Lecture 11: Guest Lecture
Modeling Social Data, Spring 2017
Columbia University

jyl2164

April 14, 2017

1 Simpson's Paradox

Selection bias can be so large that observational and causal estimates give opposite effects (e.g., going to hospitals makes you less healthy)

1.1 Comparing old vs. new algorithm

Two algorithms, A (production) and B (new) are running on the system. From the system logs, data is collected for 1000 sessions for each algorithm. Measure CTR. Which algorithm is better?

Old Algorithm (A)	New Algorithm (B)
50/1000(5%)	54/1000(5.4%)

Frequent users of the Store tend to be different from new users. So lets look at CTR separately. The Simpson's paradox:

CTR	Old Algorithm (A)	New Algorithm (B)
CTR for Low-Activity users	10/400(2.5%)	4/200(2%)
CTR for High-Activity users	40/600(5%)	50/800(6.2%)
Total CTR	50/1000(5%)	54/1000(5.4%)

Is Algorithm A better? Answer (as usual): may be, may be not. Algorithm A could have been shown at different times than B. You could have also targeted it to people with high activity. There are also much more people in the new algorithm. Essentially, you can play around and find different algorithms that work well. There are many other hidden casual variations, like income level and stratification (we think that the thing we are changing will have different effects on different groups).

1.2 Example: Simpson's paradox in Reddit

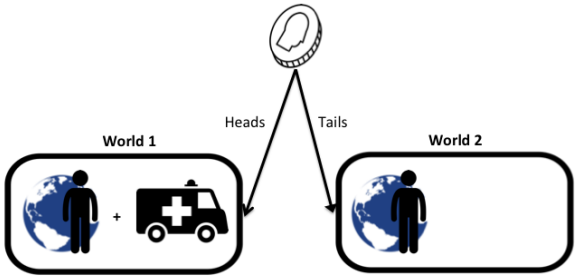
If you look at the average number of words on Reddit and condition on different years you will notice that it is decreasing over time. That is, the average comment length is decreasing over time, which could be a worrying sign if you are Reddit. This would mean that Reddit would need to find ways to make people write more. However, if you condition on when a person joined Reddit you will find that the number of comments increases over time. For some reason people who join later seem to comment less and the early adopters of Reddit are the most active. Now Reddit's problem changes. They don't want people to write more, what they want is to attract more of those "good" people.

2 Counterfactuals

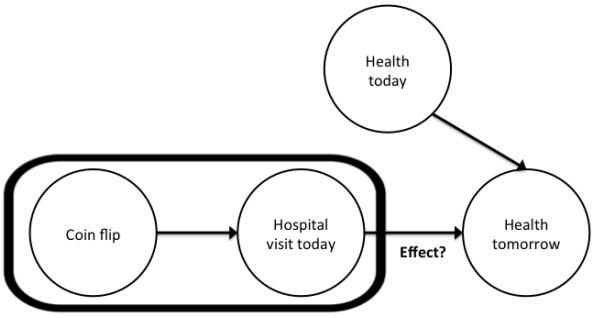
To isolate the causal effect, we have to change one and only one thing (hospitals visits), and compare outcomes. That is, in an ideal world we would take the world and make a clone world. In one case we would bring the guy to the hospital and in the other case we wouldn't. The only thing that differs between the two cases is whether or not we brought the guy to the hospital, everything else is the same. This is the core idea of what we try to do in an experiment. However, in practice it hard to do an experiment and say that world 1 and world 2 are equal except for the thing that we changed.

3 Random assignment

We can use randomization to create two groups that differ only in which treatment they receive, restoring symmetry. For each person, we will flip a coin and randomly assign them to going to the hospital or not going to the hospital. There are going to be healthy and sick people in each group. On average the people who get sent and not sent to the hospital is the same, i.e. those groups should be identical because nothing else affected whether or not they went to the hospital. We can then measure the effect of people getting treated at hospitals or not.



Random assignment determines the treatment independent of any confounds.



4 Basic identity of causal inference

The observed difference is now the causal effect: Observed difference = Causal effect - Selection bias = Causal effect
Selection bias is zero, since there's no difference, on average between those who were hospitalized and those who weren't.

5 Experiments

5.1 Caveats/limitations

Random assignment is the "gold standard" for causal inference, but it has some limitations:

- Randomization often isn't feasible and/or ethical (often happens in medical settings)
- Experiments are costly in terms of time and money
- It's difficult to create convincing parallel worlds
- Inevitably people deviate from their random assignments

5.2 Two goals for experiments

1. Internal validity: Could anything other than the treatment (i.e. a confound) have produced this outcome?
    - Did doctors give the experimental drug to some especially sick patients (breaking randomization) hoping that it would save them?
2. External validity (Generalization): Do the results of the experiment hold in settings we care about?
    - Would this medication be just as effective outside of a clinical trial, when usage is less rigorously monitored?

5.3 How we conduct behavioral experiments and write academic papers across the social sciences

1. Lab Experiment - what are held at colleges
    - Better internal validity (correctness): Greatest procedural control, can carefully curate situations
    - But less external validity (generalization): Artificial context, simple tasks, demand effects, homogeneous (WEIRD - Western Educated Industrialized Rich Democratic) subject pools, time/ scale limitation - You cant keep people there and you can only bring in a certain number of people
2. Field Experiment - run an experiment in the real world
    - Better generalization: Experiment findings apply to at least one real-world setting
    - But: Less control, more potential confounds, demand of experiment conflict with goals of real organizations, more effort to conduct and manage, more room for error

These kinds of issues also come up in systems as well. Even in the systems that we build ourselves it is not that simple. Internal validity is a problem at companies like Microsoft and Google. If your system has an A/B test you should wait at least 6 months because there are so many things that happen before your code reaches the user. Generalization also comes in. Suppose you run your code and it works well in New York, will it also work well in another city, in another country?

5.4 Natural experiments

Sometimes we get luck and nature effectively runs experiments for us e.g:

- As-if random: People are randomly exposed to water sources
- Instrumental variables: A lottery influences military service
  - Idea: An instrument independently shifts the distribution of a treatment
- Discontinuities: Star ratings get arbitrarily rounded
- Difference in differences: Minimum wage changes in just one state

5.5 Behavioral science labs are very limiting

Pros:

- High degree of procedural control, Optimized for causal inference

Cons (limitations):

- Artificial environment, Simple tasks, demand effects, Homogeneous (WEIRD) subject pools, Time/scale limitations, Expensive, difficult to set up

5.6 Experiments are underpowered

Two-thirds of psychology studies don't replicate! This shows how hard it is to do experiments properly. And if you do do it properly you should be able to run the experiment again and get the same results.

5.7 Most social science experiments aren't social

The vast majority of experiments are done with individuals as it is harder to do experiments on groups of people. Therefore, experiments are focused on individual behavior: logistically it's just easier. We actually don't know the causal effects of policies in many large-scale, collective behavior settings. Some examples include: economic inequality, systems of governance, and the black box of macroeconomic policy.

5.8 So, about them experiments

They are costly run but are limited in the types of questions they can answer. Large-scale observational data is useful for building predictive models of a static world. Randomized experiments are like custom-made datasets to answer a specific question. Moreover, published experimental research is probably wrong and they are far from answering many big important questions. So how do we fix this? By expanding the experiment design space.