HAD 5744: Quantitative Methods 1
Dr. Hoagland

Assignment #3

Late assignments will be penalized at the rate of 10% per day

Please choose one of the following two problems to complete for your assignment. I recommend choosing the method that you feel most carefully aligns with your research interests, either for the paper in this class or for other projects you are pursuing.

Please submit a knitted R-Markdown file with text and code for this assignment. Don't forget to acknowledge your group members (although each of you must submit your own assignment) and to include your AI statement.

For the problems indicated with an asterisk (*), please submit a video of no more than 3 minutes explaining both how you got to your solution and what it means. Refer to the syllabus for complete instructions here.

Recall that when it comes to grading, commenting your code is your friend. It helps you organize your thoughts and communicates to me more about what you are trying to do. Unless indicated, a complete answer to each part of the assignment will include either a regression table or a figure—both should well-formatted and easily readable—and text describing and interpreting what you are presenting.

The problem (based on Finkelstein et al., *NEJM* 2020): There is widespread interest in programs aiming to reduce spending and improve health care quality among "superutilizers," patients with very high use of health care services. Typically, superutilizers make up less than 0.5% of the population but account for over 10% of total hospital expenditures.

HAD 5744: Quantitative Methods 1
Dr. Hoagland

1. **A Matching Exercise: The Role of Cost-Sharing.** Is it true that exposing "superutilizers" to cost-sharing will reduce their utilization substantially? One hypothesis is that if a patient consumes an inordinate amount of health resources, forcing them to shoulder some of the costs of those resources may be an effective cost-containment tool.

   In this exercise we will evaluate the exposure of patients to high-deductible health insurance plans (HDHPs), a form of cost sharing in the US, in their utilization. Note that you don't need to know much about the structure of this cost-sharing to complete this assignment.

For this data set, use Dataset 3a_Claims, available in the Assignments folder. This data contains spending information for individuals in the top 1% of spending for the years 2006-2018. The main treatment variable here is hdhp, which indicates if a consumer is enrolled in an HDHP. The main outcome variable is 'pay', which captures **total costs.**

a. Select a set of covariates that you think are relevant to this relationship. Clean the data, including selecting the appropriate measurement for each variable, any transformations needed, and any procedures to deal with missing data. Describe and justify your choices.
b. Create a balance table for individuals enrolled in HDHPs compared to those not. This table should include two-way *t*-tests of any significant differences. Report and interpret your findings. Are the two groups *ex-ante* comparable? Why (or why not) would you expect there to be significant differences across groups?
c. Report a naïve regression of the impact of HDHP enrollment on total spending (i.e., without matching). What do your results from (a) suggest about how you should interpret these results. How do these results likely compare to the true causal effect?
d. Now select and perform a matching estimation using your selected covariates. Report the updated balance table you created in (b) with the matched sample, and the updated regression to (c). Discuss the key changes.
e. * What assumptions are needed for this estimate to be considered causal? Do you think they are met here? What tradeoffs are inherent in your estimation? Note that in your answer, you should create and report at least one piece of evidence to support your discussion.
f. Overall, what do you conclude about the role of cost-sharing in spending among this group?

2. **Using Random Assignment as an IV.** The "hotspotting" program is a US-based program that was (at least until recently) gaining popularity. This program uses a team of nurses, social workers, and community health workers to connect enrolled patients to outpatient care and social services, hoping to reduce excessive inpatient spending and admission rates. In 2020, a team of researchers conducted a randomized-control trial assigning superutilizers either to enroll in the program or not. Following assignment and enrollment, patient participation in the program was voluntary.

For this data set, use Dataset 3b_RCT, available in the Assignments folder. Your outcome variable will be 180-day hospital inpatient spending ("post_ie_charges_180_IP"). While the randomization variable is included in the main data set, the true treatment variable is not. You will need to construct the "treated" variable from the "link2care_duration" variable in the Dataset2b_RCT_AdditionalVaraibles dataset in the Assignments folder. Consider a treated individual as participating (i.e., truly treated) if the individual participated for at least 90 days.

a. How does voluntary participation introduce endogeneity into the RCT? Defend your answer in text as well as with a DAG or with a potential outcomes framework.
b. A common IV strategy is to use the initial randomization as an instrument for takeup of the program. Does this IV satisfy the assumptions needed? Defend your answer.
c. Present a histogram of the link2care_duration variable and a summary table over the 3 experiment groups: control, randomized but not treated, and randomized and treated. In the summary table, determine whether you will need any other covariates here, and include them as appropriate. As part of this process, make sure that you have selected the appropriate measurement for each variable, any transformations needed, and any procedures to deal with missing data. Describe and justify your choices.
d. Perform the IV estimation, reporting and interpreting both the first stage of takeup on randomization, and the 2SLS estimator of the effect of program takeup on the outcome. How do your results compare to a naïve regression that doesn't account for potential endogeneity?
e. * What estimator are we recovering here? Is it economically relevant? What would be the ideal experiment as a researcher that you would like to run or have data on?
f. What do your results suggest about the effectiveness of this program on reducing superutilization spending and readmission rates?