Overview

Submit your writeup including all code and plots as a PDF via Gradescope.[1] We recommend reading through the entire homework beforehand and carefully using functions for testing procedures, plotting, and running experiments. Taking the time to test, maintain, and reuse code will help in the long run!

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.

1 Observational Data on Infant Health

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. Review Lecture 15 to make sure you understand the details of propensity score modeling in causal inference.

(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 $\hat{e}(x)$ by fitting a logistic regression model that predicts z_i from x_i. For any x_i, $\hat{e}(x_i)$ is then the predicted probability that $z_i = 1$ made by the logistic regression model on x_i. Specifically:

1. Read the data in `ihdp.csv` (*e.g.* using the `csv` package in Python) 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.

[1] In Jupyter, you can download as PDF or print to save as PDF

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?

2 Concentration of Counts

In this problem we apply concentration inequalities to sums of independent but not identically distributed random variables. If the university has n students, let X_j denote the indicator that student j is on CalCentral. We will (unrealistically) assume the X_j are independent for $j = 1, \ldots, n$, and we would like to get high-probability bounds on the total number of students on CalCentral $S = X_1 + X_2 + \cdots + X_n$, since if too many students are on the site at once it will crash. Assume we know from historical data that $\mathbb{P}(X_j = 1) = p_j$.

(a) (2 points) Write $\mu = \mathbb{E}[S]$ and $\sigma^2 = Var(S)$ in terms of $p_1, p_2, \ldots, p_n$.

(b) (2 points) Find Markov's bound on $\mathbb{P}(S \geq K\mu)$ for $K > 1$.

(c) (3 points) Find Chebyshev's bound on $\mathbb{P}(S \geq K\mu)$ in terms of μ, K and σ.

(d) (2 points) If all the p_j's are equal to p, what is the value of the bound in (c)? How does the dependence on n compare to the bound you got in part (b)?

(e) (1 point) Show that the moment generating function of X_j is given by $M_{X_j}(t) = 1 + p_j(e^t - 1)$ for all t. (Recall the definition: $M_{X_j}(t) = \mathbb{E}[e^{tX_j}]$)

(f) (2 points) Show that $M_S(t) \leq \exp\big(\mu(e^t - 1)\big)$ for all t. *Hint:* use the fact that S is the sum of independent random variables, and that $e^x \geq 1 + x$ for all $x \geq 0$.

(g) (3 points) Use Chernoff's method and the bound in (f) to show that

$$\mathbb{P}\big(S \geq K\mu\big) \;\leq\; \exp\Big(-\mu(K\log K + 1 - K)\Big)$$

If $p_j = p$ for all j, how does this bound depend on n? *Hint:* You may use the fact that $K\log K + 1 - K \geq 0$ for all $K > 0$.