Introduction to Statistical Programming

Session 1

Dr Giacomo Vagni
Max Weber Fellow
EUI

Nov, 2021

This course

- Instructor: Dr Giacomo Vagni
- Teaching Assistant: Inês Azevedo

4 sessions of 2 hours

- Potential Office Hours (if needed upon request)
- Marking (pass/fail). A research brief (1-3 pages).
  - Articulate a research question
  - Find data
  - Run small empirical analyses **using R** to address the question

Let's learn and have fun!

Plan

- Short introduction to causal inference
- R Statistical Programming

Why Study Statistical Programming?

- Public Policy need facts
- Facts need data
- Data need to be processed to be usable

Statistical Programming = **data processing**

1. Cleaning data
2. Analysing data (statistical analysis)

Data ALWAYS need to be cleaned and processed

Why Study Causality?

- Causality is fundamental tool for public policies and interventions
- No efficient policy without causality

Causal Questions are Everywhere!

1. **Do COVID lockdowns reduce mortality?**
   - Born, B., Dietrich, A. M., & Müller, G. J. (2021). The lockdown effect: A counterfactual for Sweden. *Plos one*, 16(4),
2. **Should malaria bed nets be free or paid for?**
   - Cohen, J., & Dupas, P. (2010). Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. *The Quarterly Journal of Economics*, 1-45.
3. **Does economic liberalisation foster economic growth?**
   - Billmeier, A., & Nannicini, T. (2013). Assessing economic liberalization episodes: A synthetic control approach. *Review of Economics and Statistics*, 95(3), 983-1001.
4. **Do revolutions help or hinder development?**
   - Korolev, I. (2021). How Could Russia Have Developed without the Revolution of 1917?

But what is causality

- Causality is about
  - Manipulating
  - “Doing’ ’
  - Intervening
  - Treating

How an outcome Y changes in a predictible fashion given a specific intervention/treatment/action T

$$\mathrm{T} \xrightarrow{\text{causes}} \mathrm{Y}$$

Potential Outcomes Framework

Formalised by Donald Rubin

- Let Y be an outcome of interest (e.g. GDP, health, education, ...)
- Let T be the treatment/intervention/policy, with $T = 1$ treatment and $T = 0$ no treatment.
- $Y_i^{T=1}$ is the outcome under treatment for individual i
- $Y_i^{T=0}$ is the outcome under no treatment for the *same* unit i

The goal of causality is to estimate the difference between

$$\delta_i = Y_i^{T=1} - Y_i^{T=0} \tag{1}$$

Sliding Doors

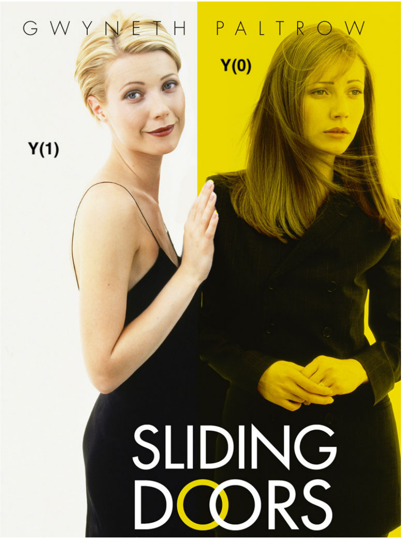


Figure 1: Potential Outcomes

It's a Wonderfuly Life

Figure 2: Potential Outcomes

Example

- $T =$ Studying for an exam
- $Y =$ Exam score
- C**asu**al question: what is the association between study time and exam scores?
- C**aus**al question: does studying for an exam **cause** higher scores?
- Potential outcomes question: what **would the exam scores have been** for those who studied if they had not studied?

But what is the problem with potential outcomes questions?

Fundamental Problem of Causal Inference

- We can only either observe $Y_i^{T=1}$, $Y_i^{T=0}$
- We can never observe the same person both studying and not studying
- We only observe the **realised outcomes**

Realised (observed) outcome for those who studied

$$\overbrace{Y_i^{T=1}}^{\text{potential}} \underbrace{\mid}_{\text{given}} \overbrace{T=1}^{\text{observed}} \tag{2}$$

Realised outcome for those who did not study

$$Y_i^{T=0} \mid T=0 \tag{3}$$

Sliding Doors Y0

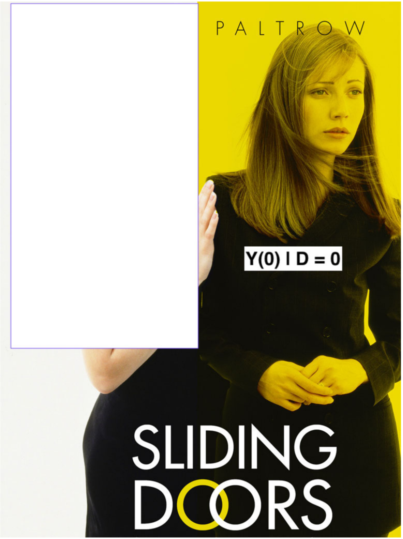


Figure 3: Potential Outcomes

Sliding Doors Y1

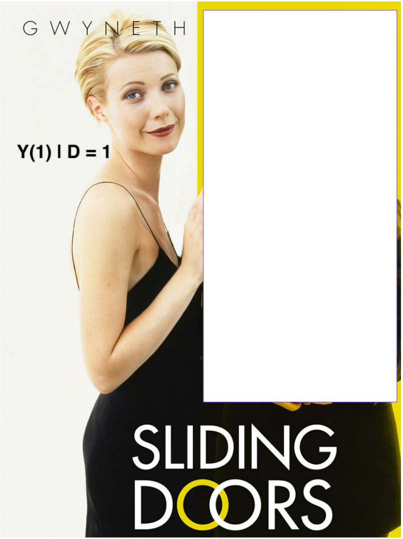


Figure 4: Potential Outcomes

We do not observe for those who studied $T = 1$, the score they would have had $T = 0$ without studying

$$\underbrace{\overbrace{Y_i^{T=0}}^{\text{potential outcome}} \quad \mid T = 1}_{\text{unobserved}} \tag{4}$$

And for those who did not study $T = 0$, their potential score

$$\overbrace{Y_i^{T=1}}^{\text{potential outcome}} \quad \mid T = 0 \tag{5}$$

4 students and their potential outcomes

E.g. exam scores from 0 to 100

Table 1: Omniscient Potential Outcomes Table

	Potential Outcomes		δ_i
	$Y^{t=0}$	$Y^{t=1}$	$\underbrace{Y^{t=1} - Y^{t=0}}_{\text{causal effect}}$
Natalie	40	60	+20
Francesca	80	80	0
Donatella	20	20	0
Sierra	0	50	+50

Table 2: Realised Outcomes Table

	T	Y_{obs}	Realised Outcomes $Y^{t=0}$	$Y^{t=1}$	δ_i $\underbrace{Y^{t=1} - Y^{t=0}}_{\text{causal effect}}$
Natalie	$T = 1$	60	?	60	?
Francesca	$T = 0$	80	80	?	?
Donatella	$T = 0$	20	20	?	?
Sierra	$T = 1$	50	?	50	?

Can we simply compare who got the treatment and who did not?

$$\text{Causal Effect ?} = \frac{\overbrace{\text{Natalie} + \text{Sierra}}^{\text{Treated}}}{2} - \frac{\overbrace{\text{Francesca} + \text{Donatella}}^{\text{Control}}}{2} \tag{6}$$

The Leap – in average (expected value) terms

We want the **average treatment effect** with

$$\begin{aligned} \mathbb{E}[\delta] =& \mathbb{E}[Y^1 - Y^0] \\ =& \mathbb{E}[Y^1] - \mathbb{E}[Y^0] \end{aligned} \tag{7}$$

$\mathbb{E} \rightarrow$ "expected value" (similar to the average/mean)

But we have

$$\mathbb{E}[\delta] = \mathbb{E}[Y^1 | T = 1] - \mathbb{E}[Y^0 | T = 0] \tag{8}$$

Let's recap the potential quantities

For those who received the treatment, what would their outcome be **without** (in our example Natalie and Sierra)

$$\overbrace{\mathbb{E}[Y^1|T=1]}^{a.Obs} - \overbrace{\mathbb{E}[Y^0|T=1]}^{b.Unobs} \tag{9}$$

For those who did **not** receive the treatment, what would their outcome be **with** treatment (in our example Francesca and Dona)

$$\overbrace{\mathbb{E}[Y^1|T=0]}^{c.Unobs} - \overbrace{\mathbb{E}[Y^0|T=0]}^{d.Obs} \tag{10}$$

$$\text{Selection Effect} = \underbrace{\mathbb{E}[Y^0 \mid T=1]}_{b.Unobs} - \underbrace{\mathbb{E}[Y^0 \mid T=0]}_{d.Obs} \tag{11}$$

Selection Effect

$$\text{Selection Effect} = \overbrace{\mathbb{E}[Y^0 \mid T = 1]}^{b} - \overbrace{\mathbb{E}[Y^0 \mid T = 0]}^{d} \tag{12}$$

If those who received the treatment *would* (potential) not have fared better than those who did not, **then** the selection effect is 0 and we can compare the treated and control groups.

$\rightarrow$ in observational studies, this is not generally the case

Selection Effect

- People who benefit from a treatment because of X, will seek the treatment (select themselves).

$\rightarrow$ next session we will see come back to this and see in more details the different forms of bias.

Relates to

Compare apples to apples

Figure 5: Apple vs Orange

Figure 6: Apple Heterogeneity

Ex. Effect of temperature (D, treatment) on apple growth (Y, outcome)

Need to be able to make meaningful comparison to estimate causal effect!

Figure 7: Apple Growth

Naive comparison of countries and policies

Ex. Effect of Welfare State on GDP

- France vs USA
- USA is much richer than France, but has smaller weflare system
- Is welfare therefore bad for growth?
- The **true question** is the *potential outcome* question: Would the USA be even better off with a larger welfare system? or would it be worse off?

Table 3: Observed Outcomes Table for France and US, GDP and Welfare

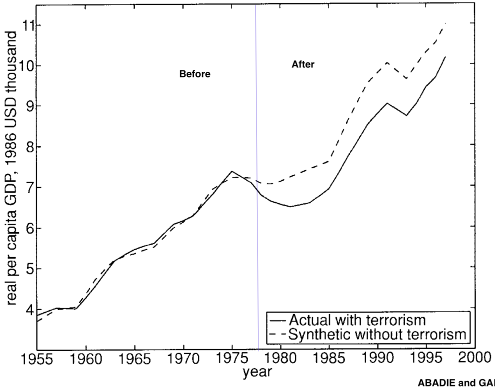
	Potential Outcomes		Observed	δ_i
	$Y^{t=0}$	$Y^{t=1}$	Y_{obs}	$\underbrace{Y^{t=1} - Y^{t=0}}_{\text{causal effect}}$
France	?	2.6 trillion	GDP 2.6	?
USA	20.9 trillion	?	GDP 20.9	?

The whole goal of causality is to estimate the unobserved potential outcome / counterfactual

Figure 8: Counterfactual

Effect of Terrorism and GDP in Spain

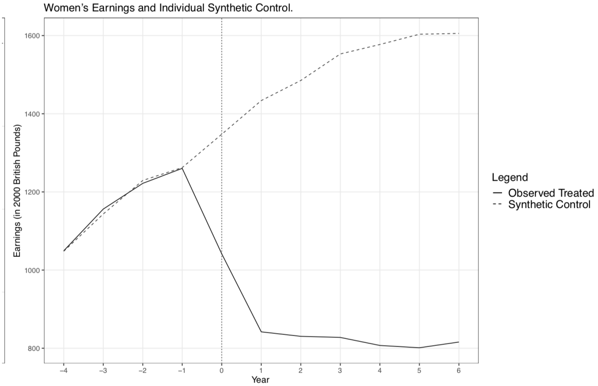
- Basque Country experienced outbreak of terrorism in the 1960-1970s.
- Counterfactual estimate: GDP declined 10% relative to a synthetic control region without terrorism



ABADIE and GARDEAZABAL 2003

Effect of Motherhood on Women's Earnings

- Counterfactual estimate: 45% motherhood earning penalty



- G. Vagni and R. Breen (2021) *European Sociological Review* https://doi.org/10.1093/esr/jcab014

Next Session

- Different types of bias
- How to remove bias
- How to express bias graphically