Segment 4: Observational Studies

Section 02: Confounding and Omitted Variable Bias

Example: Zero Causal Effect

But treatment predicts outcome

- 100 patients receive new medical treatment, 100 receive control
- Healthier patients tend to receive the treatment, sicker tend to receive the control
- Treatment indicator can be predictive of the outcome even if there is zero causal effect

- Health status *confounds* the estimate of the treatment effect $\Rightarrow$ bias

Example: Confounding

In both cases, the problem is that:

1. Previous health status is related to who gets treatment
   - Previous health status is not *balanced* across treatment groups
2. Previous health status is related to future health status
   - Observed association between Y and Z is attributable (at least in part) to previous health status (and not any causal effect)

$\Rightarrow$ The comparison between treated/non treated units is *confounded* by previous health status

Example: Adjusting for "Previous Health"

- Simple comparisons health status between treatment groups do not estimate a causal effect of treatment
  - Confounded by previous health status
- But previous health status is observed in the data
- Can we conduct our treated/control comparisons *within* patients who have the same previous health status?
  - Like a randomized block design that blocks on previous health status
  - Could be accomplished with a regression model
- $Z \not\perp\!\!\!\perp Y^t, Y^c$
- $Z \perp\!\!\!\perp Y^t, Y^c |$ "previous health"?

"blocks"

Omitted Variable Bias

Treatment Covariate

True model: $Y_i = \beta_0 + \beta_1 Z_i + \beta_2 x_i + \varepsilon_i$

$\downarrow$

τ_{SATE}

Fitted model: $Y_i = \beta_0^* + \beta_1^* Z_i + \varepsilon_i^*$

What is the relationship between "truth" and "fitted"

Without randomization, $\beta_1 \neq \beta_1^*$

Omitted Variable Bias

Relationship between β_1 and β_1^* depends on:

1. Relationship between X and Y "predictive"
2. Relationship between X and Z "imbalance"

Assume: $X_i = \gamma_0 + \gamma_1 Z_i + v_i$

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 (\gamma_0 + \gamma_1 Z_i + v_i) + \varepsilon_i$$

$$= \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1) Z_i + \varepsilon_i + \beta_2 v_i$$

$$(\beta_0^*) + (\beta_1^*) Z_i + (\varepsilon_i^*)$$

$$\beta_1^* = (\beta_1 + \beta_2\gamma_1)$$

Omitted Variable Bias

"bias"

$$\beta_1^* = \beta_1 + \beta_2\gamma_1$$

- β_1 is the "true" causal effect
- $\beta_2\gamma_1$ is the bias due to leaving out the confounder X
  1. β_2 describes the relationship between X and Y
     - What happens if $\beta_2 = 0$?
  2. γ_1 describes the relationship between X and Z
     - What happens if $\gamma_1 = 0$?
- So, in order for the omission of X to induce bias in the estimation of the effect of Z, both (1) and (2) above must be nonzero

$\Rightarrow$ **Working Definition:** A *confounder*, which is a variable that:

1. Is associated with Z
2. Is associated with Y

Adjusting for Confounding in Observational Studies

- We know that, without randomization, there can be systematic differences between units across treatment groups
  - Systematic differences in *confounders* will distort the interpretation of observed-data comparisons as causal effects
  - Imbalance on confounders $\Rightarrow$ Imbalance in potential outcomes
  - Can't separate differences due to confounding factors from differences due to the treatment effect
- To estimate causal effects, we need some way to adjust for confounding factors
  - Try to recreate balance
- When confounders are *observed* in X
  - Treat balance in X as proxy for balance in Y^t, Y^c
- When confounders are *unobserved*....

Adjusting for Confounders

In general, there could be many confounders....

- A regression model *could* adjust for all of the confounders to estimate a causal effect
- But we would need to have actually *observed* all of the confounders in the data
- Unlike in randomized studies, the regression model has to be "correct"
  - (other circumstances would need to hold as well, more later)

But maybe we didn't (or couldn't) include all of the confounders in the model....then what?

Revisiting the Electric Company Example

Recall: Randomized experiment of an educational TV program (vs. control) on elementary school classrooms investigated for its effect on reading scores

But there's actually an **observational** study embedded in the experiment

- Among those assigned to the TV program, classrooms had the option to:
  - *Replacing* the standard reading program ($Z = 0$)
  - *Supplementing* the standard reading program ($Z = 1$)
  - **Not** randomized!
  - Based on the decision of the teacher
- Question: Did supplementing (vs. replacing) cause higher post-test reading scores (Y)?
- Covariate: Pre-test reading score (X)

Confounding in the Electric Company Example

- Adjusting for pre-test (X) $\approx$ a block-randomized assignment
  - Imagine having grouped classrooms into blocks according to pre-test score, and then randomizing to supplement or replace
  - Regression adjustment approximates having done this for every level of X
- Assumes that pre-test score is the *only* confounder
  - Only thing that relates to decision to supplement (vs. replace) and have any bearing on test scores
- **Validity** of this assumption relates to the true assignment mechanism
  - The true information that dictated the decision to supplement vs. replace
  - We don't know this!
  - Could be a very strong assumption!
  - We may know little about how these decisions were actually reached
  - E.g., decision could have been based on teacher experience, classroom temperament, motivation, etc.

Unmeasured confounders