HAD 5744: Health Econometrics 1
Dr. Hoagland

Assignment #2
DUE: November 4th before class (1:00 PM)
Late assignments will be penalized at the rate of 10% per day

Total points: 100 possible

Please choose one of the following three 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.

You may use either R or STATA to complete this assignment. Please submit a compiled version of your code that includes the output. In R, this can be achieved by "Compile Report" button in RStudio; in STATA this can be done by creating a .log file. On the first line of your code after the .log or has been started (if using STATA) please also include the following:

In R:	`name <- Sys.info() name[7]`
In STATA:	`display "`c(username)'"`

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: Health Econometrics 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? In this exercise we will evaluate the exposure of patients to high deductible health plans (HDHPs) in their utilization.

For this data set, use Dataset 2a_Claims, available in the Assignments folder. This data contains spending and plan 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 in HDHPs compared to all other individuals. This table should include two-way *t*-tests of any significant differences. What do you see? Are these two groups *ex-ante* comparable? Why (or why not) might there be significant differences across these 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 (specifically, do you think that the reported coefficient is an over- or under-estimate of the true causal effect)?
c. Now, try exact matching the sample only on # of inpatient hospitalizations. What back-door in a DAG between HDHP and spending is being closed with this single matching variable? 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 (Ho, Imai, King, and Stuart, 2011)
d. Next, perform kernel-based 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)?
    i. Use a non-propensity score kernel of your choice as the distance function
e. Repeat the exercise in (d) with a propensity score approach. Also include a histogram of the propensity scores for the treatment and control groups. Are the assumptions of PSM satisfied?
f. An important question with propensity score or nearest-neighbor matching is the bandwidth size of the match (sometimes referred to as the "caliper"). Repeat the regression results in (e) with caliper sizes of c(0.01, 0.1, 0.2, 0.5, 1). How do each of the following change: the effective sample size, the balance of covariates, the estimated treatment effect? What is the tradeoff between a larger/smaller caliper?
    i. See Austin (2011, *Pharm Stat*) for a more in-depth discussion of optimal caliper selection.
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 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. Following assignment and enrollment, patient participation in the program was voluntary.

For this data set, use Dataset 2b_RCT, available in the Assignments folder.

   a. How does voluntary participation introduce endogeneity into the RCT? Defend your answer either 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.
      i. 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.
   d. What is the first stage of takeup on randomization? What does this coefficient mean in context (hint: think limited dependent variables)? What does the *F*-stat tell you about this instrument?
      i. Note: Think carefully about what regression you run for the first stage.
   e. 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?
   f. What estimator are we recovering here? Is it economically relevant?
   g. The main outcome of interest in this RCT wasn't actually inpatient spending, but rather the rate of 180-day hospital readmissions ("Ireadmit2_180_100"), which is a binary variable. So we have a binary outcome variable *and* a binary endogenous regressor/IV. This is a case where plain ol' IV is particularly imprecise, so here we will walk through the implementation of the **bivariate probit** model, a "control function" approach to doing IV in a limited dependent variable setting (see HK for more details on this, as well as Chiburis, Das, and Lokshin, 2012):
      i. First convert this variable to a true binary variable (in data, it is scaled to 100%).
      ii. Use the "gjrm" package in R[1] with the options Model = "B" (for bivariate model) and margins = c("probit","probit") (to estimate two probits). You will need to specify two equations: the main regression of interest and the first stage relationship.
      iii. Compare the estimated treatment effect to two naïve regressions: an OLS and a 2SLS. What is the effect of using the bivariate probit?
      iv. A system of equations like this is best identified when there are different covariates in each regression. Add some covariates of your choice to both equations in your model and compare how your answers to (iv) change. Note that there should be at least one covariate in each stage that does not appear in the other (an "excluded variable").
   h. What do your results suggest about the effectiveness of this program on reducing superutilization spending and readmission rates?

[1] HK has some details about how to do this in Stata if you are using Stata instead of R.

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 2b_RCT, available in the Assignments folder. You should use these covariates as controls in your regressions throughout this problem.[2]

a. First, estimate the effect of treatment ("Itreatment") on the main outcome of interest: "Ireadmit2_180_100" with a linear probability model.[3] 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. Finally, we can evaluate how treatment through this program may have affected decisions about *where* individuals sought care:
    i. Use the Dataset2b_RCT_AdditionalVariables dataset to merge in the variable "carechoice". This variable represents an answer to a survey question (180 days post-treatment) about where the individual is most likely to go for treatment.[4]
    ii. Estimate and report the odds ratios from a multinomial logit regression. Defend your choice of reference group. (Note: your table may end up looking funky here. Try to make it nice but don't spend too much time on it).
    iii. Interpret your coefficients and discuss. How might the program have led to these results? Does it match your intuition?
    iv. How accurate is your model in predicting choices? Present a classification table as well as a visualization of the predictive power of your model. Do you believe your model?
h. Taken together, what do your results suggest about the effectiveness of this program on reducing superutilization spending and readmission rates?

---

[2] Imale_100, Ihispanic_100, Iwhite_100, Iagebin3539, Iagebin4044, Iagebin4549, Iagebin5054, Iagebin5559, Iagebin6064, Iagebin6569, Iagebin7074, Iagebin75, I_diabetes_100, Idepression_100, IHBP_100, Iobesity_100.
[3] 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.
[4] The choices for this variable are 0 for "Stay home", 1 for "Outpatient", and 2 for "Inpatient."