# Propensity Score Matching
# Hebrew University of Jerusalem
# Morning Session

Alberto Stefanelli and Prof. Sharon Gilad

2019-12-11

Section 1

# The Workshop

# About me: Alberto

- PhD Candidate at University of Leuven (Ku Leuven), Belgium
- Previously...
  - MA at Central European University
- My research:
  - Radial Beliefs System: Measurement, Causes, Consequences
  - Polarization: Measurement, Consequences
  - Methods: Causality, experimental and semi-experimental design, SEM etc.
- Contact: alberto.stefanelli@kuleuven.be
- Website: www.albertostefanelli.com
- Twitter: @sergsagara

# About me: Sharon

- Professor at Hebrew University, Jerusalem
- Previously...
  - PhD at Oxford, Nuffield Collage
  - Researcher at King Collage
- My research:
  - Public Administration
  - Regulation and agenda setting
  - Response of civil servants and bureaucrats to public pressures
- Methods:
  - Mixed-Methods
- Contact: sharon.gilad@mail.huji.ac.il

# Your turn

- Name?
- Affiliation? Country?
- Research interests?
- Previous experience with experimental designs?
- Previous experience with R?
- Why are you taking this workshop?

# Structure (1/2)

**Morning (tentative)**

1. 10.00 - 10.30 Get to know each other and WS Presentation
   - Who are we ?
   - Your turn
   - Workshop presentation and rational
2. 10.30 - 11.00 "The problem of Causality"
   - Terminology
   - Equifinality and Manipulation
3. 11.00 - 11.30 Potential outcomes framework
   - General concept
   - The fundamental problem
4. 11.30 - 11.45 Coffee Break
5. 11.45 - 12.30 Casual Estimands
   - Differences between ATE ATT ATC ?
   - Exercise 1

# Structure (1/2)

**Afternoon (tentative)**

1. 14.00 - 14.15 Recap
2. 14.15 - 15.45 Propensity Matching Score
   - Matching, PS, Weights: Making sense of the terms
   - Why we should use matching
3. 14.45 - 15.30 How it works
   - Steps
   - What is happening under the hood
4. 15.30 - 16.00
   - Selection on the observables
   - Covariate Balancing
5. 16.00 - 16.15 Break
6. 16.15 - 16.45 Lalonde dataset
   - Description
   - Assessing covariate Balancing
   - Estimating the ATE
7. 16.45 - 17.00 Setting up R Studio
8. 17.00 - 17.15 Look at some code
9. 17.15 - 18.15 Exercise: Casual inference with Observational data

Section 2

# Where to find the material

# Where to find the material

1. On my website
    - Slides Morning:
    - Slides Afternoon:
    - Set up R: https://albertostefanelli.com/files/set_up.html
    - Coding Example:
    - Script for the Exercise:
2. Data
    - From my github
    - From the replication dataset https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TAXCB3

# Goal of the Workshop

1. Theoretical foundations
   - Understanding Causal Analysis
   - Understanding Propensity Score Matching
2. Technical skills
   - Gain visual understanding of data
   - Learning by doing in R
3. Have fun!

# Reading list I

**Recommended:**


- Morgan, S. L., & Winship, C. (2015). Counterfactuals and Causal Inference. 526. Chapter 2 p.37-68
- Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC." Chapter 7 p.83-98
- Bauer, P. C. (2015). Negative Experiences and Trust: A Causal Analysis of the Effects of Victimization on Generalized Trust. European Sociological Review, 31(4), 397–417. https://doi.org/10.1093/esr/jcu096
- Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. Journal of economic surveys, 22(1), 31-72.
- Bartels, L. (2013). Your genes influence your political views. So what? Washington Post. Retrieved from https://www.washingtonpost.com/news/monkey-cage/wp/2013/11/12/your-genes-influence-your-political-views-so-what/

# Reading list II


- Bartlett, J. (2016). Why you shouldn't use propensity score matching – The Stats Geek. Retrieved November 15, 2019, from https://thestatsgeek.com/2016/09/07/why-you-shouldnt-use-propensity-score-matching/


**Optional:**


- Keele, Luke. 2015. "The Statistics of Causal Inference: A View from Political Methodology." Polit. Anal. 23 (3): 313–35
- Felix J. Thoemmes & Eun Sook Kim (2011) A Systematic Review of Propensity Score Methods in the Social Sciences, Multivariate Behavioral Research, 46:1, 90-118, DOI: 10.1080/00273171.2011.540475
- Hedström, Peter, and Petri Ylikoski. 2010. "Causal Mechanisms in the Social Sciences." Annu. Rev. Sociol. 36 (1): 49–67.

Section 3

# The Problem of Causality

# Experimental logic

1. Does this workshop (treatment X) affect your knowledge on PS (outcome Y)?
2. Discussion
    - Q: Is it sufficient to measure knowledge before the lecture?
    - Q: If yes, when should I measure knowledge after the lecture?
    - Q: Is that a randomized experiment here?
    - Q: What's the problem if someone knows everything beforehand?

# Casual analysis: The basic terminology

1. In broader terms, causality is a connection of phenomena that connects one element (the cause) with another elements (effect/outcome/response)
    - It is a **process**. The 1st element of this process is responsible for 2nd and the 2nd dependent on the 1st.
    - Causality is temporally bound. The cause(s) must precede the effect and **all lie in its past.**
2. Terminology (many terms that refers to same phenomena)
    - "Causality" = "Causation" = "Cause and Effect" = "Casual Mechanism"
    - Treatment (yes, it is a variable) = Cause = treatment and control groups = pre-treatment post-treatment
    - Outcomes/response (yes, it is a variable) = Effect

# Casual analysis: Equifinality

3. Social reality is complex. An outcome can be a cause of many other effects.
    - This is called equifinality.
    - It is what Hume (1772) summaries in the **billiard ball example.**

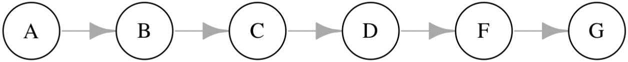


**Figure 1:** Equifinality

# Casual analysis: Identification

1. Q: At age 40 both Cory and Janneke have an income (Y) of 3000 Euros. How can this be an example for equifinality?
2. The longer the time gap between treatment and outcome
   - the more fuzzy the theory (causal chain).
   - the more likely something else, unrelated happened in between.
3. Intervening mediating variables can make the process difficult to identify (K. Imai et al. 2011)
   - Want to think more about it?
   - Read "Republic should pray for rain" Gomez, Hansford, and Krause 2007,
   - Weather —-> voter turnout (Gomez, Hansford, and Krause 2007)
   - Q: Are there intervening phenomena ?

# Casual Analysis: Manipulation (1/2)

1. Example of the causes in the social sciences
   - Income
   - Social class
   - Exposure to political news/information
   - Education
   - Genes
   - Personality traits
   - Gender
   - Ethnic background
   - Exposure to particular policy intervention
2. Q: The effect is the exam result: What is the difference between the “causes” in the following examples?
   - She did well on the exam because she is a woman.
   - She did well on the exam because she studied for it.
   - She did well on the exam because she was properly coached by her teacher.

# Casual Analysis: Manipulation (2/2)

1. **No Causation WITHOUT Manipulation"** (Holland 1986, 954–55, 959; Rubin 1975, 238)
   1. "[C]auses are only those things that could, in principle, be treatments in experiment" (Holland 1986, 954)
   2. "[C]auses are experiences that units undergo and not attributes that they possess" (Holland 2003: 8)

Section 4

# Potential Outcome Framework

# Potential Outcome Framework: A Definition

1. Developed by Fisher (1935) and Neyman (1923). After it has being forcefully advocated in a series of papers by Rubin (1974, 1977).
2. Contrasts between observed exposure to one state and **what-if counter-factual** exposure to another state
   - What would have happened if a particular unit exposed to treatment is at the same time exposed to the control (Imbens and Rubin 2015, 4)
   - Comparison of potential **(NOT observed)** outcomes of the same individual at the same moment in time post-treatment (Imbens and Rubin 2015)

# Potential Outcome framework: A visualization (1/2)

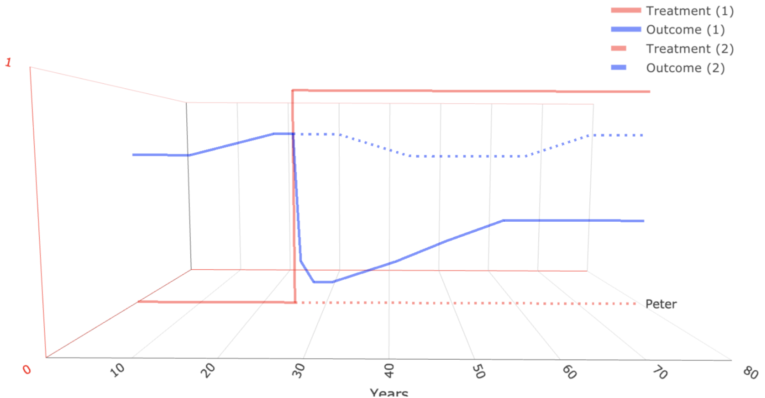


**Figure 2:** Graph POF

# Potential Outcome framework: The central question

Q: What is the **Fundamental Problem of Casual Inference**?

# Potential Outcome framework: The Answer

3. **Fundamental Problem of Casual Inference:** we can never simultaneously measure the response in case they did participate, and the response in case they did not participate. (Holland 1986)

# Potential Outcome framework: A visualization (2/2)

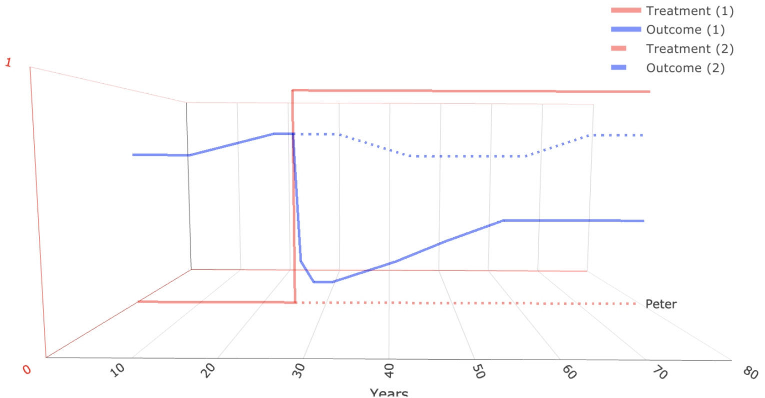


**Figure 3:** Graph POF

# Resolving the fundamental problem

1. Single Individual-level Treatment Effect: Difference in potential outcome for unit i at time t post-treatment
2. Missing data problem
3. Estimation requires **filling in the missing counterfactual** (the missing data point)
4. Basically, the potential outcomes are the data that **we wish we had** to estimate causal treatment effects

| Unit | $D_i$ (Aspirin: Yes/No) | $Y_i$ (Pain: Yes/No) | $Y_i1$ (Pain\|Aspirin:Yes) | $Y_i0$ (Pain\|Aspirin:No) |
|---|---|---|---|---|
| Alberto | Yes | No | No | **?** |

# Estimating the ITE

- Definition of "causal effect" does not require more than one unit (Imbens and Rubin 2015, 8)
- However, estimation requires multiple observations of either
  - same unit/individual (pre-post)
  - different units/individuals (control treatment)
- ITE is **unidentifiable** meaning that we cannot estimate it.

# So What?

So what?

# The Average Treatment Effect (ATE) (1/3)

- ATE= The **average** difference in the potential outcomes averaged over the entire population of interest (at a particular moment in time)
- Averaging allows us to borrow information from other units
- Meaning that we fill up the missing counterfactuals
- We circumnavigate the fundamental problem of casual inference

# The Average Treatment Effect (2/3)

- ATE $= E[Y_i 1 - Y_i 0]$
- **Naive estimate of ATE:** Difference between expected values in treatment and control
- Example: Effect of Academic college degree on a productivity

| Units | $D_i$ (Collage: Yes/No) | $Y_i$ (Level of Productivity) |
|---|---|---|
| Chris | Yes | 40 |
| Julia | Yes | 50 |
| Paul | No | 20 |
| Trump | No | 10 |
| Fred | No | 40 |
| Diego | No | 30 |

# The Naive ATE

1. Q: How can we calculate the ATE?
2. **Exercise 1:** Calculate the naive ATE
3. Hint $\sum_{i=1}^{N}(Y_i1) - \sum_{i=1}^{N}(Y_i0)$

| Units | $D_i$ (Collage: Yes/No) | $Y_i$ (Level of Productivity) |
|---|---|---|
| Chris | Yes | 40 |
| Julia | Yes | 50 |
| Paul | No | 20 |
| Trump | No | 10 |
| Fred | No | 40 |
| Diego | No | 30 |

# The Naive ATE: Results

$$
\begin{aligned}
ATE &= (40 + 50)/2 - (20 + 10 + 40 + 30)/4 \\
&= 45 - 25 \\
&= 20
\end{aligned}
$$

1. Simple mean difference

# The Average Treatment Effect (ATE)

- ATE can be **decomposed** as a function of 5 quantities (e.g. Keele 2015b, 4):
  - $\pi$=Proportion of the sample that received the treatment (e.g., 2/6=0.33)
  - $E[Y_i1|D_i = 1]$ : Average outcome under treatment for units in treatment condition
  - $E[Y_i0|D_i = 0]$ : Average outcome under control for those in the control condition
  - $E[Y_i1|D_i = 0]$ : Average outcome under treatment for those in the control condition
  - $E[Y_i0|D_i = 1]$ : Average outcome under control for units in treatment condition

| Units | $D_i$ (Collage: Yes/No) | $Y_i$ (Level of Productivity) | Productivity $Y_i1|D_i = 1(Yes)$ | Productivity $Y_i0|D_i = 0(No)$ |
|---|---|---|---|---|
| Chris | Yes | 40 | **40** | **?** |
| Julia | Yes | 50 | **50** | **?** |
| Paul | No | 20 | **?** | **20** |
| Trump | No | 10 | **?** | **10** |
| Fred | No | 40 | **?** | **40** |
| Diego | No | 30 | **?** | **30** |

# The Omniscient ATE

- Let's assume we are omniscient creatures and we can observe the quantities that were before unobservable
  - $E[Y_i1|D_i = 0]$ : Average outcome under treatment for those in the control condition
  - $E[Y_i0|D_i = 1]$ : Average outcome under control for units in treatment condition

| Units | $D_i$ (Collage: Yes/No) | Productivity $Y_i$ | Productivity $Y_i1$ | Productivity $Y_i0$ |
|---|---|---|---|---|
| Chris | Yes | 40 | **40** | **35** |
| Julia | Yes | 50 | **50** | **20** |
| Paul | No | 20 | **40** | **20** |
| Trump | No | 10 | **25** | **10** |
| Fred | No | 40 | **40** | **40** |
| Diego | No | 30 | **35** | **30** |

# The Omniscient ATE: Exercise I

2. **Exercise 2:** Calculate the ATE with these new data
   - Q: Do we have any missing information?
   - Hint:

$E[Y_i1 - Y_i0] = (\pi)\ (E[Y_i1|D_i = 1]$ - $E[Y_i0|D_i = 1]) + (1 - (\pi))\ (E[Y_i1|D_i = 0]$ - $E[Y_i0|D_i = 0])$

# The Omniscient ATE: Calculation I

1. ATT = average treatment effect for those that typically are (choose to be) treated based on counterfactual comparison.
2. ATT= $E[Y_i 1|D_i = 1]$ - $E[Y_i 0|D_i = 1]$

$$\begin{aligned} ATT &= (40 + 50)/2 - (40 + 25 + 40 + 35)/4 \\ &= (90/2) - 140/4 \\ &= 45 - 35 \\ &= 10 \end{aligned}$$

# The Omniscient ATE: Calculation II

1. ATC = average treatment effect for those that typically are (choose to be) **NOT treated** based on counterfactual comparison.
2. ATC= $E[Y_i1|D_i = 0]$ - $E[Y_i0|D_i = 0]$

$$\begin{aligned} ATC &= (35 + 20)/2 - (40 + 25 + 40 + 35)/4 \\ &= (55/2) - (100/4) \\ &= 27.5 - 25 \\ &= 2.5 \end{aligned}$$

# The Omniscient ATE: Calculation III

1. ATE is just the sum from the effect among the treated and the effect among the control
2. Easy as that !

$$
\begin{aligned}
ATE &= ATT + ATC \\
&= 10 + 2.5 \\
&= 12.5
\end{aligned}
$$

1. **ATE is 12.5 and NOT 20**
2. Q: Why is this the case?