1 Introduction

As a group, you will pick a causal question that fits your assigned topic and conduct an analysis. As a reminder, the final project includes two deliverables:

1. a short report describing the causal question of interest, the assumptions made (via a DAG), the identification strategy and the analysis findings (**due the last day of class: Dec 5**)
2. a 7-10 minute presentation video (**due Dec 18th asynchronously**)

In the next couple of weeks we will put out a detailed rubric on how you will be graded for these.

Meanwhile, your group will meet to discuss a plan for how you will complete the project. Your group can meet in-person or over Zoom. Since it's a group project, you may split up the work however you see fit, as long as everyone contributes relatively equally. You can certainly choose to do everything together, or split off into smaller groups to tackle different pieces–just make sure the end product is a cohesive representation of what your group worked on!

Your group should complete the questions/prompts in the Action Items section and submit answers to Canvas by **Sunday, Nov 17 at 11:59pm**.

1.1 Details on Task 3

Task 3 is about gathering and processing relevant data into a form which is usable for the causal analysis. This includes, at minimum, the following steps:

1. Pick a dataset you can download, and download the raw file
2. Read the file into R as a data table
3. Explore the data: are there missing values? are there outliers? What variables, besides treatment and outcome, are included in your dataset?
4. Clean the data: handle missing values and outliers
5. Process the data: Do you need to modify the treatment or outcome variable? For example, if the treatment in the study was on a scale of 1-5, your group likely needs to turn this into a binary treatment. For example, your group might decide "treatment" is a value of 3 or less and "control" is a value of 4 or above, so you'll need to modify your dataset accordingly! For any variables you decide to dichotomize, you will need to do some processing.

Be prepared to discuss this in a paragraph (or two) for the final paper. Task 3 and 4 will go together, so you may need to go back and forth between the two to complete both.

1.2 Details on Task 4

Task 4 is about applying an identification strategy and analyzing the data: Given assumptions which are reasonable for your causal question, choose an identification strategy and a method for analysis. This analysis should be made reproducible using an Rmarkdown file.

We highly suggest you use matching! This will entail the following:

1. Draw a DAG representing your causal question that includes at least three relevant variables besides treatment and outcome that are included in your dataset. This is like what you did for Task 2, but with the added piece that at least three of these extra factors should be in your dataset. (?)
2. Determine if conditional exchangeability holds i.e. does a sufficient adjustment set exist? If at first it doesn't hold, see if there is another variable in your dataset that you could add to the DAG so

conditional exchangeability holds. In some cases, you might have to make some strong assumptions in order to make this work... that's okay, just be prepared to discuss this in the final paper!

3. Consider the other identification conditions: consistency and positivity. Do these seem reasonable in your setting? Why or why not?
4. Conduct a (reproducible) matching analysis in R. Similar to Question 3 in Problem Set 4, you should:
    - Write a formula describing the factors that impact treatment. Basically, this should look something like `treatment ~ variable1 + variable2 + ...` where the variables correspond to the factors you are matching on!
    - Pick a matching strategy (e.g. exact matching, nearest neighbor matching, etc). You can find a full description of available matching strategies here https://cran.r-project.org/web/packages/MatchIt/vignettes/matching-methods.html Be prepared to discuss your choices in the final paper.
    - Assess your matching: how many units were dropped? How is the covariate balance in your matched sample? In addition to the Matching Lab from discussion, the following resource may be helpful: https://cran.r-project.org/web/packages/MatchIt/vignettes/assessing-balance.html If the balance is really bad, consider modifying your matching method to get better balance.
    - Using linear regression with `lm()`, estimate a model using a formula and your matched data. Your formula will be similar to the one you use for matching, but not exactly the same. Following the example above, it should look something like `outcome ~ treatment + variable1 + variable2 + ...`
    - Report a causal effect estimate! Interpret it.

If you do not want to use matching, be prepared to do the following. Note this is not shorter because its less work, but rather because we are providing less guidance.

- Pick a method we have discussed in class
- Justify why your chosen method is suited to your causal question and data
- Discuss the assumptions required for your method and why they are reasonable for your setting. You will also need to include a DAG (similar to Task 2). Note that even if your method does not require conditional exchangeability, you will need to discuss whether or not it might hold in your setting.
- Conduct a (reproducible) analysis in R and be prepared to discuss any choices you make, such as bias-variance trade-off, etc.

2 Action Items (15 pts)

Look for "Task 3 & 4 Check-in" on Canvas. It is a group assignment, so only one person from your group needs to submit for everyone in the group to receive credit. Either submit your answers to the following questions in the text box, or submit a PDF with your answers. No need to use RMarkdown, but you can if you want. There will be a template you can use on the website.

1. (1 pt) Has your group been able to get into contact with everyone in the group? (yes/no)
    - You get the point regardless of yes/no. However, if you are answering "no," someone in your group should email Sam and Mayleen to let us know what's going on.
2. (2 pts) List the name of the treatment and the name of the outcome your group is considering.
3. Data (3 pts):
    (a) If your group is using the ADD Health dataset, have you identified the variable names for the treatment and outcome for your casual question? (yes/no)

(b) If your group is not using ADD Health, have you picked a dataset (that includes your treatment and outcome variable)? (yes/no)

You get points regardless of yes/no, but if you are answering "no" that should be a flag to your group that this should be fixed asap!

4. (4 pts) Write 1-2 sentences describing your group's plan for Task 3 (see details in the corresponding section above). Your plan should include the following:
   - The "point people" in your group for this part, i.e. which members are responsible for taking charge and making sure this gets done
   - Since this needs to be done before your can conduct an analysis, include a tentative timeline
   - EXAMPLE: In our group, Sam and Shira will be taking the lead on the data cleaning and processing. They will check in with us via our groupchat by Monday Nov 18.
5. (4 pts) Write 1-2 sentences describing your group's plan for Task 4 (see details in the corresponding section above). Your plan should include the following:
   - If your group is doing matching: how will you choose which variables to match on?
   - If your group wants to use a method other than matching, list the method and it is suited for your causal question and data. For example, if using regression discontinuity, you should indicate the running variable and cutoff and confirm your dataset has this information.
   - Regardless of which method you are doing, list the "point people" in your group for this part, i.e. which members are responsible for taking charge and making sure this gets done and a timeline for getting this done
   - EXAMPLE: We will use matching. In our group, Mayleen and Filippo will work on identification (drawing the DAG and determining conditional exchangeability). They will work with Sam and Shira to make sure the variables are in the dataset to use in matching. They plan to finish this by Nov 18th. Juan and Xinran will work on conducting the matching in R, starting Nov 18th.
6. (1 pt) Please list the next date your group is planning to check in on progress (either via a meeting or through a groupchat). You can use this internally as a deadline to ensure things are getting done, or to start drafting the final paper, etc.