Master of Science in Statistical Practice
Portfolio

Jenna Moscaritolo

Department of Statistics
Boston University

December 2021

Table of Contents

Abstract

This report contains the eight projects that I have worked on in the first two semesters of the Master of Science in Statistical Practice (MSSP) program at Boston University (BU). Starting in September 2020, I completed one major partner project as well as three smaller client projects per semester. The two partner projects were under the supervision of Professor Luis Carvalho and the six smaller consulting projects were under the supervision of Teaching Fellow David Reynolds.

All eight of these projects are presented in chronological order of their start date. The timeline of the projects were roughly as follows:

- Intake meeting
- Introduction presentation to class
- Update documents sent to client
- Final conclusion presentation to class
- Final conclusion presentation to client
- Final report sent to client

The two partner projects were semester long discussions. The other six were about a third of the semester with some overlap between projects. The overall format of each project is Introduction, Data and Methods, Results and Discussion, and Conclusion. In addition, all projects went through an exploratory data analysis (EDA) portion and most went through a modeling section.

Noise Lab

Introduction

Erica Walker, the client for this project, runs a research lab to study noise in different parts of cities and the impacts on people's moods, ability to accomplish tasks, and more. Noise levels are professionally recorded from the lab's microphones, people's feelings and comments towards noise are inputted via the NoiseScore application, and social vulnerability levels of different demographics are provided by a public dataset from the City of Boston, Massachusetts. These datasets are explained in the next section called.

Because this project is in the beginning stages, the client wants the team to explore the data through EDA and provide feedback on improvements in her data collection methods as well as pointing out positive aspects of the research. For education and extra useful information purposes, this document includes two regressions which will be further explained towards the end. For time's sake, we only looked into two demographics: people with low to no income and people with medical illness.

Data and Methods

My team received 3 datasets detailing information about the users, measurements on the noise that was recorded, and a shapefile of social vulnerability around Boston. Upon the team's first look, we cleaned the data by removing some unusable variables. We were left with spatial data as well as variables like age, ethnicity, users' positive or negative perception of noise, their sensitivity, etc.

Next, we performed some EDA to assess the associations of users' sensitivities and feelings, ages and feelings, sensitivities and ages, as well as sensitivities and ethnicity to visualize their residual-based shadings by using the vcd package in R. I will also refer to feeling and sentiment. In the full report, there are mosaic plots that showed the different Pearson residual levels of the associations listed above. For example, the Sensitivity vs. Feeling mosaic plot has sensitivity on the y-axis and feeling on the x-axis. So, each outcome has their own rectangle with a different color. The colors represent if their residual is a large positive, large negative, or a neutral residual. Thus, for this example, the mosaic plot reads that more people with a lower sensitivity have a normal or neutral feeling whereas less people with a moderate sensitivity can be paired with a positive feeling.

We then performed more EDA to explore the distribution of noise measurement dataset from different levels, including day, interval, and hour levels, specifically, the intervals stand for daytime from 9 a.m. to 5 p.m. and the rest of the time as nighttime. At each level, we compared the frequency, proportion, and distribution of different sentiments, including negative, positive and normal. In addition, we designed some functions showing the portraits for any specific user or specific type of user.

The model that we selected for this project was a simple linear regression because we wanted to see, generally, the significant and not significant coefficients and look at their trend. Because this project is in its early stages, we wanted to give the client a solid template to build off of when she receives more data.

Results and Discussion

Noise Level Measurements

Negative sentiments, in proportion to neutral and positive sentiments, coincided with high noise levels and are the most prevalent in this study. We believe that this is due to COVID-19 because when people went into quarantine, people might have stopped recording their inputs and sound levels. Noises could have also been more irritating during this time. For the time of day, we found no significant difference.

In addition, we analyzed the distribution of frequency of users using the app. We can find that most of the users only submitted the data less than ten times. But we also have some users who submitted a lot of different data points. Therefore, we warped the EDA part into the user portrait functions to help us analyze different user behaviors intuitively.

Vulnerability

The area with the most and least low to no income population is not the area with the highest noise score. Fenway and Mission Hill areas have moderate numbers of low to no income people but have a very high average noise score. In the Brighton area, there are many low to no income people that live here, but it also has a very low average noise score. It is the same situation in Roslindale where only a small number of low income people live here in addition to having a very low average noise score.

The areas with the largest populations of people with medical illnesses are Allston, Fenway, the North End, and the West End. Also, the noise score is high in Fenway but low in Allston. The areas with the lowest noise scores are West Roxbury and the South Boston waterfront with the population of people with medical illnesses is small in West Roxbury but not very small in the South Boston waterfront.

General Discussion

In our initial discussion for choosing a model, we would have liked to fit a multinomial regression model due to the categorical outcome but, to avoid an overfit and to simplify the model, we took the outcome as a continuous variable and fit a linear regression model.

Here is a summary of the coefficients for people with low to no income. First, the coefficient of the predictor sound shows that it is not significant so we removed it from this model. Second, if the noise occurs at night, there generally are more low to no income people living in this area. Third, the coefficient of the positive feeling is not significant which makes sense because the places where the voices are pleasant are typically in expensive areas. Lastly, compared with the user's workplace, both indoor and outdoor noise corresponds to a larger density of people with low to no income.

Now, the second model, looking at people with medical illness, has similar aspects to the first model of low to no income people. First, if the noise happens at night, there may be more people with medical illnesses in the area. Second, the relationship between the sound predictor and outcome is rather vague and difficult to see clearly. Also, the p-values of this predictor are all larger than 0.15, so this is not a very significant predictor. Lastly, if the noise is not at work, both indoor and outdoor noise have fewer people with medical illnesses in the area.

Conclusion

There are many limitations to what we can generalize with this data, however, we have suggested ideas to improve the NoiseScore application to have a more even layout of data and a better user experience. We are aware that these models do now fit very well, which is why we are suggesting new methods. During our time on this project, we have discovered that feeling and time are the only two significant predictors. Based on that, here are the biases:

- Most data are from two active users
- Most users only record noise levels when noise is irritating
- Noise measurements information is unevenly distributed in parts of Boston

So in the next step, we can use a similar model to elicit associations once we adjust for confounding variables. Here are the suggestions we have for improvements to the application:

- Specifying if the place of recording (where noise is or where user can hear noise)
- Some measurement data is flawed with the maximum value being lower than the minimum value
- Broader data collection with more users and more city ground covered

Canada Mayflower

Introduction

Our client, Dr. Richard Primack, conducted a study that examines the effect of microclimate temperature variation on flowering phenology of the Canada Mayflower. In May and June of 2020, Dr. Primack visited 21 microsites (also referred to here as populations) of the flower in Hammond Woods of Newton, MA. The populations were selected with the intention to represent different environmental conditions. A population is defined as the plants living in a specific micro site and occupying an area of 2 m by 0.5 m. Within each population, 15 plants were tagged. Plants were at least 15 cm apart and were chosen at random at least a week before flowering without regard to the size of the inflorescence and when plants would flower.

Most microsites were separated by 5 m or more from other microsites. At each of these sites, and for each plant within the site, Dr. Primack collected individual plant data which consisted of soil temperature, the number of flowers, soil depth, and number of fruits. All 315 plants were visited once a day during the flowering season to collect other data related to flowering status which were the time of first flowering, time of peak flowers, flowering duration, and peak duration.

Our client came to MSSP consulting to ask for help estimating the effect of temperature variations on the time of first flowering. Additionally, our client sought to understand how the time to first flowering might be related to the size of the plant, as measured by the number of flowers.

Data and Methods

Now, here are some definitions from our client's data. Flowering time is defined as the count of days before the first flower appears on the individual plant. A flowering time of 1 would be the same as a plant whose first flower appeared on May 1st. Plants that flower in the month of June would have days counted consecutively through the month of May. For example, a plant which flowered on June 1st would have a flowering time of 32. Soil temperature is defined by the mean of a series of soil temperature measurements carried out by Dr. Primack. Soil temperature was taken at 2cm with a soil probe. A total of 3 temperatures were taken on dates at the start of the flowering period. On each day, all 315 plants were measured within a 2-3 hour time period during the middle of the day when the air temperatures were unchanging. Sun exposure was estimated by visiting each plant 4 times during the early and middle flowering period and recording whether the plant was in full sun or not. Number of flowers is defined as the number of flowers counted per plant during the middle of the flowering period.

We did some EDA and, in the report, added only scatterplots. Both plots were divided up by population by different colors and compared the first flowering date vs. the mean temperature (Plot 1) and the first flowering date vs. the number of flowers (Plot 2). The results of these plots are in the Results and Discussion section.

For our model, we chose the Cox PH survival model because the response variable of interest is time to event. The Cox model is expressed by the hazard function (ratios being called hazard ratios or HR) and can also be written as a multiple linear regression of the logarithm of the hazard with the baseline hazard being an 'intercept' term that varies with time.

We also introduced a frailty model because we noticed that there was a lack of independence in the data and we wanted to account for that. In our case, some plants are more flower-prone so they are "early bloomers" which is conceptualized differently from the hazard.

Results and Discussion

Results from the EDA scatterplots are as follows:

- Plot 1 shows that early bloomers flower within 17-20 days and range in mean temperature (55-60 °F). Most plants flower by the 23rd day and there is a noticeable increase in mean temperature variability (52-62 °F). By the 28th day of May, all plants not excluded from the study have flowered. This last group is centered around a lower mean temperature (50-54 °F).
- Plot 2 shows that early bloomers show large variability in the number of flowers they produce per plant (10-20). Most plants exhibit 10-15 flowers and the first flower between days 20-25. The plants that exhibit the most flowers (20) were on days 17 and 22. Between those days, there is a decreased variability in the number of flowers. The majority of plants have anywhere from 9 to 17 plants.

From the Cox PH model, we found that the mean temperature (Beta = 0.162, $p < 0.01$) and the number of flowers (Beta = 0.148, $p < 0.01$) were significant predictors for the time of first flowering.

From the Frailty Model, we found that the number of flowers (Beta = 0.225, $p < 0.01$) was a significant predictor for time of the first flowering but the mean temperature (Beta = 0.034, $p < 0.49$) was not a significant predictor for the time of first flowering.

Conclusion

Looking at the frailty model, the effect estimates have the same direction as the previous Cox PH model, in which observations were treated as independent. However, the significance has decreased. In effect, by taking the correlation structure within populations into account we have reduced our effective sample size and, with it, the significance of the coefficients.

Both models show there is an effect of the mean temperature on the time of first flowering based on point estimation, however, when we account for the correlation structure of the data and the errors within populations, the effect becomes less significant. Based on the data, there is an effect but we would need more data or more precise measurements to detect it.

Gender and Experience in Fundraising

Introduction

Our client, Professor Yanhua Bird, of the BU Questrom School of Business, submitted to us a paper studying the effect of gender and nonprofit work experience on venture fundraising using data from an online fundraising platform. Also, our client requested to have assistance with Coarsened Exact Matching (CEM). The CEM covariates are (a) number of pictures used, (b) number of words used, (c) number of posts by the founder, and (d) funding target. The client has attempted CEM.

Data and Methods

The study collected 425 data points from posts on a crowdsourcing platform. There are two dummy variables (gender and nonprofit work experience), and their interaction. They found that female and nonprofit work experience increases fundraising with a substitutive effect (having enough work experience as a male could offset gender differences).

Results and Discussion

Our time with this client was cut short because of the end of the semester and limited amount of contact with the client. Our suggestion was to perform an initial EDA comparing covariates with the study's variables of interest and search for signs of confounding. For example, male posters using many words, female posters using few words. This tests if the number of words is a better estimator for fundraising. Finally, we would send the client materials on CEM and how to execute with two treatments in R.

Conclusion

After performing some EDA, we researched how to conduct CEM in R, compiled all the information in an R Markdown file, and sent it over to the client. She did not have any further tasks for us.

Dennis Lab

Introduction

Our client, Alexander Saeboe, Keyi Han, Allison Dennis, from the Departments of Biomedical Engineering and Division of Materials Science and Engineering came to us with their new algorithm in detecting tumors. For respecting our client's privacy, I will not be disclosing the details of this project. So, I will summarize the report.

The client has developed a new algorithm for detecting tumors by the tumor layers and through fluorophores. Keeping the details out, the client came up with steps of their success and the mathematics behind it. They did some EDA for us, but we had a different approach which I will talk about in the next section. The client came to us with questions about comparing these analysis metrics on the basis of the goodness of fit of each metric. They also asked if it is possible to compare these two metrics and how to do that so they can use that in the future. They suggested AIC (Akaike Information Criterion).

Data and Methods

I will keep the details of the data classified, but generally, we were able to compare reference values to the depth of the tumors. Before this, we had to separate the excel sheets into multiple CSVs to be able to plot them. After this, we made six scatterplots with connective lines which showed us the reference value and how deep it could penetrate into the tumor using depth.

Results and Discussion

Generally, looking at these EDA plots, the greater the reference value, the deeper the tumor is. We did not go further past the EDA portion because of a time restraint for our client. Due to classification, I will not be disclosing anymore information.

Conclusion

In conclusion, we researched information about different tumors and other published algorithms to compare as much as possible with the Dennis Lab's algorithm. Although we did not exceed the EDA portion, we researched what AIC is and were able to provide some guidance on that.

BU Hub

Introduction

Our client, David Carballo, Amanda Urias, and Shannon Barry, came to the MSSP consulting group to learn more about the statistics of their survey methods taken by BU undergraduates to see their feelings on the BU Hub courses. The BU Hub is an extensive list of general education classes that BU undergraduates need to take a certain amount of before they graduate. The six different categories that students need to complete courses in are (a) philosophical, aesthetic, and historical interpretation, (b) scientific and social inquiry, (c) quantitative reasoning, (d) diversity, civic engagement, and global citizenship, (e) communication, and (f) intellectual toolkit. Most students can meet these requirements in 10 to 12 courses.

The client had three main questions for our team. The questions are “do course offerings allow students to meet the BU Hub requirements?”, “what is the students’ level of satisfaction?”, and “does the BU Hub encourage or discourage academic exploration and life/work skills development?”. For classification purposes, I will not be disclosing the details of the data or the results.

Data and Methods

We were provided with two forms of data: the survey built by the client and the National Survey of Student Engagement. We focused mainly on the survey our client made and sent out to students. Most questions were based on the likert scale and we focused on which questions were aimed positively, negatively, and normal written responses.

In addition to focusing on the way the questions were asked, we also looked at people pre-hub versus post-hub (freshmen and sophomores versus juniors and seniors) and STEM versus not STEM majors.

Results and Discussion

For the EDA portion of this project, we made barplots for each likert scale response and color coded them by if the students are STEM or not STEM and what undergraduate year they are (freshman, sophomore, junior, or senior). We also did some polychoric correlation between pre- and post-hub for which activities they enjoyed more (for example, study abroad versus research versus clubs). We did the same thing with STEM versus not STEM. Lastly, we did some text analysis on the written response questions. Due to classification, I will not be discussing our findings.

We used factor analysis to help us understand grouping and clustering in the input variables, since they will be grouped according to the latent variables. Also, factor analysis can uncover the trends of how these questions will move together. We did some causal inference analysis as well to determine the independence.

Conclusion

To be successful with this project, we needed to perform an incredible amount of EDA. Although I cannot disclose the details and findings of this project, we found a lot of shocking but also intended information based on these EDA figures. This helped us better understand the data we were working with and made us able to suggest improvements to the survey and offer methods of analysis for the future with our R code.

Speech Therapy

Introduction

This project seeks to understand how the usage frequency of a speech therapy application is associated with the magnitude of improvement in scores achieved within the application. The application is used primarily by stroke victims who have experienced impairment in hearing and language skills. Our client, Claire Cordella, is a Postdoctoral researcher in the Speech, Language and Hearing Sciences Department of BU. Claire provided pre-processed data obtained from the application, which is accessed from the users' iPhone or Android smart-phone.

Data and Methods

The dataset consists of weekly observations for 1,664 patients. The vast majority of patients (over 90%) are observed for 10 weeks. The application encompasses several domains and our dataset relates to user performance in the hearing and memory domain. The performance is measured by the Domain Score which is calculated as:

$$Domain\ Score\ = \frac{Highest\ Task\ Passed}{Total\ Tasks\ in\ Domain}\ x\ 100$$

This number is between 0 and 1. Based on this calculation, the variable "domain_score_weekly_average" denotes the patients' weekly average domain score obtained within the application. The predictor of interest is "dosage_group_median" which is the weekly median usage level associated with each patient across their entire observed history. Note that this number is constant across all weeks for each patient and consists of 5 levels (1 time per week, 2 times per week, ..., 5 times or more per week). Following guidance from Claire, we treat the score achieved in week 1 as the baseline score.

We then performed some EDA by using violin plots to look at the week number versus domain score difference and domain score average, looking at each dosage group separately. Then, we looked at simple scatterplots to see where each person's levels went over the weeks.

For the modeling section, we fit a linear regression model with group-specific terms to the data using the rstanarm package in R. We used the logit of the weekly average score as our response variable and used this transformation so that our response is approximately normal (recall that the weekly average score is between 0 and 1). We fit a linear mixed model with random intercepts and random slopes at the patient level.

Results and Discussion

For the first clump violin EDA plots comparing week number versus the domain score difference relative to the baseline, here are the results. They indicate that, on average, users in higher usage categories appear to achieve higher scores. However, these plots also indicate that the distribution of scores within each usage category is not unimodal. There appear to be different categories of users within each usage category. This is especially apparent in the '5 or more' group in week 3 where there is a group of users with little improvement and another subset with significant improvement relative to baseline. This could be explained by patient level characteristics such as age or initial level of impairment.

Like the first violin plot, the second clump of violin EDA plots show that the weekly average domain scores of each dosage group continue to increase from the start to the end of the study period. Also, it is apparent that the distribution of user performance within each usage group is not unimodal. Further, the differences across Dosage Groups 1, 2, 3, and 4 do not appear to be practically significant; however, it does seem apparent that Dosage Group 5 achieves significantly higher scores.

Onto the modeling portion, we did a coefficient plot which indicates that the interval estimates, on average, using the application more than once a week improves the average domain score. This increase is most pronounced in the 4 and 5+ categories (median weekly usage). Additionally, the model fit for the interaction terms indicate that the change in domain score over time does not vary across dosage groups.

We also did a posterior predictive check which indicates that there are modes in the observed data that are not captured by the model. We believe that this is due to unobserved confounding variables. These may include patient-level characteristics such as baseline age and speech impairment level.

The multimodal nature of the observations is especially pronounced in the 5+ usage group where there appear to be 3 patient types: those that show essentially no improvement over time, those who improve moderately, and those who display marked improvement. In the plot below, we display 3 specific patients who fall into these categories, along with a line that indicates the model fit using mean estimates.

Conclusion

Our analysis indicates that, on average, users in higher dosage categories (that is, those who use the application more frequently). However, our EDA indicates that the distribution of scores within usage groups is highly multimodal. This feature of the data could potentially be explained by expanding the analysis with patient-level information.

Bank Regression

Introduction

Our client, Abbas Attarwala, from the BU Computer Science Department came to us for help with looking at the association between the *efficiency* and *effectiveness* of banks versus their Tobin's Q. A Tobin's Q is a ratio that equals the market value of a company divided by its assets' replacement cost. It basically shows if the business is over or undervalued. Abbas is trying to test the efficient market hypothesis (EMH) which is when the share prices reflect all information and consistent alpha generation is possible.

Data and Methods

Abbas's data is from eight countries: Brazil, Canada, India, Mexico, China, South Korea, the USA, and Japan. Each country has 16 time points from 1998 to 2015. All variables averaged across countries' 70 largest banks.

Here are the definitions we used for *efficiency* and *effectiveness*. Efficiency (between 0 and 1) is calculated by Capital IQ, a market research firm and relates to the banks' ability to turn assets into revenue. *Effectiveness* (between 0 and 1) is also calculated by Capital IQ which relates to the organization's ability to gain prearranged objectives and goals.

We were only able to do EDA on this data, but we did many scatterplots with connective lines to see the time trends with the countries for both *efficiency* and *effectiveness*. Then, we did the same thing for time versus Tobin's Q. Lastly, we did regular scatterplots for efficiency and effectiveness versus Tobin's Q.

Results and Discussion

The scatterplots with connective lines were fairly all over the place, in that it was even for rising in some years and lowering for some certain years. For the time versus the Tobin's Q, all countries were low in their Tobin's Q except for some years in Japan, South Korea, and Brazil.

Again, the scatterplots for *efficiency* and *effectiveness* versus Tobin's Q were inconclusive with these plots, we were able to see that each country had low Tobin's Qs until *efficiency* and *effectiveness* were increased.

Conclusion

We were planning on fitting a basic multilevel model to the data with a random intercept for country and random slope for time component but we ran low on time to get back to our client. In addition, we were going to gather general finance information to learn more about how banks operate as well as research country level data and confounders. This was because the fit of a simple linear model to the data with the current covariate set (*effectiveness* and *efficiency*) explains very little of the variation in Tobin's Q.

Child Therapy

Introduction

Our client, Laura Darling and Alicia Fenley, from the BU Psychological and Brain Science Department are looking for us to help with their analysis of how shared decision making (SDM) is associated with therapist and patient attributes. They have data from therapy sessions from children (people under the age of 18) that have suffered from some sort of trauma or mental illness. The data will be elaborated in the next section.

Data and Methods

To be cognisant of the clients' classified data, I will generally summarize what we have been working with. The client has raw data of the transcripts from the pediatric sessions and provided us with a processed form of the data which is transcribed using this written down methodology used to quantify the SDM of a therapy session.

There are about seven therapists and 49 pediatric patients which are each assigned to a single therapist. The data contained most of these patients' first two sessions (a few only had one session recorded). The SDM is quantified from 0-4 where 0 is that there is no collaboration between the patient and therapist and 4 is that there was a lot of collaboration.

Results and Discussion

This data is extremely biased and very tough to make any sort of statistical findings. Even though there is a manual of how to quantify the therapy sessions, it is very biased to whomever made the manual. There are also countless other factors that we needed to assume which were not accounted for in the data and study.

Conclusion

Although we had to stop this project due to not having the data we needed, we were planning on doing EDA and looking into fitting a multilevel model and a GEE (Generalized Estimating Equation) which was asked by our client. We did not know how to do a GEE, so we would have had to research that.