Segment 4: Observational Studies

Section 03: Ignorability and the Relationship with Randomized Studies

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?

$\Rightarrow$ Leverage methods from randomized studies

Ignorability and Unconfoundedness

In observational studies, the **key assumption** about whether the observed data can be used to estimate causal effects will be whether the assignment mechanism is (conditionally) ignorable/unconfounded:

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

$$Z \perp\!\!\!\perp Y^c, Y^t | X$$

- Will not be known by design, will have to be **assumed**
- Do we have all of the right X such that, conditional on X, the treatment is as though it had been randomly assigned?
- The X needed to satisfy the unconfoundedness assumption are called *confounders*
  - Can conditioning on X "recreate" a block randomized study?

Conditional Ignorability and Confounding

Earlier we informally defined a *counfounder* as a pre-treatment variable that:

1. Is associated with Z "blame"
2. Is associated with Y "predictive"

More formally, *confounders* are covariates required to satisfy

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

$$Z \perp\!\!\!\perp Y^c, Y^t | X$$

Note: definition is for a specific treatment/outcome. If a single study has multiple outcomes, a confounder for one outcome may not be a confounder for another.

Assumption of (Conditional) Ignorability

Causal validity of an observational study will rely heavily on whether the **assumption** of conditional ignorability is plausible

- Viability of the assumption relates to what we know (or assume) about the assignment mechanism
- Cannot be empirically verified
- Goes by many other names
- Has key (and familiar) implications for estimating causal effects with observed data

What Do We Know About the Assignment Mechanism?

In order to assess the circumstances under which ignorability:

$$Z \perp\!\!\!\perp Y^c, Y^t | X$$

holds, we need to understand *why* certain units received treatment vs. control

- The more we understand about why certain units were treated, the more we can reason about ignorability
  - What are the right confounders?
  - Did we measure them?
  - If not, did we measure anything that is a reasonably proxy?
- E.g., Why do we think doctors tend to assign treatment A to some people and treatment B to others?
- E.g., What made some teachers decide to supplement (vs. replace) the reading program?
- E.g., What might make a person choose to take a dietary supplement vs. not?

Ignorability Cannot be Empirically Verified

In an observational study, since we don't *know* the assignment mechanism, we can only *assume* that it is ignorable:

- Because this conditional independence depends on *potential outcomes* we can never "check" if it holds
- Even if we think we have all of the right X to satisfy this assumption, we cannot rule out some other X that we didn't think about
- E.g., May assume that doctor prescribes a treatment based on recorded "previous health status," but the decision might also be based on other intuition about patient's health
- E.g., May assume that teachers choose to supplement (vs. replace) the reading program based on pre-test scores, but maybe the teachers with the most experience were more likely to supplement

---

Ignorability, AKA...

$$Z \perp\!\!\!\perp Y^c, Y^t | X$$

- Selection on observables
  - Observed factors dictate which units *select* treatment
- Conditional independence
- Exchangeability
  - Conditional on covariates, units across treatment groups are exchangeable with one another
- No unmeasured confounders
- Violations of ignorability: hidden bias, omitted variable bias, unmeasured confounding

Implications of Ignorability for Analysis

$$E[Y|Z=1, X=x] = E[Y^1|Z=1, X=x] = E[Y^1|X=x]$$ (ignorability)

$$E[Y|Z=0, X=x] = E[Y^0|Z=0, X=x] = E[Y^0|X=x]$$

↑ conditional ↑

↓ marginal ↓

$$E[Y^1 - Y^0] = E[Y^1] - E[Y^0]$$

$$E[Y^1] = E[Y^1|Z=1, X=1]P(X=1) + E[Y^1|Z=1, X=0]P(X=0)$$

$$E[Y^1] = \sum_x E[Y^1|Z=1, X=x]P(X=x) = E_x E[Y^1|Z=1, X=x]$$

Ways to Analyze Observational Studies

Under the assumption of conditional ignorability $Z \perp\!\!\!\perp Y^c, Y^t | X$

Two ideas that we have already seen:

1. Subclassification
   - Create strata of the sample based on $X = x$
   - $\tau_{CATE|X=x} = E[Y|Z=1, X=x] - E[Y|Z=0, X=x]$
   - $\tau_{SATE} = \sum_{\mathbf{x}} \tau_{CATE|X=x} Pr(X=x)$
   - But what if there are many possible values of X?

2. Regression
   - $E[Y^z|\mathbf{X}] = \beta_0 + \mathbf{X}\beta + \tau Z$
   - $\hat{\tau}$ represents, in principle, a weighted average of $\hat{\tau}_{CATE|X=x}$
   - This is not fool proof...requires a bit more conditions that held by design in randomized studies but may not hold in observational studies