Randomized Experiment Simulation Homework

Andrea Cornejo, Ray Lu & Zarni Htet

Introduction

Randomized experiments are called the "gold standard" due to their ability to unbiasedly answer causal questions. This is achieved by creating two (or more) groups that are virtually identical to each other on average, in terms of **distribution** of all pre-treatment variables. So, if each group receives a different treatment and the groups have different outcomes, we can safely attribute these differences due only to the systematic difference between groups: the treatment.

In a randomized experiment, units are assigned to treatments using a known probabilistic rule. Each unit has nonzero probability of being allocated to each assignment class. In class, we went through two major types of randomized experiments based on different assignment rules: **completely randomized assignment** and **randomized block assignment**.

Problem Statement 1

Recall that in Assignment 1 we created a simulated dataset that could have manifested as a result of a completely randomized experiment. In that assignment, we asked about the difference between estimating ATE by using the difference in means versus using linear regression with pretest score as a covariate. In this exercise, we will explore the properties of these two different approaches to estimating our ATEs more deeply through simulation. We would like you to compare these approaches with respect to both **bias** and **efficiency**.

(a) Write a function to generate the data generating process (DGP) from Assignment 1 with arguments for sample size, the coefficient on the covariate, and the random seed. Then use this function to simulate a data set with sample size equal to 100, seed equal to 1234, and the coefficient on the covariate set to 1.1.

(b) We will now investigate the properties of two estimators of the SATE.

- difference in means
- linear regression estimate of the treatment effect using the pretest score as a covariate

For now we will only consider the variability in estimates that would manifest as a result of the randomness in who is assigned to receive the treatment (this is sometimes referred to as "randomization based inference"). Since we are in Statistics God mode we can see how the observed outcomes and estimates would change across a distribution of possible treatment assignments. We simulate this by repeatedly drawing a new vector of treatment assignments and then for each new dataset calculating estimates using our two estimators above. We will use these estimates to create a "randomization distribution" (similar to a sampling distribution) for these two different estimators for the SATE. Obtain 100,000 draws from this distribution. [Hint: Note that the only thing that will be different in each new dataset is the treatment and observed outcome; the covariate value and potential outcomes will remain the same.]

(c) Plot the (Monte Carlo estimate of the) randomization distribution for each of the two estimators: difference in means and regression. Either overlay the plots (with different colors for each) or make sure the xlim on both plots is the same. Also add vertical lines (using different colors) for the SATE and the mean of the randomizaton distribution.

(d) What is the bias and efficiency of each of these two methods? What is the difference between them?

(e) Re-run the simuation with a small coefficient (even 0) for the pretest covariate. Does the small coefficient lead to a different bias and efficiency estimate compared to when the coefficient for pretest was at **1.1** from before?

Problem Statement 2

In a randomized block design, randomization occurs separately within blocks. In many situations, the ratio of treatment to control observations is different across blocks. In addition, the treatment effect may vary across sites. For this problem, you will simulate data sets for a randomized block design that includes a binary indicator for female as a blocking variable. You will then estimate the ATE with two estimators: one that accounts for the blocking structure and one that does not. You will compare the bias and efficiency of these estimators. We will walk you through this in steps.

(a) First simulate the blocking variable and potential outcomes. In particular:

- Set the seed to by 1234
- Generate female as blocking variable (Female vs. Other Ratio (30:70)
- Generate Y(0) and Y(1) with the following features: – the intercept is 70 – the residual standard deviation is 1.
  – treatment effect varies by block: observations with female=1 have treatment effect of 7 and those with female=0 have a treatment effect of 3. [Hint: Note that we are assuming that being female affects treatment effect but does not directly affect the average test score otherwise.]

(b) Calculate the overall SATE and the SATE for each block

Now create a function for assigning the treatment In particular, within each block create different assignment probabilities:

- Pr(Z=1 | female=0) = .6
- Pr(Z=1 | female=1) = .4

Generate the treatment and create a vector for the observed outcomes implied by that treatment.

We will use this to create a randomization distribution for two different estimators for the SATE. Obtain 100,000 draws from that distribution.

(c) Plot the (Monte Carlo estimate of the) randomization distribution for each of the two estimators: difference in means and regression. Either overlay the plots (with different colors for each) or make sure the xlim on both plots is the same.

(d) Calculate the bias and efficiency of each estimator. Also calculate the root mean squared error.

(e) Why is the estimator that ignores blocks biased? Is the efficiency meaningful here? Why did I have you calculate the RMSE?

(f) Describe one possible real-life scenario where treatment assignment probabilities and/or treatment effects vary across levels of a covariate.

(g) How could you use a regression to estimate the treatment effects separately by group? Calculate estimates for our original sample and treatment assignment (with seed 1234).

Challenge Question 1:

(a) We could have also evaluated the properties of the estimators above using sampling distributions that take into account uncertainty in all of the variables in the DGP. Simulate sampling distributions (with 100,000 draws) for the DGP and associated estimators from Problem 1 of this assignment.

(b) Create histograms of the sampling distributions just as you did above for the randomization distributions.

(c) What is the difference between a sampling distribution and a randomization distribution?

Challenge Question 2:

Redo everything above in comparison to PATE rather than SATE. Does your preferred mode of inference depend at all on the estimand that you care most about?