Problem Set 3 (35 Points)

MLHC 2023

March 29, 2023

Submission Instructions

Edit 3/28 8:30pm: Wording in Problem 2.1 clarified

Edit 3/25 4pm: Table 1 in Problem 3.2 has been changed due to an error.

Edit 3/22 1am: We revised Problem 2 and Problem 3 for clarity. Please use the new versions of these problems for your submissions, if you previously completed the old version your results should translate easily.

This problem set is due 4-06-23 at 2:00pm EST. Please submit your write-up to Gradescope. When you write up your report, put all writing into one file and name it `mit_email_username_pset3.pdf` (e.g. `lehmer16_pset3.pdf`), furthermore make sure your report follows the same structure as the problem set. You must write up their problem sets individually. You should not share your code or solutions (i.e., the write up) with anyone inside or outside of the class, nor should it be posted publicly to GitHub or any other website. You are asked on problem sets to identify your collaborators. If you did not discuss the problem set with anyone, you should write "Collaborators: none." If in writing up your solution you make use of any external reference (e.g. a paper, Wikipedia, a website), both acknowledge your source and write up the solution in your own words. It is a violation of this policy to submit a problem solution that you cannot orally explain to a member of the course staff. Plagiarism and other dishonest behavior cannot be tolerated in any academic environment that prides itself on individual accomplishment. If you have any questions about the collaboration policy, or if you feel that you may have violated the policy, please talk to one of the course staff.

Late Day policy:

- (4 "slack" days) We understand that sometimes things outside one's control prevent submitting by the deadline. As such, each student is given 4 "slack" days that they can use throughout the semester (e.g. you could submit two psets one day late each or you could submit one pset two days late) without a late penalty. The days do not subdivide into sub-day units: 2 hours late would spend one of the slack days without 22 hours of "rollover". In your pdf writeup, specify how many slack days you are using (they cannot be used retroactively).
- (10% off per unexcused late day.) If you submit a pset a day late without slack days,
- (max 4 days late.) Homework will not be accepted after 4 days late, regardless of how many slack days are used, absent communication from S3 or OGE.
- (write on homework) In order to use a slack day, students must include it in writing on their submission pdf. Otherwise, TAs will assume no slack days used and deduct 10% for each late day.

1 Problem 1: Translating to Math (10 points)

In this problem, we will present you with several free text scenarios. For each scenario, you must answer first whether or not causal inference is required in this scenario, and, if so, you must identify the relevant covariates (X), treatments (T), outcomes (Y), and any hidden confounders (H) that pose particular concern in this setting.

1.1 (5 points)

You have collected physiological time series data (in particular, EEG traces, heart rate, SpO2, and blood pressure) from patients during a set of surgical procedures which are controlled by the application of anesthetic drug D at various dose levels continuously throughout the procedure. While most patients survive the procedure, a small minority do not. You want to design a model to automatically control the drug dosage level over time to maximize chance of patient survival.

1.2 (5 points)

You work for a hospital and have just been assigned to help a new payer (e.g., insurance agent) satisfy an odd request. They like to receive advanced estimates of billing codes they will eventually be charged for their patients, and have requested that you provide them with an estimate of each patient's final billing code after only their first 24 hours. Contractually, this payer is prohibited from using these estimates to affect whether they accept or dispute claims, but instead uses them to more proactively assess their revenue streams. To accommodate this request, you plan to build a model based on retrospective data to take the first 24 hours of a patient's stay and predict the final ICD10 codes they received at discharge time.

2 Problem 2: Computing ATE (10 points)

We run a randomized controlled trial to evaluate the effectiveness of a new drug for reducing migraines. We recruit 1000 patients with a diverse history of migraines and randomly assign them to either the treatment group or the control group. We measure the number of migraines each patient has in the 6 months following the trial.

Let Y denote the number of migraines in the 6 months following the trial, let T denote the treatment group, and let X denote whether each patient reported that they suffer from migraines. X is a variable that indicates whether patients reported they suffer from migraines often: i.e. you give them a survey with the question "Do you suffer from Migraines very often?" if they say yes, X=1, if they say no, X = 0

2.1 (5 points)

Draw the causal graph that represents the experiment above. Explain the connections between the vertices.

2.2 (5 points)

We collect the following data during the experiment:

X	Y	T
0	2	0
0	10	0
0	1	1
0	8	1
0	10	0
0	2	0
1	15	1
1	4	1

Compute the average treatment effect (ATE) of the drug on the number of migraines in the 6 months following the trial using the data above

3 Problem 3: Drawing Casual Graph and ATE (15 points)

We want to study the effect of a surgery fellowship on how well a surgeon performs in the operating room.

Let Z denote whether the surgeon went to Harvard Medical School. Let X denote whether the surgeon had ≥ 100 successful surgeries in 2015. Let T denote whether a surgeon is selected to the surgery fellowship program from 2016-2017. Let Y denote whether the surgeon had ≥ 200 successful surgeries in 2018. Let W denote whether the surgeon cumulatively had ≥ 500 successful surgeries in their lifetime by 2019.

We know the following:

1. (i) The number of successful surgeries in 2015 depends solely on the prior education.
2. (ii) Whether a doctor is selected to the surgery fellowship program depends on their prior education (and other factors not modelled here.)
3. (iii) The number of successful surgeries in 2018 depends on the fellowship, prior education, and number of successful surgeries in 2015.
4. (iv) The cumulative number of successful surgeries by 2019 is directly based on how many successful surgeries a doctor performs in 2015 and 2018.

3.1 (5 points)

Draw the causal graph that represents the experiment above. Explain the connections between the vertices.

3.2 (10 points)

The following is the data collected for the experiment:

Z	X	T	Y	W
0	1	0	0	1
0	0	0	0	0
0	1	1	1	1
0	0	0	1	0
1	1	1	0	0
1	0	1	1	1
1	0	1	1	1
1	**0**	**0**	**0**	**1**

Table 1: Data for experiment. Last row was added on 3/25

Calculate the Average Treatment Effect (ATE) of the fellowship (T) on 2007 success (Y). Use covariate adjustment and empirically estimate the probabilities/expectations from the observed data. You must show your calculations. In particular, the formula that you will use (explain it) and the steps.