Statistical Projects Portfolio

Maysen Pagan

Department of Mathematics & Statistics
Boston University
April 2024

Table of Contents

1 Differentiating Cellulose Tape for Trace Evidence

1.1 Introduction

Trace evidence describes those objects or materials that are often invisible to the naked eye and are transferred between two surfaces. Since special techniques like microscopy and spectroscopy are required to view and analyze these objects, trace evidence is often overlooked at crime scenes. Common examples of trace evidence include hair, glass, soil, or paint chips. Trace evidence can provide valuable clues about a crime scene, help make connections between individuals and locations, and support or refute witness testimonies.

This project aims to determine if cellulose tape, which is more commonly known as scotch tape, can be used as trace evidence. Tape is often found either on victims or used by suspects to build weapons like bombs. The client for this project is a Graduate student in the Chobanian and Avedisian School of Medicine at Boston University. As an intern at the Boston Police Department Crime Lab, the client was interested in whether or not there was a significant difference between 22 different types of tapes which would suggest that tapes are unique enough to be used as trace evidence.

1.2 Data and Methods

The client chose 23 of the most commonly used tapes including Scotch Double-Sided Tape or Staples Invisible Tape. One tape was removed from analysis as it was a paper tape leaving 22 tapes to differentiate. Each roll of tape was divided into 3 sections or sites based on each roll's length. The width and thickness of each tape were measured using a caliber tool. Three width measurements were taken at each site, resulting in a total of nine measurements per roll. Similarly, three thickness measurements were taken at each site, yielding a total of nine measurements per roll. A stereo microscope and polarized light microscopy were also used to obtain physical characteristics like the color and texture.

The client had performed a two factor ANOVA test with replicated analysis both to determine if there was a significant difference in the widths of the tapes as well as in the thicknesses of the tapes. However, the small sample size along with a violation of ANOVA assumptions suggest that the results from ANOVA would not provide accurate results. As a result, we turned to an unsupervised learning technique of visualizing the distances between tapes to identify similar groupings of tapes.

1.2.1 Tape Distances

Defining a distance between tapes provides a measure of dissimilarity between each tape. Variables for each of the tapes included three numerical variables (width, thickness of backing, and thickness of backing and adhesive) as well as 12 categorical variables. Not all of the variables are numeric and as a result, the Gower's distance is calculated between two tapes which accounts for both numerical and categorical variables. Before taking the Gower's distance between each tape, the categorical variables were first converted to numeric values by encoding each level of the variable 1 to the total number of levels of that variable. For example, the categorical variable Texture has three levels: "Rough", "Smooth", "Slightly Rough". "Rough" was coded as 1, "Smooth" was coded as 2, and "Slightly Rough" was coded as 3. Each number was then divided by the total number of levels to obtain normalized variables from 0 to 1. In the example, "Rough" now has the numerical value 0.33, "Smooth" now has the numerical value 0.67, and "Slightly Rough" now has the numerical value 1. Once all categorical variables were coded numerically, Gower's distance calculates a matrix of dissimilarities for each of the $\frac{n(n-1)}{2} = \frac{22(22-1)}{2} = 231$ pairs of tapes. This method combines the Manhattan distance for numerical variables and Hamming loss for categorical variables to get the total distance between two observations. For the three numerical variables, a range-normalized Manhattan distance is calculated:

$$|\frac{T_{i,k} - T_{j,k}}{\text{range}_k}|$$

where $i, j = 1,\ 2,\ \dots\ ,\ 22$ and $k = 1,\ 2,\ 3$. For the remaining 12 categorical variables, the Hamming loss is calculated with the following formula:

$$I(T_{i,m} \neq T_{j,m})$$

where i,j = 1,2,...,22 and m = 4,5,...,15. This indicator will equal 1 if $T_{i,m} \neq T_{j,m}$ and 0 if $T_{i,m} = T_{j,m}$. Gower's distance between two tapes is calculated by summing these variable distances and dividing by 15, the number of variables.

1.2.2 Multidimensional Scaling

Once we have the pairwise distances between tapes from the data, multidimensional scaling provides one method of visualizing these clusters and their distances from each other on a two dimensional plot. Although the data for each tape is multivariate, multidimensional scaling preserves the distances between pairwise observations when plotting. Multidimensional scaling takes in the calculated distances and returns a set of 22 points where the distances between each point is approximately equal to the dissimilarities between each point. These points are then plotted on a two dimensional scatter plot to visualize the "closeness" of the tapes.

1.2.3 Heatmap

We can also use a hierarchical clustered heatmap to visually identify similarities and dissimilarities between tapes. In a heatmap cells are color-coded to quickly compare one row or tape to another. The heatmap also allows one to aggregate the rows in clusters based on the calculated distances.

1.3 Results

The results from the multidimensional scaling plot and heatmap showed that the 22 tapes could be clustered into 5 groupings. Looking closer at these groupings, we can recognize some similar characteristics within groups along with few differences. For example, in one cluster, all of the tapes had a smooth edge texture, narrow width, clear surface color, and smooth texture. However, in this group, all tapes had different surface features such as grooves and bubbles.

1.4 Conclusion

The multidimensional scaling plot and hierarchical clustering heatmap are effective ways of visualizing clusters of tapes based on their dissimilarities. Each cluster revealed both similar and different characteristics between the tapes. Additionally, the exact tapes that were in each cluster slightly differed between the multidimensional scaling plot. The number of five clusters compared to 22 tapes suggests that there are similar characteristics amongst almost all of the tapes. It was left to the client to determine if 5 clusters was sufficient to conclude that all of the tapes could be differentiated from one another..

2 Erosion Susceptibility of the Boston Harbor Bluffs

2.1 Introduction

Bluff erosion refers to the gradual wearing away of coastal cliffs, also known as bluffs, due to various erosive forces such as waves, wind, rainfall, and fluctuating sea levels. Bluffs are often composed of relatively soft sedimentary rock or soil, making them susceptible to erosion over time. With dozens of known historical and ancient Native cultural archaeological sites on the Boston Harbor Islands, erosion poses a danger to these sites causing them to fall into the harbor.

The client for this project is a PhD candidate in the Earth and Environment Department at Boston University's Graduate School of Arts and Sciences. The client is interested in the variability of erosion rates of the bluffs located in the Boston Harbor islands. The main objective of this project is to classify 31 Boston Harbor bluffs into different erosion vulnerability categories based on measured variables from the provided data. The client is also interested in which variables have the largest impact on erosion susceptibility.

2.2 Data and Methods

The data provided by the client contains 10 columns, with the first column containing the names of the bluffs, and 35 rows representing 35 bluffs. The variables in the data include the orientation angle, retreat rate in meters per year, wave height, maximum wave height, mud composition, base and bluff elevation in meters, and a binary variable indicating the presence of a seawall. The retreat rate of each bluff was determined using a Digital Shoreline Analysis system which uses satellite imaging to measure how far back the bluff moves. The maximum wave height was measured through simulating different scenarios of wind speed and wind direction and extracting the maximum simulated wave height of all scenarios measured. The wave height variable is the wave height simulated for each bluff for one specific scenario of NNE winds at 15 meters per second, which are typical winds leading up to a winter storm. Since there are two variables representing the heights of waves hitting each bluff in the original dataset which could introduce multicollinearity, only one will be used throughout our analysis (maximum wave height).

Before applying methods to address the client's interests, there were steps taken to clean and organize the data. The original data set contained 35 bluffs from the Boston Harbor. However, four bluffs contained missing values for the retreat rate variable as the satellite imaging was unclear for those bluffs. These 4 bluffs were removed from the data set and we continued our analysis with 31 bluffs. The other step to prepare our data for the models was to scale the data. The method used is called Min-Max Scaling so that each feature scales the range of

[0,1]. The formula to scale each feature is:

$$x^{\text{new}} = \frac{x - min(x)}{max(x) - min(x)}$$

where x is each of the 7 variables.

The method used to group the bluffs based on similar characteristics and compare susceptibility to erosion involves visualizing a multidimensional scaling plot and heatmap with hierarchical clusterings of the bluffs. These clusters are determined by the "distances" between each bluff.

2.2.1 Bluff Distances

Defining a distance between bluffs provides a measure of dissimilarity between each bluff. Variables for each of the bluffs included seven numerical variables as well as one binary or categorical variable (seawall presence). Not all of the variables are numeric and as a result, the Gower's distance is calculated between two bluffs which accounts for both numerical and categorical variables.

Gower's distance calculates a matrix of dissimilarities for each of the $\frac{n(n-1)}{2} = \frac{31(31-1)}{2} = 465$ pairs of 22 bluffs. This method combines the Manhattan distance for numerical variables and Hamming loss for categorical variables to get the total distance between two observations. For the six numerical variables, a range-normalized Manhattan distance is calculated:

$$|\frac{T_{i,k} - T_{j,k}}{\text{range}_k}|$$

where i,j = 1,2,...,31 and k = 1,2,...,6. For the remaining categorical variable, the Hamming loss is calculated with the following formula:

$$I(T_{i,seawall} \neq T_{j,seawall})$$

where i, j = 1, 2, . . . , 31. This indicator will equal 1 if $T_{i,seawall} \neq T_{j,seawall}$ and 0 if $T_{i,seawall} = T_{j,seawall}$. Gower's distance between two bluffs is calculated by summing these variable distances and dividing by 7, the total number of variables.

2.2.2 Multidimensional Scaling

Once we have the pairwise distances between bluffs from the data, multidimensional scaling provides one method of visualizing the clusters and their distances from each other on a two dimensional plot. Although the data for each bluff is multivariate, multidimensional scaling preserves the distances between pairwise observations when plotting. Multidimensional scaling takes in the calculated distances and returns a set of 31 points where the distances between each point is approximately equal to the dissimilarities between each point. These points are then plotted on a two dimensional scatter plot to visualize the "closeness" of the bluffs.

2.2.3 Heatmap

We can also use a hierarchical clustering heatmap to visually identify similarities and dissimilarities between bluffs. In a heatmap cells are color-coded to quickly compare one row or bluff to another. The heatmap also allows one to aggregate the rows in clusters based on the calcualted distances.

To analyze which variables have the largest impact on erosion susceptibility, we will observe the relationships and trends between the variables and measured retreat rates of each bluff by visualizing scatterplots of quantitative varaibles against retreat rates and boxplots for the categorical variables against retreat rates.

2.3 Results

From the heatmap, the hierarchical clustering at the lowest level provides around 8 clusters or groups of bluffs. However, when comparing to the multidimensional scaling plot, 8 clear groupings can not be visualized. As a result, we move to a higher hierarchical clustering level where the bluffs are grouped into 3 clusters which more closely corresponds with the groupings from the multidimensional scaling plot. These clusters group the bluffs based on similar retreat rates as well as other characteristics.

2.4 Conclusion

The multidimensional scaling plot and heatmap suggest three clusters of bluffs with similar and different characteristics. For example, the bluffs in one grouping all have a seawall present, have higher maximum wave heights, lower orientation degrees, and higher base elevation levels. Half of the bluffs in this cluster have higher retreat rates while the other half have more medium retreat rates compared to others. Another cluster tends to have lower retreat rates while the last cluster contains bluffs with both low and high retreat rates.

While conducting data visualizations, we observe that there might be correlation between variables. Fitting linear regression is not an appropriate model since it suffers from multicollinearity. For the patterns between each variable and the retreat rate, refer to the appendix. From our analysis, it suggests that there might be correlation between the variables that is not captured in the dataset, such as vegetation, storms, etc. Another reason is that even though we think the bluffs might be independent, their locations and orientations in the region (for example, when several bluffs are close to each other) might have an impact on the retreat rate. Therefore, even with a regression model fitted, the results might be unreliable since there are other environmental and geographical factors that can have an impact in reality, which can be a discussion to have.

3 Kleefstra Syndrome Disease Concept Model Visualizations

3.1 Introduction

Kleefstra Syndrome is a rare genetic disorder characterized by developmental delay, intellectual disability, and distinctive facial features. Literature exists on the different symptoms associated with the syndrome but the lived experiences or how the disorder impacts the lives of individuals and caregivers is not fully understood.

Our client for this project is a genetic counseling graduate student from the Graduate Medical Sciences Department at Boston University. For her capstone project, the client is developing a disease concept model for Kleefstra Syndrome which is a formal framework to assess the lived experience of individuals and their families and provide a basis for generating outcome measures. This project aims to create visualizations of references to concepts and impacts mentioned by caregivers of individuals with Kleefstra Syndrome (KS). This plot would display the frequency of references to compare the most mentioned concepts and impacts across different age groups. The client referenced an area plot in a paper that conducted similar research. However, it was determined that this visualization, which has a continuous x-axis of age, does not successfully visualize a change in proportions over discrete age groups. The client hopes that the visualization will help inform the effects of KS on the lived experience of the individuals and their caregivers, which can help medical health professionals prioritize intervention strategies and certain concepts at different ages.

3.2 Data and Methods

The client conducted survey interviews where caregivers referenced the ages of individuals with KS ranging from under two to 17 years old. References were counted for mentions of KS-defining concepts like motor or neurological concepts, KS individual impacts like social or health impacts, and caregiver impacts like financial or emotional impacts.

Two visualizations were proposed to the client to address the client's interest. The first plot was a frequency heatmap and the second plot was a stacked bar plot. Heatmaps are a great way to reveal relationships between variables. By representing numerical values with colors, heatmaps make it easier to identify areas of high and low values. The frequency of each symptom/impact was calculated by dividing the number of references to each symptom/impact by the total number of references for that age group. This normalizes the data so that the different response sizes are no longer an issue, making the age groups comparable. A stacked bar plot also allows you to observe changes in the frequencies of references over different age groups.

3.3 Results

Both the heatmap and stacked bar plot showed the same patterns of changes in reference frequencies. For references to KS individual impacts, health is the most referenced individual impact for children under the age of 2. However, as the age groups increase, the references to health decreases as the references to social impacts increase. For references to KS caregiver impacts, frequencies of references to emotional caregiver impacts decreases while social impact references increase. For KS defining features, it is easier to compare references across age groups in the heatmap than in the bar plot which does not provide exact frequencies. Furthermore, with stacked bar plots, there is a difficulty in comparing concepts or impacts when they do not start at a common baseline.

3.4 Conclusion

The heatmap and stacked bar plot both are visualizations that allow one to compare the frequencies of references to concepts and impacts across different age groups. Line graphs or area graphs would not be representative of the data as the x-axis for line graphs and area graphs are continuous. The x-axis for the data provided is categorical as the ages of the children are divided into groups. As a result, a heatmap or stacked bar plot is the categorical alternative to area graphs. It is important to note that the cells of the heatmap provide the exact frequencies of each reference, a characteristic that the stacked bar plot does not have.

4 Modeling High-Strike Red Sox Pitchers vs. Yankees

4.1 Introduction

The Red Sox-Yankees rivalry is known to be the oldest and most famous Major League Baseball (MLB) rivalry dating back to the early 1900s. How well do the Red Sox perform against their biggest rival? The all-time record of these two teams playing each other has the Yankees winning about 200 more games than the Red Sox in the regular season. In this project, I seek to analyze the performance of the Red Sox pitchers against the Yankees by modeling the probability of a pitcher throwing a high proportion of strikes against a particular batter. Using this model, it may be possible to determine what pitchers should pitch against the Yankees as well as what types of pitches should be thrown to certain batters. Many researchers use machine learning techniques such as random forests, neural networks, and support vector machines to predict player strikeout rates, producing predictions with low error rates. In this analysis, I see how different logistic regression models compare.

Throughout this project description, I use the term high-strike and low-strike. High-strike or high strike proportion refers to an bat in which the proportion of strikes thrown by a Red Sox pitcher was greater than or equal to 0.5. Low strike refers to an at bat in which the proportion of strikes thrown by a Red Sox pitcher was less than 0.5.

4.2 Data and Methods

The data used in this project was taken from the Baseball Savant website. Pitch by pitch data was aggregated for every Red Sox game against the Yankees from the 2023 season. This produces a data set with 3,809 rows representing 3,809 pitches and 16 columns.

To prepare the data for analysis and modeling, a few columns were added to the dataset. The column *team* was added to include only those rows where *team* was equal to “Red Sox” (only pitches thrown by the Red Sox). This new dataset now with 1,880 rows was then merged with two other datasets containing the wins above replacement (WAR) of the Red Sox pitchers (using FIP) and of the Yankees batters.

A new column *result* was mutated representing a binary variable that takes 1 if the pitch thrown was a strike and 0 otherwise. The pitches that were considered a strike were those pitches where the result was a swinging strike, called strike, foul, foul tip, foul bunt, and missed bunt.

The final step in preparing the data for modeling was to collapse the dataset so that every row represented a unique pitcher-batter combination. For each unique combination, the averages were taken of the quantitative variables such as velocity and spin rate. A new column *prop* was then created representing the proportion of pitches thrown by the pitcher that were considered a strike. If the proportion of strikes thrown was greater than 0.5, the binary response variable *high* is 1 and is 0 if the proportion of strikes is less than 0.5.

4.2.1 Exploratory Data Analysis

Before modeling the probability Red Sox pitchers throw high strike proportions against Yankees batters, I conduct some exploratory data analysis to better understand the data. I first look at a correlation heatmap to observe the linear relationships between all of the numerical variables. The darkest tile is that of the vertical break and velocity of the pitch thrown, suggesting the presence of redundant information and multicollinearity. As a result, the vertical break of each pitch will not be included as a variable in the modeling.

Next, I look at the distribution of high-strike and low-strike at bats for different Yankees batters with different WARs using a violin plot. The distribution of pitchers who threw high strike proportions appears to take on the same shape as pitchers who threw low strike proportions. I then looked closer at how high strike proportions varies across different Red Sox pitchers. The violin plots for each pitcher showed that the probability of a pitcher throwing high-strikes against a batter varies across different pitchers suggesting that a multilevel model with varying intercepts and slopes may be an appropriate fit to the data.

4.2.2 Models

I fit and compare different models to the data to predict whether a high strike proportion is thrown to a particular batter while considering other variables. The data was randomly split into training (80%) and testing (20%) datasets where the models were produced from the training dataset and predictions were compared using the testing dataset. The first model I fit was that of the null logistic model, a model with no predictors. The second model fit to the data was a logistic model with the predictors of the batter's WAR and characteristics of the pitches thrown to the batter. These variables are the average velocity, spin, and horizontal break of pitches thrown to the batter. This model does not consider the variability by pitcher and so this model is also the complete pooling model. The next model considered was that of the partial pooling model where we take into account the varying high strike proportions by pitcher. This model varies the intercepts and slopes. Using the same partial pooling model, I consider adding a group-level predictor, *pwar*, the WAR of each Red Sox pitcher. The last model I consider is a no pooling model where a logistic model is created for each pitcher.

4.3 Results

A brief comparison of of the AIC and BIC of the partial pooling model without a group-level predictor and with the group-level predictor shows that adding the group-level predictor does not appear to add to the prediction performance of the model.

The misclassification error calculations for each model used a threshold of 0.40 to distinguish a high strike plate appearance from a low strike plate appearance (predictions greater than 0.40 are predicted to be high-strike). For the null model and complete pooling model, the average misclassification error was calculated from the posterior predictive draws of the new data. A function was created to loop through the iterations of the posterior predictions and for each iteration, classify each observation as high-strike (1) or low-strike (0) and calculate the misclassification error. The average of these errors was then taken by dividing by the number of iterations. The misclassification errors for the null, complete pooling, partial pooling, and no pooling models are 0.4367, 0.4886, 0.3913, and 0.5217 respsectively.

4.4 Conclusion

Simply comparing the misclassification errors of the four models, the partial pooling model with no group- level predictor performs the best with a misclassification error of 39.13%. The no pooling model creates a separate regression model for each pitcher and assumes that every plate appearance does not share similarities across pitchers. This model assumes no information is shared among the pitchers and each pitcher is considered to be independent from one another. Based on the misclassification error of the no pooling model, it appears that this analysis tends to overfit the data and overstate the variation for each pitcher.

On the other hand, the complete pooling model ignores the variation between pitchers and produces one model that groups all pitchers together to give one estimate of whether a plate appearance results in a high strike proportion. This model produces the second largest misclassificaiton error of 48.86% as it fails to account for the variability of the response variable. Additionally, the objective of this analysis was to determine which Red Sox pitchers could throw a high strike proportion against Yankees batters with different WARs and so it wouldn't be beneficial to "pool away" this variable.

As a result, a compromise between the extremes of no pooling and complete pooling is the partial pooling model which takes into account the variation of high-strikes within and between pitchers. With the lowest misclassification error, the varying intercept and slopes of this model suggests that the Red Sox pitcher affects the high-strike proportion result and it also influences the effect of the average velocity, average spin, and average horizontal break of the pitches thrown during an at bat as well as the effect of the batter's WAR.

5 BU Sustainability

5.1 Introduction

Carbon emissions refer to the release of carbon dioxide (CO2) and other greenhouse gases into the atmosphere as a result of human activities. The main human activity that emits CO2 is the combustion of fossil fuels, such as coal, oil, and natural gas, for energy and transportation. Scope 1 emissions are those greenhouse gas emissions that an organization owns or controls directly such as the fuel burned from a company's furnaces. Scope 2 emissions are indirect emissions that come from purchased electricity, heat, and steam. Scope 3 emissions are those emissions an organization is indirectly responsible for up and down its value chain. At Boston University, scope 3 emissions include the purchase, use, and disposal of products as well as the transportation for students and faculty to and from the university.

Boston University (BU) Sustainability has a Climate Action Plan in which the university states its role in reducing the greenhouse gas emissions that cause global warming. In the plan, Boston University proposes to have net zero carbon emissions by 2040. The goal of this project is to help calculate the scope 3 emissions from the university.

5.2 Data and Methods

BU Sustainability provided our team with 4 data sets that came from 4 different companies that BU purchases goods from. In a smaller team, I worked with others on one of these data sets with the aim of gaining insights into the emission factors of these purchased goods that BU is responsible for. The data set from this company provided purchased information over the 2022 and 2023 fiscal years as well as partial data from the 2024 fiscal year. The variables in the data included the name of the purchased product, the product ID, the quantity purchased of the item, and the total sales of this purchased product. Unfortunately, this data set did not provide the CO2 emissions that correspond with each purchased product. Furthermore, the company was unable to provide us with the emissions data as they sell products from other businesses and would need to request the emissions of all products from each business.

As a result, we used the Environmental Protection Agency's (EPA) greenhouse gas emission factors for different commodities which was established by the North American Industry Classification System (NAICS). This data set provides the supply chain emissions for larger categories of commodities such as furniture manufacturing, office supplies manufacturing, or dry cleaning and laundry services. The units of the emissions are kilograms of CO2 emitted per US dollar for all greenhouse gases.

In order to use this data set and combine it with the company's data set, we had to aggregate the purchased products from the company into larger grouped categories that matched some of the commodities from the EPA data set. This was done through web scraping. Through web scraping, the product ID was searched on the company's website and the product type was extracted. We then web scraped the larger product category that the product falls into. For some products, and extra level of scraping was needed to get an even broader category. There was not a one-to-one matching of these larger company categories and the categories from the EPA data set. As a result, only 9 categories could be used to gain some insights about emissions from this company. For each of the 9 categories, the total sales in that group was multiplied by the corresponding emissions per US dollar from the EPA dataset to get the total emissions for each category.

5.3 Results

Through visualizations, we were able to determine which category of products contributed the most to purchased goods emissions from this company. The category that produced the most emissions in the 2022 fiscal year was also the category that produced the most emissions in the 2023 fiscal year with slightly more emissions in 2023 than 2022. This suggests that each year, BU consistently purchases approximately the same amount of products in this category resulting in approximately the same amount of CO2 emissions from this category. Looking closer at this category that contributes the most to emissions from this company, we found that there was one item that contributed the most to this category's emissions in both years.

5.4 Conclusion

The lack of availability of product level carbon emissions from this company did not allow us to calculate very accurate total emissions from the company. Web scraping allowed us to combine the company's data with the EPA data set to calculate broader total emissions. Although exact CO2 emissions were not obtained, BU Sustainability can gain insights from these broader calculations and possibly find more environmentally friendly alternatives to those products that contribute the most to emissions.

6 An Evaluation of the Causal Effect of a Promotion Campaign on Software Usage

6.1 Introduction

In an era marked by the digital transformation of businesses, startups are increasingly reliant on strategic outreach efforts to both attract new customers and improve loyalty among existing ones. One such startup, specializing in software solutions, is looking to increase consumption among existing customers by offering discounts. To gauge the impact of these discounts on customer usage, the company examines historical data from 2000 customers. This project aims to identify and estimate the causal effect of offering discounts on the revenue received from customers using the potential outcomes framework.

6.2 Data and Methods

The key variables in the dataset include a binary treatment variable indicating whether a customer received a discount and a response variable representing the revenue generated from software purchases in the year following the discount. Additionally, there are several pre-treatment covariates such as the company's employee count, the company's size given by total yearly revenue, and binary indicating factors which include whether the customer has global offices, is a large consumer, is a Small Medium Corporation, or has commercial business.

6.2.1 Assumptions

For the Positivity assumption, we observed the covariate balance plots for each of our covariates, comparing the distribution of the values between the treatment and control. All plots show the pattern of an equal distribution of covariate values between those customers who received the discount and those who did not. This demonstrates that for given values of each of the covariates, the probability of receiving the treatment and control are both non-zero. Therefore, $0 < p(A = a|X) < 1,\ a \in \{0, 1\}$.

We will also perform our causal analysis under the Consistency assumption which states that a customer's potential outcome under the observed treatment assignment is equal to the observed outcome for that customer. Therefore, if $A_i = a, \text{then } Y_i = Y_i^{(a)}$.

We also thought of possible variables that could be a confounder of the treatment-outcome relation. The startup company might tend to give discounts to larger companies either to keep their loyalty or prevent them from purchasing from their competitors. In this sense, the size of the customer directly affects the treatment assignment. Additionally, the size of the customer may also directly affect the outcome because, for example, larger customers will most likely purchase more software than smaller customers and therefore bring in more revenue for the company. However, this size of the customer confounder can be observed in our data set through variables such as the number of employees the customer has or the binary variable that indicates if the customer has global offices. As a result, we continue with the assumption of Conditional Exchangeability which states that $\{Y^{(0)}, Y^{(1)}\} \perp\!\!\!\perp A|X$.

6.2.2 Estimation

Under the three main identification assumptions, our parameter of interest, the average treatment effect (ATE), can be identified and then estimated using three methods. For each method, non-parametric bootstrap was used to generate 1000 ATEs and bootstrap standard errors and confidence intervals were obtained.

We first use the simplest estimator for the ATE which is the Outcome Regression (OR) Estimator. We also choose to estimate $\hat{\mu}$ for the OR function using linear regression with ordinary least squares estimator with $(1, A, X)$ as predictors.

$$\hat{\theta}^{or}_{ATE} = \mathbb{E}_n[\hat{\mu}(1, X) - \hat{\mu}(0, X)]$$

The Inverse Probability Weighting (IPW) approach, which models the treatment assignment mechanism instead of the outcome mechanism, is not invariant to location transformation of the outcome and also is not stable leading to high variance. Therefore, we also calculate the ATE using the Hajek estimator which does not have these issues.

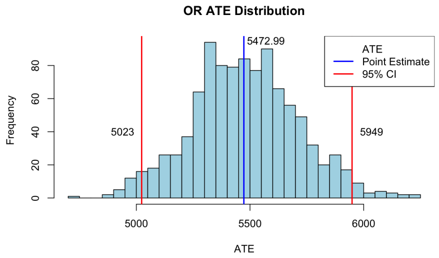
$$\hat{\theta}^{hajek}_{ATE} = \mathbb{E}_n\left[\frac{\frac{A}{\hat{\pi}(X)}}{\mathbb{E}_n[\frac{A}{\hat{\pi}(X)}]}Y - \frac{\frac{1-A}{1-\hat{\pi}(X)}}{\mathbb{E}_n[\frac{1-A}{1-\hat{\pi}(X)}]}Y\right]$$

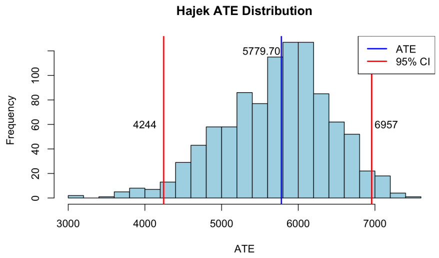
The OR function in the OR Estimator represents the outcome mechanism while the propensity score in the IPW and Hajek Estimators represents the treatment assignment mechanism. We also estimate the ATE using the Doubly Robust (DR) Estimator which combines both information and is unbiased if either the OR function or the propensity score is correctly specified.

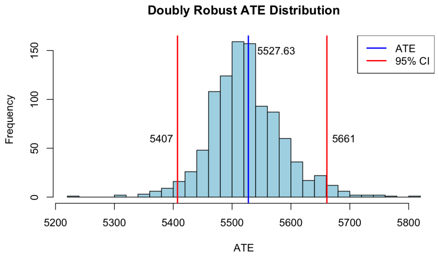
$$\hat{\theta}^{dr}_{ATE} = \mathbb{E}_n\left[\frac{A\{Y - \mu(1, X; \hat{\beta})\}}{\pi(X; \hat{\alpha})} + \mu(1, X; \hat{\beta}) - \frac{(1-A)\{Y - \mu(0, X; \hat{\beta})\}}{1 - \pi(X; \hat{\alpha})} - \mu(0, X; \hat{\beta})\right]$$

6.3 Results

The figures below display the ATE estimates as well as the bootstrap confidence intervals using the three methods above. From these figures, we can see that all of the 95% bootstrap confidence intervals do not contain zero suggesting that there is a positive causal effect between giving a customer a discount and the revenue received from that customer. We focus on the DR estimator because only one of the OR function or propensity score needs to be correctly specified for the estimator to be unbiased. We observe that the ATE is about 5,500 which means that on average, customers who received the discount would give the company $5,500 more in revenue than those who did not receive the discount.







6.3.1 Sensitivity Analysis

We then performed sensitivity analysis to observe the extent to which the Conditional Exchangeability assumption might be violated since this assumption can not be tested. We chose our sensitivity parameters to range from $\frac{1}{2}$ to 2. The signs of our ATEs are sensitive only for larger sensitivity parameters. This suggests that when customers who receive a discount tend to bring in more revenue (which means the sensitivity parameters are larger than 1), then our results are sensitive. On the other hand, for customers who receive a discount and tend to bring in less revenue, our results are not sensitive and we can conclude that there is a positive causal effect

6.4 Conclusion

Our conducted causal analysis revealed a significant increase of approximately $5,500 in revenue among customers who received the discount compared to those who did not. This indicates a positive average treatment effect of the discount on revenue generation. Through sensitivity analysis, we conclude that our estimated causal effect is primarily sensitive to only higher levels of unmeasured confounding. Despite this sensitivity, our analysis provides confidence in the existence of a causal effect, although the accuracy of the estimated ATE may be subject to some degree of inaccuracy.