"The Truth About Linear Regression"
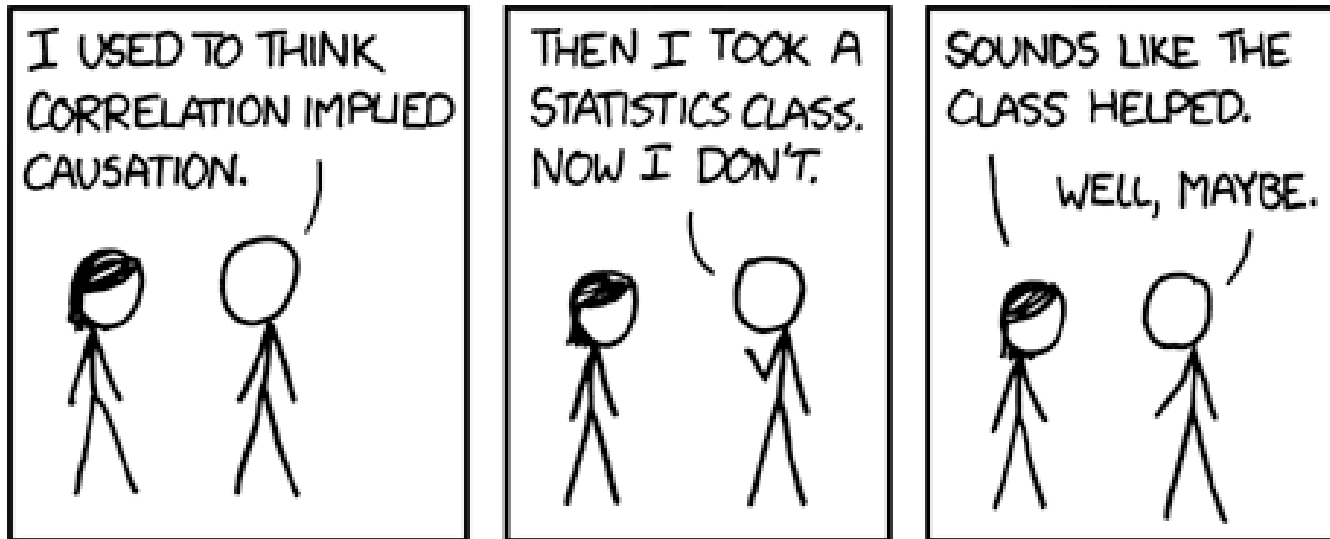

Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'.

xkcd #552

Content from Cosma Shalizi, *Advanced Data Analysis from an Elementary Point of View*

SML310: Research Projects in Data Science, Fall 2018

Michael Guerzhoy

1

"Lies about Linear Regression"

- Because a variable has a significant regression coefficient, it must influence the response

- Because a variable has an insignificant regression coefficient, it must not influence the response

- If the input variables change, we can predict how much the response will change by plugging in to the regression

Collinearity

- Two predictor variables are correlated (e.g., weight and height)
  - We will be uncertain about the coefficients for *both* weight and height
    - Could make weight matter less and height matter more and vice versa
  - We cannot say "a 1cm increase in height is associated with a 0.1 increase in GPA"
- Ways to handle
  - Remove redundant variables (dangerous)
  - PCA (will discuss later (possibly))

Omitted variables

- Variables that are not measured, but predict the response
  - Will influence coefficient estimate
  - Will make correlation look like causation
    - Examples?

Omitted variables

- The amount of ice cream consumed in a day is correlated with the number of drownings
  - Lurking variable: the weather
- Including omitted variables: "controlling for the variables"

Omitted variables: tricky cases

- Ronald Fisher (one of the founders of the field of Statistics) remained unconvinced by observational studies that showed association between smoking and lung cancer because of possible lurking variables
  - Suggested genetics might cause both smoking and lung cancer
  - Suggested illness might cause people to take up smoking
  - (Accepted funding from tabacco companies; seemed to be ideologically opposed to public health campaigns in general)
  - Is widely considered to have been wrong

Omitted variables: tricky cases

- The gender wage gap
  - Do you control for having children? How?
  - Do you control for the field of employment?
- The Harvard admissions lawsuit
  - E.g., do you control for the interview score?
  - Harvard and the plaintiffs submitted statistical analyses, arguing (among other things) for different controls
- Generally, more controls ➡ smaller effect size
  - Sometimes, controlling for a variable can be inappropriate because the variable and the outcome basically measure the same thing
    - Whether it's important or trivial that one of the variables predicts the outcome well depends on the situation
- Obviously, both of those are complex issues to which one slidedeck cannot do justice
  - And most of the issues are not necessarily statistical

Errors in variables

- Input variables measured imprecisely
  - The relationship between family income and school performance is often explored
  - But what's measured is the *reported* family income
- Tends to obscure the true relationship (and push the coefficients (effect sizes) toward 0)
  - Makes sense: more noise means it's harder to detect the trend

Significant coefficients

- All coefficients are significant if the sample size is large enough

Setting up a regression model

- Identify variables that could conceivably influence the response
  - Are there lurking variables?

- Can you theoretically justify interactions?
  - Would you want to have a hypothesis that involves the presence of interactions?

- Do model checking

- Example of an interaction?
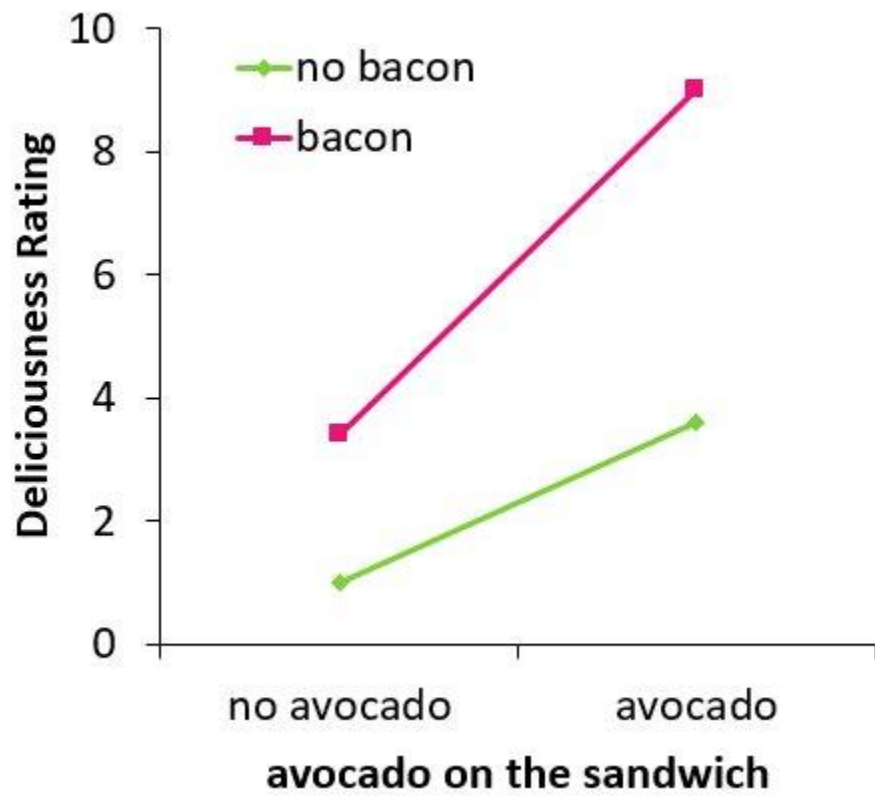
Rogue Works Progress Administration @GabrielRossman · Nov 16

The belief that describing social reality as a slope-intercept line through k-dimensional space is necessarily interesting or impressive

> **Eric Bailey** @EricVBailey
> Hi, Twitter!
>
> What did you have in 1998 that you no longer have in 2018?
> Show this thread

💬 3   🔁 3   ♡ 35

- 2005 Ph.D. Sociology Princeton University
- 2002 M.A. Sociology Princeton University
- 1999 B.A. Sociology UCLA

Rogue Works Progress Administration @GabrielRossman · Nov 17

yeah, i think regression is a useful method and i use it, but I feel like it's not interesting unless there's a plausible micro-mechanism.

💬 1   🔁 1   ♡ 2

Rogue Works Progress Administration @GabrielRossman · Nov 17

when I was in undergrad + grad school I thought it was interesting to, for example, do a regression of industry structure on artistic innovation. Now I kind of feel meh on that.

💬 3   🔁   ♡ 1

Rogue Works Progress Administration @GabrielRossman · Nov 17

This is one reason I increasingly like simulation -- it forces you to think about mechanism. That and there's no IRB.

💬   🔁   ♡ 3