactual quantities can be written in terms of these quantities based on observed data, but we still need to estimate those conditional expectations!

This Week in Review

- We introduced DAGs and explained how they help us think through confounding.
- We continued our journey into causal inference.
- The next few weeks we are going to talk about how to estimate conditional expectation functions and then we will return to other ways to use CEFs in causal inference.
- We will make liberal use of both frameworks based on whatever is the most convenient to communicate the point.

Next Time: Estimating Conditional Expectation Functions