**Final Project: Investigation on Public Facility Usage and COVID-19 Cases**

Data 102: Data, Inference, and Decisions, Spring 2021

Due Date: Monday May 10, 2021 at 11:59 PM PT

Wei Dai, Safak Simsek, Sameer Hassen, Lubah Nelson

**1 Data Overview:**

Covid-19, the novel virus has impacted every facet of society. The lingering question of whether everything will go back to normal is still of debate. With the decline of COVID cases, many facilities are opening back up, and people are returning to schools and work. Using the Google Community Mobility Data, we will determine if things are returning to what has been the norm by examining different aspects of society. This includes public transportation and various facilities.

The Google data is generated by the aggregation of anonymized user datasets whose location history settings are turned on. It is worth noting that the location setting is off by default. Given this information, we can see that the data represents a sample of the whole population. Because the data is collected from people whose location settings are turned on on their device, it should be acknowledged that this data suffers from selection bias. However, we do not have any concerns about the measurement error and convenience sampling. In addition, we do not believe that most of the users whose data were collected were not aware of this data collection because of the fact that many people do not read the user privacy agreements. Furthermore, in this mobility data, each row represents baseline changes in categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residentials for a given date and county in a specific state.

The second data we are using to perform our analysis is Covid-19 data where the rows represent the total number of cases in a given date and county. We merged this data with Google's mobility data by using the date and county fips codes.

After merging two datasets, we decided to use the subset of our data. We specifically focused on the metropolitan areas to see the mobility changes more vividly. These metropolitan areas include: Los Angeles, New York, Dallas, Miami, Philadelphia, Riverside, San Diego, and Denver. Moreover, Google indicates “for each region-category, the baseline isn’t a single value—it’s 7 individual values. The same number of visitors on 2 different days of the week, result in different percentage changes” . Therefore, we decided to use monthly averages as opposed to the daily values, in order to see the changes more clearly. The final version of our data includes the months from March 2020 to March 2021.

For our final data, we lack a feature specifying the people’s aggregate sentiments about the Covid-19 crisis. This would help us to analyze the mobility changes in more detail, since many people would have been bored by staying at home and wanted to go out to parks and use public transportation. Unfortunately, this kind of data was not available.

**2 Research Questions:**

After looking carefully at the data we have, we have decided to answer the following questions:

* Can we predict the number of total Covid-19 cases from the mobility changes?
* Is there a causal relationship between Covid-19 spread and use of public transportation?

The answers to these questions can be useful to take further measures in response to Covid-19. For example, if we find out that the number of Covid-19 cases can be predicted by the mobility changes, the whole community can be prepared in advance for various scenarios. Additionally, if we find a casual relationship between public transportation and the number of Covid-19 cases, we should consider to take more strict measures for public transportation. While the first question will be answered by the generalized linear models (GLMs) and nonparametric approaches such as decision trees, our second question is a great fit to utilize causal inference models.

**3 EDA:**

Since we look at the monthly baseline average changes in a county, the missing values for a given day were taking care of automatically. We have not taken any significant data cleaning steps except renaming some of the columns for better visualization purposes and removing columns that were only needed for county identification.

One of the first visualizations we created was the monthly logarithmic total number of cases in each county. As we can see on figure 1, the trend for monthly average total number of cases for each county are mostly similar.

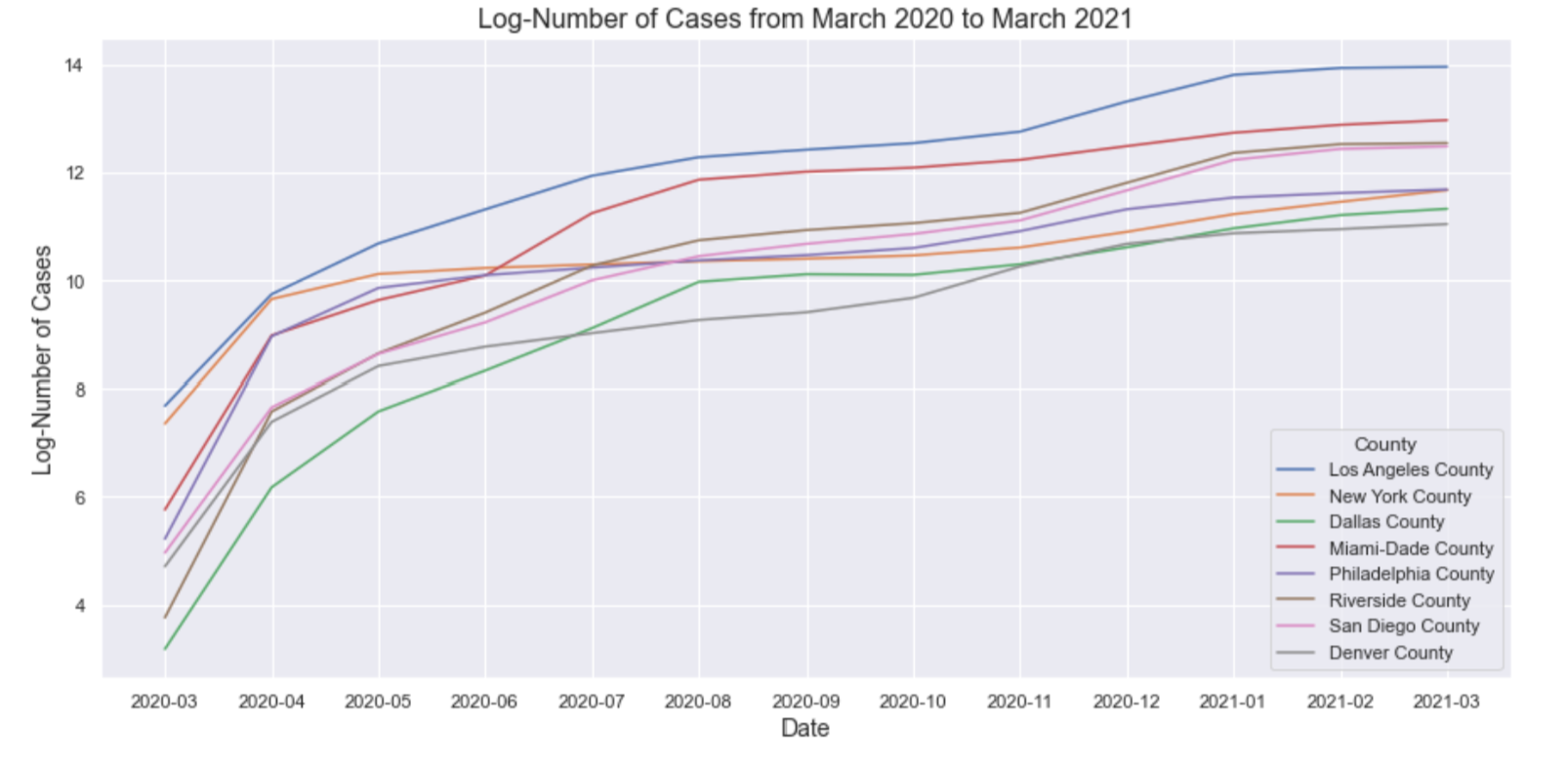


Figure 1: Monthly average log number of total cases in each county.

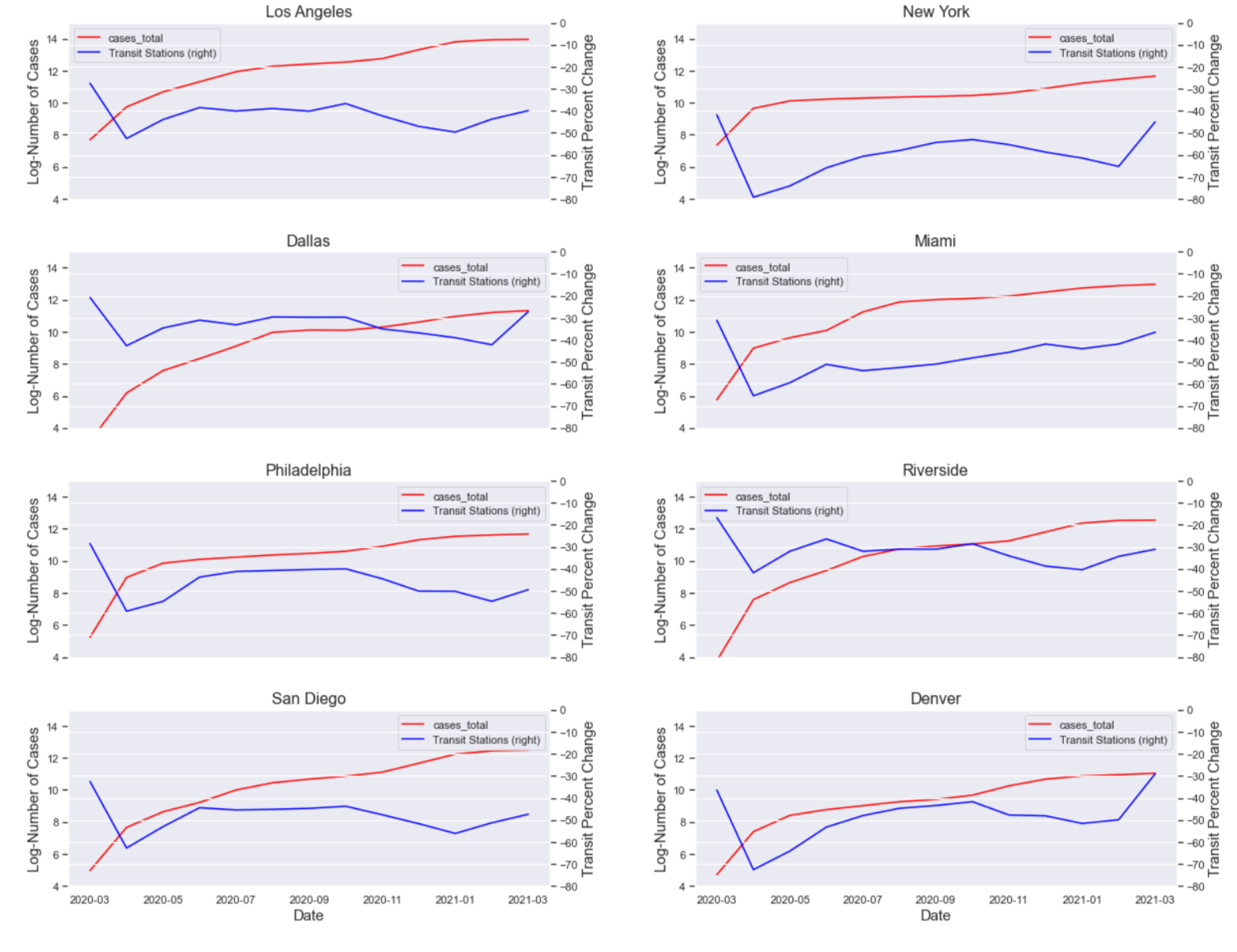
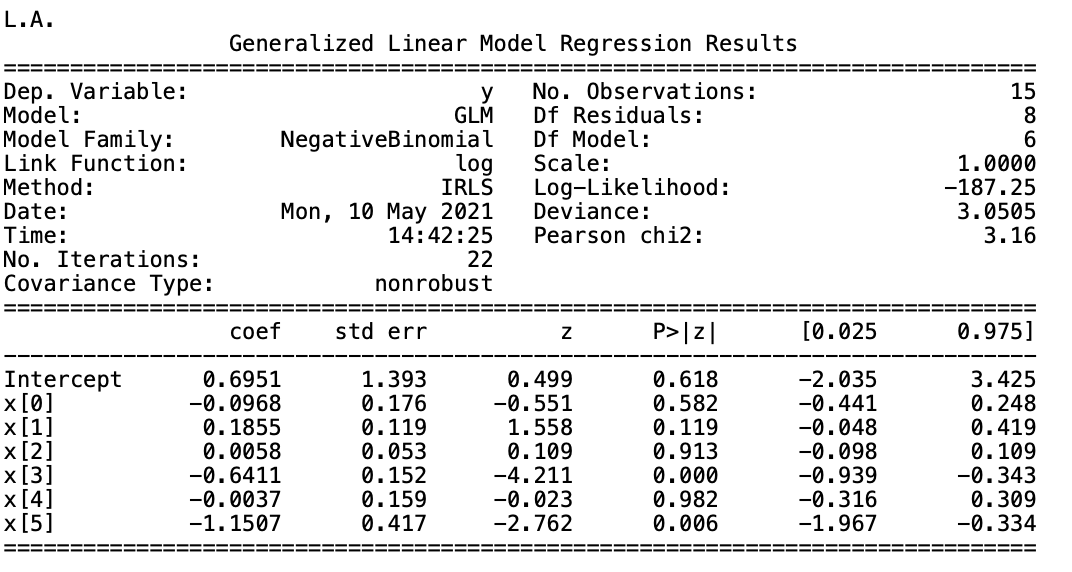
Next, we observed the mobility changes in transit stations in accordance with the Covid-19 cases in each county. Figure 2 below shows our findings. 

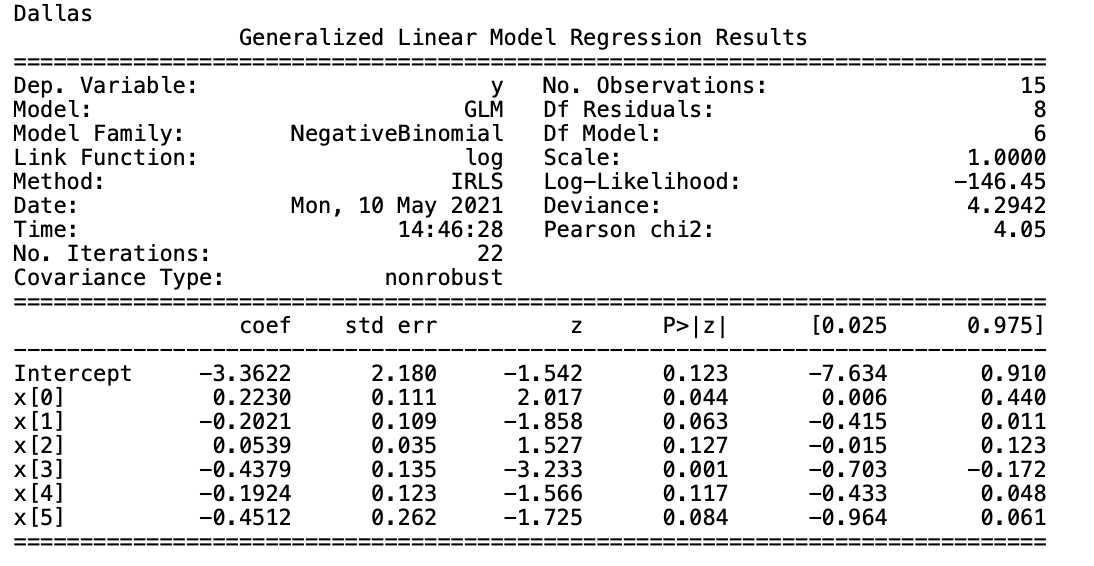
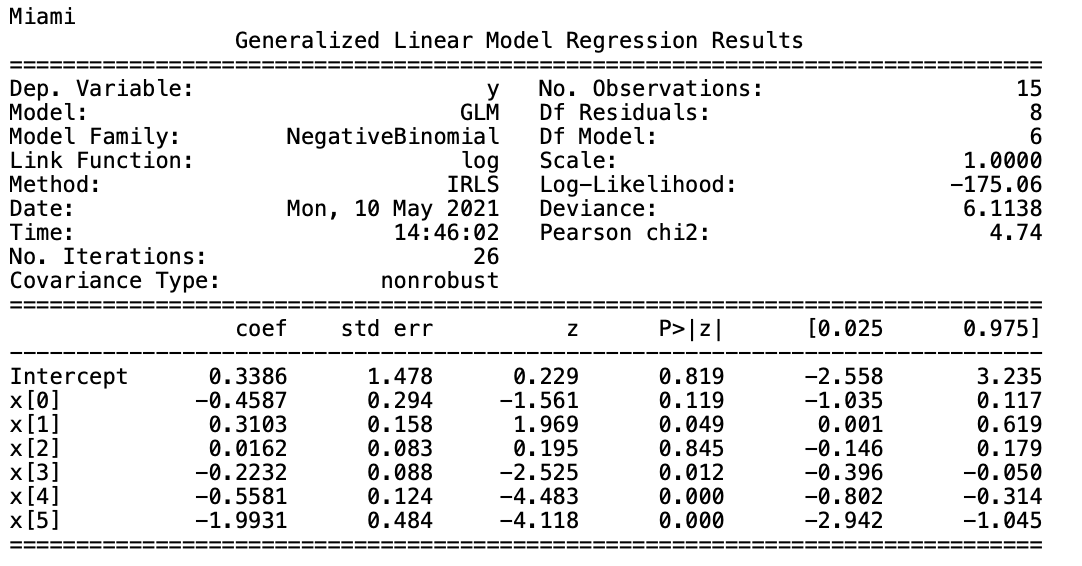
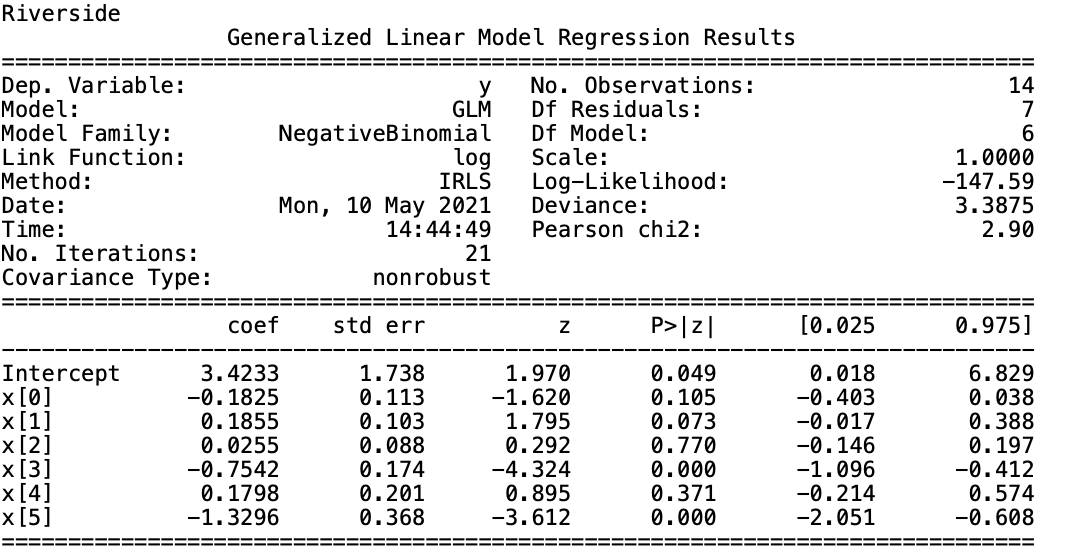
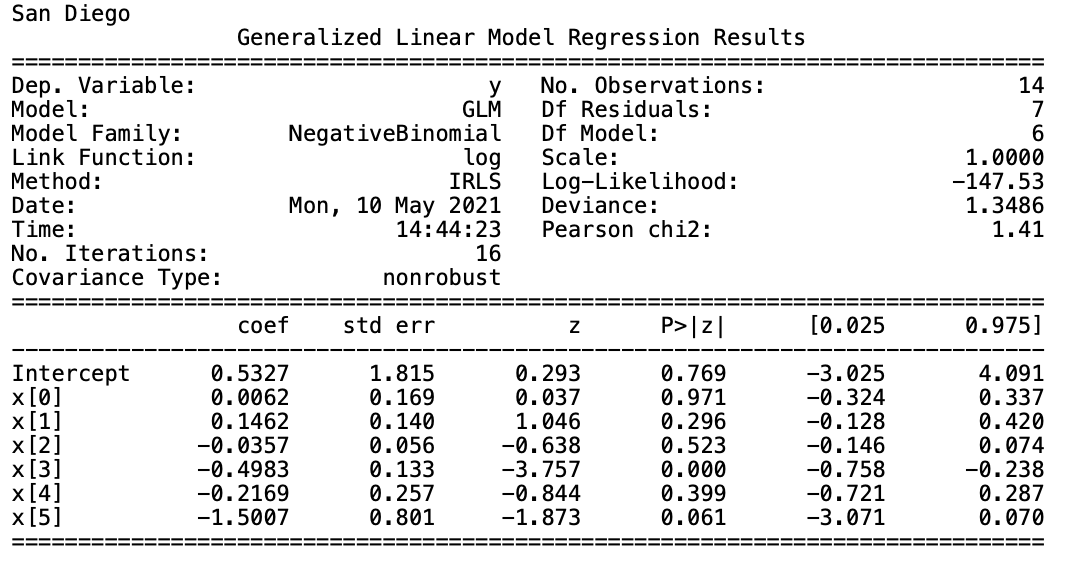
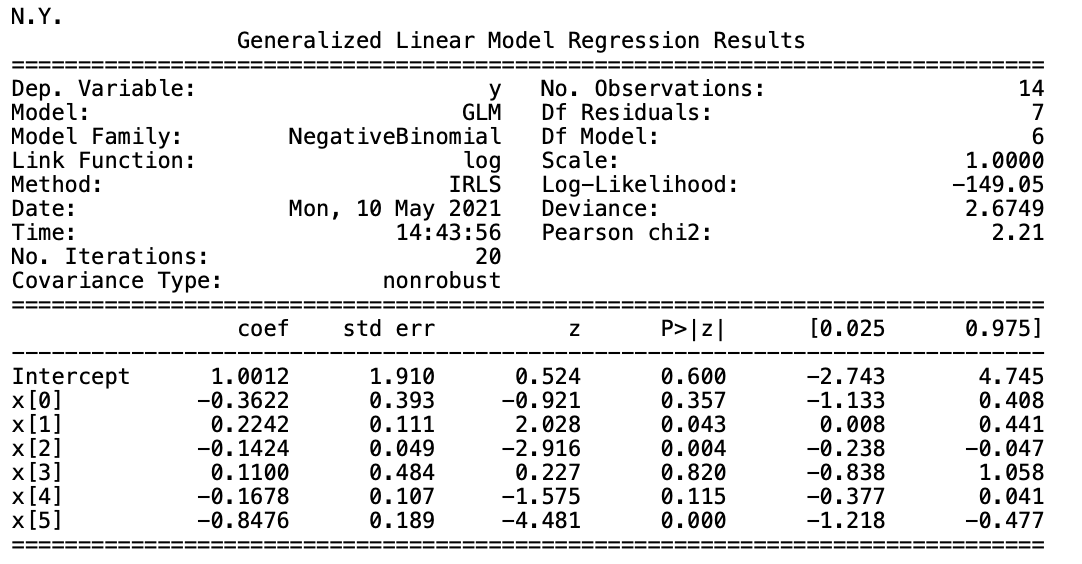
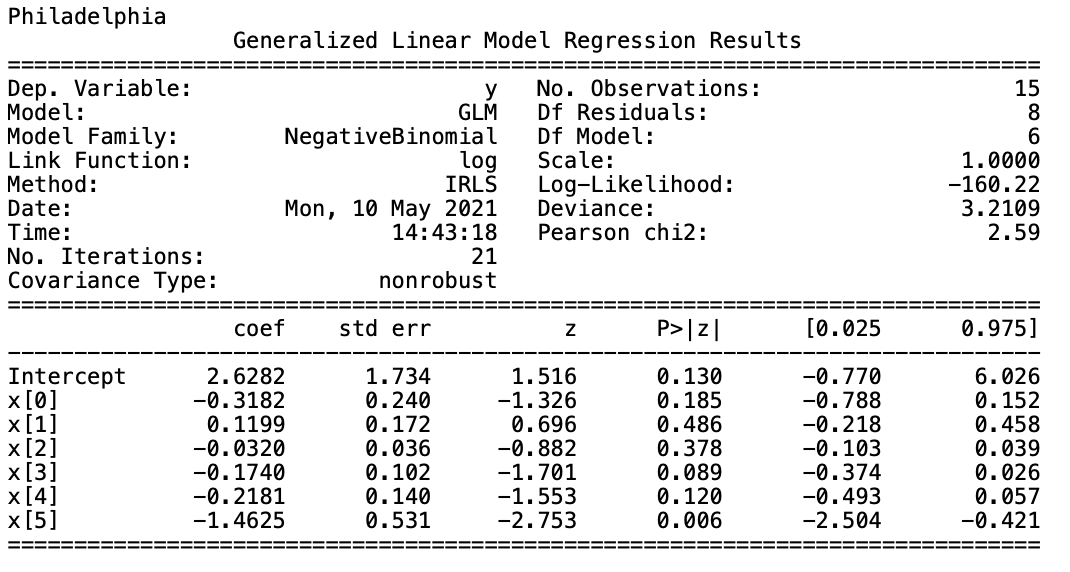
Figure 2: Transit stations mobility changes compared to average baseline value in accordance with log number of total cases in each county.

**4.1 Prediction with GLMs and nonparametric methods**

Given the trends we saw in retail and recreation, grocery and pharmacy, parks ,etc. and their correlation with the log of total COVID-19 cases, it seemed appropriate to explore whether there exists a generalized linear model that would help us predict the number of COVID-19 cases. We used Statsmodels’ GLM to model the relationship. Because the correlation was observed with the log of the total number of COVID cases, we understood that the link function had to be that of a logit. Keeping in mind the distribution of our data, it was determined that the negative binomial family was the most appropriate for the purpose of our GLM. In addition, we chose to fit a GLM for every metropolitan city, as opposed to fitting one GLM to all of the cities, since COVID restrictions and the city's population density varies drastically. Although we originally tried to fit one GLM to the metropolitan cities in our dataset, we received a very poor model. We decided that this can be an outcome of population density, and COVID restrictions, which are just two confounding variables to the number of COVID cases. After fitting a GLM model to each county, our results were improved dramatically. We believe that this improvement supports our belief of differences in each county.

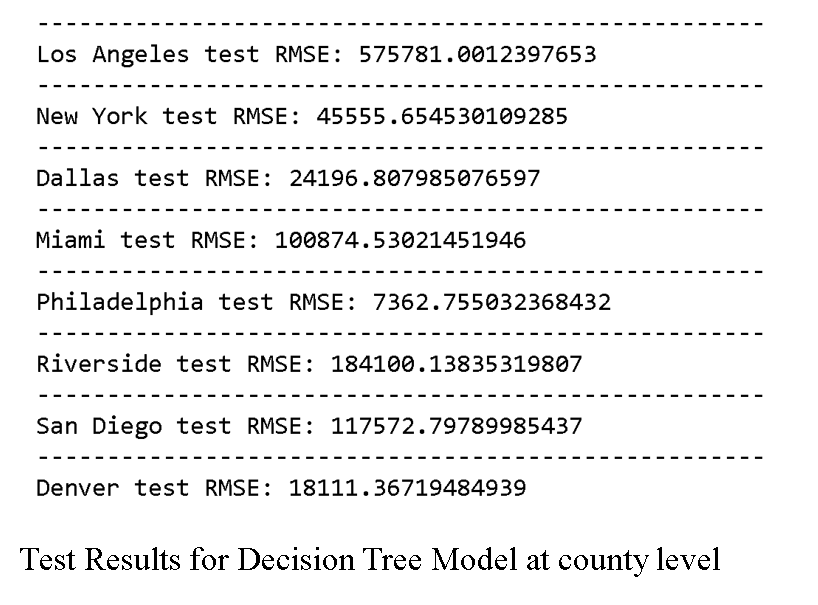
To compare our GLM model, we decided to use a nonparametric method. Initially, we had two choices for nonparametric models: Neural Networks and Decision Trees. Usually, neural networks need tons of data to work effectively. Plus, they are not as good as decision trees when it comes to interpretability. Therefore, we moved forward with the decision trees. We used scikit-learn’s Decision Tree Regression model to predict the number of total covid cases. Before fitting the data, we splitted the data into two datasets for training and testing. To test the decision tree model performance, we used root mean squared error as an error metric and we evaluated our prediction on a test data that was not trained on previously. It is worth mentioning that since decision trees are usually prone to overfitting, we used a depth 10 decision tree model for better generalizability. In addition, as in the GLM case, we performed this for each county because of the reasons stated above. Interestingly, before doing so, we actually tested a decision tree model for the whole metropolitan data. However, our results, based on the test set RMSE, were much better at the county level, further supporting our belief of differences in each county.

**4.2 Results**



Figures 3a-g

As we performed glm on all the metropolitan cities, the model yielded favorable results for all of the cities. Each of the cities had a log-likelihood in the -100s, this is a favorable range for log-likelihood value, as it indicates that the model is a good fit for our data. The less important metrics, such as deviance and Pearson Chi-square also performed favorably, with relatively low values for both, hovering around the single digit.



The figure on the left summarizes our results for a depth 10 decision tree. We can see that the decision tree performs differently at each county. Its best predictions occurs in Philadelphia .

The RMSE for each city upon performing decision tree modeling remains relatively high, this can be generally attributed to the fact that the data sets were rather small, and was not enough data for the decision tree model to begin to accurately predict the COVID-19 cases. Interestingly enough, the RMSE of the cities in their magnitude roughly follows the divergence between mobility and the log of COVID cases as seen in Figure 2.

**4.3 Discussion:**

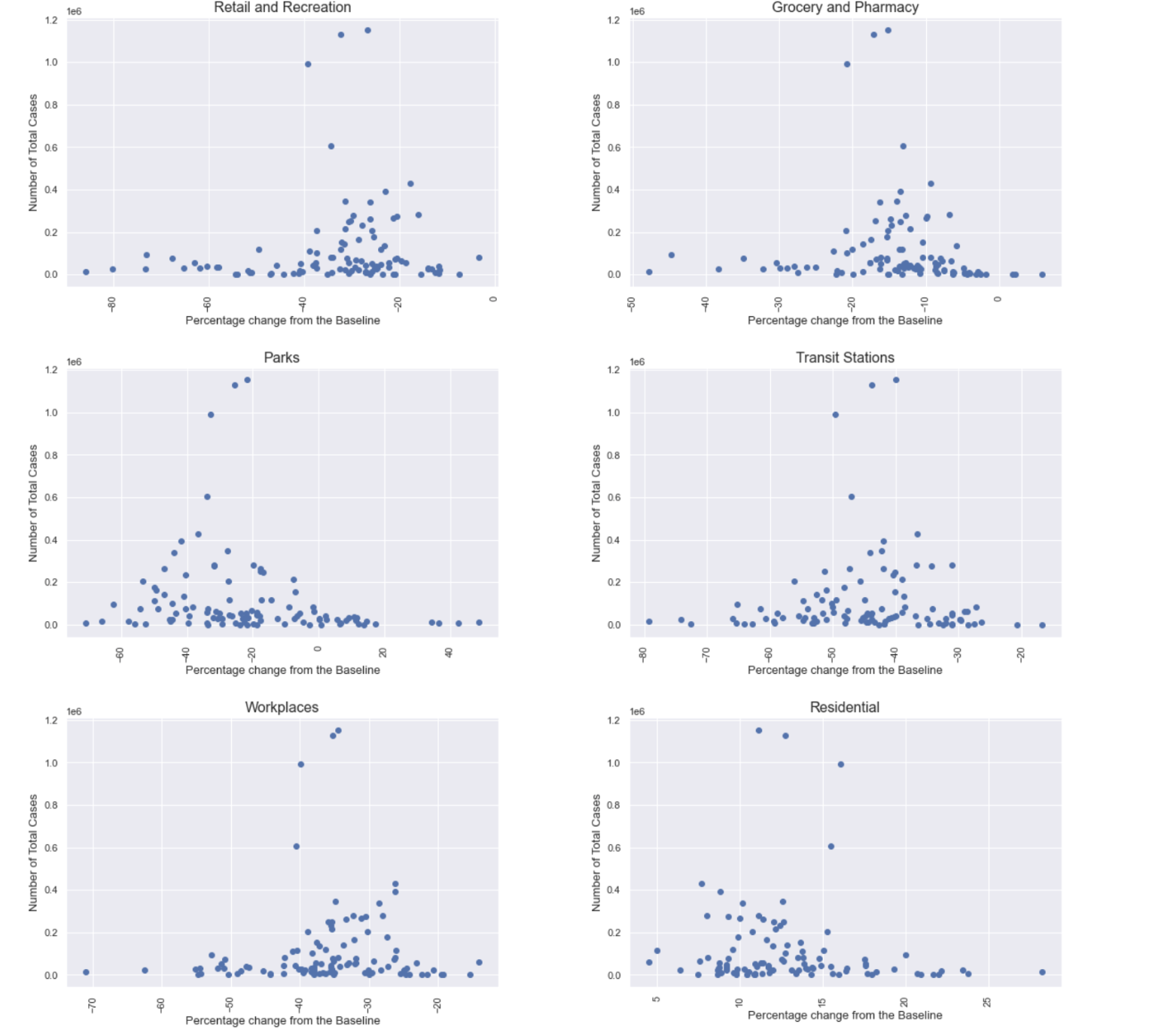
The results we obtained both from GLM and decision tree models show that the GLM model performed better than the decision tree model. We came to this conclusion by analyzing the log-likelihood and root mean square error across all counties. It can be seen that the log-likelihood does not deviate significantly across counties. However, in the decision tree context, the root mean squared error deviates significantly from county to county, leading to the conclusion that it is a poor fit and should not be used in future dataset.

However, we need to note that our both models are limited since they were trained on datasets that have few number of observations.

**5 .0 - Technique 2**

**5.1 Causal Inference**

We started by binarizing our data, this is because continuous data is difficult to perform causal inference upon. To binarize the various features, a threshold was required to delimit what was considered positive and negative. In this case, high facility utilization vs low facility utilization. We used the mean of each feature as the threshold for the binarizer, though we understand that this is perhaps an oversimplification, it was done in the sake of having a unified standard for binarizing the data. The same goes for the COVID cases, but because the total cases doesn’t inform us much of the dynamic state of the pandemic, we converted the data to the change of covid cases from day to day, and binarized that as well. Thereon, we performed causal inference on the binarized data.

From our previous GLM study, we were able to establish that transportation was among one of the highest in feature importance. Upon creating a number of visualizations for the other columns, we are able to establish correlations between transportation utility and the other non-outcome features, along with some levels of correlation between the non-primary features (non-transportation utilities) and COVID-19 cases.

The result of our causal inference is as follows:

In this scenario, each data column that deals with mobility is our treatment, and COVID-19 cases are our outcome. Each of the other independent variables (mobility columns) are confounding factors to every other variable, as any outing would contribute to the usage of transportation stations, as well as a potential rise in COVID-19 cases due to social exposure. As illustrated in the causal graph below.

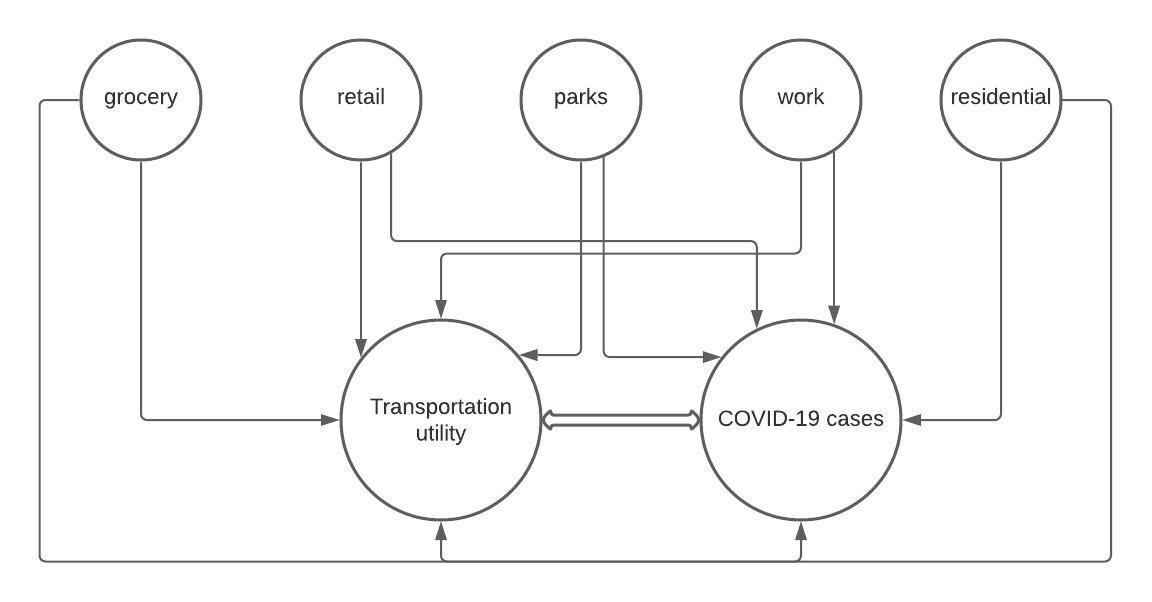


Figure 4. The causal graph for transportation utilization and COVID-19 cases,where non transportation features are treated as confounding variables, contributing to the use of transportation facilities and the rise in COVID-19 cases.

**5.2 Results:**

Our prior assumption was that COVID19 spread and use of Public transportation were correlated. This assumption was drawn from the idea that metropolitan areas utilize their public transportation systems. Although there exists some dependencies between each confounding variable, it is unclear if there exists a relationship between the outcome groups and treatment. For example, the scatter plots above, shows there the treatment and outcome groups have no relationship. But the following study contradicts these findings [[1]](#footnote-0) [[2]](#footnote-1). Specifically, public transportation can be a major vessel of spread. This can be an outcome of too many confounding factors that we are not able to take into account. For example, within the time frame that our data has been collected from, the cities went into varying levels of quarantine, which affected people’s ability to go to work, parks, and etc. Public transportation itself varied in availability as cities imposed different restrictions on public transportation[[3]](#footnote-2).

As exemplified in medical studies, with a sufficient amount of confounding variables, or a number of variables contributing to a singular outcome variable, there may be no singular variable that can be identified as uniquely causally significant, due to a relatively even spread of feature importance.

**5.3 Discussion**

We suspect that the relationship between transportation and COVID-19 is not what we have purported it to be. Through further research, we discovered that the decline in transportation utilization was likely due to the imposing of COVID-19 restriction on public transportation. Our method was still limited in discovering the direction of causality, as the numerous confounding variables exhibited correlation with our two variables of interest. Additional data regarding the chronological availability of public transportation from city to city would’ve been helpful in determining the extent to which COVID-19 guidelines impacted transportation utilization. This would also go a long way in clarifying any confusion with the confounding variables, where utilization of parks and recreation facilities have gone up while transportation utilization was stunted.

**6. Conclusion:**

Through our study of facility utilization in relation to covid-19 cases, we have found that with sufficient information regarding a city’s various utilization metrics, epidemiological prediction can be reliably produced. It is important to note this can only be reliably drawn from city/county levels, as on national level no generalized linear model is nuanced enough to account for the numerous confounding factors that impact epidemiological spread on a local level. This is further confirmed by the lack of accuracy in our non-parametric method, which relied on a larger data set, and failed to accurately predict COVID-19 cases. Our causal inference investigation led to the realization that causal relationships can be especially tricky to determine when there are a number confounding factors cannot be ignored. It also led to the consideration that between any two variables, there could exist a feedback loop of causation; as one factor influences the other, the second factor influences right back, resulting in a trend that self perpetuates for a period of time.

Considering that we found a relationship between transport facilities and the spread of COVID-19 cases, one call to action could be to conduct epidemiological studies in areas that have large populations that use or have access to transport facilities. Such studies could give us further insights on transport related policy changes that the government could make in the case of another pandemic or endemic.

We merged the provided transportation and facilities data with COVID-19 county level cases data from the Yu-group. This was a necessary step for us to study the relationships of interest in our research questions. The merging of these tables was beneficial in that they allowed us to have a chronological view of how the pandemic, transportation, and facility utilization progressed. Furthermore, the data’s merging allowed for the utilization of GLM in the answering of our research questions.

We did not account for the fact that each independent variable would be a confounding factor to every other variable in the causal analysis. If we had data that provided us with non-confounding variables we would have probably been able to carry out more concrete analysis. Regardless, we were able to establish that there was some causal relationship between each of the mobility variables and the number of COVID-19 cases.

GLM models can be built upon facility utilization and other infectious diseases such as the flu, and help us build a base model for disease spread. This is not limited to respiratory diseases, but other diseases as well, through an understanding of public venues in cities, epidemiological models can be built with an additional understanding of how a disease spreads, helping us to anticipate and prepare for potential outbreaks.

7 Appendix

<https://github.com/Yu-Group/covid19-severity-prediction/tree/master/data>

<https://covidcountydata.org/data/download>

<https://medium.com/@akelleh/causal-inference-with-pandas-dataframes-fc3e64fce5d>

<https://scikit-learn.org/stable/modules/tree.html>

1. <https://scholarcommons.usf.edu/jpt/vol22/iss1/1/> [↑](#footnote-ref-0)
2. https://www.sciencedirect.com/science/article/pii/S2590198221000737 [↑](#footnote-ref-1)
3. <https://scholarcommons.usf.edu/jpt/vol22/iss1/1/> [↑](#footnote-ref-2)