Week 9: Frameworks for Causal Inference

Brandon Stewart[1]

Princeton

November 7–9, 2022

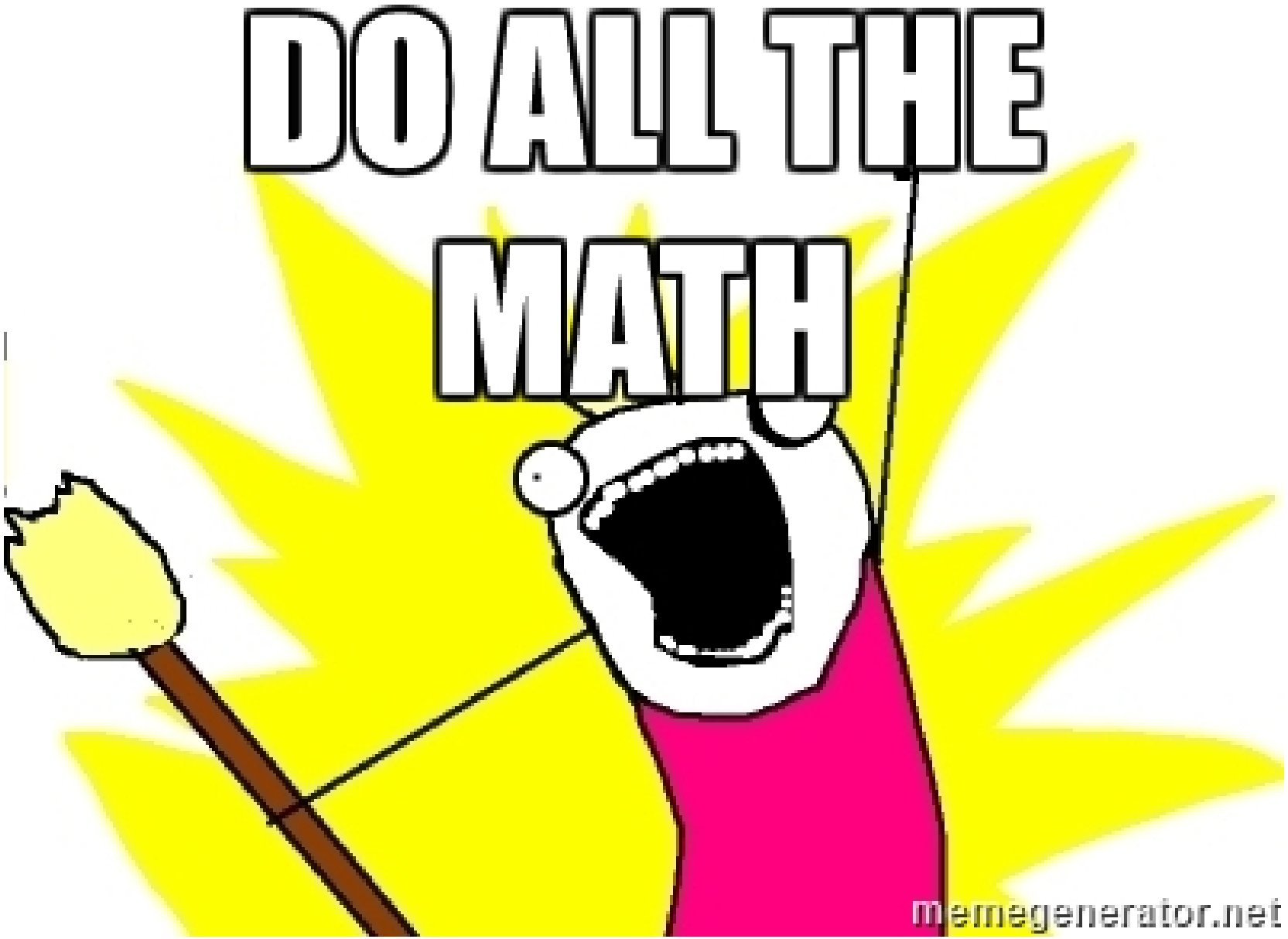
[1]These slides are heavily influenced by Matt Blackwell, Justin Grimmer, Jens Hainmueller, Erin Hartman and Kosuke Imai

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

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

Why Are We Doing All of This Again?

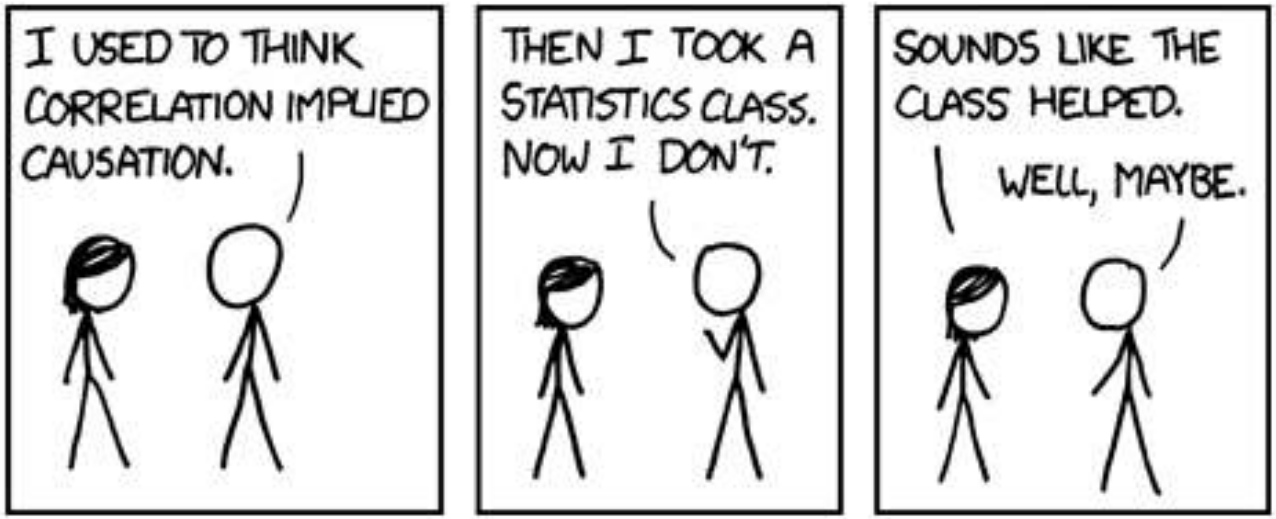
- We are all here because we are trying to do some social science, that is, we are in the business of knowledge production.
- Quantitative methods are an increasingly big part of that so whether you are reading or actively doing quantitative analysis it is going to be there.
- So why all the math?
- Reason 1: We are taking a future-oriented approach. We want to prepare you for the next big thing.
  - Methods that became popular in the social sciences since I took the equivalent of this class: machine learning, text-as-data, Bayesian nonparametrics, design-based inference, DAG-based causal inference, deep learning.
  - A technical foundation prepares you to learn new methods for the rest of your career. Trust me now is the time to invest.
- Reason 2: When you understand how the methods work you can make choices based on substantive knowledge rather than what software spits out at you.


DO ALL THE
MATH
memegenerator.net

Regression as a Tool: A Review

- Regression is a tool for approximating a conditional expectation function, we can always think of it as saying 'amongst the subgroup of units with covariates $X = x$ what is the average outcome.'
- This in turn is the best prediction of Y given X when 'best' is measured in terms of mean squared error.
- Confusion starts to creep in when we start talking about marginal effects in our prediction.
- Marginal effects are a really powerful way of summarizing differences across subgroups but they tend to lend themselves to causal interpretations that they don't necessarily have.
- This is because they are about different groups of units not about the same unit under intervention.

Causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Fundamental Problem of Causal Inference

- Causal inference is the study of counterfactuals.
- The hard thing about counterfactuals is that we never get to see all of them: Fundamental Problem of Causal Inference Holland (1986)).
- Assumptions and careful design are the only way out of this problem because we never get to see the truth.
- When it works though it can be a powerful view into the things that we care the most about.
- By convention we often call the counterfactual levels we care about treated and control and we start with binary treatment variables (continuous things get much more complicated!).

Workflow

1) Question $\leftarrow$ the thing we care about
2) Theoretical Estimand $\leftarrow$ the (often causal) quantity of interest (an ideal experiment can help!)
3) Identification Strategy $\leftarrow$ connect the theoretical estimand (can be based on unobservable distribution) to empirical estimand (feature of probability distribution of observable data).
4) Estimation $\leftarrow$ how we estimate a feature of a probability distribution from a sample of observed data.
5) Inference/Uncertainty $\leftarrow$ what would have happened if we observed a different treatment assignment? (and possibly sampled a different population)

We will read Lundberg, Johnson, and Stewart 2021 where theoretical and empirical estimand terminology comes from.

Identification

- A quantity of interest is identified when (given stated assumptions) access to infinite data would result in the estimate taking on only a single value.
- For example, having all dummy variables in a linear model is not statistically identified because they cannot be distinguished from the intercept.
- Causal identification is what we can learn about a causal effect from available data.
- If an effect is not identified, no estimation method will recover it.
- 'What's your identification strategy?' means 'what are the assumptions that allow you to claim that the association you've estimated has a causal interpretation?'
- Identification depends on assumptions not statistical models.
- As we will see this is not a conversation about estimation: in other words, if someone answers "regression" they have made a category error

Identification vs. Estimation

- Identification: How much can you learn about the estimand if you have an infinite amount of data?
- Estimation: How much can you learn about the estimand from a finite sample?
- Identification precedes estimation

The role of assumptions:

- Often identification requires (hopefully minimal) assumptions
- Even when identification is possible, estimation may impose additional assumptions (i.e. that the linear approximation to the CEF is good enough)
- Law of Decreasing Credibility (Manski): The credibility of inference decreases with the strength of the assumptions maintained

1 Core Ideas in Causal Inference

2 Potential Outcomes
- Framework
- Estimands
- Three Big Assumptions
- What Gets to Be a Cause

3 Nonparametric Structural Equation Models and Causal Directed Acyclic Graphs

4 The Experimental Ideal

The Potential Outcomes Framework

- Potential Outcomes is one of two major frameworks that we will consider for doing causal inference.
- It is a way of thinking about counterfactuals and the assumptions required to make statements about them.
- We will first step through the framework, then discuss estimands, three big assumptions and finally what counts as a cause.

Potential Outcomes

Definitions:

T_i: Dichotomous Treatment assignment for unit i (multi-valued treatments are possible too–just more potential outcomes for each unit)

$$T_i = \begin{cases} 1 & \text{Unit is assigned to treatment} \\ 0 & \text{Unit is not assigned to treatment} \end{cases}$$

Y_i: Outcome for unit i

Potential outcomes for unit i:

$$Y_i(T_i) = \begin{cases} Y_i(1) & \text{Potential outcome for unit } i \text{ with treatment} \\ Y_i(0) & \text{Potential outcome for unit } i \text{ without treatment} \end{cases}$$

Pre-treatment covariates X_i

τ_i: The treatment effect

$$\tau_i = Y_i(1) - Y_i(0)$$

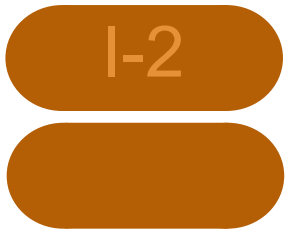
Potential Outcomes – Aspirin Example

Definitions:

T_i: Unit assigned to:

$$T_i = \begin{cases} 1 & \text{Receive Aspirin} \\ 0 & \text{Receive Placebo} \end{cases}$$

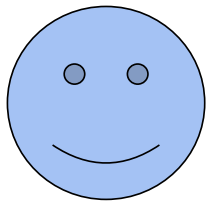
$(T_i = 1)$



$(T_i = 0)$

Y_i: Outcome for unit i – Patient has headache, or not

$(Y_i = 1)$



$(Y_i = 0)$

Potential outcomes for unit i:

$$Y_i(T_i) = \begin{cases} Y_i(1) & \text{Headache (or not) for unit } i \text{ with Aspirin} \\ Y_i(0) & \text{Headache (or not) for unit } i \text{ with placebo} \end{cases}$$

Pre-treatment covariates X_i

Illustrated potential outcomes here and later courtesy of Erin Hartman

What is random in the potential outcomes framework?

Note that potential outcomes are thought of as fixed, and that they, and the difference between them, can vary by arbitrary amounts for each unit i. There is some true distribution of potential outcomes across the population.

Treatment assignment is the source of randomness

Causal Inference is a Missing Data Problem

Definition: Observed Outcome

$$Y_i = T_i * Y_i(1) + (1 - T_i) * Y_i(0)$$

Inherently, since we cannot observe both treatment and control for unit i, thus we only observe Y_i, causal inference suffers from a **missing data problem**.

No methodology allows us to simultaneously observe both potential outcomes, $Y_i(1)$ and $Y_i(0)$, making τ_i unobservable–and unidentifiable without additional assumptions (Fundamental Problem of Causal Inference Holland (1986))

Causal Inference is a Missing Data Problem

Example: Asprin's Impact on Headaches

Patient i		Pill T_i	Headache Status $Y_i(0)$	$Y_i(1)$	Y_i	Age X_{1i}	Academic X_{2i}
1		1	0	0	0	25	Y
2			0	1		55	N
3			1	1		62	Y
4		0	1	1	1	80	N
5		1	0	1	1	32	Y
6			1	0		45	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n		0	0	0	0	71	N

Some Estimands of Interest

- **Sample average treatment effect (SATE)**
  $\frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$
- **Population average treatment effect (PATE)**
  $\frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$
- **Population average treatment effect for the treated (PATT)**
  $\mathbb{E}(Y_i(1) - Y_i(0) \mid T_i = 1)$
- **Population conditional average treatment effect (CATE)**
  $\mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = x)$
- Treatment effect heterogeneity: Zero ATE doesn't mean zero effect for everyone

1 Core Ideas in Causal Inference

2 Potential Outcomes
- Framework
- Estimands
- Three Big Assumptions
- What Gets to Be a Cause

3 Nonparametric Structural Equation Models and Causal Directed Acyclic Graphs

4 The Experimental Ideal

Built in Assumptions

The notation implies three related assumptions:

- No simultaneity
- No interference
  - We are implicitly stating that the potential outcomes for that unit are unaffected by the treatment status of other units
  - If this is not true, the number of potential outcomes for unit i grows
  - Ex: in an experiment with 3 units, if the potential outcomes for unit i depend on the treatment assignment of units j and k, the potential outcomes for unit i are defined by $Y(i, j, k)$:

$$\begin{array}{ll} Y(1,0,0) & Y(0,0,0) \\ Y(1,1,0) & Y(0,1,0) \\ Y(1,0,1) & Y(0,0,1) \\ Y(1,1,1) & Y(0,1,1) \end{array}$$

- Same version of the treatment

How do we proceed?

Combined, the previous assumptions give us

- **S**table **U**nit **T**reatment **V**alue **A**ssumption (SUTVA)
- Potential violations:
  - feedback effects
  - spill-over effects, carry-over effects
  - different treatment administration

We also need to assume Positivity $0 < P(T_i = 1) < 1 \; \forall \; i$ with probability 1.

Ignorability

Identification by randomization:

- If treatment is randomized, then treatment is unrelated to any and all underlying characteristics, observed and unobserved (and even unknown)
- Randomization therefore means treatment assignment is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$, i.e.

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i$$

- This is sometimes called unconfoundedness or ignorability
- Another way of thinking of it: The distributions of the potential outcomes $(Y_i(1),\ Y_i(0))$ are the same for the treatment and control group.
- Yet another way of thinking of it: The treatment and control group are exchangeable, or balanced (on observables and unobservables) on average

How do we proceed?

Identification by randomization:

- Assume a completely randomized trial (fixed number of units assigned to treatment)
- Assumption: for all $i = 1, \ldots, n, \sum_{i=1}^{n} T_i = n_t$ and

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i, \qquad Pr(T_i = 1) = \frac{n_t}{n}, \qquad n = n_t + n_c$$

- Estimand of interest: SATE
- Estimator: Difference-in-means:

$$\begin{aligned}\hat{\tau} &= \frac{1}{n_t}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_c}\sum_{i=1}^{n}(1 - T_i) Y_i \\ &= \frac{1}{n_t}\sum_{i=1}^{n} T_i Y_i(1) - \frac{1}{n_c}\sum_{i=1}^{n}(1 - T_i) Y_i(0)\end{aligned}$$

How can we estimate the difference in means with regression?
`lm(y ~ T)`. Recall that, when our only regressor is binary, $\hat{\beta}$ estimates the difference-in-means.

Unbiasedness holds in sample just by randomization and SUTVA!

Design-based Inference for SATE

Variance of $\hat{\tau}$:

$$V[\hat{\tau}] = \frac{1}{n}\left(\frac{n_c}{n_t}S_1^2 + \frac{n_t}{n_c}S_0^2 + 2S_{0,1}\right)$$

where for $t = 0, 1$,

$$S_t^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(t) - \overline{Y(t)})^2 \qquad \text{sample variance of } Y_i(t)$$

$$S_{01}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\Big(Y_i(1) - \overline{Y(1)})(Y_i(0) - \overline{Y(0)})\Big)$$

Is the variance of our estimator is identifiable?

Conservative estimator:

$$\hat{V}[\hat{\tau}] \leq \frac{\hat{S}_1^2}{n_t} + \frac{\hat{S}_0^2}{n_c}$$

How do we proceed?

Identification by conditional independence:

- If treatment is not randomized, then treatment may be related underlying characteristics, observed and unobserved, which are related to the potential outcomes
- Therefore, we need to assume that treatment assignment is independent of the potential outcomes $Y_i(1)$ and $Y_i(0)$, conditional on some pre-treatment characteristics X, i.e.

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$$

- Conditioning set should yield $Y_i(0)$, $Y_i(1)$ and T_i conditionally independent. (This is next week's topic).
- This is conditional ignorability.

The Selection Problem

- Why is this difficult? selection bias
- The core idea is that the people who get treatment might look different from those who get control and thus they are not good counterfactuals for each other.
- Let's look at what we get from a naive difference in means with a binary treatment:

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

- Naive estimator = Average Treatment Effect on Treated + Selection Bias
- Selection bias: how different the treated and control groups are in terms of their potential outcome under control.

Selection Makes Us Care About Assignment Mechanisms

Assignment Mechanism

"The process that determines which units receive which treatments, hence which potential outcomes are realized and thus can be observed, and, conversely, which potential outcomes are missing."
(Imbens and Rubin, 2015, p. 31)

Key Assumptions:

- Individualistic assignment: Limits the dependence of a particular unit's assignment probability on the values of the covariates and potential outcomes for other units
- Probabilistic assignment: Requires the assignment mechanism to imply a non-zero probability for each treatment value, for every unit
- Unconfounded assignment: Disallows dependence of the assignment mechanism on the potential outcomes

The Assignment Mechanism

- Since missing potential outcomes are unobservable we must make assumptions to fill in, i.e. **estimate** missing potential outcomes.
- In the causal inference literature, we typically make assumptions about the **assignment mechanism** to do so.

Types of Assignment Mechanisms

- random assignment
- selection on observables
- selection on unobservables

Most statistical models of causal inference attain identification of treatment effects by restricting the assignment mechanism in some way.

What Gets to Be a Cause?

We can imagine a world where individual i is assigned to treatment and control conditions

What is the Hypothetical Experiment?

Problem: Immutable (or difficult to change) characteristics

- Effect of gender on promotion
- Effect of race on traffic stops

Consider causal effect of race on traffic stops:

- Do we mean effect of officer perceiving a certain race?
- Do we mean randomly assigning race at birth?
- manipulating perceptions is a lot different from manipulating the characteristic

No Causation Without Manipulation

Caveats and Implications

- Does not dismiss claims of discrimination on immutable characteristics as illegitimate
  - Pervasive effects of racism/sexism in society
  - Suggests: we need a different empirical strategy to evaluate claims
  - What facet of institutionalized racism (or its consequences) causes racial disparities?
- Correlation problem :
  - Regression models can estimate coefficients for immutable characteristics
  - But are necessarily imprecise: what do scholars have in mind in models?
- Design Principle:
  - Pretend you could design any experiment
  - If that experiment does not exist, be concerned about interpretation

Summing Up: Neyman-Rubin causal model

- Useful for studying the "effects of causes", less so for the "causes of effects".
- No assumption of homogeneity, allows for causal effects to vary unit by unit
  - No single "causal effect", thus the need to be precise about the target estimand. (This is true even for perfect experiments.)
- Distinguishes between observed outcomes and potential outcomes.
- Causal inference is a missing data problem: we typically make assumptions about the assignment mechanism to go from descriptive inference to causal inference.

1. Core Ideas in Causal Inference

2. Potential Outcomes
   - Framework
   - Estimands
   - Three Big Assumptions
   - What Gets to Be a Cause

3. Nonparametric Structural Equation Models and Causal Directed Acyclic Graphs

4. The Experimental Ideal

Nonparametric Structural Equation Models

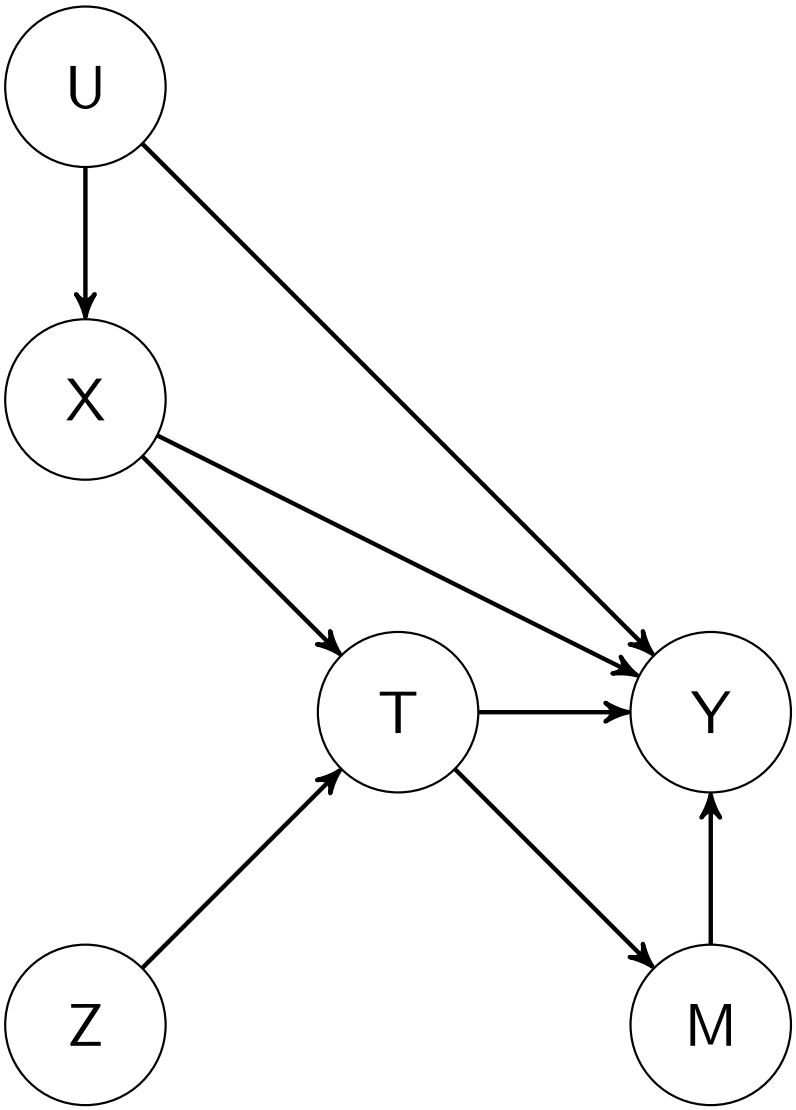
- The work we discuss here comes out of developments by Judea Pearl and others
- Idea of nonparametric structural equations (NPSEMs), is that we will write each variable as a general function of the variables that it "listens to" (i.e. could affect it directly if intervened on).
  e.g. $T = f(X, U_T)$ 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).

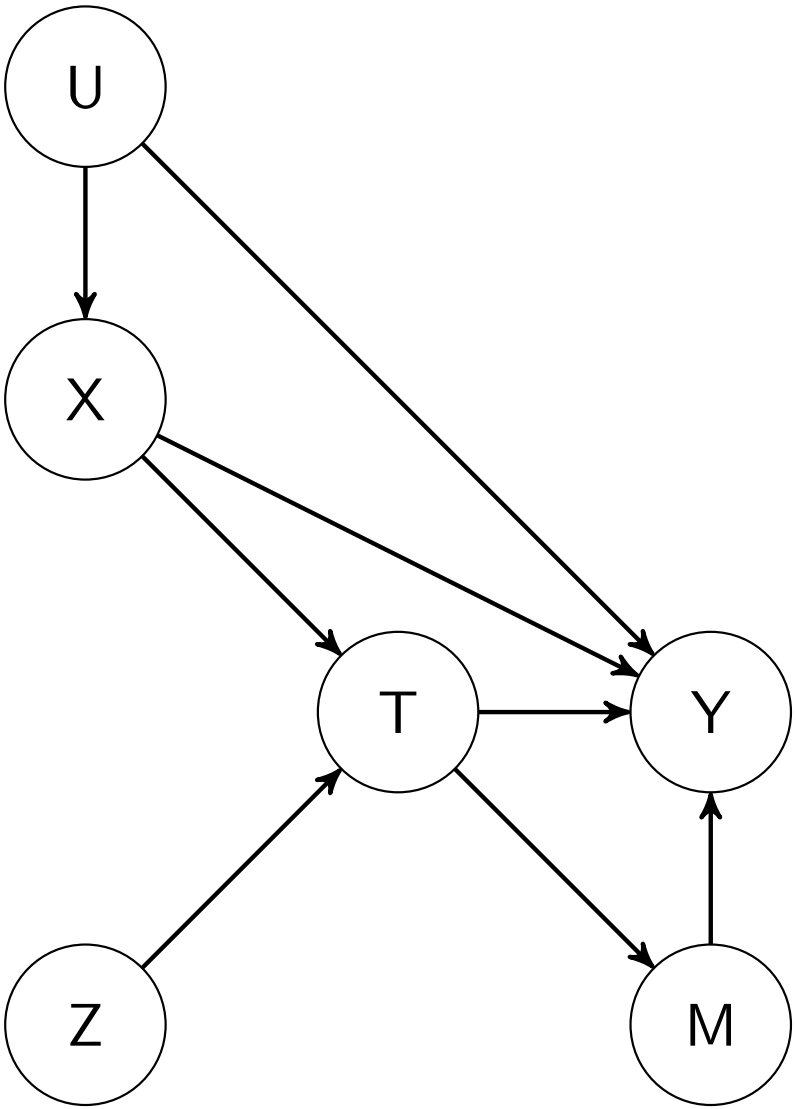
Code Idea: causal reasoning before statistical reasoning

Components of a DAG



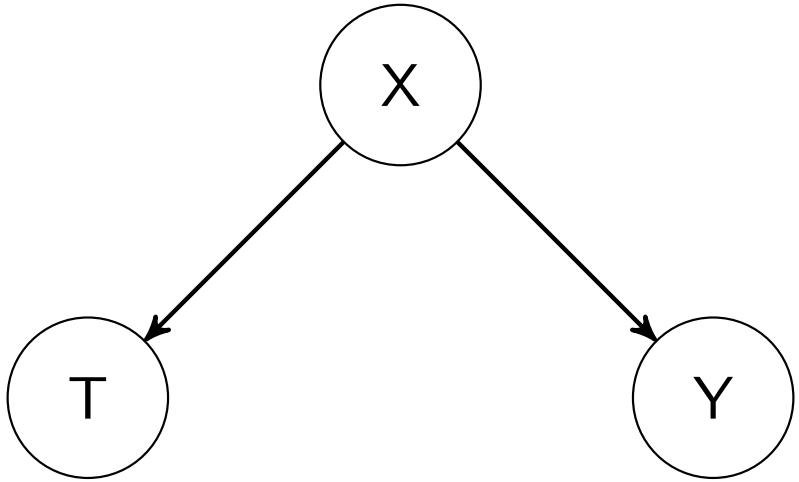
- nodes represent variables (unobserved typically called U or V)
- (directed) arrows represent causal effects
- absence of nodes represents no common causes of any pair of variables
- absence of arrows represents no causal effect
- positioning conveys no mathematical meaning but often is oriented left-to-right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

Relationships in a DAG



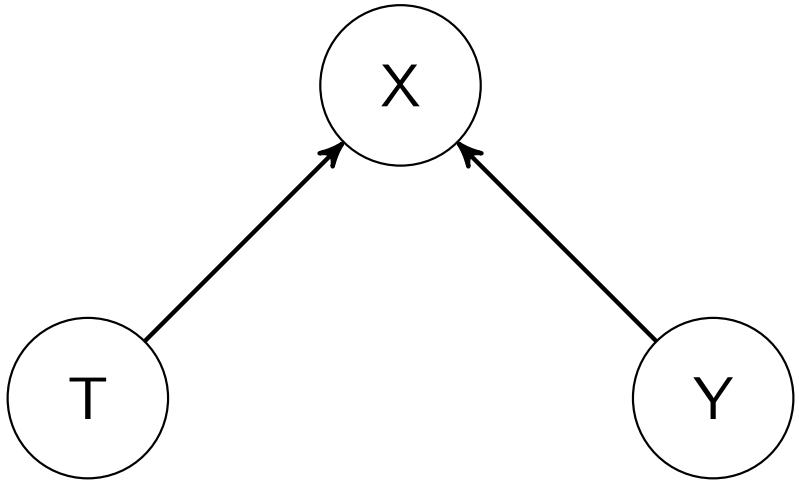
- Parents (Children): directly causing (caused by) a node
- Ancestors (Descendants): directly or indirectly causing (caused by) a node
- Path: a route that connects the variables (path is causal when all arrows point the same way)
- Acyclic implies that there are no cycles and a variable can't cause 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

Confounders



- X is 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 a confounder as blocking the flow of association.

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 they become associated and create a connection between T and Y

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

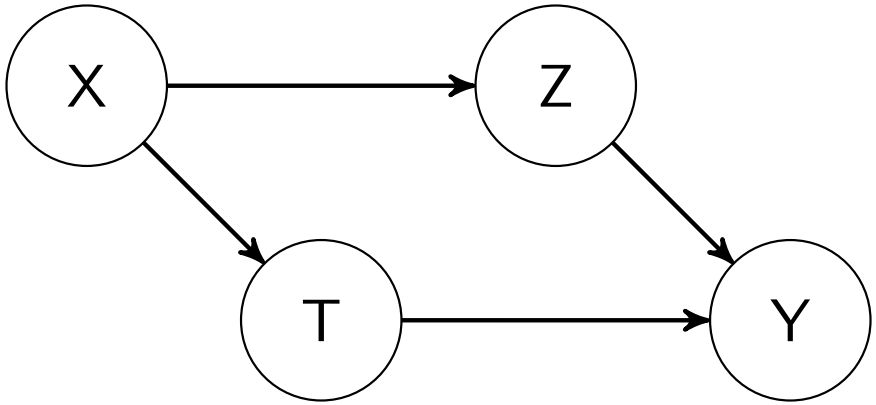
Keywords

causality, directed acyclic graphs, identification, confounding, selection

Abstract

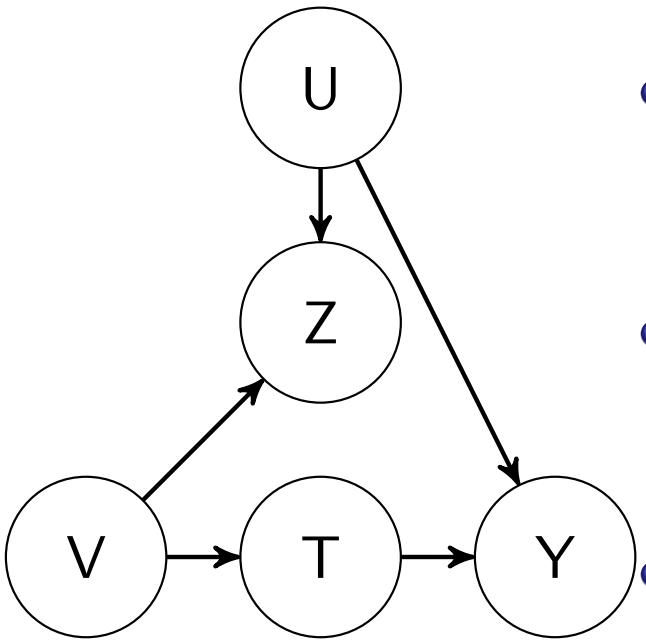
Endogenous selection bias is a central problem for causal inference. Recognizing the problem, however, can be difficult in practice. This article introduces a purely graphical way of characterizing endogenous selection bias and of understanding its consequences (Hernán et al. 2004). We use causal graphs (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.

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?

D-Separation

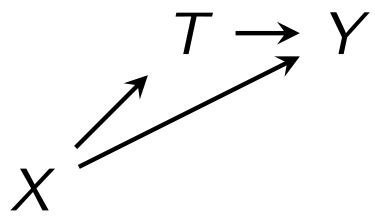
- Graphs provide us a way to think about conditional independence statements. Consider disjoint subsets of the vertices A, B and C
- A is D-separated from B by C if and only if C blocks every path from a vertex in A to a vertex in B
- A path p is said to be blocked by a set of vertices C if and only 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

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.

Blocking backdoor paths: College and earnings

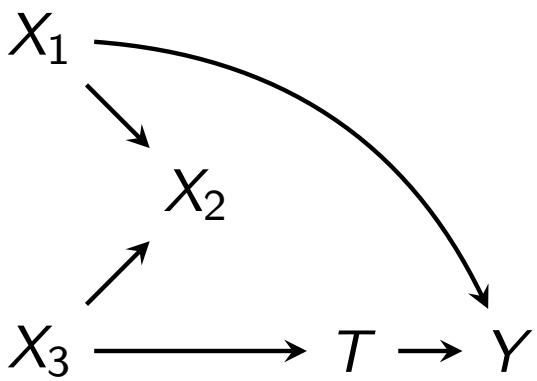
What do we need to include to block all backdoor paths between college and earnings?



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!

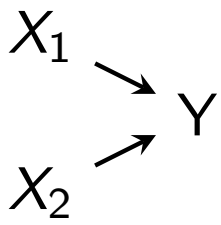
Non-causal paths: Part 2

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.

Colliders: Be careful!



Y is a collider. X_1 and X_2 are not associated, but they are when we hold Y constant.

What situations might produce this?

- X_1 being in a car accident. X_2 is having cancer. Y is being in a hospital.
- X_1 is living in a warm climate. X_2 is being an elite swimmer. Y is going swimming in January.
- X_1 is family income. X_2 is religiosity. Y attendance at a Catholic high school.

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:

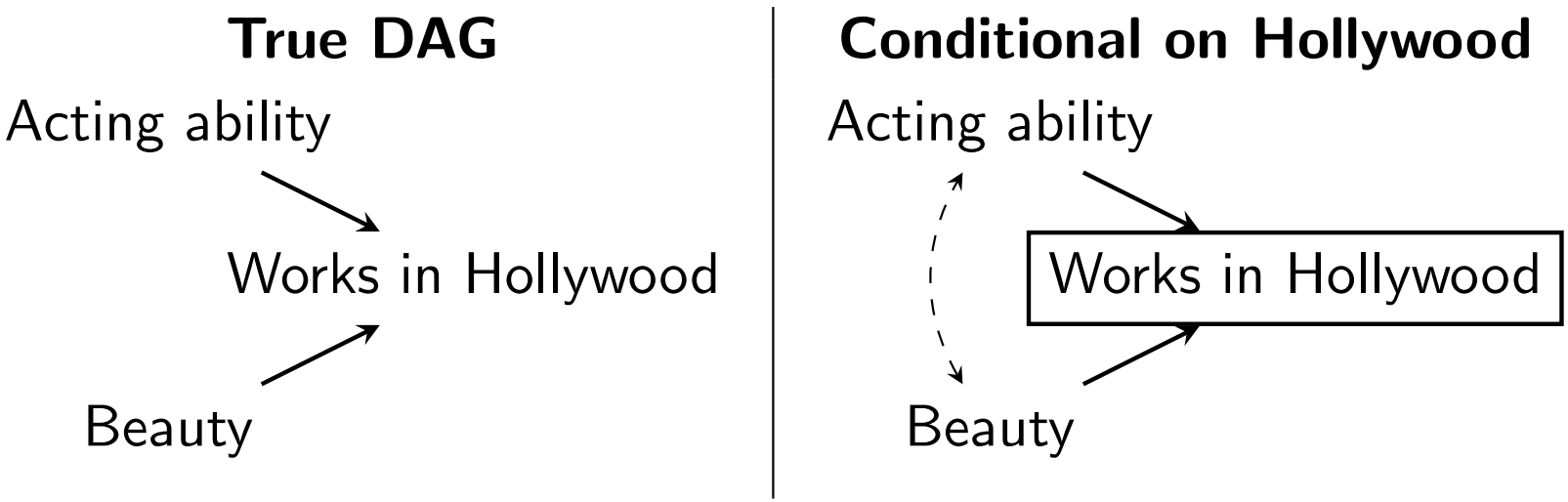
Acting ability $\longrightarrow$ Marriage

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

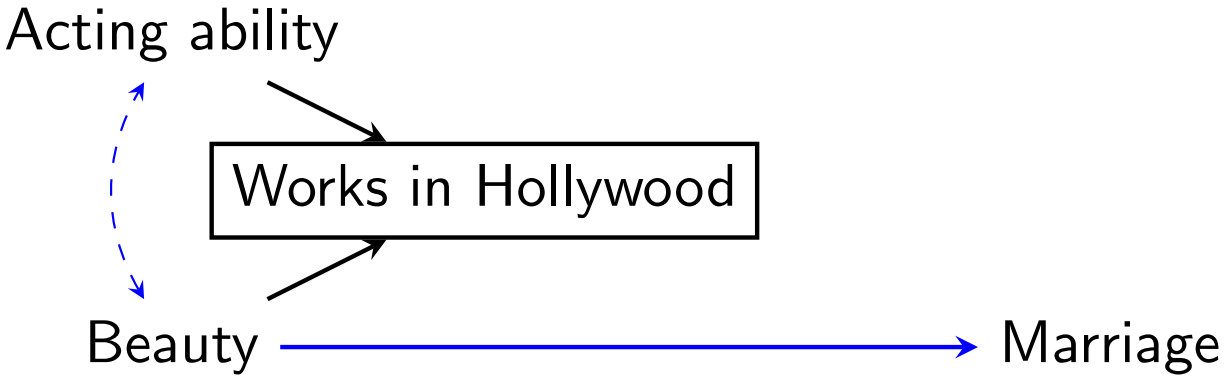
Suppose working in Hollywood is a function of two factors: acting ability and beauty. In the general population, these two are uncorrelated. However, among those who work in Hollywood, those who are bad at acting must be beautiful.



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!

Thoughts on DAGs and Potential Outcomes

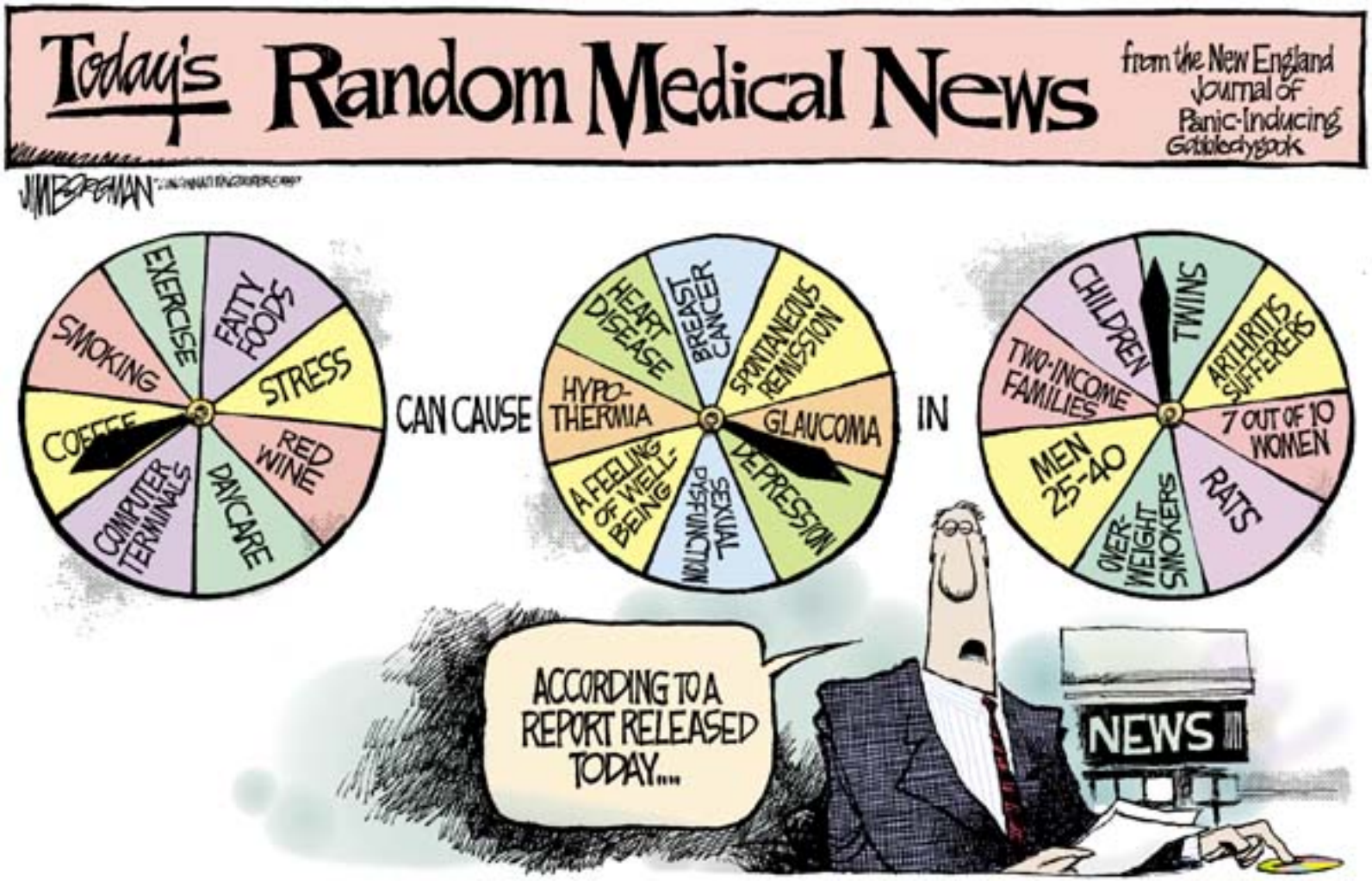
- Two very different languages for talking about and thinking about causal inferences.
- Potential outcomes is very focused on thinking about the treatment assignment mechanism and helpful for heterogeneity of treatment effects.
- Potential outcomes is also less of a "foreign language" for most statisticians, 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!)

1. Core Ideas in Causal Inference

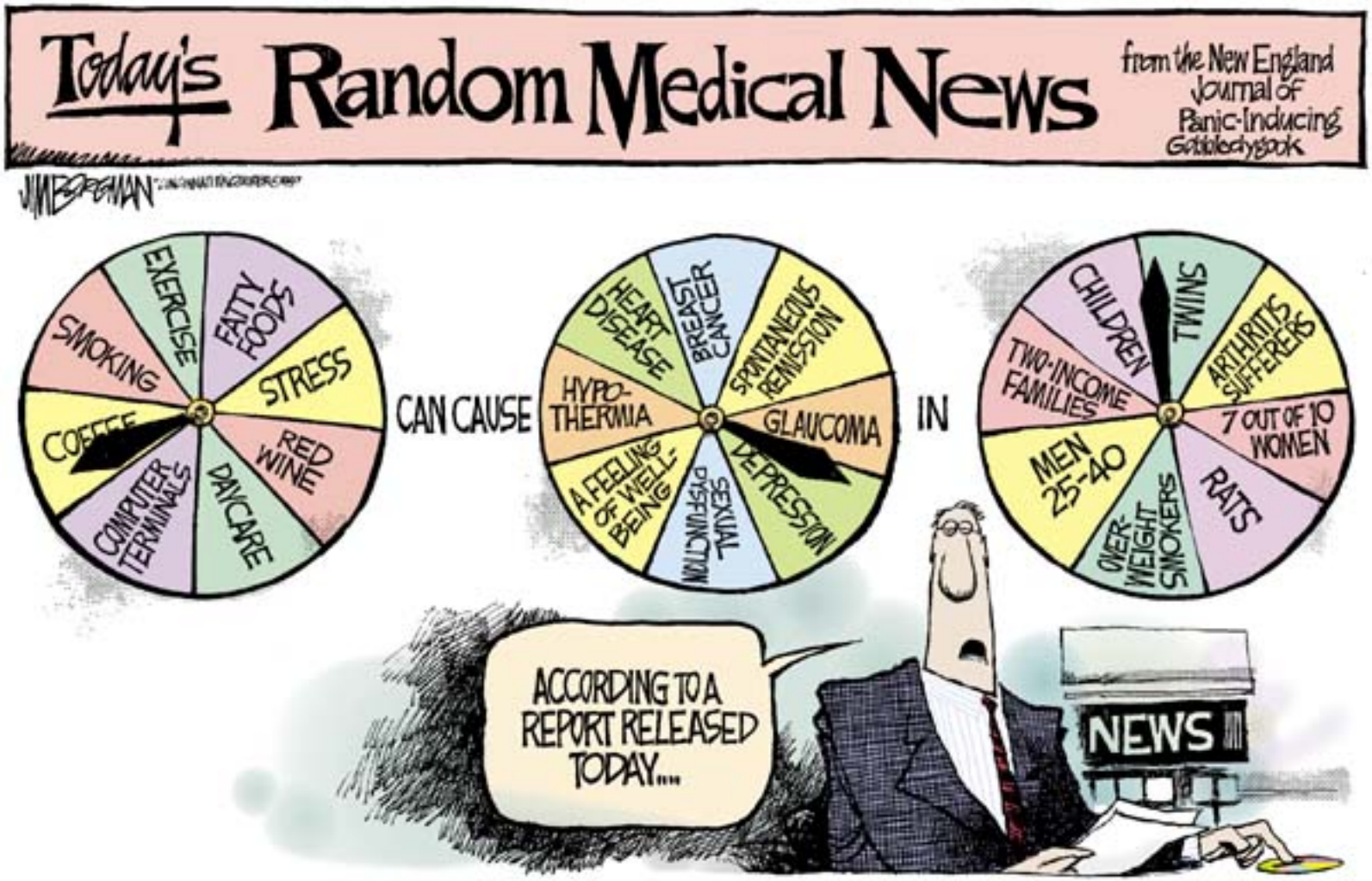
2. Potential Outcomes
   - Framework
   - Estimands
   - Three Big Assumptions
   - What Gets to Be a Cause

3. Nonparametric Structural Equation Models and Causal Directed Acyclic Graphs

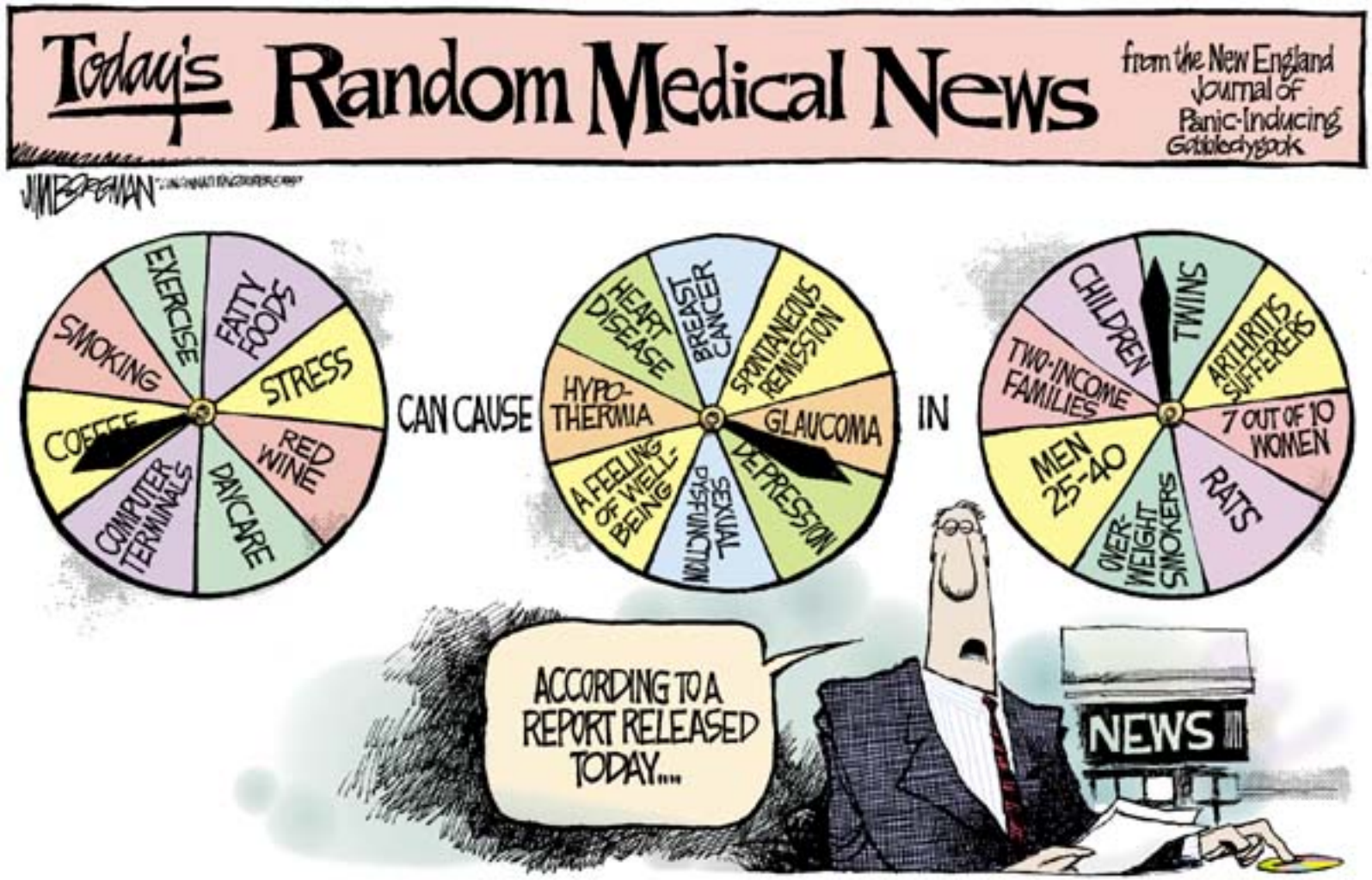
4. The Experimental Ideal



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)



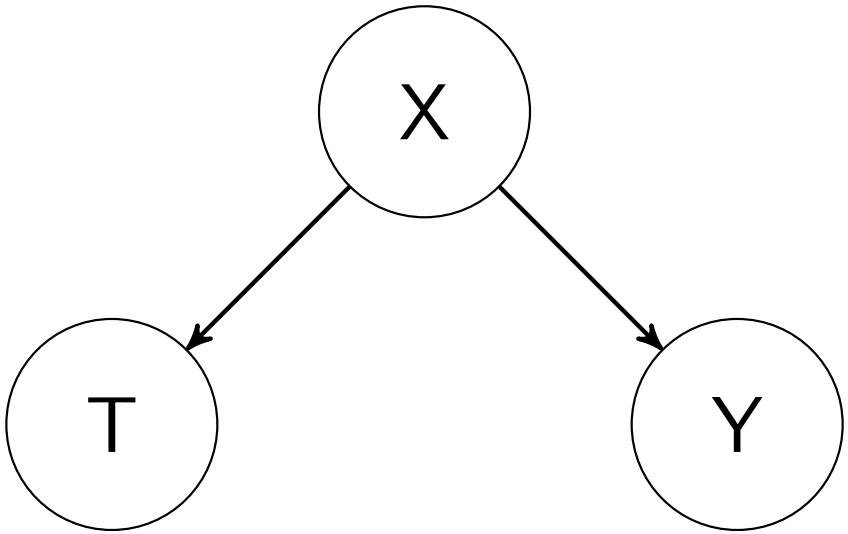
Lancet 2002: no effect of vitamin C on mortality in controlled placebo trial (controlling for nothing)



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?

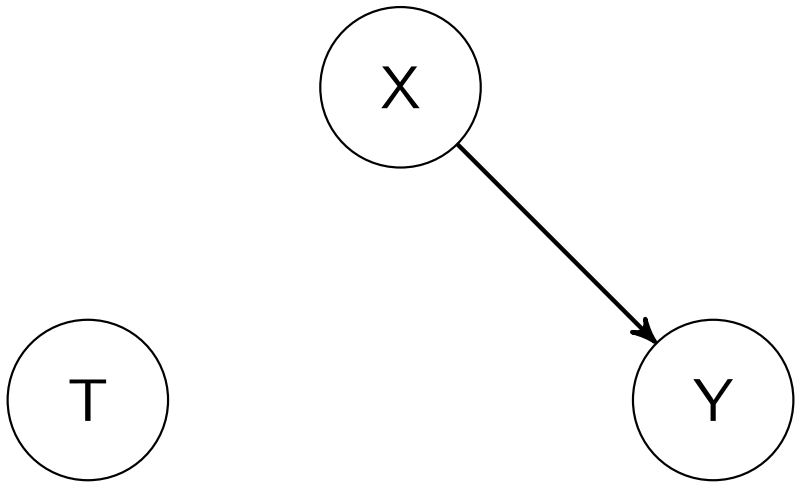
$$
\begin{aligned}
&E[Y_i|T_i=1] - E[Y_i|T_i=0] \\
&\quad = E[Y_i(1)|T_i=1] - E[Y_i(0)|T_i=0] \\
&\quad = E[Y_i(1)|T_i=1] - E[Y_i(0)|T_i=1] + E[Y_i(0)|T_i=1] - E[Y_i(0)|T_i=0] \\
&\quad = \underbrace{E[Y_i(1) - Y_i(0)|T_i=1]}_{\text{Average Treatment Effect on Treated}} + \underbrace{E[Y_i(0)|T_i=1]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

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

$$
\begin{aligned}
E[Y_i(1)|T_i=1] - E[Y_i(0)|T_i=0] &= E[Y_i(1)|T_i=1] - E[Y_i(0)|T_i=1] \\
&= E[Y_i(1)] - E[Y_i(0)]
\end{aligned}
$$

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

This Week in Review

- We talked about what regression is doing and how we go about making an argument.
- This week we began our journey into causal inference.
- The next few weeks we are going to talk about how to use these frameworks to estimate causal effects across a wide variety of scenarios.
- We will make liberal use of both frameworks based on whatever is the most convenient to communicate the point.
- You want to have some familiarity with the core concepts of the frameworks—but don't worry, we will review them more in coming weeks.

Next Time: Causality with Measured Confounding