Lecture 5 - Causality

Dr. L.S. Sanna Stephan

2023

RUG Groningen

Learning objectives

You will learn

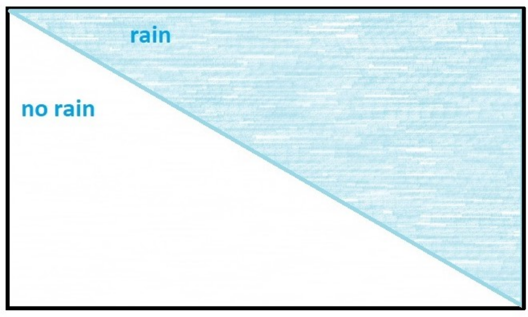
1. ... what it means when two variables are dependent,
2. ... why causality is important and how we try to identify it and
3. ... what is endogeneity and why we it hinders causal analysis.

Statistical Dependence

(further reading: DABEP Ch. 4.1-4.6)

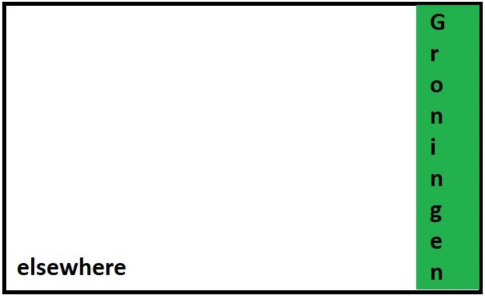
(In)dependent events

In the Netherlands, half of the time, it rains!



(In)dependent events

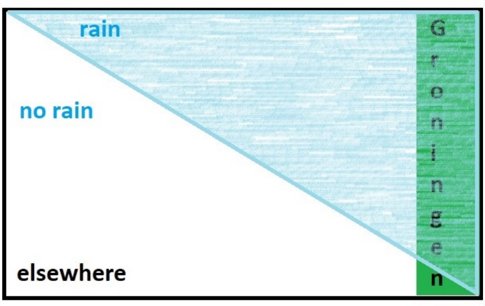
In the Netherlands, you can be in Groningen, or elsewhere!



(In)dependent events

Event A: it rains $\Rightarrow P(rain) = 0.5$

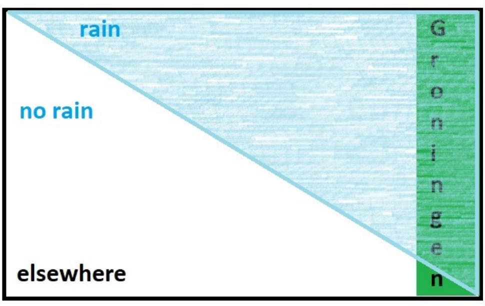
Event B: we are in Groningen $\Rightarrow P(Groningen) = \frac{1}{12}$



(In)dependent events

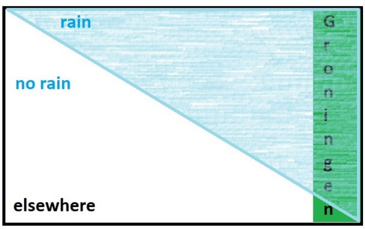
Event A: it rains $\Rightarrow P(rain) = 0.5$

Event B: we are in Groningen $\Rightarrow P(Groningen) = \frac{1}{12}$



Event A and B are **dependent**: in Groningen, the probability of rain is higher than elsewhere!

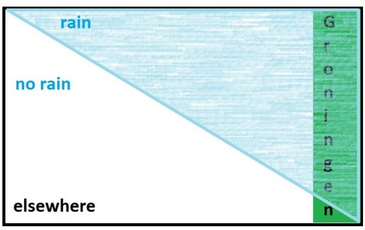
(In)dependent events



- Event A and B are **dependent**
- **Conditional** on being in Groningen, the probability of rain is higher!

The information that B has occurred affects the probability of A to occur

(In)dependent events



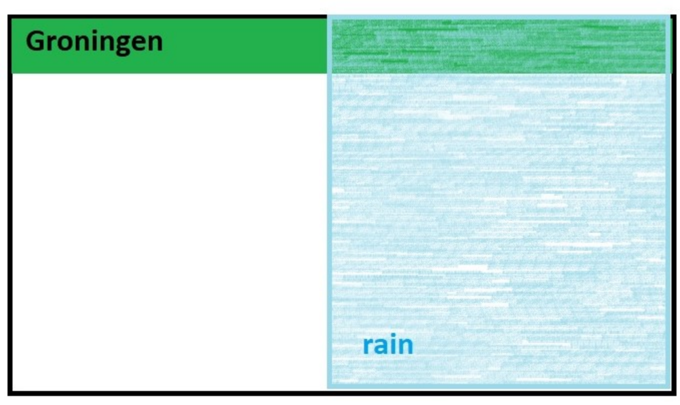
- Event A and B are **dependent**
- **Conditional** on being in Groningen, the probability of rain is higher!

$$P(rain|Groningen) = 0.92 \neq P(rain) = 0.5$$

The information that B has occurred affects the probability of A to occur

(In)dependent events

If A and B were independent, it would look like this:



In Groningen (and in NL) it rains half of the time!

⇒ Knowing you are in Groningen does **not help** predicted probability of rain

Estimating Causal Effects

Estimating Causal Effects

Introduction

(further reading: DABEP Ch. 19-19.1)

Why care about causality?

- We care about causality when there is a **choice**.
- We care about the **outcome** y, but we cannot influence it directly!
- **⇒ What is the effect of x on y?**
- x can include **treatment** and **control** variables

Causal questions

What is the effect of schooling on earnings?

1. Define Y: hourly wage.
2. Define the counterfactual.
3. Consider causal channels.

What is the effect of schooling on earnings?

2. Define the counterfactual. e.g.

    ... schooling versus no schooling,

    ... one additional year of schooling,

    ... completing a bachelor degree versus apprenticeship,

    ... attending a training or

    ...

What is the effect of schooling on earnings?

3. Consider causal channels. e.g.

   ...hard skills (specific knowledge/ competencies),
   ...soft skills (time management, communication, teamwork etc.),
   ...develop passion/ interest/ motivation/ purpose and
   ...(social/professional) network

Ceteris paribus

Ceteris paribus means **“all else being equal”**.

Just as in lab, we want to control for everything else.

Two ways to do so:

1. **treatment group** and **control group** drawn at random from same population
2. include **control variables**.

Estimating Causal Effects

Control Group

(further reading: DABEP Ch. 19.2-19.6, 19.13)

Treatment and control group

For simplicity, assume binary treatment ($T_i \in [0;1]$)

Basic idea:

$$\underbrace{y_i}_{\substack{\textit{observed}\\ \textit{outcome}}} = (1 - T_i) \underbrace{y_{i,0}}_{\substack{\textit{outcome if}\\ \textit{untreated}}} + T_i \underbrace{y_{i,1}}_{\substack{\textit{outcome if}\\ \textit{treated}}}$$

Treatment effect:

$$y_{i,1} - y_{i,0}$$

☠ ☠ ☠ this is never observed ☠ ☠ ☠

ATE

If treatment is **randomly** assigned (treatment and control group sampled from the same population), the **average treatment effect**

$$ATE = E(y_{i,1} - y_{i,0})$$

can be estimated by

$$\widehat{ATE} = \bar{y_1} - \bar{y_0} = \frac{\sum_{i=1}^{N} T_i y_i}{\sum_{i=1}^{N} T_i} - \frac{\sum_{i=1}^{N}(1 - T_i) y_i}{\sum_{i=1}^{N}(1 - T_i)}$$

(treatment group average minus control group average).

Estimating Causal Effects

Regression *(further reading: DABEP Ch. 7.1-7.3, 7.10, 7.1, 7.12, Ch. 19.7)*

Regression

Set of statistical methods to estimate the relationships between a variable Y and one or more variables X.

- The **regression equation** specifies (a characteristic of) the conditional distribution of Y given X as a function of
- the variables of interest X (variable of interest and controls),
- the unobserved error ε and
- potentially parameters.

Three types of regression

How much structure do we impose?

- **Parametric regression:** everything known except (finite number of) parameters
- **Semiparametric regression:** everything known except (infinite number of) parameters
- **Nonparametric Regression:** only some characteristic of the conditional distribution of Y is known

Parametric regression

$$Y = g(X, \beta_0, \varepsilon) \qquad \text{unknown: } \beta_0 \text{ (finite)}$$

Example: linear regression

$$\mu_{Y|X} = X'\beta_0 \qquad \rightarrow y_i = x_i'\beta_0 + \varepsilon_i \qquad \mu_{\varepsilon_i} = 0$$

$$Q_{1,Y|X} = X'\beta_0 \qquad \rightarrow y_i = x_i'\beta_0 + \varepsilon_i \qquad Q_{1,\varepsilon_i} = 0$$

Example from lecture 2:

$$grade_i = \beta_0 + \beta_1 hours_i + \beta_2 iq_i + \underbrace{\varepsilon_i}_{\substack{motivation \\ test\ day\ condition \\ luck\ \ldots}}$$

Semiparametric regression

$$Y = g(X, \beta_0, \varepsilon_i, \eta_0(.)) \qquad \text{unknown: } \theta_0, \eta_0(.) \text{ (infinite)}$$

$$\mu_{Y|X} = \eta_0(X'\beta_0) \qquad \rightarrow y_i = \eta_0(x_i'\beta_0) + \varepsilon_i \qquad \mu_{\varepsilon_i} = 0$$

$$Q_{1,Y|X} = \eta_0(X'\beta_0) \qquad \rightarrow y_i = \eta_0(x_i'\beta_0) + \varepsilon_i \qquad Q_{1,\varepsilon_i} = 0$$

More flexibility!

Example from lecture 2:

$$grade_i = \eta_0(\beta_0 + \beta_1 hours_i + \beta_2 iq_i) + \underbrace{\varepsilon_i}_{\substack{\textit{motivation} \\ \textit{test day condition} \\ \textit{luck ...}}}$$

"grading function" not known

Semiparametric regression

$$Y = g(X, \beta_0, \varepsilon_i, \eta_0(.)) \qquad \text{unknown: } \theta_0, \eta_0(.) \text{ (infinite)}$$

$$\mu_{Y|X} = X'\beta_0 + \eta_0(Z) \qquad \rightarrow y_i = x_i'\beta_0 + \eta_0(z_i) + \varepsilon_i \qquad \mu_{\varepsilon_i} = 0$$

$$Q_{1,Y|X} = X'\beta_0 + \eta_0(Z) \qquad \rightarrow y_i = x_i'\beta_0 + \eta_0(z_i) + \varepsilon_i \qquad Q_{1,\varepsilon_i} = 0$$

More flexibility!

Example from lecture 2:

$$grade_i = \beta_0 + \beta_1 hours_i + \eta_0(iq_i) + \underbrace{\varepsilon_i}_{\substack{\textit{motivation} \\ \textit{test day condition} \\ \textit{luck ...}}}$$

don't know relationship between IQ score and test score

Nonparametric regression

$$Y = \eta_0(X, \varepsilon) \qquad \text{unknown: } \eta_0(.) \text{ (infinite)}$$

$$\mu_{Y|X} = \eta_0(X) \qquad \rightarrow y_i = \eta_0(x_i) + \varepsilon_i \qquad \mu_{\varepsilon_i} = 0$$

$$Q_{1,Y|X} = \eta_0(X) \qquad \rightarrow y_i = \eta_0(x_i) + \varepsilon_i \qquad Q_{1,\varepsilon_i} = 0$$

Even more flexibility! Cost of flexibility:

before we could make some general statements (β_0) valid **for all X**. Now we can only make statement **for each** X

Example from lecture 2:

$$grade_i = \eta_0(hours_i, iq_i) + \underbrace{\varepsilon_i}_{\substack{motivation \\ test\ day\ condition \\ luck\ \ldots}}$$

the relationship may be specific for each (type of) student.

Regression

1. Choose type of regression.
2. Specify regression equation.
3. Choose estimator.
4. Estimate.
5. Draw conclusions/test hypothesis.

Spurious Correlation

Spurious Correlation

Introduction

You can always estimate

...but your result may not have a causal interpretation.

Statements like

"x has a significant (positive/negative) effect on y."

require causality! Estimation detects correlation, not causation.

False generalisation

- ✓ There is a relationship in the data...
- ☠ ... but only in **that particular data.**
- ☠ Mistake: one (few) particular event(s) used to derive general relationship!
- ☠ Remember: variance decreases in N!

False generalisation

"The election of Joe Biden (a new US president) lead to a run up in bitcoin prices"

☠ Joe Biden's election coincided with period during which retail investors FOMOd into bitcoin.

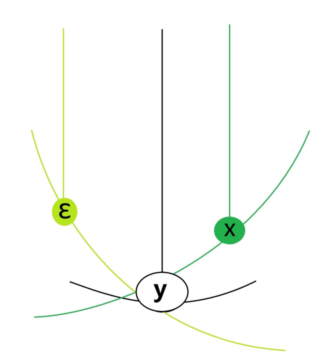
Spurious Correlation

Endogeneity

(further reading: DABEP Ch. 7.1, 7.2, 7.3, 7.10, 7.11, 7.12, Ch. 19.12, 19.14, 19.15)

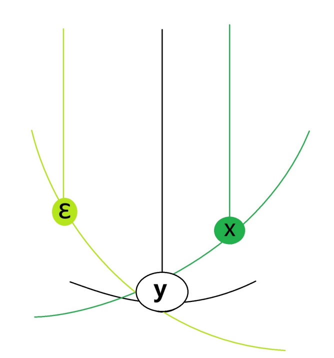
Exogeneity

An independent variable x is exogenous if it is determined **outside** the system.



Exogeneity

An independent variable x is exogenous if it is determined **outside** the system.



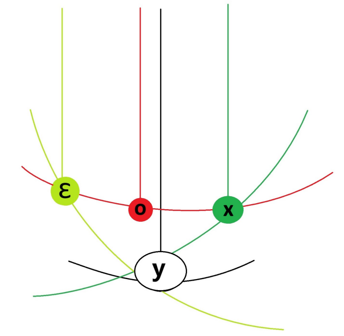
$x \longrightarrow y$ and $\varepsilon \longrightarrow y$, but $x \not\longleftrightarrow \varepsilon$

Endogeneity

An endogenous variable is determined **inside** the system

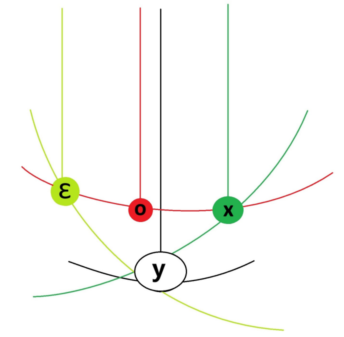
- because it is also caused by the outcome
- because both are determined by something else
- because it is systematically measured with error

Omitted variable bias



$x \rightarrow y$ and $\varepsilon \rightarrow y$, but $o \rightarrow x$ and $o \rightarrow \varepsilon$, therefore $x \longleftrightarrow \varepsilon$

Omitted variable bias



$x \rightarrow y$ and $\varepsilon \rightarrow y$, but $o \rightarrow x$ and $o \rightarrow \varepsilon$, therefore $x \longleftrightarrow \varepsilon$

Many (!) endogeneity problems are actually OVB!!!

What is the effect of schooling on earnings?

What is the effect of schooling on earnings?

Omitted variables, e.g.

- ☠ IQ, socioeconomic status, gender, health, age
- ☠ motivation/aspiration, preferences, talent

impact both earnings and schooling decision.

Sample selection is OVB

What is the effect of class size on test scores?

Sample selection is OVB

What is the effect of class size on test scores?

Assignment of students to classes not random, if

- ☠ assign high performers to small classes (max. benefit)
- ☠ assign low performers to small classes (min. variation)

then omitted variable is (past) performance/ talent/ attitude.

Simpson paradox is OVB

Recall the **Simpson paradox**: a trend is present in a sub sample but absent or reversed in the entire population.

- ☠ Drug against prostate cancer will show no effect in entire population.
- ☠ Omitted variable: gender

Variable definition problem is OVB

" GDP growth causes increase in smoking!"
Evidence: in countries with larger GDP, the overall cigarette consumption is higher.

- ☠ Both GDP and cigarettes smoked increase in population size.
- ☠ Omitted variable: population size (solution: use per capita variables).

Common time trend creates OVB

"A higher fraction of vegans/vegetarians in the population causes a higher per capita GDP."

- ☠ Both variables have been increasing over time in the past.
- ☠ Omitted variable: time (solution: use deviation from time trend).

Some false generalisation is OVB

✓ There is a statistically significant relationship in the data...

☠ ... but only in **that particular data.**

Some false generalisation is OVB

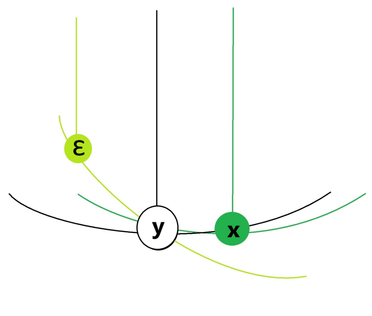
✓ There is a statistically significant relationship in the data...

☠ ... but only in **that particular data.**

"Better online banking opportunities cause individuals to adhere to special diets (ketogenic, intermittent fasting)."

☠ Omitted variable: technological progress and more information dissemination.

Reverse causality



$x \rightarrow y$ and $\varepsilon \rightarrow y$, but $y \rightarrow x$, therefore $x \longleftrightarrow \varepsilon$

Reverse causality

Reverse causality creates the "reflection problem": akin to a hall of mirrors.

$$x \rightarrow y \rightarrow x \rightarrow y \ldots$$

Reverse causality - supply and demand

price impacts both supply and demand

→ government regulates rental price

→ less new construction, decrease in supply

→ prices go up again

Exogenous shock to supply: subsidized construction.

Exogenous shock to demand: subsidized mortages.

Reverse causality examples

"The impact of **tweets** on **product popularity**"

Reverse causality examples

"The impact of **tweets** on **product popularity**"

More tweets → more popularity,

BUT

more popularity → more tweets!

Reverse causality

Why could this research question be problematic?

"Do countries with **better childcare facilities** have **higher birth rates**?"

Reverse causality

Why could this research question be problematic?

"Do countries with **better childcare facilities** have **higher birth rates**?"

☠ better childcare facilities → cost of raising child ↓
BUT democracy with many children → political pressure for better childcare ☠

Reverse causality

Why could this research question be problematic?

"What is the effect of **hourly wage** on **hours worked**?"

Why could this research question be problematic?

"What is the effect of **hourly wage** on **hours worked**?"

☠ Higher hourly wage → work more (income effect)
OR higher hourly wage → work less (substitution effect)
BUT unwillingness to work full time excludes jobs with highest hourly wage ☠

Reverse causality

Why could this research question be problematic?

"What is the effect of **healthy diet** on **chronic health condition**?"

Why could this research question be problematic?

"What is the effect of **healthy diet** on **chronic health condition**?"

☠ Healthy diet → better health
BUT chronic health condition → need/ recommendation to eat healthily ☠

Sample selection can be reverse causality

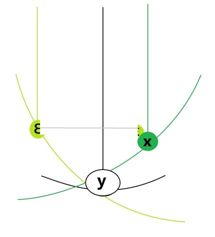
Treatment: government introduces new training for self employed to help them develop their business

Sample selection can be reverse causality

Treatment: government introduces new training for self employed to help them develop their business

☠ ☠ ☠ problem if participants are chosen (e.g. those most in need or those benefiting most)

Measurement error



$$x = \underbrace{\tilde{x}}_{\substack{\textit{unobserved}\\ \textit{truth}}} + \underbrace{\varepsilon_x}_{\substack{\textit{measurement}\\ \textit{error}}}$$

true model: $y = f(\tilde{x}, \varepsilon_y)$, estimated model: $y = f(x, \underbrace{\varepsilon_y, \varepsilon_x}_{\varepsilon})$,

consequence: $x \longleftrightarrow \varepsilon$

Why could this research question be problematic?

"What is the effect of **IQ** on **education choices**?"

Why could this research question be problematic?

"What is the effect of **IQ** on **education choices**?"

☠ More educated individuals/ native speakers perform better on IQ tests! ☠

(Systematic) measurement error examples

Why could this research question be problematic?

"What is the effect of **excessive drinking** on **health outcomes**?"

Why could this research question be problematic?

"What is the effect of **excessive drinking** on **health outcomes**?"

☠ Embarrassing question → high values under-reported ☠

Why could this research question be problematic?

"Why do countries with good treatment/ high awareness feature **higher** rates of mental illness?"

(Systematic) measurement error examples

Why could this research question be problematic?

"Why do countries with good treatment/ high awareness feature **higher** rates of mental illness?"

☠ No treatment possibilities & un-awareness → cases not diagnosed ☠

Dealing with endogeneity

- In applied research, endogeneity is a key concern.
- In your studies, you will learn much (!) more about ways to estimate models even in presence of endogeneity.

Recap

Today we learned

- ... what is statistical dependence,
- ... how to estimate causal effects and
- ... reasons for which your estimation results may **not** have a causal interpretation.