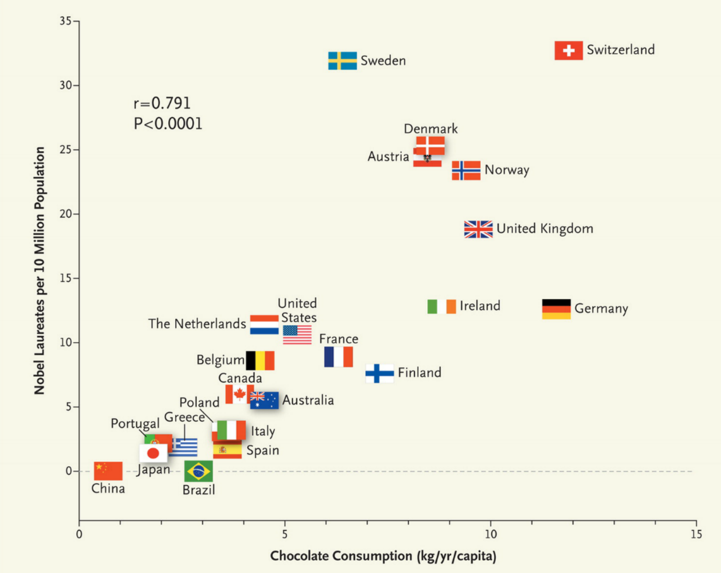
All of causal inference

Joshua Loftus

August 5, 2015

These are a few of my favorite things



Source: NEJM

The real reason Pope Benedict XVI abdicated? Rugby

From *Rugby (the religion of Wales) and its influence on the Catholic church*

> *"every time Wales win the rugby grand slam, a Pope dies, except for 1978 when Wales were really good, and two Popes died" . . . "our model for the general theory of papal rugby predicts that 0.62 of a Pope will die this year"*

A few other fun sources:

- Spurious correlations
- Google correlate

What if it’s not so obvious?

Consider the relationship between artificial sweeteners (e.g. diet soda) and obesity. There is a positive correlation between diet soda consumption and obesity.

- Maybe people who are obese switch to diet soda
- Maybe artificial sweeteners affect our gut bacteria
- Or maybe they give us a sweet tooth and affect the rest of our diet

It may not be so easy to rule out causation when the example isn’t an absurd one. Proving causation is also very challenging (impossible in some philosophical sense)

RCTs are the answer!

- The problem with observational data is that the comparisons may be unfair.
- The group drinking diet soda might be different from the group not drinking diet soda in some other important ways. One of those other differences might explain obesity.
- **R**andomized **c**ontrolled **t**rials make two (usually) random groups, treatment and control, which should yield fair comparisons.
- e.g.: Suppose some subset of the participants are smokers. Because they are assigned randomly, on average half of them will be in the treatment group and half in the control group. So whatever effect smoking has on the outcome, it should be (close to) the same in both groups. If the outcome is significantly different, it's probably not because of smoking, so it's probably because of the treatment.

But RCTs aren't always feasible. . .

Study on parachutes and "gravitational challenge"

> *As with many interventions intended to prevent ill health, the effectiveness of parachutes has not been subjected to rigorous evaluation by using randomised controlled trials. Advocates of evidence based medicine have criticised the adoption of interventions evaluated by using only observational data. We think that* ***everyone might benefit if the most radical protagonists of evidence based medicine organised and participated in a double blind, randomised, placebo controlled, crossover trial of the parachute****.*

Causal inference with–*gasp!*–observational data

Common approaches and methods:

- Matching (and propensity scores).
- (Neyman-Rubin) Potential outcome models. Jerzy Neyman (1923) and Donald Rubin (1974).
- Structural equation models (**SEM**). Common in educational testing and social sciences, e.g. *economics*.
- Instrumental variables (**IV**). Philip Wright (1928).

There are similarities/relationships and some of these methods can be used together.

Matching: balance groups for fair comparison

Given data on people who were exposed to some risk factor, and a larger (usually) group who have not been exposed. Want to estimate the effect of the risk factor on some outcome.

For each person who was exposed, find a control with similar covariates. This pair should be similar in *every relevant way* except that one was exposed and one was not. If some *relevant* information is not included in the data, this method (and pretty much all methods) will fail.

Propensity scores: match based on $\mathbb{P}(\text{exposed}|\text{covariates})$ (and/or include these scores in the regression model)

Potential outcomes: what *is* a causal effect?

Notation: $Y_i(1)$ is the potential outcome of person i if they are given the treatment, and $Y_i(0)$ is the potential outcome if they are given the control.

We only get to observe one of $Y_i(1)$ or $Y_i(0)$ for each i (the actual outcome), but not both. Treat the other as an unobserved random variable. Connection with missing data. RCT = missing at random.

Additive causal effect: $Y_i(1) - Y_i(0) = \tau_i$. For simplicity assume $\tau_i = \tau$ for all i. Can we estimate τ?

Use averages of observed: $\bar{Y}_t(1) - \bar{Y}_c(0)$ where t and c stand for the indices i that were assigned to treatment/control respectively.

Is it unbiased? What assumptions do we need?

Assumptions in Neyman-Rubin causal model

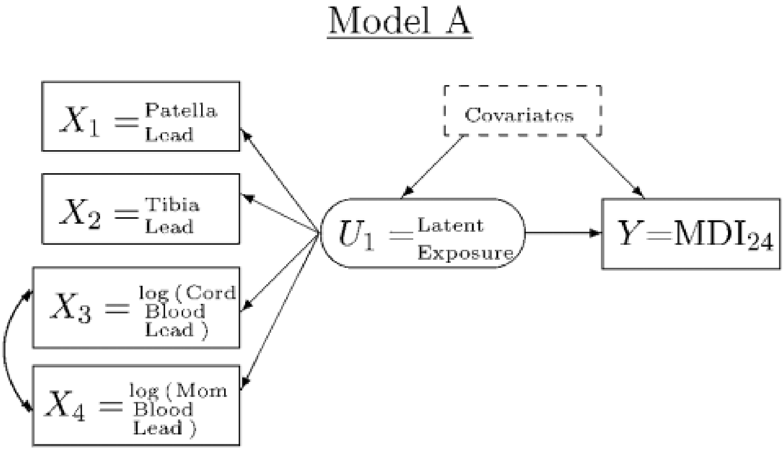
The following two assumptions make the previous estimator unbiased:

- Stable unit treatment value assumption (**SUTVA**): treatment assignment of one person does not affect potential outcomes of others (non-interference), and treatments are stable (i.e. some patients don't receive a different dose)
- Ignorability: treatment assignment is independent of potential outcomes

Confounders are variables that cause violations of the assumptions. If they are measured, we can use matching to adjust. If they are not measured. . .

Sensitivity analysis: either there is a causal effect of some magnitude, or an unmeasured confounder with some magnitude.

Structural equation modeling


Model A
X_1 = Patella Lead
X_2 = Tibia Lead
X_3 = log (Cord Blood Lead)
X_4 = log (Mom Blood Lead)
Covariates
U_1 = Latent Exposure
Y = MDI_{24}

Latent variables (missing data connection) estimated through imperfect "instruments." Directed arrows represent conditional dependence relations, "path analysis" multi-stage models.

Instrumental variables

Find (or generate) a new variable Z which affects Y *only* through the treatment assignment.

Assumptions: Z is assigned randomly, or its assignment is independent of the outcomes, and a few other things. . .

Regress Y on Z and treatment assignment on Z. If Z increases likelihood of being treated by a little bit, and increases Y a little bit, maybe the treatment increases Y a lot.

Challenge: finding/creating Z involves domain knowledge, creativity, getting the right data by hook or by crook, etc.