Problem Set 4

Professor: Christopher Lucas

Due Monday, February 17

You are required to use either `knitr` or `Markdown`, these are fully compatible with `R` and LaTeX.

Please (1) bring a printed version of your answers to class and (2) upload an electronic version on Canvas. Hard copies of your solutions should be handed in at the start of class and electronic copies are due before class begins at 4:00 pm. Late submissions will not be accepted. Please, comment our code and show your work in questions that require calculations.

Please set your seed using set.seed(02139) when the question requires simulations.

1 Election Campaign

You have been brought on as a data scientist for a large election campaign. The campaign manager is interested in determining whether a recent fundraising drive on a large social networking site was effective in increasing campaign donations. The manager explains, "The drive was designed primarily to maximize donations, but having heard about the importance of experiments, we introduced some randomness into the process. Specifically, we sent a 'please donate' message along with a funny-but-endearing video featuring our candidate to half of the supporters who had 'liked' our campaign page. Then, we randomly selected one friend of each selected supporter and sent the same message to them, too."

Download and import the file `donations.RData` from the course webpage. The file contains two objects:

- `dat`: A data frame containing a unique identifier for each user of the network `id`, a binary indicator for whether the individual is a supporter (specifically whether he or she had 'liked' the campaign page prior to the experiment) `supporter`, the binary treatment indicator `D` indicating whether each user received the fundraising message, the binary outcome `donate`, and a feeling thermometer score `thermo` from a subsequent network-wide survey (which, amazingly, everyone in the network answered)
- `G`: An $N \times N$ matrix, where the $\{i, j\}$-th element takes on a value of `1` when i and j are "friends" on the social networking site (measured pre-treatment). Row and column names in `G` correspond to the user identifiers `dat$id`.

(a) The campaign manager, who has a fancy data science degree but didn't take Causal Inference at WashU, tells you he already ran OLS on the data with `donate ~ D + supporter` and found *HUGE* results. He continues, "After all, the treatment is randomized, and we can control for whether people were supporters." Control your righteous anger, but justify your consulting fees. Under what circumstances would this model produce misleading results? Why? Among which subgroups? What about `donate` $\sim$ `D + supporter`?

(b) Calculate the *degree* (number of friends) of each individual in the network and add this vector to your data frame. Assess mean and distributional balance on degree between treatment and control groups among (i) the whole network, (ii) supporters, and (iii) non-supporters. Use a statistical test to assess mean balance for each of these groups. You may assess distributional balance visually or using an appropriate statistical test. Explain your findings.

(c) You present your concerns about the experimental design to the campaign manager, but he insists that you provide him with a valid assessment of whether the campaign had an effect. The funny-but-endearing video was expensive to produce, and high-level campaign donors are pressuring him for answers! You think back to Causal Inferece and decide you can use your knowledge of randomization inference to provide him with an answer.

Write a function in `R` that re-randomizes treatment assignment using the same process the campaign used in the actual experiment and returns a simulated treatment assignment vector D^*. Your function should assign treatment using the same logic as the original experiment; that is, to exactly half of all supporters and then to one non-supporter who is friends with each treated supporter. Include your function in your write-up. Generate one simulated treatment vector D^*, using `set.seed(02139)`. Assess the mean balance on *degree* between the simulated treatment group and the actual treatment group among (i) the whole network, (ii) supporters, and (iii) non-supporters. What do your results suggest?

Note that there are no overlaps in the friend-groups of supporters; this will make the sampling process much easier. Also note that, by convention, $G_{i,i} = 0$. Finally, be aware that `sample()` behaves differently when `length(x) == 1`, so be sure to take into account the fact that some supporters have only one friend.

(d) Consider the test statistic $T_{\text{sum}} = \sum_i D_i Y_i$, where Y_i corresponds to the binary outcome measure `donate`. Use or modify your function from part (c) to compute T_{sum} for each simulated treatment vector. Generate a distribution of T_{sum} using at least 1,000 simulated treatment vectors and display your results in a histogram or density plot.

i) What does this distribution represent? Formally state the null hypothesis you are using this distribution to evaluate, and its alternative.

ii) Calculate the realized value of the test statistic (i.e. the value of T_{sum} for the actual treatment assignment). Add a vertical line representing this value to your plot.

iii) Using the distribution of test statistics you created, calculate and report a **two-tailed** p-value for the realized test statistic.

(e) Explain your findings from part (d) to the campaign manager in your own words, assuming no prior knowledge of statistics. Describe and justify the procedure you used and provide a clear and concise interpretation of your results.

(f) The campaign manager is impressed by your skills, and now wants to know whether the fundraising drive affected perceptions of the candidate. Again, use randomization inference to test this claim. To do this, you will use a new test statistic: the average outcome among treated units $T_{\text{avg}} = \frac{1}{\sum D_i} \sum_i D_i Y_i$, where Y_i now corresponds to the feeling thermometer score `thermo`. Generate a reference distribution of your new test statistic T_{avg} under the sharp null hypothesis, using the same procedure as in part (d) (*Hint:* You may use the same set of simulated treatment vectors you used in part (d)). Compare the realized test statistic with this distribution and report a two-tailed p-value. What does this tell you about the effect of the fundraising message on perceptions of the candidate?

(g) Some of the variation in feeling thermometer score may be due to observed pre-treatment covariates (in this case, being a supporter and degree). Adjust for this by first fitting a model to the data and then calculating the test statistic using the estimated residuals rather than the outcome itself. Report your findings.

In doing covariate adjustment, use a fully saturated model (i.e., `thermo` $\sim$ `supporter * factor(degree)`). How does covariate adjustment affect your findings?

(h) The campaign manager comes back to you again and says he wants more than just a p-value. In order to understand whether it is worth the cost to pursue similar messaging campaigns in the future, he wants some information about the *magnitude* of the effect on perceptions of the candidate. Can you grant his request just using the test you performed in part (f)? Why or why not? Explain briefly.

2 DAGs - Part 2

After months of deep thought, you decide that your dissertation will focus on the relationship between X and Y. In preparation for your first colloquium, you read all the relevant literature, conduct field interviews, and finally write down a directed acyclic graph that you believe captures all of the relevant variables, and their inter-relationships. You proudly present Figure 1 to your committee.

(a) Your committee chair, who is not used to DAGs, asks you to explain the purpose of your DAG. As a researcher, what does it buy you, and what does it not buy you?

(b) Explain, in your own words, Pearl's back-door criterion.

(c) For Figure 1, enumerate all paths from X to Y.

(d) In the path $\{X \leftarrow Z_4 \rightarrow Z_1 \leftarrow Z_2 \rightarrow Y\}$, what type of node is Z_1? Does conditioning on Z_1 block or unblock this path from X to Y?

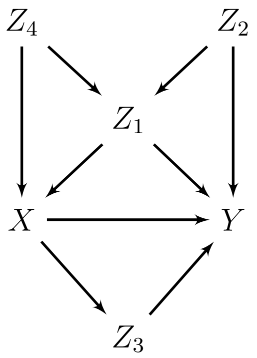

Figure 1: Your Dissertation In A Graph

(e) Now consider path $\{X \leftarrow Z_1 \rightarrow Y\}$. In this path, what type of node is Z_1? Does conditioning on Z_1 here block or unblock this path?

(f) Previous research in your area has exclusively conditioned on Z_1, claiming that this is necessary and sufficient to identify the relationship between X and Y. Given Figure 1, does conditioning on Z_1 satisfy the back-door criterion for identifying the effect of X on Y? Why or why not?

(g) Based on your DAG, enumerate the minimum conditioning sets that satisfy the back-door criterion for identifying the effect of X on Y.