Research Article
Volume 12:1, 2021

Journal of
Biometrics & Biostatistics

ISSN: 2155-6180 Open Access

Population Statistics that Contribute to COVID Mortality

Gautham Pavar, Vedaank Tiwari, Namrata Kantamneni*

University of California, Berkeley. 101 Sproul Hall, Berkeley, CA, 94720, United States

Abstract

Our primary objective in this paper was to determine the impact of various factorsaffecting disproportionate COVID mortality rates between counties in the United States. We primarily relied on the CDC's demographics data and the CDC's data on COVID andcomorbidities in US counties. We used these datasets to visualize mortality rates andco-morbidity rates. Exploratory data analysis was then performed to attempt to find trends. Afterwards, we fit our data to a linear regression model to identify the factors that contributed most to the model. The most important features of our model was the proportion of the population that was male and the median age. We found that the median age of the population was a stronger predictor of COVID mortality than presence of comorbidities like diabetes and heart disease. More analysis has yet to be done on the intersection of various comorbidities and median age.

Keywords: Statistics component • Population

Introduction

COVID-19 is an infectious viral respiratory disease that originated in Wuhan, China and spread to become a pandemic. The WHO declared the COVID-19 outbreak a "Public HealthEmergency of International Concern" on January 30th [1]. Self-quarantine orders were put in place worldwide to curb the spread of the disease and have saved many lives in the absence of COVID vaccines and COVID tests. See (Figure 1) for a line graph showing the drastic rise in COVID cases.

However, periods of self-quarantine spanning months pose a serious risk to jobs and the economy. While there are many workers who have the ability to work from home, many others work in jobs where that is not an option (waitressing, plumbing, etc.). Even those who can work from home have lost productivity because of unsuitable work environments [2]. It's estimated that the global economy will lose $2.7 trillion dollars in productivity because of this pandemic [3]. Self-quarantine is by no means a sustainable solution to the COVID pandemic nor future pandemics when vaccines are unavailable.

We want to identify what factors predisposed a county in the US to be more negatively affected due to the COVID pandemic.

This analysis will help policymakers identify which counties are at risk and what policies to enact to prevent these counties from contributing to the spread of future respiratory pandemic diseases. Taking these measures may curb the spread of future pandemics to buy enough time for vaccine development without paying the cost of lost economic productivity.

Description of Methods and Data

In addition to the explanation below, please see our notebook for all the code we ran.

***Address for Correspondence:** Namrata Kantamneni, University of California, Berkeley. 101 Sproul Hall, Berkeley, CA, 94720, United States, and Tel: 7408033660; E-mail: namratakantamneni@berkeley.edu*

Received 19 January 2021; **Accepted** 29 January 2021; **Published** 05 February 2021

Databases

We used data provided by the Yu Group. For the purposes of this assignment, we used data that was last updated on Jan 6th 2021. The dataset has granularity on the county level and is called "county_data_abridged.csv". A full list of the columns described by the dataset is here.

We also obtained state level COVID data from the Center for Systems Science and

Engineering at Johns Hopkins University's COVID-19 dataset. Like the Yu Group data, we used data that was last updated on Jan 6th 2021.

We also obtained provisional COVID death counts from the CDC. This is data last uploaded on Jan 6th 2021 and has a granularity on the county level. The CDC data set provides the presumed deaths due to COVID in counties that have at least 10 presumed deaths due to COVID.

Data cleaning

We made several deletions:

- All columns and rows where null values comprised at least half of the values. Forcibly interpolating the values of more than half of any column or row would compromise the data analysis.
- We also ignored the data regarding when federal guidelines stipulated that gatherings should be limited and data regarding when the foreign travel ban was initiated since the data were the same for all counties in the USA.
- Rows where the STATEFP was more than 56. All 50 states in the USA are retained if we restrict STATEFP to be under 56. Additionally, the data for the American Territories

Were much more spotty (more null values) than the data for the US states.

We interpolated the values of remaining null values by inserting the mean of each column.

Exploratory data analysis

We performed Principal Component Analysis (PCA) on our dataset. We normalized and subtracted the means of our data's columns before starting PCA. We found the first principal

Component accounts for about 60% percent of the variation of the dataset and is primarily determined by population (Figures 2 and 3):

From Figures 3 and 4, we interpreted that the first principal component

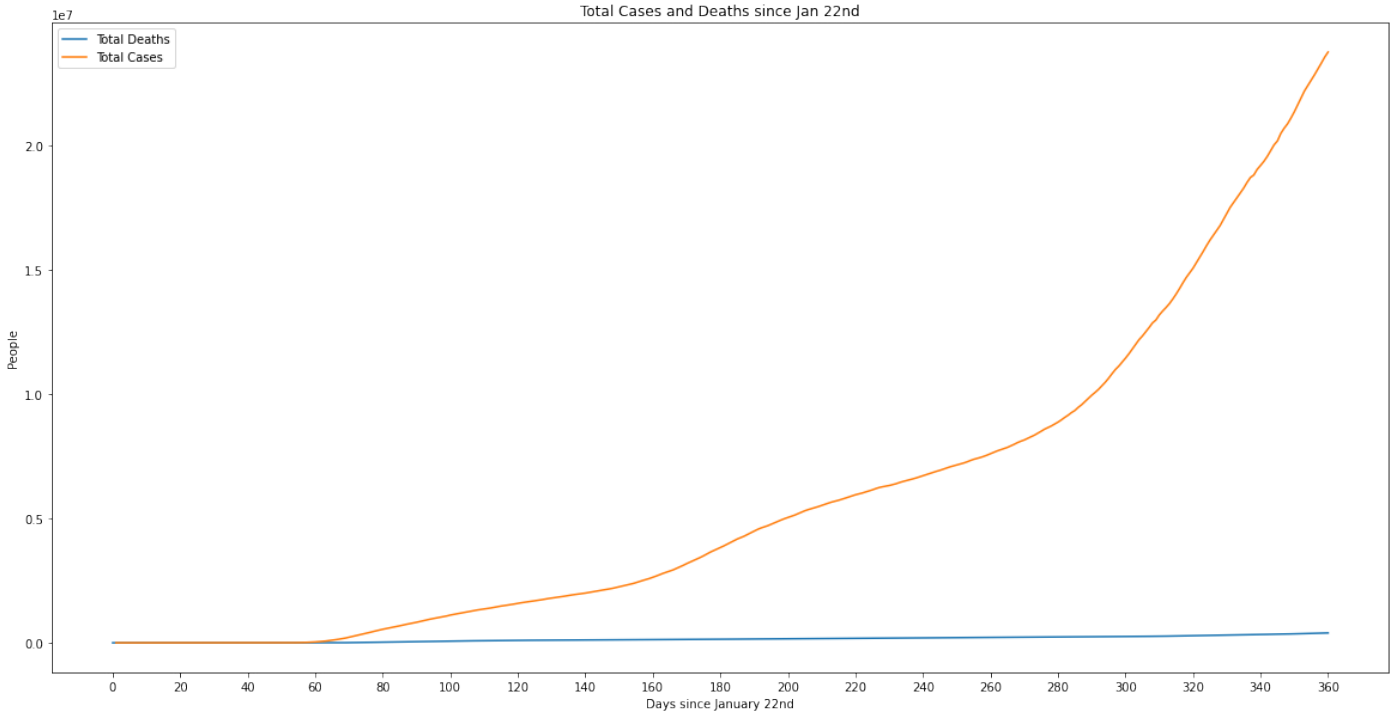

Figure 1. Rise in COVID cases since the first case in the US January 22nd.

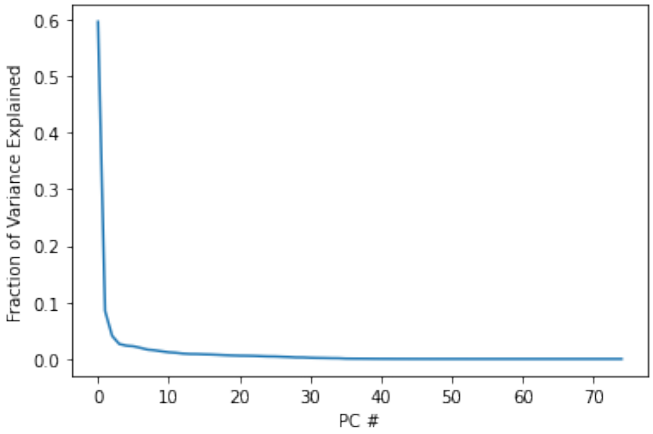

Figure 2. Scree plot of PCA performed with all population columns.

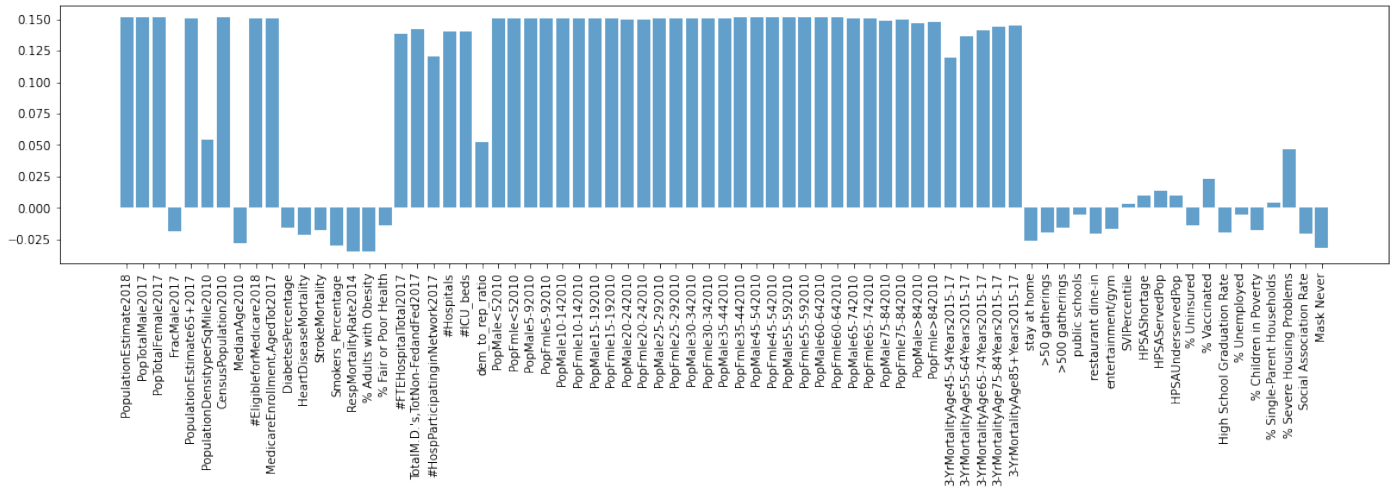

Figure 3. Variance of each column of the dataset accounted for by the principal component.

describes population in our dataset and the second principal component describes comorbidities (Heart Disease, Diabetes, etc.) in our dataset. Given that many of the columns that are tied to the first principal component are all

Population-based (they have format "Pop…") we decided to drop all the population based columns except for the first one ("Population Estimate 2018"). By dropping around 30 of these columns, we were able to get this scree plot (Figure 5):

In this scree plot (Figure 5), the first principal component accounts for around 33% of the variation in the data and is still based primarily on population (Figure 6). By reducing our dataset's reliance on the first principal component we hope that our model will be able to better analyze the effect of non-population factors on mortality rate.

We chose to focus on examining comorbidities since the CDC says that health conditions affect severity of infection. We quantify the severity of COVID-19 by looking at mortality rate since Case Fatality Rate is not readily

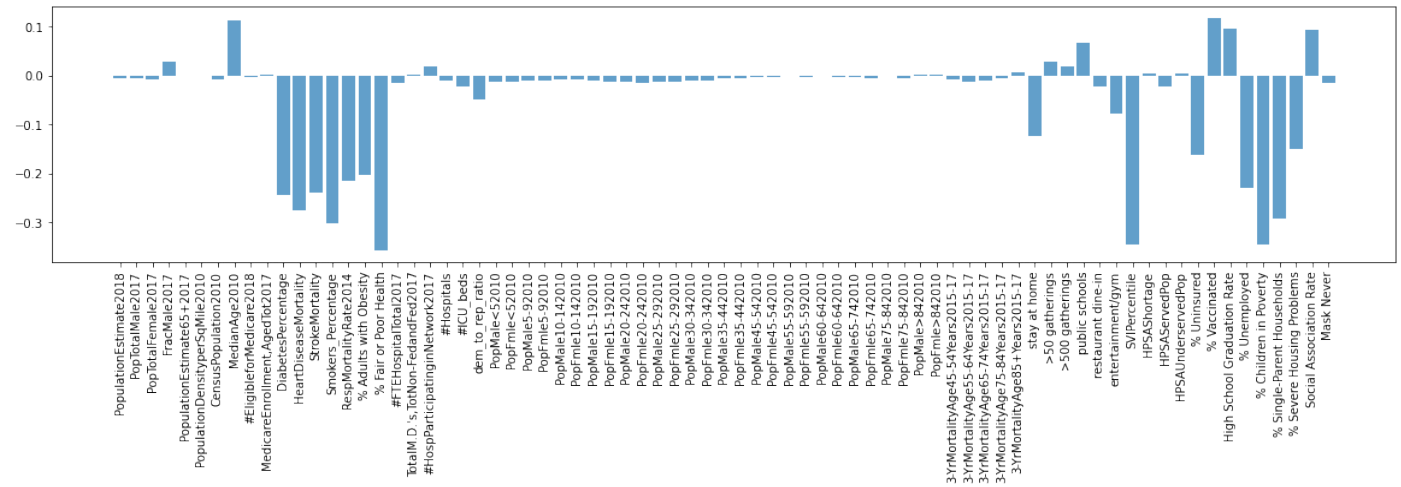

Figure 4. Variance of each column of the dataset accounted for by the second principal component.

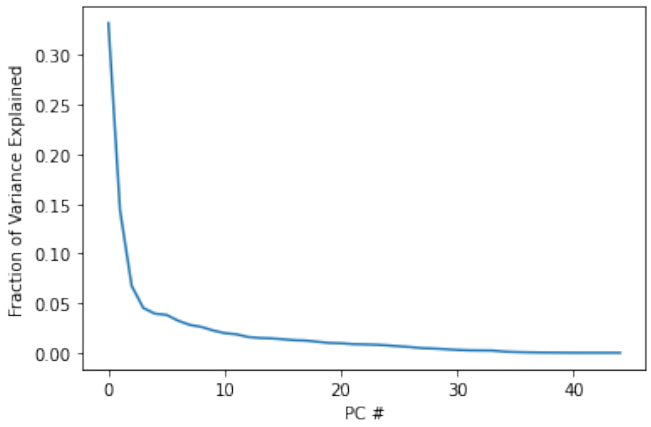

Figure 5. Scree plot of PCA after deleting most of the population related columns.

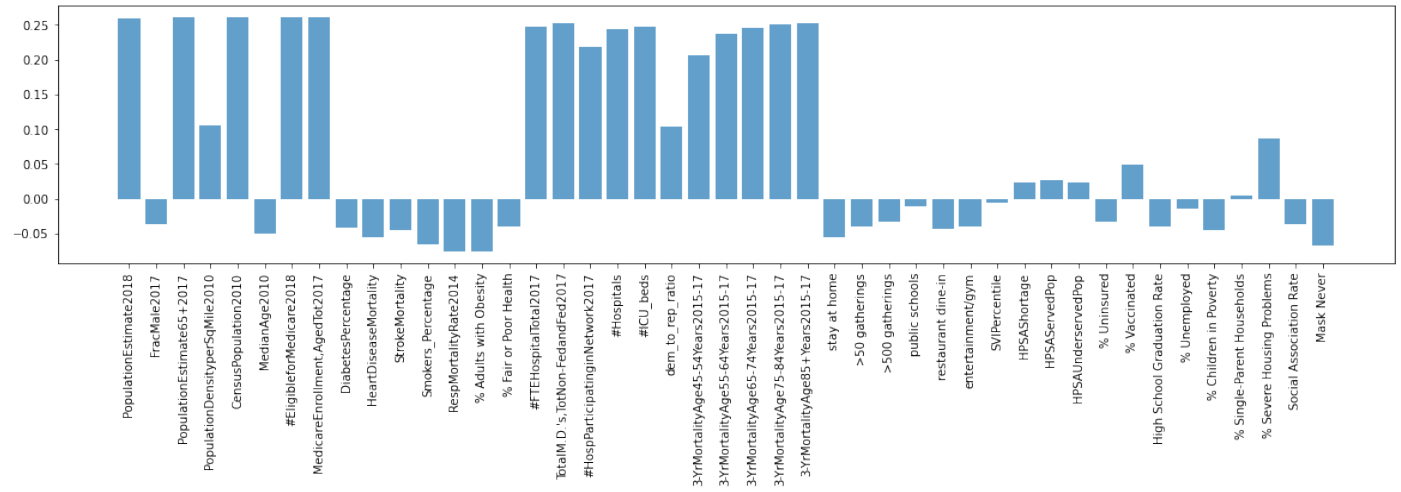

Figure 6. Variance of each column of the dataset accounted for by the first principal component, after feature reduction.

available (due to COVID testing inadequacies). The CDC describes several factors that increase risk of severe COVID infection such as age, heart disease, smoking, etc.

Based on the CDC guidelines, we thought that the Population over 65 column would be a significant factor in our dataset and be significantly different from the general population column of our dataset ("Population Estimate 2010"); however, it turns out that the two population

Columns are similar enough to be described by the same Primary component (see Figure 2). We ended up not using the population over 65 columns even though we had thought it would be a crucial portion of our analyses.

Regardless, population and comorbidities were two of the most useful and interesting features of our model since our dataset's first two principal components described population and comorbidities.

Methods

Initially, we wanted to aggregate all the county level descriptive data provided by the Yu Group by state so that we can easily compare them to the state-level COVID mortality data provided by Johns Hopkins. However, we were not able to properly compute a weighted average (this was needed to obtain a state's Male fraction of the population from each county's Male

Fraction of the population and each county's total population, for example) so this fell through. We were not able to train a model since we could not calculate state-level data.

However, we later found the CDC's dataset that contained county-level COVID mortality, this could more readily be combined with the county level descriptive data from the Yu Group. After performing the data cleaning methods mentioned in the above section, we decided to fit our model to a Linear Model using ridge regression. We chose ridge regression since we have a lot of potential features (over 60). We used a linear model since we were modeling mortality rate, which is a quantitative variable.

Using the our PCA analysis we found that a number of features had a high effect on the variance when it comes to mortality rates including the male proportion of the population, the median age, the percent of the population with diabetes, the heart disease mortality, the stroke mortality, the percentage of smokers, the respiratory disease mortality, the length of the stay at home order, the length of the ban on >50 gatherings, the length of the ban of >500 gatherings, time public school had been closes, as well as restaurants, entertainment, and gyms. We sampled 20% of this data randomly for our training data (X).

We pulled data from the CDC about the number of deaths in each county, and divided this by the population of the county, giving us the mortality rate for both COVID and all causes. We used this as our training data (y)

We trained 2 separate models, one which included the CDC's covid mortality rate for each county and the other which used the overall mortality rate for each county. We did this because we wanted to see how accurately our model would be able to predict overall deaths, especially considering the fact that COVID may indirectly be causing many deaths (perhaps people are not going to hospitals as often, perhaps they run out of money to buy their everyday drugs because they lost their job) and that many COVID deaths may be not counted properly.

We analyzed each model using the RMSE, and it wasn't a surprise to us that the model predicting COVID deaths had a lower error than our model that predicted all deaths. We realized that our errors were extremely small numbers, so we went back and multiplied the CDC mortality rate figures by 100,000, effectively giving us a mortality rate per 100,000. That made the errors much easier to interpret.

Results, Interpretation and Discussion

The two most important features for our model were:

- The proportion of the population that was male
- The median age

These two features had the biggest weights in our models.

Analysis

Looking at the results, we can clearly see that the comorbidities were not as predictive of the overall mortality rate as we expected. Factors like heart disease, respiratory disease, smoking, etc. had relatively small weights in our linear regression equation. We did not expect such

Results. It's not clear why the proportion of the population that is male would affect COVID mortality. However, other researchers have noted that males tend to get more severe COVID infections than women.

The one thing that was expected was the dependence on median age. That is part of the CDC guidance and is also a significant feature in the model.

In the future, it would be interesting to create a model based on the same principalcomponents, subtracting out the COVID related deaths just to see how accurate our model is at predicting non-COVID related deaths (giving us a base death rate). That would give us a good baseline to compare with our models that include the COVID deaths to see how good our models really are at predicting mortality. Of course, we would have to use pre-COVID data to make sure that there would be no possible COVID-related deaths influencing the model.

Data Limitations

The CDC dataset is limited in that it relies on passive surveillance. The CDC has labeled their death counts due to COVID as 'provisionary' since the most recent death certificates have probably not been processed and coded by the National Center for Health Statistics.

We have inner joined the CDC dataset with the Yu Group data set on the county FIPS. Since the CDC dataset only describes the COVID related deaths of those counties that had at least 10 COVID related deaths, our model is not able to predict the factors that contribute to having an extremely low number of COVID related deaths in a county. However, our model does have the ability to characterize the factors that contribute to having a high number of COVID related deaths in a county. Of the almost 3000 counties in the United States, our data analysis only looks at the 366 counties that had the most COVID related deaths.

Our model would be useful for high-risk counties to lower the possible deaths in the next pandemic; however, it would not be useful for low-risk counties to learn how they can

Completely mitigate the risk of pandemic related deaths.

There are some future directions for the project. If we can get the CDC's COVID-related death data for all counties in the US, then we can understand what factors contribute to having no COVID-related deaths in counties. Ideally, we would train a model after the pandemic is

Resolved based on data that has been verified to be accurate (instead of just being 'provisional'). We could also use data collected by online software like the How We Feel app to understand COVID infection rates, but that would be dependent on the availability and accuracy of COVID tests.

Ethics

From an ethical standpoint, making a model based on limited data could be a disservice to policymakers and their constituents who are affected by our model. For example, our model may support incorrect policy decisions at the cost of the county's constituents without reducing the risk of pandemic-related deaths that much. Regardless, we believe that producing an imperfect model would still benefit more people than it would harm, so our actions are approved of from a utilitarian standpoint.

Another issue with our project is one that is common with public health interventions.

According to social contract theory, individuals give up rights (like certain freedoms) in order to receive the protection of a society. Public health interventions upend the old social contract and replace it with one where people are required to give up more freedoms in order to get more protection. For example, self-quarantine orders and contact tracing may require that people give up autonomy and privacy in order for everyone to be safer. From a rights framework, it's unjust if people are forced to give up these rights during an emergency. One solution, which our model may contribute to, is proposing potential public health interventions well in advance of the next pandemic. Individuals in a society will have the time to understand what rights they need to give up and may even be able to propose alternate solutions that take away less rights but have the same level of protection. This is necessary so that everyone will follow the amended and stricter social contracts that are required during public health emergencies.

References

1. Coronavirus Disease (COVID-19). "Events as they happen" (2020).
2. Gorlick Adam. "The productivity pitfalls of working from home in the age of COVID-19. "*The Productivity Pitfalls of Working from Home in the Age of COVID-19; Stanford University* (2020).
3. Tom Orlik, Jamie Rush, Maeva Cousin and Jinshan Hong. "Coronavirus Could Cost the Global Economy $2.7 Trillion. Here's How." *Bloomberg* (2020).

How to cite this article: Gautham Pavar, Vedaank Tiwari, Namrata Kantamneni. Population Statistics that Contribute to COVID Mortality. *J Biom Biostat* 12 (2021): 448.