SciencesPo
DEPARTMENT OF ECONOMICS

ScPoEconometrics

Intro To Causality

Florian Oswald, Gustave Kenedi and Pierre Villedieu
SciencesPo Paris
2021-02-24

Quick "Quiz" on Last Week's Material

1. From your ***computer*** 👉 connect to ***www.wooclap.com/SCPOSLR***

   OR

2. From your ***phone*** 👉 flash QR code below

Today - Introduction to Causal Inference

- ***Causality*** versus ***correlation***
- The ***Potential Outcome Framework*** a.k.a. Rubin's Causal Model
- ***Randomized controlled trials*** (RCTs)
- Follow up on the empirical application of *class size* and *student performance*

Why do we care about causality?

Many of the *interesting questions* we might want to answer with data are causal

Why do we care about causality?

Many of the *interesting questions* we might want to answer with data are causal

- *Understanding* the world
  - *Social sciences*: Why do people behave in the way they do?
  - *Health sciences*: Why do people get sick? Which medicine can cure them?

Why do we care about causality?

Many of the *interesting questions* we might want to answer with data are causal

- ***Understanding*** the world
  - *Social sciences*: Why do people behave in the way they do?
  - *Health sciences*: Why do people get sick? Which medicine can cure them?
- Causal understanding is also of first interest to **policymakers**
  - How to lower unemployment?
  - How to improve student learning?
  - Whether governments should care about the level of public debt?

Why do we care about causality?

Many of the *interesting questions* we might want to answer with data are causal

- *Understanding* the world
    - *Social sciences*: Why do people behave in the way they do?
    - *Health sciences*: Why do people get sick? Which medicine can cure them?
- Causal understanding is also of first interest to **policymakers**
    - How to lower unemployment?
    - How to improve student learning?
    - Whether governments should care about the level of public debt?
- Note that some questions we might want to answer are not causal
    - Most *Artificial Intelligence* tasks only care about **prediction**
    - *Example*: predicting whether a photo is of a dog or a cat is vital to how Google Images works, but it doesn't care what *caused* the photo to be of a dog or a cat.

Causality and Economics

- Making causal inference from data can be seen as economists' *comparative advantage* among the social sciences!
- Plenty of fields do statistics. But very few make it standard training for their students to understand causality.
- Economists' endeavour to establish causal relationships is also what makes them useful both in the private (e.g. tech companies) and public sector (e.g. policy advisors).

Causality and Economics

- Making causal inference from data can be seen as economists' *comparative advantage* among the social sciences!
- Plenty of fields do statistics. But very few make it standard training for their students to understand causality.
- Economists' endeavour to establish causal relationships is also what makes them useful both in the private (e.g. tech companies) and public sector (e.g. policy advisors).
- Ok, that's enough preaching 😅

The Concept of Causality

Causality: what are we talking about?

- We say that `X` *causes* `Y`

The Concept of Causality

Causality: what are we talking about?

- We say that `X` *causes* `Y`
  - if we were to intervene and *change* the value of `X` ***without changing anything else***...

The Concept of Causality

Causality: what are we talking about?

- We say that `X` *causes* `Y`
  - if we were to intervene and *change* the value of `X` ***without changing anything else***...
  - then `Y` would also change ***as a result***.

The Concept of Causality

Causality: what are we talking about?

- We say that `X` *causes* `Y`
  - if we were to intervene and *change* the value of `X` ***without changing anything else***...
  - then `Y` would also change ***as a result***.
- The key point here is the ***without changing anything else***, often referred as the **ceteris paribus (*all else equal*) assumption**.
  (*latin* makes things seem more complicated 🤓)

The Concept of Causality

Causality: what are we talking about?

- We say that `X` *causes* `Y`
  - if we were to intervene and *change* the value of `X` ***without changing anything else***...
  - then `Y` would also change ***as a result***.
- The key point here is the ***without changing anything else***, often referred as the **ceteris paribus (*all else equal*) assumption**.
  (*latin* makes things seem more complicated 🤓)
- ⚠️ It does **NOT** mean that `X` is the only factor that causes `Y`.

Correlation vs Causation

Correlation does not equal causation has become a ubiquitous mantra, but can you tell why it is true?

Correlation vs Causation

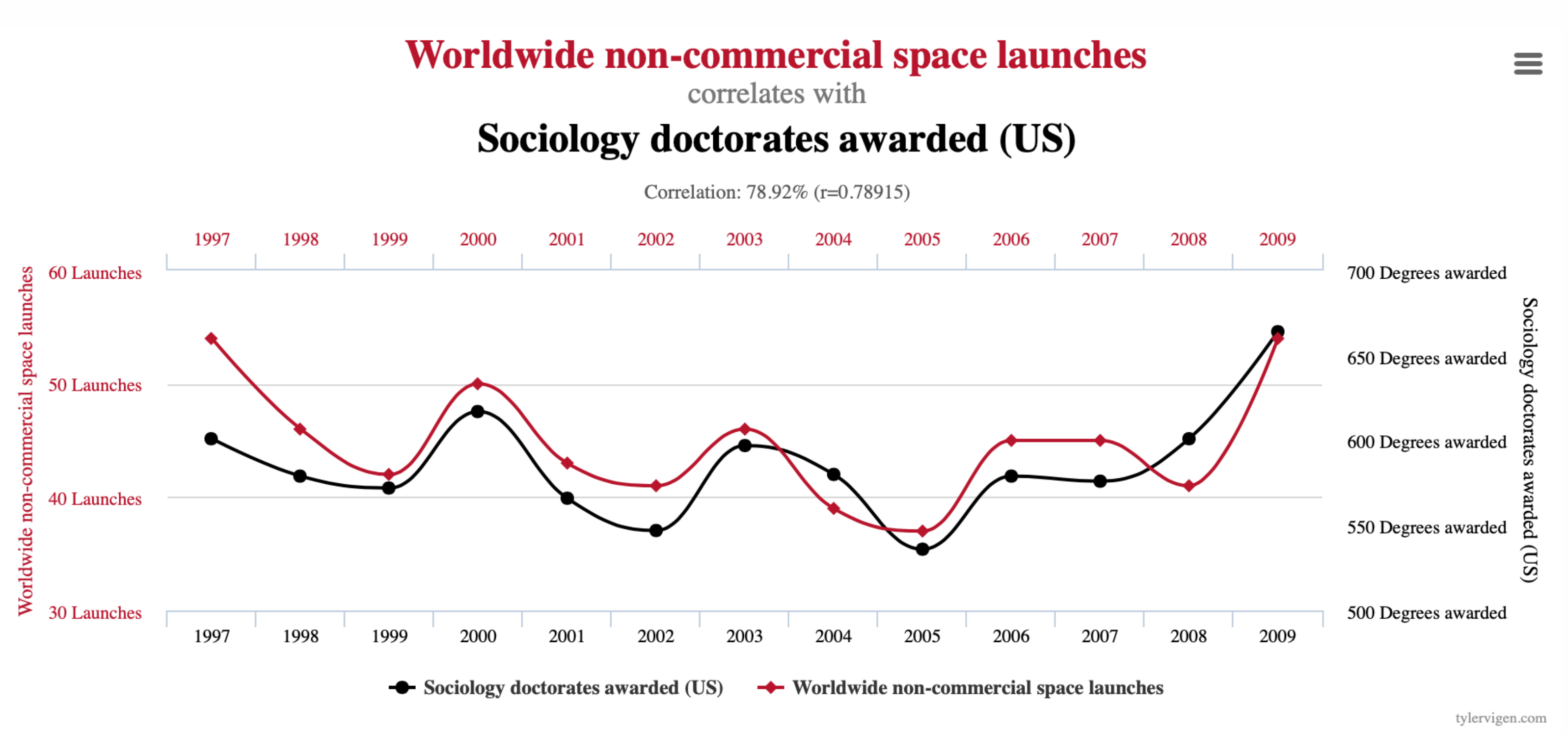
Correlation does not equal causation has become a ubiquitous mantra, but can you tell why it is true?

Some correlations obviously don't imply causation (e.g. spurious correlation website).

Correlation vs Causation

Correlation does not equal causation has become a ubiquitous mantra, but can you tell why it is true?

Some correlations obviously don't imply causation (e.g. spurious correlation website).



Correlation vs Causation: Smoking and Lung Cancer

But not all correlations are so easy to rule out

Correlation vs Causation: Smoking and Lung Cancer

But not all correlations are so easy to rule out

Does smoking cause lung cancer?

Correlation vs Causation: Smoking and Lung Cancer

But not all correlations are so easy to rule out

Does smoking cause lung cancer?

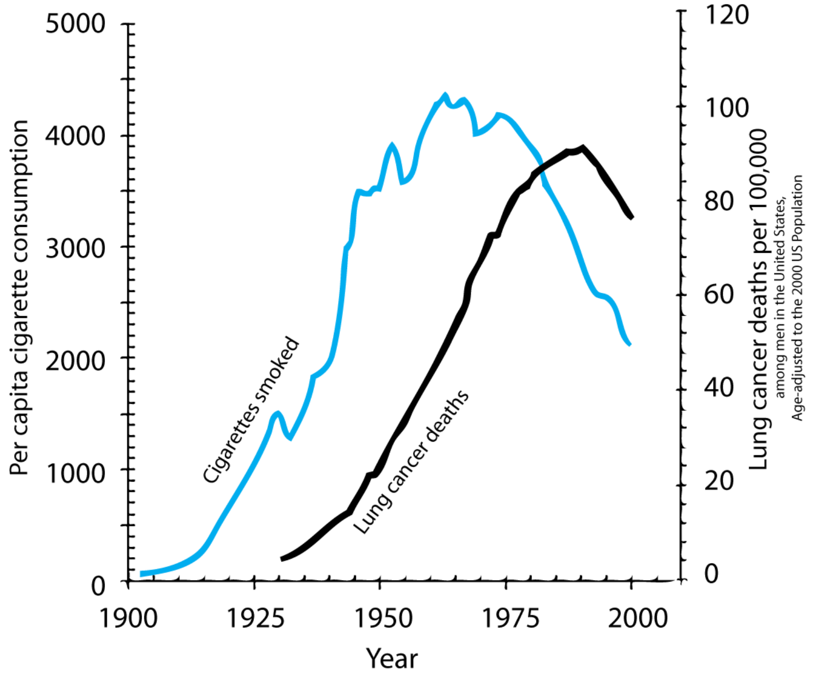
- Today, we know the answer is *YES*!
- But let's go back in the 1950's
  - We are at the start of a big increase in deaths from lung cancer...
  - ... which is happening after a fast growth in cigarette consumption

Correlation vs Causation: Smoking and Lung Cancer

But not all correlations are so easy to rule out

Does smoking cause lung cancer?

- Today, we know the answer is *YES*!
- But let's go back in the 1950's
    - We are at the start of a big increase in deaths from lung cancer...
    - ... which is happening after a fast growth in cigarette consumption

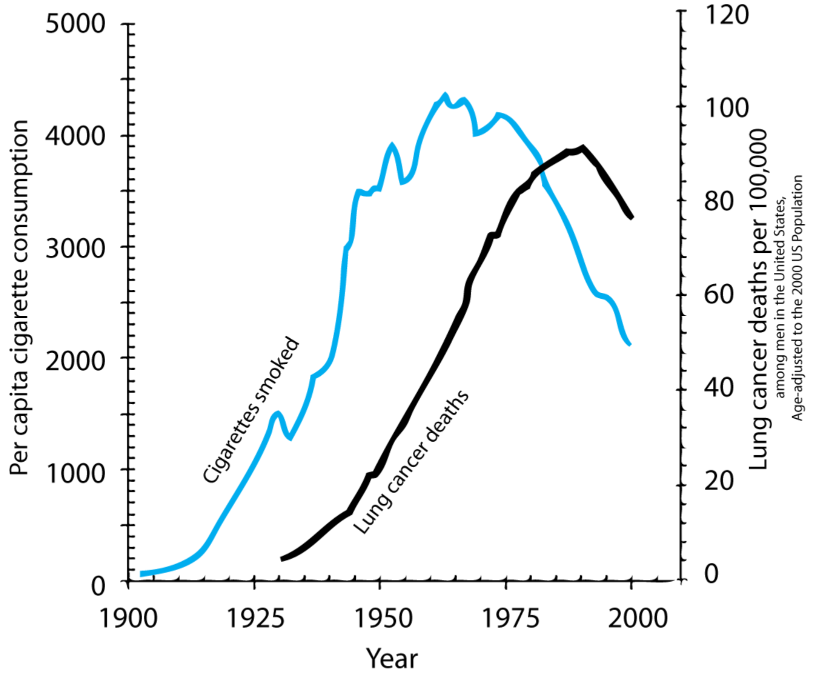


Correlation vs Causation: Smoking and Lung Cancer

But not all correlations are so easy to rule out

Does smoking cause lung cancer?

- Today, we know the answer is *YES*!
- But let's go back in the 1950's
  - We are at the start of a big increase in deaths from lung cancer...
  - ... which is happening after a fast growth in cigarette consumption



- It's very tempting to claim that smoking causes lung cancer based on this graph.

Correlation vs Causation: Smoking and Lung Cancer

At the time many people were still skeptical, including some famous statisticians:

Correlation vs Causation: Smoking and Lung Cancer

At the time many people were still skeptical, including some famous statisticians:

Macro confouding factors:

Other macro factors which can cause cancers also changed between 1900 and 1950:

- Tarring of roads,
- Inhalation of motor exhausts (leaded gasoline fumes),
- General greater air pollution.

Correlation vs Causation: Smoking and Lung Cancer

At the time many people were still skeptical, including some famous statisticians:

Macro confouding factors:

Other macro factors which can cause cancers also changed between 1900 and 1950:

- Tarring of roads,
- Inhalation of motor exhausts (leaded gasoline fumes),
- General greater air pollution.

Self selection:

Smokers and non-smokers may be different in the first place:

- **Selection on observable characteristics**: age, education, income, etc.
- **Selection on unobservable characteristics**: genes (the hypothetical confounding genome theory of Fisher).

How Can We Tell?

- Sometimes correlations are just pure *artefacts*: there is no causal relationship between the variables of interest.
- In some other cases, there are both correlation and causality but not of the same **magnitude**, or even the same **direction**.

How Can We Tell?

- Sometimes correlations are just pure *artefacts*: there is no causal relationship between the variables of interest.
- In some other cases, there are both correlation and causality but not of the same **magnitude**, or even the same **direction**.
- So how can we make causal inference then?
- The ***Potential Outcomes Framework*** will be our guide.

Causal Inference

The Potential Outcomes Framework

Often called the **Rubin Causal Model** in memory of the statistician **Donald Rubin** who generalised and formalized this model in the 1970's.

The Potential Outcomes Framework

Often called the **Rubin Causal Model** in memory of the statistician **Donald Rubin** who generalised and formalized this model in the 1970's.

Key idea: Each individual can be exposed to **multiple alternative treatment states**.

- smoking cigarrettes, smoking cigars or not smoking,
- growing up in a poor vs a middle class neighborhood vs a rich neighborhood,
- being in a small or a big class.

The Potential Outcomes Framework

Often called the **Rubin Causal Model** in memory of the statistician **Donald Rubin** who generalised and formalized this model in the 1970's.

Key idea: Each individual can be exposed to **multiple alternative treatment states**.

- smoking cigarrettes, smoking cigars or not smoking,
- growing up in a poor vs a middle class neighborhood vs a rich neighborhood,
- being in a small or a big class.

For practicality, let this treatment variable D_i be a binary variable:

$$D_i = \begin{cases} 1 \text{ if individual } i \text{ is treated} \\ 0 \text{ if individual } i \text{ is not treated} \end{cases}$$

The Potential Outcomes Framework

Often called the **Rubin Causal Model** in memory of the statistician **Donald Rubin** who generalised and formalized this model in the 1970's.

Key idea: Each individual can be exposed to **multiple alternative treatment states**.

- smoking cigarrettes, smoking cigars or not smoking,
- growing up in a poor vs a middle class neighborhood vs a rich neighborhood,
- being in a small or a big class.

For practicality, let this treatment variable D_i be a binary variable:

$$D_i = \begin{cases} 1 \text{ if individual } i \text{ is treated} \\ 0 \text{ if individual } i \text{ is not treated} \end{cases}$$

Treatment group
all the individuals such that $D_i = 1$.

Control group
all the individuals such that $D_i = 0$.

The Potential Outcomes Framework

- In this framework, each individual has two ***potential outcomes***, but only one ***observed outcome*** Y_i:
    - Y_i^1: *potential outcome if individual i receives the treatment* $(D_i = 1)$,
    - Y_i^0: *potential outcome if individual i does not receive the treatment* $(D_i = 0)$.

The Potential Outcomes Framework

- In this framework, each individual has two ***potential outcomes***, but only one ***observed outcome*** Y_i:
  - Y_i^1: *potential outcome if individual i receives the treatment* $(D_i = 1)$,
  - Y_i^0: *potential outcome if individual i does not receive the treatment* $(D_i = 0)$.
- From these we can define the ***individual treatment effect*** δ_i:

$$\delta_i = Y_i^1 - Y_i^0$$

- δ_i measures the **causal effect of the treatment** (D_i) on outcome Y for individual i.

The Potential Outcomes Framework

- In real life we only observe Y_i which can be written as:

$$Y_i = D_i * Y_i^1 + (1 - D_i) * Y_i^0$$

The Potential Outcomes Framework

- In real life we only observe Y_i which can be written as:

$$Y_i = D_i * Y_i^1 + (1 - D_i) * Y_i^0$$

- ***Fundamental Problem of Causal Inference***: for any individual i, we only observe one of either potential outcomes, and thus we cannot compute δ_i (Holland, 1986).

The Potential Outcomes Framework

- In real life we only observe Y_i which can be written as:

$$Y_i = D_i * Y_i^1 + (1 - D_i) * Y_i^0$$

- ***Fundamental Problem of Causal Inference***: for any individual i, we only observe one of either potential outcomes, and thus we cannot compute δ_i (Holland, 1986).

- The potential outcome that is not observed exists in principle, it is called the ***counterfactual outcome***.

The Potential Outcomes Framework

- In real life we only observe Y_i which can be written as:

$$Y_i = D_i * Y_i^1 + (1 - D_i) * Y_i^0$$

- ***Fundamental Problem of Causal Inference***: for any individual i, we only observe one of either potential outcomes, and thus we cannot compute δ_i (Holland, 1986).

- The potential outcome that is not observed exists in principle, it is called the ***counterfactual outcome***.

Group	Y_i^1	Y_i^0
Treatment group $(D_i = 1)$	Observable as Y_i	Counterfactual
Control group $(D_i = 0)$	Counterfactual	Observable as Y_i

The Potential Outcomes Framework

- In real life we only observe Y_i which can be written as:

$$Y_i = D_i * Y_i^1 + (1 - D_i) * Y_i^0$$

- ***Fundamental Problem of Causal Inference***: for any individual i, we only observe one of either potential outcomes, and thus we cannot compute δ_i (Holland, 1986).

- The potential outcome that is not observed exists in principle, it is called the ***counterfactual outcome***.

Group	Y_i^1	Y_i^0
Treatment group $(D_i = 1)$	Observable as Y_i	Counterfactual
Control group $(D_i = 0)$	Counterfactual	Observable as Y_i

- Since the treatment effect *cannot* be observed at the individual level, we estimate population averages.

Average Treatment Effect (ATE)

Broadest possible average effect: **A**verage **T**reatment **E**ffect (***ATE***)

$$
\begin{aligned}
ATE &= \mathbb{E}(\delta_i) \\
&= \mathbb{E}(Y_i^1 - Y_i^0) \\
&= \mathbb{E}(Y_i^1) - \mathbb{E}(Y_i^0)
\end{aligned}
$$

- The ATE simply measures the ***average of individual treatment effects over the whole population***.
  - The $\mathbb{E}(.)$ operator stands for **expectation** or *population mean*.
  - The $\mathbb{E}(.)$ operator is linear, in other words, $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ with X and Y being two random variables.

Average Treatment on the Treated (ATT)

Other ***conditional*** average treatment effects may be of interest:

- The **A**verage **T**reatment Effect on the **T**reated (***ATT***)

$$\begin{aligned} ATT &= \mathbb{E}(\delta_i | D_i = 1) \\ &= \mathbb{E}(Y_i^1 - Y_i^0 | D_i = 1) \\ &= \mathbb{E}(Y_i^1 | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 1) \end{aligned}$$

- The ATT measures the ***average treatment effect conditional on being in the treatment group***.
  - The $\mathbb{E}(. | D = x)$ operator stands for **conditional expectation**. It refers to the expectation over a subcategory of the entire population, namely people who satisfy the condition $D = x$.
  - The $\mathbb{E}(. | D = x)$ operator is also linear.

Average Treatment on the Untreated (ATU)

Other ***conditional*** average treatment effects may be of interest:

- The **A**verage **T**reatment Effect on the **U**ntreated (***ATU***)

$$\begin{aligned} ATU &= \mathbb{E}(\delta_i | D_i = 0) \\ &= \mathbb{E}(Y_i^1 - Y_i^0 | D_i = 0) \\ &= \mathbb{E}(Y_i^1 | D_i = 0) - \mathbb{E}(Y_i^0 | D_i = 0) \end{aligned}$$

- The ATU measures the ***average treatment effect conditional on being in the control group***.

Average Treatment on the Untreated (ATU)

Other ***conditional*** average treatment effects may be of interest:

- The **A**verage **T**reatment Effect on the **U**ntreated (***ATU***)

$$
\begin{aligned}
ATU &= \mathbb{E}(\delta_i | D_i = 0) \\
&= \mathbb{E}(Y_i^1 - Y_i^0 | D_i = 0) \\
&= \mathbb{E}(Y_i^1 | D_i = 0) - \mathbb{E}(Y_i^0 | D_i = 0)
\end{aligned}
$$

- The ATU measures the ***average treatment effect conditional on being in the control group***.
- In the majority of cases, ATE $\neq$ ATT $\neq$ ATU!

Example

Example: Potential outcomes for 10 students of being in a class of 7 students (Y^1) or 15 students (Y^0) on GPA (0-10)

Student	Y^1	Y^0	δ
1	5	2	3
2	6	4	2
3	3	6	-3
4	5	4	1
5	10	8	2
6	2	4	-2
7	5	2	3
8	6	4	2
9	2	9	-7
10	8	2	6

Example

Example: Potential outcomes for 10 students of being in a class of 7 students (Y^1) or 15 students (Y^0) on GPA (0-10)

Student	Y^1	Y^0	δ
1	5	2	3
2	6	4	2
3	3	6	-3
4	5	4	1
5	10	8	2
6	2	4	-2
7	5	2	3
8	6	4	2
9	2	9	-7
10	8	2	6

$$\begin{aligned} \text{ATE} &= \mathbb{E}(Y^1) - \mathbb{E}(Y^0) \\ &= 5.2 - 4.5 \\ &= 0.7 \end{aligned}$$

$\rightarrow$ the ***average*** causal effect of being in small relative to large class on GPA is 0.7 points.

⚠️ not all students benefited equally from the treatment!

Example

Now, imagine a benevolent and omniscient school director assigns students to the treatment that maximizes their GPA

Example

Now, imagine a benevolent and omniscient school director assigns students to the treatment that maximizes their GPA

Student	Y	D	δ
1	5	1	3
2	6	1	2
3	6	0	-3
4	5	1	1
5	10	1	2
6	4	0	-2
7	5	1	3
8	6	1	2
9	9	0	-7
10	8	1	6

Example

Now, imagine a benevolent and omniscient school director assigns students to the treatment that maximizes their GPA

Student	Y	D	δ
1	5	1	3
2	6	1	2
3	6	0	-3
4	5	1	1
5	10	1	2
6	4	0	-2
7	5	1	3
8	6	1	2
9	9	0	-7
10	8	1	6

$$\begin{aligned} \text{ATT} &= \mathbb{E}(\delta|D=1) \\ &\approx 2.71 \end{aligned}$$

$$\begin{aligned} \text{ATU} &= \mathbb{E}(\delta|D=0) \\ &= -4 \end{aligned}$$

$$\begin{aligned} \text{ATE} &= 7/10 \times \text{ATT} + 3/10 \times \text{ATU} \\ &= 7/10 \times 2.71 + 3/10 \times (-4) \\ &= 0.7 \end{aligned}$$

The Problem of Causal Inference

- In practice, we have the same **missing data problem** for computing the ATE, ATT or ATU as we did for δ_i. Either Y_i^1 or Y_i^0 is missing for each i.

The Problem of Causal Inference

- In practice, we have the same **missing data problem** for computing the ATE, ATT or ATU as we did for δ_i. Either Y_i^1 or Y_i^0 is missing for each i.

- From the data, we can compute the **S**imple **D**ifference in mean **O**utcomes (***SDO***) for both groups:

$$
\begin{aligned}
SDO &= \mathbb{E}(Y_i^1 | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0) \\
&= \frac{1}{N_T} \sum_{i=1}^{N_T} (Y_i | D_i = 1) - \frac{1}{N_C} \sum_{i=1}^{N_C} (Y_i | D_i = 0)
\end{aligned}
$$

The Problem of Causal Inference

- In practice, we have the same **missing data problem** for computing the ATE, ATT or ATU as we did for δ_i. Either Y_i^1 or Y_i^0 is missing for each i.

- From the data, we can compute the **S**imple **D**ifference in mean **O**utcomes (***SDO***) for both groups:

$$
\begin{aligned}
SDO &= \mathbb{E}(Y_i^1 | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0) \\
&= \frac{1}{N_T} \sum_{i=1}^{N_T} (Y_i | D_i = 1) - \frac{1}{N_C} \sum_{i=1}^{N_C} (Y_i | D_i = 0)
\end{aligned}
$$

- From our example, we obtain: $SDO \approx 6.43 - 6.33 \approx 0.1$ (much smaller than the 0.7 ATE)

The Problem of Causal Inference

- In practice, we have the same **missing data problem** for computing the ATE, ATT or ATU as we did for δ_i. Either Y_i^1 or Y_i^0 is missing for each i.

- From the data, we can compute the **S**imple **D**ifference in mean **O**utcomes (***SDO***) for both groups:

$$SDO = \mathbb{E}(Y_i^1 | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0)$$
$$= \frac{1}{N_T} \sum_{i=1}^{N_T} (Y_i | D_i = 1) - \frac{1}{N_C} \sum_{i=1}^{N_C} (Y_i | D_i = 0)$$

- From our example, we obtain: $SDO \approx 6.43 - 6.33 \approx 0.1$ (much smaller than the 0.7 ATE)

- Almost always, such a difference **will fail to capture the causal treatment effect**.

The Problem of Causal Inference

- In practice, we have the same **missing data problem** for computing the ATE, ATT or ATU as we did for δ_i. Either Y_i^1 or Y_i^0 is missing for each i.

- From the data, we can compute the **S**imple **D**ifference in mean **O**utcomes (***SDO***) for both groups:

$$\begin{aligned} SDO &= \mathbb{E}(Y_i^1 | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0) \\ &= \frac{1}{N_T} \sum_{i=1}^{N_T} (Y_i | D_i = 1) - \frac{1}{N_C} \sum_{i=1}^{N_C} (Y_i | D_i = 0) \end{aligned}$$

- From our example, we obtain: $SDO \approx 6.43 - 6.33 \approx 0.1$ (much smaller than the 0.7 ATE)

- Almost always, such a difference **will fail to capture the causal treatment effect**.

- Notice that this kind "naive" comparison is often done by journalists, politicians, badly trained scientists (but not you now! 😉).

Problems with Naive Comparisons

Let's rewrite the SDO to make the individual treatment effect (δ_i) appear in the equation.

$$\begin{aligned} SDO &= \mathbb{E}(Y_i^1 | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0) \\ &= \mathbb{E}(Y_i^0 + \delta_i | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0) \end{aligned}$$

Problems with Naive Comparisons

Let's rewrite the SDO to make the individual treatment effect (δ_i) appear in the equation.

$$\begin{aligned} SDO &= \mathbb{E}(Y_i^1 | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0) \\ &= \mathbb{E}(Y_i^0 + \delta_i | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0) \end{aligned}$$

For now, suppose **treatment effect is constant** across people: for all i, $\delta_i = \delta$.

Problems with Naive Comparisons

Let's rewrite the SDO to make the individual treatment effect (δ_i) appear in the equation.

$$\begin{aligned} SDO &= \mathbb{E}(Y_i^1 | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0) \\ &= \mathbb{E}(Y_i^0 + \delta_i | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0) \end{aligned}$$

For now, suppose **treatment effect is constant** across people: for all i, $\delta_i = \delta$.

Then,

$$SDO = \delta + \mathbb{E}(Y_i^0 | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0)$$

And because $ATE = \mathbb{E}(\delta_i) = \mathbb{E}(\delta) = \delta$ (by assumption), we get:

$$SDO = ATE + \underbrace{\mathbb{E}(Y_i^0 | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0)}_{\text{Selection bias}}$$

Problems with Naive Comparisons

Let's now relax the assumption that the treatment effect is constant among all individuals.

After some tedious calculations that we skip, the SDO can now be decomposed as:

$$SDO = ATE + \underbrace{\mathbb{E}(Y_i^0|D_i = 1) - \mathbb{E}(Y_i^0|D_i = 0)}_{\text{Selection bias}} + \underbrace{(1 - \pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}$$

where $1 - \pi$ denotes the share of people in the control group.

Problems with Naive Comparisons

Let's now relax the assumption that the treatment effect is constant among all individuals.

After some tedious calculations that we skip, the SDO can now be decomposed as:

$$SDO = ATE + \underbrace{\mathbb{E}(Y_i^0 | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0)}_{\text{Selection bias}}$$
$$+ \underbrace{(1 - \pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}$$

where $1 - \pi$ denotes the share of people in the control group.

So there is a novel source of bias that comes from the potential ***heterogeneity in the individual treatment effect*** δ_i.

Problems with Naive Comparisons

Let's now relax the assumption that the treatment effect is constant among all individuals.

After some tedious calculations that we skip, the SDO can now be decomposed as:

$$SDO = ATE + \underbrace{\mathbb{E}(Y_i^0|D_i = 1) - \mathbb{E}(Y_i^0|D_i = 0)}_{\text{Selection bias}}$$
$$+ \underbrace{(1-\pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}$$

where $1-\pi$ denotes the share of people in the control group.

So there is a novel source of bias that comes from the potential ***heterogeneity in the individual treatment effect*** δ_i.

- ***Selection bias***: those who attend university are likely to have higher baseline cognitive skills (regardless of whether they actually attend college).
- ***Heterogeneous treatment effect bias***: those who attend university may improve their cognitive skills more at university because they are more motivated.

Task 1: SDO, ATE and Randomization

10:00

Let's compute these various quantities and biases with some toy data (i.e. data we generated ourselves).

1. Load the data here. The `group` variable corresponds to whether the individual has been treated or not, the `Y0` variable corresponds to the potential outcome if the individual does not receive the treatment (Y_i^0) while `Y1` corresponds to the potential outcome if the individual receives the treatment (Y_i^1). Create a new variable containing the observed outcome (Y_i) and the individual treatment effect (δ_i). (Recall that $Y_i = D_i * Y_i^1 + (1 - D_i)Y_i^0$ and $\delta_i = Y_i^1 - Y_i^0$.)

2. Compute the ***ATE*** and the ***SDO***. Is there is any *bias*? Is it large in magnitude?

3. In this new dataset we've randomly assigned the same individuals to the treatment and control groups. Compute the ***SDO under randomization***. Remember that you need to recompute Y_i because the assignment is not the same anymore. If you got it right, the bias should be very close to 0. Why is it not exactly 0?

4. *To do at home*: Compute the value of the ***selection bias*** and the ***heterogenous treatment effect bias*** and check that we have

Randomization solves the problem of causal inference!

- ***Randomized experiments***: you ***randomly*** assign people to a treatment and a control group.
- In this case, the treatment assignement is **independent** of the potential outcomes.

Randomization solves the problem of causal inference!

- ***Randomized experiments***: you ***randomly*** assign people to a treatment and a control group.
- In this case, the treatment assignement is **independent** of the potential outcomes.
- In particular, there is no reason for $\mathbb{E}(Y_i^0|D_i = 1)$ to be different from $\mathbb{E}(Y_i^0|D_i = 0)$
  - Therefore the ***selection bias is equal to 0***.

Randomization solves the problem of causal inference!

- ***Randomized experiments***: you ***randomly*** assign people to a treatment and a control group.
- In this case, the treatment assignement is **independent** of the potential outcomes.
- In particular, there is no reason for $\mathbb{E}(Y_i^0|D_i = 1)$ to be different from $\mathbb{E}(Y_i^0|D_i = 0)$
    - Therefore the ***selection bias is equal to 0***.
- In the same way, there is no reason for $\mathbb{E}[\delta_i]$ to be different in the treatment and control group.
    - There $ATT = ATU$, implying the ***heterogenous treatment effect bias will also be 0***.

Randomization solves the problem of causal inference!

With random assignment we have:

$$SDO = \mathbb{E}(Y_i^1 | D_i = 1) - \mathbb{E}(Y_i^0 | D_i = 0) = ATE$$

👉 We can directly estimate the ATE from the data!

Randomized Experiments

Randomized Experiments

- Often called **R**andomized **C**ontrolled **T**rials (RCT).
- The first RCTs were conducted a long time ago (18th and 19th century), mainly in **Medecine**.
- In the beginning of the 20th century they were popularized by famous statisticians like **J. Neyman** or **R.A. Fisher**.
- Since then they have had a growing influence and have progressively become a reliable tool for public policy evaluation.
- As for economics, the **2019 Nobel Price in Economics** was awarded to three exponents of RCTs, Abhijit Banerjee, Esther Duflo and Michael Kremer, "for their experimental approach to alleviating global poverty".

Back to class size and students' achievement

Last week we regressed average student math or reading scores on class size.

$$\text{math score}_i = b_0 + b_1 \text{class size}_i + e_i$$

We briefly discussed why b_1^{OLS} could only establish an ***association*** and not a ***causal relationship***.

Back to class size and students' achievement

Last week we regressed average student math or reading scores on class size.

$$\text{math score}_i = b_0 + b_1 \text{class size}_i + e_i$$

We briefly discussed why b_1^{OLS} could only establish an ***association*** and not a ***causal relationship***.

- **Student sorting**: There is selection into schools with different sized classes. Suppose parents have a prior that smaller classes are better, they will try to get their kids into those schools.

Back to class size and students' achievement

Last week we regressed average student math or reading scores on class size.

$$\text{math score}_i = b_0 + b_1 \text{class size}_i + e_i$$

We briefly discussed why b_1^{OLS} could only establish an ***association*** and not a ***causal relationship***.

- **Student sorting**: There is selection into schools with different sized classes. Suppose parents have a prior that smaller classes are better, they will try to get their kids into those schools.

- **Teacher sorting**: Teachers may sort in schools with smaller classes because it's easier to teach a small rather than a large class, and if there is competition for those places then higher quality teachers will have an advantage.

Back to class size and students' achievement

Last week we regressed average student math or reading scores on class size.

$$\text{math score}_i = b_0 + b_1 \text{class size}_i + e_i$$

We briefly discussed why b_1^{OLS} could only establish an ***association*** and not a ***causal relationship***.

- **Student sorting**: There is selection into schools with different sized classes. Suppose parents have a prior that smaller classes are better, they will try to get their kids into those schools.
- **Teacher sorting**: Teachers may sort in schools with smaller classes because it's easier to teach a small rather than a large class, and if there is competition for those places then higher quality teachers will have an advantage.
- **Location effect**: Large classes may be more common in wealthier and bigger cities, while small classes may be more likely in poorer rural areas.

Back to class size and students' achievement

Last week we regressed average student math or reading scores on class size.

$$\text{math score}_i = b_0 + b_1 \text{class size}_i + e_i$$

We briefly discussed why b_1^{OLS} could only establish an ***association*** and not a ***causal relationship***.

- **Student sorting**: There is selection into schools with different sized classes. Suppose parents have a prior that smaller classes are better, they will try to get their kids into those schools.
- **Teacher sorting**: Teachers may sort in schools with smaller classes because it's easier to teach a small rather than a large class, and if there is competition for those places then higher quality teachers will have an advantage.
- **Location effect**: Large classes may be more common in wealthier and bigger cities, while small classes may be more likely in poorer rural areas.

An RCT would take care of all these biases!

The Project STAR Experiment

Tennessee **S**tudent/**T**eacher **A**chievement **R**atio Experiment (see Krueger (1999))

- Funded by Tennesse legislature for a total cost of approx. $12 million.
- The experiment started in the 1985-1986 school year and lasted four years.

The Project STAR Experiment

Tennessee **S**tudent/**T**eacher **A**chievement **R**atio Experiment (see Krueger (1999))

- Funded by Tennesse legislature for a total cost of approx. $12 million.
- The experiment started in the 1985-1986 school year and lasted four years.
- 11,600 students and their teachers where **randomly assigned** to one of the following 3 groups from kindergarten through third grade:
    1. ***Small class***: 13-17 students per teacher,
    2. ***Regular class***: 22-25 students,
    3. ***Regular/aide class***: 22-25 students with a full-time teacher's *aide*.

The Project STAR Experiment

Tennessee **S**tudent/**T**eacher **A**chievement **R**atio Experiment (see Krueger (1999))

- Funded by Tennesse legislature for a total cost of approx. $12 million.
- The experiment started in the 1985-1986 school year and lasted four years.
- 11,600 students and their teachers where **randomly assigned** to one of the following 3 groups from kindergarten through third grade:
    1. ***Small class***: 13-17 students per teacher,
    2. ***Regular class***: 22-25 students,
    3. ***Regular/aide class***: 22-25 students with a full-time teacher's *aide*.
- Randomization occurred within schools.
- Students' math and reading skills were tested around March each year.

The Project STAR Experiment

Tennessee **S**tudent/**T**eacher **A**chievement **R**atio Experiment (see Krueger (1999))

- Funded by Tennesse legislature for a total cost of approx. $12 million.
- The experiment started in the 1985-1986 school year and lasted four years.
- 11,600 students and their teachers where **randomly assigned** to one of the following 3 groups from kindergarten through third grade:
    1. ***Small class***: 13-17 students per teacher,
    2. ***Regular class***: 22-25 students,
    3. ***Regular/aide class***: 22-25 students with a full-time teacher's *aide*.
- Randomization occurred within schools.
- Students' math and reading skills were tested around March each year.
- There was a problem of ***non-random attrition*** but we will ignore it.

Task 2: STAR data

10:00

1. Load the *STAR* data from here and assign it to an object called `star_df`.
2. Read the help for `AER::STAR` to understand what the variables correspond to. (Note: the data has been *reshaped* so don't mind the "k", "1", etc. in the variable names in the help.)
3. What's the unit of observation? Which variable contains: (i) the (random) class assignment, (ii) the student's class grade, (iii) the outcomes of interest?
4. How many observations are there? Why so many?
5. Why are there so many `NA`s? What do they correspond to?
6. Keep only cases with no `NA`s with the following code:
   `star_df <- star_df[complete.cases(star_df),]`
7. Let's check how well the randomization was done by doing ***balancing checks***. Compute the average percentage of girls, african americans, and free lunch qualifiers by grade *and* treatment class. (*Hint*: The following computes the percentage of girls (without the relevant `dplyr` verbs): `share_female = mean(gender == "female") * 100`.)

The Project STAR Experiment

We just saw that in an RCT the Average Treatment Effect is obtained by computing the differences in outcomes between the treatment and control groups.

Let's only focus on:

- One treatment group: **small classes**,
- One control group: **regular classes**,
- One grade: **kindergarten** (k).

The Project STAR Experiment

We just saw that in an RCT the Average Treatment Effect is obtained by computing the differences in outcomes between the treatment and control groups.

Let's only focus on:

- One treatment group: **small classes**,
- One control group: **regular classes**,
- One grade: **kindergarten** (k).

grade	test	mean regular	mean small	ATE
k	math	484.45	493.34	8.9
k	read	435.76	441.13	5.37

What's the interpretation for these ATEs?

The Project STAR Experiment

We just saw that in an RCT the Average Treatment Effect is obtained by computing the differences in outcomes between the treatment and control groups.

Let's only focus on:

- One treatment group: **small classes**,
- One control group: **regular classes**,
- One grade: **kindergarten** (k).

grade	test	mean regular	mean small	ATE
k	math	484.45	493.34	8.9
k	read	435.76	441.13	5.37

What's the interpretation for these ATEs?

That's nice but can't we put this in regression form?

RCT in Regression Form

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

RCT in Regression Form

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

Factoring by D_i and replacing $Y_i^1 - Y_i^0$ by δ_i, we get:

$$\begin{aligned} Y_i &= Y_i^0 + D_i(Y_i^1 - Y_i^0) \\ &= Y_i^0 + D_i \delta_i \end{aligned}$$

RCT in Regression Form

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

Factoring by D_i and replacing $Y_i^1 - Y_i^0$ by δ_i, we get:

$$\begin{aligned} Y_i &= Y_i^0 + D_i (Y_i^1 - Y_i^0) \\ &= Y_i^0 + D_i \delta_i \end{aligned}$$

Assuming $\delta_i = \delta$, for all i,

$$Y_i = Y_i^0 + D_i \delta$$

RCT in Regression Form

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

Factoring by D_i and replacing $Y_i^1 - Y_i^0$ by δ_i, we get:

$$\begin{aligned} Y_i &= Y_i^0 + D_i(Y_i^1 - Y_i^0) \\ &= Y_i^0 + D_i \delta_i \end{aligned}$$

Assuming $\delta_i = \delta$, for all i,

$$Y_i = Y_i^0 + D_i \delta$$

Adding $\mathbb{E}[Y_i^0] - \mathbb{E}[Y_i^0] = 0$ to the right-hand side:

$$\begin{aligned} Y_i &= \mathbb{E}[Y_i^0] + D_i \delta + Y_i^0 - \mathbb{E}[Y_i^0] \\ &= b_0 + \delta D_i + e_i \end{aligned}$$

where $b_0 = \mathbb{E}[Y_i^0]$ and $e_i = Y_i^0 - \mathbb{E}[Y_i^0]$

The Project STAR Experiment: Regression

The last equation looks exactly like the simple regression model we saw last week! (with $\delta = b_1$)

Let's therefore estimate the ATE of being assigned to a small class size on math scores.

The Project STAR Experiment: Regression

The last equation looks exactly like the simple regression model we saw last week! (with $\delta = b_1$)

Let's therefore estimate the ATE of being assigned to a small class size on math scores.

We want to estimate the following model: $\text{math score}_i = b_0 + \delta \text{small}_i + e_i$, with

$$\text{small}_i = \begin{cases} 1 \text{ if assigned to a small class} \\ 0 \text{ if assigned to a regular class} \end{cases}$$

```
star_df_k_small <- star_df %>%
  filter(
    star %in% c("regular", "small") &
      grade == "k") %>%
  mutate(small = (star == "small"))
```

```
star_df_k_small %>% count(star, grade)

star_df_k_small %>% count(small)
```

```
## # A tibble: 2 x 3
##   star    grade     n
##   <chr>   <chr> <int>
## 1 regular k      1781
## 2 small   k      1578
```

```
## # A tibble: 2 x 2
##   small     n
##   <lgl> <int>
## 1 FALSE  1781
## 2 TRUE   1578
```

The Project STAR Experiment: Regression

Regression model we want to estimate: $\text{math score}_i = b_0 + \delta \text{small}_i + e_i$

```
lm(math ~ small, star_df_k_small)
```

```
##
## Call:
## lm(formula = math ~ small, data = star_df_k_small)
##
## Coefficients:
## (Intercept)    smallTRUE
##     484.446        8.895
```

The Project STAR Experiment: Regression

Regression model we want to estimate: math score$_i = b_0 + \delta \text{small}_i + e_i$

```
lm(math ~ small, star_df_k_small)
```

```
##
## Call:
## lm(formula = math ~ small, data = star_df_k_small)
##
## Coefficients:
## (Intercept)    smallTRUE
##     484.446        8.895
```

Recall that: $b_0 = \mathbb{E}[Y_i^0]$ and $\delta = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$

```
b_0 = mean(star_df_k_small$math[
    star_df_k_small$small == FALSE])
b_0
```

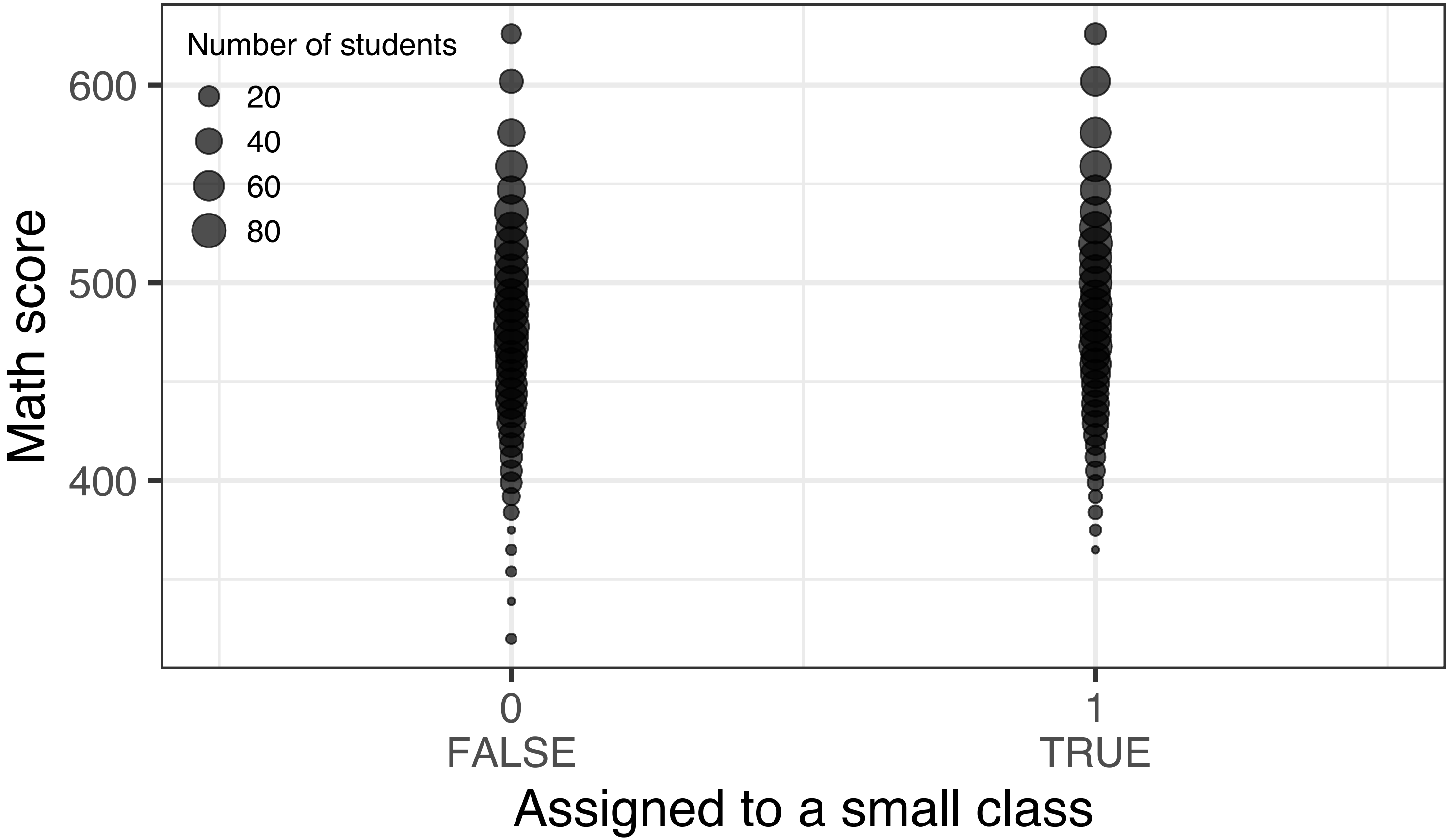
```
## [1] 484.4464
```

```
delta = mean(star_df_k_small$math[
    star_df_k_small$small == TRUE]) -
  mean(star_df_k_small$math[
    star_df_k_small$small == FALSE])
delta
```

```
## [1] 8.895193
```

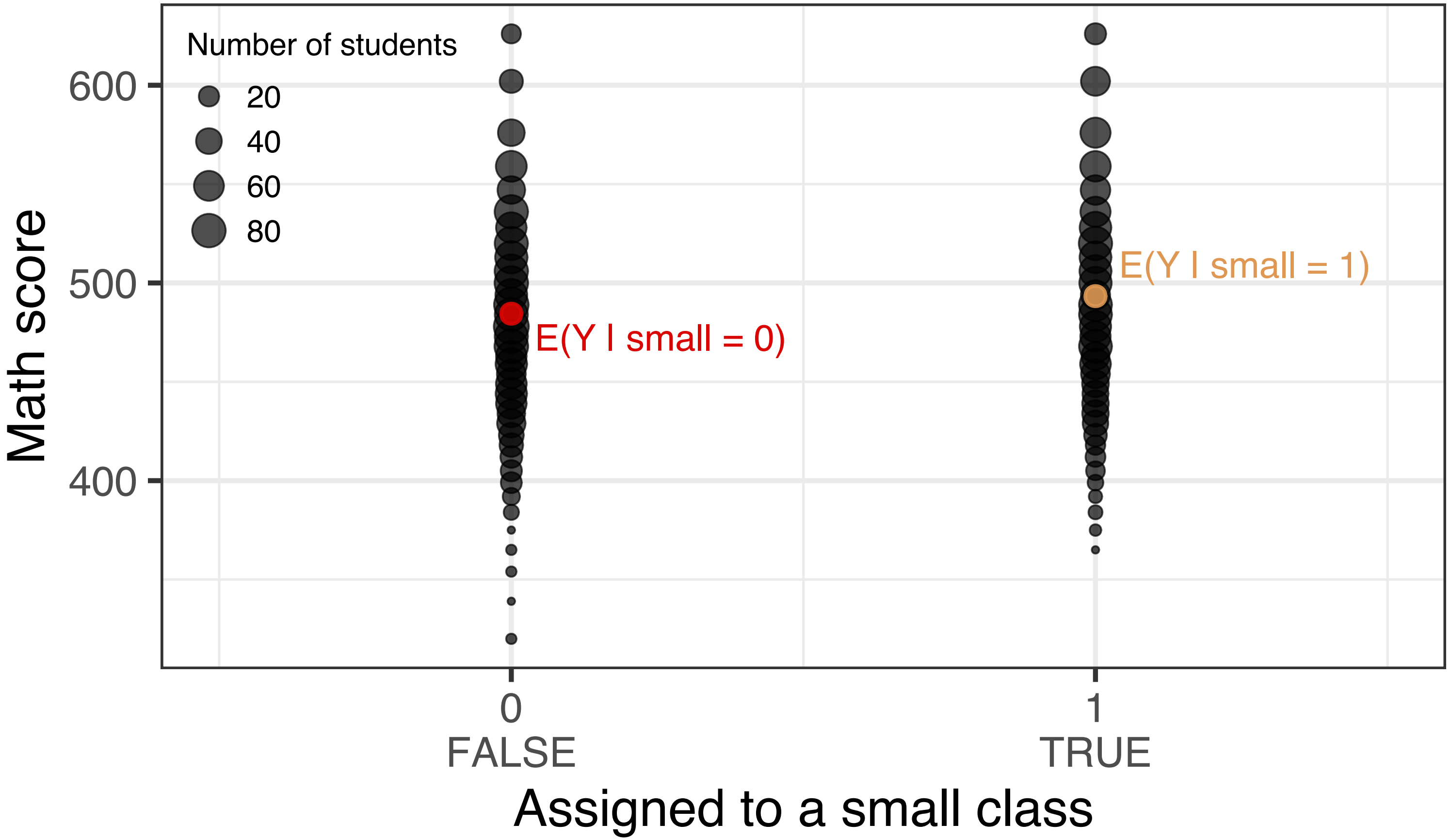
Regression with a Dummy Variable: Graphically

Contrary to last week, the regressor in our regression is a ***dummy variable***, i.e. a variable that takes the values TRUE or FALSE (1 or 0).



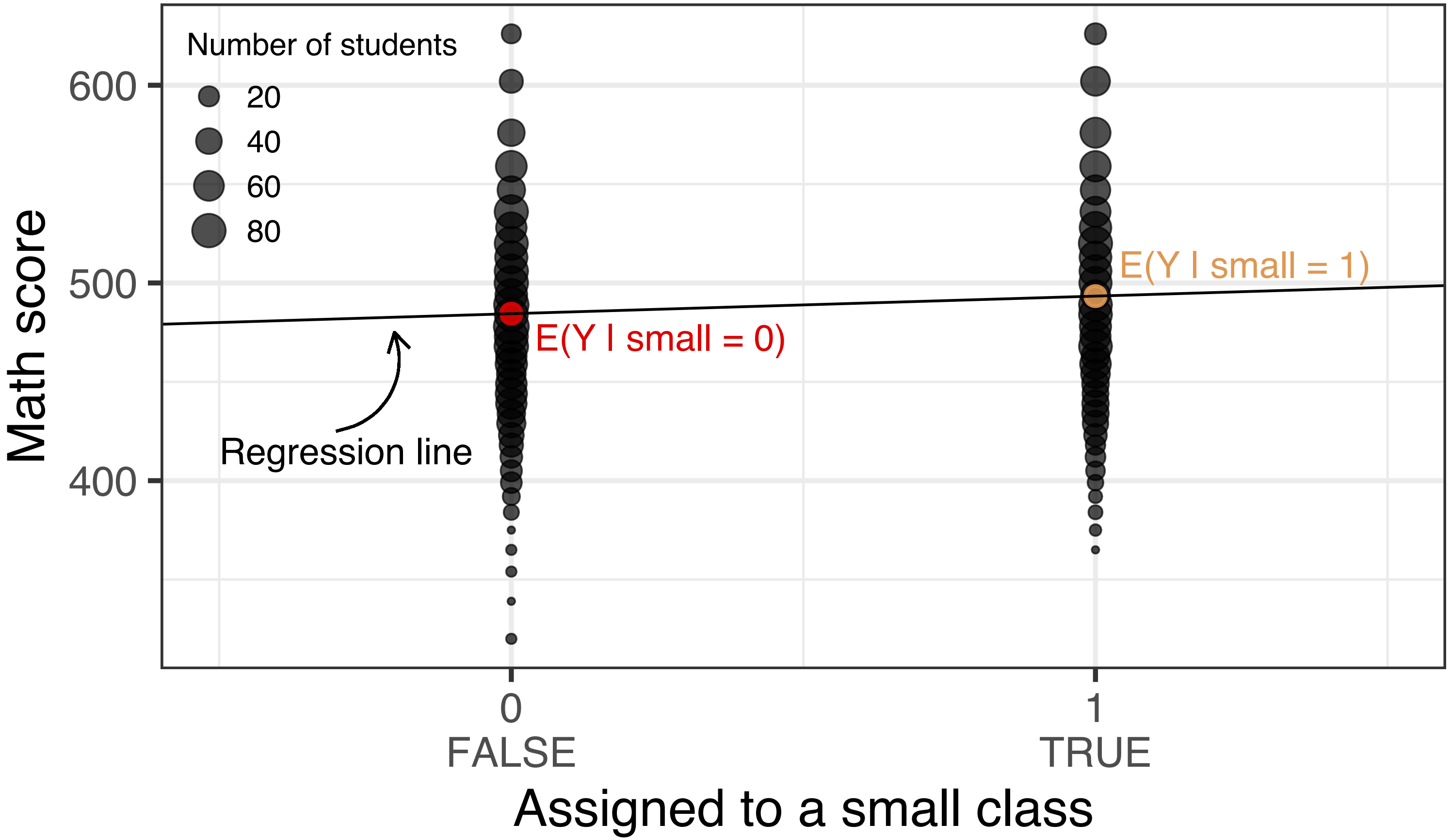
Regression with a Dummy Variable: Graphically

Contrary to last week, the regressor in our regression is a ***dummy variable***, i.e. a variable that takes the values TRUE or FALSE (1 or 0).



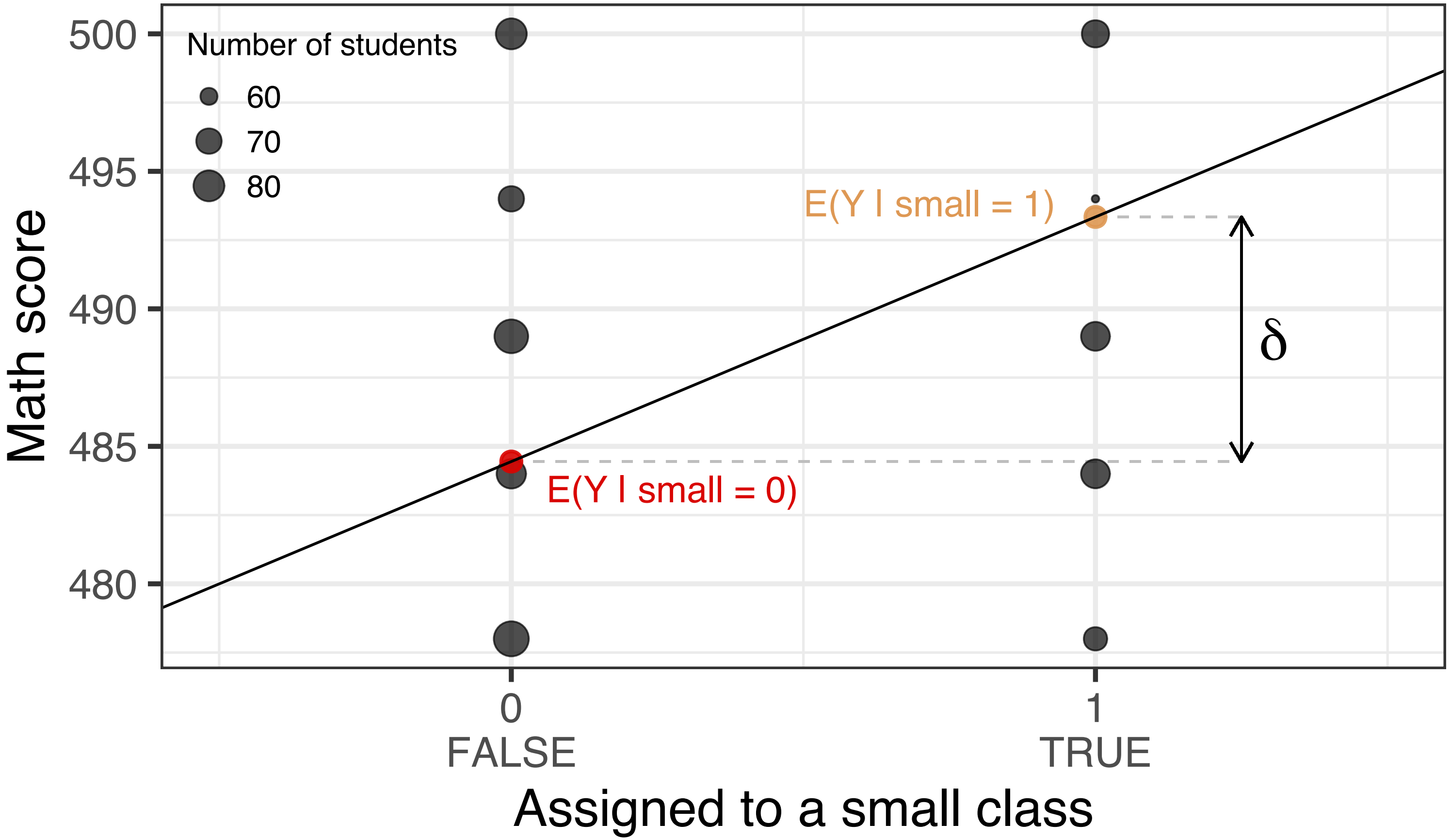
Regression with a Dummy Variable: Graphically

Contrary to last week, the regressor in our regression is a ***dummy variable***, i.e. a variable that takes the values TRUE or FALSE (1 or 0).



Regression with a Dummy Variable: Graphically

Contrary to last week, the regressor in our regression is a ***dummy variable***, i.e. a variable that takes the values TRUE or FALSE (1 or 0).



Regression with a Dummy Variable: Formally

Recall the regression model: $\text{math score}_i = b_0 + \delta \text{small}_i + e_i$

$$\begin{aligned}
\mathbb{E}[\text{math score}|\text{small}_i = 0] &= \mathbb{E}[b_0 + \delta \text{small}_i + e_i|\text{small}_i = 0] \\
&= b_0 + \delta \mathbb{E}[\text{small}_i|\text{small}_i = 0] + \mathbb{E}[e_i|\text{small}_i = 0] \\
&= b_0
\end{aligned}$$

Regression with a Dummy Variable: Formally

Recall the regression model: $\text{math score}_i = b_0 + \delta \text{small}_i + e_i$

$$\begin{aligned}
\mathbb{E}[\text{math score}|\text{small}_i = 0] &= \mathbb{E}[b_0 + \delta \text{small}_i + e_i|\text{small}_i = 0] \\
&= b_0 + \delta \mathbb{E}[\text{small}_i|\text{small}_i = 0] + \mathbb{E}[e_i|\text{small}_i = 0] \\
&= b_0
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\text{math score}|\text{small}_i = 1] &= \mathbb{E}[b_0 + \delta \text{small}_i + e_i|\text{small}_i = 1] \\
&= b_0 + \delta \mathbb{E}[\text{small}_i|\text{small}_i = 1] + \mathbb{E}[e_i|\text{small}_i = 1] \\
&= b_0 + \delta
\end{aligned}$$

Regression with a Dummy Variable: Formally

Recall the regression model: $\text{math score}_i = b_0 + \delta \text{small}_i + e_i$

$$
\begin{aligned}
\mathbb{E}[\text{math score}|\text{small}_i = 0] &= \mathbb{E}[b_0 + \delta \text{small}_i + e_i|\text{small}_i = 0] \\
&= b_0 + \delta \mathbb{E}[\text{small}_i|\text{small}_i = 0] + \mathbb{E}[e_i|\text{small}_i = 0] \\
&= b_0
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[\text{math score}|\text{small}_i = 1] &= \mathbb{E}[b_0 + \delta \text{small}_i + e_i|\text{small}_i = 1] \\
&= b_0 + \delta \mathbb{E}[\text{small}_i|\text{small}_i = 1] + \mathbb{E}[e_i|\text{small}_i = 1] \\
&= b_0 + \delta
\end{aligned}
$$

$$
\begin{aligned}
ATE &= \mathbb{E}[\text{math score}|\text{small}_i = 1] - \mathbb{E}[\text{math score}|\text{small}_i = 0] \\
&= b_0 + \delta - b_0 \\
&= \delta
\end{aligned}
$$

Regression with a Dummy Variable: Formally

Recall the regression model: $\text{math score}_i = b_0 + \delta \text{small}_i + e_i$

$$\begin{aligned}\mathbb{E}[\text{math score}|\text{small}_i = 0] &= \mathbb{E}[b_0 + \delta \text{small}_i + e_i|\text{small}_i = 0] \\ &= b_0 + \delta\mathbb{E}[\text{small}_i|\text{small}_i = 0] + \mathbb{E}[e_i|\text{small}_i = 0] \\ &= b_0\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\text{math score}|\text{small}_i = 1] &= \mathbb{E}[b_0 + \delta \text{small}_i + e_i|\text{small}_i = 1] \\ &= b_0 + \delta\mathbb{E}[\text{small}_i|\text{small}_i = 1] + \mathbb{E}[e_i|\text{small}_i = 1] \\ &= b_0 + \delta\end{aligned}$$

$$\begin{aligned}ATE &= \mathbb{E}[\text{math score}|\text{small}_i = 1] - \mathbb{E}[\text{math score}|\text{small}_i = 0] \\ &= b_0 + \delta - b_0 \\ &= \delta\end{aligned}$$

We knew this already but we now understand why this is true ✌️

Task 3: Your Turn!

10:00

1. Filter the dataset to only keep first graders and the small class and regular class groups.

2. Compute the average math score for both groups, and the difference between the two. (Use base `R`.)

3. Create a dummy variable `treatment` equal to `TRUE` if student is in treatment group (i.e. small class size) and `FALSE` if in control group (i.e. regular class size). See slide 33 for how to create a dummy variable.

4. Regress math score on the treatment dummy variable. Are the results in line with question 2?

5. How do you interpret these coefficients?

Shortcomings of RCTs

RCTs have very strong ***internal validity***, that is they can convincingly establish causal links.

However, they have some shortcomings:

- RCT are often **infeasible**:
  - RCTs are **costly**,
  - RCTs may face some **ethical issues**: some *treatments* simply cannot be given to people,
  - RCTs take time and we may be **time constrained**.

Shortcomings of RCTs

RCTs have very strong ***internal validity***, that is they can convincingly establish causal links.

However, they have some shortcomings:

- RCT are often **infeasible**:
    - RCTs are **costly**,
    - RCTs may face some **ethical issues**: some *treatments* simply cannot be given to people,
    - RCTs take time and we may be **time constrained**.
- **Interpretation** of the results:
    - ***External validity***: To what extent can the results from a given RCT be generalized to other contexts (countries, populations,...)?
    - Uncovering the mechanisms that are at stake may be difficult,
    - Imperfect randomization, attrition, ...

What comes next?

- So if we cannot rely on RCTs to make our life easy, it means we have to find a way to make causal inference from ***observational data*** (as opposed to ***experimental data***).

What comes next?

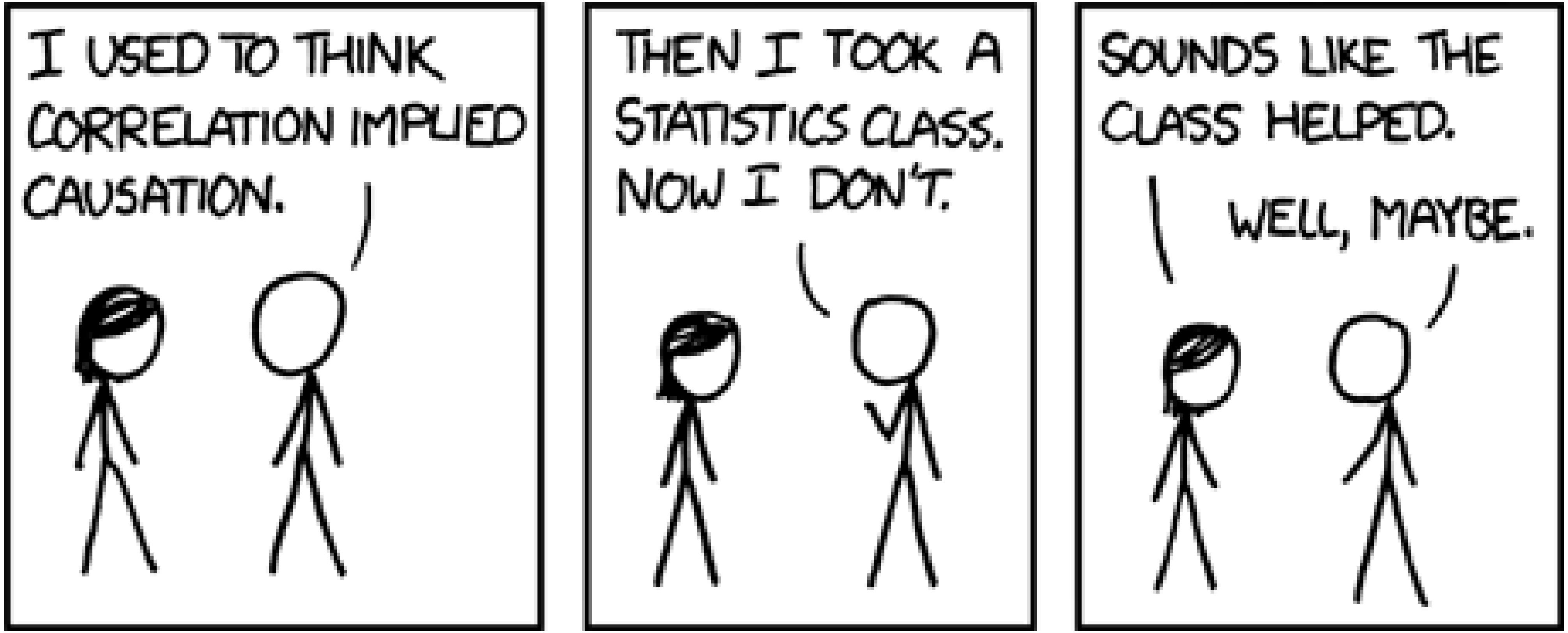
- So if we cannot rely on RCTs to make our life easy, it means we have to find a way to make causal inference from ***observational data*** (as opposed to ***experimental data***).
- It brings us back to models
  - In causal inference, the *model* is our idea of what the process that *generated the data* is.
  - We have to make some assumptions about what this is!

What comes next?

- So if we cannot rely on RCTs to make our life easy, it means we have to find a way to make causal inference from ***observational data*** (as opposed to ***experimental data***).
- It brings us back to models
  - In causal inference, the *model* is our idea of what the process that *generated the data* is.
  - We have to make some assumptions about what this is!

2 broad cases:

- ***selection occurs on observable characteristics***: *multiple regression* (next week!)
- ***selection occurs on unobservable characteristics***: *regression discontinuity design* (lecture 10) or *difference-in-differences* (not covered but set of slides online!)


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

SciencesPo
DEPARTMENT OF ECONOMICS

SEE YOU NEXT WEEK!

	florian.oswald@sciencespo.fr
	Slides
	Book
	@ScPoEcon
	@ScPoEcon