Statistics 2

Model assumptions and violations. Causality & Association

Casper Albers & Jorge Tendeiro

Lecture 3, 2019 – 2020

university of groningen

Overview

Assumptions

Causality and association

Literature for this lecture

Contents:

- Model assumptions and violations
  Causality & Association

Read:

Agresti, Section 9.6 , Ch. 10

Model assumptions

$$\boxed{y_i = \alpha + \beta x_i + \varepsilon_i} \quad \boxed{\varepsilon \sim \mathcal{N}(0, \sigma)}$$

Assumptions of the model:

1. Independent observations:
   - All observations are independent of each other.
   - True random sampling.
2. Linear relations:
   - Relation between x and $E(y)$ is a straight line.
3. Homoscedasticity:
   - The conditional standard deviation σ is constant
4. Residuals follow a normal distribution:
   - y_i follows a normal distribution around $E(y)$.

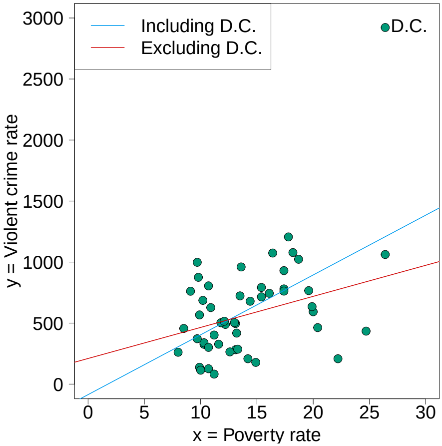
Model assumptions

What if the assumptions are invalid?

- The analyses are no longer guaranteed the best approach or, worse, not even valid anymore.
- Tests and CI's can lead to misleading and incorrect conclusions.
- Inferences are no longer justified.
- Checks and corrections are necessary.

Thus, the validity of the model is at play here.
More on assumptions in Lecture 7.

Influential points

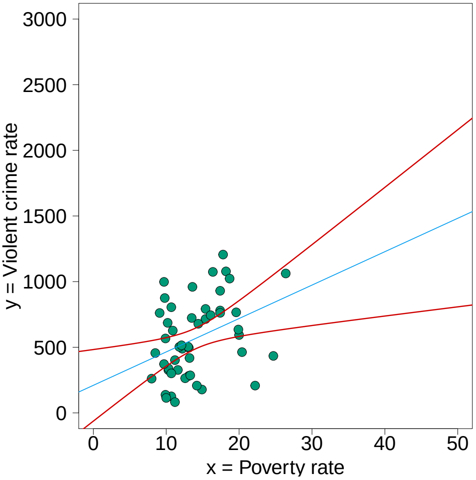


Be careful with influential points.

Points with Cook's distance>1 are deemed 'influential'.

(You do not need to be able to compute Cook's distance yourself.)

Extrapolation is dangerous



Avoid making predictions of violent crime rate (y) for poverty rates (x) of, say, 0 or 50.

Causality and association

There are many ways in which two variables can be associated (i.e., correlated). Some are more interesting than others.

Association (correlation):

Variable A — Variable B

Causal relation:

Variable A $\longrightarrow$ Variable B

Post Hoc Ergo Propter Hoc

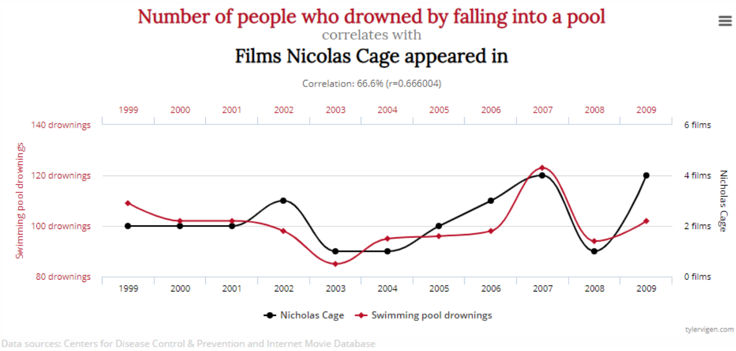
- Correlation does not imply causation, but
- Causation does imply correlation

Requirements for causality

A relation $A \rightarrow B$ can be causal if, at least, three requirements are met:

1. There is an association ($cor(A, B) \neq 0$).
2. There is an appropriate time ordering.
3. All other alternative explanations have been eliminated.

No causality



1. Very strong association ($r = .67$)
2. Time-ordering somewhat unclear.
   But also $\{$movies in year x, drownings year $x+1\}$ correlates strongly.
3. Obvious alternative explanation: Coincidence (n is only 11)

Randomized experiments

- Optimal way of establishing causality.
- Assign participants at random to control and experiment group.
  $\rightarrow$ This minimizes the possibility of alternative explanations.
- It does not guarantee causal relation (e.g., chain relationship.)

Oftentimes, randomization is not possible, e.g. effect of education level on political views.

Statistical control

To study whether '$A \to B$' or 'A — B', the effects from other variables need to be removed.

- Lab experiments: Keep variables (e.g., temperature) constant. Experimental control
- Observational studies: Experimental control is impossible. We need statistical control

Example:

- Dataset, $n = 100$ school children. Shoe size (x) and reading ability (y) correlate: $r(x, y) > 0$ ($p < .01$).
- Controlling for age: $y = b_0 + b_1 x + b_2 \mathrm{age}$: Partial regression effect b_1 is very small.

Not controlling might lead to incorrect conclusions

Study[1] on grant applications with science council NWO:

	Men	Women
Awarded	290	177
Not awarded	1345	1011
Success rate (%)	17.7	14.9

A χ^2 test yields a just significant results ($p = .045$): 'Compelling evidence' of gender bias in allocating research money by NWO.

[1] Van der Lee & Ellemers, PNAS, 2015

Not controlling might lead to incorrect conclusions

Post-publication analysis[2]: No gender bias but lack of statistical control.

Success rates (%):

Field	Men	Women
Chemistry	26.5	25.6
Physical sciences	19.3	23.1
Physics	26.9	22.2
Humanities	14.3	19.3
Technical sciences	15.9	21.0
Interdisciplinary	11.4	21.8
Earth sciences	24.4	14.3
Social sciences	15.3	11.5
Medical science	18.8	11.2

- The correlation between gender and success rate non-significant after controlling for field.

[2] Albers, PNAS, 2015

Types of multivariate relationships

There are different ways in which A and B can be related:

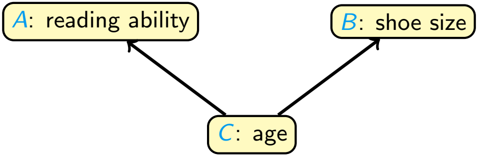
1. Direct causal relation
2. Spurious association (shoe size, NWO examples)
3. Chain relations
4. Interacting variables
5. Coincidence

From $r_{A,B} \neq 0$ alone we can never infer which type of association we have.

Direct causal relation

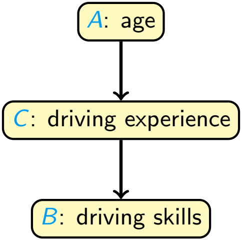
$$\text{Variable } A \longrightarrow \text{Variable } B$$

Spurious associations



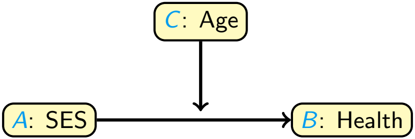
- A and B do not directly influence each other but have a common cause C.
  - When not controlling for C, $r_{A,B} \neq 0$: Spurious correlation.
  - C is called a lurking variable or hidden moderator.

Chain relations



- A is not directly causing B, but A causes C which in turn causes B.
  - Older people are, on average, better drivers than young people.
  - Yet, age is not the direct cause.

Interacting variables



- SES is positively correlated with health;
- The strength of the relation is moderated by age: Stronger correlation for older people.
- More on this in Section 11.4 (Lecture 5).

Suppressor variables

Lurking variable:

A present correlation disappears when a third variable is taken into account.

Supressor variable:

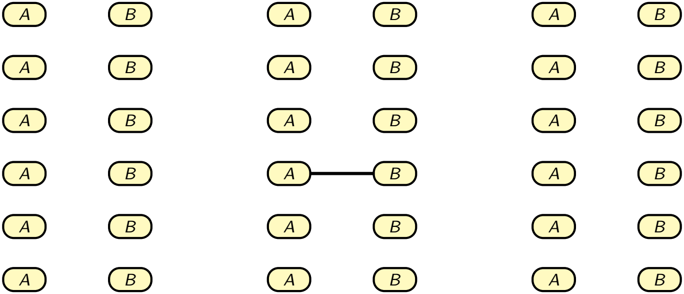
An absent correlation appears when a third variable is taken into account.

	All	people	Young	people	Old	people
Education vs. Income	High	Low	High	Low	High	Low
High	250	250	125	225	125	25
Low	250	250	25	125	225	125

In this example, age is a suppressor variable.

Note that categorizing continuous variables (age) is, in general, not a good idea.

Coincidence



- Coincidence: Every now and then you make a Type I error

Cheating and violating the rules of statistics

Another way to get A and B associated: Cheating, fraud, honest mistakes, etc.

Not part of today's lecture but of Lecture 13.

Next week: Multiple regression

Agresti, Sections 11.1 – 11.3