Stat 215A - Week 11

Causal inference - Zoe Vernon
Based on notes from Peng Ding's 239A course

Upcoming schedule

Today: causal inference basics

Friday Nov. 9

- ❏ Lab 4 due 9am
- ❏ Midterm review (email me what questions you have)

Tuesday Nov. 13: midterm

Friday Nov. 16: final project assigned (3 weeks)

Neyman-Rubin causal model

Motivation: Simpson's paradox

"Correlation does not imply causation"

Many ways to measure association, but are often hard to interpret.

Consider the example of kidney stone treatment (Charig et.al 1986)

- Get a risk difference (RD) of -5% (think less aggressive treatment is better)
- After conditioning on size of kidney stones
  - RD for small stones: 6% > 0: more aggressive surgery is better
  - RD for large stones 4% > 0: more aggressive surgery is better
- What could be causing this to happen?

Neyman-Rubin causal model

Rubin: “no causation without manipulation”

Need a well defined intervention. In this case of observational studies this may be hypothetical.

Can we study the causal effects of race or gender?

Neyman-Rubin causal model

Rubin: “no causation without manipulation”

Need a well defined intervention. In this case of observational studies this may be hypothetical.

Can we study the causal effects of race or gender? Not in general, from this perspective it is ill-defined.

Neyman-Rubin causal model: notation

Experimental units: $i = 1, \ldots, n$

Treatment levels (e.g. patient receives aspirin or not): $\begin{cases} 1 \\ 0 \end{cases}$

Outcome of interest, *Y*, (e.g. number of headaches in a week) hypothetically can take two **fixed** values

1. $Y_i(0)$ if unit *i* receives treatment
2. $Y_i(1)$ if unit *i* receives control

Neyman-Rubin causal model: hidden assumptions

Stable Unit Treatment Value Assumption (SUTVA)

1. **No interference**: unit *i*'s potential outcomes do not depend on any other unit's treatment
2. **No ambiguity of treatment**: there are no other versions of the treatment

Neyman-Rubin model: causal effects

Our goal is to estimate the causal effect

Individual level effects:

$$\tau_i = Y_i(1) - Y_i(0)$$

Average treatment effect:

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \{Y_i(1) - Y_i(0)\}$$

Neyman-Rubin model: what do we observe?

Treatment assignment: $Z = (Z_1, \ldots, Z_n)$

Observed outcome: $$Y_i = \begin{cases} Y_i(1) & Z_i = 1 \\ Y_i(0) & Z_i = 0 \end{cases} = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$$

The **key component** is the probability distribution of the treatment assignment indicators. This is the only source of randomness.

Neyman-Rubin causal model: randomized experiment

In a completely randomized experiment (with fixed number of treated units)

$$\mathbb{P}(Z = z) = \binom{n}{n_1}^{-1}$$

Total number of units

Number of treated units

Due to randomization

1. Treatment group approximately the same as the control group.
2. Reasoned basis for statistical inference

Neyman-Rubin model: how do we make inference?

Two frameworks

1. Fisher Randomization Test (FRT)
    a. Explicitly uses our knowledge of the randomization procedure
2. Neymanian repeated sampling
    a. More interested in estimation
    b. Use CLT for inference

Neyman-Rubin model: FRT

Test **sharp null hypothesis** that the potential outcomes are the same under treatment and control for all units. That is for a test statistic T

$$T = T(Z, Y(1), Y(0)) = T(Z, Y)$$

We can compute T under all possible realizations of our treatment assignment (or a random sample if *n* is large) and compare to our observed value.

$$p = \mathbb{P}(T \geqslant T_{obs})$$

Neyman-Rubin model: FRT

Example in FRT.R

Neyman-Rubin model: Neymanian inference

Estimate average treatment effect (unbiased for true ATE)

$$\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}(1 - Z_i)Y_i = \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i(1) - \frac{1}{n_0}(1 - Z_i)Y_i(0)$$

Estimate variance

$$\hat{V} := \widehat{\mathrm{Var}}(\hat{\tau}) = \frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n} Z_i(Y_i - \bar{Y}_1)^2$$

$$s_0^2 = \frac{1}{n_0 - 1}\sum_{i=1}^{n}(1 - Z_i)(Y_i - \bar{Y}_0)^2$$

Apply CLT to get confidence intervals and compute p-values

Neyman-Rubin model: randomized experiments cont'd

There is lots more to consider with randomized experiments under this model...

We can use different statistics for FRT

We can have different designs... like stratifying our experiment (e.g. a male group and a female group and do separate analyses)

We can adjust for covariates post-randomization through regression

Causal inference in observational studies

Observational studies

When we don’t randomize, what is issue with using the same estimators as a randomized experiment (e.g. difference in means between treatment and control)?

Observational studies

When we don’t randomize, what is issue with using the same estimators as a randomized experiment (e.g. difference in means between treatment and control)?

The differences between the treatment and control groups may be the cause of the difference in outcomes

We can correct for this difference in measured covariates

But how can we deal with unmeasured confounding?

Observational studies

When we don’t randomize, what is issue with using the same estimators as a randomized experiment (e.g. difference in means between treatment and control)?

The differences between the treatment and control groups may be the cause of the difference in outcomes

We can correct for this difference in measured covariates

But how can we deal with unmeasured confounding? We assume it away...

Observational studies: additional assumptions

Assumption 1: $\{X_i, Z_i, Y_i(1), Y_i(0)\}_{i=1}^n$

Assumption 2 (ignorability): $Z_i \perp\!\!\!\perp \{Y_i(1), Y_i(0) | X_i\}$

Assumption 3 (overlap): $\exists \eta > 0$ such that $\eta \leqslant \mathbb{P}(Z_i = 1 | X_i) \leqslant 1 - \eta$ for $\eta \in (0, 1/2)$

Observational studies: additional assumptions

Under the previous assumptions we can treat our data as a randomized experiment conditional on the the covariates.

This is because when we condition on $\boldsymbol{X}$ we are assuming that the treatment is independent of the potential outcomes

Essentially, we are assuming away the unmeasured confounding issue

Observational studies: estimating causal effect

Because we are considering our data to be random draws from a super population we are now estimating

$$\mathbb{E}\left\{Y_i(1) - Y_i(0)\right\}$$

Under the three assumptions there are a number of ways estimate the causal effect by adjusting for the measured covariates

See observational_study.R for examples