SciencesPo
DEPARTMENT OF ECONOMICS

ScPoEconometrics

Intro To Causality

Florian Oswald, Gustave Kenedi and Pierre Villedieu
SciencesPo Paris
2020-03-03

Recap from last week

- *Simple Linear Regression Model*: $y_i = b_0 + b_1 x_i + e_i$
- *Ordinary Least Squares* (OLS) estimation: minimize the sum of squared errors
- R command to estimate a linear model: `lm(dependent variable ~ independent variable, data)`

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

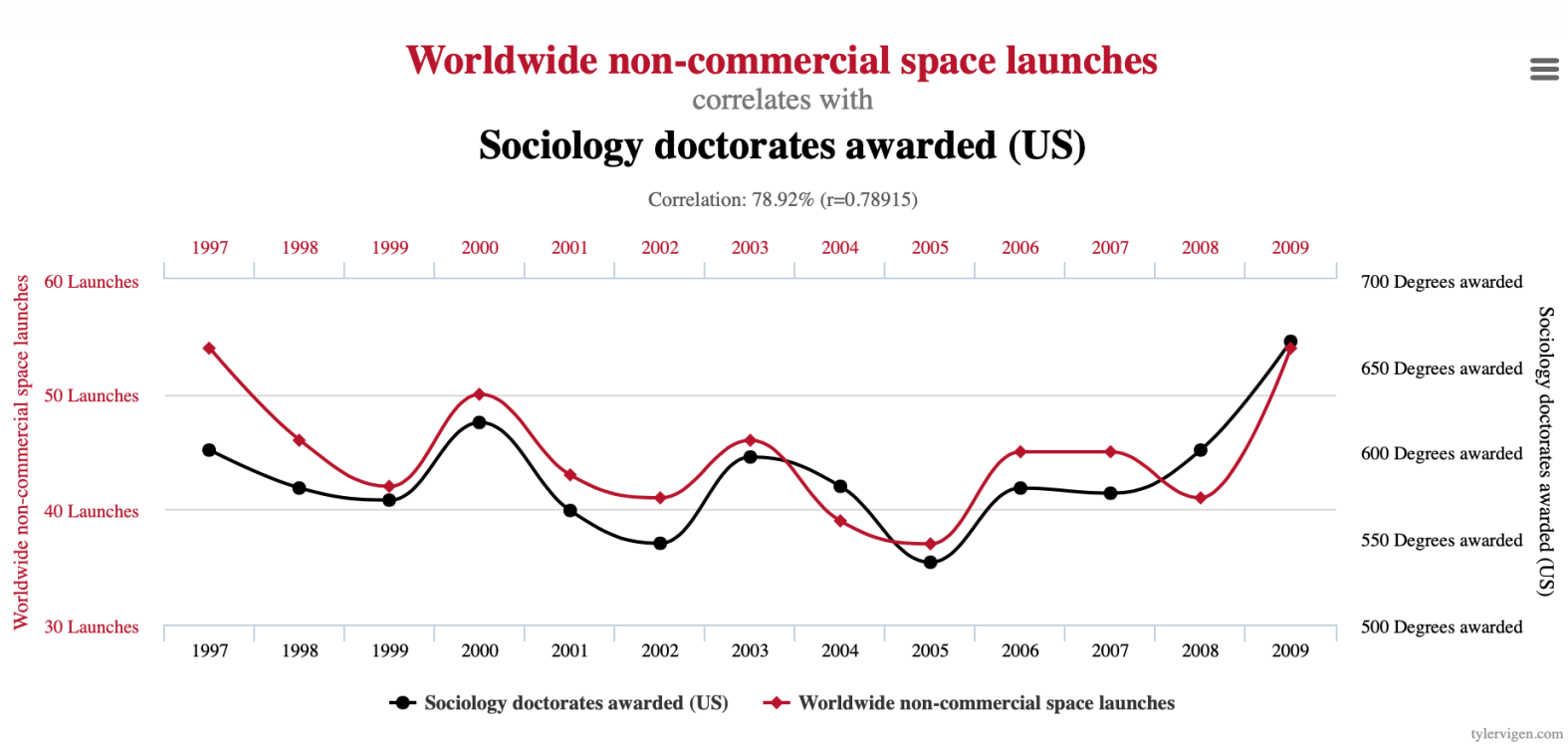
Correlation does not equal causation has become a ubiquitous mantra, but can you tell why it is true?

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?

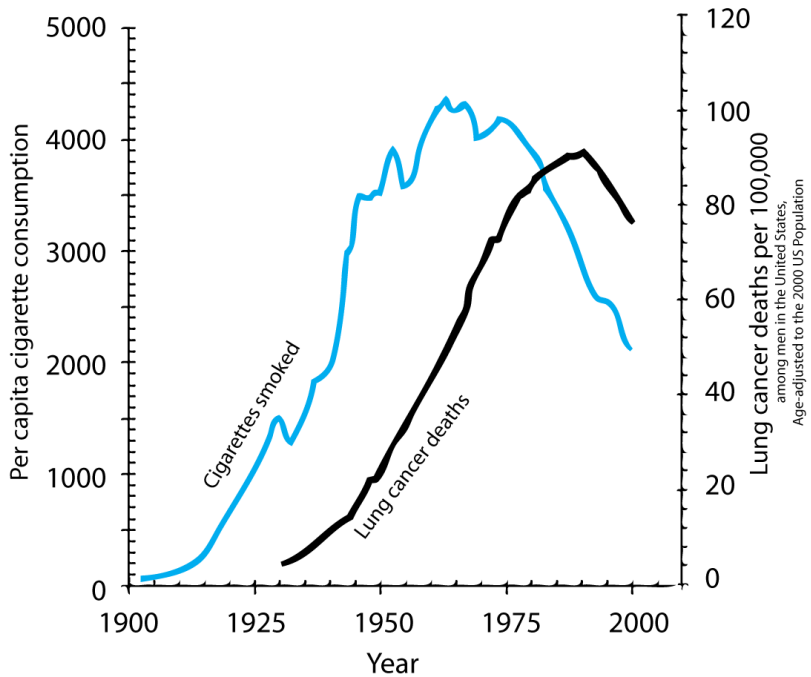
- Today, we know the answer is *YES*!
- But let's go back in the 1950's
  - We are at the start of a big increase in deaths from lung cancer...
  - ... which is happening after a fast growth of cigarette consumption

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 of cigarette consumption

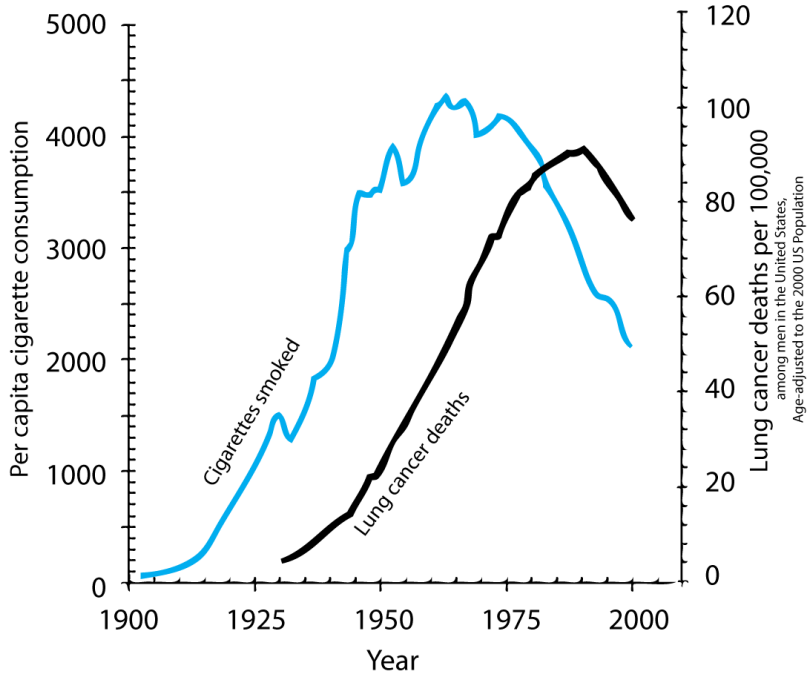


Correlation vs Causation: Smoking and Lung Cancer

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 of 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).

Correlation vs Causation: Smoking and Lung Cancer

- Let's focus on one of these potential confounding characteristics: **age**.

Correlation vs Causation: Smoking and Lung Cancer

- Let's focus on one of these potential confounding characteristics: **age**.

<<<<<<< HEAD

- Based on Cochran (1968), we will use death rates in Canada, the U.K. and the U.S.
- For each country 3 groups of men were studied for 20 months (U.S.) to 5 five years (U.K.).
||||||| merged common ancestors
- Based on Cochran (1968), we will use death rates from lung cancer in Canada, the U.K. and the U.S.

- **Add simple details about sample size and years!**

- Based on Cochran (1968), we will use death rates in Canada, the U.K. and the U.S.

| | | | | | | 8e41343abd4d3b1a2406ce7df22377c133dac098

<<<<<<< HEAD

Correlation vs Causation: Smoking and Lung Cancer

- Let's focus on one of these potential confounding characteristics: **age**.

<<<<<<< HEAD

- Based on Cochran (1968), we will use death rates in Canada, the U.K. and the U.S.
- For each country 3 groups of men were studied for 20 months (U.S.) to 5 five years (U.K.).
||||||| merged common ancestors
- Based on Cochran (1968), we will use death rates from lung cancer in Canada, the U.K. and the U.S.

- **Add simple details about sample size and years!**

- Based on Cochran (1968), we will use death rates in Canada, the U.K. and the U.S.

||||||| 8e41343abd4d3b1a2406ce7df22377c133dac098

<<<<<<< HEAD||||||| merged common ancestors

Correlation vs Causation: Smoking and Lung Cancer

- But the age *distribution* is also very different depending on smoking status.
- The following table gives the mean age by smoking status

Smoking group	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

- Because health is likely to deteriorate with age the previous table could be far from giving causal estimates.

<<<<<<< HEAD

Task 1 (5 minutes)

||||||| merged common ancestors

Correlation vs. causality : smoking and lung cancer #3

=======

Task 1 (5 min)

| | | | | | | 8e41343abd4d3b1a2406ce7df22377c133dac098

<<<<<<< HEAD Let's adjust our Canadian statistics **taking the age distribution into account**
. ||||||| merged common ancestors

Let's consider adjust our statistics taking into account the

<<<<<<< HEAD

Correlation vs Causation: Smoking and Lung Cancer

||||||| merged common ancestors

Correlation vs. causality : smoking and lung cancer #4

=======

Correlation vs. causality : smoking and lung cancer #3

8e41343abd4d3b1a2406ce7df22377c133dac098

Here is the age-adjusted death rates table found by Cochran (1968).

Smoking group	Canada	U.K.	U.S.

<<<<<<< HEAD

Correlation vs Causation: Smoking and Lung Cancer

||||||| merged common ancestors

Correlation vs. causality : smoking and lung cancer #4

=======

Correlation vs. causality : smoking and lung cancer #3

8e41343abd4d3b1a2406ce7df22377c133dac098

Here is the age-adjusted death rates table found by Cochran (1968).

Smoking group	Canada	U.K.	U.S.

<<<<<<< HEAD

Correlation vs Causation: Smoking and Lung Cancer

||||||| merged common ancestors

Correlation vs. causality : smoking and lung cancer #4

=======

Correlation vs. causality : smoking and lung cancer #3

8e41343abd4d3b1a2406ce7df22377c133dac098

Here is the age-adjusted death rates table found by Cochran (1968).

Smoking group	Canada	U.K.	U.S.

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 cigarettes, 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.

[1]: The potential outcomes framework was first proposed by Jerz Neyman in his 1923 Master's thesis.

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 $D_i = 0$.

<<<<<<< HEAD ***Treatment group***: all the individuals such that $D_i = 1$.

Control group: all the individuals such that $D_i = 0$. ||||||| merged common ancestors **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***:
  - Y_i^1: *potential outcome in the treatment state* $(D_i = 1)$ for individual i,
  - Y_i^0: *potential outcome in the control state* $(D_i = 0)$ for individual i.

The Potential Outcomes Framework

- In this framework, each individual has two ***potential outcomes***:
  - Y_i^1: *potential outcome in the treatment state* $(D_i = 1)$ for individual i,
  - Y_i^0: *potential outcome in the control state* $(D_i = 0)$ for individual i.
- From these we can define the ***individual treatment effect***:

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

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

The Potential Outcomes Framework

- In this framework, each individual has two ***potential outcomes***:
  - Y_i^1: *potential outcome in the treatment state* $(D_i = 1)$ for individual i,
  - Y_i^0: *potential outcome in the control state* $(D_i = 0)$ for individual i.
- From these we can define the ***individual treatment effect***:

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

- δ_i measures the **causal effect of D_i (treatment)** for individual i on outcome Y.
- 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 this framework, each individual has two ***potential outcomes***:
  - Y_i^1: *potential outcome in the treatment state* $(D_i = 1)$ for individual i,
  - Y_i^0: *potential outcome in the control state* $(D_i = 0)$ for individual i.
- From these we can define the ***individual treatment effect***:

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

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

Table : The Fundamental Problem of Causal Inference

Group	Y^1	Y^0
Treatment group $(D=1)$	Observable as Y	Counterfactual
Control group $(D=0)$	Counterfactual	Observable as Y

- The potential outcome that is not observed exists in principle, it is called the ***counterfactual outcome***.
  - What your test score would have been if you had been in a big class, knowing that you were in a small one.

The Potential Outcomes Framework

Table : The Fundamental Problem of Causal Inference

Group	Y^1	Y^0
Treatment group $(D=1)$	Observable as Y	Counterfactual
Control group $(D=0)$	Counterfactual	Observable as Y

- The potential outcome that is not observed exists in principle, it is called the ***counterfactual outcome***.
  - What your test score would have been if you had been in a big class, knowing that you were in a small one.
- Since the treatment effect *cannot* be observed at the individual level, we estimate treatment effects at the population level.

This table is from Morgan and Winship (2015, p.46).

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!

The Problem of Causal Inference

- 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

- 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:

<<<<<<< HEAD

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

The Problem of Causal Inference

- 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:

<<<<<<< HEAD

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

||||||| merged common ancestors

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

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

The Problem of Causal Inference

- 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:

<<<<<<< HEAD

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

||||||| merged common ancestors

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

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

The Problem of Causal Inference

- 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:

<<<<<<< HEAD

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

||||||| merged common ancestors

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

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

The Problem of Causal Inference

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

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

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

Then,

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

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

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

<<<<<<< HEAD

The Problem of Causal Inference

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

After tedious calculations that we skip, we get that the SDO is now 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.

<<<<<<< HEAD

The Problem of Causal Inference

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

After tedious calculations that we skip, we get that the SDO is now 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.

<<<<<<< HEAD

The Problem of Causal Inference

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

After tedious calculations that we skip, we get that the SDO is now 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).

|||||| merged common ancestors

The Problem of Causal Inference

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

After tedious calculations that we skip, we can get that the SDO is now 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.

|||||| merged common ancestors

The Problem of Causal Inference

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

After tedious calculations that we skip, we can get that the SDO is now 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.
- **Intuition** : Suppose that *education* is our treatment and this is people who will most

8e41343abd4d3b1a2406ce7df22377c133dac098
class:inverse

Task 2 (10 minutes)

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

1. Load the data here. Notice that `Di_random` is a treatment status we would have under random assignment.

2. Create the following variables: Y_i and δ_i. Recall that $Y_i = D_i * Y_i^1 + (1 - D_i)Y_i^0$ and $\delta_i = Y_i^1 - Y_i^0$.

3. Compute the ***ATE*** and the ***SDO***. (Use base `R`.) Is there is any *bias*? Is it large in magnitude?

4. Using `D_i_random`, 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?

<<<<<<< HEAD

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 status **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 status **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 status **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:

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

👉 We can directly estimate the ATE from the data!

Randomized Experiments

Randomized experiments

<<<<<<< HEAD

- 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". ||||||| merged common ancestors

- **Often called Random Controlled Trial (RCT).**

Back to class size and students' achievement

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

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

We briefly discuss 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 class size on average student math and reading scores.

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

We briefly discuss 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 class size on average student math and reading scores.

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

We briefly discuss 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 class size on average student math and reading scores.

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

We briefly discuss 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 class size on average student math and reading scores.

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

We briefly discuss 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 3 (10 minutes)

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_long),]`

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

<<<<<<< HEAD We just saw that in an RCT the ATE is obtained by computing the differences in outcomes between the treatment and control groups.

Let's only focus on: ||||||| merged common ancestors

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

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

Let's only focus on:

- One treatment group: **small classes**,

The Project STAR Experiment

<<<<<<< HEAD We just saw that in an RCT the ATE is obtained by computing the differences in outcomes between the treatment and control groups.

Let's only focus on: ||||||| merged common ancestors

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

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

Let's only focus on:

- One treatment group: **small classes**,

The Project STAR Experiment

<<<<<<< HEAD We just saw that in an RCT the ATE is obtained by computing the differences in outcomes between the treatment and control groups.

Let's only focus on: ||||||| merged common ancestors

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

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

Let's only focus on:

- One treatment group: **small classes**,

The Project STAR Experiment

<<<<<<< HEAD We just saw that in an RCT the ATE is obtained by computing the differences in outcomes between the treatment and control groups.

Let's only focus on: ||||||| merged common ancestors

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

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

Let's only focus on:

- One treatment group: **small classes**,

RCT in Regression Form

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

RCT in Regression Form

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

Rewriting this equation, 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

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

Rewriting this equation, 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 $\delta_i = \delta, \forall i$,

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

RCT in Regression Form

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

Rewriting this equation, 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 $\delta_i = \delta, \forall i$,

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

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!

Let's therefore estimate the ATE of small class size on math scores using a regression.

The Project STAR Experiment: Regression

The last equation looks exactly like the simple regression model we saw last week!

Let's therefore estimate the ATE of small class size on math scores using a regression.

We want to estimate the following model:
$mathscore_i = b_0 + \delta D_i + e_i$, with

$$D_i = \begin{cases} 1 \text{ if small class} = 1 \\ 0 \text{ if small class} = 0 \end{cases}$$

<<<<<<< HEAD

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

The Project STAR Experiment: Regression

The last equation looks exactly like the simple regression model we saw last week!

Let's therefore estimate the ATE of small class size on math scores using a regression.

We want to estimate the following model: `(mathscore_i = b_0 + \delta D_i + e_i)`, with

$$D_i = \{ 1 \text{ if small class} = 1 \; 0 \text{ if small class} = 0$$

<<<<<<< HEAD

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

|||||| merged common ancestors

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

]

8e41343abd4d3b1a2406ce7df22377c133dac098

<<<<<<< HEAD .pull-right[

The Project STAR Experiment: Regression

The last equation looks exactly like the simple regression model we saw last week!

Let's therefore estimate the ATE of small class size on math scores using a regression.

||||||| merged common ancestors
=======

We want to estimate the following model:
$mathscore_i = b_0 + \delta D_i + e_i$, with

$$D_i = \begin{cases} 1 \text{ if small class} = 1 \\ 0 \text{ if small class} = 0 \end{cases}$$

<<<<<<< HEAD

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

$r \star_d f_k \_ small < - \star_d f \% > \% fi < er(\star \% \in \% c(\text{regular, small}) \& \nabla e = = \text{k}) \% >$
]
>>>>>>> 8e41343abd4d3b1a2406ce7df22377c133dac098
<<<<<<< HEAD
.pull-right[
|||||||

merged common ancestors
=======

The Project STAR Experiment: Regression

From the estimation output we get the following:

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

The Project STAR Experiment: Regression

From the estimation output we get the following:

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

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

The Project STAR Experiment: Regression

From the estimation output we get the following:

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

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

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

The Project STAR Experiment: Regression

From the estimation output we get the following:

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

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

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

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

Task 4 (10 minutes)

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

<<<<<<< HEAD

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

2. 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. ||||||| merged common ancestors

3. Compute the average math score for both groups.

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

5. Go back to the original data and this time keep first graders but don't get rid of the regular+aide treatment group.

6. Compute the average math score for the regular+aide treatment group.

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**.

<<<<<<< HEAD

- **Interpretation** of the results: ||||||| merged common ancestors

Interpretating the results

- **Interpretation** of the results

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?

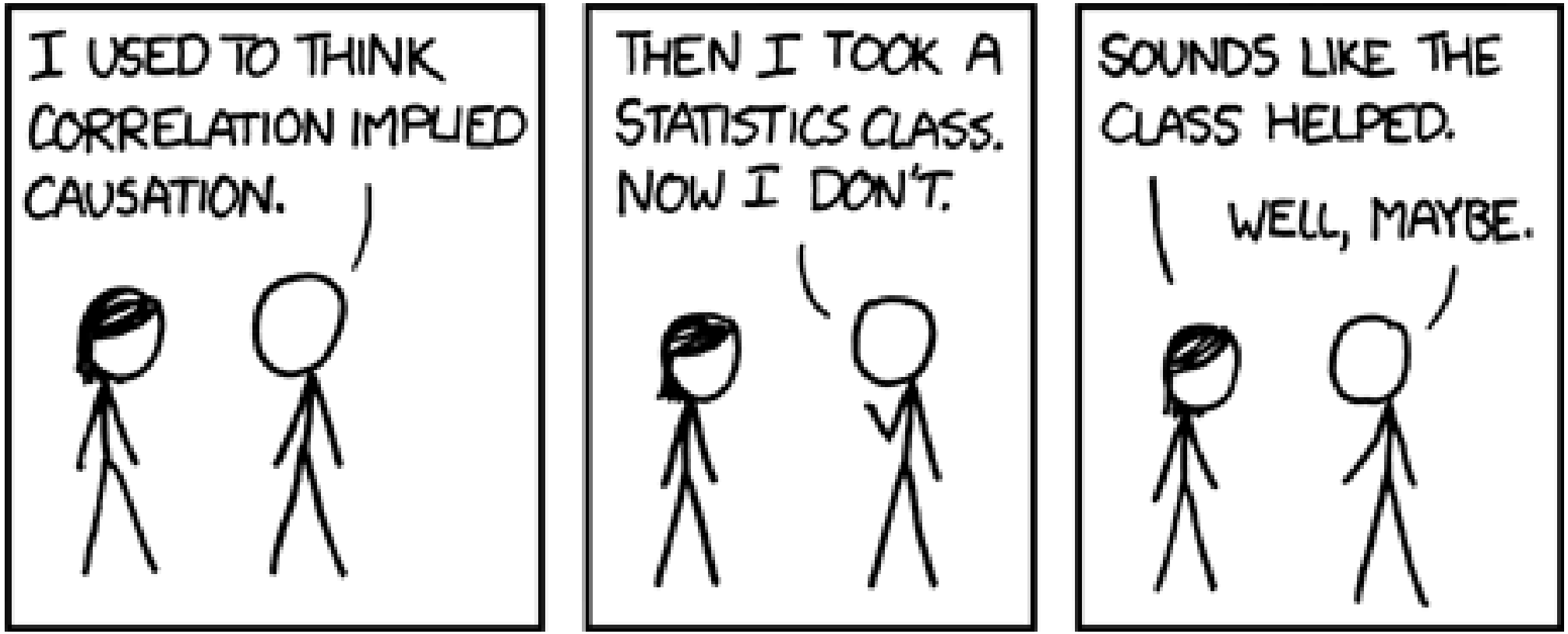
- So if we cannot rely on RCTs to make our life easy, it means we have to find a way to make causal inference from ***observational data*** (as opposed to ***experimental data***).
- It brings us back to models
  - In causal inference, the *model* is our idea of what the process that *generated the data* 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* or *difference-in-differences* (lectures 10 and 11!)


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