Potential Outcomes Simulation Homework

Andrea Cornejo, Ray Lu & Zarni Htet

Objective

In this exercise, you will be tasked with simulating an intervention study with a pre-determined average treatment effect. The goal is for you to understand the **potential outcome framework**, and the properties of **completely randomized experiments** through simulation.

Problem Statement

The goal is to simulate a data set with a treatment effect of τ = 5.

The setting for our hypothetical study is Professor Jennifer Hill's Casual Inference class. After the first attempt at Quiz I, Professor Hill decides to give students an opportunity to take the quiz again. Before the second attempt of the quiz, Professor Hill randomly assigns half the class to attend an extra tutoring session with graduate students (who are equally talented in **All Ways**: Andrea, Ray and Zarni) to half of the class. The other half of the class does not receive any additional help. Consider the half of the class that receives home tutors as the treated group. The goal is to estimate the effect of the extra tutoring session on average test scores for the retake of Quiz 1. We are assuming that SUTVA is satisfied.

Question 1: Calculating ATE (all seeing/omniscient)

For this section, you are a god of Statistics. That is, assume you are omniscient. You know the potential outcome of **Y(0)** and **Y(1)** for everyone.

(a). Please simulate a dataset consistent with the assumptions below while demonstrating an average treatment effect (ATE) of approximately **5**.

Simulation assumptions

The Data Generating Process (DGP) has the following features:

- Population size N is 1000.
- The pretest causal quiz I score is independent and identically distributed with a Normal distribution with mean of 65 and standard deviation of 3.
- The potential outcomes for the post-treatment causal quiz I score should be linearly related to the pretest quiz score. In particular they should take the form:

$$Y(0) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} X + 0 + \boldsymbol{\epsilon}$$

$$Y(1) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} X + \boldsymbol{\tau} + \boldsymbol{\epsilon}$$

where β_0 is the intercept taking the value of **10**. β_1 is set to **1.1.** Set τ to be 5 and draw ϵ from a N(0,1) distribution. Please also set seed at 1234 before generating these.

Answer the following questions based on the DGP or using your simulated data set. Remember that you are still all-seeing.

(b) What is your interpretation of tau?

Students who received the treatment effect (home tutors) have a score of 5 points higher on average than had the students **NOT** received the treatment.

(c) Please calculate SATE.

(d) Why is SATE different from tau?

(e) How would you interpret the intercept in the DGP for Y(0) and Y(1)?

(f) How would you interpret the β_1 coefficient?

Question 2: Estimating ATE (not all seeing/researchers'view)

For Questions 2 and 3, you are a **mere** researcher! Return your god-vision goggles and use only the data available to the researcher (that is, you will not have access to the counterfactual outcomes for each student).

(a) Using the same simulated dataset used in the previous case where τ = **5**, please randomly assign students to treatment and control groups. Then, create the observed data set which must include pretest scores, treatment assignment and observed y.

- Hint: sample() is the command in R to draw out random samples.

(b). You can also use rbinom function to assign treatment. What's the difference between rbinom and sample function?

(c) Now please estimate SATE using a difference in means.

(d) Is this estimate close to the true SATE? Divide the difference between SATE and estimated SATE by the standard deviation of the observed outcome, Y.

(e) Why is $S\hat{A}TE$ different from SATE and τ ?

Question 3: Use Linear Regression to estimate the treatment effect

(a) Now we will use linear regression to estimate SATE for the observed data set created by Question 2. With this set up, we will begin to better understand some fundamental assumptions crucial for the later R homework assignments.

(b) What is gained by using linear regression to estimate ATE instead of the mean difference estimation from above?

(c) What assumptions do we need to make in order to believe this estimate?

Challenge Question (optional): Treatment Effect Heterogenity (a). Based on the following function: Simulate the following response surfaces E[Y(0) | X] and E[Y(1) | X] and plot them. Also simulate Y(0) and Y(1).

Note: X is the same pretest score used before.

$$\begin{aligned}
\mathrm{E}[Y(0) \mid X] &= \boldsymbol{\beta_0^0} + \boldsymbol{\beta_1^0} X \\
Y(0) &= \mathrm{E}[Y(0) \mid X] + \boldsymbol{\epsilon}^0 \\
Y(0) &= \boldsymbol{\beta_0^0} + \boldsymbol{\beta_1^0} X + \boldsymbol{\epsilon}^0 \\
\mathrm{E}[Y(1) \mid X] &= \boldsymbol{\beta_0^1} + \boldsymbol{\beta_1^1} X \\
Y(1) &= \mathrm{E}[Y(1) \mid X] + \boldsymbol{\epsilon}^1 \\
Y(1) &= \boldsymbol{\beta_0^1} + \boldsymbol{\beta_1^1} X + \boldsymbol{\epsilon}^1
\end{aligned}$$

where β_0^0 is set to **35**, β_1^0 is set to .6, β_0^1 is set to **15**, β_1^1 is set to 1. First generate a vector of predicted Y(0) and Y(1) (that is E[Y(1) | X]. Then generate Y(0) and Y(1) with noise added as ϵ^0 or ϵ^1 from a distribution of N(0,1). Again, please also set seed at 1234.

(b) Comment on your findings. In particular, note that there is no longer a tau included in the DGP. Is there still a SATE? Can we calculate SATE? What is it? How do we interpret the average treatment effect in this setting?

(c) Is the treatment effect the same for all students? If not, is there a pattern to the way it varies? Why do we care about treatment effect heterogeneity?

(d) Now generate a similar plot from the initial DGP in Question 1 to reinforce the differences between a setting with constant treatment effect and a setting with heterogeneous treatment effects.