Randomization

Silje Synnøve Lyder Hermansen

03-12-2019

Where are we? And what are we at?

We've completed the first part of the course: Congrats!

Where are we? And what are we at?

We've completed the first part of the course: Congrats!

- Our focus has been on *describing* data: GLMs

Where are we? And what are we at?

We've completed the first part of the course: Congrats!

- Our focus has been on *describing* data: GLMs
- Now, we'll focus on *research design*: causal inference

Where are we? And what are we at?

We've completed the first part of the course: Congrats!

- Our focus has been on *describing* data: GLMs
- Now, we'll focus on *research design*: causal inference

The goal of the social sciences

Why do we run regressions?

We run regressions to learn about the world, which means. . .

Why do we run regressions?

We run regressions to learn about the world, which means. . .

- To describe data

Why do we run regressions?

We run regressions to learn about the world, which means...

- To describe data $\rightarrow$ observe the world

Why do we run regressions?

We run regressions to learn about the world, which means...

- To describe data $\rightarrow$ observe the world
  - ... but how do we know if it's not an illusion?

Why do we run regressions?

We run regressions to learn about the world, which means...

- To describe data $\rightarrow$ observe the world
  - ... but how do we know if it's not an illusion?
- To make causal claims

Why do we run regressions?

We run regressions to learn about the world, which means. . .

- To describe data $\rightarrow$ observe the world
  - . . . but how do we know if it's not an illusion?
- To make causal claims $\rightarrow$ manipulate the world

Why do we run regressions?

We run regressions to learn about the world, which means...

- To describe data $\rightarrow$ observe the world
  - ... but how do we know if it's not an illusion?
- To make causal claims $\rightarrow$ manipulate the world
  - ... in the social sciences, that's not always possible

Why do we run regressions?

We run regressions to learn about the world, which means...

- To describe data $\rightarrow$ observe the world
  - ... but how do we know if it's not an illusion?
- To make causal claims $\rightarrow$ manipulate the world
  - ... in the social sciences, that's not always possible

$\Rightarrow$ *We design studies to approximate manipulation*

We want to make causal claims

Two (compatible) approaches.

We want to make causal claims

Two (compatible) approaches.

- **Logic of inference:** (King, Keohane and Verba, 1994)

We want to make causal claims

Two (compatible) approaches.

- **Logic of inference:** (King, Keohane and Verba, 1994)
  - We can only imperfectly observe the world

We want to make causal claims

Two (compatible) approaches.

- **Logic of inference:** (King, Keohane and Verba, 1994)
    - We can only imperfectly observe the world
    - . . . but we can theorize (causal mechanism)

We want to make causal claims

Two (compatible) approaches.

- **Logic of inference:** (King, Keohane and Verba, 1994)
  - We can only imperfectly observe the world
  - . . . but we can theorize (causal mechanism)
  - . . . and test hypotheses (observable implications)

We want to make causal claims

Two (compatible) approaches.

- **Logic of inference:** (King, Keohane and Verba, 1994)
  - We can only imperfectly observe the world
  - . . . but we can theorize (causal mechanism)
  - . . . and test hypotheses (observable implications)

$\Rightarrow$ *A closer connection between theory and statistics (e.g. EITM).*

We want to make causal claims

Two (compatible) approaches.

We want to make causal claims

Two (compatible) approaches.

- **Logic of inference:** (King, Keohane and Verba, 1994)

We want to make causal claims

Two (compatible) approaches.

- **Logic of inference:** (King, Keohane and Verba, 1994)
- **Potential outcomes** (Donald Rubin)

We want to make causal claims

Two (compatible) approaches.

- **Logic of inference:** (King, Keohane and Verba, 1994)
- **Potential outcomes** (Donald Rubin)

What is causation?

A sequence of events in which – if the first didn't happend – the second wouldn't occur either.

What is causation?

A sequence of events in which – if the first didn't happend – the second wouldn't occur either.

- We can manipulate the first event

What is causation?

A sequence of events in which – if the first didn't happend – the second wouldn't occur either.

- We can manipulate the first event $\rightarrow$ what happens then?

What is causation?

A sequence of events in which – if the first didn't happend – the second wouldn't occur either.

- We can manipulate the first event $\rightarrow$ what happens then?
- Can we infer what *would have* happened if we did not manipulate?

What is causation?

A sequence of events in which – if the first didn't happend – the second wouldn't occur either.

- We can manipulate the first event $\rightarrow$ what happens then?
- Can we infer what *would have* happened if we did not manipulate?

$\Rightarrow$ *Potential outcomes*

We want to make causal claims

Two (compatible) approaches.

- **Logic of inference:** (King, Keohane and Verba, 1994)
- **Potential outcomes** (Donald Rubin)

We want to make causal claims

Two (compatible) approaches.

- **Logic of inference:** (King, Keohane and Verba, 1994)
- **Potential outcomes** (Donald Rubin)
  - *causal effect*: difference between what is and could have been

We want to make causal claims

Two (compatible) approaches.

- **Logic of inference:** (King, Keohane and Verba, 1994)
- **Potential outcomes** (Donald Rubin)
  - *causal effect*: difference between what is and could have been

$\Rightarrow$ *a set of methods designed for causal inference with observational data*

The conundrum

The true causal effect

What is causal effect?

Imagine two versions of me.

What is causal effect?

Imagine two versions of me.

- I have a headache and I take an aspirine ($Y_{1,Silje}$).

What is causal effect?

Imagine two versions of me.

- I have a headache and I take an aspirine ($Y_{1,Silje}$).
- I have a headache but receive no treatment ($Y_{0,Silje}$).

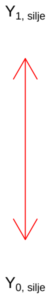
What is causal effect?

Imagine two versions of me.

- I have a headache and I take an aspirine ($Y_{1,Silje}$).
- I have a headache but receive no treatment ($Y_{0,Silje}$).

$\Rightarrow$ *the causal effect is* $Y_1 - Y_0$

True causal effect

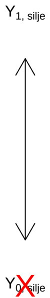


A causal effect is the difference between two potential outcomes

A causal effect is the difference between two potential outcomes

- ... but – at best – I can only observe one outcome.

True causal effect is NOT POSSIBLE to observe



A causal effect is the difference between two potential outcomes

A causal effect is the difference between two potential outcomes

- ... but – at best – I can only observe one outcome.

⇒ *We have to compare two different individuals*

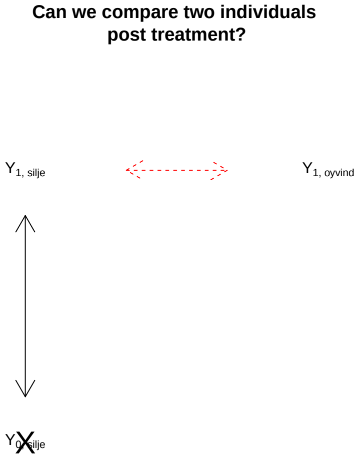
Plan B

Plan B: Can we compare across cases?

Let's compare my headache now with Øyvind's current headache ($Y_{1,Silje} - Y_{1,Oyvind}$)

Let's compare my headache now with Øyvind's current headache ($Y_{1,Silje} - Y_{1,Oyvind}$)

Let's compare my headache now with Øyvind's current headache ($Y_{1,Silje} - Y_{1,Oyvind}$)

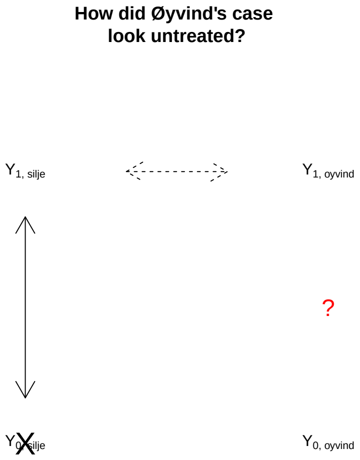


Let's compare my headache now with Øyvind's current headache ($Y_{1,Silje} - Y_{1,Oyvind}$)

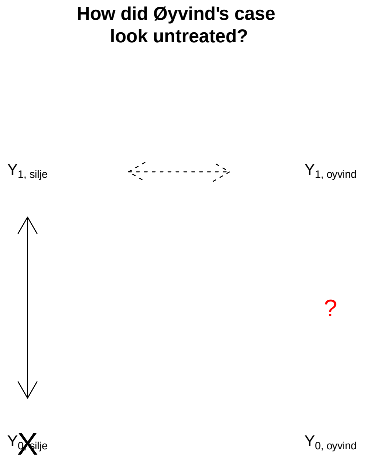
Let's compare my headache now with Øyvind's current headache ($Y_{1,Silje} - Y_{1,Oyvind}$)

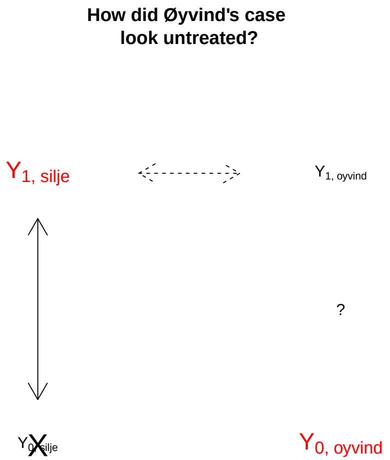
- ... but did he even have a headache before?

Is there a selection bias?

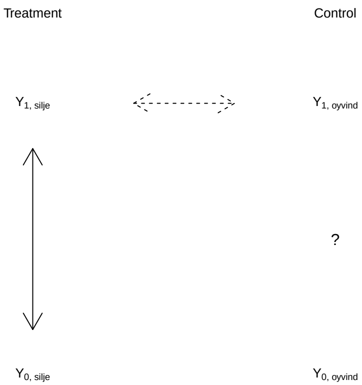


Is there a selection bias?




How did Øyvind's case
look untreated?
Y1, silje
Y1, oyvind
?
Y0, silje
X
Y0, oyvind

What do we compare?



Where's the selection bias?

The solution

We have to observe Øyvind's untreated headache ($Y_{0,Oyvind}$) and compare with treated me ($Y_{1,Silje}$)

The solution

We have to observe Øyvind's untreated headache ($Y_{0,Oyvind}$) and compare with treated me ($Y_{1,Silje}$)

$$\begin{aligned} Y_{Silje} - Y_{Oyvind} &= Y_{1,Silje} - Y_{0,Oyvind} \\ &= Y_{1,Silje} - Y_{0,Silje} + Y_{0,Silje} - Y_{0,Oyvind} \end{aligned} \tag{1}$$

- **Causal effect**: $Y_{1,Silje} - Y_{0,Silje}$
- **Selection bias**: $Y_{0,Silje} - Y_{0,Oyvind}$

How to do it?

We use statistics

We cannot observe two potential outcomes, but we can rely on the law of large numbers (LLN).

We use statistics

We cannot observe two potential outcomes, but we can rely on the law of large numbers (LLN).

- We use **average** causal effect

Average causal effect = Differences in means - Selection bias

Differences in means

- We create a **dummy** for treated vs. untreated observations:

$$D_i = \begin{cases} 1 & \Leftrightarrow & treated \\ 0 & \Leftrightarrow & untreated \end{cases} \tag{2}$$

- We calculate the **differences in means**

$$= Avg_n[Y_i|D_i = 1] - Avg_n[Y_i|D_i = 0] \tag{3}$$

Differences in means

- We create a **dummy** for treated vs. untreated observations:

$$D_i = \begin{cases} 1 & \Leftrightarrow & treated \\ 0 & \Leftrightarrow & untreated \end{cases} \tag{4}$$

- We calculate the **differences in means**

$$\begin{aligned} &= Avg_n[Y_i|D_i = 1] - Avg_n[Y_i|D_i = 0] \\ &= Avg_n[Y_{1,i}|D_i = 1] - Avg_n[Y_{0,i}|D_i = 0] \end{aligned} \tag{5}$$

Basic assumption

We have to assume that the treatment has the same effect accross all units

Basic assumption

We have to assume that the treatment has the same effect accross all units

- then we can compare across units

Basic assumption

We have to assume that the treatment has the same effect accross all units

- then we can compare across units
- contrast that with the effect of β in OLS vs GLM

Selection bias

Now we have to get rid of the selection bias!

Selection bias

Now we have to get rid of the selection bias!

- **A priori** selecting units without bias:

Selection bias

Now we have to get rid of the selection bias!

- **A priori** selecting units without bias: randomization

Selection bias

Now we have to get rid of the selection bias!

- **A priori** selecting units without bias: randomization
- **A posteriori** assessing the bias and extract it:

Selection bias

Now we have to get rid of the selection bias!

- **A priori** selecting units without bias: randomization
- **A posteriori** assessing the bias and extract it: Rubin's contribution

Why not just compare?

Consider the fate of young mothers

https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)31411-8/fulltext

The gold standard

Randomization

Randomization is the gold standard. This requires

Randomization

Randomization is the gold standard. This requires

- manipulation

Randomization

Randomization is the gold standard. This requires

- manipulation $\rightarrow$ experiments

Randomization

Randomization is the gold standard. This requires

- manipulation $\rightarrow$ experiments
- a sufficient number of units (LLN)

Randomization

Randomization is the gold standard. This requires

- manipulation $\rightarrow$ experiments
- a sufficient number of units (LLN) $\rightarrow$ statistical power

$\Rightarrow$ *Randomization eliminates bias*

Checking on observables

Even when we randomize, we check for signs of selection bias

Checking on observables

Even when we randomize, we check for signs of selection bias

- we cannot observe the bias

Checking on observables

Even when we randomize, we check for signs of selection bias

- we cannot observe the bias
- but we can check the balance of possible correlates (of bias)

Checking on observables

Even when we randomize, we check for signs of selection bias

- we cannot observe the bias
- but we can check the balance of possible correlates (of bias)

$\Rightarrow$ *Here comes the social science theories back in!*

Checking on observables

Even when we randomize, we check for signs of selection bias

Checking on observables

Even when we randomize, we check for signs of selection bias

$\Rightarrow$ *We verify the balance of pre-treatment variables*

The post hoc fixes