# Masters of Science in Statistical Practice Portfolio

**Jason Lu**

Department of Statistics
Boston University
May 2021

## Table of Contents

# Noise Lab

## Abstract

We helped Boston's domestic Noise Lab with a start-up project to map out noise readings in the city and visualize its presence towards vulnerable populations. We utilized various visualization techniques such as mosaic plots and heatmaps alongside fitting a linear regression model to measure the effects of noise. Age appears to be a confounding variable and the majority of observations tend to have negative attitudes in the area they are in, however there are a myriad of biases in the data which make our current set of results highly inconclusive. Despite this, we hope further improvements to the data collection process will eventually be able to meet the core objective and that the work we have done will set the framework for future iterations of the project. Eventually, our client would like to accurately see the special and temporal patterns of noise score reports across the city of Boston and eventually expand to other parts of the world.

## Introduction

Our client, Erica Walker, 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. Increased noise exposure can lead to stress and health issues and potentially cultivate into diseases. This can be especially harmful to vulnerable populations such as the elderly, low-income, or racial neighborhoods.

General noise levels are professionally recorded from the lab's microphones while people's feelings and comments towards noise are inputted via the NoiseScore application, where they can also record noise levels in their current area. Social vulnerability levels of different demographics are provided by a public dataset from the City of Boston.

Because this project is in the beginning stages, our client wants us 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, we included two regressions which will be covered at the end of this report.

## Data and Methods

We received 3 datasets detailing information about the users, measurements on the noise that was recorded, and a shapefile of social vulnerability around Boston.

The first data source contained information regarding users of the NoiseScore application such as their ethnicity, year born, and recordings of noise levels in their area. After cleaning the data, we focused on the ethnicity, age group, sensitivity, and feeling (taken from the measurements dataset) to see if there are any confounding relationships between the variables so we can conduct a proper analysis later on. We visualize their residual-based shadings by using the VCD package and mosaic plots.

The second dataset has the noise level recordings from users when they used the app. We utilized this data to compare different sentiments at different points in time as well as finding biases in the app's usage. Specifically, we used average sound levels, location, time of day, sentiment, and userID. We performed the exploratory data analysis to explore the distribution of noise measurement dataset from different levels, such as day and hour levels. and each level, we compared the frequency, proportion and distribution of different sentiments. In addition, we designed some functions showing the portraits for any specific user or specific type of user.

The final dataset were shapefiles containing the density of differing demographics in Boston such as age, income, people of color, and English proficiency. We mapped out this vulnerability data across each of the demographics alongside the area's noise score to get a general idea of where people lie.

We concluded this project with exploring the relationship between the variables related to noise and the density of vulnerability via a linear regression model. For noise intensity, we chose *Sound* as the predictor because it can generalize the noise scores in the data with just five categories. For human perception, sentiment is an appropriate variable because it shows people's three different attitudes towards noise (negative, neutral, and positive). Furthermore, different times and places correspond closely to noise measurement and perception.

## Results and Discussion

### Noise Level Measurements

In general, older people tend to be more sensitive to noise alongside having negative feelings. Younger people, on the other hand, are less sensitive alongside having less negative feelings. Increased sensitivity can be associated with negative feelings, and lower sensitivity can be associated with more neutral feelings. As such, a person's age can be a confounding variable, in that it affects both sensitivity and feelings so we must take this variable into account when analyzing sensitivity or feelings.

### Vulnerability

We found that the largest proportion of people harbor negative sentiment, followed by neutral and then positive sentiment. There were no significant differences in sentiment for the time of day, although more data points were collected in the night interval. Spikes in the app's usage at midnight are most likely due to app's nature of being a tool for complaints. Most users post less than 10 times, and there are not that many unique users. After organizing the data by unique users, it turns out only a few users make up 80% of the data pool.

### General Discussion

In our initial discussion, we would like to fit a multinomial regression model due to the categorical outcomes of each vulnerability response variable. However, to avoid overfitting and to simplify the model, we took the outcomes as continuous variables and fit a linear regression model. In the end, we measured the relationship between noise measurements and density of people with low to no income and the relationship between noise measurements and density of medical illness in the Boston area. For our noise measurements, we chose sound to represent noise intensity, sentiment to represent human perception, and time and place since they correspond closely to noise measurement and perception.

In our first model, noise occurring at night is related to more low to no income people living in this area. Having positive feelings is not significant, but makes sense because places where sound is pleasing are typically expensive for low to no income people. Both indoor and outdoor noise corresponds to a larger density of people with less income. In the second model, noise occurring at night relates to more people with medical illnesses in the area, similar to the first model. More positive feelings interestingly correspond to higher illness and noises generated indoors and outdoors would have fewer people with medical illness living around.

For both models, the relationship between the sound predictor and outcome is rather vague and difficult to see clearly, leading to be not a very significant predictor.

## Conclusion

The models we made seem to not fit very well because we have discovered sentiment and time of day are the only two significant predictors. There are also substantial biases in the data such as mainly being from two active users, the app used as a complaint tool, and data points unevenly distributed in Boston with vacancy in many places, although fewer data points available could also be due to Covid-19 and its quarantine policies. We suggested a few things that can be fixed for future iterations of the project such as clearer meaning of some variables as well as numerical corrections to column values to make noise values more viable.

We know there are very difficult to solve constraints for this project that we also do not currently have answers for such as people using the app as a complaint machine and a way to get consistent, unbiased recordings of noise (since each person's phone is different). In the end, there are many limitations to what we can generalize with Noise Lab's data, however we hope to have set the groundwork for future data points to come.

# Academic Outcomes Between Two Student Groups

## Abstract

Using mixed method growth curve models to estimate the academic growth in students identified as deaf and hard hearing compared to students without disability. For this study, mixed method quadratic growth curve models were used to explore the sample's growth from Grades 2 through 8. The study was conducted with students from different schools, so both student id and school id are treated as random effects in the model. Growth outcome was measured through math and reading related scores (3 tests each subject for a total of 6 tests per student per year and 6 total models). Note that the test scores might be incomplete for some students, since some students might join the study later or drop out of study before it ended. We will verify the model and provide mathematical equations for the model.

## Introduction

Our client, Johny Daniel, came to us interested in seeing differences in growth for students typically developing (no disabilities) and those deaf/hard-of-hearing. The timeline of growth is from Grades 2 through 8. He had already fit a mixed effect growth curve model for the data and wanted the mathematical formulas for the growth curve as well as how to get the term values.

## Data and Methods

The data contained multiple schools and in each school there were a group of students who are classified as non-disabled and difficulties with hearing. Each student in the group took math and reading examinations for 3 terms over each grade 2 through 8. The initial model our client proposed listed student ID as a random intercept and school ID incorrectly listed as a random slope with the rest of the variables as predictors to the outcome of score. Furthermore, it was not nested in any way, as each student belongs to a unique school and cannot be under more than one school.

## Results and Discussion

After much research and reading of literature, we improved the model as a mixed effect model with adjustments. The term grade and its quadratic component were random slope because different students would have different growth rate in academic performance. The random intercept component is student ID nested in school ID since the academic performance varies between different schools and students.

The mathematical formula for the equations were displayed hierarchically with school random effects being independent of student random effects. There was a variable representing the baseline for all schools and another variable representing the baseline for each student, which is based off the school baseline due to the nested design. School is the highest level so it only contains intercept and random effect terms.

## Conclusion

In conclusion, we sent over our work and references and closed the consulting project.

# Students' Understanding of the Scoring Criteria

## Abstract

Using a quasi-binomial model to determine if student self-assessed score is related to teacher assessment. Students were to take a written exam and self-grade their scores before their teacher graded them without access to student self-assessed scores. We used regression over the traditionally used correlation in this field as we could check model validation assumptions. There is no strong evidence for the association between teacher total and self total.

## Introduction

In China, high school students will be taking the college entrance examination, and with it comes a guided writing portion. Our client is interested in the comparison of student self-assessment and teacher assessment on this writing task. For this study, 12th grade students from the same school, taught by the same teacher, were requested to self-assess their writing based on a rubric (10 points for Content, 10 points for Language, 5 points for Organization and Structure, for a total of 25 points). Then, the teacher will grade the writing without access to student self-assessed scores.

After taking the writing examination once, students were given the option to take the exam again, following the same procedure as the first time. Only some of the students retook the assessment.

Our client is concerned with the best way to analyze this data, whether it be regression or correlation, as well as meeting the assumptions for the method we chose, how much of the data to use, and if there are any other better ways to analyze the data.

## Data and Methods

There were two datasets given to us: one contained scores of all students who took it the exam the first time and another one containing scores of all students who took it a second time. They include the total score given from both students and teachers as well as a breakdown of the points from each scoring criteria. She also provided the average scores of two examinations the students took in the months of June and September as a baseline indicator for a student's overall English proficiency.

We first had to decide how much of the data we can use, whether it was all students taking the exam the first time or all students who took it twice, but not both. We initially ran a regression to see if the second session affected scores, which it did, so we could not run a model that used the data together. This is because students from the second session will score themselves higher than they did on the first session so session will become a confounding variable to answering whether there is a difference between student and teacher assessment.

After taking these factors into consideration, we fitted a model. Because our response variable had a ceiling of 25 points, and some under-dispersion concerns in previous models, we chose to fit a quasi-binomial generalized linear model with the teacher total and average examination scores as predictors and student self-assessment total as the response.

## Results and Discussion

An initial look of student self-evaluation score vs. teacher score as well as correlation plot of each variable against each other suggested no clear relationship at first.

Although most researchers in our client's field use correlation to answer these types of questions, we used regression instead so we can incorporate more variables such as the average examination score as a baseline for students. We can also check model validation assumptions for regression which we cannot do for correlation.

Our coefficients in our model were not significant at the 0.05 alpha level, so we can conclude that there is no strong evidence for the association between teacher total and self total. To examine the validity of our model, we plotted the residual plot of the quasibinomial model. Those points on the residual plot are evenly distributed along average residual equals zero, and most of the residuals are near to zero without any specific trends (ex. fanning pattern), indicating that the model fits well. We also looked at binned residual plots for the individual self-assessment categories versus teacher total, as they have a limited number of levels, and most points fit within the bounds.

## Conclusion

To sum up, we could not say there is a correlation between teachers' total assessment and students' self-total assessment. Lastly, we can generalize only to this teacher-student class since the class is taught by the same teacher. Thus, this statement cannot be made in general to all classes.

# Mango Sciences

## Abstract

Investigating named entity recognition software and natural language processing models to organize unstructured medical discharge summaries from patients. We worked with various forms of MetaMap, i2b2 data, BERT models, and ICD-10 codes to work towards accomplishing this task.

## Introduction

Mango Sciences is a data science company accelerating affordability and access to global precision medicine in emerging markets. They have a substantial amount of data regarding patient clinical journeys and capturing all elements of this is essential for robust patient analytics. Our group was asked to explore natural language processing (NLP) models for extracting clinical information from unstructured text contained in their electronic medical records into standard medical reporting schemes like ICD-10 codes. This includes standardizing the discharge summaries and running it through a process to draw out information such as what the patient came in for, procedure done on the patient, and identifying cancer statuses.

## Data and Methods

The data given to us came in the form of raw medical discharge summaries containing information about their demographic, principal diagnosis, medical history details, treatment, medication, and the like. Due to the vast amount of directions we could take the project, we split up our focus into multiple groups taking on various datasets and natural language processing models.

The Metamap group worked on utilizing the program, which maps biomedical text to the unified medical language system Metathesaurus, and apply it to the discharge summary data. The i2b2 group (Informatics for Integrating Biology and the Bedside) focused on utilizing their open-source clinical data warehouse to convert medical text to BIO labeling (Beginning, Inside, Outside) for the BERT team to utilize. The BERT group (Bidirectional Encoder Representation from Transformers) trained and finetuned a model from i2b2 data to extract discharge summary information. Finally, the ICD-10 group (International Classification of Diseases) will try to create a machine learning model to convert the summaries to a standardized ICD code for future use.

## Results and Discussion

We managed to run MetaMap through a docker image, which is an efficient and portable way to use it as you only need docker installed, as well as PyMetaMap, which is a python wrapper for the program. There is documentation to implement the program through docker and python we will deliver to Mango.

The i2b2 and BERT group fine-tuned two models, one with 2009 data and another with 2012 data, to extract medicine and event information respectively with different markers and keywords describing the input summaries. This information includes dosage, reasons, temporal information, clinical departments, and event occurrences to the patient. There is also code and documentation to convert Mango’s discharge summaries into a predictive format and to do prediction on preprocessed discharge summaries.

Finally, the ICD-10 group cultivated a convolutional neural network to classify information into ICD-10 codes. The network has an embedding layer followed by a convolutional layer, max pooling, dropout, and a dense layer.

## Conclusion

In conclusion, the project was a heavy learning experience to us. We learned a lot about different forms of named entity recognition and medical text data to assist Mango with structuring their summaries for ensuing tasks. A large part of this project for us was learning how to run these natural language processors and cleaning the biomedical text to insert into these programs while heavily documenting our progress as this is something new to Mango Sciences as well.

# Practicum Class Project Cancer

## Abstract

Cancer has been the forefront of the medical scene and treatment options are known to be limited by a patients' insurance. For this project, we delved into SEER data to see if claims regarding insurance holds true and created a boosted tree model to predict a patients' surgery decision.

## Introduction

It has been known that treatment and management of patients with chronic and severe illnesses, including cancer, have been affected by different aspects of their medical and social care. Our clients work alongside the Surveillance, Epidemiology, and End Results (SEER) Program with a large database compiled across several states containing cancer care statistics for people, deidentified, with their social background, socioeconomic status, cancer type, and other information. An initial look found that recommendations for cancer treatment given to patients are different depending on their insurance type. Although their data focuses on head and neck cancer, they are looking to see if this treatment of patients is prevalent in their dataset and cancer as a whole. The National Comprehensive Cancer Network (NCCN) has recommendations and guidelines for treatments depending on the type and stage of cancer, so it will prove useful in determining if treatment deviates for differently insured patients.

## Data and Methods

We received an extensive dataset containing information of patients encountering head or neck cancer after post-processing of the data which filtered patients with no insurance information, as SEER only has insurance status of patients from 2010 onwards. The data included standard demographic information alongside their registry state, from which they can garner the percentage of people below a specific education, poverty level, unemployment, and median income. The remaining portion of the data specifies cancer and tumor information the patient has. Because the client was interested in how insurance affects surgery being performed, we looked into EDA containing insurance with surgery as well as anything else we found interesting. We also chose to fit a machine learning model to see what variables affect a surgery action.

## Results and Discussion

In our heterogeneous correlation matrix, Surgery Performed is mostly correlated with Chemotherapy, Radiation, Lymph nodes, AJCC 7 stage, and Mets. For the most part, Insurance categories were evenly distributed when against other variables, but it was interesting to find that larger tumor sizes corresponded to more patients having some sort of Medicaid and less forms of other insurance. Against our mosaic plots, it was found that patients who have insurance are more likely to have surgery compared to the uninsured patients. There are also fewer than expected patients with Medicaid who received surgery. Patients under the age of 45 are more frequently receiving surgery than predicted and more patients over 45 are not receiving surgery than expected in general.

Our boosting model, tuned for AUC which maximizes true positives and true negatives, predicting Surgery Performed and underwent 5-fold cross-validation yielded 77% accuracy on prediction. The importance plot shows that Oral Cavity, Salivary Gland, and Sinonasal sites as well as size of tumor were key decision factors for the boosting tree decisions.

## Conclusion

We found that site location and size were impactful when it comes to insurance and surgery decision. We will create a report detailing our EDA and present our findings to our client.