Segment 4: Observational Studies

Section 01: What is an Observational Study?

Operating Question: Segment 4

So I didn't do an experiment...how can I still estimate causal effects with my data?

Why Randomization Was So Important

In randomized experiments, the *assignment mechanism*

$$Pr(\mathbf{Z}|\mathbf{X},\mathbf{Y}^c,\mathbf{Y}^t) = Pr(\mathbf{Z}) \text{ or } Pr(\mathbf{Z}|blocks)$$

was known because a study was prospectively designed an implemented with control over which units were assigned to which treatments

Therefore, we *knew* the following:

- Ignorability: $Y^t, Y^c \perp\!\!\!\perp Z$
  - Block Randomized: $Y^t, Y^c \perp\!\!\!\perp Z|blocks$
- Design would achieve *balance* between treatment groups
  - Y^t, Y^c, X
- Straightforward to estimate causal effects
  - Outcomes in one treatment group were "good substitute" for what would have happened to the units in the opposite treatment group

But we can't always randomize...

But there are *many* reasons why we may be interested in a problem where we cannot control the assignment mechanism:

- Ethical concerns, practical considerations, financial constraints....
- For whatever reason, we observe data on units that receive different treatments, but not necessarily due to randomization

This means that the assignment mechanism is *unknown*

$$Pr(\mathbf{Z}|\mathbf{X}, \mathbf{Y}^c, \mathbf{Y}^t) = ??$$

The following then are not guaranteed

- ~~Ignorability: $Y^t, Y^c \perp\!\!\!\perp Z$~~
- ~~Design would achieve *balance* between treatment groups~~
- ~~Straightforward to estimate causal effects~~

Observational Studies

In an observational study, the assignment mechanism:

$$Pr(\mathbf{Z}|\mathbf{X}, \mathbf{Y}^c, \mathbf{Y}^t)$$

is *unknown* because we never actually controlled it in an experiment

- Could be prospective or retrospective
- Might the treatment assignment have been *ignorable*?
  - $Y^t, Y^c \perp\!\!\!\perp Z$
- Is there any reason to expect (Y^t, Y^c) are *balanced* between treatment groups?
- Was the "treatment" actually something that was manipulated?

Example: Smoking and Lung Cancer

Clearly unethical to randomize people to **smoke vs. not** in order to study the effect on lung cancer

- Obtain observed data on smokers and nonsmokers
- Measure the lung cancer rates in each group
- Without randomization, what are some of the differences between smokers and nonsmokers that could distort any comparison?
  - Smokers tend to be older
  - Smokers tend to have less healthy lifestyles
  - ...
- Is smoking "ignorable"?
  - Do we think potential outcomes would be balanced across smokers and nonsmokers?
  - Is lung cancer in nonsmokers a "good substitute" for what we think would happen to smokers if they didn't actually smoke Y^0?

Example: Fellowships and Faculty Promotion

Junior faculty at a university can apply for a **research fellowship** to support their work, with the receipt of the fellowship decided based on a combination of merit and need-based criteria.

Does receipt of the fellowship cause faculty to be promoted earlier than those who did not receive the fellowship?

- Without randomization, what are some of the differences between faculty who did and did not receive the fellowship?
  - Those who receive the fellowship may tend to be in certain types of departments where it is easier to get promoted
  - Those who receive for the fellowship are high achieving faculty who are likely to be promoted anyway
  - Those who receive the fellowship have some hardship, making it less likely that they'll get promoted regardless of the fellowship
- Can we conclude that observed promotion rates in fellowship recipients (vs. non-recipients) represent an effect of the fellowship?

Example: Air Pollution

People are exposed to different levels of **outdoor air pollution** depending on where they live.

Does living in an area with high (vs. low) levels of air pollution cause cardiovascular disease?

- Highly polluted areas my also be in hotter climates
- Cities tend to be more polluted than rural areas, and people who live in cities have different health-related behaviors
- Certain regions of the country have higher pollution, and these regions also have different demographic makeups than other less polluted parts of the country

Would we conclude that the higher rates of cardiovascular disease in polluted areas are actually *caused* by the pollution?

Correlation is Not Causation!

without randomized assignment

In all of the previous examples, any correlation between Z, Y is not likely to represent a causal effect

- Smokers have higher rates of lung cancer...
  - But is it *because* of the smoking or attributable to other lifestyle differences?
- Fellowship recipients get promoted faster than non-recipients...
  - But is it *because* of the fellowship or because the people who got the fellowship were particularly talented high achievers?
- Polluted areas have higher rates of cardiovascular disease...
  - But is it *because* of the pollution or is it because unhealthier people tend to live in highly polluted areas?
- These are all examples of *confounding*

Systematic Differences Between Comparison Groups

- Without randomization, observational studies may exhibit *systematic* differences between the treatment and control groups
  - Not just "unlucky" assignments
  - Differences would persist "on average" over repeated sampling
- Difficult to disentangle whether observed differences are due to treatment or systematic differences
  - Outcomes may be different across groups even when there is zero causal effect
  - Outcomes may be the same across groups even when there is a causal effect
- Need to have some ways to "adjust" for these systematic differences in order to recover estimates of causal effects

Frame Observational Studies as (Approximate) Randomized Experiments

Motivating Principle: View observational studies *approximate* randomized studies

- Randomized studies are the "gold standard" for estimating causal effects
- Try to relate circumstances of the observational study to those of an experiment
  - Language of experiments: treatment/control groups, treatment assignment, balance, etc.
  - Reason about the *assignment mechanism*
  - Think about what an "ideal" experiment *might have* looked like for the problem
  - Evaluate how well the circumstances of the observational study might approximate a randomized experiment
- Is the observational study an *approximate* experiment?

⇒ Leverage methods from randomized studies