HAD 5744: Quantitative Methods 1
Dr. Hoagland

Assignment # 3

DUE:

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 covariates of interest are: age, sex, family size, if the individual is the policy holder, # of inpatient hospitalizations per year, # of chronic conditions, and # of active prescriptions per year.

a. 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?
b. Report a naïve regression of the impact of HDHP enrollment on total spending (i.e., without matching). Remember to use 'pay' (not 'oop') as the outcome variable of interest and control for all covariates. 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?
c. Now, try exact matching the sample only on # of inpatient hospitalizations. How does this address your concern in (a) and (b)? Is it a complete way to do so? Evaluate and report the quality of your match, and report (and discuss) regression results on the matched sample.
    i. Due to the large sample size here, I recommend using the "matchit" package in R (Ho, Imai, King, and Stuart, 2011)
d. * Next, perform propensity-score nearest neighbor matching with the full set of covariates. Report the updated balance table you created in (a) with the matched sample. How has the number of observations changed? How does your new match change your answers to (c)?
e. As a researcher, which method do you prefer? What are the relevant tradeoffs? Are there other options you think you should consider?
f. Report the histogram of the propensity scores for the treatment and control groups. Discuss and/ or justify whether or not the assumptions needed for PSM are satisfied.
g. * Suppose we had data on all spenders, not just the superutilizers. What are the advantages and disadvantages of including these in your matching? Would you choose to utilize them in this research project if you were leading it?
h. Overall, what do you conclude about the role of cost-sharing in spending among this group? Does this match your intuition?

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.

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. While the randomization variable is included in the main data set, the true treatment variable is not. 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.
d. Present a histogram of the link2care_duration variable and a summary of the 3 experiment groups: control, randomized but not treated, and randomized and treated.
e. What is the first stage of takeup on randomization? How do you interpret this regression coefficient in context, and how do you interpret the overall significance of the regression?
f. * Report an IV estimator of the effect of program takeup on a continuous outcome variable: 180-day hospital inpatient spending ("post_ie_charges_180_IP"). How does this compare to a naïve regression that doesn't account for the potential endogeneity introduced by voluntary participation?
g. * 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?
h. What do your results suggest about the effectiveness of this program on reducing superutilization spending and readmission rates?

3. Evaluating an RCT with LDVs. The "hotspotting" program is a US-based program that is 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 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.

For this data set, use Dataset 3b_RCT, available in the Assignments folder. You should use these covariates as controls in your regressions throughout this problem.[1]

   a. First, estimate the effect of treatment ("Itreatment") on the main outcome of interest: "Ireadmit2_180_100" with a linear probability model.[2] Report and discuss the treatment effect.
   b. Evaluate the performance of the LPM in this setting. You should use a well-chosen histogram as well as a description of the significance of the regression (i.e., what are we using it for).
   c. What alternative model might you use given this dependent variable? Perform an adequate regression and report it.
   d. To interpret your results, construct a density plot of all marginal effects for the treatment variable, and a histogram of all the associated p-values. Add the AME and the MEM to your density plot. Is there evidence of heterogeneous treatment effects? Were there any individuals for whom the policy appeared to reduce spending meaningfully?
   e. Let's investigate the effect of the treatment on the number of inpatient hospitalizations ("post_ie_admit_cnt_180_IP"). Report a histogram of this variable to defend a specific regression model, then report the results of that regression. Compare the results to a naïve OLS regression. How do you interpret your coefficients (particularly for the treatment effect)?
   f. Does your model rest on any assumptions? If so, test them, and modify your reported regression in (e) if needed. Compare the new results to (e) and interpret how they've changed.
      i. Note: think about the construction of this variable and the people in your sample before jumping to extra complications.
   g. Taken together, what do your results suggest about the effectiveness of this program on reducing superutilization spending and readmission rates?

[1] Imale_100, Ihispanic_100, Iwhite_100, Iagebin3539, Iagebin4044, Iagebin4549, Iagebin5054, Iagebin5559, Iagebin6064, Iagebin6569, Iagebin7074, Iagebin75, I_diabetes_100, Idepression_100, IHBP_100, Iobesity_100.
[2] Note that in the data, this is scaled so that it takes on values {0,100}. Please rescale it so that it is a true binary variable.