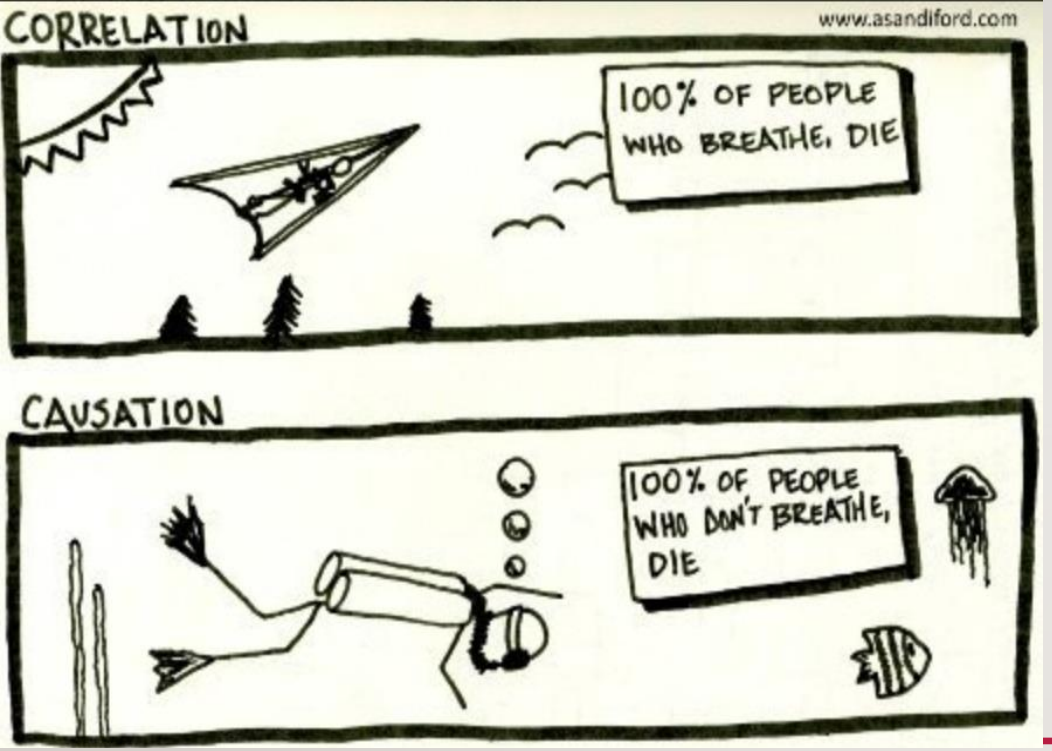


CAUSAL INFERENCE IN NATURAL LANGUAGE PROCESSING FOR ECONOMICS

WEEK 1: CAUSAL INFERENCE WITH TEXT DATA OVERVIEW

SUMMER SEMESTER 22

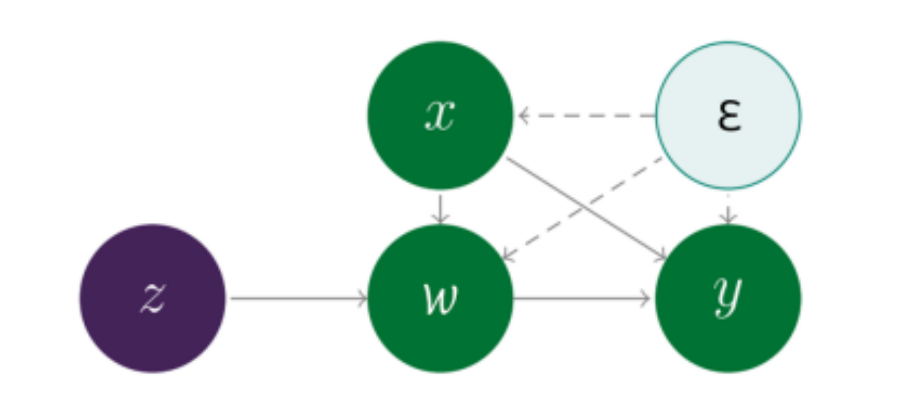
DR.HUYEN NGUYEN

HTTP://HUYENTTNGUYEN.COM

CENTRAL POLICY QUESTION

What happens if someone receives a treatment/policy X is implemented vs. what would happen if they had not received that treatment/X is not implemented?

THE EMPIRICAL PROBLEM

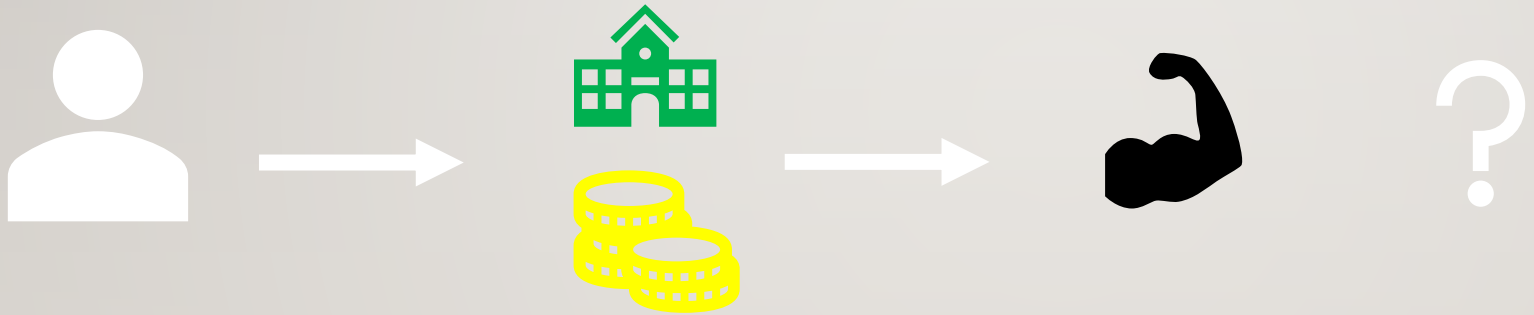


y = outcome
w = (text) treatments
x = observable covariates
ε = confounders
z = instruments

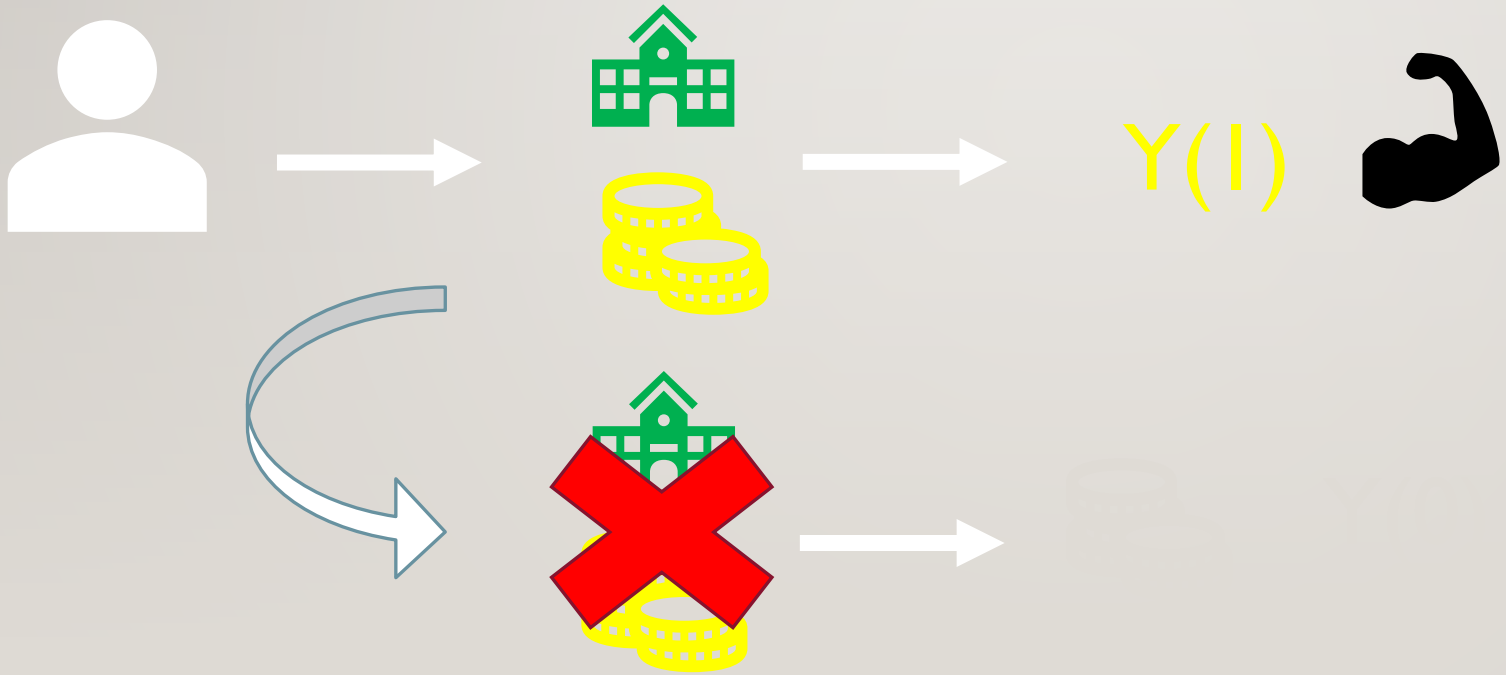
(!) spurious correlation by confounders i.e. variables that are correlated with both the treatment and outcome.

What are the likely (text) confounders ε in Example 1?

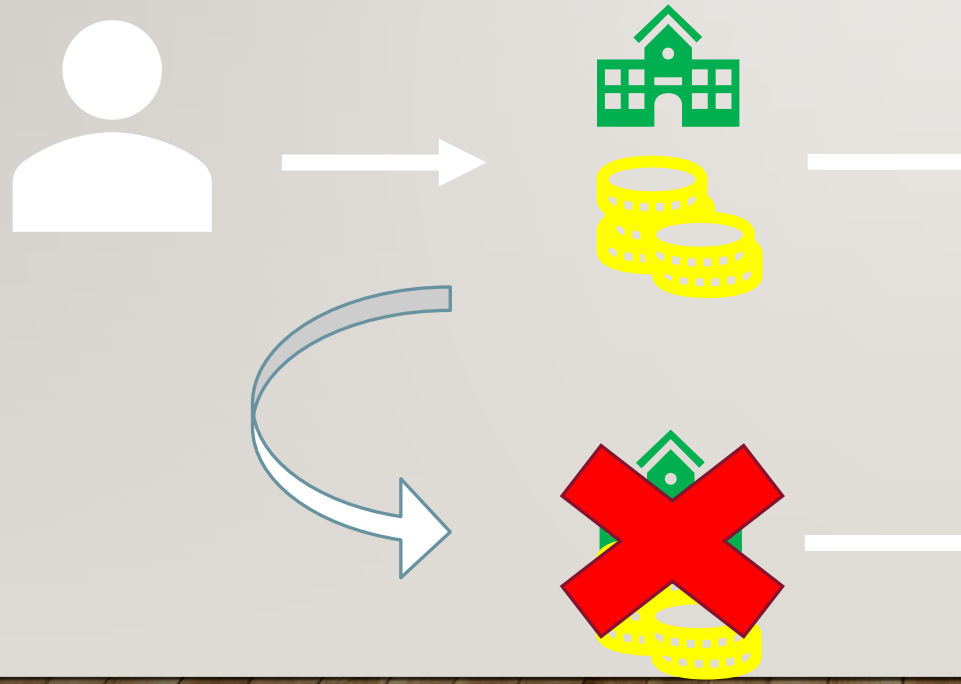
SUPPOSE YOU WANT TO KNOW WHETHER BEING RICH INCREASES LONGEVITY….



That can only be answered if you have identical individual, with one having a lot of money, and the other not….



And the difference between longevity outcomes is the causal impact (i.e. treatment effect) of being rich.



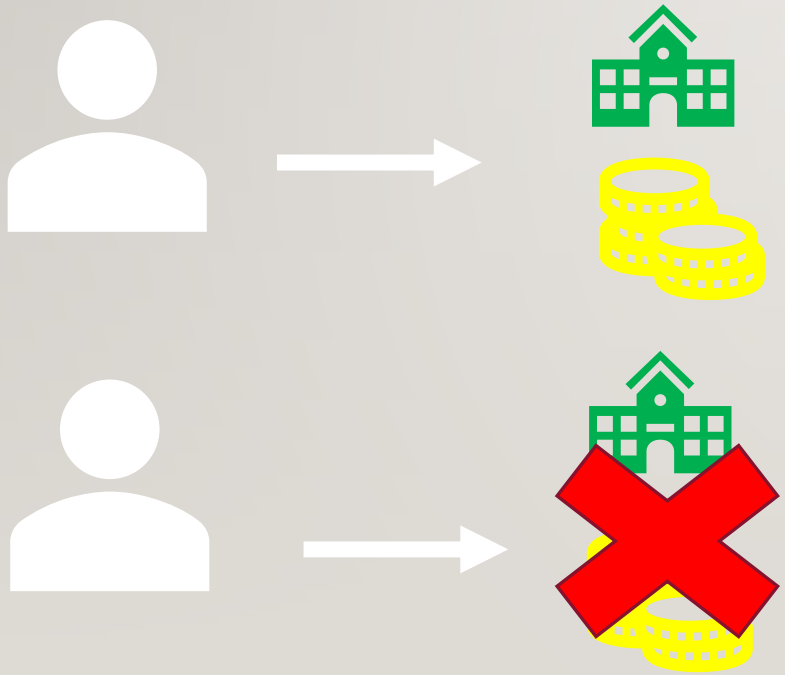
Y(1)

Y(0)

$$\Delta = |Y(1) - Y(0)|$$

BUT EVERYONE IS DIFFERENT.
AND THERE ARE OTHER FACTORS AFFECTING LONGEVITY.

AND IT COULD BE THAT PEOPLE WITH MORE MONEY OPT FOR HEALTHIER LIFESTYLES…



Y(1)

Y(0)

$$\Delta = |Y(1) - Y(0)|$$

FUNDAMENTAL CHALLENGE OF CAUSAL INFERENCE…

- We can NEVER observe both counterfactual outcomes for an individual/policy.
- → Impossible to make both interventions to the treatment and observe the resulting outcomes.
- → This problem is what makes causal inference harder than statistical inference and impossible without assumptions.

AVERAGE TREATMENT EFFECT (ATE)

- Based on the defined counterfactuals, an analyst must specify a quantity of interest that involves the distribution of counterfactuals.

$$ATE = E[Y(1)] - E[Y(0)]$$

- This is the average difference in counterfactual individuals who do not obtain university degrees, if we could intervene to change the obtain-university- degree "treatment", with which the individual is assigned to.

PITFALLS OF NLP

- NLP systems are often black boxes → hard to understand how human interpretable features of the text lead to the observed predictions.

- In this setting, we want to know if some part of the text (e.g., some sequence
- of tokens) causes the output of an NLP method (e.g., classification prediction).

ASSUMPTIONS OF CAUSAL INFERENCE

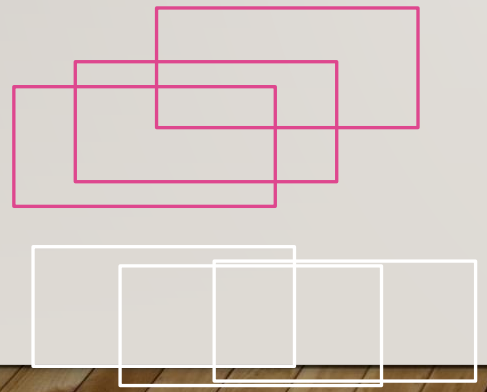
- Ignorability: $\big(Y(1), Y(0)\big) \perp T$
- Positivity: $0 < \Pr(T = 1 | X = x) < 1$
- Consistency: $\forall i \; s.t \; T_i = t, Y_i(t) = Y_i$

CI WITH TEXT AS CONFOUNDERS

- Where? Observational data sets with texts
- Challenge: How to condition on the text such that it blocks confounding with NLP methods.
- Unsupervised dim reduction techniques. E.g. topic models (Roberts et al. 2020), embedding methods, auto-encoders
- → Supervised models by learning low-dimensional variables that can predict
- treatment and outcome well, confounding properties can be found in text.
- . E.g:Veitch, Sridhar and Blei (2020) adapted pre-trained language models and supervised topic models to predict the treatment and outcome.

CI WITH TEXT AS OUTCOMES

- Where? Open-ended responses from randomized experiments
- Challenge:
- Distill high-dimensional data into low-dimensional quality of interest
  - E.g: The effect of intensive English skill course intervention on readability in student essays



Readability (1)

Readability (0)

Readability is a latent variable of language! → Unknown outcome measure at the start of the experiment

CI WITH TEXT AS TREATMENTS

- Where?
- 1) Study of treatment discovery:
- e.g: topics &latent dimensions of candidate biographies that drove voter evaluations Fong & Grimmer 2016
- n-gram influential writing style in marketing materials that increase sales figures Pryzant et al., 2017)
- →producing interpretable features of the text that can be causally linked to outcomes.

CI WITH TEXT AS TREATMENTS (CONT)

- Where?
- 2) estimate the causal effects of specific properties extracted from text:
- e.g: how appealing to civic duty via mails impacts on voter turnout in Michigan 2006 primary election (Gerber et al 2008)
- → factors are latent properties of the text for which we need a measurement model.

CI WITH TEXT AS TREATMENTS (CONT)

- Goal? The causal relationship that (specific aspects of the) text has on downstream decisions, behaviors, and outcomes.

- Challenges?
- 1) Ignorability is typically violated.
- Why? People often choose the text they read for reasons related to outcome of interests. (e.g.: political affiliation, mood, education level, gender)

CI WITH TEXT AS TREATMENTS (CONT)

Challenges?

- 2) Positivity is often violated.
- Why? Suppose we want to know if politeness causally affect e-mail response times. Even if the text contains all confounders (e.g.: topic, tone, writing styles), a polite email is unlikely to contain a certain writing style (e.g., profane).

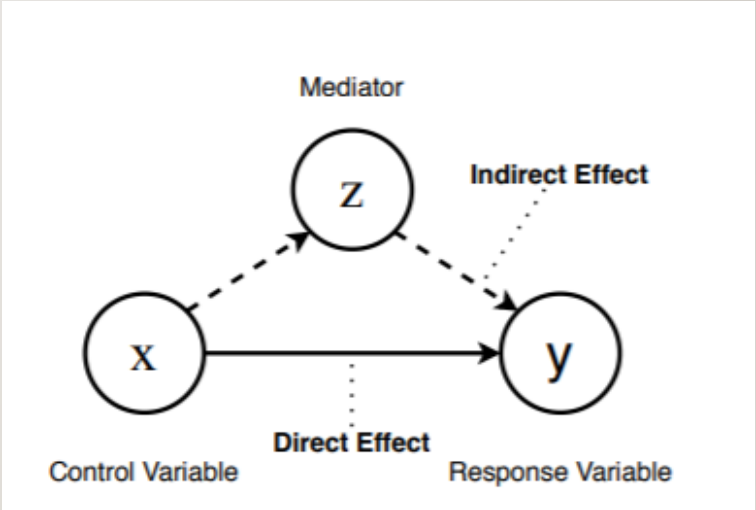
CI WITH TEXT AS MEDIATOR

Causal mediation analysis (Hicks and Tingley 2011 for Stata, Tingley et al., 2014 for R, Statsmodel in Python) introduces a mediator into the causal graph.

The mediator decouples the total effect into a direct and indirect effect (of intermediaries).

E.g: Vaccine causes headaches -> subjects take aspirin (mediators) -> impact on protection level

→ To understand the process in which the treatment causally affects the outcome.



CI WITH TEXT AS MEDIATOR: AN APPLICATION

- How grammatical gender bias in large pre-trained Transformer-based language models. (Vig et al., 2020)

Major concern in word representations, both static word embeddings and contextualized word representations.

(You can check their Github code here: https://github.com/sebastianGehrmann/CausalMediationAnalysis)

- Given a prompt u : The nurse said that…, a language model is asked to generate a continuation:
- Stereoptypical candidate: SHE . Anti-stereotypical candidate: HE

- → A biased model assigns a higher likelihood to SHE than HE

$$p_\theta(SHE|u) > p_\theta(HE|u)$$

→ A measure of grammatical gender bias in the model:

$$y(u) = \frac{p_\theta(HE|u)}{p_\theta(SHE|u)}$$

$y(u)$ >1 ?

$y(u)$ <1 ?

$y(u)$ =1 ?