Text Analysis for Social Sciences in Python

Week 12: Causal Inference – Introduction & Application Overview

WINTER SEMESTER 21/22

HUYEN NGUYEN

HTTP://HUYENTTNGUYEN.COM

Course Feedback survey ☺

Kindly take 10’ to fill them in – your inputs are hugely appreciated!
https://evasys-online.uni-hamburg.de/evasys/online.php?pswd=1S46L

Final presentation date

Monday 31.01: 14.00 – 17.00 (20-25’ presentation/group + 10’ Q&A)

(i.e. code & slides submission by Thursday 27.01 23.59 to Open Olat folder)

https://www.openolat.uni-hamburg.de/url/RepositoryEntry/211256038/CourseNode/102299299236229

Time slot allocation (Google sheet link on your Slack channel #group_matching):

https://docs.google.com/spreadsheets/d/1E-IFBNDNk7-5GxALoYRfK-P8GO5aYJ8v9j0uEjUsZ5M/edit#gid=548845839

Today’s agenda

- Causal Relationships
- Identification vs. Estimation vs. Inference
  - Assumptions for Causal Inference
- The Empirical Problem
  - Econometrics ft. Machine Learning
  - Causal strategies with Text Applications
  - Text in the Causal Pipeline

Central Policy Question

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?

Causal relationships

- Labor economics: Do better-looking people earn more?

Causal relationships

- Labor economics: Do better-looking people earn more?
- Political economy: Is democracy good for economic growth?

Causal relationships

- Labor economics: Do better-looking people earn more?
- Political economy: Is democracy good for economic growth?
- Development: Did the slave trade lead to underdevelopment in Africa?

Causal relationships

- Labor economics: Do better-looking people earn more?
- Political economy: Is democracy good for economic growth?
- Development: Did the slave trade lead to underdevelopment in Africa?
- Health: Does deworming improve child school attendance rates?

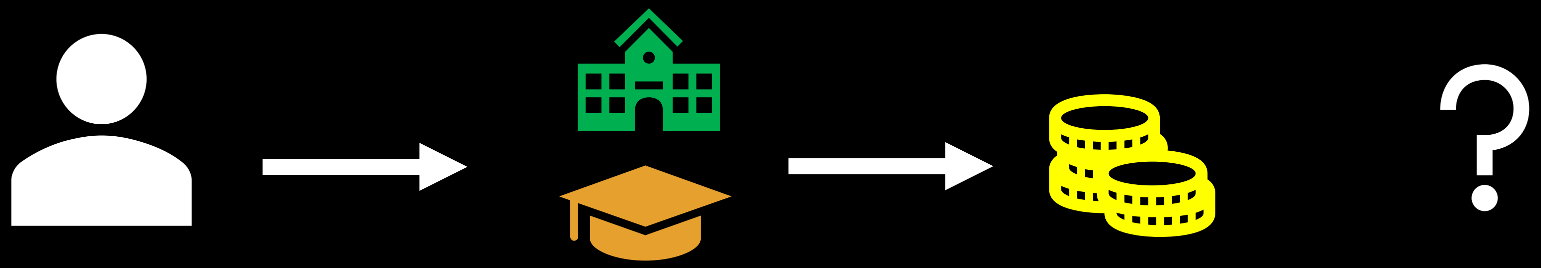
Causal relationships

- Labor economics: Do better-looking people earn more?
- Political economy: Is democracy good for economic growth?
- Development: Did the slave trade lead to underdevelopment in Africa?
- Health: Does deworming improve child school attendance rates?
- Industrial Organization: Does environmental regulation dampen production?

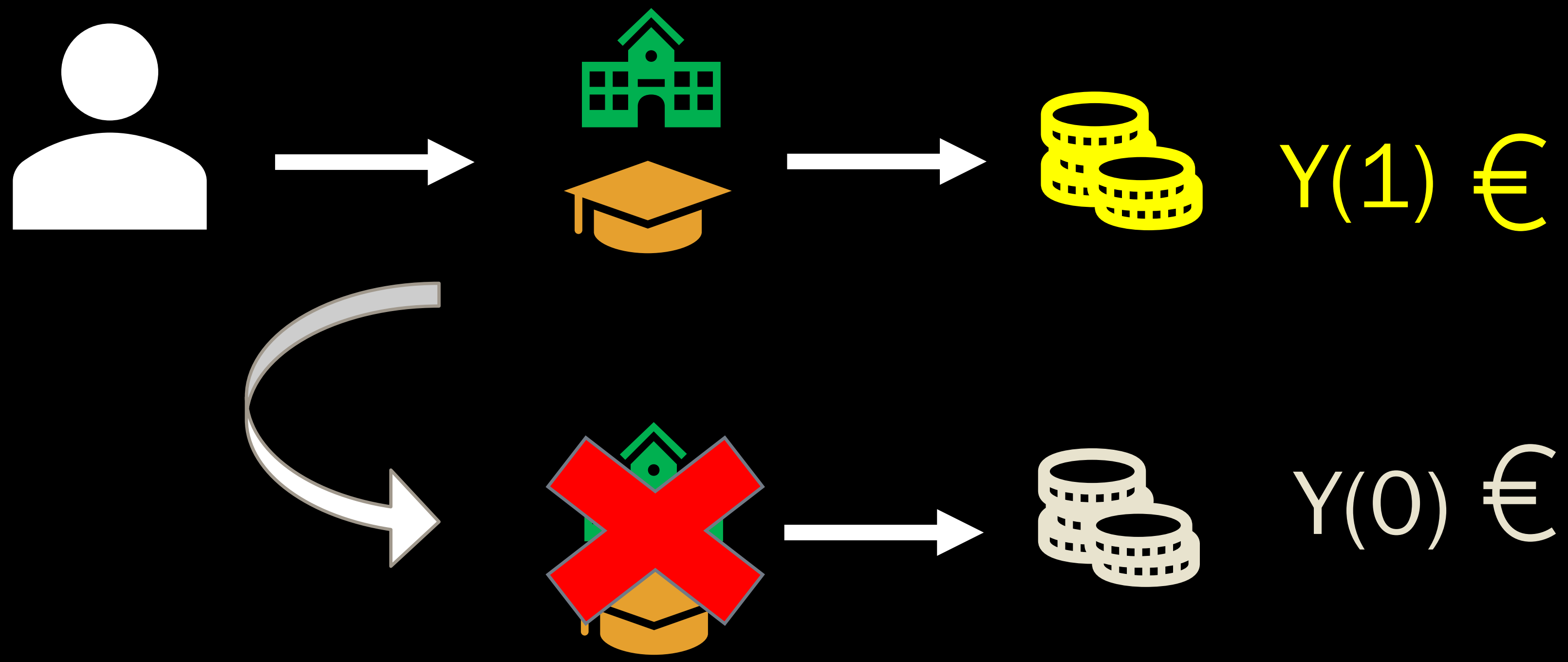
Causal relationships

- Labor economics: Do better-looking people earn more?
- Political economy: Is democracy good for economic growth?
- Development: Did the slave trade lead to underdevelopment in Africa?
- Health: Does deworming improve child school attendance rates?
- Industrial Organization: Does environmental regulation dampen production?
- Finance: Does the FED release of FOMC minutes affect the stock markets?

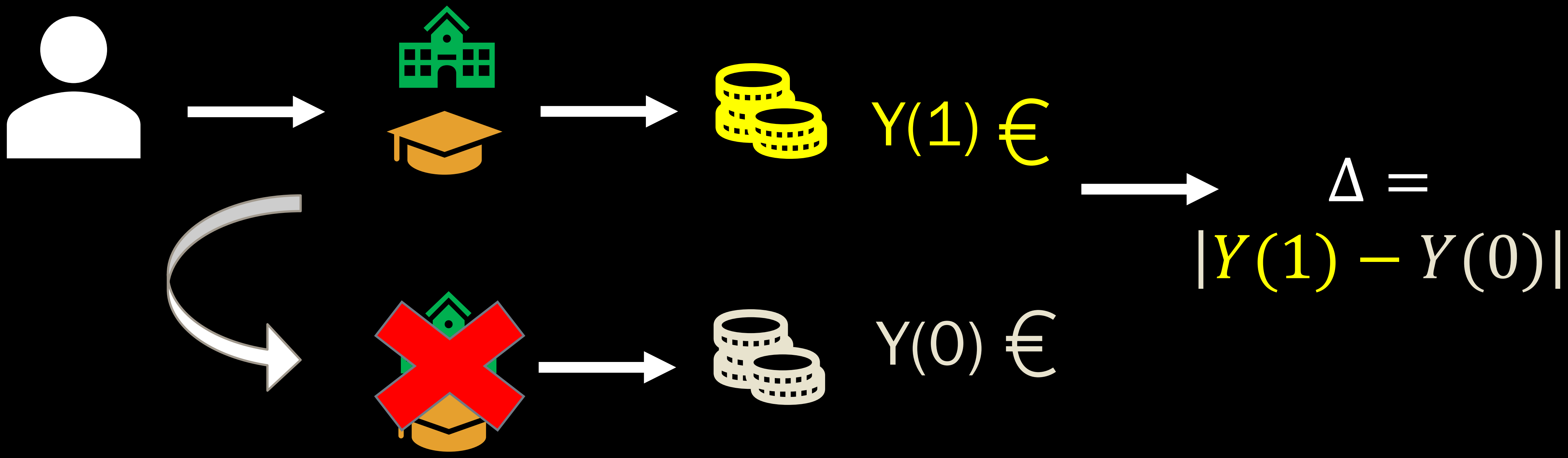
Suppose you want to know whether obtaining university degrees give higher life expected income….



That can only be answered if you have identical individual, with one go to university, and the other not….



And the difference between the outcomes is the causal impact (i.e. treatment effect) of obtaining university degrees.

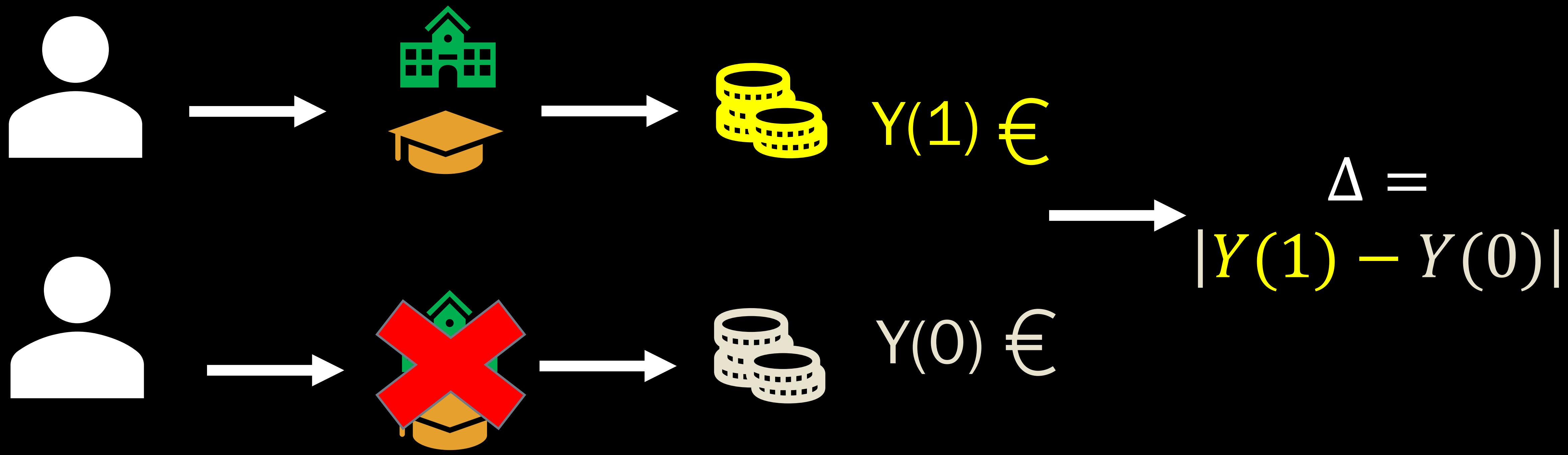


Y(1) €

Y(0) €

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

BUT everyone is different.
AND there are other factors affecting life expected income.
AND it could be that people with higher life expected income opt for university degrees…



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.

Ideal experiment

If you had any resources you need at your disposal, what would you do to answer your question of interest?

What would you need to do to get random assignment of treatment?

Ideal experiment

1)Natural experiment: Someone else/An institution/Nature does the experiment for you.

- E.g: The impact of WWII and Nazi “purge” of scientists on knowledge production (Iaria, Schwarz and Waldinger QJE 2018)

Ideal experiment

1)Natural experiment: Someone else/An institution/Nature does the experiment for you.

- E.g: The impact of WWII and Nazi “purge” of scientists on knowledge production (Iaria, Schwarz and Waldinger QJE 2018)

2)Clever identification strategies (instrumental variables, regression discontinuity design, difference-in-difference, matching/high dimension controls)

- E.g: The effect of liberalization on economic performance (Giavani & Tabellini JME 2005)

For detailed, easy-to-understand readings on these techniques, I recommend the book of Angrist, Joshua and Jörn-Steffen Pischke. 2009. Mostly Harmless Econometrics.

Identification vs. Estimation vs. Inference

Given the underlying data generating process (DGP), if I could take as large a sample as I want, would I eventually learn the causal effect of interest? = IDENTIFICATION question

→ The answer to question of causal interest is identified IF there exists an ideal experiment that can answer it.

Causal questions for which there is no ideal experiment that can answer it are Fundamentally Unidentified Questions (in short, FUQ'ed questions >_<)

Identification vs. Estimation vs. Inference

In reality, we only have access to a finite number of observations (sampling is costly).

→ What method will we use to estimate the causal effect of interest (OLS,

Maximum likelihood, etc.)?

→ How do we summarize our confidence in the estimate (standard errors etc.)?

= ESTIMATION questions

(!) There is NO point in considering questions of estimation if your answers to identification questions are inadequate (e.g. if your question is FUQ'ed).

Assumptions for Causal Inference (1)

In short: How to make sure we can identify the ATE?

1) Ignorability: requires that treatment T assignment is independent of the realized counterfactual outcomes.

$$\big(Y(1), Y(0)\big) \perp T$$

(!) the most important and difficult to justify assumption.

Assumptions for Causal Inference (1)

Randomizing treatment assignment is one way to satisfy ignorability. Why?

Assumptions for Causal Inference (1)

Randomizing treatment assignment is one way to satisfy ignorability. Why?

In expectation there are no systematic pretreatment differences between treated and untreated samples.

For example, an online forum could run an A/B test, a randomized trial where some readers are randomly assigned posts labeled with a woman icon and some are randomly assigned posts with other icons.

Assumptions for Causal Inference (1)

Randomized assignment may NOT always be feasible.

In this case, we may need to rely on conditional ignorability:

$$\left(\left(Y(1), Y(0)\right) \perp\!\!\!\perp T\right) | X$$

where X is a set of variables such that treatment T assignment and the potential outcomes is unconfounded within levels of X.

Assumptions for Causal Inference (1)

Randomized assignment may NOT always be feasible.

In this case, we may need to rely on conditional ignorability:

$$\left(\big(Y(1), Y(0)\big) \perp\!\!\!\perp T\right) | X$$

where X is a set of variables such that treatment T assignment and the potential outcomes is unconfounded within levels of X.

Caveat: Conditional ignorability requires that there are no unobserved confounders. → strong assumption that analyst must carefully assess.

Assumptions for Causal Inference (2)

2) Positivity: requires that the probability of receiving the treatment is bounded between 0 and 1.

$$0 < \Pr(T = 1|X = x) < 1$$

Under conditional ignorability, this must hold for all x such that

$$\Pr(X = x) \neq 0$$

Assumptions for Causal Inference (2)

Intuitively, this requires that in order to learn about causal effects, it needs to be possible to observe units under <u>both</u> types of treatment statuses.

It also implies that the treatment status cannot be perfectly predicted given the variables X.

Randomized treatment assignment also ensures positivity is satisfied.

Assumptions for Causal Inference (3)

3) Consistency: requires that the observed outcome at a given treatment status for a given unit, is the same as we would observe if that unit was assigned to the treatment.

$$\forall i \; s.t \; T_i = t, Y_i(t) = Y_i$$

Assumptions for Causal Inference (3)

3) Consistency: requires that the observed outcome at a given treatment status for a given unit, is the same as we would observe if that unit was assigned to the treatment.

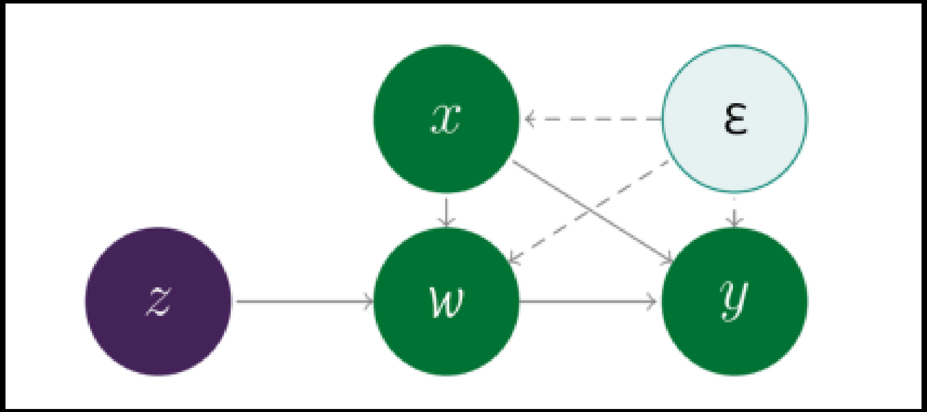
$$\forall i \ s.t \ T_i = t, Y_i(t) = Y_i$$

Consistency relates counterfactual outcomes to observed ones and captures two main assertions:

→First, no interference: the outcome for a sample i is affected only by sample i's treatment status, and NOT the treatment status of other samples.

→Second, there is ONLY ONE version of treatment.

The Empirical Problem



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

Econometrics ft. Machine Learning

$$f(w,\theta) = \mathbb{E}\{y|w\}$$

We want to learn the conditional expectation for y, where θ represents the true parameter vector.

If we assume linearity and run OLS, the estimates for $\hat{\theta}$ are biased because of the confounder ε.

Similarly, we could take an ML approach and learn a nonlinear approximation

$\hat{f}(w,x,\theta)$to predict y in the hold-out data.

- If ↑ text w_i for new individual i = ↑ prediction power about the associated y_i .
- BUT the ML estimates $\hat{\theta}$ are NOT causal: it cannot be used to make a counterfactual prediction about the effect on y of exogenously altering the text features w.

Lab/field experiments with texts

Gold standard to obtain causal estimates. For instance:

- Subjects fill out open-ended survey responses before and after the experiment
- Subjects can talk to each other in a chatroom
- Subjects view randomly assigned text treatments

Difference in difference with text

Estimate changes in an outcome due to changes in text.

(!) a lot of strong assumptions for this, probably infeasible in reality.

-> More realistic approach? measure changes in a text-based metric due to a treatment

- Include fixed effects and trends for the individuals.
- Try to illustrate effect with event-study graph
  - Ash, Morelli, and Van Weelden (2017): effect of senate elections on divisiveness
  - Ash (2016): effect of political control on tax code

Regression Discontinuity with text

Local impact of a threshold treatment on a text-based metric.

- E.g: Electoral RD effect on type of speech used by Congressmen

Local impact of a threshold text treatment.

- E.g: Electoral RD effect on economy of ballot referenda

Text in the causal pipeline

Text as outcome: assess a text-based response

- E.g. text-predicted ideology changes according to senate election schedule (Ash, Morelli, and Van Weeldden 2017)

Text as treatment: assess the effect of a text on outcomes

- E.g: how do different information provided to students by alumni affect career choices (Gallen & Wasserman 2021)

Text as confounder: condition on text

- E.g: adjusting for confoundings with text matching (Roberts, Stewart and Nielsen 2017)

Text as source of heterogeneity: estimate conditional average treatment effects as function of text

- E.g: Wager and Athey 2017

Practice time ☺

Open your Stata and download the dataset bm.dta.

https://github.com/httn21uhh/Text-Analysis-for-Social-Sciences-in-Python

→ W12_causal_inference.pdf+ bm.dta

- Download and put them in the same directory/environment. Open the task.

This week…

- Causal analysis – simple exercises and applications
- Unmute yourself to ask questions at any point, including when you think things go too fast/slow for you.