Causal Inference - Mini Course

session 1 — intro: identification, estimation, and inference

Simon Heß

August 23

Intro

About this course

One question: When does correlation imply causation?

Two books: I do not strictly follow a textbook, but useful references are

- *Causal Inference: The Mixtape* (Cunningham 2021)
- *Mostly Harmless Econometrics* (Angrist and Pischke 2008)

Three sessions:

1. selection bias and how randomization solves it
   - why care about causality and why we **cannot** simply use descriptive statistics
   - randomized control trials (RCTs) aka experiments
2. regression discontinuity (RD)
   - sometimes policies are naturally akin to randomized experiments
3. difference-in-differences (DD)
   - a fallback if other methods are unavailable?

Problem sets with data for self-study will be shared after classes

Learning goals

1. Understanding of the concept of causality
2. Basic knowledgie of 3 canonical research designs (RCT, RD, DD)
3. Ability to apply these designs to own work
4. Ability to critically assess other work using these strategies

Your background

I assume familiarity with linear regression and conditional expectations

- the material is not deeply technical, but it will help

Who are you? (by show of hands)

1. Who has taken an econometrics class? ("Introductory econometrics")
2. Who knows what *selection bias* is?
3. Who has worked with data (in R, Stata, python, . . .)?
4. Who has heard of randomized experiments, A-B-testing, or medical trials?
5. Who has heard of regression discontinuity, or difference-in-differences?

The problem

Terminology: Treatment effects, Counterfactual

Ex.: Job training program

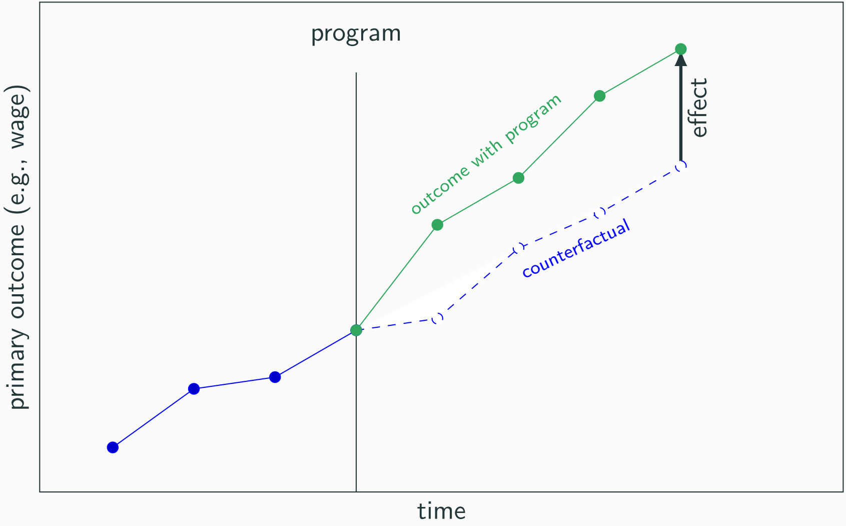
A training for low-wage workers to improve their skills.

We want to know the treatment effect of the program on wages later wages.

- What is a treatment effect?
  - The difference in outcomes between what happened and what would have happened without the program.
  - Problem: We never observe both states $\Rightarrow$ need to know the "**counterfactual**"
- What is a counterfactual?
  - What would have happened if the program had not been implemented
  - Never directly observed, has to be estimated

All causal inference is about finding credible answers to **"what if?"-questions**

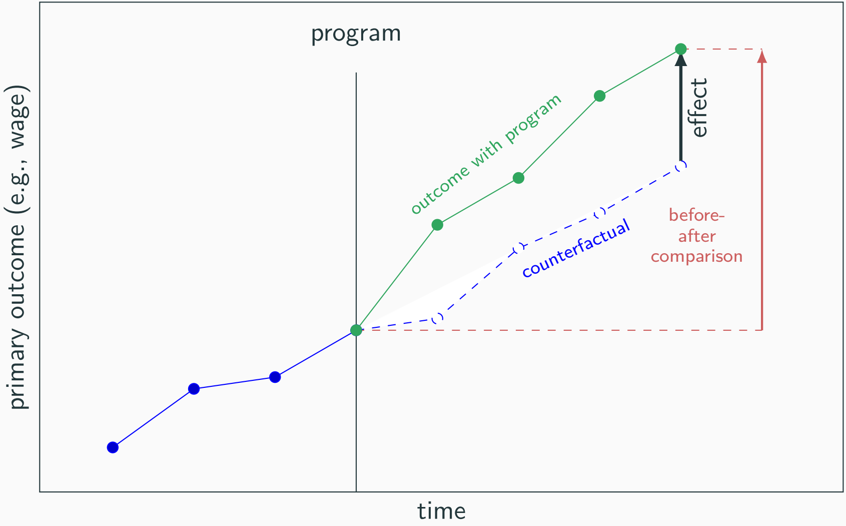
Measuring effects



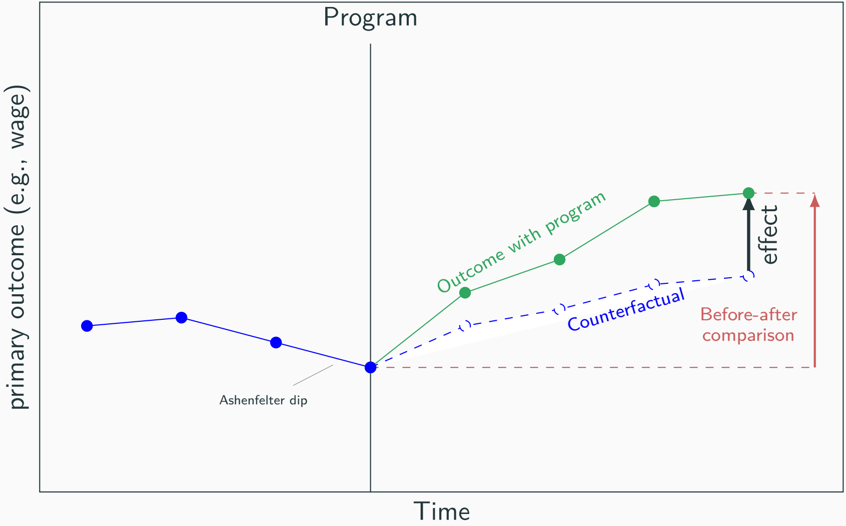
Estimating the counterfactual

1. Do participants prior to the program make a good counterfactual?
   - **Generally no!**
2. Do people who choose not to participate (are not assigned to participation) make a good counterfactual?
   - **Generally no!**

Measuring effects: Why not compare before and after? Trends



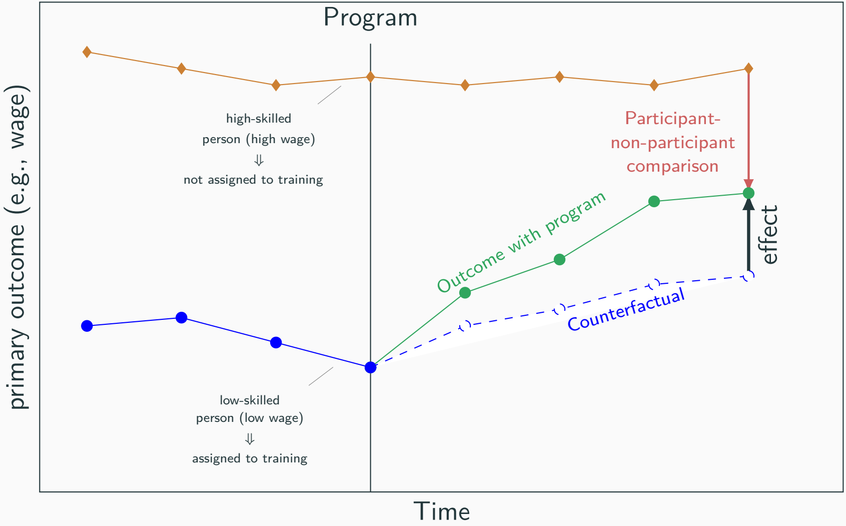
Why not compare before and after? Ashenfelter Dip



Estimating the counterfactual

1. Do participants prior to the program make a good counterfactual?
    - **Generally no!**
2. Do people who choose not to participate (are not assigned to participation) make a good counterfactual?
    - **Generally no!**

Why not compare participants and non-participants?



Recap

- to estimate effects, need to estimate the counterfactual ("what if"-scenario)
- observable outcomes (pre-intervention baseline outcomes, or non-participants) provide poor counterfactuals

Why care about causality

many interesting econometric questions are causal questions

Q: do people send their kids to school if they have a more stable income?

"kids of parents in formal employment have on average x more years of education"

Q: does microfinance reduce poverty?

"people receiving microfinance are $x\%$ less likely to be poor"

Q: does more policing reduce crime?

"states with 1% more police have $x\%$ less/more burglaries"

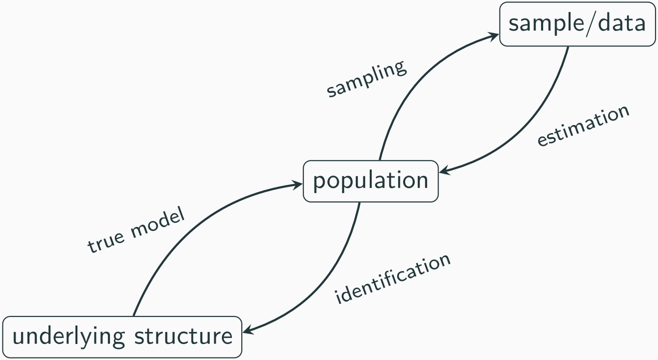
Q: do agricultural development projects increase deforestation?

"villages with more development interventions have less forest"

questions are about the underlying structure of the observable world

- answers are about observable distributions (correlations, etc)
- do they answer our questions about underlying structure? maybe (not).

Identification, estimation, inference



observed

observable

unobservable

Identification

- learning about underlying structures (causal relationships)
- from a population distribution
- identification is not directly related to data
    - this is a question of what is **knowable**.

Ex.: identification

if we have a randomized experiment, the causal effect is identified by a difference in population means of the treated and untreated population (more on this later.)

Estimation (and inference)

- learning about a population distribution
- from finite sample observations

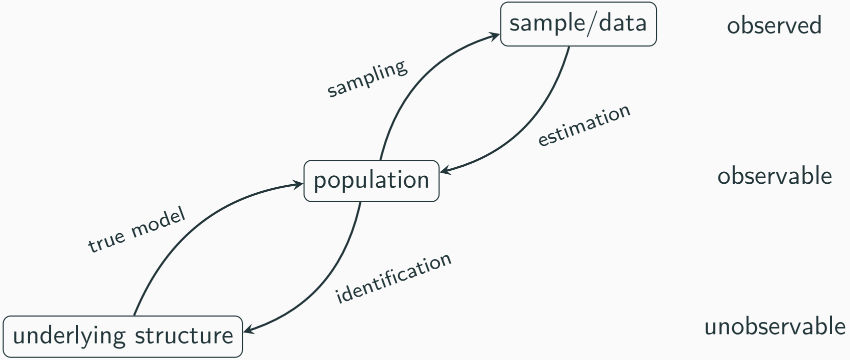
Ex.: estimation

To estimate a difference in population means for two groups, we can use the differences in sample means and perform inference using a t-test.

Note:

- "inference" sometimes refers to the last step (e.g., conducting a test) and sometimes to the whole process (as in the title of this class)

Identification, estimation, inference



Recap

- examples of causal questions
- identification: linking population characteristics to causal mechanisms
- estimation: learning about population characteristics from a sample

Next:

- identification
  - potential outcome framework
  - selection bias
  - a tour through **identification strategies**
- some words on estimation

Potential outcomes

Counterfactuals

Causal analysis tries to answer 'what if'-questions

Ex.: Job training

Cal took a job training and later on earned US$40k.

- Did the training improve Cal's wage?
- What would Cal earn in the *counterfactual* world where Cal did not take the training?

- Central problem: **We never observe the counterfactual**
- Let's formalize the problem to see if we can solve it.
  - I.e., if we can learn something about the counterfactual.

The potential outcome framework ...

... conceptualizes the idea of counterfactuals

- binary treatment: $D_i = 1$ if treated, $D_i = 0$ if not
- every unit i has two potential outcomes: y_i^0 and y_i^1
- for each i, only one of the two outcomes is observed

$$y_i = \begin{cases} y_i^0 & \text{if } D_i = 0 \\ y_i^1 & \text{if } D_i = 1 \end{cases} \qquad = D_i y_i^1 + (1 - D_i) y_i^0.$$

- the individual-level treatment effect we are want to know:

$$\Delta_i = y_i^1 - y_i^0.$$

- this is never observable. but summary measures of its distribution can be identified, e.g.:
  - the average treatment effect (ATE): $\mathbb{E}[\Delta_i]$

Average treatment effect (ATE)

$$\text{ATE} = \mathbb{E}[\Delta_i] = \mathbb{E}[y_i^1 - y_i^0]$$

or the conditional version:

$$= \mathbb{E}[y_i^1 - y_i^0 | x_i].$$

where x_i is a vector of observed characteristics (e.g., age, gender, etc.).

ATE measures average effects of treatment on a unit in the population

- average effect of job training on wages among all unemployed
- average effect of smoking on the probability of developing cancer

A note on heterogeneity

- ATE looks at *average effects*
- treatment effects can be heterogeneous:
  - a development intervention may help some but leave others worse off
  - a pill could heal some but be detrimental to others
  - a job training could affect only junior workers

$$\mathbb{E}[\Delta_i | x_i = x'] \neq \mathbb{E}[\Delta_i | x_i = x'']$$

- ATE might be positive even if the majority has a negative Δ_i
- studying heterogeneous treatment effects is a large field of research
  - looking at heterogeneity may help understand **how** a treatment works (mechanism)

Selection bias (1)

Typically, we cannot identify the ATE from differences in observable means

If we compare means in a treated and an untreated group, we estimate:

$$
\begin{aligned}
&\mathbb{E}[y_i|D_i=1]-\mathbb{E}[y_i|D_i=0]\\
=&\,\mathbb{E}[y_i^1|D_i=1]-\mathbb{E}[y_i^0|D_i=0]\\
=&\,\mathbb{E}[y_i^1|D_i=1]-\mathbb{E}[y_i^0|D_i=0]+\mathbb{E}[y_i^0|D_i=1]-\mathbb{E}[y_i^0|D_i=1]\\
=&\,\underbrace{\mathbb{E}[y_i^1-y_i^0|D_i=1]}_{\text{ATE among all with } D_i=1}+\underbrace{\mathbb{E}[y_i^0|D_i=1]-\mathbb{E}[y_i^0|D_i=0])}_{\text{selection bias}}
\end{aligned}
$$

- first term is the treatment effect
- second term is a confounding selection bias
  - zero if potential outcomes are independent from treatment ($\mathbb{E}[y_i^0|D_i=1]=\mathbb{E}[y_i^0|D_i=0]$)

Selection bias (2)

Ex.: selection bias in job training

People who enroll in job training differ in terms of unobservable characteristics (motivation, mindset, etc.) from people who do not.

- they also might differ in their expected income without training
- comparing participants to non-participants does not give the causal effect

Ex.: selection bias in smoking

Smokers may differ in terms of unobservable characteristics such as (lifestyle choices, risk behavior) from non-smokers.

- they also might differ in their cancer risk without smoking
- comparing smokers to non-smokers does not give the effect of smoking

Iff D_i is assigned independently from potential outcomes (e.g., by coin toss) **then** comparing means between groups identifies the causal effect

Recap

- Potential outcomes conceptualize the idea of counterfactuals.
- An ATE is a summary of "underlying structure" that is useful in identification arguments and to describe causal effects.
- Selection bias implies that simple comparisons between treated and untreated observations do not identify the ATE.

Next:

- Identification strategies that overcome selection bias.

RCTs

Independence of treatment and potential outcomes: RCTs

- no selection bias if D_i and potential outcomes are independent
- easiest way to ensure independence is to flip a coin for each person to decide if they get treatment:
  - a randomized control trial (RCT)
  - selection bias is 0 and ATE is **identified** by the difference in expected outcomes
  - difference in expected outcomes can be **estimated** from differences in sample means, or a regression

$$y_i = \alpha + \beta D_i + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d}{\sim} \cdot(0, \sigma^2)$$

Ex.: An RCT on microfinance

100 village; 50 are randomly selected to open a microfinance bank
5 years later, we measure incomes in types of villages and compare them

- long history in medical sciences
- shorter but successful track record in social sciences (Econ Nobel Prize 2019 for Banerjee, Duflo, Kremer)

RCTs

- in RCTs we know that D_i is random
- then identification of the ATE is straightforward
- randomization is not always feasible
- other popular approaches *resemble* RCTs under specific **identifying assumptions**

Rest of this class:

- outlook on other identification strategies

Causal identification strategies as generalizations of RCTs

Generalizing from D_i random ...

RCTs

- in RCTs we know that D_i is random
- then identification of the ATE is straightforward
- randomization is not always feasible
- other popular approaches *resemble* RCTs under specific **identifying assumptions**

Rest of this class:

- outlook on other identification strategies

Causal identification strategies as generalizations of RCTs

Generalizing from D_i random ...

RCTs

- in RCTs we know that D_i is random
- then identification of the ATE is straightforward
- randomization is not always feasible
- other popular approaches *resemble* RCTs under specific **identifying assumptions**

Rest of this class:

- outlook on other identification strategies

Causal identification strategies as generalizations of RCTs

Generalizing from D_i random ...

(RD) ... to D_i random conditional on $x_i \in (\bar{x} - c, \bar{x} + c)$ for $c \to 0$.

RCTs

- in RCTs we know that D_i is random
- then identification of the ATE is straightforward
- randomization is not always feasible
- other popular approaches *resemble* RCTs under specific **identifying assumptions**

Rest of this class:

- outlook on other identification strategies

Causal identification strategies as generalizations of RCTs

Generalizing from D_i random . . .

(RD) ... to D_i random conditional on $x_i \in (\bar{x} - c, \bar{x} + c)$ for $c \to 0$.

(DD) ... to D_i random relative to $y^0_{i,t} - y^0_{i,t-1}$.

RD

Regression discontinuity designs – Two identifying assumptions

1. Discontinuous assignment of treatment: Treatment is determined based on whether an observable continuous running variable x exceeds some threshold $\bar{x}$.

$$D_i = \begin{cases} 1 \text{ if } x \geq \bar{x} \\ 0 \text{ if } x < \bar{x} \end{cases}$$

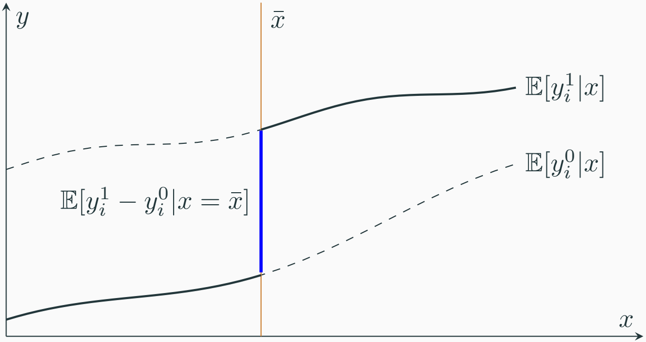
   - e.g., students scoring $> 95\%$ get a stipend ...

2. Continuous mean of potential outcomes:

   - $\mathbb{E}[y_i^1|x]$ and $\mathbb{E}[y_i^0|x]$ are continuous in x.

Then, the conditional ATE at $x = \bar{x}$, $\mathbb{E}[y^1 - y^0|x = \bar{x}]$, can be identified.

Regression discontinuity designs – graphical illustration



- vertical distance between the lines is (unobservable) conditional ATE, $\mathbb{E}[y^1 - y^0|x]$
- observations with $x < \bar{x}$ are not treated; others are
- at $x = \bar{x}$, ATE becomes observable.

DD

Difference-in-differences – Setup

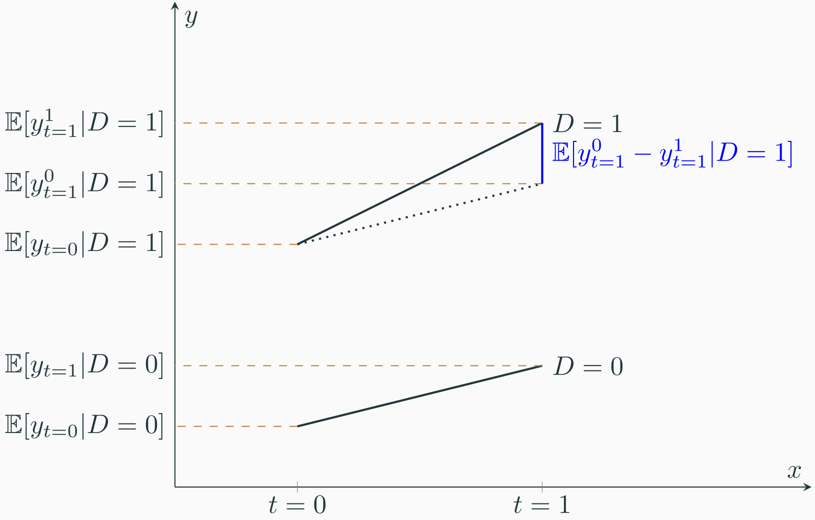
- 2 groups $\times$ 2 periods:
  - $t = 0$: pre-treatment period, $t = 1$: post-treatment period
  - $D = 1$: treated units, $D = 0$: control units
- treatment occurs after $t = 0$, so:
  - in period 0, no one is treated
  - in period 1, one group will be treated the other not
  - from those who are never treated, we see what the trend without treatment is
  - so we can extrapolate from $t = 0$ for the untreated, to know what they would have been like without treatment (counterfactual)
- If we assume common trends:

$$\mathbb{E}[y^0_{t=1} - y^0_{t=0} | D = 1] = \mathbb{E}[y^0_{t=1} - y^0_{t=0} | D = 0]$$

  (i.e, without treatment, treated and untreated exhibit same trend)
- then, the treatment effect is identified by the difference of two differences:

$$\underbrace{(\mathbb{E}[y_{t=1}|D = 1] - \mathbb{E}[y_{t=1}|D = 0])}_{\text{post difference}} - \underbrace{(\mathbb{E}[y_{t=0}|D = 1] - \mathbb{E}[y_{t=0}|D = 0])}_{\text{pre difference}}$$

Difference-in-differences – Graphical representation



DD remarks (more later)

- treatment allowed to be correlated with the potential outcomes
- but treatment needs be uncorrelated with **change in potential outcomes** ("parallel trends assumption")
  - assumes treated and control observations would have developed in parallel without treatment
- algebraically, DD 'nets out' pre-existing differences in outcomes

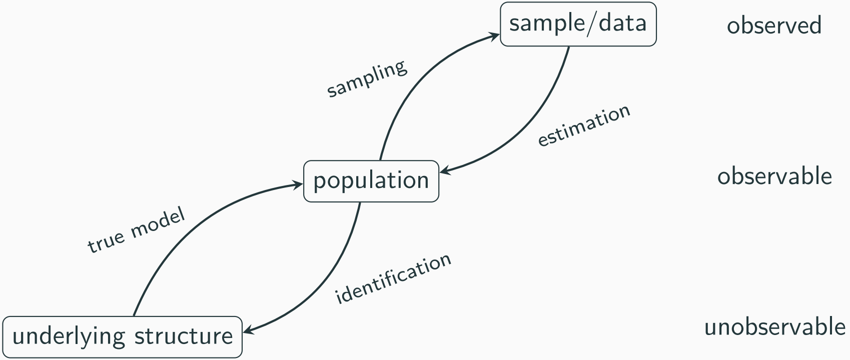
Recap

1. In randomized experiments: Treatment is randomly assigned
   - effect is identified by the difference in means
2. RD: Treatment is discontinuous along some running variable
   - effect is identified by the jump in outcomes at the cutoff
   - RDs often arise from administrative or legal rules
3. DD: 2 groups, 2 periods. Only one group gets treated in the second period
   - effect is identified by the difference-in-differences

Next

- Estimation

Identification, estimation, inference



Estimation

Estimation and inference

The first part was on identification:

- How do things we care about (causal effects) relate to population moments (differences is means).

Not covered:

- How to estimate these?
  - While the population is hypothetically observable, we usually only have a sample of observations
- Need to estimate population moments from the sample
  - Estimation: Obtaining "best guesses" for population moments.
    - E.g., using sample means to estimate population means.
  - Inference: Test hypotheses, assess uncertainty in estimates.
    - E.g., checking if an estimated difference could be the result of chance (during sampling) or is an actual difference in population means

Usually estimation is done by means of some regression.

Further topics

Basics on estimation

▸ Estimation ▸ Estimators in the most basic forms

Inference

▸ Sources of uncertainty ▸ Two ways to think about uncertainty ▸ Inference

Bootstrapping

▸ Inference – Sampling-based uncertainty ▸ Bootstrap - Example

Randomization inference

▸ Randomization inference – Design-based uncertainty ▸ Randomization inference - Example (1)

▸ Randomization inference - Example of an insignificant estimate ▸ Randomization inference - Example of a significant estimate

▸ More on randomization inference

Recap and outlook

Section recap

- Many relevant research questions are causal questions
- Comparing groups in observational data says little about causality
  - Because of selection bias
- Identification results imply certain relationships between population moments and underlying structure
  1. RCTs imply that means between groups correspond to causal effects
  2. RD and DD are alternative identification strategies where (under certain identifying assumptions) causal effects can be identified

Outlook

- Rest of today
  - RCTs
- Session 2: Regression discontinuity
  - Identification and estimation
  - Suri, Bharadwaj, and Jack (2021)
- Section 3: Difference-in-differences

Let's have a break.

Additional resources

Books:

- Cunningham (2021)
- Angrist and Pischke (2008)
- Imbens and Rubin (2015)

Videos:

- Videos by Josh Angrist on RCTs and related topics
  mru.orgcourses/mastering-econometrics/introduction-randomized-trials

Bibliography

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2020. "Sampling-Based Versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88 (1): 265–96.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics*. Princeton university press.

Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. Yale University Press, free via https://mixtape.scunning.com.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Suri, Tavneet, Prashant Bharadwaj, and William Jack. 2021. "Fintech and Household Resilience to Shocks: Evidence from Digital Loans in Kenya." *Journal of Development Economics* 153: 102697.

Appendix

Back

Estimation

- Identification results are statements linking underlying structure to the population distribution
- The easiest approach to estimation is to replace properties of the population distribution by sample analogues. E.g.,

$$
\begin{aligned}
E[y] &\rightarrow \bar{y} &&= \frac{1}{n}\sum_i y_i && \text{(sample means)} \\
E[y|D=1] &\rightarrow \bar{y}|_{D=1} &&= \frac{1}{\sum_i D_i}\sum_i D_i y_i && \text{(subsample means)} \\
E[y|x] &\rightarrow \hat{y}|_x &&= \widehat{\alpha} + \hat{\beta}x && \text{(fitted regression values)} \\
& \ldots
\end{aligned}
$$

Back

Estimators in the most basic forms

- Randomized experiments

$$\widehat{ATE} = \bar{y}|_{D=1} - \bar{y}|_{D=0}$$

- Sharp regression discontinuity, choose bandwidth b:

$$\widehat{ATE}|_{x=\bar{x}} = \bar{y}|_{x-b>x>\bar{x}} - \bar{y}|_{x-b<x<\bar{x}}$$

- Difference-in-differences:

$$\widehat{ATT} = (\bar{y_1}|_{D=1} - \bar{y_1}|_{D=0}) - (\bar{y_0}|_{D=1} - \bar{y_0}|_{D=0})$$

Back

Sources of uncertainty

Ex.: Evaluation of a job training program

- Assignment to the program was random (coin flip).
  - Tails: Person participates, Heads: Person does not receive training.
- Interview 20 random people from 5 random cities
  - 100 in total, 50 treated 50 control
- Finding: those who participated earn *on average* US$1/day more.
- Interpretation
  - Since the program was randomized, we can say that difference in pop means $\mathbb{E}[y|D=1] - \mathbb{E}[y|D=0]$, identifies the ATE.
  - Since the sample was randomly selected, we can say that difference in means, $\bar{y}|_{D=1} - \bar{y}|_{D=0}$, estimates the difference in expectations
- Thus: US$1 is our estimate for the ATE.

Where is uncertainty in coming from?

- random sampling
- random treatment assignment

Back

Two ways to think about uncertainty

1. Uncertainty about individuals
   - There is a population (say 4m working-age Austrians) half are treated, half control
   - Our random sampling only draws 100 from those
2. Uncertainty about other potential outcomes
   - Even if there is no sampling uncertainty (we observe the whole population) maybe we randomly gave treatment to those who had a good outcome anyways.;

Often that distinction make a negligible difference for results.

- in some cases (small samples) it matters for how we think about uncertainty (see Abadie et al. 2020 for a discussion)

Inference

Typically standard OLS asymptotics are sufficient to give us decent ...

- standard errors
- p-values
- confidence bands

For other cases we might resort to

1. bootstrapping

Back

Inference – Sampling-based uncertainty

- Goal:
  - Quantify the extent to which an estimate is a result of the sample.
  - If we took a new sample, how much would findings differ?
- Problem:
  - We only have one sample
- Infeasible solution:
  - Repeat the whole data collection 1000 times.
  - This would give us the distribution of the effect estimate, given the effect.
- Bootstrap solution:
  - Pretend the sample is the population and repeatedly draw new samples (with replacement) from it.
    - Mimics the infeasible solution.
    - Allows to study how conclusions (i.e. estimates) vary across draws.

Back

Bootstrap - Example

- Recall: 5 cities were randomly sampled and in each city 20 random people where interviewed.
- Bootstrap "Algorithm":
  1. Draw, with replacement, 20 people for each of the 5 cities from the sample, to obtain a bootstrap sample of 100 people.
  2. Estimate the treatment effect in the bootstrap sample, $\widehat{\tau}_b$.
  3. Repeat steps 1-2 B times (e.g., 10,000): $\{\widehat{\tau}_b\}_{b=1,\ldots,B}$.
  4. Compute summary statistics for the distribution of bootstrap estimates.
- standard deviation of $\{\widehat{\tau}_b\}$ $\Rightarrow$ standard error of the estimate.
- 2.5% and 97.5% percentile of $\{\widehat{\tau}_b\}$ $\Rightarrow$ the 95% confidence interval.
- This quantifies the sampling-based uncertainty.
  - This ignores sampling uncertainty from the selection of the 5 villages: $\Rightarrow$ Alternative: Draw (with repl.) a sample of 5 villages in step 1

Back

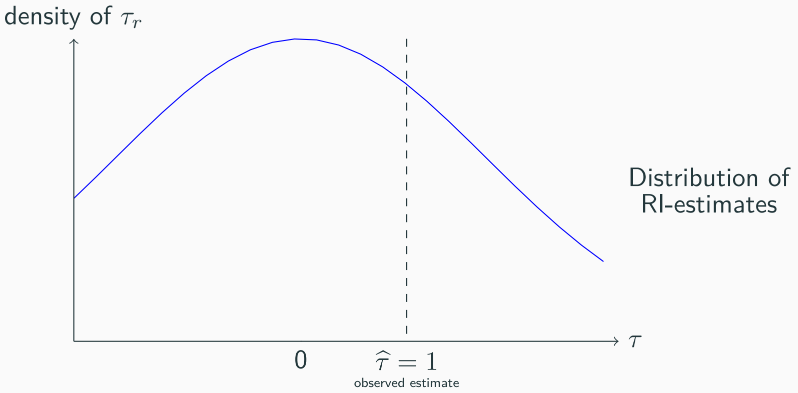
Randomization inference – Design-based uncertainty

- Goal:
  - Quantify the extent to which an estimate is a result of the realization of the treatment assignment.
  - How likely would we observe a certain estimate if there was no effect?
- Problem:
  - We only observe data generated under the true (unknown) effect.
- Infeasible solution:
  - Repeat the whole experiment 1000 times giving placebo treatments.
  - This would give us the distribution of the estimate, given no effect.
- Randomization inference solution (aka, permutation tests):
  - Estimate the effect for 1000 hypothetical treatment assignments.
    - Mimics the distribution of effect estimator if there's no effect.
    - Allows to study if our true estimate is [un]likely to be the result of chance.

Back Randomization inference - Example (1)

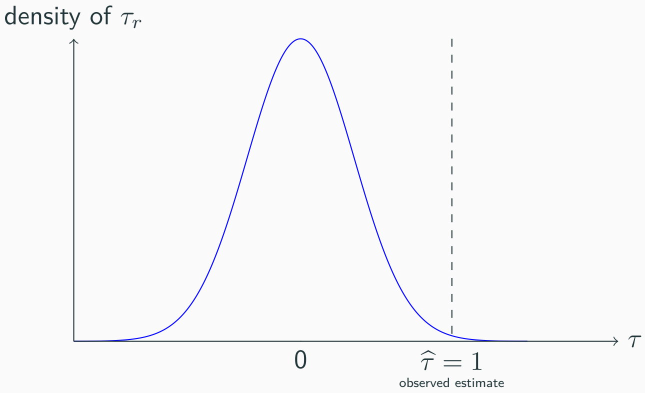
- Recall: For each person a coin flip determined participation.
- Randomization inference (RI) "Algorithm":
  1. Simulate a new coin flip for each participant
  2. Estimate the treatment effect (difference in means) using the real outcome data but the "fake" treatment dummy, $\widehat{\tau}_r$.
  3. Repeat steps 1-2 R times (e.g., 10,000): $\{\widehat{\tau}_r\}_{r=1,\ldots,R}$.
  4. Compare the 'true' estimate $\widehat{\tau}$ against the distribution of $\{\widehat{\tau}_r\}$.
- Recall: We started RI off imposing that there is no treatment effect
- If the true estimate falls "outside" the distribution of RI-estimates:
  - The estimate is not what we would expect if there was no effect.
  - We contradicted our imposed assumption, so there must be an effect.
- If the true estimate lies "well within" the distribution:
  - The estimate is consistent with what we expect if there was no effect.

Randomization inference - Example of an insignificant estimate



- The estimated treatment effect, $\widehat{\tau}$, is not very different from the R "treatment effects" we estimated using "fake" coin tosses.
  - i.e, the observed difference is not significant.

Back Randomization inference - Example of a significant estimate



- The estimated treatment effect, $\widehat{\tau}$, is very different from the R "treatment effects" we estimated using "fake" coin tosses.
  - the observed difference is inconsistent with the H_0 of no effect.

Back

More on randomization inference

- Goal: Understand if we would observe our estimate $\hat{\beta}$ under $H_0 : \beta = 0$.
- Basic idea behind randomization inference straightforward:
    - If H_0, then D does not matter. I.e., values for D should explain our data equally well.
    - If we reshuffle D we mimic data from "parallel universes".
    - If H_0, these are as 'valid' as our actual data.
- Draw R alternative treatment assignments and compute the treatment effect estimate on those data.
- If H_0, then our actual estimate $\hat{\beta}$ can is a draw from the distribution of the R estimates.
- If not H_0, then our actual estimate may be entirely different.
- $\Rightarrow$ Reject the H_0, if our estimate is far outside the distribution.