Causal Data Science for Business Decision Making

Counterfactuals

Paul Hünermund

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

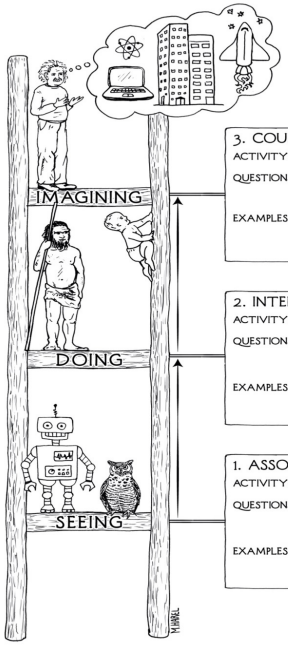
EFMD EQUIS ACCREDITED

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:

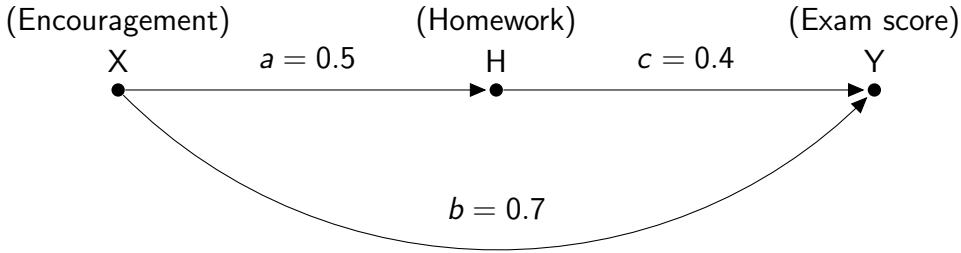
$$Y_x \perp\!\!\!\perp X | Z.$$

- 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



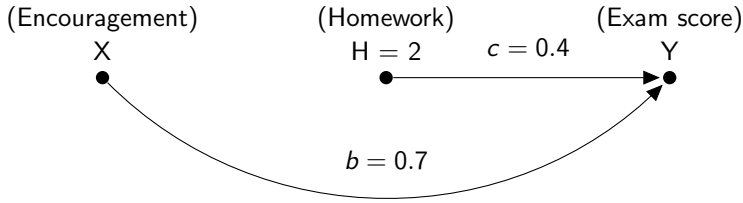
$$X = U_X$$
$$H = a \cdot X + U_H$$
$$Y = b \cdot X + c \cdot H + U_Y$$

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

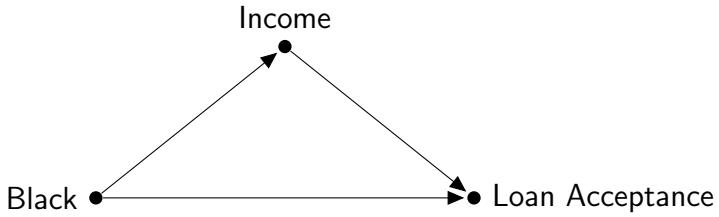
1. **Abduction**: Use evidence $E = e$ to determine the value of U.
2. **Action**: Modify the model M, by removing teh structural equations for the variables in X and replacing them with the appropriate functions $X = x$, to obtain the modified model, M_x.
3. **Prediction**: Use the modified model, M_x,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

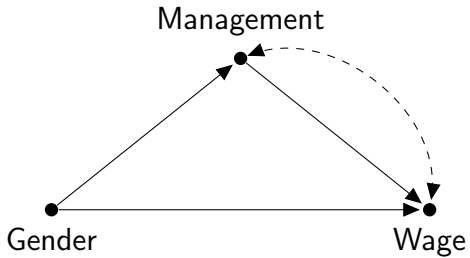
- While the total effect is equal to

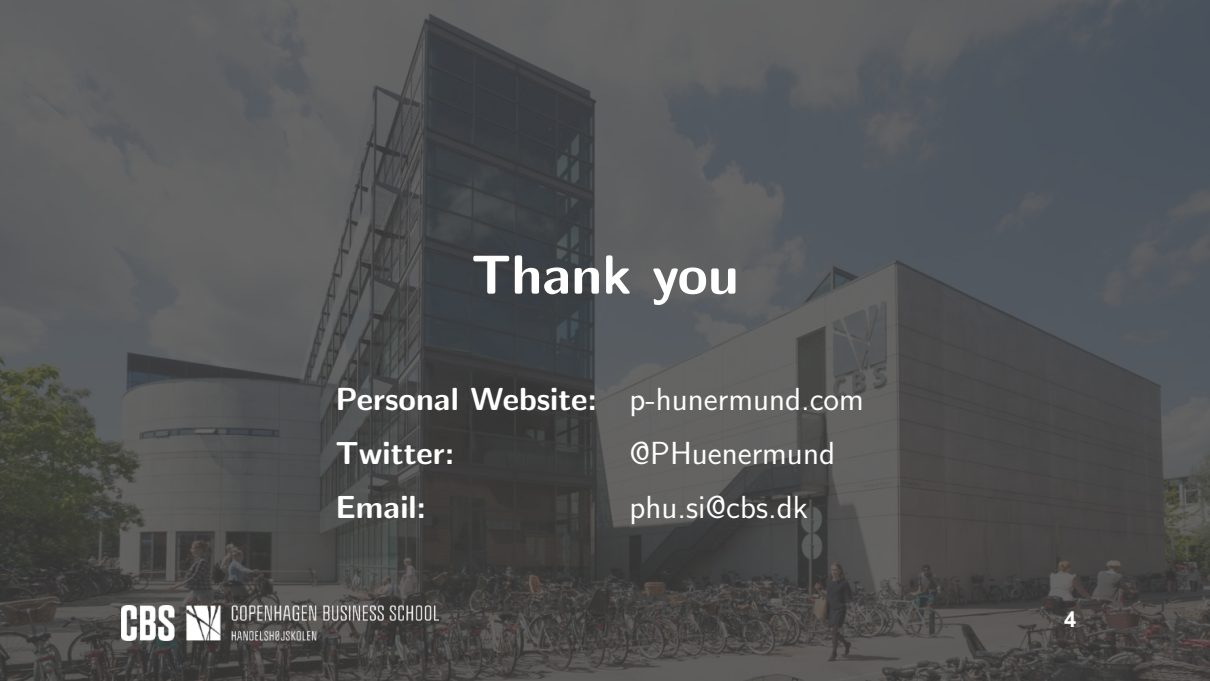
$$\begin{aligned} TE &= E[Y|X=1,M=1] \times P(M=1|X=1) \\ &\quad + E[Y|X=1,M=0] \times P(M=0|X=1) \\ &\quad\quad - \{E[Y|X=0,M=1] \times P(M=1|X=0) \\ &\quad\quad\quad + E[Y|X=0,M=0] \times P(M=0|X=0)\} \\ &= 0.30 \times 0.40 + 0.20 \times 0.60 - \{0.80 \times 0.75 + 0.40 \times 0.25\} \\ &= -0.46 \end{aligned}$$

- $NDE/TE \approx 0.924$ 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.