Problem Set 2

Your Name - Net ID - Section Number

Due Oct 26, 2022

This homework must be turned in on Brightspace by Oct 26 2022. It must be your own work, and your own work only – you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone.

Your homework submission must be written and submitted using Rmarkdown. No handwritten solutions will be accepted. You should submit:

1. A compiled PDF file named yourNetID_solutions.pdf containing your solutions to the problems.

2. A .Rmd file containing the code and text used to produce your compiled pdf named your-NetID_solutions.Rmd.

Note that math can be typeset in Rmarkdown in the same way as Latex. Please make sure your answers are clearly structured in the Rmarkdown file:

1. Label each question part

2. Do not include written answers as code comments.

3. The code used to obtain the answer for each question part should accompany the written answer. Comment your code!

Problem 1 (25 Points Total)

Consider the following Directed Acyclic Graph:
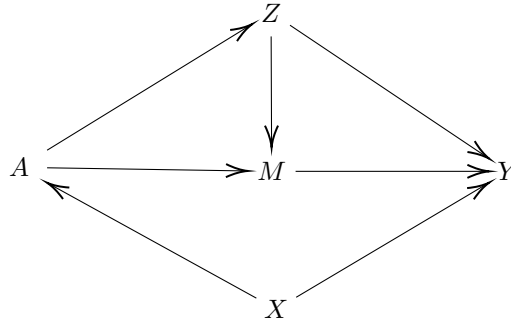
Part A (15 points)

Of the five variables in the graph, 2 are colliders and 3 are non-colliders. Which variables are colliders and which are non-colliders? Explain why?

Part B (5 points)

Suppose that we wanted to estimate the effect of A on Y. Indicate if we should or should not condition on X, and why. Also, indicate if we should or should not condition on Z and explain why or why not.

Part C (5 points)

Suppose that we wanted to estimate the effect of M on Y. List all of the backdoor paths between M and Y, and indicate which variable(s) we should condition on to close each path. There may be multiple valid options for each path.

Problem 2 (75 Points Total)

Consider again the GOTV data from last problem set by Gerber, Green and Larimer (APSR, 2008). Although it is not specified in the paper, it is highly possible that the authors created subgroups based on the turnout history for 5 previous primary and general elections (number of times the individual voted), and number of registered voters in the household. In this problem, we will create subgroups based on the turnout history, and investigate the CATE (conditional average treatment effect) and the effect modifications in each subgroup. We denote the turnout history/number of times voted as a covariate X_i for individual i.

Part A. Data Preparation (20 Points Total)

Construct a new dataset for this problem using the individual level dataset provided below.

1. Create a new column *num_voted* to represent the number of times the individual has voted in previous 5 elections by summing the variables g2000, p2000, g2002, p2002 and p2004 (exclude g2004 because the experiment filtered out people who didn't vote in g2004), the resulting column should be an integer ranging from [0,5]. (5 points)

2. In the following problems, we are using the individual data with num_voted as different subgroups. To simplify the problem, we investigate only the "Neighbor" treatment effect. Construct a cleaner dataset with $id, hh\_id, hh\_size, num\_voted, voted, treatment$ as columns and filter out treatment groups besides $Neighbor, Control$. (5 points)

3. Construct a household-level dataset by taking the means of hh_size, num_voted, and $voted$ in each household (the other variables are all equal within the same household and can simply be left as they are). Round the mean of num_voted **up** to the nearest integer. Your resulting dataset should have one household per row, and hh_id, hh_size, num_voted, $voted$, and $treatment$ as columns. The variable num_voted should have only values 0, 1, 2, 3, 4, 5. (5 points)

4. Report number of households in each subgroup for both treatment and control, what do you observe? (5 points)

Part B. CATE for subgroups (25 points total)

We define conditional average treatment effects as the ATE for different subgroups defined by the num_voted variable:

$$\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x], x \in \{0, 1, 2, 3, 4, 5\}$$

Since treatment was randomized at the household level, positivity and ignorability hold both unconditionally, and conditionally, within each subgroup. For each subgroup:

1. Estimate the CATE and report the variance of your estimates. (5 points)

2. Construct a 95% confidence interval around your estimates. (5 points)

3. What conclusion can you draw from these statistics? (15 points)

You can skip subgroups that either do not have members in them or do not have any treated/control members

Part C. Effect Modification (15 points total)

Suppose we want to estimate whether these is a difference in effects for two extreme groups, individuals who always vote ($X_i = 5$) and individuals who never vote ($X_i=0$), we construct an estimator $\hat{\Delta}$ to estimate the difference. We can estimate this difference as:

$$\hat{\Delta} = \hat{\tau}(0) - \hat{\tau}(5)$$

Calculate the variance of $\hat{\Delta}$ and construct a 95% confidence interval around it. Can we say that there's a significant difference in the treatment effect for people who always vote and people who never vote? (15 points)

Part D (15 Points)

In the experiment, the authors claimed no significant differences between groups, one possible reason may be that the sample size for each subgroup is too small. This is a practical problem we may encounter in experimental designs when we are testing multiple hypothesis or we are having too many subgroups. Explain in your own words why having more hypothesis/subgroups would make significant effect harder to detect for each group, assuming the overall sample size is fixed.