Introduction to Causal Inference, week 2

Daniel Stegmueller

Duke University

The problem of *causal* inference

Causal inference questions and statistics

Typical questions of interest:

- Could government action have prevented the crisis?
- How effective is this treatment in preventing this disease?
- Is there evidence for discrimination in these policing records?

Do changes in one variable cause changes in another. And, if so, how much?

(Until recently) relative silence in statistical education

- Focus on data summary / parameter "inference".
- Often silent on if the estimated parameters reflect the wanted quantity (large fraction of statistics books do not even have the term causal in their index)
- Rather, well-worn, but not very helpful advice that "correlation does not imply causation" etc.

Why study causal effects?

"WHAT IS?"
is central to many scientific questions, to guide action, policies

- Effect of smoking on lung cancer
- Effect of institutional features on behavior
- Effect of education on future salaries
- ...

"WHY and HOW is it?"

- Understanding how and why causes influence their effects (think 'mal-air' vs. mosquitos and what remedial action they imply...)

Causal inference and statistics

But how do considerations of causality differ from “statistics”?

Consider this famous puzzle; Simpson’s paradox (after E. Simpson, 1951)

- Effect of drug on disease outcome
- A group of sick patients is offered a novel drug
- Among those who took the drug, outcomes are *worse*! (a lower percentage recovers)
- When partitioning by gender, one finds:
  - Men who took drug have higher recovery rate than those who did not
  - Women who took drug have higher recovery rate than those who did not
- Thus, drug helps men and women; hurts the general population

Simpson's paradox

- 700 patients, 350 take drug, 350 do not

	Drug		No Drug	
Men	81 of 87 recovered	93%	234 of 270 recovered	87%
Women	192 of 263 recovered	73%	55 of 80 recovered	69%
Combined	273 of 350 recovered	78%	289 of 350 recovered	83%

- Is the drug safe? Can we only prescribe it knowing gender (nonsensical)? What is the relevant causal effect?

Simpson's paradox

- 700 patients, 350 take drug, 350 do not

	Drug		No Drug	
Men	81 of 87 recovered	93%	234 of 270 recovered	87%
Women	192 of 263 recovered	73%	55 of 80 recovered	69%
Combined	273 of 350 recovered	78%	289 of 350 recovered	83%

- Is the drug safe? Can we only prescribe it knowing gender (nonsensical)? What is the relevant causal effect?
- Consider two details of the *data generating process* (common cause):
  - Estrogen has a negative effect on recovery (women are less likely to recover regardless of the drug)
  - More women take the drug compared to men
- A randomly selected drug user is more likely to be a women and less likely to recover
- The disaggregated data shows the true causal effect

Simpson's paradox

However, disaggregation does not always work to reveal the causal effect

- Extend drug experiment by recording blood pressure at end of experiment
- Drug effect recovery by lowering blood pressure (but also has a toxic side effect)

	No Drug		Drug	
Low BP	81 of 87 recovered	93%	234 of 270 recovered	87%
High BP	192 of 263 recovered	73%	55 of 80 recovered	69%
Combined	273 of 350 recovered	78%	289 of 350 recovered	83%

- Again, understanding of this pattern stems from understanding the data generating process
  - In general population, drug improves recovery by lowering BP
  - Among those with low or high (post-treatment) blood pressure, we only see the drug's toxic effect
  - Thus, correct answer lies in using the aggregate data

Causal inference and statistics

- Note that both tables contain the same numbers
- No statistical model can tell us if one or the other strategy should be used
- Extraneous information was the determinant (timing of the measurements, treatment affects BP, etc..)

Encoding such information is the role of a causal language – whether in potential outcomes or graphical form. Next week(s).

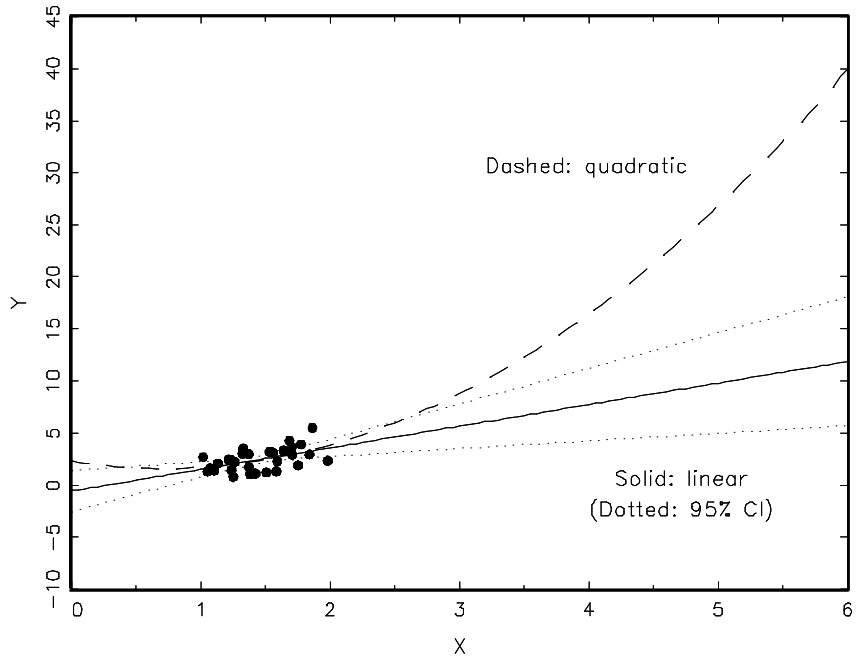
The problem of counterfactuals and model dependence

Counterfactuals

- Types of counterfactuals
  - Forecasts (will candidate win election?)
  - What-if (what would have happened if US had not signed free-trade agreement)
  - Effects of cause (what is the causal effect of protest on court decisions?)
- Central to scientific inquiry
- Usually based on models
- How to identify *sensible* counterfactuals?

Dangerous counterfactuals....

- Which model to choose? Both fit the data rather well



- Prediction at x = 1.5 vs. at x = 3 or x = 6

Dangerous counterfactuals....

- Some answers do not exist in the data
- Results / estimates of quantities of interest can be highly model dependent

Model dependence

- A model free estimate of $E(Y|X = x)$: average 'many' observed Y with value x
- Model dependence at x is a function of the distance from the counterfactual to the data

Model dependence illustration

A hypothetical experiment on education and earnings

- Take a random sample of children
- Assign them at random to obtain different years of schooling [0, 6, 10, 12, 16 years]
- (assume away many complications, such as non-compliance, measurement error, etc.)
- Outcome of interest: yearly earnings in year 20
- Regression of earnings on education yields estimate of 2,000 USD

Model dependence

- Suppose you have data on Y: earnings and X: education with 10 categories
- Estimation of $E(Y|X)$ uses 10 parameters: $E(Y|X = x_j)$, $j = 1, /dots, 10$
- Model-free estimate: average observations of Y for each x_j
- Model-based estimate: (typically) linear regression of Y on X
- Difference in parameters between the two approaches (10 vs 2) is simple due to an *assumption* (linear functional form)

[What if X were continuous instead of discrete?]

Model dependence, now with two explanatory variables

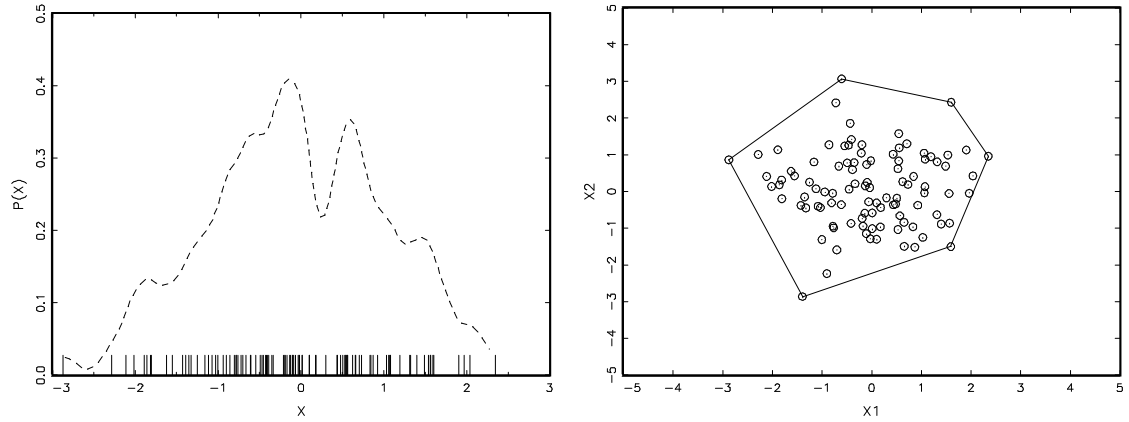
- Add new variable, Z: parental social class (with 7 categories)
- How many parameters do we need? (17? 70?)
- With more explanatory variables, model complexity (number of parameters) increases *geometrically*; a.k.a. the curse of dimensionality
- A linear regression as typically specified reduces this to 3 parameters
- Again, this reduction of the parameter space is (usually) arrived at via assumptions, not theory or prior evidence

Model dependence, realistically

- Common model specifications easily include 10 or more variables ("controls")
- 10 variables each with 5 categories
- Number of parameters: $5^{10} = 9,765,625$
- Simple linear regression? 11 parameters
- Number of observations?

Defensible simplicity?

- Is the counterfactual quantity produced by your statistical model close (enough) to the data to provide empirical answers?
- If no, answers likely depend on hard-to-defend model assumptions



- *Interpolation*: "inside" the data (convex hull)
- *Extrapolation*: "outside" the data (convex hull)

[WhatIf R package provides approx. convex hull checking]

Impact on naive linear regression analyses

- Decomposition of bias in a typical regression

$$\tau = mean(Y|D=1) - mean(Y|D=0)$$

$$bias = \Delta_o + \Delta_p + \Delta_i + \Delta_e$$

  - Δ_o: omitted variable bias
  - Δ_p: post-treatment bias
  - Δ_i: interpolation bias
  - Δ_e: extrapolation bias

Summa summarum

Model dependence

- Both *intrapolation* and *extrapolation* is possible and unproblematic if the statistical model is correct (or "true")
- That is unlikely to be the case. Less ambitiously, the model might be locally correct
- Thus likely to get very different counterfactuals the further we move away from the data.