REMINDER: DO NOT DOWNLOAD OR COPY ANY DATA FROM THE IBM SERVER.

Version 1 **Last Updated**: Thursday April 18 @ 2am

Edits:

- n/a

6.S897/HST.956
Problem Set 6: Causal Inference for Opioids
Due: April 25, 2019 at 11:59pm by midnight through Stellar

Instructions

Log in to your account on the IBM server from PS2. You will find a *ps6materials* directory in your home folder on the IBM server. This contains starter code and data. You do not need to submit your code on stellar.

In the body of this exercise, blue text describes what specifically you should be submitting and orange text describes fourth-wall-breaking commentary from the TAs. If you have any questions about what the wording of the questions mean, ask us on piazza!

2.0: Setting the Scene

You feel like you've learned a lot this semester from working with hospitals, researchers, and public servants. You decide to start your own company to help identify problems in healthcare, suggest solutions, and to make the word a better place. You wander down the street to Sloan and start a new company that uses analytics to suggest ways for making care more effective and efficient. You are confident that you can use causal inference on observational data to make recommendations to providers about the impact of their decisions.

Today, you and your business partner are meeting with a large fast food chain as your first potential client. Anya Jenkins, the CEO of Doublemeat Palace, invites you to her office to discuss how her company would save money by using your services (as opposed to their current strategy of self-insuring and hoping their employees don't get too sick). She requests that you follow-up the meeting with a standalone report of your analyses including relevant plots and interpretations, named ${mit_user}.pdf (e.g. *wboag.pdf*). Do not copy your code from the IBM server or submit it.

In your write-up, please make your answers as easy to identify as possible (e.g. highlight with colors, label and number sections clearly, etc).

REMINDER: DO NOT DOWNLOAD OR COPY ANY DATA FROM THE IBM SERVER.

Q1. Chart Review *(5 points)*

Ms. Jenkins is concerned that an increasing number of her employees have been getting addicted to opioids after their doctor gave them a legitimate prescription. She would really like your help investigating what is causing this behavior. Use opioids.csv for your cohort this assignment. This cohort was used in this initial IBM study on opioids using MarketScan data. In this cohort, the covariates include demographics, diagnoses, and procedures. All of these patients received opioid prescriptions. The “treatment” is whether their prescription was for more than 7 days. The “outcome” is whether the patient continued opioid use for a year or sought addiction treatment.

The first thing she asks you to do is to get to know the people behind the data.

- Do a chart review (fill in the table below) for the following 5 patients.
- Try predicting (manually) which patients will become addicted.[1]

ENROLID	Description	Correct Prediction?
2413297502	47 year old female from the Northeast.	
1270285805		
28315945401		
1805427001		
2956002301		

Q2. Average Treatment Effects (*20 points*)

Ms. Jenkins’s main priority is learning what is causing opioid abuse among her employees. She has read that prescription patterns can be a driving source of the problem, and she asks you do identify how much of an impact those prescriptions have.

Q2a. You decide that the easiest way to demonstrate the Average Treatment Effect (ATE) of the intervention is to start simple and then refine.

- Report the unadjusted opioid abuse rate among patients who received the treatment (i.e. >7 days) and report the rate among patients who did not receive the treatment (i.e. received <= 7 days).

[1] Your predictions will not be graded for accuracy; this is just a good way to get acquainted with a new dataset.

- Report the difference between these two numbers.

Q2b. Now that your audience is thinking along the right lines, you explain that the difference between those two rates is not necessarily all attributable to the intervention.

- List 3 potential confounders that could be biasing the naive estimate.

Q2c. This is easy! You think back to PS5's inverse propensity weighting methods and decide to to estimate the propensity score (use sklearn's LogisticRegression with C=1.0) and re-weight the samples accordingly.

- Report the re-weighted opioid abuse rate among patients who received the treatment and among patients who did not receive the treatment.
- Report the difference between these two rates.

Q2d. Ms. Jenkins is very surprised by the findings of your analysis. She asks if you could double check whether you are confident of the conclusions. Wanting to do a thorough job, you check the literature and realize other researchers used a non-linear model on data like this. You decide to do the same (use sklearn's GradientBoostingClassifier with n_estimators=500).

- Report the re-weighted opioid abuse rate among patients who received the treatment and among patients who did not receive the treatment.
- Report the difference between these two rates.

Q2e. Which method/conclusion is more appropriate for this scenario? You need to reconcile why there are such stark the differences reported by these methods.

- Show the calibration plot of the two models.
- Report the heldout AUC of each model on the data in heldout_opioids.csv.
- Using both of these pieces of information, justify which model is more appropriate for this task.

Q3. CATE / Heterogeneity *(5 points)*

Pleased with your analysis, Ms. Jenkins asks you a follow-up question. Do opioids have the same amount of addictiveness for all people? Doublemeat Palace has workers across the United States, and it might make sense to pay especially close attention to areas at higher level of risk.

- Report the Conditional Average Treatment Effect (CATE) among the rural population and among the non-rural population.
- Are there differences between the two populations?

Q4: Checking Causal Assumptions *(15 points)*

Ms. Jenkins is very impressed and would like you to send your analysis to her so that she can run it by her statisticians. Knowing that your work will be scrutinized (for whether to hire you), you decide to go even further to justify your methods.

Q4a. You explain that it is very important to have “common support” in order to draw any reasonable conclusions about how a certain kind of patient would respond to a treatment. You think it would be persuasive to show the 2d projection of each population and that the re-weighting scheme gives a more balanced dataset.

- Plot the 2-dimensional t-SNE of the first 2000 patients in the cohort. Color patients differently based on whether they received the treatment or not.
- Report the mean and standard deviation of {the absolute differences in corresponding covariate means} between the un-weighted treated and untreated populations.There are 44 covariates, so you should compute 88 means and 44 differences.
- Report the mean and standard deviation of {the absolute differences in corresponding covariate means} among the weighted treated and untreated populations.

Q4b. You also remind the reader that ignorability is an important assumption in causal inference. If there are hidden confounders, then this could make the model of the world inaccurate, and therefore susceptible to invalid conclusions from probabilistic reasoning. If there are additional confounders, then our propensity score estimates may be wrong. How much would it affect the estimated ATE if our estimated scores were ε% off (in relative, not absolute terms[2]) from the “real” propensity scores?

For a given ε, the largest estimation of ATE comes from all treated patients being assigned a lower score (and therefore higher weight) and vice versa for untreated patients. Likewise, for the lowest estimation, treated patients have a higher-than-predicted propensity score.

$$\hat{ATE}_{\text{HIGH}} = \frac{1}{n}\sum_{i \text{ s.t. } t_i=1} \frac{y_i}{\hat{p}(t_i=1|x_i)\,(1\text{-}\varepsilon)} - \frac{1}{n}\sum_{i \text{ s.t. } t_i=0} \frac{y_i}{\hat{p}(t_i=0|x_i)\,(1\text{+}\varepsilon)}$$

$$\hat{ATE}_{\text{LOW}} = \frac{1}{n}\sum_{i \text{ s.t. } t_i=1} \frac{y_i}{\hat{p}(t_i=1|x_i)\,(1\text{+}\varepsilon)} - \frac{1}{n}\sum_{i \text{ s.t. } t_i=0} \frac{y_i}{\hat{p}(t_i=0|x_i)\,(1\text{-}\varepsilon)}$$

- Complete out the following table below.
- Describe in a sentence what the upper and lower bounds mean about the estimated ATE if the estimated propensity scores are wrong by 20% (i.e. ε=0.20).

[2] Absolute terms would be p±ε as opposed to p*(1±ε). This, however, quickly runs into issues of initially small probabilities becoming negative (from subtractions).

ε	**LOW**	**HIGH**
0.001		
0.01		
0.05		
0.10		
0.20		

Ms. Jenkins is very grateful for all of your hard work. She asks you to send your write-up of the discussion within a week. She is optimistic that this will be the beginning of a rich partnership. There is plenty of more data on the way about Doublemeat Palace employees. You are excited to be making a difference!

Q5: Not graded but please answer *(0 points)*

a) How many hours did you spend on this problem set?
b) What would you change about this pset for future iterations of this class?