Data 102, Fall 2024
Homework 4

Due: **5:00 PM** Friday, November 8th, 2024

Submission Instructions

Homework assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

Written Portion:

- Every answer should contain a calculation or reasoning.
- You may write the written portions on paper or in $\LaTeX$.
- If you type your written responses, please make sure to put it in a markdown cell instead of writing it as a comment in a code cell.
- Please start each question on a new page.
- It is your responsibility to check that work on all the scanned pages is legible.

Code Portion:

- You should append any code you wrote in the PDF you submit. You can either do so by copy and paste the code into a text file or convert your Jupyter Notebook to PDF.
- Run your notebook and make sure you print out your outputs from running the code.
- It is your responsibility to check that your code and answers show up in the PDF file.

Submitting:

You will submit a PDF file to Gradescope containing all the work you want graded (including your math and code).

- When downloading your Jupyter Notebook, make sure you go to File → Save and Export Notebook As → PDF; do not just print page from your web browser because your code and written responses will be cut off.
- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so. As a Berkeley student, you get free access to Adobe Acrobat, which you can use to merge as many PDFs as you want.
- Please see this guide for how to submit your PDF on Gradescope. In particular, for each question on the assignment, please make sure you understand how to select the corresponding page(s) that contain your solution (see item 2 on the last page).

Late assignments will count towards your slip days; it is your responsibility to ensure you have enough time to submit your work.

Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you discuss the homework with your peers, please include their names on your submission. Please make sure any handwritten answers are legible, as we may deduct points otherwise.

Causal Inference Potpourri

1. A research team wants to estimate the effectiveness of a new veterinary drug for sick seals. They ask aquariums across the country to volunteer their sick seals for the experiment. Since the team offers monetary compensation for volunteering, zoos with less income decide to volunteer their sick seals, whereas zoos with more income are less compelled to volunteer their seals.

   It turns out that zoos with less income feed their seals less nutritious diets (regardless of whether they are sick or healthy), due to budgetary constraints. Less nutritious diets prevent seals from recovering as effectively.

   (a) (2 points) **Draw** a causal graph between variables Z, Y, I and N which denote receiving the drug, recovering, the income level of the zoo, and how nutritious a seal's diet is, respectively. Justify each edge in your graph.

   (b) (3 points) We saw in lecture that if we can identify and condition on (adjust for) all confounding variables, then we can use the unconfoundedness assumption to compute the average treatment effect (ATE).

   The *backdoor criterion* provides a way to determine which variables are confounders. In particular, we simply need to "block" all the confounding pathways in the graphical model between Z and Y.

   In a causal graph, we define a *path* between two nodes X and Y as a sequence of nodes beginning with X and ending with Y, where each node is connected to the next by an edge (pointed in either direction).

   Given an ordered pair of variables (Z, Y), a set of variables S satisfies the backdoor criterion relative to (Z, Y) if no node in S is a descendant of Z (to prevent us from conditioning on colliders), and S blocks every path between Z and Y that contains an arrow into Z.

   Using the causal graph in the previous part, **determine all possible sets of variables** that satisfy the backdoor criterion relative to (Z, Y).

   (c) (3 points) Read the following paper: *Safety of the BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Setting* and answer the following questions. As when reading any academic paper, you don't have to understand every single thing in the paper as you read: focus on connecting what you're reading to what we learned in class.

   (i) Does the paper use one of the causal inference techniques we learned in class? If so, which one?

   (ii) The paper answers two distinct causal questions. For each of the two questions, what is the causal question? What is the treatment? What is the outcome(s)?

What are the confounding variables? If any instrumental variables are used, what are they?

(iii) Identify at least one potential concern or limitation with this study. For example, do you have any concerns with how causal inference techniques were applied in this study?
Hint: when reading any paper, the discussion and/or conclusion sections are often the most insightful.

Observational Data on Infant Health

2. The Infant Health and Development Program (IHDP) was an experiment treating low-birth-weight, premature infants with intensive high-quality childcare from a trained provider. The goal is to estimate the causal effect of this treatment on the child's cognitive test scores. The data *does not* represent a randomized trial with randomly allocated treatment, so there may be confounders between treatment and outcome. In this problem, we devise a propensity score model to control for observed confounders.

(a) (2 points) The CSV file `ihdp.csv` has 27 columns:

- Column 1 is the treatment $z_i \in \{0, 1\}$, which indicates whether or not the treatment was given to the infant.
- Column 2 is the outcome $y_i \in \mathbb{R}$, the child's cognitive test score.
- Columns 3-27 contain 25 features of the mother and child (*e.g.* the child's birth weight, whether or not the mother smoked during pregnancy, her age and race). Since this dataset was not collected by a randomized trial, these features could all confound z_i and y_i, and are denoted by $x_i \in \mathbb{R}^{25}$.

In this part, you'll estimate $\hat{e}(x)$ (the predicted probability that $z_i = 1$) by fitting a logistic regression model that predicts z_i from x_i. Specifically:

(1) Read the data in `ihdp.csv` (*e.g.* using `pd.read_csv`) into three arrays: $Z \in \{0, 1\}^n$ containing the treatments, $Y \in \mathbb{R}^n$ containing the outcomes, and $X \in \mathbb{R}^{n \times 25}$ containing the features.

(2) To fit a logistic regression model, use the `scikit-learn` package in Python, which is imported as `sklearn`. Start with the following two lines:

```
from sklearn.linear_model import LogisticRegression as LR
lr = LR(penalty='none', max_iter=200, random_state=0)
```

(3) Use the `lr.fit()` method to fit the logistic regression model $\hat{e}(x)$

See the documentation here.

(b) (2 points) Write a function `estimate_treatment_effect` to estimate treatment accounting for the propensity. It should take as arguments a fitted regression model (the `LogisticRegression` object `lr` from the previous part), X, Y, and Z, and output a single value, which is the estimate of the average treatment effect.

Hint: Use the inverse propensity weighted estimator:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{z_i y_i}{\hat{e}(x_i)} - \frac{(1 - z_i) y_i}{1 - \hat{e}(x_i)} \right). \tag{1}$$

See the `LogisticRegression` object's `predict_proba` method.

(c) (3 points) Use the function `estimate_treatment_effect` from the previous part to estimate the treatment effect on the IHDP dataset. Report this estimate. According to the estimate, did the treatment have a beneficial causal effect on the outcome (*i.e.* cause cognitive test scores to increase)?

(d) (3 points) The naïve estimator is the difference between the sample means:

$$\tilde{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} y_i z_i - \frac{1}{n_0}\sum_{i=1}^{n} y_i(1-z_i), \tag{2}$$

where $n_1 = \sum_{i=1}^{n} z_i$ and $n_0 = n - n_1$. Report this estimate on the IHDP dataset. Why is it different from the estimate you computed in the previous part? Are there any circumstances under which these two estimators should produce the same estimates?

Optimal Mean Estimation via Concentration Inequalities

3. Suppose we observe a sequence of i.i.d. random variables $X_1, \ldots, X_n$, each distributed according to an unknown distribution with a known variance σ^2 and an unknown mean μ that we would like to learn more about. In particular, we would like to estimate the true value of mean μ from the observations we have. To accomplish this task, the most natural choice of estimator would seem like the sample mean.

   In this question, we will investigate the sample mean as a possible estimator for μ, and show why the sample mean isn't always the best estimator in a scenario like this one. To do this, we'll use concentration inequalities to understand how many samples n are required to estimate μ to a given precision ϵ for a confidence threshold δ.

   (a) (2 points) Let $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Use Chebyshev's inequality to show that $n = \lceil \frac{\sigma^2}{\delta\epsilon^2} \rceil$ samples are sufficient for $|S_n - \mu| \leq \epsilon$ with probability at least $1 - \delta$.
   Note: $\lceil x \rceil$ is the ceiling function, which rounds x to the nearest larger integer.

   (b) (3 points) Now assume that each X_i is bounded between a and b. Use Hoeffding's inequality to compute the number of samples n sufficient for $|S_n - \mu| \leq \epsilon$ with probability at least $1 - \delta$. In particular, show that the dependence of n on δ is $O(\log(1/\delta))$.

   (c) (2 points) Now, let's compare the bounds that we derived in parts (a) and (b). For this part only, assume that $\sigma^2 = 1$, $a = -1, b = 1$, and $\epsilon = 0.1$. Make a plot that shows, for particular values of δ, the number of samples n required based on your answers from parts (a) and (b). Your plot should show a range of ten δ values between $1/2$ and $1/1000$, using `np.geomspace(1/2,1/1000,10)`, and should be shown on a log-log scale. What do you observe? In situations where we cannot assume that each X_i is bounded, how might this be a problem?

To overcome this problem, we'll replace the sample mean with another estimator, and construct bounded random variables that will help us reason about the new estimator. To construct this estimator, we'll start by considering m groups of X_i, each with fixed size n_0. We'll compute the sample mean for each group, and call these sample means $S^{(1)}, \ldots, S^{(m)}$. Then, we'll use the median of all these group means as our estimate for the mean. The diagram below summarizes our approach.

$$\underbrace{\underbrace{X_1, X_2, \ldots X_{n_0}}_{\text{Sample mean } S^{(1)}} \quad \underbrace{X_{n_0+1}, X_{n_0+2}, \ldots, X_{2n_0}}_{\text{Sample mean } S^{(2)}} \quad \cdots \quad \underbrace{X_{n-n_0+1}, X_{n-n_0+2}, \ldots, X_n}_{\text{Sample mean } S^{(m)}}}_{\textbf{Median } S_n}$$

We do this because even though one such sample mean $S^{(i)}$ might be far from the true mean μ, we hope (and will show) that the median of all of them is more likely to be close to the true mean μ.

(d) (2 points) Fix a sample size $n_0 = \lceil \frac{4\sigma^2}{\epsilon^2} \rceil$. For each of the group means i, we define a binary random variable Z_i:

$$Z_i = \mathbb{1}(|S^{(i)} - \mu| \geq \epsilon).$$

In other words, Z_i is 0 if the corresponding group mean is close to the true mean μ (within ϵ), and 1 otherwise.

Show that $\mathbb{E}[Z_i] \leq 1/4$.

Hint: Z_i is a Bernoulli random variable.

(e) (2 points) We set $S_{\text{Med}} := \text{Median}(\{S^{(1)}, \ldots, S^{(m)}\})$. This is called the *median-of-means estimator*. Explain in words why having $|S_{Med} - \mu| \geq \epsilon$ implies that $\sum_{i=1}^{m} Z_i \geq \frac{m}{2}$.

Hint: If y is the median of m numbers, it means that $\lceil m/2 \rceil$ of the numbers are greater than or equal to y, and similarly $\lceil m/2 \rceil$ of the numbers are less than or equal to y.

(f) (2 points) By taking probabilities, part (e) implies

$$\mathbb{P}(|S_{\text{Med}} - \mu| \geq \epsilon) \leq \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i \geq \frac{1}{2}\right).$$

If we combine this fact with the result of (d), we can show that

$$\mathbb{P}(|S_{\text{Med}} - \mu| \geq \epsilon) \leq \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} \left(Z_i - \mathbb{E}[Z_i]\right) \geq \frac{1}{4}\right).$$

Now use Hoeffding's inequality to compute what number m is sufficient to ensure that $|S_{\text{Med}} - \mu| \leq \epsilon$ with probability at least $1 - \delta$. What is the final number of samples of X required?