Predicting Cause Specific Mortality Rate of Covid-19 in U.S. Counties

Teresa Bodart & Anna Litskevitch
Department of Data Science, University of California, Berkeley
Data 100: Principles and Techniques of Data Science
Joey Gonzalez and Ani Adhikari
May 13, 2020

Abstract

For our final project in Data 100, we are analyzing the COVID-19 data current to April 18, 2020 to determine what factors influence the impact of COVID-19 on specific counties in the continental United States. The purpose of our exploration is to find patterns and create a model that predicts cause-specific mortality rate in counties across the U.S. using data from 'time_series_covid19_deaths_US.csv', 'time_series_covid19_confirmed_US.csv' and 'abridged_couties'. The cause-specific mortality rate is the proportion of the population that has died due to a specific cause, in our case COVID-19. We wanted to focus on cause-specific mortality as we believe that it is a better representation of the impact that COVID-19 is having on a specific county than the absolute number of deaths because it accounts for the size of the population. Through this project, we found that the most important features in predicting cause-specific mortality rate were related to the susceptibility of the population and existing health care infrastructure.

Introduction

We chose the COVID-19 data sets because we felt they were the best opportunity to apply our data science knowledge gained from this course to a pressing, real world issue. The other topics are of course interesting, but the chance to work with data that data scientists across the world are using in real time to make projections about the pandemic was too good to pass up. The datasets are really interesting as they contain an incredible amount of information on the American public. Additionally, it is intriguing to know that major news sources such as *The New York Times* are using this very same data for the dissemination of COVID-19 information across the U.S. We will be drawing from multiple datasets to create visualizations showing where the coronavirus is hitting across the country, and where it is hitting the hardest. From there, we will attempt to create a model which shows the factors, of the ones provided in the datasets and new features created from them, that contribute the most to the cause-specific mortality rate. We initially thought about assessing the impact of COVID-19 based on the number of deaths that occurred in each county, but we found that this measure of outcome would mostly be based upon the population of the county itself since COVID-19 has already spread to every part of America. A better estimate of the impact is the cause-specific mortality rate, as having a larger proportion of a population die from COVID-19 would be more indicative of the severity of the outbreak than just the absolute number of deaths.

Description of Methods

Because we wanted to see what characteristics of a county affect the cause-specific mortality rate, we used the datasets with county-level granularity, specifically 'time_series_covid19_deaths_US.csv', 'time_series_covid19_confirmed_US.csv' and 'abridged_couties'. Initially, we wanted to use a hexbin plot, similar to the one we created in Project 1, in order to illustrate geospatial trends (See note about why they look different in the discussion section). We found this plot misleading as it represented each county with a single dot, giving the illusion that it was representative of a single city. We instead decided to use geopandas. We downloaded shapefiles from data.gov, which provides public access to datasets from the federal government.

Since our question is focused specifically on counties in the continental U.S., we filtered out data points that were state-level or described U.S. territories. We included the District of Columbia, as it is in the continental U.S. and is similar in size to a county. We felt that using the number of confirmed cases and deaths on dates prior to 4/18/20 as features would not give us interesting information about the specific characteristics of the counties, so we opted to only include the latest count of the deaths and confirmed cases. However, we did calculate the number of days since the first case of COVID-19 was reported in a county and the rate at which the epidemic was growing. We added these columns, as we believed that knowing how much time COVID-19 had to spread in a county would influence how its current impact. Along with those, we kept the latest reported number of deaths and cases and the code to identify the county from the time-series datasets.

Most of the data is contained within 'abridged_couties' and as such, it required the most cleaning. When observing the data frame, we noticed a good amount of columns that we could tell by eye had very sparse data. We got rid of these columns, as it would be difficult to decode any useful associations with such a limited amount of information. We also noticed that two counties in this data frame barely had any data. One of these counties was Yellowstone County with FIPS 30113. This county was not included in the time-series data frames. After researching, we found that this county had been integrated into Gallatin County 30031 in 1970, meaning that we could drop this county from our data frame. Another county with limited data was Shannon County 46113, which also does not appear in the time-series data frames. After another round of research, we found that this county was renamed Oglala Lakota County 46102, which does appear in the confirmed and deaths datasets. We wanted to include this county, as there is data available on it from the deaths and confirmed data frames. We then filled the null values of this county with the mean values of counties located in South Dakota, specifically those with a Rural-Urban continuum code of 9, the same as Oglala Lakota County (taken from the United States Department of Agriculture), as they likely resemble this county the closest.

To make sure that no null values were left in the data frame, we made a bar graph to represent the columns that still contained null values (See Fig. 1). We noticed that there were a large number of null values in columns that describe the time certain restrictions were put into place in the county. When we looked back on the data, the null values corresponded to counties that had not issued such guidelines. We decided it would be fine to replace those values with a 0, which would represent that no guidelines were put into place. We also dropped the '3 year mortality for 45-54 year olds' column, as almost a third of the values are missing. The rest of the columns have an acceptable amount of missing values, so replacing the values with the mean of the column should be an acceptable guess for the missing value.

The last data frame we used was the one created from the downloaded county shapefile converted to shapefile (I switched this to a csv so that the file was less than 100 mb). I left in the columns necessary to plot the counties and the county code to make the data frame more manageable.

Finally, we merged all our data frames together on the county codes to create one comprehensive data frame to be used for modeling and visualizations. Then we added a few features related to epidemiology such as incidence rate, case fatality, and finally, cause-specific mortality.

Our first visualization mapped the number of deaths to each county (Fig. 2), but we found that the only thing showing up was New York City on account of how many deaths there are. This made the visualization less than useful, so we instead calculated the log of the number of deaths in order to compensate for New York City, and that provided us a much better visualization (Fig 3.). From this, we observed that most of the deaths were concentrated around major cities and more urban centers, such as New York City and Los Angeles. We then wanted to compare that trend with cause-specific mortality rate, so we plotted cause-specific mortality rate for the different counties (Fig. 4). New York caused us difficulties again, so we took the log of the cause-specific mortality rate. Here, we still found higher values around highly populated areas, but also a few new counties that were not as prominent in the previous plot (Fig. 5). It seems, then, that the largest counties are not the only ones that are being heavily impacted by COVID-19. To illustrate this further we mapped the log of the population density of each county and saw that it more closely resembles the number of deaths than the cause-specific mortality rate plots (Fig. 6).

We then began to explore the data available to us about each county and how it could be used predictively. We noticed that we had data on stroke, heart disease, and respiratory mortality rates, and wanted to make sure that they were not linearly dependent. From our 3D scatter plot

(Fig. 7) we concluded that all three of the columns had an overall linear relationship and that we should refrain from using more than one in our model.

Modeling is one of the most important tools for answering our question because we can assess which combinations of features create the greatest predictive power for cause-specific mortality rate. We used a linear regression model with RMSE as the error function and 5 fold cross-validation to assess the model. We decided to use a linear regression model given the fact that we wanted to predict a continuous outcome and our data was not classification-based as it was in Project 2.

For our first model, we took a look at how the model worked when we used all the numerical features we had and then by plotting the coefficients (Fig. 9), which features it used the most. We then created another model that exclusively used the features we saw used in the previous coefficient chart (Fig. 10) to see if they would combine to a good model (Model 2). This did not lead to a good reduction in error, so we changed our approach to a more trial and error approach that focused on our knowledge of epidemiology and the previous data exploration we did. This approach got our cross validation error under 100 (Model 3).

Summary of Results

We used cross-validation with RMSE to assess the success of our linear regression model. RMSE weights large errors heavily, and so it is a good measure to use when we want to avoid large errors. Our RMSE, calculated with five-fold cross validation, went from 106.79 to 105.30 to 99.64, showing improvement as we refined our model. We feel that our final RMSE, which is under 100, is an acceptable error for the following reasons: the range of cause_specific_mortality_million is very large, having values in the thousands, so 100 is comparatively a small error; our residual plot shows that many of the errors are clustered around 0, even if the residual plot also shows a pattern meaning that a linear model might not be the most appropriate model for the data; the standard deviation of the cause-specific mortality rate was around 170, which means than our model is better than random chance at predicting the outcome. Since we are content with the cross-validation error we achieved, we feel more confident about the inferences we will describe in our discussion.

Discussion

One interesting feature that we found was proportion_65+ in each county. We created this feature using the PopulationEstimate65+2017 and PopulationEstimate2018 columns from abridged_couties. Our thinking was that a proportion rather than count would be more meaningful in our model, and indeed when we examined the coefficients for our linear model in Fig. 10 and Fig. 11, proportion_65+ was one feature with a relatively big impact on the model. ICU beds per person was also really important for our model, something which we had to calculate from the columns provided. It was interesting to see how these features became important to our model given they were simply manipulations of features already in the datasets.

One feature we assumed would be important was PopulationDensityperSqMile2010. This was because there was still a tendency for more densely populated places to have a greater cause-specific mortality rate, but when we added or removed it from the model our loss function, cross_validation_rmse, was not improved, and our graph of coefficients showed that it had minimal impact on the model. This indicates that it is not so much the population size or density that determines the spread of COVID-19, but a feature like ICU beds per person, which indicates the preparedness of the area, that better predicts the cause-specific mortality rate of the virus.

We initially had many issues creating visualizations which showed the confirmed and death cases of COVID-19 across the U.S. At first we attempted to create hexadecimal plots similar to the one we made of San Francisco in Project 1b. However, we quickly realized that the counties across the U.S. are much more complex than the simple dots we used in that project, and those dots created a misleading model. From there, we moved into geopandas, where, after learning the module, we implemented shapefiles from the government of counties in order to create our geovisualizations. However, the shapefiles data downloaded from data.gov was too large a file to submit to gradescope, so we had to work around that, downloading it into a csv to make it smaller, which led to us losing the beautiful county shapes. You can see what one of our original plots looked like in Fig. 12. Originally we wanted to simply do case_fatality_rate, but this model was too impacted by the counties where one person had COVID-19 and one person died, leading to a huge overestimation of the fatality predictions of the model. Finally, as described in the description of methods, we also had difficulties with Null values, and deciding which were okay to drop, which were okay to send to zero, and which needed to somehow be filled in. Another issue we came across

One main limitation of our analysis was using the log function to reshape our data for more useful visuals. While doing this was the right decision to improve our plots, this led to some errors in that by filling certain Null values with zero, they were artificially inflated compared to other values which had been smaller decimals. Though we tried our best to not use any sparse columns, there were still a few that had null values in them. We believe that we dealt with them appropriately, but the lack of data is still a limitation we should be aware of. Another limitation of our model, though it was the best type of model from the ones we know how to use, is that it is clear based on our plot of the residuals that a linear regression model might not be the best for our data.

In terms of ethical concerns over our data, we used mean values from surrounding counties in the same Rural-Urban continuum code as Shannon County aka Oglala Lakota County to fill in Null values in abridged_couties. While this was the best option we could find while still including the county in our model, we realize that this county lies within the Pine Ridge Indian Reservation, and by assuming values based on counties which might not be in that same reservation, we could be grossly misrepresenting conditions in that county.

Some ethical concerns we have with our overall question are mostly concerned with the way features that indicate a higher cause-specific mortality rate of COVID-19 might be used to single out certain populations. One example, even though it does not appear in our model, is obesity. Experts have noted obesity as a risk factor for COVID-19. This could lead to an even greater stigma against certain populations in the U.S. Additionally, this question, and any "answers" that come from it could be leveraged in politics, one side or another using this data and COVID-19 to push certain agendas, again perhaps dealing with certain vulnerable populations. We also noticed in our model that democratic to republic voter ratio played a role in our model, and this knowledge can definitely be used by political parties to stoke animosity. We can start to prevent some of these ethical concerns by selecting for a variety of features in a model, not singling out any specific populations. We can also remember to look into biases which may have affected data collection in certain regions. If we are training a model based on data collected in unethical or ethically questionable ways, we have to be sure to watch out for the perpetuation of such things in the model itself. We should also always preface the result of our model by stating that the correlations we find may be confounded by other variables and should not be used as proof of direct causal relationships, but rather suggestions for further research.

We think that additional healthcare data would have strengthened our analysis, which rested heavily on one or two features. Specifically, as it is known that ventilators would be key in treating patients with COVID-19, data on the number or proportion relative to the population of ventilators in each county might have proved to be indicative of cause-specific mortality rate, as this would be another type of preparedness like ICU beds per person. Other hypotheses which seem relevant to follow up with would be more specific data on the 65+ populations in each county. As we know, COVID-19 has been severely impacting elderly care homes, and so data on where the over 65 population in a county resides—a private residence or a nursing home—might help predict the death rate in that county.

Figures

Fig. 1: Null Values per column in Cleaned abridged_counties

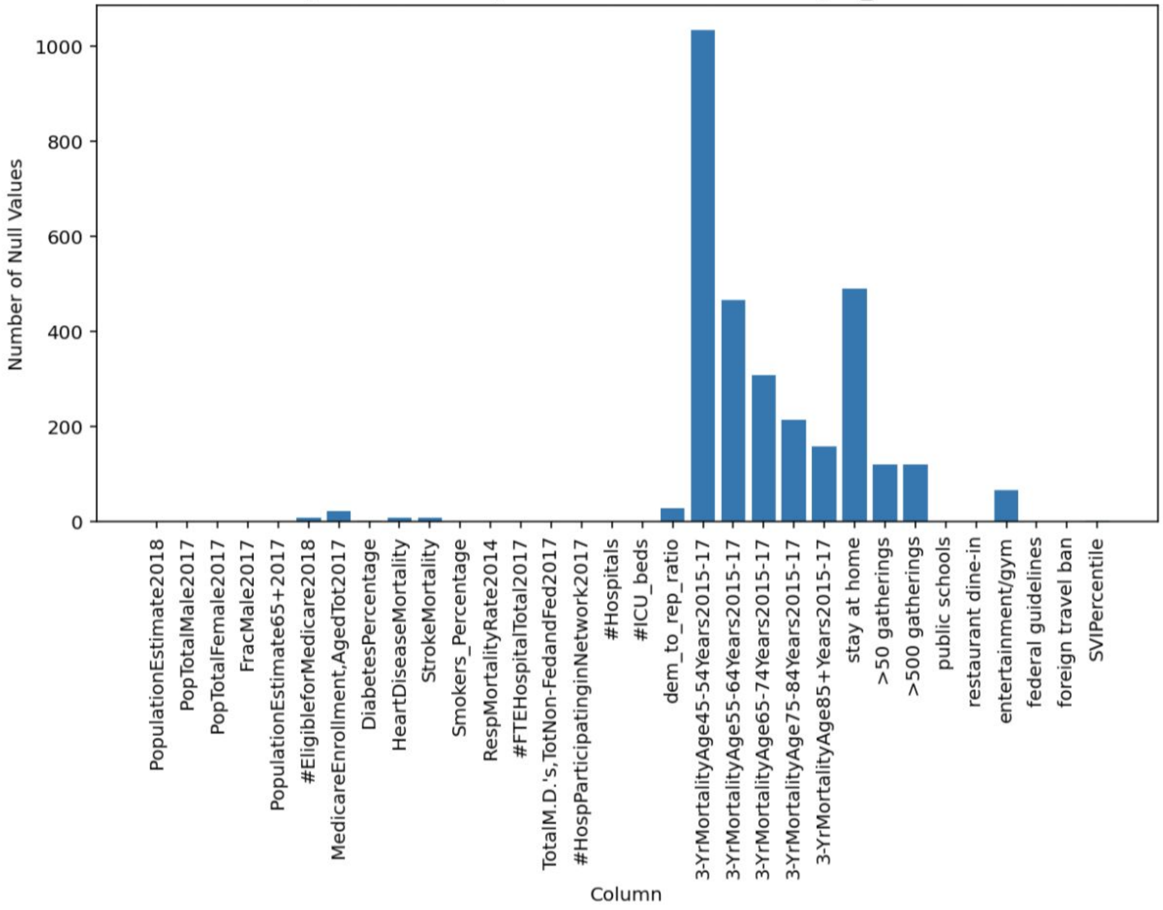
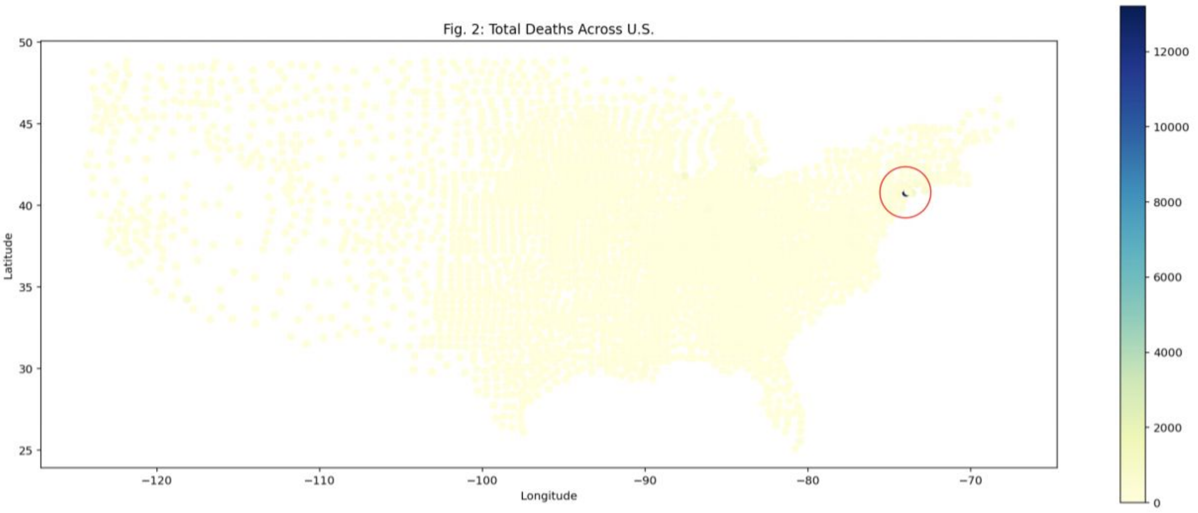


Fig. 2: Total Deaths Across U.S.



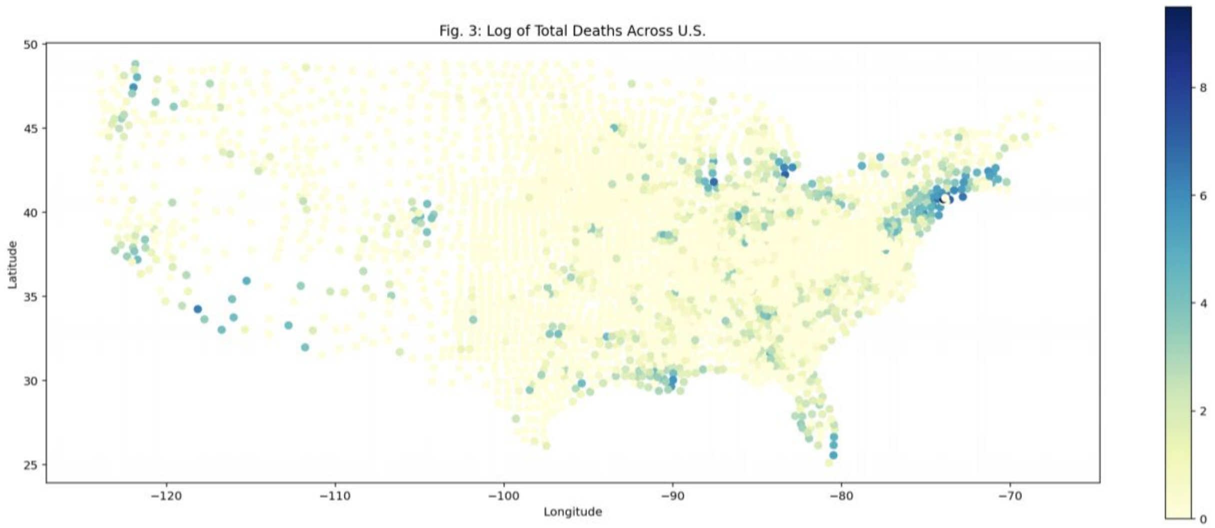


Fig. 3: Log of Total Deaths Across U.S.

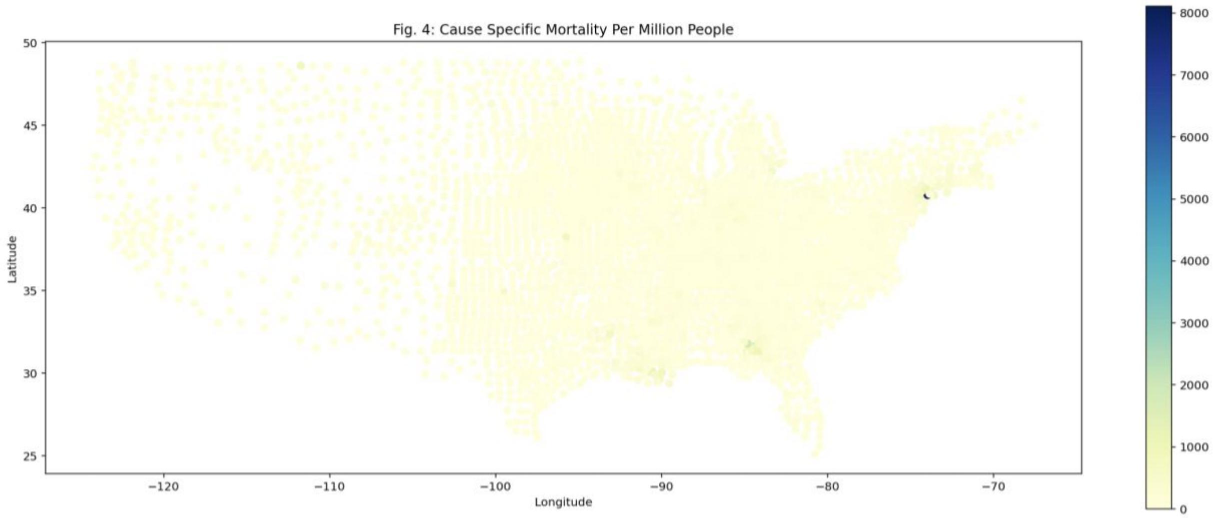


Fig. 4: Cause Specific Mortality Per Million People

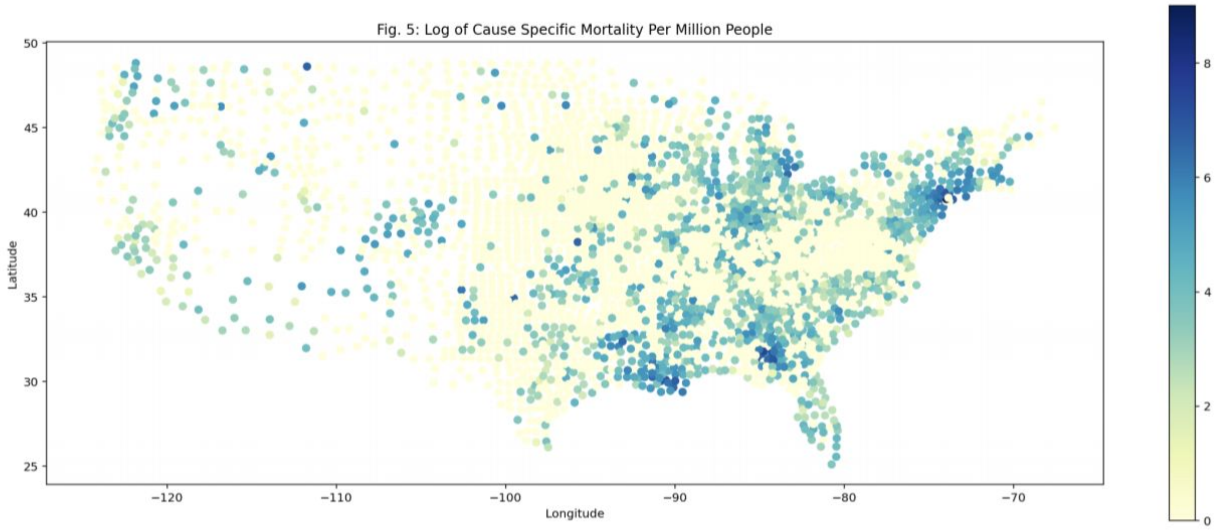


Fig. 5: Log of Cause Specific Mortality Per Million People

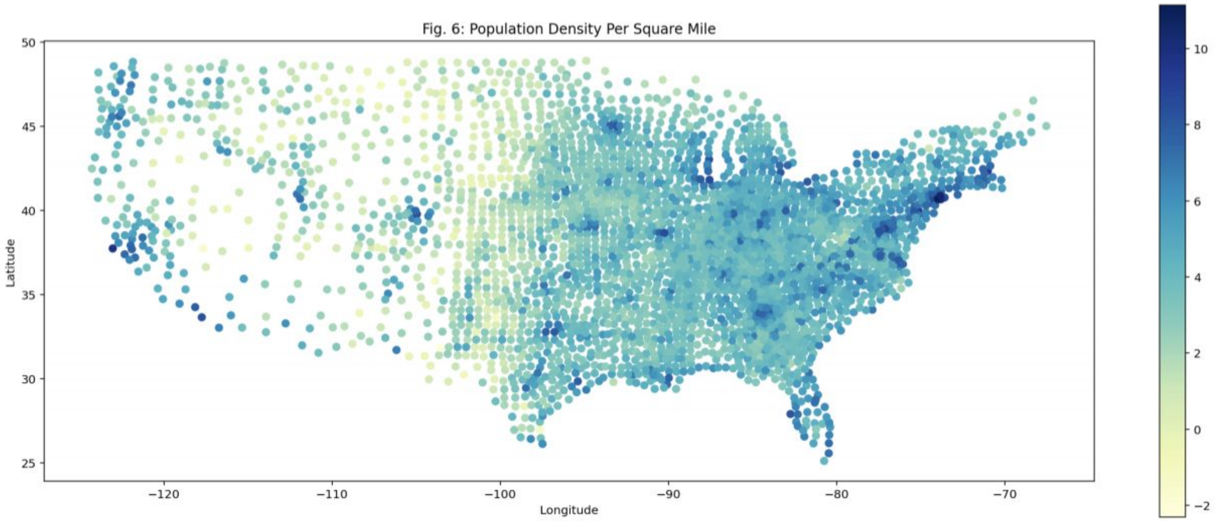


Fig. 6: Population Density Per Square Mile

Fig. 7: Correlation Between Stroke Mortality, Respiratory Mortality and Heart Disease Mortality

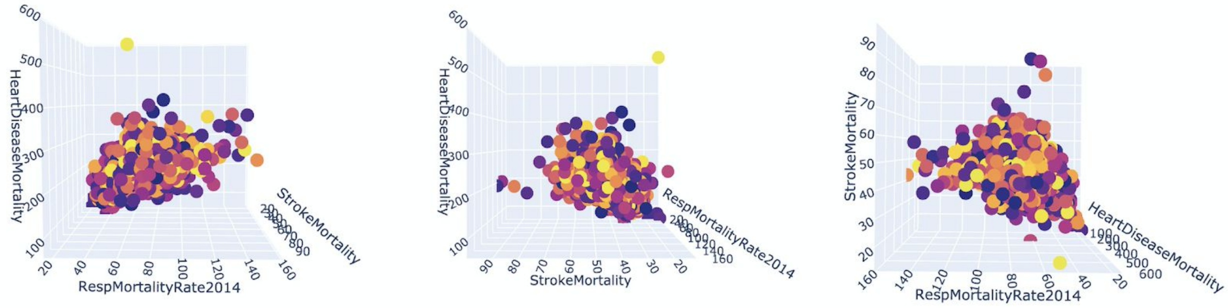


Fig. 8: Residuals vs Observed

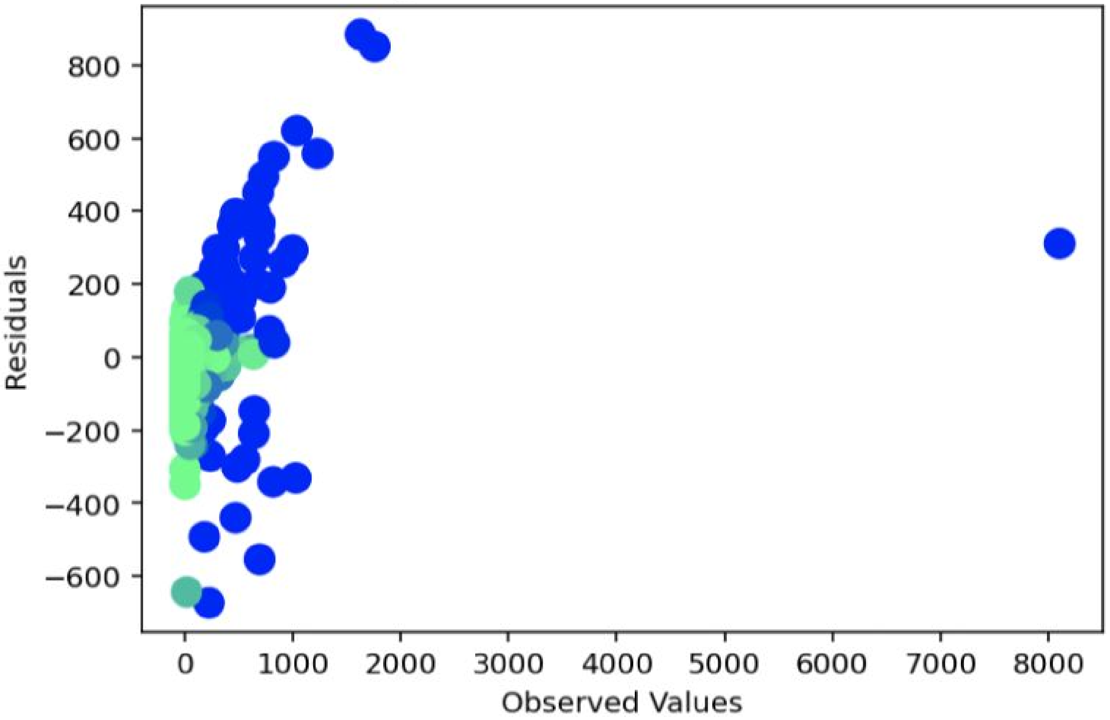
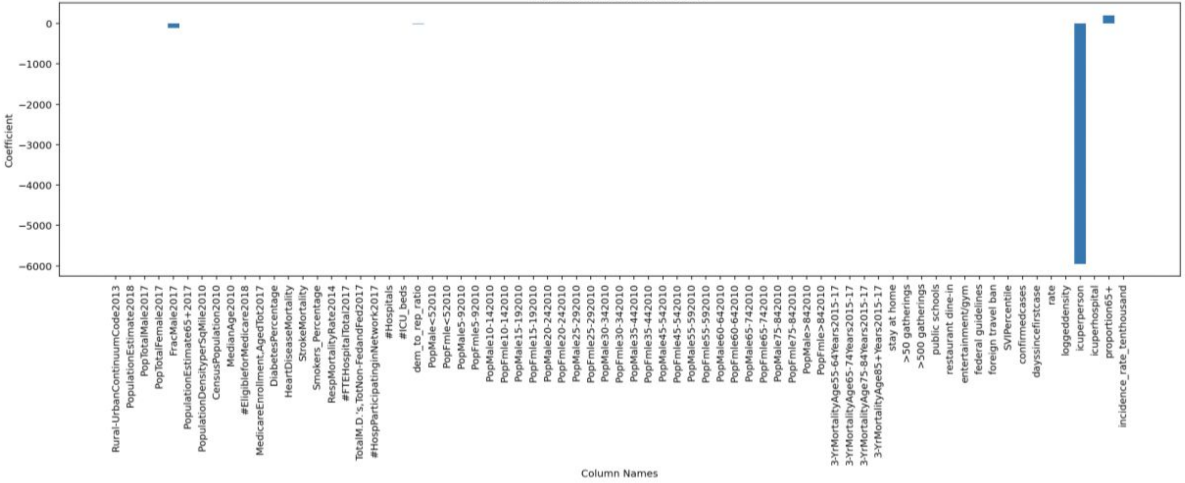


Fig 9. First Model Coefficients



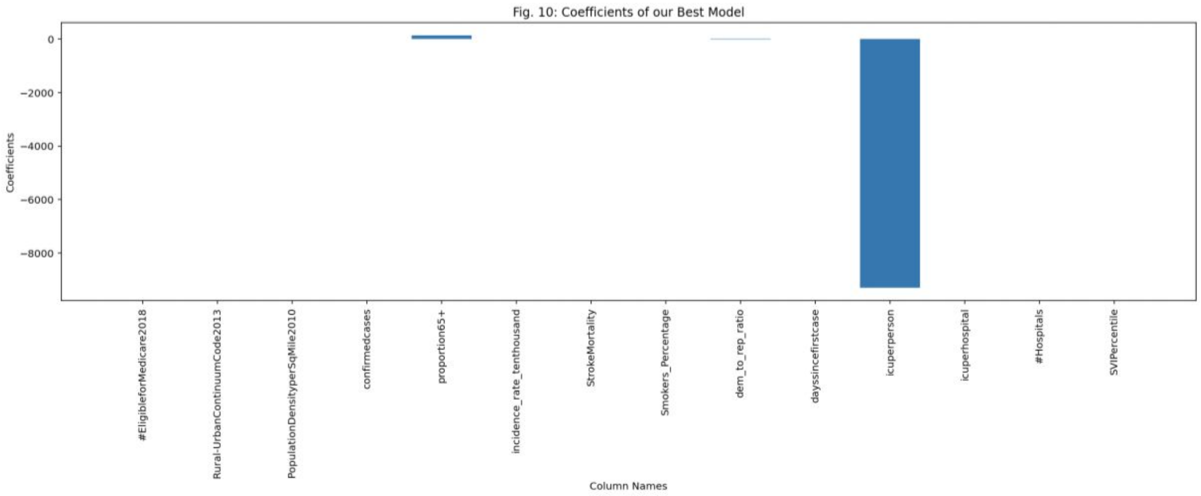


Fig. 10: Coefficients of our Best Model

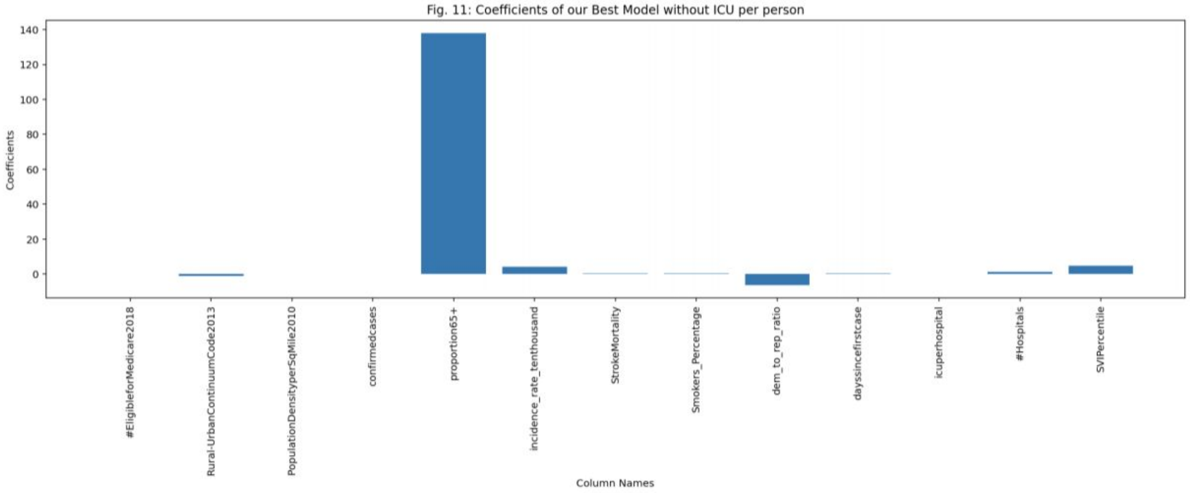


Fig. 11: Coefficients of our Best Model without ICU per person

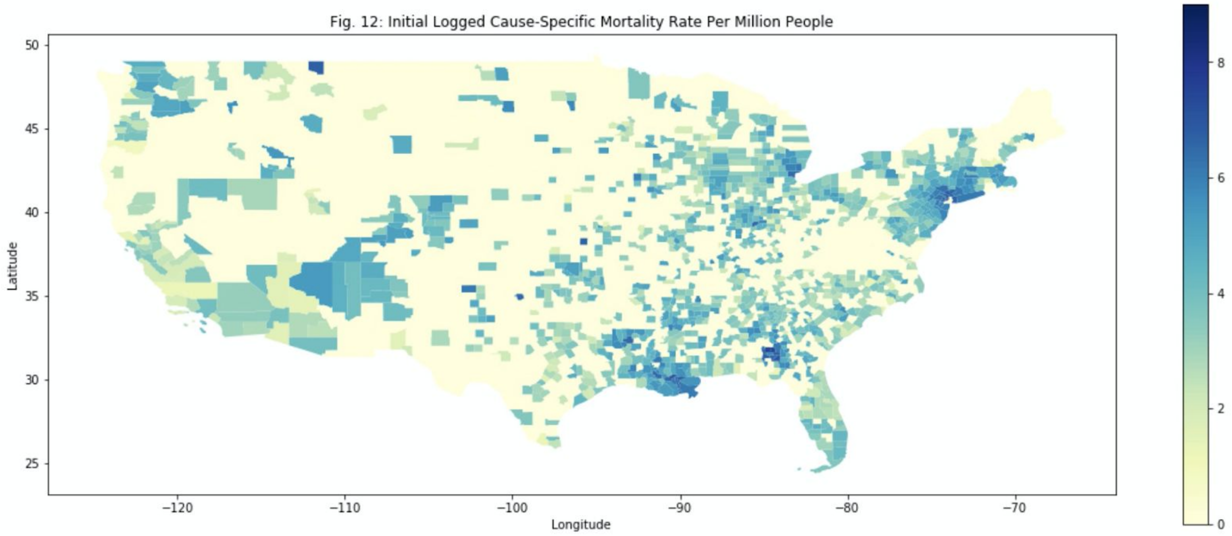


Fig. 12: Initial Logged Cause-Specific Mortality Rate Per Million People