Causal Inference

I used to think correlation implied causation.

Then I took a statistics class. Now I don’t.

Sounds like the class helped. Well, maybe.

https://xkcd.com/552/

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

SML310: Research Projects in Data Science, Fall 2018

Michael Guerzhoy
Directed Acyclical Graphs (DAG) for representing causal structure

- If we know the value of Smoking (0 or 1), we can generate the value of Yellow teeth and Tar in lungs
- If we know the value of Tar in lungs and Asbestos, we can generate the value of Cancer
- A datapoint is generated by first generating Smoking and Asbestos, then Yellow teeth and Asbestos, then Cancer
Directed Acyclical Graphs (DAG) for representing causal structure

- The graph encodes our knowledge (or assumptions) about the causal structure of the data
- Can help with inferences
Directed Acyclical Graphs (DAG) for representing causal structure

- Tar is independent of Asbestos
  - They are independently generated
- Tar is not independent of Asbestos given Cancer
  - Intuition: if Cancer = 1 and Asbestos = 1, then Tar = 1 is less likely than otherwise, since the cancer is already explained
  - This called “Explaining away”
Directed Acyclical Graphs (DAG) for representing causal structure

- Yellow teeth is not indep. of Tar
  - Both caused by Smoking
- Yellow teeth is indep. of Tar conditioned on Smoking
Directed Acyclical Graphs (DAG) for representing causal structure

- The right way to think of the arrows: *Tar might* cause *Cancer*, or have no effect
- *Smoking* is *definitely* independent of *Asbestos*
Causation

• Define “A caused B”
Causation

• Counterfactual: we say that $A$ causes $B$ if $B$ would not have happened if $A$ had not happened

• Causal inference: trying to answer causal questions from empirical data
  • Difficult to derive counter-factual conclusions from factual premises
Effect of causes

- **What is the causal relationship between exposure to asbestos and yellow teeth?**
  - There is none!
  - *Yellow* and *Asbestos* are not indep. conditioned on Cancer
    - Explaining away phenomenon
    - *Yellow~Asbestos + Cancer* will have significant coefficients for a large dataset
Effect of causes

- What is the causal relationship between exposure to asbestos and yellow teeth?
  - Want to know what to control for and not to control for
Effect of causes

A way of thinking about this:

\[ P(\text{Yellow}|\text{Asbestos} = 1, \text{Cancer}) \neq P(\text{Yellow}) \]
\[ P(\text{Yellow}|do(\text{Asbestos}), \text{Cancer}) = P(\text{Yellow}) \]

\textit{do(Asbestos)} sets \textit{Asbestos} to 1, and changes the causal graph eliminating the mechanism that generates \textit{Asbestos}
Effect of causes

- To do causal inference, compare $P(\text{Yellow}|\text{do}(\text{Asbestos} = 1))$ and $P(\text{Yellow}|\text{do}(\text{Asbestos} = 0))$ (and verify the model says they are the same)
- A logistic regression would give a significant coefficient for \textit{Asbestos} ("keeping the cancer diagnosis constant, exposure to Asbestos increases the log-odds of yellow teeth by 0.1")
Causal inference: Take a step back

• $P(Yellow|\text{do}(Asbestos = 1))$ is computable if we know how to generate the dataset
  • This is difficult!
  • Possible if we know the mechanisms that generate the data and if we study each step in the mechanism
    • On the whiteboard
do(brushing)

Again, \( P(Heart\ disease|Brushing = b) \neq P(Heart\ disease|do(Brushing = b)) \)
Identifying Causal Effects from Observations

• The most straightforward way to compute $P(Y|do(X = x))$ is to manipulate $x$ physically and see what happens to $Y$
  • Run an experiment
    • Hold all other variables constant
    Or
    • Randomize all other variables
Identification

• Want to calculate the causal effect of $X$ on $Y$ (i.e., $P(Y|do(X = x))$, but can’t run an experiment.

• Can do this if we have the causal graph and observe all the variables
  • Saw this before

• Can sometimes do this if not all variables are observed
  • Need to carefully look at the graph
Identification

- Do regression right
  - Control for variables which so that all the “explaining away” phenomena are eliminated
- Find all causal paths, and figure out exactly how they work
- Instrumental variables
Instrumental Variables

\[ P(Y|do(I = i)) = \sum_x P(Y|do(X = x))P(X = x|do(I = i)) \]

Intuition: manipulating I is like manipulating X since I causes X

**Figure 22.6:** A valid instrumental variable, I, is related to the cause of interest, X, and influences Y only through its influence on X, at least once control variables block other paths. Here, to use I as an instrument, we should condition on S, but should not condition on B. (If we could condition on U, we would not need to use an instrument.)
Example of Instrumental Variables

• X: smoking
• I: cigarette taxes
• Y: health

I can be "manipulated" (maybe) by looking at different states with different taxes. We can then claim to be doing causal inference.
Causal inference with Matching

• Match “like” objects which differ only on $X$
  • Lots of techniques
Summary

• Causal inference from observational data is sometimes possible when we can figure out the causal graph
  • This is very difficult in general

• Correlation really doesn’t imply causation
  • Often the best you can do is say “An increase in X is associated with an increase in Y”