Causal Data Science for Business Decision Making

Introduction

Paul Hünermund

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

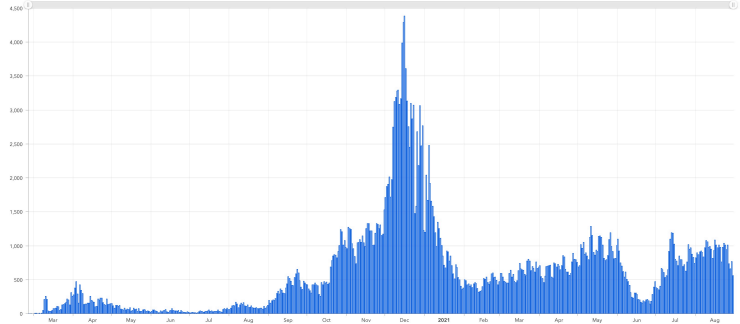
ASSOCIATION AMBA ACCREDITED

CEMS

PIM

COVID-19 & Safety

- To secure a safe learning environment and make sure that we can stay on campus until the end of the semester, let's agree on a few rules
  - Let's keep the windows open to ensure proper ventilation
  - Try to maintain proper distancing
  - If you want to hang out in groups, please do so outside
  - If you're sick, please stay at home (and get tested)



A few words about myself...

- Assistant professor at Strategy & Innovation (SI)
- Joined in September 2021
  - Before 3 years in the Netherlands at Maastricht University
- M.Sc. in economics from Mannheim University and Ph.D. in business economics from KU Leuven
- Research interests:
  - Innovation
  - Science, technology & innovation policy
  - Causal inference & applied econometrics
- Associate editor at the Journal of Causal Inference
- Teaching: causal inference, digital economics, industrial organization, innovation strategy

Why this course?

- Causal data science is becoming a more and more important topic in industry
  - Not least because of the "Book of Why" which was published in 2018
- Standard econometrics and statistics courses teach causal inference in a very process-oriented way
- A conceptual framework for causality is often missing
- On the other hand, data science and machine learning courses teach data analytics as something purely data-driven
- Importance of theory and background knowledge for causal inference
  - Bring together people skilled in data science techniques and business domain experts for effective business decision making
  - This course is meant to be a step in this direction

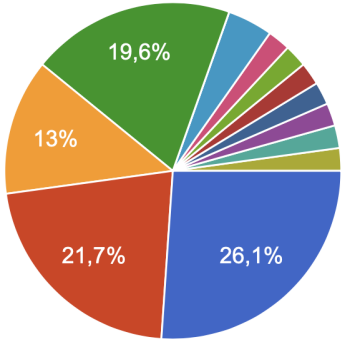
What kind of course is this?

- A little bit of everything
  - Management & Strategy
  - Economics (& Finance)
  - Computer science and AI
  - Statistics
  - Mathematics
  - Philosophy of science
- Focus lies on conceptual understanding of causality and how to infer it from data
- This will involve some statistics and math, but is also **no rocket science**...
- We will see many examples and cases where these concepts are very relevant for decision-making and strategy formulation

You are quite a heterogeneous group...

What is your main background of studies?

46 Antworten

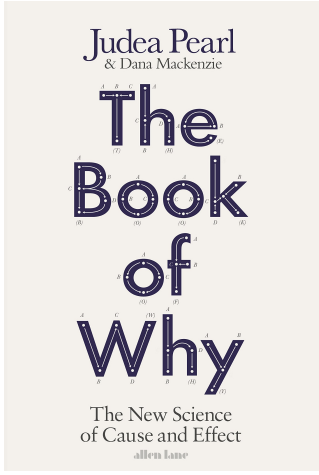


- Economics
- Management
- Marketing
- Finance
- Sociology
- Data Science
- IT/IS and Economics
- American studies

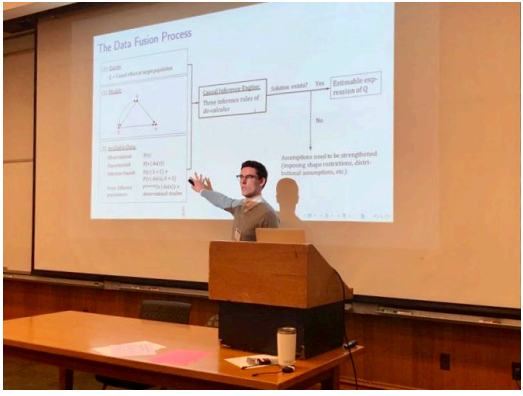
1/2

Course readings

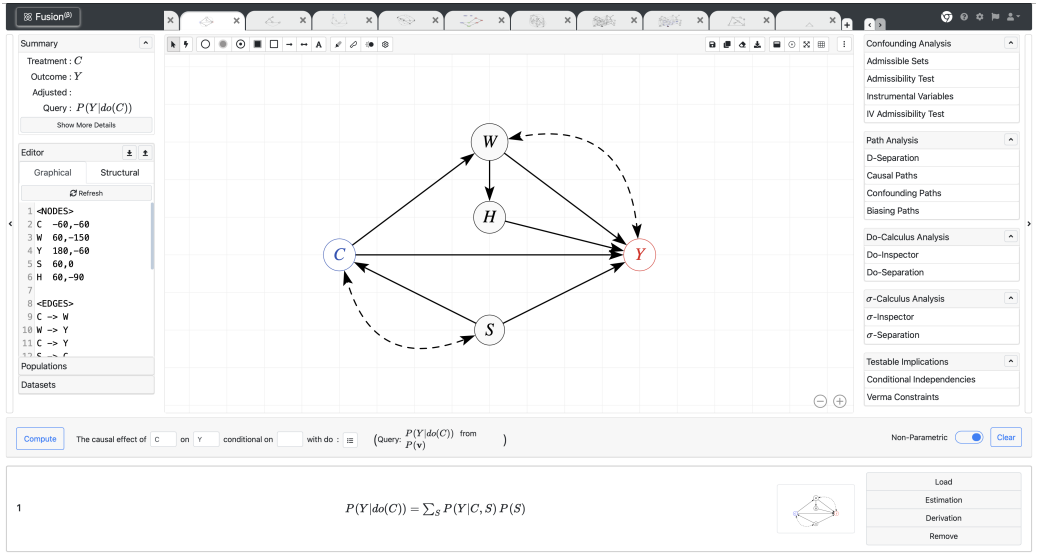
- Mandatory readings
  - The Book of Why (Judea Pearl & Dana Mackenzie)
  - Accompanying papers
  - Blog posts and online material, if it fits
- Additional references
  - Additional sources to dive deeper into a specific topic (and find relevant references there)
  - Classic papers in that area
  - "Causal Inference in Statistics – A Primer" (technical compendium to BoW, full ebook access at CBS library)



WHY-19, AAAI Spring Symposium, Stanford, CA



Causal Fusion



Examination

- Individual written product (**max.** 15 pages)
- Two weeks in December
- Conceptual "essay-style" piece
- Focus on highlighting the relevance of the concepts discussed in course for strategic and business decision making
- Software exercises with Fusion will be relevant for the exam
  - Problem set as preparation (part of self-paced module in week 37)
- More info to come...

Writing Challenge

- Work in teams of 3–5 students
- Write a 800–1500 words blog post about a topic of your choice
  - E.g., pick out a technique or topic we have discussed in class and illustrate its relevance for industry
  - Try to find practically relevant examples
  - Format similar to articles on `www.medium.com` or `www.towardsdatascience.com`
- Feedback opportunity for you!!
- The best three submissions will be published on `www.causalscience.org`
- Deadline: November 28, 2021

Motivating Example: How to Estimate the Gender Pay Gap?

- The New York Times reported in March 2019:
  - *"When Google conducted a study recently to determine whether the company was underpaying women and members of minority groups, it found, to the surprise of just about everyone, that men were paid less money than women* for doing similar work.*"*

    https://www.nytimes.com/2019/03/04/technology/google-gender-pay-gap.html
- The study led Google to increase the pay of its male employees to fight this blatant discrimination of men
- What's going on here? Wasn't Google just recently accused of discriminating against women, not men?
  - *"Department of Labor claims that Google systematically underpays its female employees"*

    https://www.theverge.com/2017/4/8/15229688/department-of-labor-google-gender-pay-gap

Simpson's Paradox

- Suppose we collected data on wages payed to 100 women and 100 men in company X. We observe the following distribution of average monthly salaries for women and men in management and non-management positions (case numbers in parentheses). And our goal is to estimate the magnitude of the gender pay gap in company X. How should we tackle this problem?

	Female	Male
Non-management	\$3163.30 (87)	\$3015.18 (59)
Management	\$5592.44 (13)	\$5319.82 (41)

Simpson's Paradox (II)

- On average, women earn less in this example

$$\left(\frac{87}{100}\cdot \$3163.30 + \frac{13}{100}\cdot \$5592.44\right) - \left(\frac{59}{100}\cdot \$3015.18 + \frac{41}{100}\cdot \$5319.82\right)$$
$$\approx -\$481$$

- But in each subcategory women actually have higher salaries?
  - Non-management: $\$3163.30 - \$3015.18 = \$148.12$
  - Management: $\$5592.44 - \$5319.82 = \$272.62$
- Conditioning on job position gives adjusted gender pay gap

$$\frac{87+59}{200}\cdot \$148.12 + \frac{13+41}{200}\cdot \$272.62 \approx \$181.74$$

- Which estimate gives us a more accurate picture of the gender pay gap?

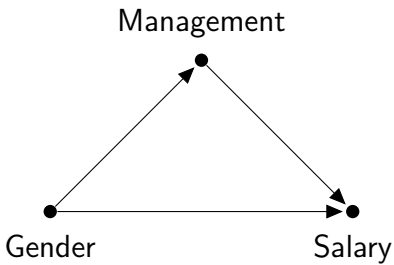
Simpson's Paradox (III)

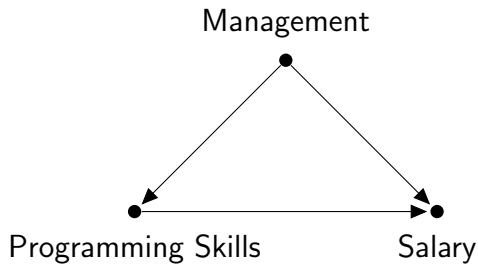
- The phenomenon that a statistical association, which holds in a population, can be reversed in every subpopulation is named after the British statistician Edward Simpson
- Simpson's paradox well-known, for example, in epidemiology and labor economics
- Here, the unadjusted gender pay ($-\$481$) gap gives the right answer
- But what about this example?

	Programming Skills	No Programming Skills
Non-management	\$3163.30 (87)	\$3015.18 (59)
Management	\$5592.44 (13)	\$5319.82 (41)

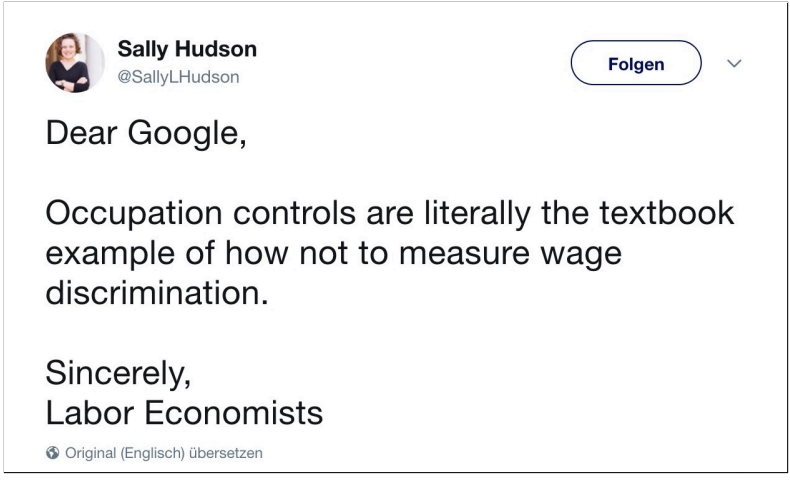
Simpson's Paradox (IV)

- Here we would correctly infer that people with programming skills earn more on average ($181.74). What is the difference between the two examples?





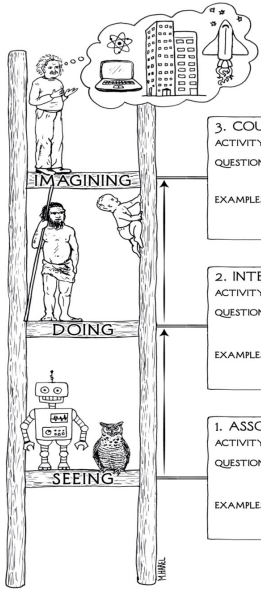
Simpson's Paradox (V)



Simpson's Paradox (VI)

- Statistics alone doesn't help us to answer this question
- Note that the joint distribution of salaries is the same in both cases
- Both problems are thus identical from a statistical point of view
- Instead, we need to make causal assumptions in order to come to a conclusion here
  - Gender affects both a person's salary level and job position
  - Whereas, programming skills increase salaries but persons in high-ranking positions usually have less of it
- After the course you will know how to incorporate this kind of causal knowledge in your analysis in order to solve all sorts of practical problems of causal inference

The Ladder of Causation



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

Simpson's Paradox and Covid-19 Vaccination

Age	Population (%)		Severe Cases (per 100k)		Efficacy
	Not Vax	Fully Vax	Not Vax	Fully Vax	
All ages	18.2%	78.7%	16.4	5.3	67.5%
<50	23.3%	73.0%	3.9	0.3	91.8%
>50	7.9%	90.4%	91.9	13.6	85.2%

- Vaccine effectiveness defined as $1 - V/N$ (e.g., $1 - 5.3/16.4 = 0.675$)
- Vaccine effectiveness is higher in every age group than in the general population. How can that be?
- Vaccination status and risk of severe disease are systematically higher in the older age group $\Rightarrow$ Simpson's Paradox (full story here)
- Lesson: **Get vaccinated!!**

Thank you

Personal Website: p-hunermund.com

Twitter: @PHuenermund

Email: phu.si@cbs.dk

Causal Data Science for Business Decision Making

Graphical Causal Models I

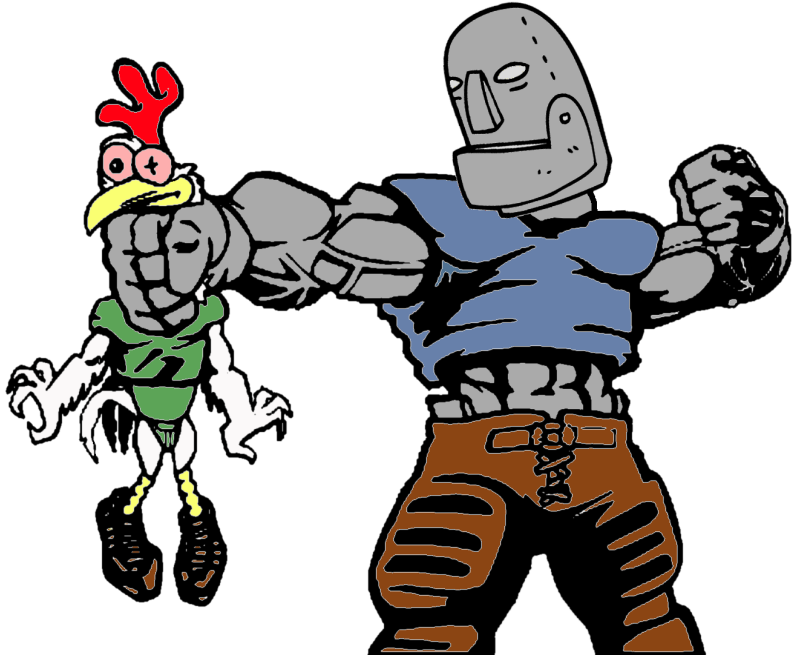
Paul Hünermund

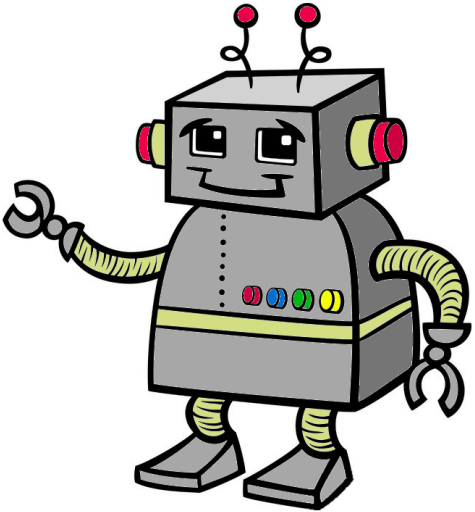
CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Historical Background

- Causal inference is arguably the most important goal in econometrics
  - Inform policy-makers, legislators, and managers about the likely impact of their actions by uncovering quantitative relationships in statistical data
- Since the end of the 1980s, an extensive literature on causal inference was developed in the computer science and artificial intelligence field
  - Builds on the graph-theoretic approach to causality developed by
  - Interest emerged from older AI techniques such as Markov random fields and Bayesian nets
  - Shares several mutual intellectual roots with econometrics
- Why do AI scholars care about causality?

How can we prevent a future robot from trying to make the rooster crow at 3am in order to make the sun come up?





"Beyond Curve Fitting" in Machine Learning and AI

"To Build Truly Intelligent Machines, Teach Them Cause and Effect"

— Judea Pearl

- The notion of causality is a fundamental concept in human thinking
- Current ML / AI techniques remain purely prediction-based
- In other words: machine learning is very good at high-dimensional pattern recognition ("is this a cat or dog?)
- But nothing in the theoretical basis of ML allows to capture the asymmetry inherent to causal relationships
- If we want machines to be able to interact meaningfully with us, they should be equipped with a notion of cause and effect

"If you can't explain it to a ~~six year old~~ ROBOT,
you don't understand it yourself."

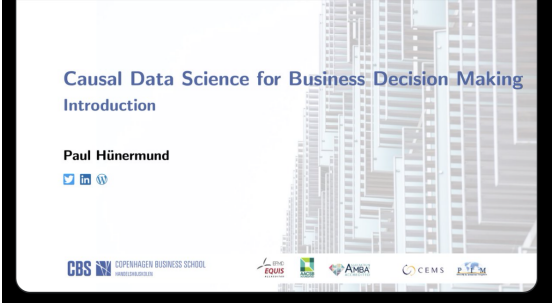
— attributed to *Albert Einstein*

Judea Pearl ✓
@yudapearl

If "Data Science" was truly data science (as defined eg here ucla.in/3iEDRVo), you wouldn't need to add "Causal" ahead of the title. But, given what it is today, this is an effective way of telling students: "This is not another function-fitting class."

Tweet übersetzen

> **Paul Hünermund** @PHuenermund · 14 Std.
> First session of my new elective in the @CBScph master program today. What a great feeling to be back in the class room and teach the stuff you're really excited about!



10:50 vorm. · 3. Sep. 2021 · Twitter Web App

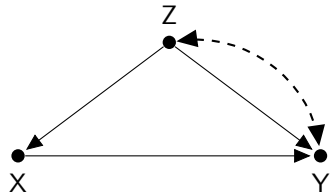
Structural Causal Models

$$z \leftarrow f_Z(u_z)$$
$$x \leftarrow f_X(z, u_x)$$
$$y \leftarrow f_Y(x, z, u_Y)$$

- The f_i's denote the causal mechanisms in the model
  - Are not restricted to be linear as in traditional structural equation models
- The u_i's refer to background factors that are determined outside of the model
- Assignment operator ($\leftarrow$) captures asymmetry of causal relationships
  - $x \leftarrow a \cdot z \neq z \leftarrow x/a$
- Similar to definition of "structure" according to Cowles foundation

Directed Acyclic Graphs

$$z \leftarrow f_Z(u_z)$$
$$x \leftarrow f_X(z, u_x)$$
$$y \leftarrow f_Y(x, z, u_Y)$$



- In a fully specified SCM, every counterfactual quantity is computable
- In most social science contexts it's hard to know the causal mechanisms f_i and distribution of background factors $P(U)$
- Therefore, restrict attention to qualitative causal information of the model, which can be encoded by a graph G
  - Nodes V: variables in the model
  - Directed edges E: causal relationships in the model

Directed Acyclic Graphs

- No functional form or distributional assumptions means that framework remains fully nonparametric
  - Particularly helpful in fields where theory is purely qualitative and no shape restrictions on can be derived
- $Z \dashleftarrow\!\dashrightarrow Y$ is a shortcut notation for unobserved common causes $Z \leftarrow U \rightarrow Y$
- Acyclicity
  - Directed cycles such as $A \rightarrow B \rightarrow C \rightarrow A$ are excluded
  - This means there are no feedback loops
  - Otherwise A could be a cause of itself
  - Gives rise to what economists call a *recursive* model
  - Extensions of the SCM framework to cyclic graphs exist

Recursive vs. Interdependent Systems

D-Separation

- DAGs are such a useful tool because they are able to efficiently encode conditional independence relationships:

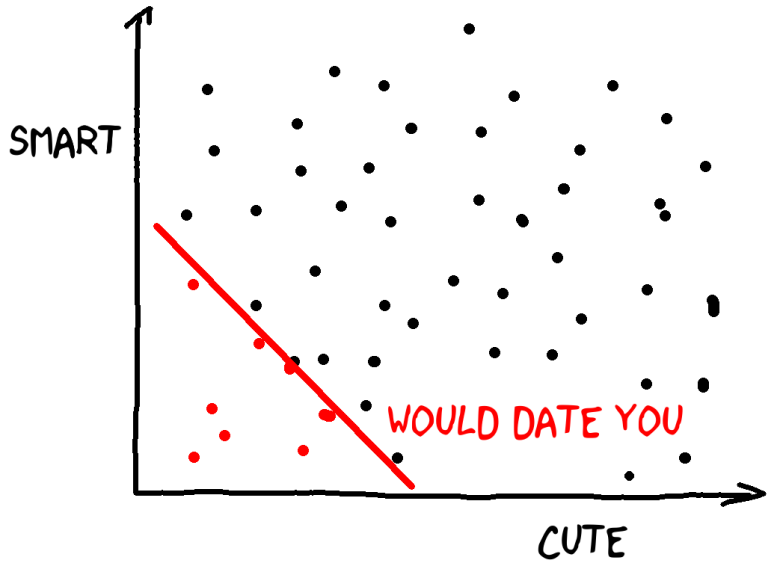
Chain:	$A \to B \to C$	$\Rightarrow$	$A \not\perp\!\!\!\perp C$ and $A \perp\!\!\!\perp C \mid B$
Fork:	$A \leftarrow B \to C$	$\Rightarrow$	$A \not\perp\!\!\!\perp C$ and $A \perp\!\!\!\perp C \mid B$
Collider:	$A \to B \leftarrow C$	$\Rightarrow$	$A \perp\!\!\!\perp C$ and $A \not\perp\!\!\!\perp C \mid B$

- Independence: knowledge that B occurred gives no additional information about the probability of A

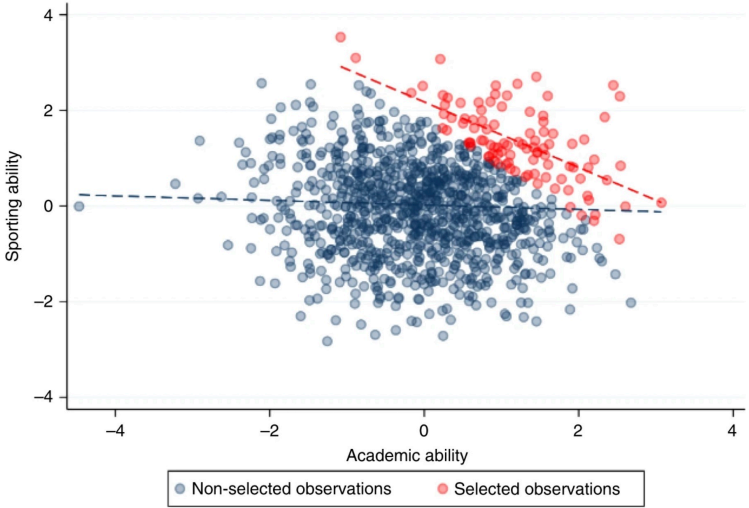
$$A \perp\!\!\!\perp B \Rightarrow P(A|B) = P(A)$$
$$A \perp\!\!\!\perp B|C \Rightarrow P(A|B, C) = P(A|C)$$

- The same holds for longer paths in the graph
  - Conditioning on a variable along a chain or fork blocks (*"d-separates"*) the path
  - Conditioning on a collider opens the path

Collider Bias Example



Collider Bias Example II



Source: "Collider bias undermines our understanding of COVID-19 disease risk and severity" (2020, Nature Comm.)

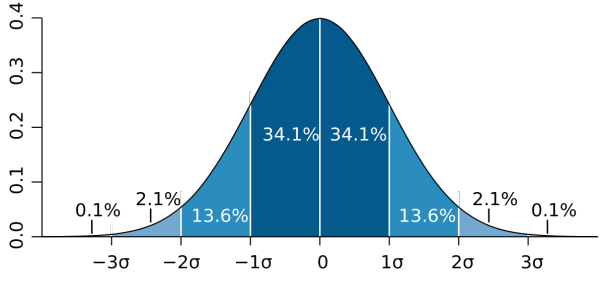
Refresher: Variance and Standard Deviation

- The variance of a variable X is a measure for how "spread out" values of X are

$$Var(X) = E((X - \mu)^2)$$

where $\mu = E(X)$

- The standard deviation σ_X is defined as the square root of the variance



Refresher: Covariance and Correlation

- The covariance between two variables X and Y measures the degree to which they *covary*

$$\sigma_{XY} = E[(X - E(X)(Y - E(Y))]$$

- More specifically, it measures the degree to which they *linearly* covary
- The covariance depends on the scale of the variables, that's why we often normalize it by the standard deviation to arrive at the *correlation coefficient*

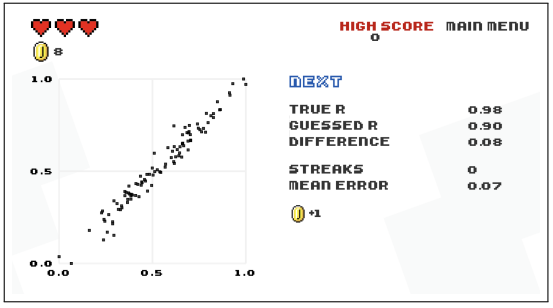
$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

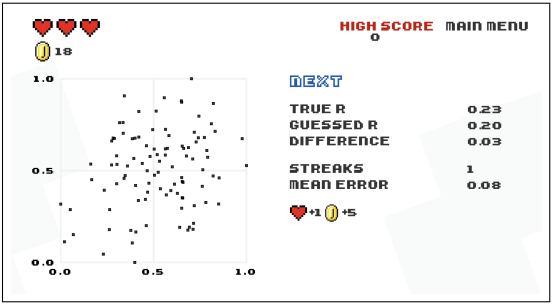
- It holds that

$$X \perp\!\!\!\perp Y \Rightarrow \sigma_{XY} = 0,$$

but not the other way round

Guessthecorrelation.com



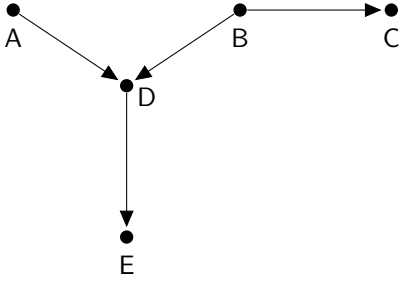


Collider Bias – R Example

```r
# Create two independent uniformly distributed variables
cute <- runif(1000)
smart <- runif(1000)

# Plot
plot(scute, smart, pch=20)

# Construct date equal to one if smart + cute < 0.5, and zero
    otherwise
date <- 1*(smart + cute > 1)

# By design, there is no correlation between the two variables
cor(smart, cute)

# But if we condition on date==1, we find a negative correlation
cor(smart[date==1], cute[date==1])
```

Testable Implications

- D-separation provides testable implications of a model



Testable implications:

$A \perp\!\!\!\perp B$ $\quad$ $A \perp\!\!\!\perp C$

$A \perp\!\!\!\perp E|D$ $\quad$ $B \perp\!\!\!\perp E|D$

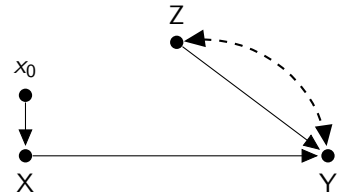
$C \perp\!\!\!\perp D|B$ $\quad$ $C \perp\!\!\!\perp E|D$

$C \perp\!\!\!\perp E|B$

- If one of these conditional independence relations do not hold in the data, the model can be rejected
- *"Causal discovery"*: try to learn compatible model from conditional independence relations found in the data
  - We will talk about that in week 38

Interventions in Structural Causal Models

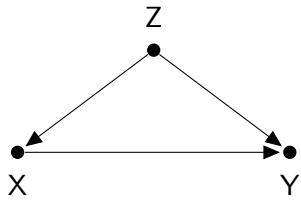
$$z \leftarrow f_Z(u_z)$$
$$x \leftarrow x_0$$
$$y \leftarrow f_Y(x, z, u_Y)$$



- Causal inference $\triangleq$ predict the effects of interventions (policy initiatives, social programs, management initiatives, etc.)
- Interventions in SCMs amount to *"wiping out"* of causal mechanisms, an idea that originally came from econometrics (Strotz and Wold 1960)
  - Delete naturally occurring causal mechanism $f_X(\cdot)$ from model and set X to constant value x_0
  - This operation is denoted by *do-operator*: $do(X = x_0)$
- Query of interest: post-interventional distribution $P(Y = y|do(X = x))$

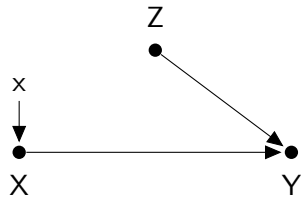
Pre- versus Post-intervention Distribution

Pre-intervention



$$Z = f_z(u_z)$$
$$X = f_x(Z, u_x)$$
$$Y = f_y(X, Z, u_y)$$

Post-intervention



$$Z = f_z(u_z)$$
$$X = x$$
$$Y = f_y(X, Z, u_y)$$

- The intervention changes the data-generating process; thus, $P(Y|X)$ (pre-intervention) is generally not equal to $P(Y|do(X))$ (post-intervention)

Thank you

Personal Website: p-hunermund.com

Twitter: @PHuenermund

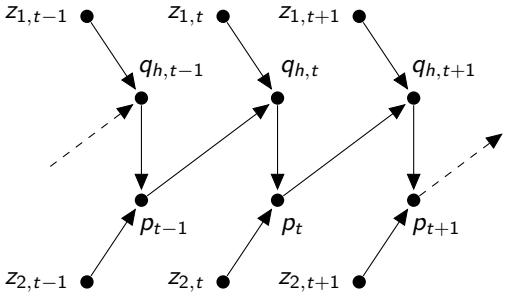
Email: phu.si@cbs.dk

Recursive versus Interdependent Systems

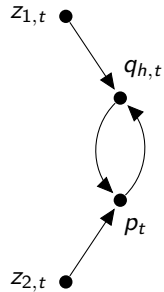
- DAGs represent recursive systems, but many standard models in economics are interdependent (Marshallian cross, game theory, etc.)
- This connects to an old debate within econometrics about the causal interpretation of recursive versus interdependent models that emerged in the aftermath of Haavelmo's celebrated 1943 paper
- One central argument (Strotz and Wold, 1960):
  - Individual equations in an interdependent model do not have a causal interpretation *in the sense of a stimulus-response relationship*
  - Interdependent systems with equilibrium conditions are regarded as a *shortcut* description of the underlying dynamic behavioral processes

Recursive versus Interdependent Systems

- In this context, Strotz and Wold (1960) discuss the example of the cobweb model:



$$p_{t-1} = p_t$$
$$\Rightarrow$$



$$q_{h,t} \leftarrow \gamma + \delta p_{t-1} + \nu z_{1,t} + u_{1,t},$$
$$p_t \leftarrow \alpha - \beta q_{h,t} + \varepsilon z_{2,t} + u_{2,t}.$$

$$q_{h,t} \leftarrow \gamma + \delta p_t + \nu z_{1,t} + u_{1,t}$$
$$p_t \leftarrow \alpha - \beta q_{h,t} + \varepsilon z_{2,t} + u_{2,t}$$

Recursive versus Interdependent Systems

- However, equilibrium assumption $p_{t-1} = p_t$ carries no behavioral interpretation
- Individual equations in interdependent system do not represent autonomous causal relationships in the stimulus-response sense
  - Endogenous variables are determined jointly by all equations in the system
  - Not possible, e.g., to directly manipulate p_t to bring about a desired change in $q_{h,t}$
- Equilibrium models can of course still be useful for learning about causal parameters
- But, if individual mechanisms are supposed to be interpreted as stimulus-response relationships, cyclic patterns need to be excluded

Back

Causal Data Science for Business Decision Making

Graphical Causal Models II

Paul Hünermund

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

CEMS

PIM

Causal Effects

Definition: Causal Effect (Pearl, 2009, p. 70)

Given two disjoint sets of variables, X and Y, the causal effect of X on Y, denoted either as $P(y|\hat{x})$ or $P(y|do(x))$ is a function from X to the space of probability distributions on Y. For each realization x of X, $P(y|\hat{x})$ gives the probability of $Y = y$ induced by deleting from the [structural causal] model [...] all equations corresponding to variables in X and substituting $X = x$ in the remaining equations.

- Sometimes, causal effects of interventions are defined as

$$E(Y = y|do(X = x'')) - E(Y = y|do(X = x'))$$

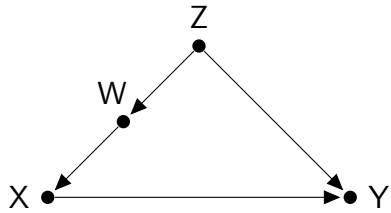
which can always be computed from the general function $P(y|do(x))$

The Identification Problem

- Carrying out the intervention ourselves, in a randomized control trial, is not always feasible (too expensive, impractical, or unethical)
- How can we then identify the effect of interventions purely from observational data?
  - We want to know $P(y|do(x))$ but all we have is data $P(x, y, z)$
  - And we know that $P(y|do(x)) \neq P(y|x)$ ("correlation is not causation")
  - No fancy machine learning algorithm will ever solve this problem
- We need to find a way to transform $P(y|do(x))$ into an expression that only contains, observed, "do-free" quantities
- What if we only have data that is measured with selection bias or that stems from a different population (topic of later lectures)?

Confounding Bias

- Problem of confounding: Paths between treatment X and outcome Y that are not emitted by X and which create an association between X and Y that is not causal



- In this example there is one confounding path: $X \leftarrow W \leftarrow Z \rightarrow Y$
  - Because confounding paths always point into X, they are also called *backdoor paths* (i.e., they "enter through the backdoor")
- The other path between X and Y ($X \rightarrow Y$) is emitted by X and therefore causal

Blocking Backdoor Paths

- So confounding paths create spurious correlations between treatment and outcome. But remember the d-separation criterion from the previous lecture

Definition: *d*-separation (Pearl et al., 2016, p. 46)

A path p is blocked by a set of nodes Z if and only if

1. p contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that the middle node B is in Z (i.e., B is conditioned on), or
2. p contains a collider $A \rightarrow B \leftarrow C$ such that the collision node B is not in Z, and no descendant of B is in Z

- We can block biasing paths by conditioning on intermediate variables on these paths that are not colliders or descendants of colliders

Backdoor Adjustment

Definition: The Backdoor Criterion (Pearl et al., 2016, p. 61)

Given an ordered pair of of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.

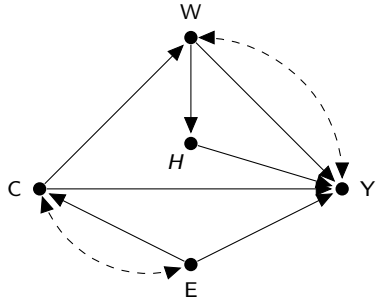
- If a set of variables Z satisfies the backdoor criterion fo X and Y, then the causal effect is given by

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

- I.e., condition on the values of Z and average over their joint distribution (= adjusting for Z)

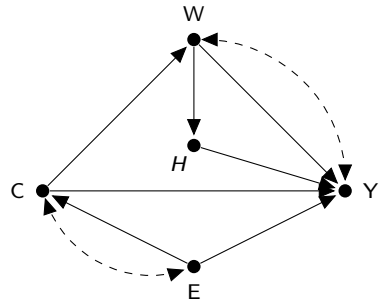
Example 1: College Wage Premium

- Take the stylized example of the college wage premium
  - C: college degree
  - Y: earnings
  - W: occupation
  - H: work-related health
  - E: other socio-economic factors



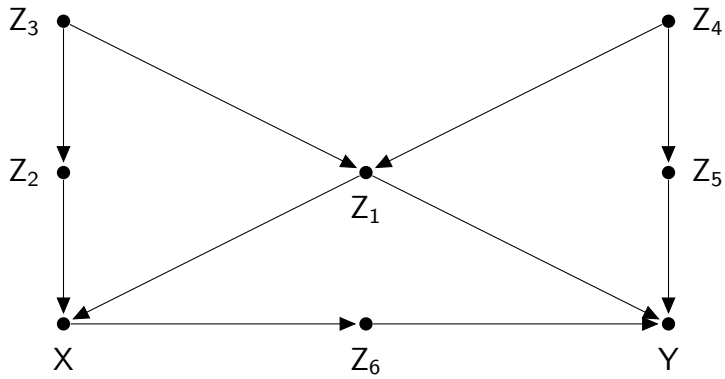
Example 1: College Wage Premium

- Take the stylized example of the college wage premium
  - C: college degree
  - Y: earnings
  - W: occupation
  - H: work-related health
  - E: other socio-economic factors

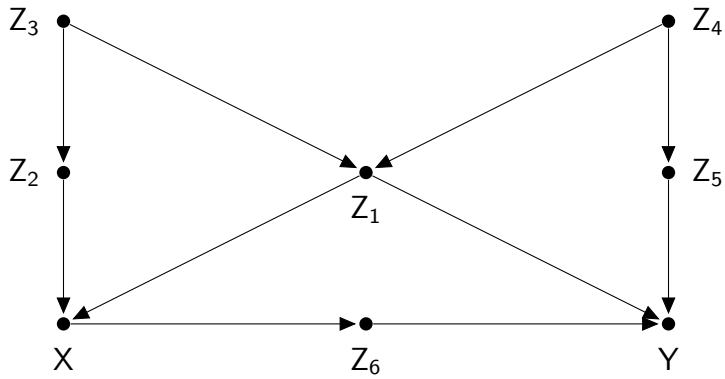


- The graph contains two backdoor paths
  1. $C \leftarrow E \rightarrow Y$
  2. $C \dashleftarrow\dashrightarrow E \rightarrow Y$
- We can close both of these backdoor paths by adjusting for E

Example 2

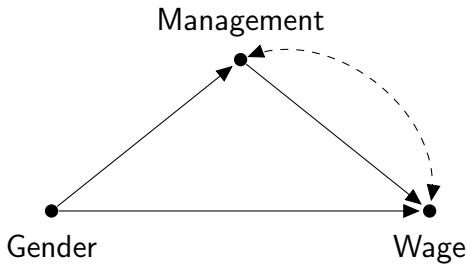


Example 2

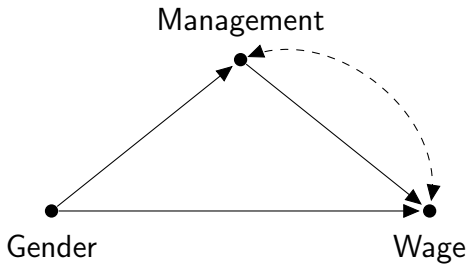


- Minimum sufficient adjustment sets: $\{Z_1, Z_2\}$, $\{Z_1, Z_3\}$, $\{Z_1, Z_4\}$, $\{Z_1, Z_5\}$

Example 3: Gender Wage Gap

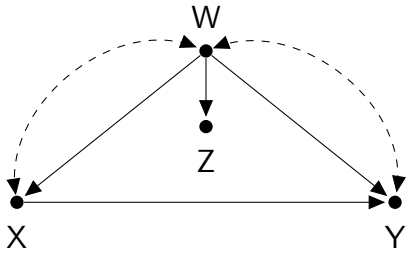


Example 3: Gender Wage Gap

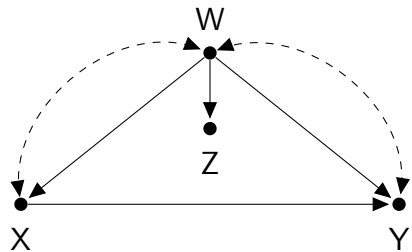


- Minimum sufficient adjustment sets: $\emptyset$ (empty set)
  - Causal effect is identified without adjusting for any covariate
- In fact, conditioning on *Management* would lead to collider bias in this case

Example 4

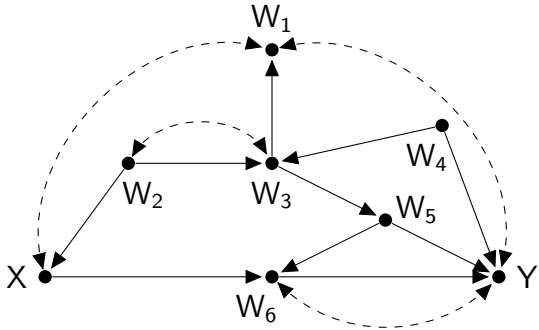


Example 4

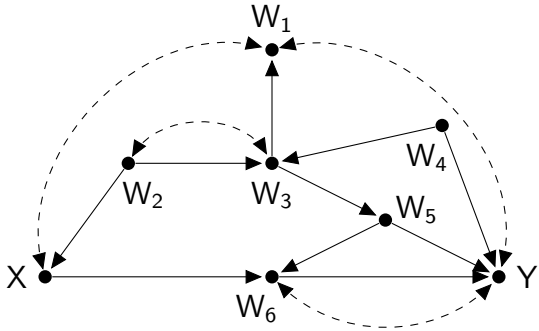


- There is no admissible adjustment set in this graph
  - W is a confounder on the path $X \leftarrow W \rightarrow Y$
  - But conditioning on W or Z leads to collider bias
- The causal effect of X on Y is not identifiable in this graph

Example 5



Example 5



- Backdoor-admissible adjustment sets:

$$Z = \{\{W_2\}, \{W_2, W_3\}, \{W_2, W_4\}, \{W_3, W_4\}, \{W_2, W_3, W_4\}, \{W_2, W_5\}, \{W_2, W_3, W_5\}, \{W_4, W_5\}, \{W_2, W_4, W_5\}, \{W_3, W_4, W_5\}, \{W_2, W_3, W_4, W_5\}\}$$

Estimation

- We have already seen that once we have found a backdoor-admissible adjustment set Z, the causal effect is identified via the adjustment formula

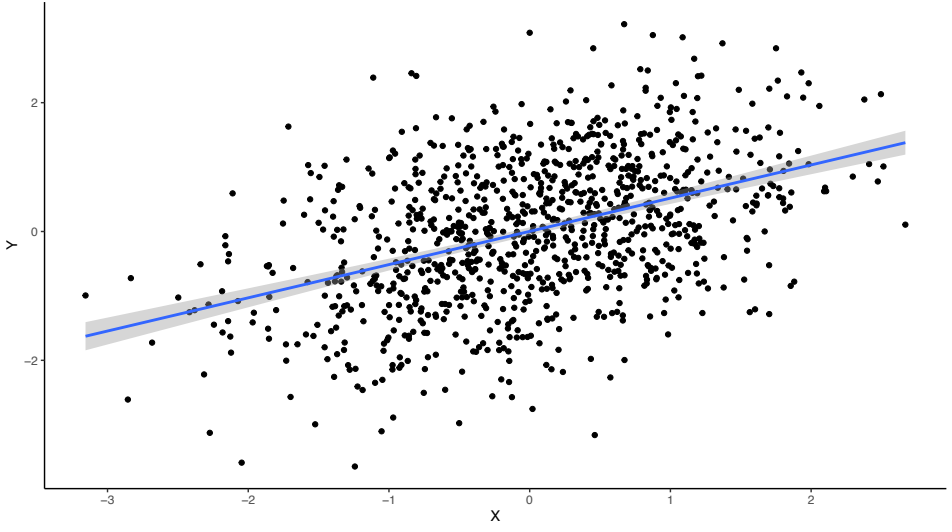
$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

- This non-parametric regression formula can be hard to estimate directly
- In practice, we can relay on simpler estimation algorithms such as
  - Nearest-neighbor matching
  - Ordinary least squares (if we are willing to additionally assume $E[Y|X, Z]$ to be linear)

Nearest-neighbor matching

- The basic idea is very simple
  - Assume X takes two values: one (treated) and zero (untreated)
  - For every treated unit find an untreated neighbor in the data that has a similar value of Z
  - Estimate the correlation between X and Y in this matched sample
  - Due to the matching, the distribution of Z is balanced between treated and untreated units and thus cannot produce confounding anymore
- In practice, a few more complications arise
  - Do you always find a suitable neighbor for every treated unit (problem of common support)?
  - Should you match one or more neighbors (bias-variance tradeoff)
  - How to deal with continuous variables?

Refresher: Linear Regression



Refresher: Linear Regression (II)

- We assume that the conditional expectation of Y can be approximated by a linear function

$$y_i = a + bx_i$$

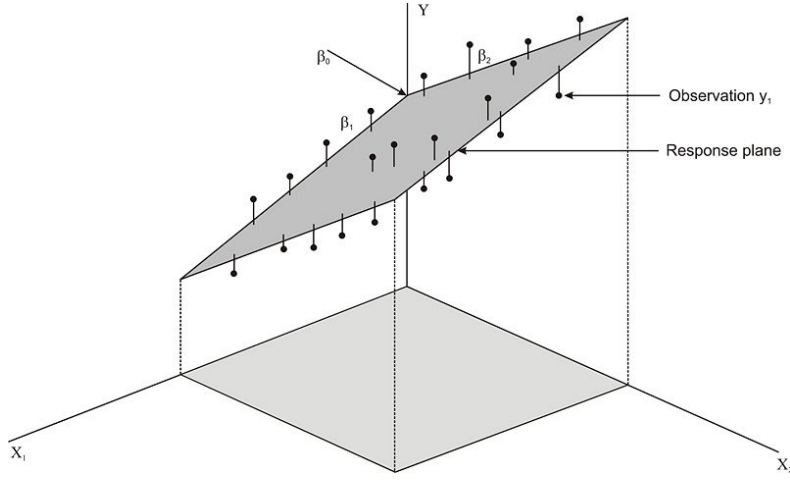
- To find the line of best fit we minimize the loss function

$$\sum_i (y_i - y_i')^2 = \sum_i (y_i - a - bx_i)^2$$

- The name *rdinary least squares* (OLS) derives from the square loss function
- Solution of the minimization problem for the bivariate case
  - $\hat{b} = R_{YX} = \frac{\sigma_{XY}}{\sigma_X^2}$ and $\hat{a} = \bar{y} - \hat{b}\bar{x}$
- We can similarly find solution if there are more than one explanatory variable

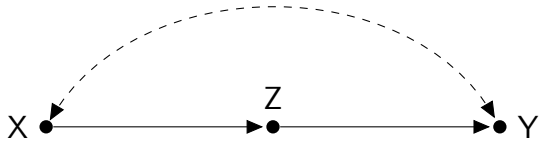
$$y_i = a + b_1 x_{1i} + b_2 x_{2i} \ldots + b_k x_{ki}$$

Multiple Regression



Source: https://www.ck12.org/section/multiple-regression-of-regression-and-correlation/

Front-door Criterion



- What happens if cannot measure all variables that we need for backdoor adjustment?
- We will see several solutions for dealing with unobservables throughout the course, but one particularly elegant one is the so-called front-door criterion (FC)
- For the FC to work, Z has to transmit the entire effect of X on Y
- Bellemare et al. (2020) use the FC to estimate whether sharing a ride on Uber and Lyft (X) leads to a lower propensity to tip (Y)
- The mediator Z in this case is whether a ride is actually shared after it has been requested by the app user

Front-door Criterion (II)

Definition: The Frontdoor Criterion (Pearl et al., 2016, p. 69)

A set of variables Z is said to satisfy the frontdoor criterion relative to an ordered pair of variables (X, Y) if

1. Z intercepts all directed paths from X to Y
2. There is no unblocked path from X to Z
3. All backdoor paths from Z to Y are blocked by X

Theorem: Frontdoor Adjustment (Pearl et al., 2016, p. 69)

If Z satisfies the frontdoor criterion relative to (X, Y) and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by the formula

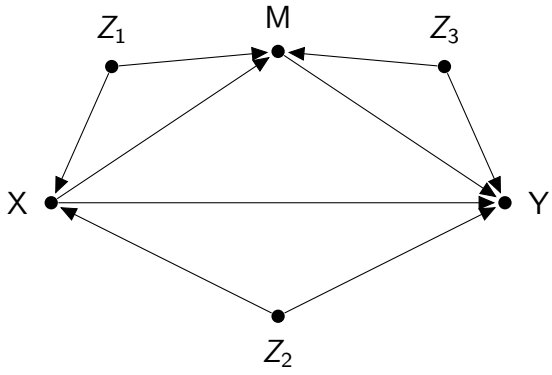
$$P(Y = y|do(X = x)) = \sum_{z} \sum_{x'} P(Y = y|Z = z, X = x')P(X = x')P(Z = z|X = x)$$

Where do DAGs come from?

Causal Modeling

- Causal diagrams are a model of how we think the world works
- We arrive at such a model by using our knowledge about the particular context under study
  - E.g., by consulting the relevant scientific literature for a topic
  - Or by interviewing domain experts
- The ladder of causation tells us that there is no way around these theoretical assumptions for causal inference: *"no causes in, no causes out"* (Cartwright, 1989)
- There some data-driven approaches, which are known under the rubric of *causal discovery*
  - They rely on the d-separation criterion we have already encountered and try to infer a compatible DAG from the conditional independence relationships found in the data
  - You can show, however, that you will never be able to perfectly determine a DAG from data alone. Some ex-ante causal assumption will always be needed

Causal Discovery and D-Separation



This graph implies the following conditional independence relationships in the data:

$$M \perp\!\!\!\perp Z_2 | X, Z_1$$
$$X \perp\!\!\!\perp Z_3$$
$$Y \perp\!\!\!\perp Z_1 | M, X, Z_2, Z_3$$
$$Z_1 \perp\!\!\!\perp Z_2$$
$$Z_1 \perp\!\!\!\perp Z_3$$
$$Z_2 \perp\!\!\!\perp Z_3$$

- The conditional independence relationships on the right can actually be used to learn the graph on the left from data

Thank you

Personal Website: p-hunermund.com

Twitter: @PHuenermund

Email: phu.si@cbs.dk

References I

Marc F. Bellemare, Jeffrey R. Bloom, and Noah Wexler. The paper of how: Estimating treatment effects using the front-door criterion. Working Paper, 2020.

Nancy Cartwright. *Nature's Capacities and Their Measurement*. Clarendon Press, Oxford, 1989.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, United States, NY, 2nd edition, 2009.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons Ltd, West Sussex, United Kingdom, 2016.

Causal Data Science for Business Decision Making

A/B Testing and Experimentation

Paul Hünermund

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

CEMS

PIM

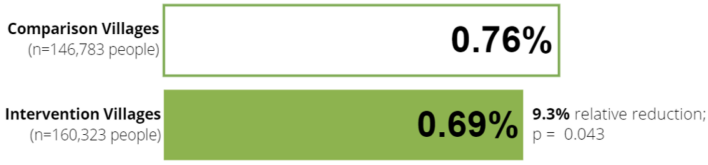

A/B Testing

Recent Example: Does Mask Wearing Reduce Covid Spread?

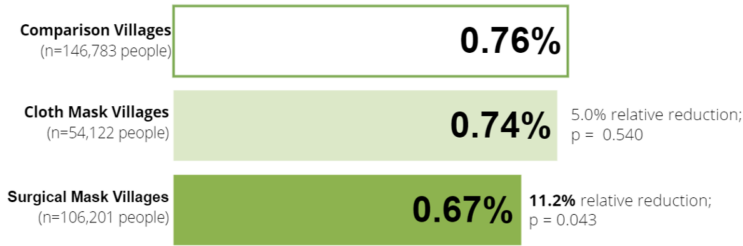
- Large cluster-randomized trial of community-level mask promotion in rural Bangladesh from November 2020 to April 2021 ($N = 600$ villages, $N = 342,126$ adults)
- Cross-randomized mask promotion strategies at the village and household level, including cloth vs. surgical masks
- Intervention: all intervention arms received
  - Free masks
  - Information on the importance of masking
  - Role modeling by community leaders
  - In-person reminders for 8 weeks
- The control group did not receive any interventions
- Neither participants nor field staff were blinded to intervention assignment
- More info: `https://www.poverty-action.org/publication/impact-community-masking-covid-19-cluster-randomized-trial-bangladesh`

Results

(a) Intervention Effect on Symptomatic Seroprevalence



(b) Intervention Effect on Symptomatic Seroprevalence by Mask Type



How Booking.com Experiments with Site Improvements

SCENARIO #1

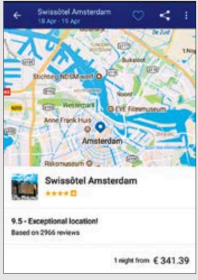
Hypothesis

Highlighting a neighborhood's walkability helps users make better decisions about property location.

A

The Control

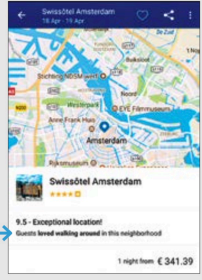
Shows the site's current practice



B

The Treatment

Adds walkability information



The Result

The treatment had no significant impact on the key metric. The current practice is kept in place.

SCENARIO #2

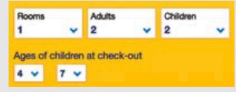
Hypothesis

Displaying the checkout date when users select the age of children in their party improves their experience.

A

The Control

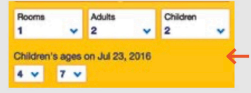
Shows the site's current practice



B

The Treatment

Adds the checkout date above children's ages



The Result

The treatment had a significant positive impact on the key metric, and the change is implemented.

Experimentation doesn't only happen online

A Look Inside Walmart's Next-Gen Test Stores

By Jeff Muench | May 12, 2017

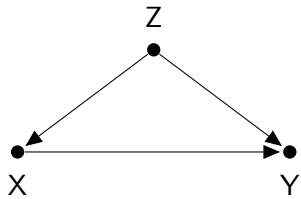
The world is navigating a cultural revolution into the digital age.

Meeting customers' needs is critical as they adopt more digitally-driven lifestyles, expectations increase and increasingly shopping options do not require a trip inside a store.

With this in mind, Walmart is testing new approaches in two recently opened supercenters in Tomball, Texas, and Lake Nona, Florida. These stores were fully reimagined from a new layout to building and environmental enhancements to added technology that all improve the shopping experience.

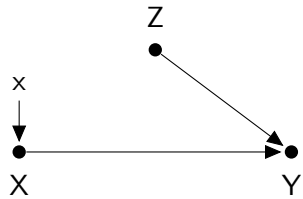
Reminder: Interventions in SCMs

Pre-intervention



$$Z = f_z(u_z)$$
$$X = f_x(Z, u_x)$$
$$Y = f_y(X, Z, u_y)$$

Post-intervention



$$Z = f_z(u_z)$$
$$X = x$$
$$Y = f_y(X, Z, u_y)$$

- If we have the possibility to physically intervene on X, the post-interventional distribution $P(y|do(x))$ becomes **in principle** directly observable

Interventions and the Causal Hierarchy

- An experiment cuts all incoming arrows into the treatment variable X from the model
- This works even if the confounding influence factors Z are unobservable
  - You do not need a theory about confounding influence factors anymore
  - This is much more convenient and convincing than in the backdoor adjustment case where you need to rule out all unobservable confounders
- For this reason, many call experiments or randomized control trials (RCTs) the "gold standard" of causal inference
  - Instead of painstakingly go through all possible confounding variables, just run an experiment and you're good
  - You can see this attitude in Thomke's "let the data speak" approach, which he compares to the alleged wisdom of design experts at Bing
- How does that fit to the causal hierarchy, which stated that we always need ex-ante theoretical assumptions to identify causal effects?


FINE
GOLD
NET WT
2000

Are RCTs Really the "Gold Standard"?

- Experiments are often not possible in practice (too costly, unethical, etc.)
- Long-term versus proxy metrics
- External validity requires you to make a theoretical claim about the transportability of results (see lecture in week 48)
- Observational methods allow you to analyze actual field data versus a controlled lab experiments
- You can sometimes learn about causal effects without a theory as to why something works
  - Citrus fruits $\rightarrow$ (Vitamin C) $\rightarrow$ Scurvy
  - But you need a good theory in order to formulate hypotheses that you can later test
  - "Theory-based view of the firm" (Felin and Zenger, 2009, 2017)
- Experiments are a great tool for causal learning, but there is no such thing as a "gold standard"
  - Which method is most suitable will always depend on the specific context

Some Practical Considerations for Experimental Design

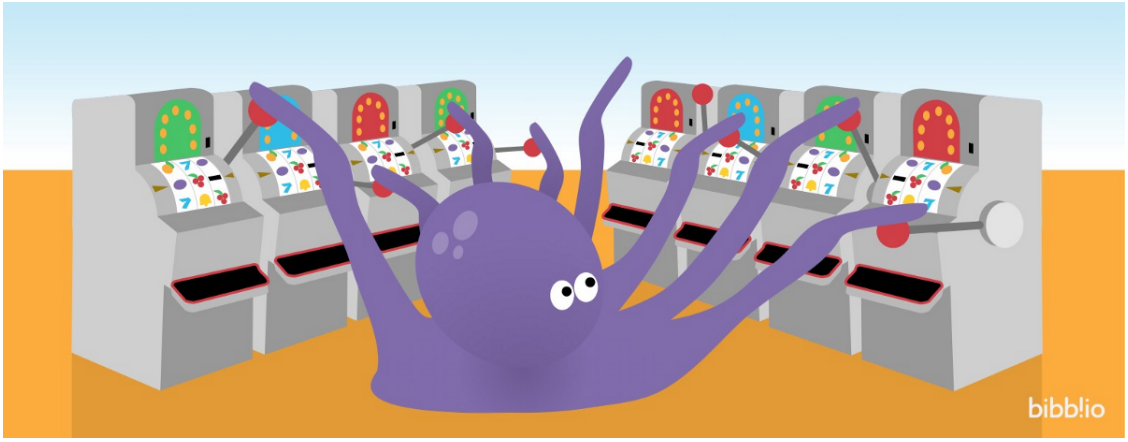
- Beware, this is an incomplete list! There are entire books written on experimental design
- Most A/B tests in practice do not show such spectacular results as suggested in Thomke's HBR piece
- Sometimes, just the awareness of taking part in an experiment can change people's behavior, even if no treatment is administered (Hawthorne effect)
- We can test whether treatment and control group is balanced with respect to observable covariates, but be aware of type-1 error
- You need to make sure that there is sufficient sample size in each treatment arm (power calculation)
- Effects can occur in clusters
  - In this case, treated an untreated units are not independent of each other $\Rightarrow$ SUTVA (stable unit treatment value assumption) violation
- Averages can hide treatment effect heterogeneity
- Consider to pre-register your experiments

P-Hacking

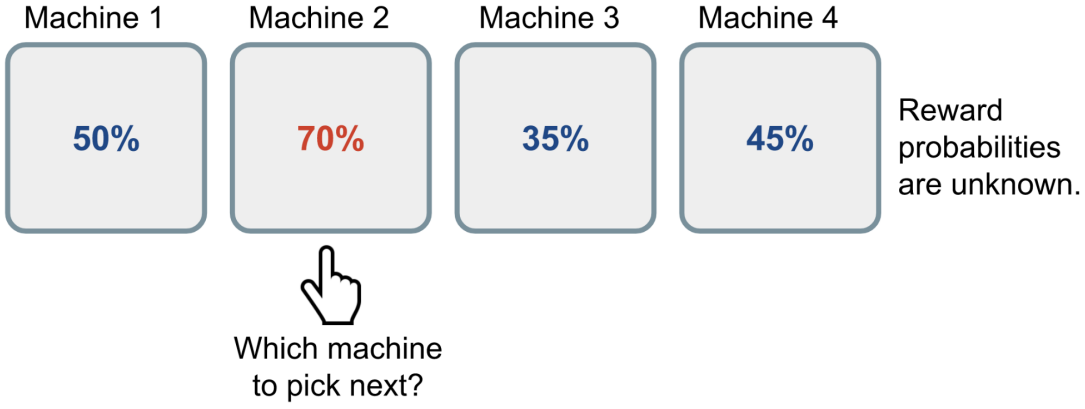
P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥0.1	

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Can we do better than simple A/B testing?

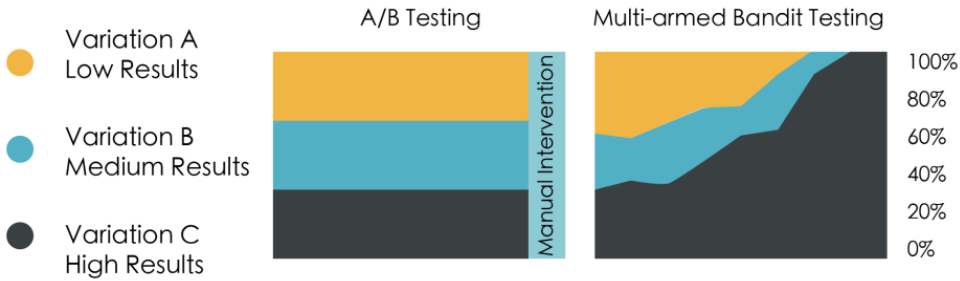


Multi-armed Bandits



- Trade-off between exploration versus exploitation
  - Do you play the currently best strategy or do you try to find out a possibly better strategy?

A/B vs. MAB


Variation A
Low Results
Variation B
Medium Results
Variation C
High Results
A/B Testing
Manual Intervention
Multi-armed Bandit Testing
100%
80%
60%
40%
20%
0%

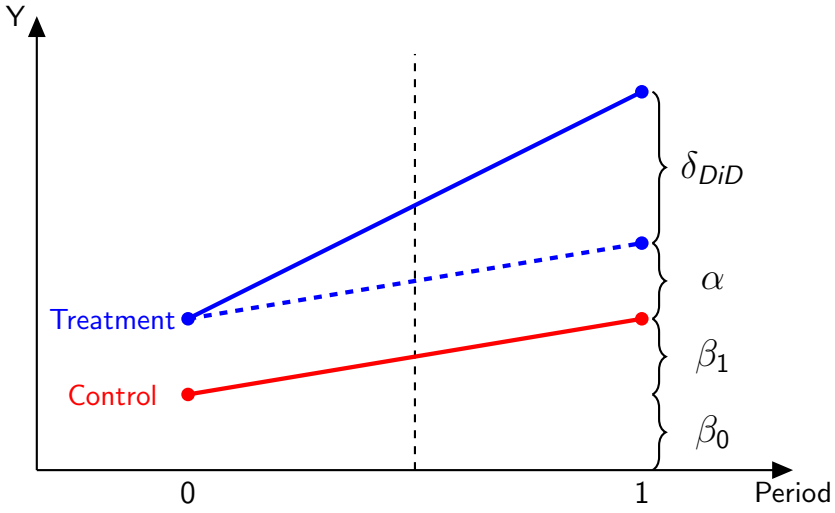
Difference-in-differences

- Randomization ensures that treatment and control group are on average similar
- If we have data over time, we can allow for differences across groups as long as they stay constant over time
- Take the following linear model with G indicating whether an observation belongs to the treatment group and T a dummy for the post-treatment period

$$Y = \beta_0 + \beta_1 T + \delta_{DiD}(G \cdot T) + \alpha G + \varepsilon$$

  - Example: test the effect of a new educational approach on grades at CBS with students at KU as control group
  - The model allows for a common time trend β_1, a fixed effect for the treatment group α, and a separate time trend for the treatment group which is the treatment effect of interest
  - Because DiD allows for differences across groups, it is also a popular technique for non-experimental data

Difference-in-differences (II)



Difference-in-differences (III)

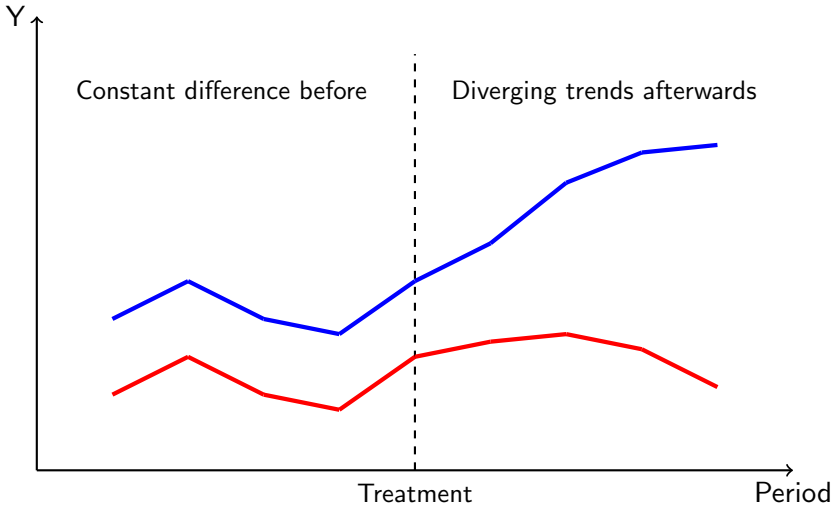
- For the linear DiD model take averages per group and time period

$$\bar{Y}_{treat,post} = \beta_0 + \beta_1 + \delta_{DiD} + \alpha$$
$$\bar{Y}_{treat,pre} = \beta_0 + \alpha$$
$$\bar{Y}_{control,post} = \beta_0 + \beta_1$$
$$\bar{Y}_{control,pre} = \beta_0$$

- Then we can see where the estimator has its name from

$$\begin{aligned}
&(\bar{Y}_{treat,post} - \bar{Y}_{treat,pre}) - (\bar{Y}_{control,post} - \bar{Y}_{control,pre}) \\
&= (\beta_0 + \beta_1 + \delta_{DiD} + \alpha - (\beta_0 + \alpha)) - (\beta_0 + \beta_1 - \beta_0) \\
&= (\beta_1 + \delta_{DiD}) - \beta_1 \\
&= \delta_{DiD}
\end{aligned}$$

More than two time periods



Democratizing online Controlled Experiments at Booking.com

Case Discussion

- Take 20 minutes to read the following paper: `https://arxiv.org/abs/1710.08217`
- Discussion Questions:
  1. What are they trying to achieve specifically with this decentralized approach to experimentation?
  2. What are the practical challenges for implementation?
  3. How can you make sure that experiments are well aligned with the overall strategy of the organization?
  4. How can you ensure organizational learning and retention of causal knowledge?
  5. How do you ensure continued stakeholder support for the experimentation approach?

Thank you

Personal Website: p-hunermund.com

Twitter: @PHuenermund

Email: phu.si@cbs.dk

References I

Teppo Felin and Todd R Zenger. Entrepreneurs as theorists: On the origins of collective beliefs and novel strategies. *Strategic Entrepreneurship Journal*, 3(2):127–146, 2009.

Teppo Felin and Todd R Zenger. The theory-based view: Economic actors as theorists. *Strategy Science*, 2(4):258–271, 2017.

Causal Data Science for Business Decision Making

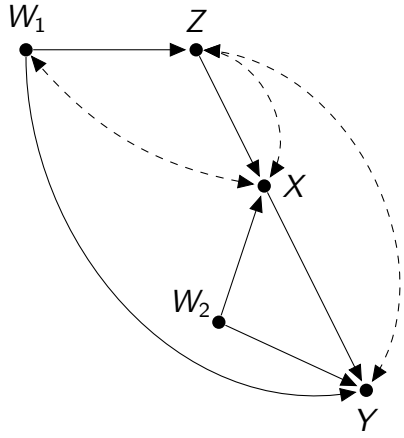
Surrogate Experiments

Paul Hünermund

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Surrogate Experiments

- Let's have a look at a causal graph like the following



Surrogate Experiments (II)

- There is no way to identify $P(y|do(x))$ via backdoor adjustment
- The set $\{W_1, W_2, Z\}$ is not backdoor-admissible because Z is a collider on the path $X \dashleftarrow\dashrightarrow Z \dashleftarrow\dashrightarrow Y$
- If we could run an experiment in which we manipulated X, we could delete all the incoming arrows into X and immediately read of $P(y|do(x))$ from the post-intervention distribution
- But what if that's not possible? Could we use experimental variation in another variable instead to get at the causal effect of interest?
- It turns out we can: In the above graph, if we are able to manipulate Z, we can transform $P(y|do(x))$ into an expression that only contains $do(z)$ (Bareinboim and Pearl, 2012)
  - Solution: $P(y|do(x)) = \sum_{w_1,w_2} P(y|do(z), x, w_1, w_2)P(w_1)P(w_2)$

Identification by Surrogate Experiment

$\mathcal{Z}$-identification (graphical criterion; Bareinboim and Pearl, 2012)

Let X, Y, Z be disjoint sets of variables and let G be the causal graph. The causal effect $Q = P(y|do(x))$ is zID in G if one of the following conditions hold:

(i) Q is identifiable in G; or

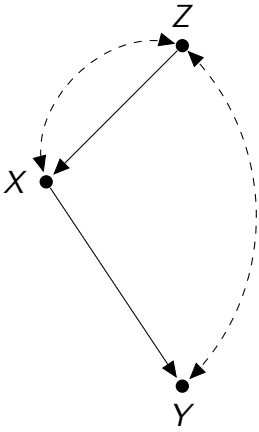
(ii) There exists $Z' \subseteq Z$ such that the following conditions hold,

a. X intercepts all directed paths from Z' to Y, and

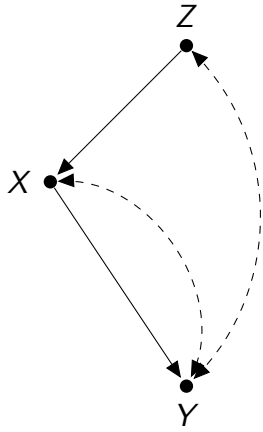
b. Q is identifiable in $G_{\overline{Z'}}$.

- Since the entire post-interventional distribution is identified, also other quantities such as the average causal effect are
- This graphical criterion is only a sufficient condition but not necessary for identification (i.e., there exist solutions that do not fulfill these criteria)

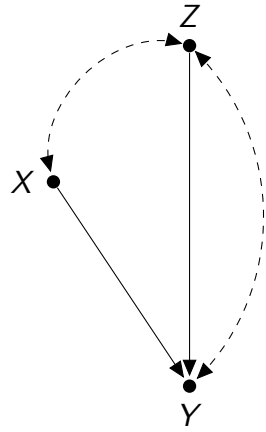
Test: In which causal graphs is $P(y|do(x))$ z-identifiable?



A



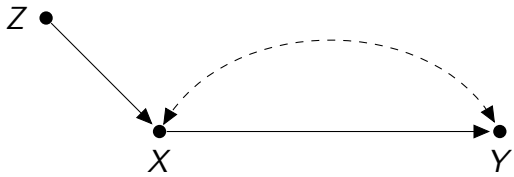
B



C

Instrumental Variables

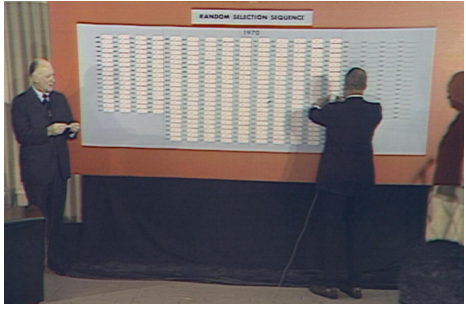
- $\mathcal{Z}$-identification does not allow for unobserved confounders that directly affect treatment and outcome
- With such a direct unobserved confounder, there is no way to identify $P(y|do(x))$ nor the average causal effect (Manski, 1990; Balke and Pearl, 1995)
- There is, however, a way to obtain some causal insights if we are willing to introduce an additional *monotonicity* assumption (Imbens and Angrist, 1994)
  - Monotonicity $\hat{=}$ every individual's treatment status X is affected by the instrument Z in the same direction



Example: Vietnam Draft

Vietnam Draft Lottery

- Conscription for serving in the Vietnam war was organized as a (somewhat macabre) lottery of birthdates for men born between 1944 and 1950
- From an urn with all dates of the year September 14 was drawn first and got assigned the number 1, the second date drawn got assigned the number 2 and so on. The first 195 birthdates drawn were eventually drafted.
- The lottery creates exogenous variation in military service, which can be used to estimate the labor market effects of veteran status (Angrist, 1990)



Local Average Treatment Effect

- For a binary instrument and binary treatment, we can divide the population in four subgroups depending on how their treatment status reacts to the instrument

$$
\begin{aligned}
\text{Compliers:} \quad & X^{Z=0} = 0 \text{ and } X^{Z=1} = 1 \\
\text{Defiers:} \quad & X^{Z=0} = 1 \text{ and } X^{Z=1} = 0 \\
\text{Always takers:} \quad & X^{Z=0} = 1 \text{ and } X^{Z=1} = 1 \\
\text{Never-takers:} \quad & X^{Z=0} = 0 \text{ and } X^{Z=1} = 0
\end{aligned}
$$

- Compliers only serve in the military ($X = 1$) if they get drafted ($Z = 1$)
- Always-takers do military service ($X = 1$) irrespective of wether they get drafted or not $()Z$), and so forth
- Monotonicity assumption rules out the existence of defiers

Local Average Treatment Effect (II)

- If there are no defiers we can identify the causal effect of X on Y for the subgroup of compliers (Imbens and Angrist, 1994)
  - But only for this subgroup! The literature therefore calls this estimand a **"local average treatment effect"**
  - We can't say anything about the always- and never-takers, unless everyone has the same (homogenous) treatment effect, then LATE = ATE (special case)
- Problem: It's often hard to tell who the compliers are
  - Are compliers representative for the entire population?
  - The estimated LATE might thus not tell us much about the likely effect for non-compliers
- Problem 2: If the instrument doesn't effect treatment status X by much, the subgroup of compliers will be small
  - A small complier group can render effect estimates very unstable (small effective sample size), which is the so-called "weak instrument problem" stated in causal terms

Regression Discontinuity Design

Strategic Management Journal
Strat. Mgmt. J., **38**: 1827–1847 (2017)
Published online EarlyView 7 February 2017 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/smj.2629
Received 1 July 2015; Final revision received 10 October 2016

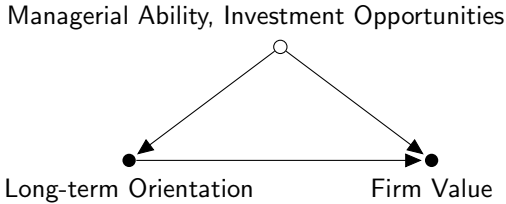
DOES A LONG-TERM ORIENTATION CREATE VALUE? EVIDENCE FROM A REGRESSION DISCONTINUITY

CAROLINE FLAMMER[1]* and PRATIMA BANSAL[2]
[1] *Questrom School of Business, Boston University, Boston, Massachusetts, U.S.A.*
[2] *Ivey Business School, Western University, London Ontario, Canada*

Research summary: *In this paper, we theorize and empirically investigate how a long-term orientation impacts firm value. To study this relationship, we exploit exogenous changes in executives' long-term incentives. Specifically, we examine shareholder proposals on long-term executive compensation that pass or fail by a small margin of votes. The passage of such "close call" proposals is akin to a random assignment of long-term incentives and hence provides a clean causal estimate. We find that the adoption of such proposals leads to (1) an increase in firm value and operating performance—suggesting that a long-term orientation is beneficial to companies—and (2) an increase in firms' investments in long-term strategies such as innovation and stakeholder relationships. Overall, our results are consistent with a "time-based" agency conflict between shareholders and managers.*

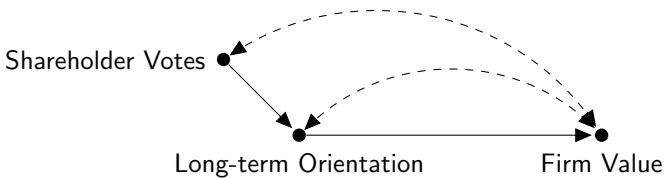
Introduction

- Research question:
  - Do companies face time-based agency problems?
  - Does the provision of long-term incentives to managers increase firm value and stimulate innovation activities?
- Confounding problem:
  - Managerial ability, investment opportunities, etc. drive both the long-term orientation and firm value
  - These confounding influences are unobserved at the firm-level



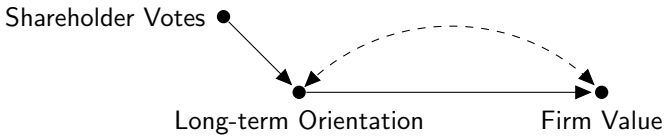
Regression Discontinuity Design (II)

- Research design:
  - Long-term executive compensation affects long-term orientation of a firm by incentivizing managers to create long-term value
  - By itself, shareholder votes on executive compensation plans are likely driven by the same unobservables though



Regression Discontinuity Design (III)

- Discontinuity:
  - Shareholders in public firms vote on executive compensation plans that incentivize long-term orientation
  - If we look at very "close call" votes, let's say between 49% and 51% for the proposal, we can reasonably assume that the respective firms do not differ systematically below and above the cutoff of 50%
  - At the same time, making the cut leads to a large impact on long-term orientation
  - I.e., in a close area around the cutoff, shareholder votes are a good instrument for long-term orientation

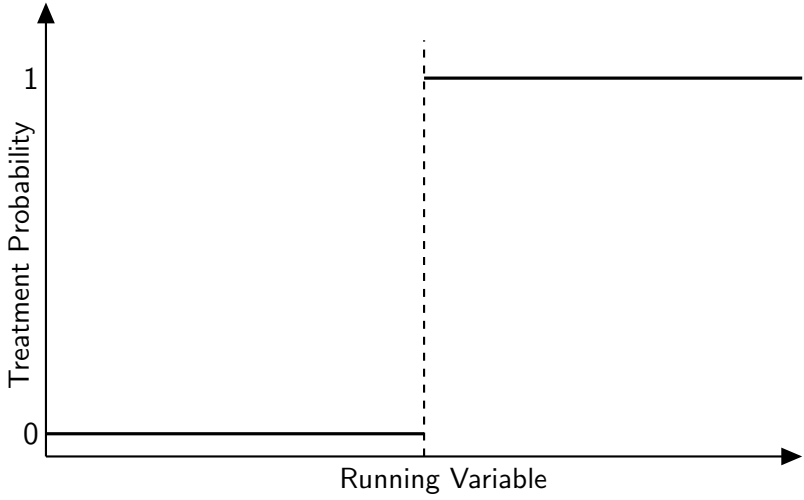


Sharp vs. Fuzzy RDD

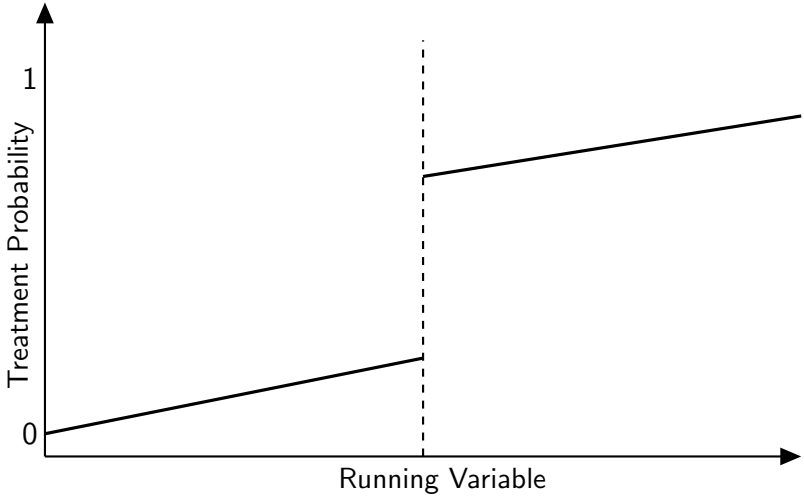
- There are two types of RDDs

1. Sharp
   - Probability to receive treatment jumps from zero to one at the discontinuity
   - Everyone above the threshold is treated and no-one below
2. Fuzzy
   - Probability to receive treatment jumps discontinuously but from a value above zero to a value below one
   - It's more likely to be treated if you're above the discontinuity, but this is not certain
   - The specification in Flammer and Bansal (2017) corresponds to a sharp RDD design

Sharp RDD



Fuzzy RDD



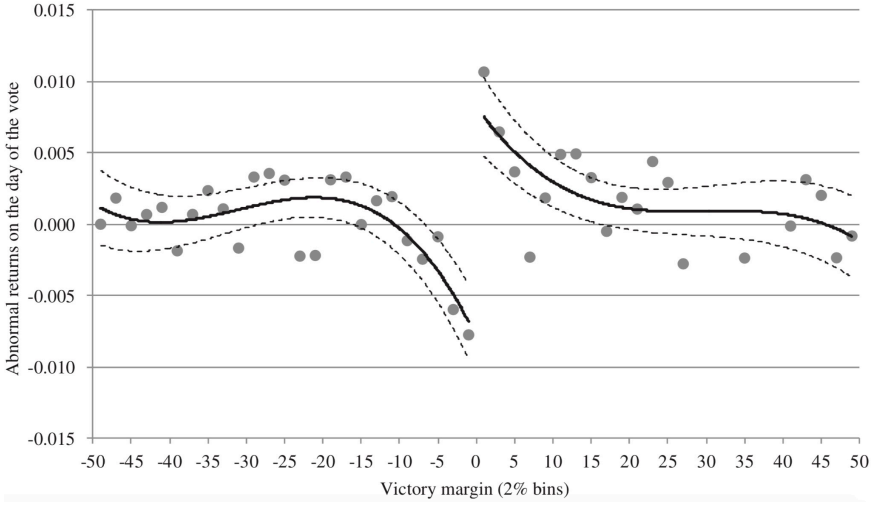
Estimation

- The basic estimation idea of an RDD could not be easier
  - Go as close to the discontinuity z_0 as possible and compare means below and above the threshold
- Problem with this approach:
  - The closer we make the window around z_0, the more data we lose, which makes our estimates unreliable
  - The wider we make the window, the more bias we possibly buy in
- Almost all practical issues with implementing RDDs revolve around this variance-bias trade-off
- Another drawback is external validity: an RDD only allows us to say something about a very specific population around the threshold
  - E.g., firms in which shareholder proposals on long-term compensation barely pass, might be very different from those where the proposal fails with zero votes in favor

Estimation (II)

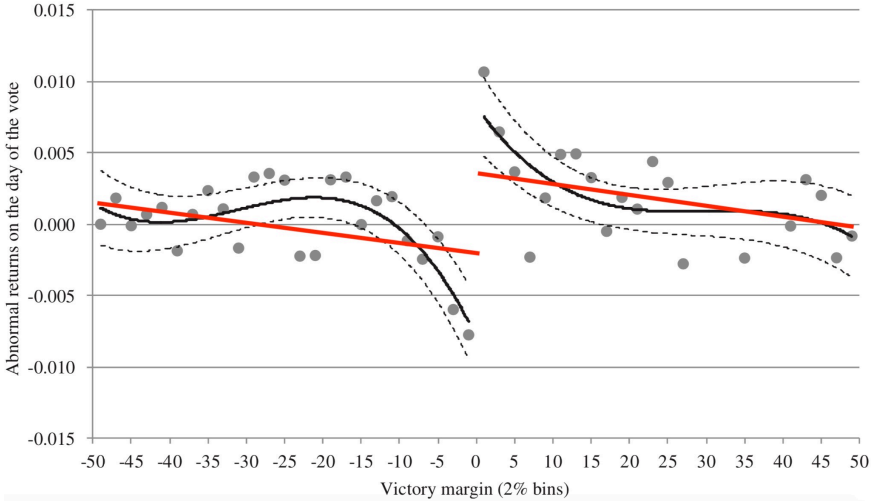
- First step in an RDD analysis is usually to plot the data to see what's going on
  - Is the jump at z_0 visually detectable?
- Instead of comparing means in a close window around z_0, we can fit two straight regressions lines below and above
  - The causal effect estimate is then just the difference between the two lines at z_0
- This functional form assumption is often too rigid though, because we can't be sure that everything is nicely linear
- Alternatively, we can fit more flexible polynomial regressions (including quadratic, cubic, etc, terms of the running variable) or use nonparametric regression techniques

Sensitivity to Functional Form Assumptions



Source: Flammer and Bansal (2017)

Sensitivity to Functional Form Assumptions

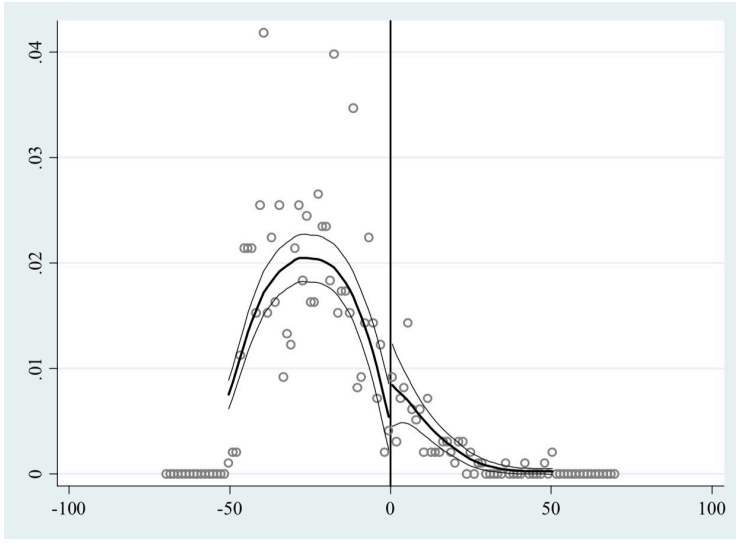


Source: Flammer and Bansal (2017)

Diagnostics

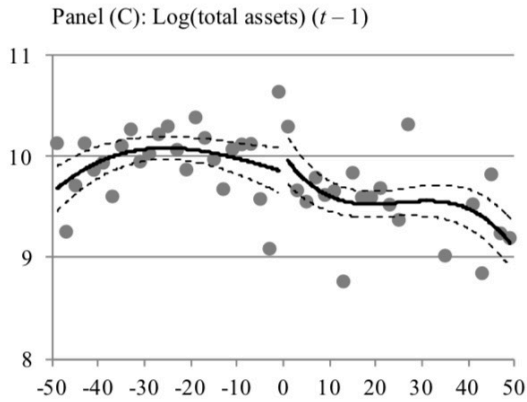
- The entire identification strategy in an RDD depends on the notion that there are no systematic differences between treatment and control group above and below the threshold
- What can go wrong?
- We might see "bunching" below or above the threshold
  - This would be an indication that individuals are somehow able to manipulate their running variable
  - Example: persuade teachers to still give minimum passing grade to go to college
  - This raises concern about self-selection: are those individuals that manage to manipulate their running variable different from the others?
- Do other covariates change discontinuously too?

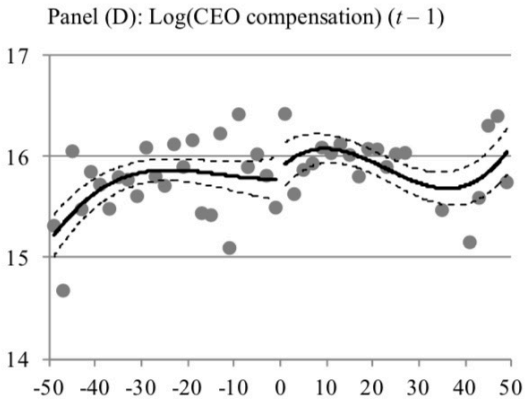
McCrary Test



Source: Flammer and Bansal (2017), Online Appendix

Smoothness of Covariate Distribution





Source: Flammer and Bansal (2017), Online Appendix

Thank you

Personal Website: p-hunermund.com

Twitter: @PHuenermund

Email: phu.si@cbs.dk

References I

Joshua D. Angrist. Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *The American Economic Review*, 80(3):313–336, 1990.

A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92:1171–1176, 1995.

Elias Bareinboim and Judea Pearl. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 113–120, 2012.

Esther Duflo, Rachel Glennerster, and Michael Kremer. Using randomization in development economics research: A toolkit. In *Handbook of Development Economics*, volume 4, chapter 61. Elsevier, 2008.

Caroline Flammer and Pratima Bansal. Does a long-term orientation create value? Evidence from a regression discontinuity. *Strategic Management Journal*, 38:1827–1847, 2017.

Guido W. Imbens and Joshua D. Angrist. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475, 1994.

Charles F. Manski. Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80:319–323, 1990.

J. McCrary. Manipulation of the running variable in the regression discontinuity design: a density test. *Journal of Econometrics*, 142(2):698–714, 2008.

Causal Data Science for Business Decision Making

Causal Artificial Intelligence

Paul Hünermund

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

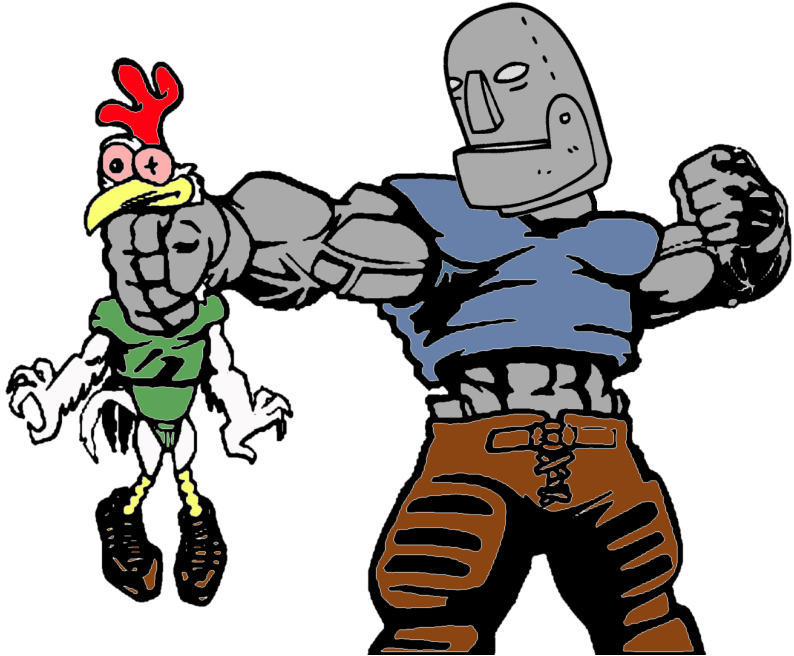
ASSOCIATION AMBA ACCREDITED

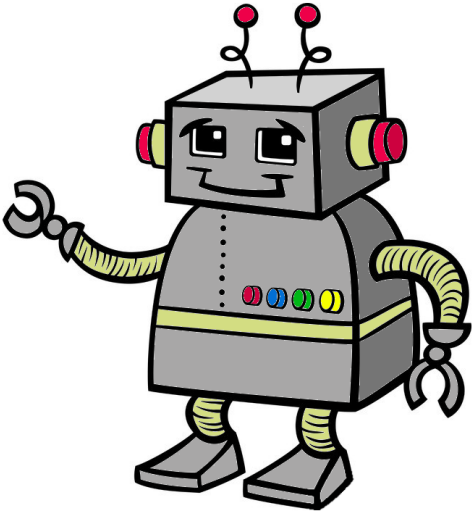
CEMS

PIM

Remember: Writing Challenge

- Work in teams of 3–5 students
- Write a 800–1500 words blog post about a topic of your choice
  - E.g., pick out a technique or topic we have discussed in class and illustrate its relevance for industry
  - Try to find practically relevant examples
  - Format similar to articles on `www.medium.com` or `www.towardsdatascience.com`
- Feedback opportunity for you!!
- The best three submissions will be published on `www.causalscience.org`
- Deadline: November 28, 2021





Why Causality and AI?

The way you talk about curve fitting, it sounds like you're not very impressed with machine learning.

No, I'm very impressed, because we did not expect that so many problems could be solved by pure curve fitting. It turns out they can. But I'm asking about the future — what next? Can you have a robot scientist that would plan an experiment and find new answers to pending scientific questions? That's the next step. We also want to conduct some communication with a machine that is meaningful, and meaningful means matching our intuition. If you deprive the robot of your intuition about cause and effect, you're never going to communicate meaningfully. Robots could not say "I should have done better," as you and I do. And we thus lose an important channel of communication.

Do-Calculus

- Remember: Identification task (with observational data) $\hat{=}$ we need to transform our target query $Q = P(y|do(x))$ into an expression that only contains standard probability objects
- For that we need to know the rules on how to deal with these do-objects: *do-calculus*
  - Do-calculus is a powerful symbolic machinery that provides a set of inference rules by which sentences involving interventions can be transformed into other sentences (Pearl, 2009; Pearl et al., 2016)
- Do-calculus can be used to solve the confounding problem
  - Apply the rules of *do*-calculus repeatedly until a do-expression is translated into an equivalent expression involving only standard probabilities of observed quantities

Do-Calculus (II)

Theorem: Rules of Do-Calculus (Pearl, 2009, p. 85)

Let G be the directed acyclic graph associated with a [structural] causal model [...], and let $P(\cdot)$ stand for the probability distribution induced by that model. For any disjoint subset of variables X, Y, Z, and W, we have the following rules.

Rule 1 (Insertion/deletion of observations):

$$P(y|do(x), z, w) = P(y|do(x), w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}.$$

Rule 2 (Action/observation exchange):

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}.$$

Do-Calculus (III)

Rule 3 (Insertion/deletion of actions):

$$P(y|do(x), do(z), w) = P(y|do(x), w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{XZ(W)}}},$$

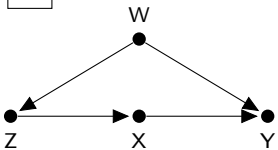
where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

- $G_{\overline{X}}$ denotes the graph obtained by deleting from G all arrows pointing to nodes in X
- $G_{\underline{X}}$ denotes the graph obtained by deleting from G all arrows emerging from nodes in X

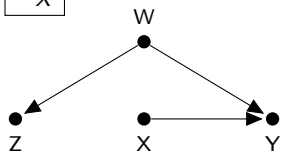
Do-Calculus Rule 1

$$P(y|do(x), z, w) = P(y|do(x), w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}$$

G



$G_{\overline{X}}$



The graph that results from deleting all arrows pointing into X in G is denoted by $G_{\overline{X}}$.

In $G_{\overline{X}}$, W blocks the only backdoor path between Z and Y: $Z \leftarrow W \rightarrow Y$.

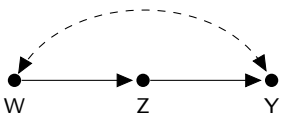
By d-separation $(Y \perp\!\!\!\perp Z|W)_{G_{\overline{X}}}$ we can therefore ignore Z in the conditional distribution of Y.

$$P(y|do(x), z, w) = P(y|do(x), w)$$

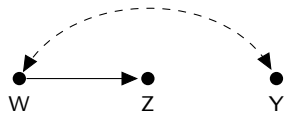
Do-Calculus Rule 2

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \qquad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}$$

G



$G_{\underline{Z}}$



Assume we are interested in the query $P(y|do(z), w)$. The graph that results from deleting all arrows emitted by Z in G is denoted by $G_{\underline{Z}}$.

In $G_{\underline{Z}}$, W blocks the only backdoor path between Z and Y: $Z \leftarrow W \dashleftarrow\dashrightarrow Y$.

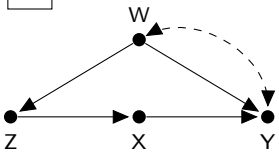
Thus, by d-separation $(Y \perp\!\!\!\perp Z|W)_{G_{\underline{Z}}}$ and therefore the second rule of do-calculus applies.

Consequently, we can get rid of the do-operator by setting $P(y|do(z), w) = P(y|z, w)$. The latter expression is estimable from observational data.

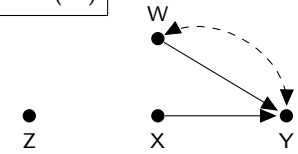
Do-Calculus Rule 3

$$P(y|do(x), do(z), w) = P(y|do(x), w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{XZ(W)}}},$$

G



$G_{\overline{XZ(W)}}$



Z is not an ancestor of W, therefore we delete all arrows pointing into Z and X from G to arrive at $G_{\overline{XZ(W)}}$.

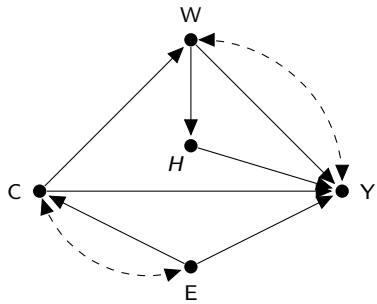
In $G_{\overline{XZ(W)}}$, Z and Y are d-separated: $(Y \perp\!\!\!\perp Z)_{G_{\overline{XZ(W)}}}$

Thus, we can delete the intervention on Z because it is not relevant anymore after we have already intervened on X

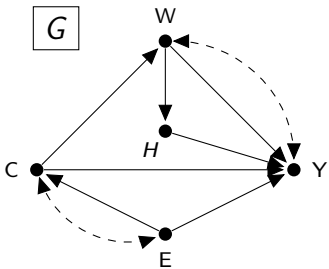
$$P(y|do(x), do(z)) = P(y|do(x))$$

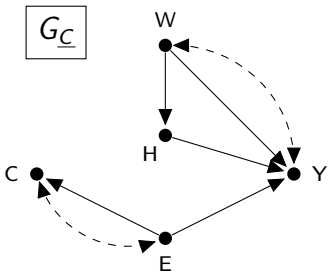
Example: Applying Do-Calculus

- Take the example of the college wage premium from one of the previous lectures
  - C: college degree
  - Y: earnings
  - W: occupation
  - H: work-related health
  - E: other socio-economic factors
- Task: Transform $P(y|do(c))$ into a do-free expression by using the rules of do-calculus



Example: Applying Do-Calculus (II)





There are two backdoor paths in G, which can both be blocked by E. Conditioning and summing over all values of E yields (law of total probability)

$$P(y|do(c)) = \sum_e P(y|do(c), e)P(e|do(c)).$$

By rule 2 of do-calculus

$$P(y|do(c), e) = P(y|c, e), \quad \text{since } (Y \perp\!\!\!\perp C|E)_{G_{\underline{C}}}.$$

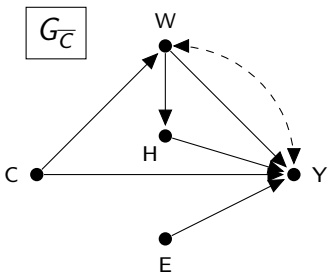
Refresher: Law of Total Probability

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_k)P(B_k)$$
$$= \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

- Suppose there are 60% men in the class and 40% women
- Among women, the probability to get an A in the class is 25% while for men it is only 20%
- Then the overall probability to get an A in the class is

$$P(A) = P(A|male)P(male) + P(A|female)P(female)$$
$$= 0.2 \cdot 0.6 + 0.25 \cdot 0.4$$
$$= 0.22$$

Example: Applying Do-Calculus (III)



By rule 3 of do-calculus

$$P(e|do(c)) = P(e), \quad \text{since } (E \perp\!\!\!\perp C)_{G_{\overline{C}}}.$$

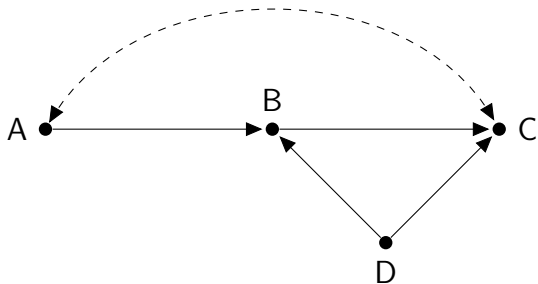
It follows that

$$P(y|do(c)) = \sum_{e} P(y|c,e)P(e).$$

The right-hand-side expression is do-free and can therefore be estimated from observational data.

Exercise: Do-Calculus

- Using the rules of do-calculus, show that $P(C|do(B))$ is identifiable via the backdoor adjustment formula in the following graph



Completeness of Do-Calculus

- Do-calculus is shown to be *complete*, meaning that if a causal effect is identifiable there exists a sequence of steps applying the rules of do-calculus that transforms the causal effect formula into an expression that includes only observable quantities (Shpitser and Pearl, 2006; Huang and Valtorta, 2006)
  - Put differently, if do-calculus fails, the causal effect is guaranteed to be unidentifiable
- Completeness proofs are notoriously difficult and showing this for the case of do-calculus was a major breakthrough in the literature (Pearl and Mackenzie, 2018)

Automatizing the Identification Task

- We know that do-calculus is complete, but the theorem is only procedural and does not tell us which series of steps leads to the desired solution
- Shpitser and Pearl (2006), building on work by Tian and Pearl (2002), propose an algorithm that automates this task
  - The algorithm takes a description of a DAG as input an returns an expression for a queried causal effect, if it exists
  - Since the algorithm is based on *do*-calculus and therefore complete, if it doesn't return a causal effect expression involving only observable quantities, no such expression exists
  - This approach is extremely user-friendly
    - Basically the computer is doing everything for you
    - But you should be aware of what's going on "under the hood"
    - To get a gist of how this algorithm works, see: `https://david-salazar.github.io/2020/07/31/causality-testing-identifiability/`

Identification Algorithms

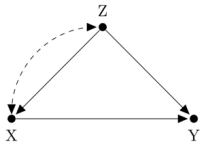
- The algorithm by Shpitser and Pearl (2006) is for the observational data case where $P(y|do(x))$ is supposed to be transformed into an expression that does not contain a do-operator (with unobservables included in the model)
- But similar algorithms have been developed for the other causal inference tasks we will discuss throughout this course
  - $\mathcal{Z}$-Identification (Bareinboim and Pearl, 2012)
  - Selection bias (Bareinboim and Tian, 2015)
  - Transportability (Bareinboim and Pearl, 2013, 2014)
- Input:
  1. A causal query Q
  2. The model in form of a diagram
  3. The type of data available
- Output: an estimable expression of Q
  - Most algorithms inherit *completeness* property from do-calculus

The Data Fusion Process

(1) Query:

Q = Causal effect at target population

(2) Model:



(3) Available Data:

Observational:	P(v)
Experimental:	P(v | do(z))
Selection-biased:	P(v | S = 1) + P(v | do(x), S = 1)
From different populations:	$P^{(source)}$(v | do(x)) + observational studies

Causal Inference Engine:

Three inference rules of *do-calculus*

Solution exists? Yes → Estimable expression of Q

No → Assumptions need to be strengthened (imposing shape restrictions, distributional assumptions, etc.)

Does Your Dog Understand Cause and Effect?

Article | Open Access | Published: 15 September 2017

The effects of domestication and ontogeny on cognition in dogs and wolves

Michelle Lampe ✉, Juliane Bräuer, Juliane Kaminski & Zsófia Virányi

Scientific Reports **7**, Article number: 11690 (2017) | Cite this article

Abstract

Cognition is one of the most flexible tools enabling adaptation to environmental variation. Living close to humans is thought to influence social as well as physical cognition of animals throughout domestication and ontogeny. Here, we investigated to what extent physical cognition and two domains of social cognition of dogs have been affected by domestication and ontogeny. To address the effects of domestication, we compared captive wolves (n = 12) and dogs (n = 14) living in packs under the same conditions. To explore developmental effects, we compared these dogs to pet dogs (n = 12) living in human families. The animals were faced with a series of object-choice tasks, in which their response to communicative, behavioural and causal cues was tested. We observed that wolves outperformed dogs in their ability to follow causal cues, suggesting that domestication altered specific skills relating to this domain, whereas developmental effects had surprisingly no influence. All three groups performed similarly in the communicative and behavioural conditions, suggesting higher ontogenetic flexibility in the two social domains. These differences across cognitive domains need to be further investigated, by comparing domestic and non-domesticated animals living in varying conditions.

Thank you

Personal Website: p-hunermund.com

Twitter: @PHuenermund

Email: phu.si@cbs.dk

References I

Elias Bareinboim and Judea Pearl. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 113–120, 2012.

Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134, 2013.

Elias Bareinboim and Judea Pearl. Transportability from multiple environments with limited experiments: Completeness results. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances of Neural Information Processing Systems*, volume 27, pages 280–288, November 2014.

Elias Bareinboim and Jin Tian. Recovering causal effects from selection bias. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Yimin Huang and Marco Valtorta. Pearl's calculus of interventions is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI2006)*, 2006. URL `https://arxiv.org/ftp/arxiv/papers/1206/1206.6831.pdf`.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, United States, NY, 2nd edition, 2009.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 2018. ISBN 9780465097609.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons Ltd, West Sussex, United Kingdom, 2016.

References II

Ilya Shpitser and Judea Pearl. Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models. In *Twenty-First National Conference on Artificial Intelligence*, 2006.

J. Tian and J. Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 567–573, Menlo Park, CA, 2002. AAAI Press/The MIT Press.

Causal Data Science for Business Decision Making

Sample Selection Bias

Paul Hünermund

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

CEMS

PIM

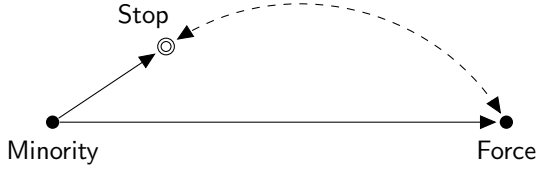
Administrative Records Mask Racially Biased Policing

DEAN KNOX *Princeton University*
WILL LOWE *Hertie School of Governance*
JONATHAN MUMMOLO *Princeton University*

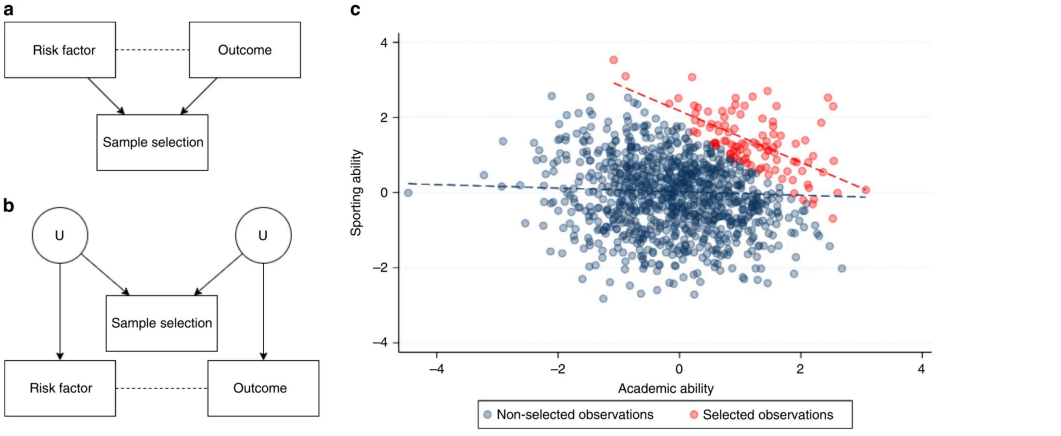
Researchers often lack the necessary data to credibly estimate racial discrimination in policing. In particular, police administrative records lack information on civilians police observe but do not investigate. In this article, we show that if police racially discriminate when choosing whom to investigate, analyses using administrative records to estimate racial discrimination in police behavior are statistically biased, and many quantities of interest are unidentified—even among investigated individuals—absent strong and untestable assumptions. Using principal stratification in a causal mediation framework, we derive the exact form of the statistical bias that results from traditional estimation. We develop a bias-correction procedure and nonparametric sharp bounds for race effects, replicate published findings, and show the traditional estimator can severely underestimate levels of racially biased policing or mask discrimination entirely. We conclude by outlining a general and feasible design for future studies that is robust to this inferential snare.

The Problem of Selection Bias

- Fryer (2019) studied the extent to which racial minorities (blacks, Hispanics) in the U.S. experience police violence more frequently (while controlling for context and civilian behavior) and found relatively small effects
- Knox et al. (2020) criticize this and similar papers that try to measure the degree of racial-bias in policing with the help of administrative records
  - Problem: An individual only appears in the data, if it was stopped by the police
  - If there is a racial bias in policing, stopping can be the result of minority status
  - There are unobserved confounders, such as officers' suspicion, between the selection variable and outcome
  - In a reanalysis they find effects that are at least four times larger



Sample Selection and 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).

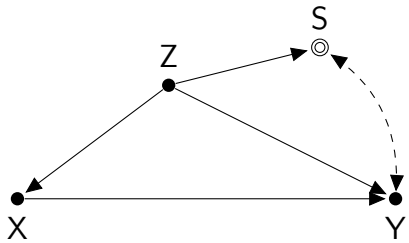
Two Special Cases

- Traditionally, economists knew two techniques for dealing with selection bias
  1. Selection propensity scores
  2. (Heckman) selection models
- The selection propensity score is the probability of being included in the sample, S, given the parent nodes of S: $e(PA) = P(s|PA)$
  - In the policing example this would be $P(stop|minority)$
- If we are willing to make certain functional form assumptions about e
  - e.g., that $e(PA)$ is monotonically increasing (stopping probability increases with minority status)
  - Then by conditioning on $e(PA)$ in the analysis, we can control for selection bias (Angrist, 1997)

Heckman Selection Model

- Heckman (1976, 1979) was interested in studying the hourly wage of women
- His sample included 2,253 working women interviewed in 1967
- Problem: Wages are not observed for women who choose not to work
  - If the wages they are able to obtain on the market are too low, some women might decide to stay at home
  - Working women are not a representative sample of the population (keep in mind, this is the 60s)

Heckman Selection Model (II)



$$s_i \leftarrow \quad \mathbb{1}[Z_i'\delta - \eta_i > 0]$$

$$y_i \leftarrow \begin{cases} x_i\beta + Z_i'\gamma + \varepsilon_i & \text{if } s_i = 1, \\ \text{unobserved} & \text{if } s_i = 0. \end{cases}$$

X: hours worked

Y: wages

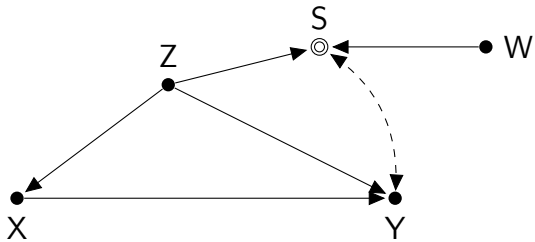
Z: other socio-economic factors

S: Sample selection indicator

with $Corr(\eta, \varepsilon) \neq 0$, i.e., there are unobservables that jointly affect whether a women chooses to work and her market wage (e.g., financial situation)

Heckman Selection Model (III)

- By making specific distributional assumption, such as joint normality of η and ε, we can recover from selection bias
- The standard version of the Heckman selection model, where Z_i is the same in both equations, relies heavily on distributional assumptions
- To avoid this, you should generally have at least one observed variable W that affects S but not Y in the model ("exclusion restriction", similar to an instrument)



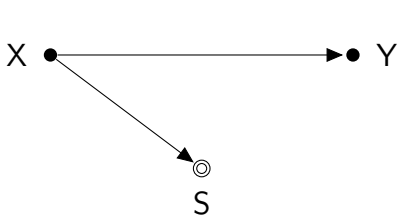
Selection Bias in Causal Diagrams

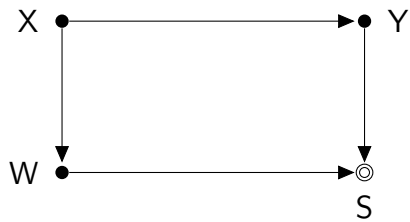
Task: Use the rules of do-calculus to transform $Q = P(y|do(x))$ into an expression that only contains probabilities conditional on $S = 1$

- There is a principled solution for dealing with selection bias in DAGs based on do-calculus (Bareinboim and Pearl, 2012; Bareinboim et al., 2014; Bareinboim and Tian, 2015)
- Compared to standard approaches in econometrics, these results do not rely on
  - functional-form assumptions about the selection propensity score $P(s|PA)$ (Heckman, 1979)
  - or, ignorability of the selection mechanism (Angrist, 1997)
- In addition, there is the possibility to combine biased and unbiased data in order to increase identifying power (Bareinboim et al., 2014; Correa et al., 2017)
  - In many applied settings, unbiased records of covariates are available from secondary data sources (e.g., census data)

Selection Diagram

- We can model the selection mechanism by incorporating a variable S in the DAG, which denotes whether we observe an observation ($S = 1$) or not ($S = 0$)
  - S receives an incoming arrow from those variables in the model that cause the sample selection
  - We denote a DAG that has been augmented with a selection node by G_S





Recovering from Selection Bias

- How can we recover the causal effect from a selected sample without invoking functional form assumptions such as linearity and normality (as in the famous Heckman selection model Heckman, 1979)?
- In order to recover $P(y|do(x))$, we need to translate it into a do-free expression that also conditions on $S = 1$, because that's all we can observe
- Bareinboim and Pearl (2012), Bareinboim et al. (2014) and Bareinboim and Tian (2015) give this problem a full formal treatment and derive graphical criteria and algorithms for recovering causal effects from selected samples

Recovering from Selection Bias (II)

Theorem 1 Bareinboim and Tian (2015)

The conditional distribution $P(y|t)$ is recoverable from G_S (as $P(y|t, S = 1)$) if and only if $(Y \perp\!\!\!\perp S|T)$.

Corollary 1 Bareinboim and Tian (2015)

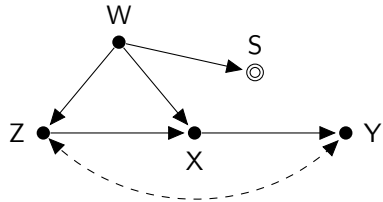
The causal effect $Q = P(y|do(x))$ is recoverable from selection-biased data if using the rules of the do-calculus, Q is reducible to an expression in which no do-operator appears, and recoverability is determined by the previous Theorem.

Recovering from Selection Bias (III)

- In the previous selection diagram with $X \rightarrow Y$ and $X \rightarrow S$, the causal effect is recoverable because $P(y|do(x)) = P(y|x)$ and S is d-separated from Y
  - Thus, $P(y|do(x)) = P(y|x, S = 1)$, which can be estimated from selection-biased data
- An immediate consequence of Theorem 1 is that if S is directly connected to Y, as in the racial policing case, the causal effect will not be recoverable (without strengthening assumptions)
- Note that corollary 1 is a sufficient but not necessary condition for recoverability
  - We can use do-calculus to reduce $P(y|do(x))$ to a do-free expression that conditions on $S = 1$ in cases when corollary 1 is not applicable
  - Bareinboim and Tian (2015) develop an algorithm which automatizes this step

Do-Calculus Example: Selection Bias

Take the following DAG augmented with selection node S:



By the first rule of do-calculus, since $(S, W \perp\!\!\!\perp Y)$ in $G_{\overline{X}}$ (the resulting graph when all incoming arrows in X are deleted),

$$P(y|do(x)) = P(y|do(x), w, S = 1),$$

$$= \sum_z P(y|do(x), z, w, S = 1)P(z|do(x), w, S = 1),$$

where the second line follows from the law of total probability.

Do-Calculus Example: Selection Bias (II)

Applying rule 2, since $(Y \perp\!\!\!\perp X | W, Z)$ in $G_{\underline{X}}$ (the resulting graph when all arrows emitted by X are deleted), we can eliminate the do-operator in the first term

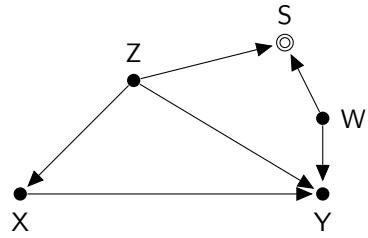
$$= \sum_z P(y|x, z, w, S = 1) P(z|do(x), w, S = 1).$$

Finally, because $(Z \perp\!\!\!\perp X | W)$ in $G_{\overline{X}}$, it follows from rule 3 that

$$= \sum_z P(y|x, z, w, S = 1) P(z|w, S = 1).$$

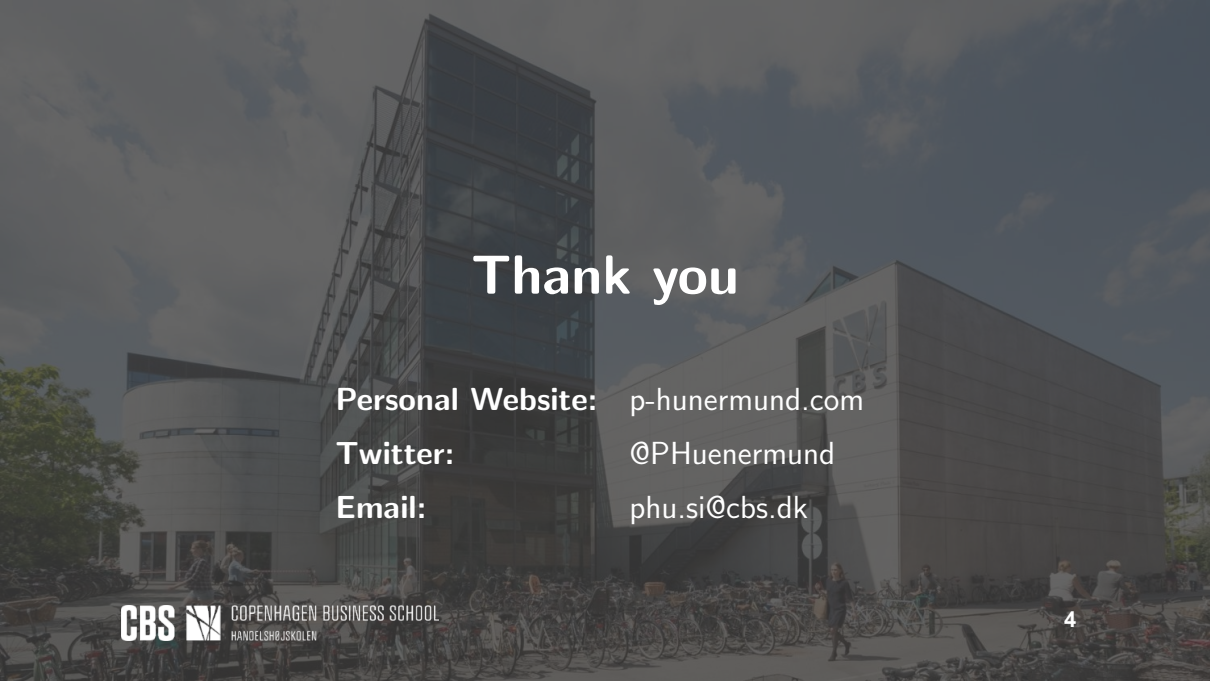
Combining Biased and Unbiased Data

- Sometimes we can get at unbiased measurements of covariates, e.g., from census data
- Take the graph on the right. Conditioning on the set $\{Z, W\}$ closes all backdoor paths and d-separates Y from S. Thus



$$P(y|do(x)) = \sum_{z,w} P(y|x,z,w)P(z,w)$$
$$= \sum_{z,w} P(y|x,z,w,S=1)P(z,w)$$

- $P(Z=z, W=w)$ is not recoverable according to Theorem 1. But if we can get unbiased measurements of Z and Y, this expression is estimable
- Bareinboim et al. (2014) provide more general criteria for this idea



Thank you

Personal Website: p-hunermund.com

Twitter: @PHuenermund

Email: phu.si@cbs.dk

References I

Joshua D. Angrist. Conditional independence in sample selection models. *Economics Letters*, 54:103–112, 1997.

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 100–108, 2012.

Elias Bareinboim and Jin Tian. Recovering causal effects from selection bias. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Juan D. Correa, Jin Tian, and Elias Bareinboim. Generalized adjustment under confounding and selection biases. Technical Report R-29-L, 2017.

Roland G. Fryer. An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, 127(3), 2019. doi: 10.1086/701423.

James Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, 5:475–492, 1976.

James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

Dean Knox, Will Lowe, and Jonathan Mummolo. Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637, 2020.

Causal Data Science for Business Decision Making

Counterfactuals

Paul Hünermund

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

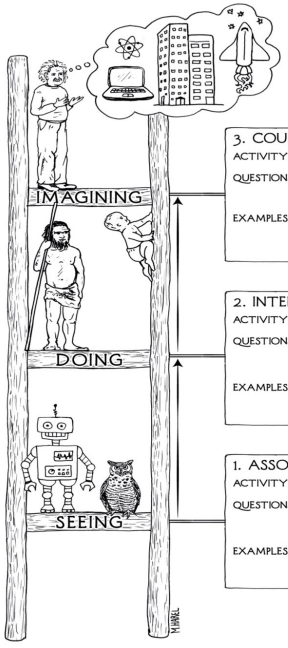
EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

CEMS

PIM



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see …?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?


Watching TikTok
Studying

"The fundamental problem of causal inference" (Holland, 1986)

Individual	Treatment	Y_1	Y_0
1	1	2.6	–
2	0	–	1.7
3	0	–	1.3
4	1	2.5	–
5	0	–	0.1
6	1	1.8	–
7	1	1.2	–
...			

- We do not observe *potential outcomes* Y_0 for treated individuals ($X = 1$) and, vice versa, we do not observe Y_1 for untreated ($X = 0$)

Unconfoundedness

- We can however impute potential outcomes for treated individuals by the average of observed outcomes for untreated individuals, and vice versa, if the treatment is random

$$(Y_1, Y_0) \perp\!\!\!\perp X$$

- In this case, we expect no systematic difference between the treated and untreated group such that no other influence factors lead to biased results
- Often we only require the treatment to be independent of potential outcomes conditional on covariates: $(Y_1, Y_0) \perp\!\!\!\perp X|Z$
- Other names for unconfoundedness: ignorability, exogeneity, exchangeability
- Violation of unconfoundedness: e.g., older people are more liley to get vaccinated, but we fail to account for age in the analysis

Counterfactual Backdoor

Counterfactual interpretation of backdoor (Pearl, 2009)

If a set Z of variables satisfies the backdoor criterion relative to (X, Y), then for all x, the counterfactual Y_x is conditionally independent of X given Z:

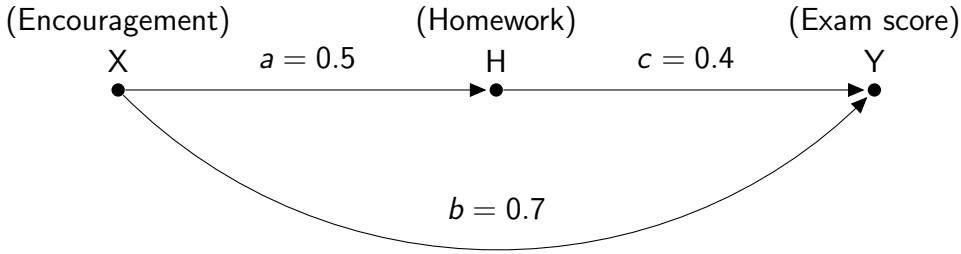
$$Y_x \perp\!\!\!\perp X|Z.$$

- I.e., if we can find a backdoor admissible adjustment set Z based on a causal diagram, then unconfoundedness holds
- We have already seen how we can identify the average causal effect $E[Y|do(X)] = E[Y_1] - E[Y_0]$ in this case

Individual-level Counterfactuals

- But the ACE does not help us to answer questions such as:
  - *"Was it the aspirin that stopped my headache?"*
  - *"Would Kennedy be alive if Oswald had not killed him?"*
  - *"What if I had not smoked for the last 2 years?"*
- Those questions require to move from the population- to individual-level counterfactuals
- Individual-level counterfactuals are generally not identified from the information encoded in causal diagrams alone, for that we need to make assumptions about the *functions* and *background factors* in the structural causal model

Computing Counterfactuals – Example



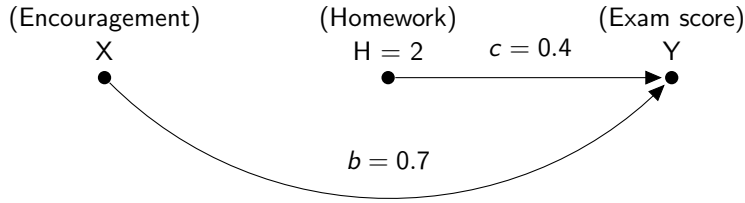
$$X = U_X$$
$$H = a \cdot X + U_H$$
$$Y = b \cdot X + c \cdot H + U_Y$$

Computing Counterfactuals – Example (II)

- Let us assume all nachground factors U are independent and that we are given the values for the coefficients ($a = 0.5$, $b = 0.7$, $c = 0.4$)
  - We can estimate them from population-level data, but we need to make a functional-form assumption (in this case: linearity)
- Query: *"What would Joe's score have been had he doubled his study time?"*
- Assume we record the data for Joe: $X = 0.5$, $H = 1$, and $Y = 1.5$
- From the data we can infer the value of the background variables as:

$$
\begin{aligned}
U_X &= 0.5, \\
U_H &= 1 - 0.5 \cdot 0.5 = 0.75 \\
U_Y &= 1.5 - 0.7 \cdot 0.5 - 0.4 \cdot 1 = 0.75
\end{aligned}
$$

Computing Counterfactuals – Example (III)



- Now we can infer the effect of Joe doubling his study time by replacing the structural equation for H with the constant $H = 2$

$$Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) = 0.5 \cdot 0.7 + 2.0 \cdot 0.4 + 0.75 = 1.9$$

- Joe's grade would increase by 0.4 points or $\sim 27\%$
- This is an indiviual-level counterfactual, because we have computed Joe's individual background factors U_i which in a standard DAG at rung 2 we would not know

Three Steps in Computing Counterfactuals

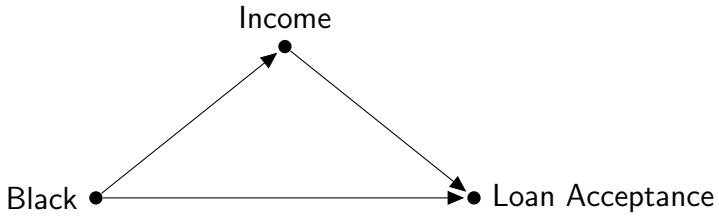
1. **Abduction**: Use evidence $E = e$ to determine the value of U.
2. **Action**: Modify the model M, by removing teh structural equations for the variables in X and replacing them with the appropriate functions $X = x$, to obtain the modified model, M_x.
3. **Prediction**: Use the modified model, M_x, and the value of U to compute the value of Y, the consequence of the counterfactual

Mediation

- Throughout the course, we have talked about total causal effect of a treatment X on an outcome Y
- In many situation, we are furthermore interested in how an effect actually comes about
- What are the intermediate variables M (mediators) that transmit an effect of X on Y?
- In other words, we are interested not only in total but also *path-specific* effects
- This the ability to make this kind of attribution is important for ethical decision-making and in the legal realm

Example: Algorithmic Bias

- Imagine an insurance company that applies an automated credit rating model based on observable customer characteristics
- Although race is not explicitly used as a characteristic in the training algorithm, it is detected that the credit rating model denies loan applications of black customers much more often than for white customers
- Confronted with the claim of biased decision-making, the company responds that black customers systematically report lower income levels, which is why their loan applications are denied more often



Direct and Indirect Effects

- In the mediation setting, we can define four types of effects (assume binary X)
- **Total effect**

$$\begin{aligned} TE &= E[Y_1 - Y_0] \\ &= E[Y|do(X=1)] - E[Y|do(X=0)] \end{aligned}$$

- **Controlled direct effect**

$$\begin{aligned} CDE &= E[Y_{1,m} - Y_{0,m}] \\ &= E[Y|do(X=1, M=m)] - E[Y|do(X=0, M=m)] \end{aligned}$$

  - The controlled direct effect would be the effect of race, if we could set everyone in the population to the same income level
  - While this is certainly a very desirable state from an equity point of view, it is not very relevant for the insurance case (individual income levels do differ)

Direct and Indirect Effects (II)

- **Natural direct effect**

$$NDE = E[Y_{1,M_0} - Y_{0,M_0}]$$

  - Measures the difference in loan acceptance probabilities between black and white customers, if income of whites were set to those levels they *would have attained* if they were black

- **Natural indirect effect**

$$NIE = E[Y_{0,M_1} - Y_{0,M_0}]$$

  - Measures the change in loan acceptance rate if race is held constant, and income is changed to whatever value it would have attained under $X = 1$
  - *NIE* captures the effect that can be explained by mediation alone

Direct and Indirect Effects (III)

- Note that TE and $CDM(e)$ contain do-expressions and can therefore be estimated from population-level data (rung 2), e.g., either from experiments or with the help of the backdoor or frontdoor criterion
- NDE and NIE contain cross-world (nested) counterfactuals, which leads to the fundamental problem of causal inference (we cannot observe the income level that a white customer would have obtained if he were black, and vice versa)
- In our hypothetical example, the insurance company basically claims that the total effect (black customers experience lower acceptance rates) can be fully explained by an indirect effect via income levels
- If there is, however, a direct effect of race on acceptance rates, this would be highly problematic from both an ethical and legal point of view
  - This can occur, even though race is not used explicitly to train the credit rating model, if the algorithm picks up on other characteristics that proxy for race (e.g., name, address, etc.)

Mediation Formula

- In the no confounding case (as in the previous causal diagram), the *NIE* and *NDE* can be identified from observational data

$$NDE = \sum_m \big[E[Y|X=1, M=m] - E[Y|X=0, M=m]\big] P(M=m|X=0)$$

$$NIE = \sum_m E[Y|X=0, M=m]\big[P(M=m|X=1) - P(M=m|X=0)\big]$$

Numerical Example

Black X	Household Income $>$ \$80.000 p.a. M	Acceptance Rate $E[Y\|X=x, M=m]$
0	1	0.80
0	0	0.40
1	1	0.30
1	0	0.20

Black X	Household Income $>$ \$80.000 p.a. $E[M\|X=x]$
0	0.75
1	0.40

Numerical Example (II)

- We get the following numbers for the direct and indirect effects

$$
\begin{aligned}
NDE &= \big[E[Y|X=1,M=1] - E[Y|X=0,M=1]\big] \times P(M=1|X=0) \\
&\quad + \big[E[Y|X=1,M=0] - E[Y|X=0,M=0]\big] \times P(M=0|X=0) \\
&= (0.30 - 0.80) \times 0.75 + (0.20 - 0.40) \times (1 - 0.75) \\
&= -0.425
\end{aligned}
$$

$$
\begin{aligned}
NIE &= E[Y|X=0,M=1] \times \big[P(M=1|X=1) - P(M=1|X=0)\big] \\
&\quad + E[Y|X=0,M=0] \times \big[P(M=0|X=1) - P(M=0|X=0)\big] \\
&= 0.80 \times (0.40 - 0.75) + 0.40 \times (0.60 - 0.25) \\
&= -0.14
\end{aligned}
$$

Numerical Example (II)

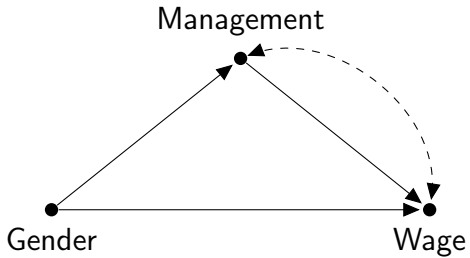
- While the total effect is equal to

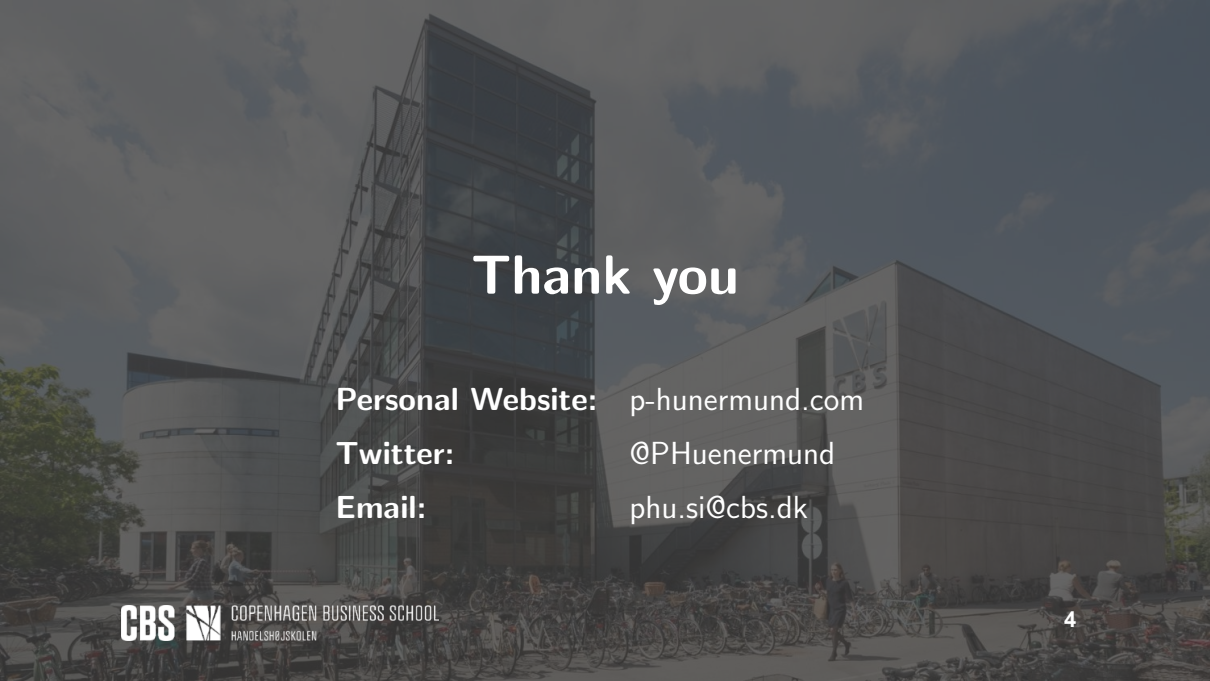
$$\begin{aligned} TE &= E[Y|X=1,M=1] \times P(M=1|X=1) \\ &\quad + E[Y|X=1,M=0] \times P(M=0|X=1) \\ &\quad\quad - \{E[Y|X=0,M=1] \times P(M=1|X=0) \\ &\quad\quad\quad + E[Y|X=0,M=0] \times P(M=0|X=0)\} \\ &= 0.30 \times 0.40 + 0.20 \times 0.60 - \{0.80 \times 0.75 + 0.40 \times 0.25\} \\ &= -0.46 \end{aligned}$$

- $NDE/TE \approx 0.924$ and $1 - NDE/TE \approx 0.076$
  - Only ca. 7.6% of the total effect can be explained by differences in income levels

Confounded Mediation

- Under certain circumstances, the NDE and NIE can also be identified if there is confounding $X \dashleftarrow\dashrightarrow M$ and $M \dashleftarrow\dashrightarrow Y$
- But the mediation formula is more complicated in this case (see Pearl et al., 2016, sec. 4.5.2)
- Unobserved confounding is often a serious obstacle for mediation analysis in practice!





Thank you

Personal Website: p-hunermund.com

Twitter: @PHuenermund

Email: phu.si@cbs.dk

References I

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, December 1986.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, United States, NY, 2nd edition, 2009.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons Ltd, West Sussex, United Kingdom, 2016.

Causal Data Science for Business Decision Making

Transportability & External Validity

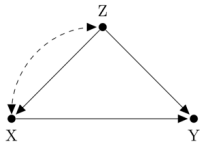
Paul Hünermund

The Data Fusion Process

(1) Query:

Q = Causal effect at target population

(2) Model:



(3) Available Data:

Observational:	P(v)
Experimental:	P(v | do(z))
Selection-biased:	P(v | S = 1) + P(v | do(x), S = 1)
From different populations:	$P^{(source)}$(v | do(x)) + observational studies

Causal Inference Engine:

Three inference rules of *do-calculus*

Solution exists? Yes → Estimable expression of Q

No → Assumptions need to be strengthened (imposing shape restrictions, distributional assumptions, etc.)

Motivating Example: Banerjee et al. (2007)

- Banerjee et al. (2007) study the effect of a randomized remedial education program for third and fourth graders in two Indian cities: Mumbai and Vadodara
  - They find similar effects on math skills, but effect positive impact on language proficiency is much smaller in Mumbai compared to Vadodara
- Banerjee et al. (2007) explain this result by baseline reading skills that were higher in Mumbai, because families are wealthier there and schools are better equipped
- What do we do if we do not have a second experiment to validate our results?
  - Naïve extrapolation (also called *direct transportability*) clearly would have gotten them into trouble

External Validity of A/B Tests

A/B Testing and Covid-19: Data-Driven Decisions in Times of Uncertainty

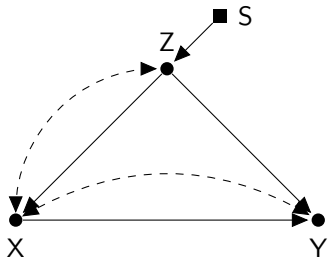
June 26, 2020

I ran an A/B test: will the results be valid once "normal" life resumes?

We can't know. User behavior always changes over time, so it is never certain whether the impact of a feature will persist 2, 3, or 12 months from now. A/B tests help us to optimize features for current usage patterns. But rarely have people's activities, work style, and concerns changed so drastically. Don't assume that usage will stabilize in the same patterns they follow now, or in the patterns you observed pre-Covid. If you run an A/B test now and decide to fully deploy, consider re-running the A/B test comparing the old variant again as circumstances evolve. Re-running the A/B test will allow you to see if the treatment effect has substantially changed. Consider this especially if:

Selection Diagram

- We can incorporate knowledge about structural differences across domains by a selection node (■) in a causal diagram
- Captures the notion that domains differ either in the distribution of background factors $P(U_i)$ or causal mechanisms f_i in the underlying structural causal model
  - Differences across domains can be in arbitrary ways (akin to the nonparametric nature of DAGs)



Selection Diagram (II)

Definition: Selection Diagram (Pearl and Bareinboim, 2011)

Let $\langle M, M^* \rangle$ be a pair of structural causal models relative to domains $\langle \Pi, \Pi^* \rangle$, sharing a causal diagram G. $\langle M, M^* \rangle$ is said to induce a selection diagram D if D is constructed as follows:

(i) Every edge in G is also an edge in D.

(ii) D contains an extra edge $S_i \rightarrow V_i$ whenever there might exist a discrepancy $f_i \neq f_i^*$ or $P(U_i) \neq P^*(U_i)$ between M and M^*.

- Switching across domains Π and Π^* is denoted by conditioning on different values of S
  - Note: $P^*(V) = P(V|S = 1)$
- Compared to selection bias case, now S points into other variables

Transportability Task

- The transportability problem: We have experimental results from a source domain Π, how can we transport them to a target Π^* where we only have passive observations?
  - I.e., we know $P(y|do(x))$ but would like to know $P^*(y|do(x))$
- Note that Π and Π^* share the same causal diagram G. Thus, if the causal effect in Π would be identified from observational data alone (i.e., no experimental data needed) then it would also be identified in Π^* and there would be no need for transportation ("trivial transportability", Pearl and Bareinboim, 2011)
- So transportability theory is (mainly) concerned with transporting experimental results across domains
  - Although observational / statistical transportability can be useful to economize on data collection efforts (Pearl and Bareinboim, 2011)

S-Admissibility

Theorem 2 in Pearl and Bareinboim (2011)

Let D be the selection diagram characterizing two populations, Π and Π^*, and S the set of selection variables in D. The strata-specific causal effect $P^*(y|do(x), z)$ is transportable from Π to Π^* if Z d-separates Y from S in the X-manipulated version of D, that is, Z satisfies $(Y \perp\!\!\!\perp S|Z)_{D_{\overline{X}}}$.

- The set of variables Z is then called *s-admissible*
- Note that $D_{\overline{X}}$ is the result of experimentally manipulating X in the source domain
  - Since S has to be d-separated from Y in $D_{\overline{X}}$, it follows that every S pointing into X does not threaten transportability and can be ignored

S-Admissibility (II)

Corollary 1 in Pearl and Bareinboim (2011)

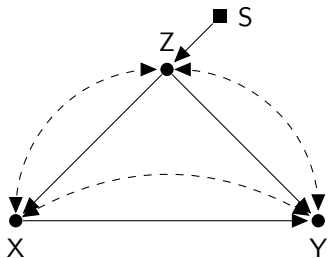
The causal effect $P^*(y|do(x))$ is transportable from Π to Π^* if there exists a set Z of observed pretreatment covariates that is s-admissible. Moreover, the transport formula is given by the weighting

$$P^*(y|do(x)) = \sum_z P(y|do(x), z)P^*(z).$$

- This *transport formula* says that we reweight the z-specific causal effect in the source domain by the distribution of z in the target domains
  - E.g., find experimental results for several income levels in Mumbai and weight by income distribution in Vadodara

S-Admissibility (III)

- Consider this selection diagram, which is the same as before expect for the added edge $Z \dashleftarrow\dashrightarrow Y$
- Is the causal effect transportable in this case?



Transportability – The General Case

- Transportability formula is well-known in economics (Hotz et al., 2005; Andrews and Oster, 2018), but these papers focus on s-admissible *pretreatment* variables. Solutions based on do-calculus are more general

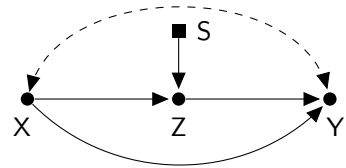
Theorem 1 in Pearl and Bareinboim (2011)

Let D be the selection diagram characterizing two populations, Π and Π^*, and S as set of selection variables in D. The relation $R = P^*(y|do(x))$ is transportable from Π to Π^* if the expression $P(y|do(x), s)$ is reducible, using the rules of do-calculus, to an expression in which S appears only as a conditioning variable in do-free terms.

- Bareinboim and Pearl (2013a) develop a complete algorithm for automatization

Example: Selection Affecting Post-Treatment Variables

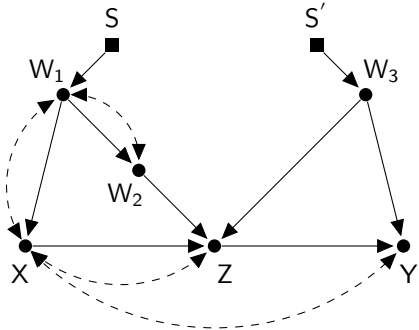
- Gordon et al. (2019) discuss an example in which the effectiveness of a social media advertising campaign, X, is mediated by users' exposure to ads, Z



- Imagine a company that runs advertising campaigns on the desktop version of a social media platform
- Exposure to ads differs across the desktop and mobile version of the platform
- How can experimental results be transported from desktop to mobile?
  - With the help of the transportability algorithm developed by Bareinboim and Pearl (2013a) we find the transport formula

$$P^*(y|do(x)) = \sum_z P(y|do(x), z)P^*(z|x)$$

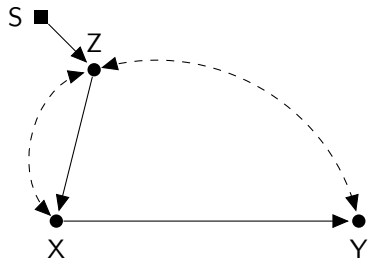
Example: A Complex Selection Diagram



- Here the relevant transport formula is found to be

$$P^*(y|do(x)) = \sum_{z,w_2,w_3} P(y|do(x), z, w_2, w_3) P(z|do(x), w_2, w_3) P^*(w_2, w_3)$$

z-Transportability



- What if we do not have the possibility to run experiments on the treatment X, but we can conduct a surrogate experiment on Z instead (as in the encouragement design from development economics discussed previously)
- This gives rise to the idea of z-transportability

z-Transportability (II)

Theorem 1 in Bareinboim and Pearl (2013b)

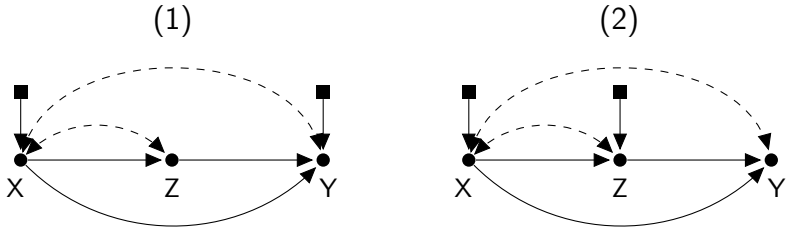
Let D be the selection diagram characterizing two populations, Π and Π^*, and S as set of selection variables in D. The relation $R = P^*(y|do(x))$ is z-transportable from Π to Π^* in D if the expression $P(y|do(x), s)$ is reducible, using the rules of do-calculus, to an expression in which all do-operators apply to subsets of Z, and the S-variables are separated from these do-operators.

- Again, this theorem is only procedural. Bareinboim and Pearl (2013b) develop a complete algorithm for the z-transportability case

Meta-Transportability

- Transportability techniques are particularly valuable we can combine results from several source domains
- This strategy is generally known under the rubric of "meta-analysis"
  - Meta-analyses become increasingly popular in economics (Card et al., 2010; Dehejia et al., 2015)
- The problem with standard meta-analytic tools is that they do not take domian heterogeneity into account but instead aim to "average out" differences across populations
- Bareinboim and Pearl (2013c) extend the transportability idea, which captures domain-specific heterogeneity by the ■-nodes in the causal diagram, to the case with multiple source domains

Meta-Transportability (II)

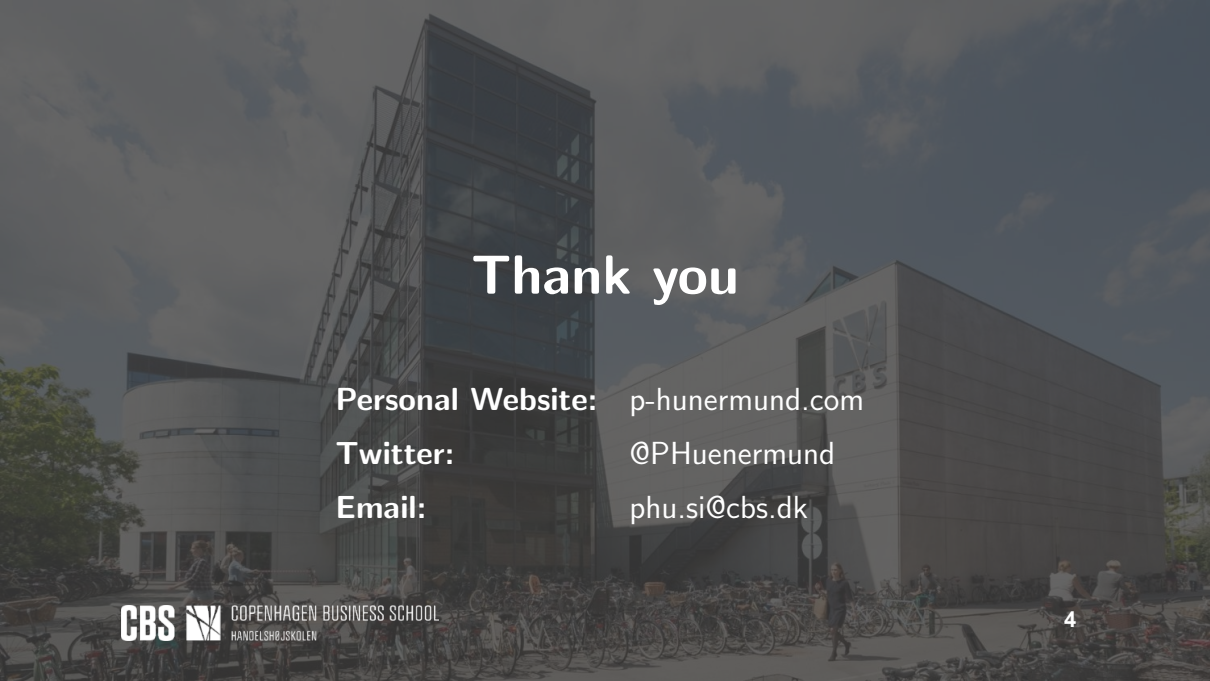


- Both selection diagrams D_1 and D_2 depict how domains π_1 and π_2 differ from the target domain π^*
- Here, the causal effect would not be individually transportable from a single source domain. But it is transportable using combined information from both domains as

$$P^*(y|do(x)) = \sum_z P^{(2)}(y|do(x), do(z))P^{(1)}(z|do(x)).$$

Meta-Transportability (III)

- Bareinboim and Pearl (2013c) develop a complete algorithm for deciding about meta-transportability
- Bareinboim and Pearl (2014) combine the idea of meta-transportability with z-transportability to what they call *mz*-transportability
  - Bareinboim and Pearl (2014) develop a complete algorithm for *mz*-transportability
- These results will hopefully allow to combine more flexibly and make more effective use of results from a whole body of empirical literature



Thank you

Personal Website: p-hunermund.com

Twitter: @PHuenermund

Email: phu.si@cbs.dk

References I

Isaiah Andrews and Emily Oster. Weighting for external validity. NBER Working Paper No. 23826, 2018.

Abhijit V. Banerjee, Shawn Cole, Esther Duflo, and Leigh Linden. Remedying education: Evidence from two randomized experiments in india. *The Quartely Journal of Economics*, 122(3):1235–1264, 2007.

Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134, 2013a.

Elias Bareinboim and Judea Pearl. Causal transportability with limited experiments. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 95–101, 2013b.

Elias Bareinboim and Judea Pearl. Meta-transportability of causal effects: A formal approach. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 31, Scottsdale, AZ, 2013c.

Elias Bareinboim and Judea Pearl. Transportability from multiple environments with limited experiments: Completeness results. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances of Neural Information Processing Systems*, volume 27, pages 280–288, November 2014.

David Card, Jochen Kluve, and Andrea Weber. Active labour market policy evaluations: A meta-analysis. *The Economic Journal*, 120:452–477, 2010.

Rajeev Dehejia, Cristian Pop-Eleches, and Cyrus Samii. From local to global: External validity in a fertility natural experiment. NBER Working Paper No. 21459, 2015. URL `https://www.nber.org/papers/w21459`.

Brett R. Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 38:193–364, 2019. URL `https://www.kellogg.northwestern.edu/faculty/gordon_b/files/fb_comparison.pdf`.

References II

V. Joseph Hotz, Guido W. Imbens, and Julie H. Mortimer. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125:241–270, 2005.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 2011.