Exploring Factors of COVID-19 Mortality Rates in US Counties

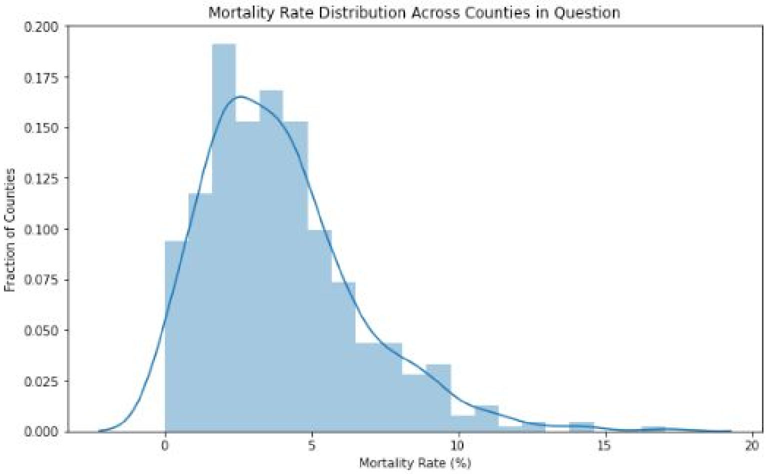
Jacob Bland & Daniel Covelli
UCB DS100 Spring 2020

Abstract

In this paper we took an exploratory look at the mortality statistics of COVID-19 in US counties. After initial exploration of the data, we asked *what characteristics of a county, if any, were the most directly related with its COVID mortality rate*. Our hypothesis was that there are important county-level characteristics that affect COVID mortality rates. We found there seem to be some characteristics, like number of cases, median age, and overall respiratory mortalities, that do affect the COVID mortality rate. However, our model never achieved a desirable level of goodness-of-fit, therefore, our results are inconclusive.

Introduction

The global community is being forced to reckon with COVID-19, the disease caused by the novel coronavirus SARS-CoV-2. While the virus has sent nearly all Americans into lockdown, it's deadly effects do seem to vary across areas. For example, in New York, NY, the mortality rate has been nearly 10% according to the Johns Hopkins data. San Diego, CA, on the other hand, has seen a mortality rate of only around 3%. We wanted to know why. The variation in this statistic is further evidenced in the below plot, where we see that while the mortality rate of the disease falls between 0% and 5% for the majority of the counties, there are those for which it rises above that boundary.



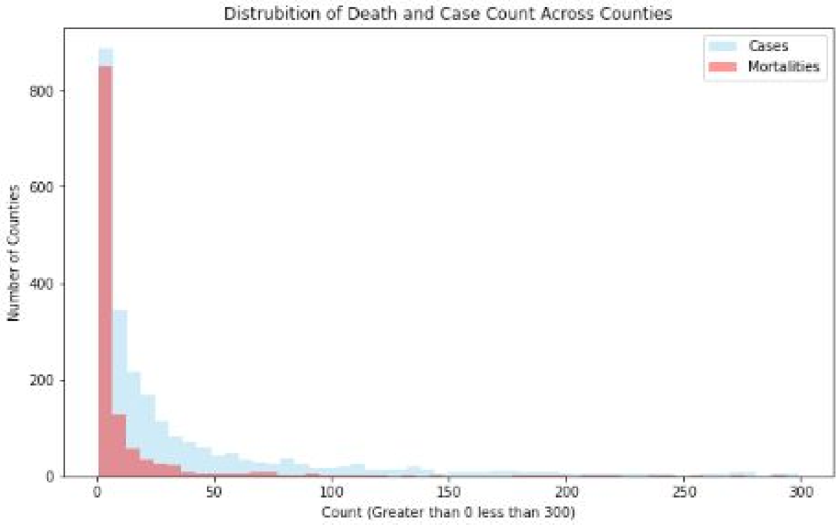
Of course, there could be things at play here such as testing availability, and perhaps the insight that a place with fewer cases is possibly better able to track the case count, or is also possibly just not as far along in this devastating journey.

Dataset

The analysis of this paper relied upon the data provided by two repositories. The first, provided by the Yu group at UC Berkeley, holds information about each and every county in the United States, from such statistics as its median age, to the percentage of its population eligible for Medicare (this dataset was named ***counties*** in the notebook). The second, provided by John Hopkins, holds time series data on the aggregated counts of COVID-19 cases and deaths in US counties (these datasets were named ***cases*** and ***deaths*** in the notebook). Putting the data from these two groups together, we were able to gain insight into some of the characteristics of a county that make it more susceptible to a high mortality rate from this disease.

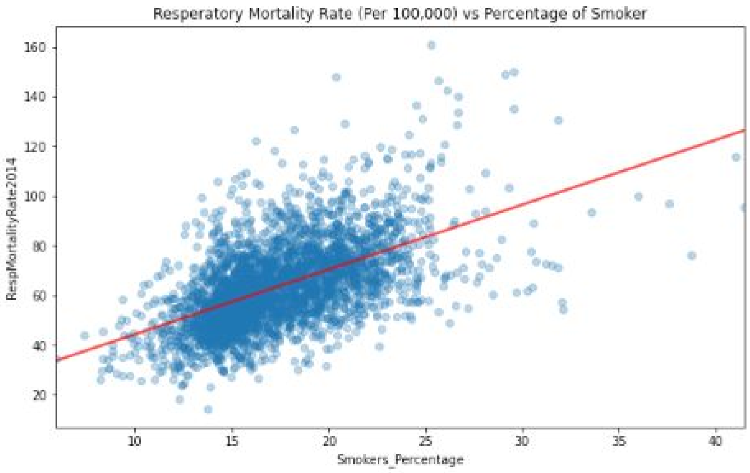
Exploratory Data Analysis

Before beginning this modeling pipeline, we investigated the general structure of data and searched for underlying relationships. Our first step was to print out the datasets and began researching the different explanatory variables we had at our disposal. In doing this we also realized that there were more than 50 states in the ***cases*** and ***deaths*** datasets, indicating that other observations, like cruise ships and US territories, were present in our data. After recognising some important characteristics, we moved on to visualizing the distribution of cases and deaths across counties. Without having a mortality rate statistic already built, we began by visualizing the case and death counts across counties.



Upon visualizing this distribution, we discovered that many counties had zero COVID cases. This helped us realize that in order to calculate the COVID mortality rate, this subset of counties with 0 cases wouldn't be applicable.

We also sought to visualize some of the underlying relationships in our data as it presented. For example, we reasoned that the Respiratory Mortality Rate variable from our dataset might behave similarly to this new covid mortality rate variable. Under this assumption, we were able to explore how our mortality rate variable might relate to other features, such as the percentage of smokers in the population.



We see that indeed, there is a positive correlation here. While this did not involve the response variable around which we would originally build our model, it did give us early insight into a few of the relationships between features present in our datasets.

<u>Data Cleaning and Transformations</u>

With the initial round of EDA complete, we were prepared to manipulate our data. In EDA, we had identified that the ***cases*** and ***deaths*** datasets held very similar structure. Namely, they had an ‘Admin2’ column, which usually represented the county name, and a ‘Province_State’ column. Together, these two columns specified a primary key for each time series table which allowed us to join with the county name and state name columns of the ***counties*** tables. Upon inner joining our three datasets into one set, we intentionally dropped all of the unimportant US territory and cruise ship observations in the ***cases*** and ***deaths*** datasets.

With all of our desired data now in one table, we moved onto the task of wrangling it into a form suitable for modeling. The most straightforward of these decisions was dropping duplicate or otherwise redundant columns from our main dataframe. Additionally, the question arose of what to do with columns that held a majority of null values. After a quick inspection, the decision was made to drop them, realizing that any attempt at filling the null values would be a misrepresentation of the observation.

Another restriction that we imposed which narrowed the scope of our dataset actually emerged from a realization made during our feature engineering process. While we were creating a new column for the mortality rate, we instantly encountered a division by zero error. It was clear that since mortality rate was being calculated as deaths per cases, those counties with zero cases would be an issue, so we dropped them. In fact, it actually helped refine the question statement, since trying to reason about the factors of the mortality rate of a disease in a place where that disease simply does not exist is an ill-defined question.

The final restriction that we imposed on our data was limiting our observation to only countries with more than 100 cases. With fewer than that many cases, we reasoned that the variance of mortality rates was likely not due to any county level features but rather the small number of cases. We reasoned that 100 cases was sufficiently large to capture general trends about county level features and associated mortality rates. There were still more than 300 cases to consider after this additional constraint, so we still had a substantially sized dataset to work with.

Methodology

Assumptions

When deciding to use linear regression (with RMSE loss), we assumed that relationships between the features and output were linear. For example, we assume that as median county age increases mortality rate also increases, and the same for most of the other variables in the available dataset.

We also assume here that mortality rates fluctuate across time, that is, as case counts increase, we expect an increasing fraction of those cases to lead to mortalities. This assumption holds when considering the effect of increased cases on overburdened hospitals. In such a scenario it would be reasonable to expect mortality rates to increase, as doctors begin to ration ventilators, hospital beds, and other vital supplies. Because of this assumption, we have added case count as an important feature in our county-level analysis.

Feature Engineering

Our feature engineering pipeline relied on three important transformations (and one transformation that we didn't end up using). The most important step was to obtain our outcome variable by dividing the number of COVID deaths in each county by the number of COVID cases then multiply by 100 to get the mortality rate in percentage. The second most important feature change was normalizing our explanatory variables so that we could interpret their comparative importance in affecting COVID mortality rates. The third feature change was one-hot-encoding each state and census region name for each county. We later discovered that this third change led to drastic overfitting, when applied. The fourth change, which we never implemented, was converting NaN values in the policy categories to 0 and the dates (representing when the policy was implemented) to 1. There turned out not to be any recognizable distinction between countries that did and did not implement mitigation policies because almost all counties did. After running our data through this feature engineering pipeline, we split our data 75/25 into training and testing sets, respectively.

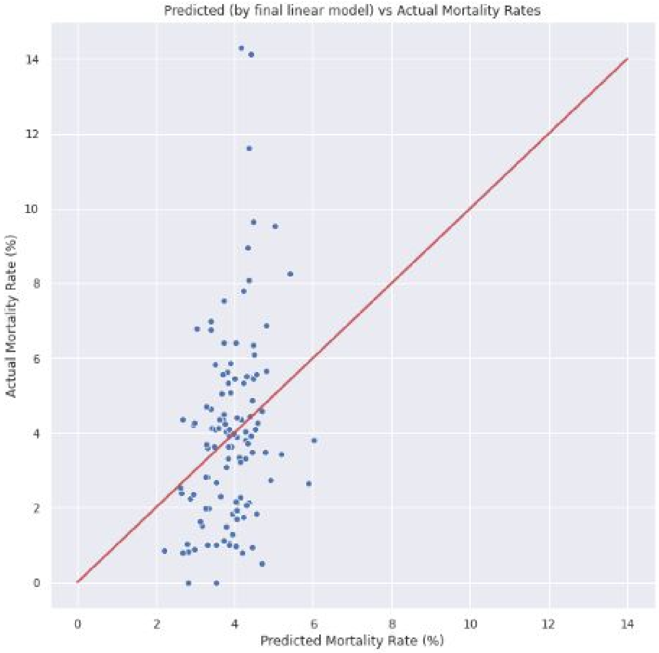
Model Selection and Iteration

After having picked linear regression as our primary tool for analysis and engineering our data, we began feature selection. We first started off by plotting each feature against our output variable to see if there was a correlation. We found many features that looked correlated with the Covid mortality rate at set them aside (features included number of cases, median age, and respiratory mortality rate). We decided to set this batch of features aside because we wanted to understand how the other features behaved as predictors before proving our suspicion that this batch was the best model. After having set aside our suspected best model, we went ahead and began composing other models.

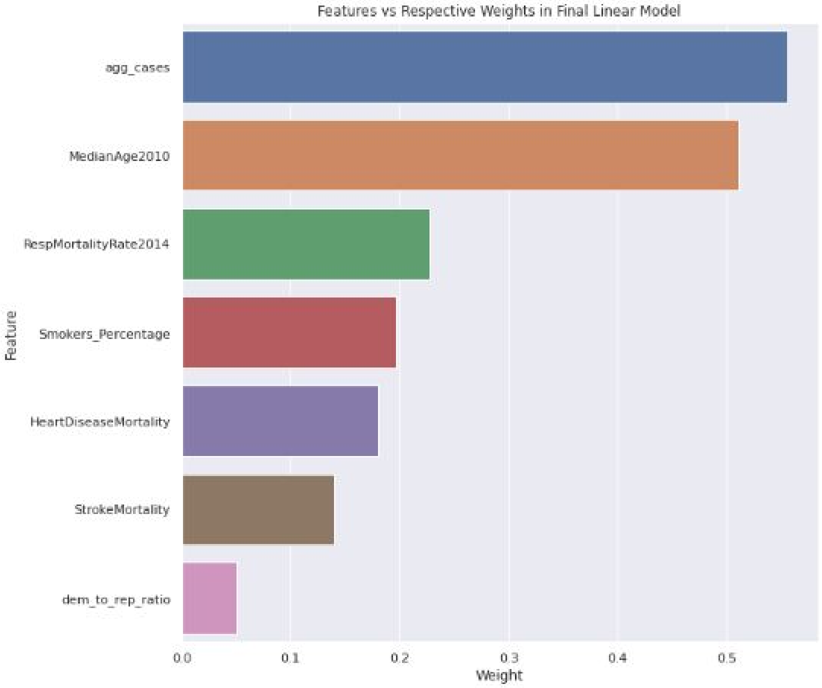
In our first model, we selected only characteristics about each county's population density, gender, and age. For our second model, we started by adding the one-hot-encoding for each state and census region to our first model. After discovering that this model drastically overfit our data, we decided to only add the latitude and longitude of each county's population center to our first model. For our third model, we added important county level healthcare features (like number of hospital beds, medicare enrollment, and number of hospitals) to our second model. For our four model, we added our set aside features to our third model. We found that while decreasing training RMSE, the addition of the set aside features to our already large third model increased cross-validation error. Understanding that we likely had overfit our data, for our fifth model we only used our set aside features and found that this model produced the best cross validation and training error. After having discovered our best model, we went ahead and produced the test error for this model.

Results

After training our numerous models and selecting the one with the lowest cross validation error, we layed out the goodness-of-fit plot below.



Ideally, graphing the actual versus predicted variables would follow the red y = x line, a 1 to 1 relationship, but that is clearly not the case. We will discuss possible reasonings and responses in the following section. Despite the model providing suboptimal predictions, we can still look at the weights that it assigned its features and try to gain insight about their relative importance.



The top three features (available in counties) at predicting Covid mortality percentage per county are:

1. agg_cases: number of Covid cases per county
2. MedianAge2010: median age in of county in 2010
3. RespMortalityRate2014: estimated respiratory mortality rate of county in 2014

where a one standard unit increase in agg_cases, MedianAge2010, and RespMortalityRate2014 results in a 0.55, 0.51, and 0.23 increase in Covid mortality percentage, respectively.

Discussion

A few of the most interesting features that we came across were the Democrat to Repubplican ratio and the size of the population eligible for medicare. We were interested because they seemed that they might reveal information that has been in all of the news headlines politicizing this pandemic. Despite our interest, they did not end up in our final linear model.

We actually had an entire set of features we thought would be useful, but proved ineffective. This was the one hot encodings that we made of county membership in particular states. In fact, a model that included these additional features fared fine enough on the training data, but ballooned our cross validation error. Considering that this one hot encoding introduced 50 features into our model, it shouldn't have come as a surprise that this model was overfitting, and thus deemed ineffective.

One challenge that we faced with our data was that generally the reporting data for both case counts and death counts were skewed toward zero. That is, the distribution was nowhere near uniform. We got stuck in the process of actually choosing what to model. The sheer amount of information available to us essentially gave us choice anxiety.

Some of the limitations of our analysis include that we only considered the mortality rates of the disease in the counties with over one hundred cases. While we believe that this did allow us to move forward with a better model, it is possible that in so doing, it disregarded counties with very small populations, where that many cases comprises a much greater percentage of the population than it would in a dense area. Another limitation that we will explore further as we discuss ethical issues is that our analysis ignored the mortality rates in US territories, despite having the case and death counts for those areas.

One ethical dilemma that we faced with this data was deciding which regions (and consequently, demographics) to include in our analysis. In the end, we choose to forgo analysis of some of the data included in the Johns Hopkins time series information. As mentioned that dataset included statistics on non-county US locations, such as US territories and correctional facilities. In fact, these are some of the places that we are seeing being ravaged by this disease, and are often the ones left most defenseless. We urge future researchers to apply similar methodologies to the specific data of these communities.

Additional data that would allow us to do deeper analysis include statistics on the average household income in each county. We believe it might be highly correlated with the mortality rate, and while some of the other statistics already available (Medicare Eligibility, etc.) might capture a bit of that space, it has the potential to allow for exploration of entirely new relationships. Another set of data that would allow us to do not necessarily deeper, but broader analysis is the same sort of demographic information to which we had access for the counties, but now for the territories. This would allow us to explore the relationship between demographic statistics and covid mortality in these areas that we previously admitted to ignoring.

In studying this problem, there are a few ethical concerns that might arise. For one, the dataset on counties provided by the Yu group includes statistics that could reflect wealth disparities, such as the percentage of the population on Medicare, or the relative ratio between people well served by health care practitioners and those underserved. In studying this problem with this data it could be possible to find a correlation between the mortality rate of this disease and wealth. Frankly, the best way to address might be to dive deeper. Exploring and exposing these inequalities is the first step towards correcting them. Secondly, this is a grim subject. Many are dying. With that, the audience might be personally affected, and it is important to keep this in mind through the routinely cold and sharp presentation of information.

To give a frank evaluation of our model, it was far from perfect at predicting the mortality rate in a given county. While disappointing, this may simply reveal a truth about the underlying data. Looking back to our distribution of mortality rates, we saw that it was skewed heavily towards 0, with only a few rates venturing out higher. The maximum mortality rate reached is still less than 18%, and this is an outlier. Especially since the loss is squared, any model is strongly inclined to overreport lower values. One major limitation of our model is that by nature, it was constrained to linearity. If in fact, the underlying relationship was non-linear, then our current model would be unable to capture it.

One surprising discovery that we made was that in fact the number of cases of COVID-19 in a county carried the most weight in our linear model, while we expected the rates of other diseases in the county to be the top indicators. One interpretation of this surprising result is that the number of cases was capturing when a county had its first infection -- a statistic that we ignored in our analysis. Reasoning about this, it would make sense for the counties that were earlier infected to appear to suffer higher mortality rates, since the disease often takes a multiple weeks to kill.

To anyone conducting future research, we recommend finding data case territory on other US territories and doing a similar type of analysis as was done here with the counties. Additionally, the inclusion of more and varied features could allow for the discovery of new and important factors of the mortality rate of this disease.