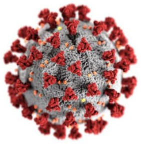
ORIE 3120 Data Visualization Project

An Analysis of current United States COVID-19 crisis data



Caleb Berman

Introduction

As COVID-19 has arguably become one of the largest Public Health crises the world has faced in the last 100 years, our group decided to leverage the Data visualization tools we'd gained so far in ORIE 3120 to better understand how this virus has affected the United States and its people. We hope by visually analyzing the current data we have of the present that we will be able to further make some meaningful predictions of what the effects of COVID-19 on the United States will look like in the Future. As there is an immense amount of data regarding Coronavirus, we focused the scope of this project down to the following guiding questions:

- What factors are most responsible for the change in mortality rates in each state?
- What American populations are most at risk for severe COVID-19 related illness?
- How will the number of COVID-19 cases in these States evolve over time?

These were chosen as we thought information on how COVID-19 varies across space, time, and population demographics could be helpful in providing insight towards future policy actions, as well as creating a robust infrastructure for areas and populations that are at a particularly high risk for future viral infections.

Dataset

The metadata we chose to examine was compiled for the UNCOVER COVID-19 challenge, curated by the Roche Data Science Coalition located on Kaggle. It's made up of a diverse collection of over 20 COVID-19 related datasets. We chose to focus only on United States state-level datasets within this collection. Through inner joining tables by state, we assembled the following data from the following sources:

Data	Source
Hospital available beds by State	Harvard Global Health Institute
Positive tests and deaths by state and time	New York Times
State Prevalence of Respiratory issues by state	Behavioral Risk Factor Surveillance System
Population Demographics	United States Census
Geographic mobility	Google

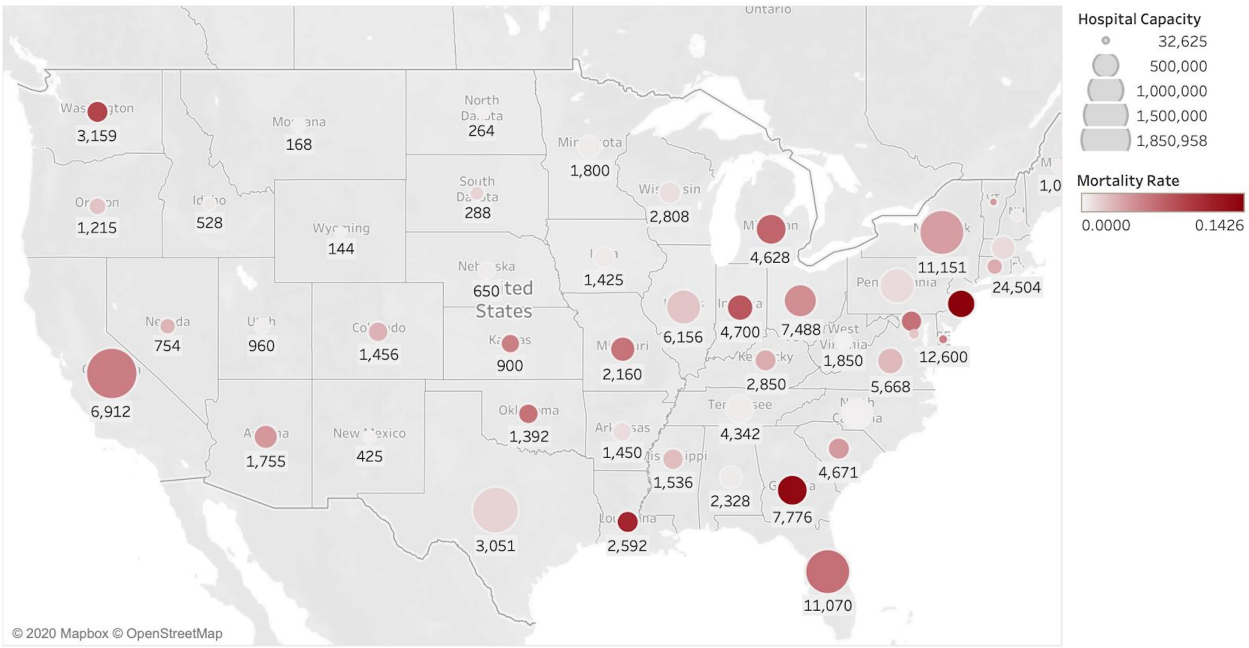
Through combining infection data, available hospital capacity data, and population data we sought to figure out how COVID-19 was impacting the United States population and the factors that influence that impact.

Geospatial analysis of States public health performance during the COVID-19 Crisis

While Coronavirus continues to spread nationwide, we wanted to better understand how each State is performing against the virus. To better understand each State's performance, we compared mortality rates for each State. Mortality Rate was calculated by the number of Total Deaths divided by the number of Total Confirmed Cases in that state. In order to answer why each State exhibits a different response to the COVID-19, we then compared the mortality rate to the State Density and the Hospital Capacity.

Mortality Rate vs Hospital Capacity by State

Labled with 2020 State Density.



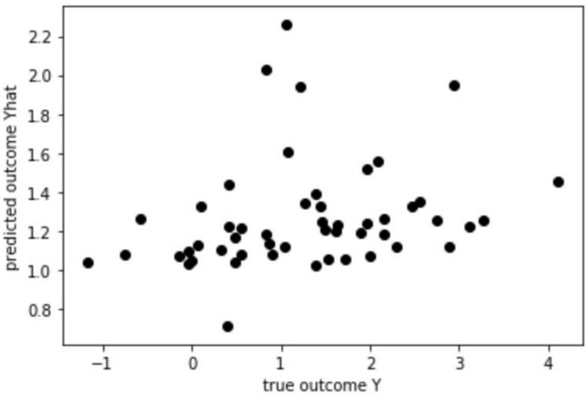
Map based on Longitude (generated) and Latitude (generated). Color shows sum of Mortality Rate . Size shows sum of Hospital capacity. The marks are labeled by sum of 2020 State Density. Details are shown for State.

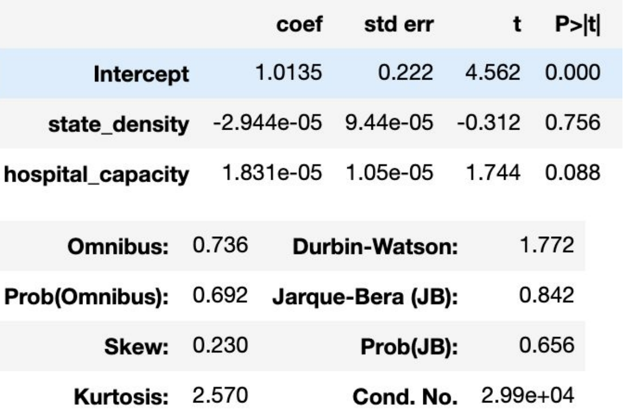
In order to better understand which factors affect the mortality rates in each state, we compared the neighboring states such as Georgia and Florida, Washington and California, and New York and Delaware. Looking at Georgia and Florida, two neighboring states, we can observe that Florida, whose state density is greater and hospital capacity is larger than Georgia, in fact, exhibits a smaller mortality rate. The same trend appears to other pairs of neighboring states. Therefore, we can assume that the hospital capacity is more correlated to mortality rate than state density.

According to the geographical visualization, we can make a hypothesis that the smaller hospital capacity is more responsible for higher mortality rates. To test the hypothesis, we utilized linear regression to better understand how each factor affects the mortality rates from COVID-19. We believe this modelling would help each state to implement a better solution to fight against the virus.

To further analyze how each state's hospital capacity and state density affect the change in mortality rates, we designed a linear regression model to see the correlations.

Linear Regression model between the change in mortality rates and hospital capacity, state density (R^2=0.258)





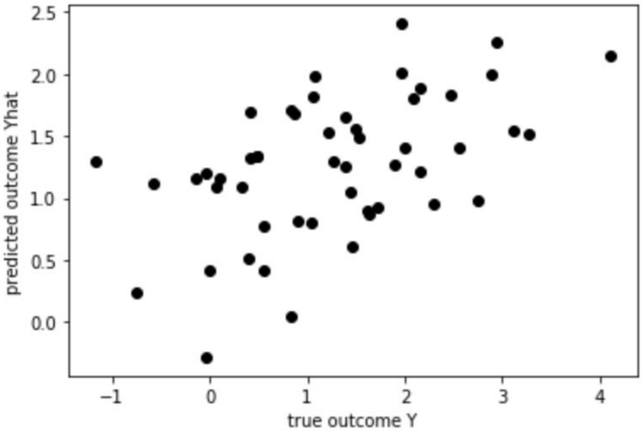
	coef	std err	t	P>|t|
Intercept	1.0135	0.222	4.562	0.000
state_density	-2.944e-05	9.44e-05	-0.312	0.756
hospital_capacity	1.831e-05	1.05e-05	1.744	0.088

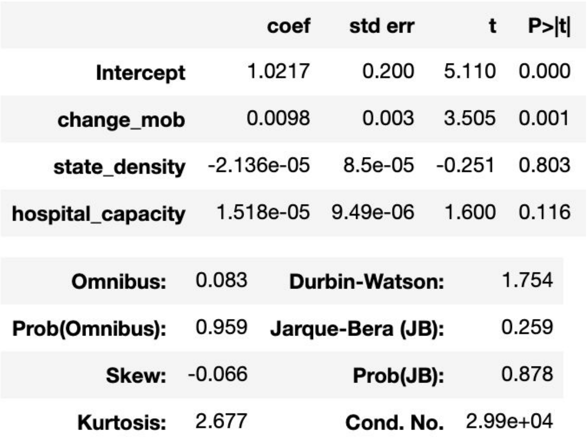
Omnibus:	0.736	**Durbin-Watson:**	1.772
Prob(Omnibus):	0.692	**Jarque-Bera (JB):**	0.842
Skew:	0.230	**Prob(JB):**	0.656
Kurtosis:	2.570	**Cond. No.**	2.99e+04

However, according to the graph, both state density and hospital capacity do not seem to strongly predict the mortality rate change as the linear model has an R^2 value of 0.258. Moreover, state density's p value is 0.756 which fails to reject the null hypothesis. On the other hand, hospital capacity has the p value of 0.088, which is slightly greater than 0.05(the default confidence interval from python is 95%), so there is more confidence in some relationship between the hospital capacity and the change in mortality rates, but definitely not strong.

Since both factors do not seem to influence the change in mortality rates as strongly as expected, we decided to add another geographic factor, change in mobility. This was calculated by taking the sum of the change in mobility with respect to baseline as created by google as of April 11th.

Linear Regression model between the change in mortality rates and [change in mobility, hospital capacity, state density](R^2=0.258)





	coef	std err	t	P>|t|
Intercept	1.0217	0.200	5.110	0.000
change_mob	0.0098	0.003	3.505	0.001
state_density	-2.136e-05	8.5e-05	-0.251	0.803
hospital_capacity	1.518e-05	9.49e-06	1.600	0.116

Omnibus:	0.083	**Durbin-Watson:**	1.754
Prob(Omnibus):	0.959	**Jarque-Bera (JB):**	0.259
Skew:	-0.066	**Prob(JB):**	0.878
Kurtosis:	2.677	**Cond. No.**	2.99e+04

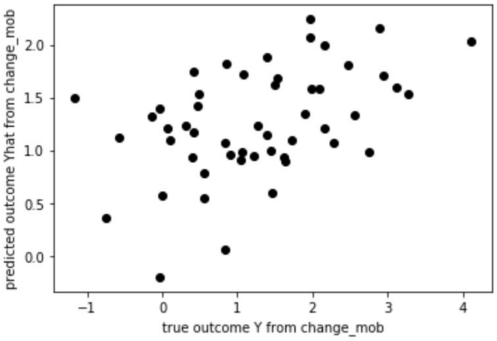
The addition of change in mobility as another independent variable improved the fit of our linear model. The coefficient of change in mobility is 0.0098, which is much greater than that of other two variables. Moreover, p values of change in mobility is 0.001 which is less than 0.05; therefore, we can conclude there is likely a strong relationship between the change in mobility and the COVID-19 mortality rate.

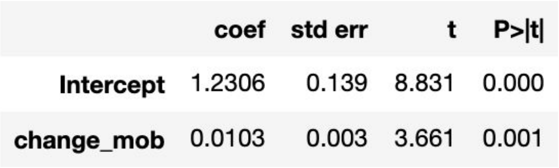
Unfortunately, the condition number is very high as 2.99e+04, indicating that there might be some multicollinearity between the variables. Therefore, we conducted the collinearity test of each variable to determine whether or not it should be deleted from the model due to overly high collinearity.

Below is a collinearity test with respect to the change in mortality rates for each variable depicts a graph and has the coefficient and p values as below.

Collinearity model

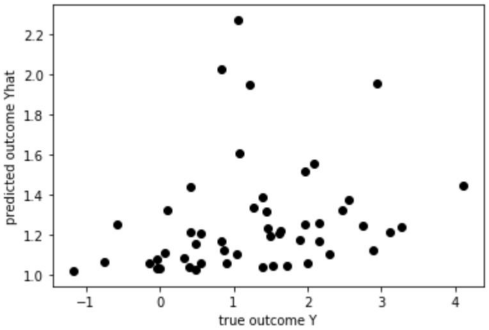
1. *Change in mobility (Condition Number = 49.7)*

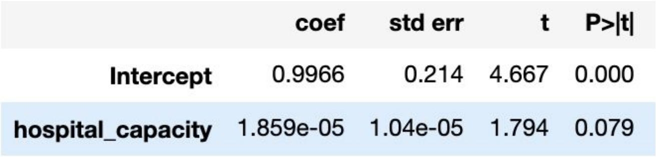




	coef	std err	t	P>|t|
Intercept	1.2306	0.139	8.831	0.000
change_mob	0.0103	0.003	3.661	0.001

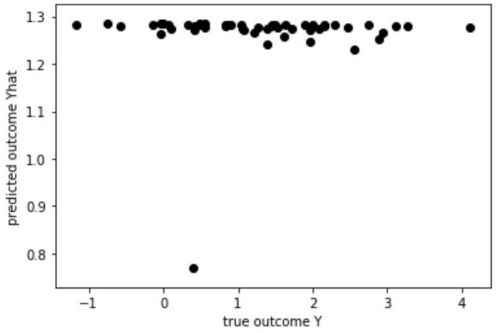
2. *Hospital Capacity (Condition Number = 2.90e+04)*

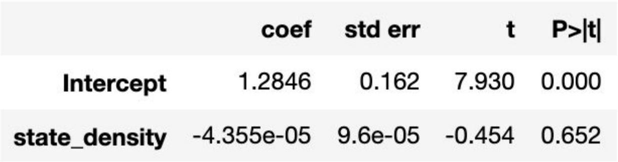




	coef	std err	t	P>|t|
Intercept	0.9966	0.214	4.667	0.000
hospital_capacity	1.859e-05	1.04e-05	1.794	0.079

3. *State Density (Condition Number = 1.75e+03)*

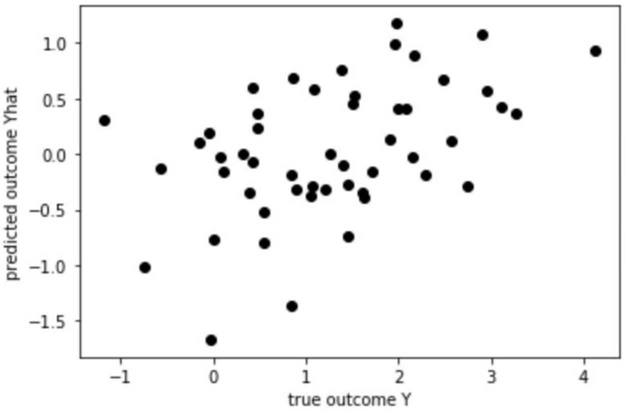


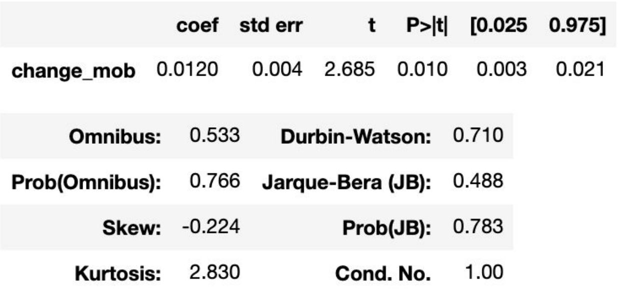


	coef	std err	t	P>|t|
Intercept	1.2846	0.162	7.930	0.000
state_density	-4.355e-05	9.6e-05	-0.454	0.652

Therefore, we can conclude that both hospital capacity and state density have high multicollinearity. After deleting the two variables from our initial regression, and making a new linear regression model solely between the change in mobility and the change in mortality rate, we created the graph and statistical values below.

Linear regression model between the change in mortality rates and the change in mobility





	coef	std err	t	P>|t|	[0.025	0.975]
change_mob	0.0120	0.004	2.685	0.010	0.003	0.021

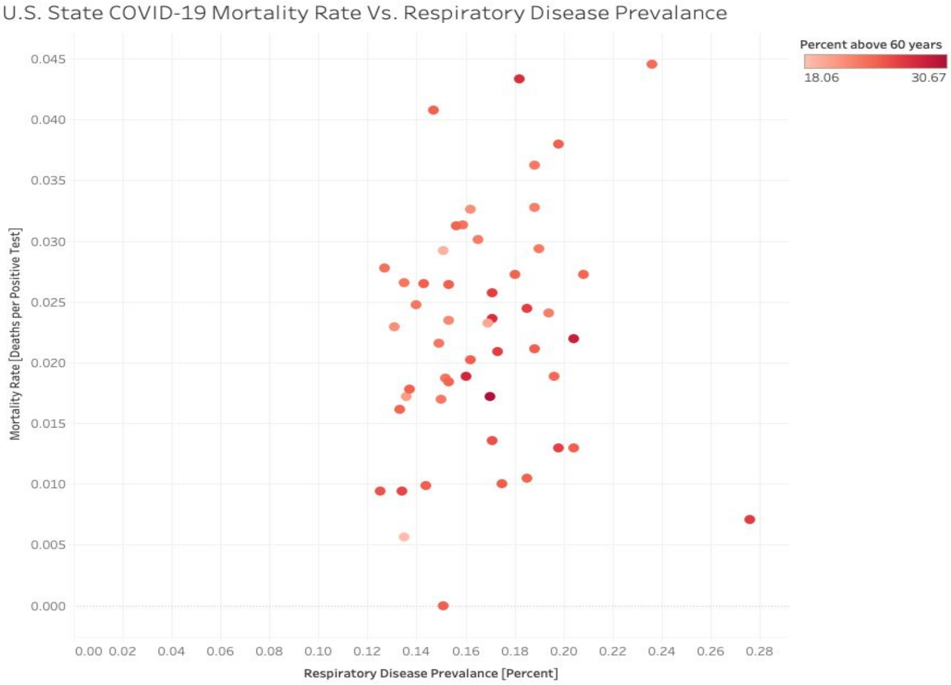
Omnibus:	0.533	**Durbin-Watson:**	0.710
Prob(Omnibus):	0.766	**Jarque-Bera (JB):**	0.488
Skew:	-0.224	**Prob(JB):**	0.783
Kurtosis:	2.830	**Cond. No.**	1.00

In the end, we can conclude that change in mobility, in other words the amount of people conducting the social distancing, poses great impact on the change in mortality rates from COVID-19.

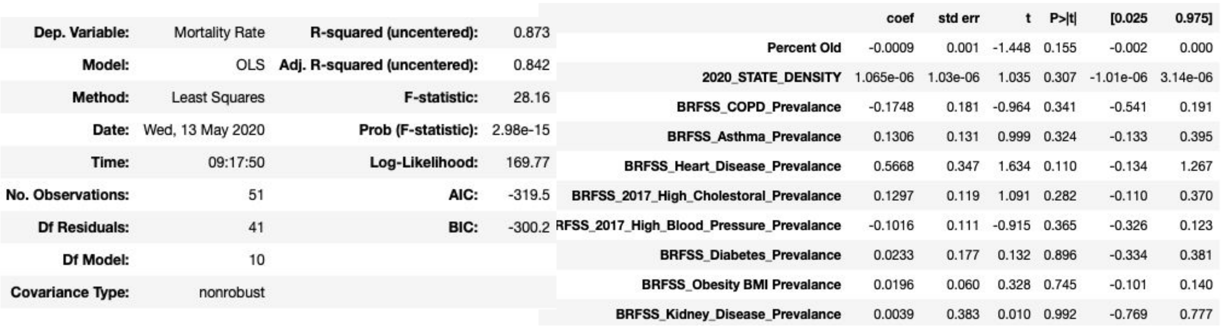
Possible underlying causes of COVID-19 case severity

Since COVID-19 produces pneumonia-like symptoms in patients, negatively impacting their respiratory system we wanted to better understand how populations with pre existing respiratory problems were affected by the virus. We were also interested in how the virus affects those of more advanced age as they are likely to have weaker immune systems. Below is a scatter-plot of Mortality rate as a function of Respiratory Disease Prevalence for all 50 of the United States as of April 4th. Here respiratory disease prevalence was defined to be the sum of the prevalence of asthma and Chronic obstructive pulmonary disease, two common respiratory affiliations that occur on similar scales. The color of the points represents the percentage of those in a population above the age of 60 (as of 2010 Census).

As you can see, the plot suggests some form of linear relationship between coronavirus mortality and prevalence of respiratory illness in a state's population. However contrary to preconceived notion, the proportion of elderly in a population had little to do with current mortality rates as it's very randomly distributed across the plot.



While this visualization hinted at some sort of linear relationship between respiratory disease and COVID-19 mortality we wanted to dig deeper and find out what pre-existing medical conditions might put a certain group at further risk of dying from COVID-19. In order to further understand which populations are most vulnerable to COVID-19 mathematically, we began by fitting a linear model of variables of which we thought would be both of interest and independent from one another. The features we chose to fit our linear regression to are as follows: Population percentage above 60, State Density as of 2020, COPD prevalence rate, Asthma prevalence rate, Heart Disease Prevalence rate, High Cholesterol Prevalence rate, High Blood Pressure Prevalence rate, Diabetes Prevalence rate, Obesity prevalence (according to BMI) and Kidney Disease prevalence. The linear regression gave the following statistical results:



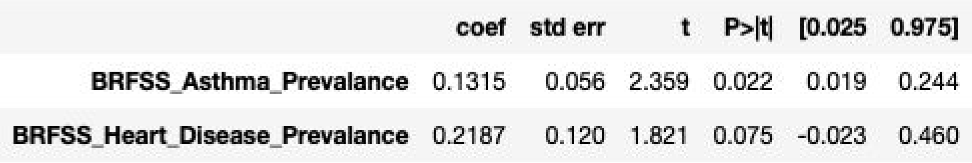
Dep. Variable:	Mortality Rate	**R-squared (uncentered):**	0.873
Model:	OLS	**Adj. R-squared (uncentered):**	0.842
Method:	Least Squares	**F-statistic:**	28.16
Date:	Wed, 13 May 2020	**Prob (F-statistic):**	2.98e-15
Time:	09:17:50	**Log-Likelihood:**	169.77
No. Observations:	51	**AIC:**	-319.5
Df Residuals:	41	**BIC:**	-300.2
Df Model:	10		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Percent Old	-0.0009	0.001	-1.448	0.155	-0.002	0.000
2020_STATE_DENSITY	1.065e-06	1.03e-06	1.035	0.307	-1.01e-06	3.14e-06
BRFSS_COPD_Prevalance	-0.1748	0.181	-0.964	0.341	-0.541	0.191
BRFSS_Asthma_Prevalance	0.1306	0.131	0.999	0.324	-0.133	0.395
BRFSS_Heart_Disease_Prevalance	0.5668	0.347	1.634	0.110	-0.134	1.267
BRFSS_2017_High_Cholestoral_Prevalance	0.1297	0.119	1.091	0.282	-0.110	0.370
RFSS_2017_High_Blood_Pressure_Prevalance	-0.1016	0.111	-0.915	0.365	-0.326	0.123
BRFSS_Diabetes_Prevalance	0.0233	0.177	0.132	0.896	-0.334	0.381
BRFSS_Obesity BMI Prevalance	0.0196	0.060	0.328	0.745	-0.101	0.140
BRFSS_Kidney_Disease_Prevalance	0.0039	0.383	0.010	0.992	-0.769	0.777

The first thing we can note from this fit is that the percent of people above 60 has much less of an effect on COVID-19 mortality as compared to the prevalence of these common pre-existing conditions within a population as evidenced by its fit coefficient being several orders of magnitude lower.

Because our goal is to find the factors most related to dying from COVID-19, we performed some model selection on the available data in order to reduce the amount of non-statistically significant features. We implemented an algorithm to select the features that would minimize the ordinary least squares regression's Aikake Information Criterion (AIC) value and then modified the linear fit to only include those features.

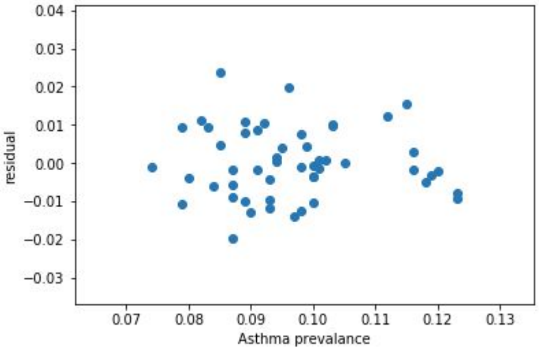
Due to the fact that by looking at state-wide data we only had access to 51 distinct data points, we did not split our data into a testing set and a training set for model selection as a training set would not be large enough to represent the dataset as a whole and would be prone to overfitting. This was deemed an acceptable choice for 2 reasons. Firstly, our initial linear regression has prob(F-statistic) on the order of magnitude of 10^{-15}. This gives us reason to believe it is unlikely that the initial data is just due to pure random chance. Secondly, minimizing the AIC will still inform us of the most significant features and give us an accurate predictive model even if the statistics on those features are less accurate.

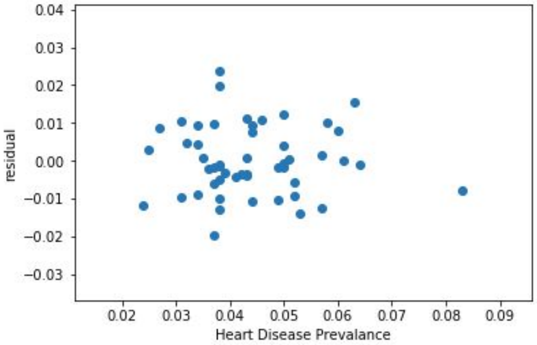


	coef	std err	t	P>|t|	[0.025	0.975]
BRFSS_Asthma_Prevalance	0.1315	0.056	2.359	0.022	0.019	0.244
BRFSS_Heart_Disease_Prevalance	0.2187	0.120	1.821	0.075	-0.023	0.460

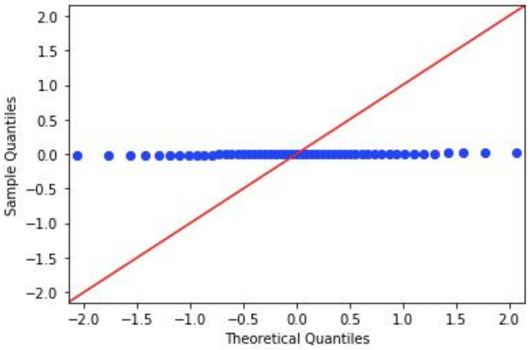
The minimization of the AIC value selected Asthma Prevalence and Heart Disease Prevalence as the two most significant predictors of COVID-19 mortality. This new model had a condition number of 10.9 so Collinearity was of little concern.

In order to confirm the linear nature of these predictors the plots of the predictors against their residuals are found below and due to their seemingly random distribution we concluded that there's no evidence to suggest nonlinearity.





The QQ plot of the data on the right additionally suggests that the data is normally distributed, albeit very light tailed. Giving further evidence that we were justified in generating this linear model.



Thus the two most significant health factors that predicted COVID-19 mortality were statewide Asthma prevalence and Heart Disease prevalence. We concluded that those people with Asthma and Heart Disease are among the most at risk from dying from COVID-19 and recommend that they take extra precautions.

Reference

URL: **https://www.kaggle.com/roche-data-science-coalition/uncover**

API Command: kaggle datasets download -d roche-data-science-coalition/uncover