Predicting and Preventing the Spread of Coronavirus

By Victoria Nguyen, Gathenji Njoroge, and Andrea Seet

CS 200A Final Project

Abstract

Since the first case of coronavirus disease 2019 (COVID-19) was confirmed in the United States on January 21, 2020, over one million people have developed confirmed cases of COVID-19 according to the National Public Radio.[1] On April 12, the United States became the nation with the most deaths globally and, with these high mortality rates, there is a significant need for determining what pre-existing conditions and demographic characteristics can predict the cases of COVID-19. Through analyzing the existing data, predictions can be used as a tool to further mitigate the spread of the virus, especially in states that are most vulnerable. According to The Washington Post, there is evidence that the responses of everyday Americans to the coronavirus differ based on political affiliations, though these differences are smaller in states with higher rates of cases.[8] As Republicans have consistently reported less concern about the virus and have taken less precautionary measures to control the spread, we aimed to explore the potential association between the states' political party and cases of coronavirus. Our analyses aim to determine whether a state's overall political leanings are associated with the number of cases of coronavirus it has; identify correlations between features used in our models; utilize PCA for dimensionality reduction and feature extraction; and, via OLS linear regression, identify which features are statistically significant predictors of case counts by state.

Introduction

By the end of March 2020, coronavirus disease 2019 (COVID-19) spread through all 50 states in the United States. The United States currently has the greatest number of confirmed cases and confirmed deaths in the world. Efforts to control this growing crisis involve analyzing the factors that have increased community transmission of this highly contagious respiratory virus.

While researchers are still learning about how COVID-19 spreads, current research indicates that the degree to which the disease impacts communities is increased in populations with greater health risks, especially older adults and those with underlying medical conditions.[2] These trends are apparent, as COVID-19 was first observed rapidly spreading through nursing and long-term care facilities.[2] The research also indicates that areas with high population density have a higher risk of COVID-19 spread.[2] Clinical indicators in a case-control study in Singapore identified individuals at greater risk of developing the disease were more likely to have higher temperature, higher respiratory rate, and gastrointestinal symptoms.[8] Other notable risk factors are older age, being male, and comorbidities such as respiratory diseases, cancer, hypertension, and arteriosclerotic diseases.[8] The evidence also supports that differences in partisanship among the states have affected the rates of transmission.[1]

Because of these trends, we are interested in examining the associations between comorbidities, age, sex, population estimates, partisanship, and the number of confirmed cases of COVID-19 at the state level by April 18, 2020.

By evaluating the variables that best predict the states with the highest cases, there could be benefits of preventing the spread of the coronavirus through public health efforts of addressing the upstream determinants of unhealthy outcomes. For example, if there is evidence that certain pre-existing conditions predict high coronavirus cases, these diseases can be targeted in the long-run in hospital systems as coronavirus is predicted to affect states across the nation as well as internationally.

Description of Data

We analyzed two datasets: confirmed time series cases and abridged data. The time series data for each U.S. state and territory included the city and the number of confirmed coronavirus cases over time starting from January 22 to April 18. The abridged dataset provided information including the state population estimates of 2018, population estimates by gender, population density, eligibility for Medicare, diabetes percentage, heart disease mortality, stroke mortality, smokers percentage, respiratory mortality rate in 2014, Democrat to Republican ratio, and age and sex specific population rates for 3,244 counties across the U.S.

Description of Methods

First, we cleaned the data to make sure that there were no missing values or outliers that needed to be adjusted or accounted for. We aggregated and collapsed the county-level demographic data by states because we were interested in how measures taken by different states have affected their total number of COVID-19 cases. We removed American territories, the Diamond Princess cruise ship, and Alaska, and Hawaii since we did not have demographic information on these places. Therefore, our final sample only included the 48 “lower” contiguous U.S. states. The sum of the county-level data was taken for population size, population by sex and age group, number of hospitals, and number of ICU beds and saved as state-level totals. The mean was taken on all comorbidities we were interested in analyzing, the percentage of smokers in each county, and the Democrat to Republican ratios. We aggregated the values this way because these variables were presented as proportions rather than absolute numbers. These were also saved as state-level totals.

We conducted descriptive statistics of the states with the highest number of cases in order to determine which features we might want to analyze as potential predictors of the high case total. In addition to looking at the data, we also used the information from background research to inform our models. We conducted exploratory data analysis and created data visualizations such as scatterplots in order to determine if features were correlated with each other for dimensionality reduction, as including many variables could lead to overfitting. We also created scatterplots of total cases of the top ten states versus hospitals/ICU beds and total cases and the number of intensive care unit beds by top ten states. This would inform whether we would want to put the variables of hospitals/ICU beds and the number of intensive care unit beds into our

principal components analysis. As there has been significant evidence from the CDC that people who have pre-existing conditions are at a higher risk for severe COVID-19, we created scatterplots of heart disease mortality versus stroke mortality to see if there was an association between the pre-existing conditions as we do not want to include variables that are highly linearly related.[3]

In order to best predict states with high coronavirus cases using pre-existing conditions, we conducted principal components analysis to 1) determine which of the variables of diabetes percentage, heart disease mortality, stroke mortality, and respiratory mortality rate best account for the variance and 2) reduce the number of variables in our model. This method is a form of feature extraction to reduce the dimensionality of our data. We then conducted principal components analysis using the four pre-existing conditions and additional features of the ratio of Democrats to Republicans, age by sex, hospitals, and the number of intensive care unit (ICU) beds.

We created a categorical column for states and trained and tested our data using an 80%/20% split, respectively. Then, we generated two linear regression models. The target variable of the first model was the total number of cases. The target variable for the second model was proportional number of cases (total cases/population estimate in 2018). The predictors for both models were population estimates in 2018, diabetes percentage, heart disease mortality, stroke mortality, Democrat to Republican ratio, and percentage of smokers.

Furthermore, we conducted ordinary least squares regression on our variables to quantify and explore the statistical significance of the features of interest with respect to 1) the number of cases and 2) the proportional number of cases. We also plotted the top 10 states with the highest number of cases compared to their county population broken down by gender and age categories.

Summary of Results

Through our analyses, we found the top ten states with the highest coronavirus cases during the duration of January 22 to April 18 (Table 1). We also found the top ten states with the highest proportional number of coronavirus cases over the same period (Table 2).

From the scatterplot of stroke mortality versus heart disease mortality in the top ten states with the highest number of COVID-19 cases, we observe a slight positive trend (Figure 1).

People of all ages with chronic lung disease are also of a higher risk for severe illness of COVID-19, which is why we wanted to compare the percentage of smokers by states with the highest number of cases (Figure 2). We found that Louisiana (#9), Florida (#8), Pennsylvania (#4), Michigan (#5), and Illinois (#7) had populations in which 15% are smokers.

During our exploratory analysis, we also wanted to answer the question of whether the prevalence of coronavirus cases is affected by a state's political leanings. The scatterplot in Figure 3 shows the Democrat to Republican ratio and the total cases of coronavirus in the state. There is no clear trend that shows whether there is an association between a state's political

leanings and higher cases. This is a feature that we thought would be useful, but turned out to be ineffective.

In order to best predict states with high coronavirus cases using pre-existing conditions, we conducted principal components analysis. After mean centering and scaling to unit variance, we created a scree plot and found that 95% of the variance is described by the first principal component, accounting for a large amount of variance (Figure 4).

We would expect to see a negative relationship in the number of cases and hospitals/ICU beds. With a higher number of ICU beds or hospitals in a state, we would expect there would be fewer cases (Figure 5), as it would be easier for symptomatic individuals to get care (minimizing virus-spreading symptoms such as coughing) and public health information informing them of best practices for minimizing transmission. We found New York to be an extreme outlier with a higher number of cases despite having a middle range of number of hospitals (approximately 200) as compared to the other top ten states with the highest number of cases. Furthermore, we conducted a similar analysis and created a scatterplot for the total cases and the number of intensive care unit beds by top ten states and received the same results.

We also wanted to test our second part of the research question to see whether variables unrelated to pre-existing conditions were associated with high coronavirus cases. We conducted principal components analysis using variables of the ratio of Democrats to Republicans, age by sex, smoking habits, and population to see how strong they are as predictors.

Then, we trained and tested our model and used linear regression on the variables that we were interested in: population estimates in 2018, diabetes percentage, heart disease mortality, stroke mortality, Democrat to Republican ratio, and percentage of smokers. Adding smokers percentage decreased the training RMSE (30073.067 to 29991.71) and increased the testing RMSE (17379.69 to 18813.825).

We created a plot for the residuals versus actual total cases on test data and we saw a line that was not completely horizontal (Figure 6). We would ideally want to see a horizontal line of points at 0 meaning a perfect prediction.

We used ordinary least squares (OLS) regression to test whether we could identify which features were statistically significant (p-values < 0.05) and what the coefficients of significant features were. The OLS regression was used to create two models, the first model having total cases as the dependent variable and the second model having proportional cases (total cases/state population in 2018) as the dependent variable. Independent variables were selected by adding one feature at a time and observing the change in the adjusted R-squared value. If the adjusted R-squared value increased, the feature was kept in the model, otherwise, it was removed. The first OLS model had an adjusted R-squared value of 0.406. The second OLS model had an adjusted R-squared value of 0.556. As the second model performed better, we will report the results from this model.

The second OLS model found the following features to be significant: heart disease mortality (p-value: 0.004; coefficient: 0.5416), stroke mortality (p-value: 0.008; coefficient:

-0.5562), and Democrat to Republican ratio (p-value: 0.002; coefficient: 0.2958). Interestingly, diabetes percentage was not significant (p-value: 0.501) despite ample scientific literature indicating that diabetes is associated with severe COVID-19.[3] This finding led us to consider the effect that multicollinearity has on the model, as multicollinearity could cause the model's estimates of p-values and coefficients to differ from reality. With a variance inflation factor threshold of 5 (selected based on common practice described in scientific literature), we identified the features for which much of the variance in the regression coefficients was due to the presence of multicollinearity.[8] These features were diabetes percentage, heart disease mortality, and respiratory mortality rate in 2014.

In addition, when we plotted the ten highest states with coronavirus when using the 12 county population/sex categories, we received a perfect estimation line, meaning that the predictions were completely accurate in predicting the total coronavirus case values (Figure 7). When applying this model to all states, the accuracy decreased but still followed the same trend (Figure 8). This is interesting because we did not anticipate age categories stratified by males and females to be such a good standalone predictor for COVID-19 confirmed cases. After further examination with our training and test data, these 12 predictors were not actually good predictors on our test set: the linear regression model actually predicts some negative cases, which is impossible (Figure 9). These results could indicate overfitting using these 12 features, and they were particularly good for the top 10 states.

Discussion

In a pandemic situation, careful analysis detailing the number of cases, where cases are concentrated, and who is at highest risk of morbidity or mortality is essential. Our analysis shows that the four states with the greatest number of cases are New York (241,717), New Jersey (81,420), Massachusetts (36,372), and Pennsylvania (31,652). The four states with the greatest proportional number of cases are New York (0.012369), New Jersey (0.009140), Massachusetts (0.005270), and Louisiana (0.005060). These descriptive statistics are influential evidence for the benefit of more stringent quarantine or shelter-in-place regulations for these states and nearby states can help control the virus spread.

By creating a scatterplot of heart disease mortality versus stroke mortality of the top 10 highest coronavirus cases, we saw that there was a slight positive association between the two variables. Even though our research has shown that health conditions can lead to more serious cases, the results of the scatterplot show that these pre-existing conditions could be related and that we would only need to include one of the conditions in the model.

The fact that the first principal component accounts for 95% of the variance is interesting, as it is an indication that many of the features we analyzed are collinear. This interpretation is supported by the results of the variance inflation factor analysis (with a threshold of 5), which showed that only three of the eight predictors assessed in the OLS regression had little to no collinearity.

One of the variables that we would have wanted to include when conducting our principal components analysis would be the amount of resources (e.g. PPE per healthcare worker) that the

states' hospital has. We did not have this data, but this would be helpful to include in our analyses as the patient care has an effect on the spread within the states.

The fact that New York consistently had a higher number of cases than our regression models and scatterplots predicted indicates that something unusual may be occuring in the state, increasing the transmission rate there. Though this goes beyond the data contained in the dataset, our background research suggests that the high number of cases in New York are likely due to a number of factors.[4] New York City (NYC) is a popular international tourist destination, meaning those traveling from Asia shortly after the virus surfaced were more likely to have visited NYC than, for example, Jackson, Wyoming. The high population density of NYC is also likely contributing to the ease of transmission of the virus.

One important limitation of this dataset is that it only records confirmed cases of COVID-19. This means that a state with many cases but very little testing will have low numbers reported. This may be relevant for New York especially, as the state has aggressively pursued testing to limit the spread of the virus, likely inflating its reported cases.

Another limitation of the dataset is the likely influence of the ecological fallacy. As we are using aggregated state-level data rather than individual cases, there may be trends that exist at the state level that differ or disappear at the individual or city level and we are unable to report on. One example is the fact that, while coronavirus is known to have a greater effect on individuals with comorbidities, the total number of these individuals with comorbidities may be low in the general population, meaning that their influence on total morbidity and mortality is diluted. Similarly, the ratio of Democrats to Republicans may not have been statistically significant because the mayors and governors determine shelter-in-place rules. Wisconsin has a Democrat governor, but overall tends to have Republican constituents.

The ecological fallacy is important to note because it has important ethical implications. If researchers, governors, public health workers and other leaders do not take the ecological fallacy into consideration, they could take or advocate for actions that have dire consequences for how the country responds to the emerging crisis and, by extension, how much worse the pandemic gets. In the case of COVID-19, the actions and inactions that result from data analysis are truly a matter of life or death and it is important to understand the potential biases and limitations of the data available. That being said, the careful application of the results of this analysis can be important in limiting the human suffering caused by this pandemic and provide an important foundation for a meaningful and effective public health response.

Link to video here

Works Cited

[1]Bump, Philip. "Analysis | Even in States Hit Harder by the Coronavirus, Views of the Outbreak Correlate to Partisanship." *The Washington Post*, WP Company, 27 Mar. 2020, www.washingtonpost.com/politics/2020/03/27/even-states-hit-harder-by-coronavirus-views-outbreak-correlate-partisanship/.

[2]Geographic Differences in COVID-19 Cases, Deaths, and Incidence - United States, February 12-April 7, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(15):465-471. doi:10.15585/mmwr.mm6915e4

[3]Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497-506. doi:10.1016/S0140-6736(20)30183-5

[4]Kolata, Gina. "What Made New York So Hospitable for Coronavirus?" *The New York Times*, The New York Times, 26 Mar. 2020, www.nytimes.com/2020/03/26/health/coronavirus-nyc-spread.html.

[5]Renken, Elena, and Daniel Wood. "Tracking The Pandemic: How Quickly Is The Coronavirus Spreading State By State?" *NPR*, NPR, 8 May 2020, www.npr.org/sections/health-shots/2020/03/16/816707182/map-tracking-the-spread-of-the-coronavirus-in-the-u-s.

[6]"People Who Are at Higher Risk for Severe Illness." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 15 Apr. 2020, www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-at-higher-risk.html.

[7]Sun Y, Koh V, Marimuthu K, et al. Epidemiological and Clinical Predictors of COVID-19. *Clin Infect Dis*. 2020:1-7. doi:10.1093/cid/ciaa322

[8]Vatcheva, Kristina P., and Minjae Lee. "Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies." *Epidemiology: Open Access*, vol. 06, no. 02, 7 Mar. 2016, doi:10.4172/2161-1165.1000227.

Tables and Figures

Table 1. We found the top ten states with the highest number of coronavirus cases during the duration of January 22 to April 18.

State	**Coronavirus Cases**
New York	241712
New Jersey	81420

Massachusetts	36372
Pennsylvania	31652
Michigan	30791
California	30491
Illinois	29160
Florida	25492
Louisiana	23580
Texas	18704

Table 2. We found the top ten states with the highest proportional number of coronavirus cases during the duration of January 22 to April 18.

State	**Proportional Coronavirus Cases**
New York	0.012369
New Jersey	0.009140
Massachusetts	0.005270
Louisiana	0.005060
Connecticut	0.004912
Rhode Island	0.004248
Michigan	0.003080
Delaware	0.002624
Pennsylvania	0.002471
Illinois	0.002289

Figure 1. From the scatterplot of determining whether there is a potential association between heart disease mortality and stroke mortality of the top ten states with the highest cases of coronavirus, we observe a slight positive

trend meaning that these two variables are associated with each other and we may not have to include both in our model.

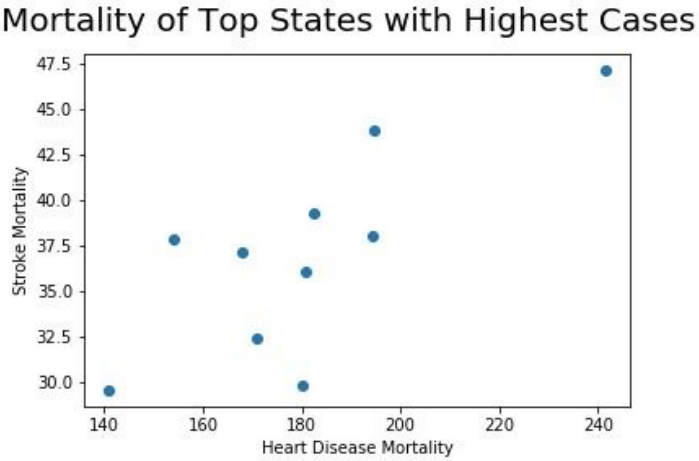

Figure 2. People of all ages with chronic lung disease are also of a higher risk for severe illness of COVID-19, which is why we wanted to compare the percentage of smokers by states with the highest number of cases.

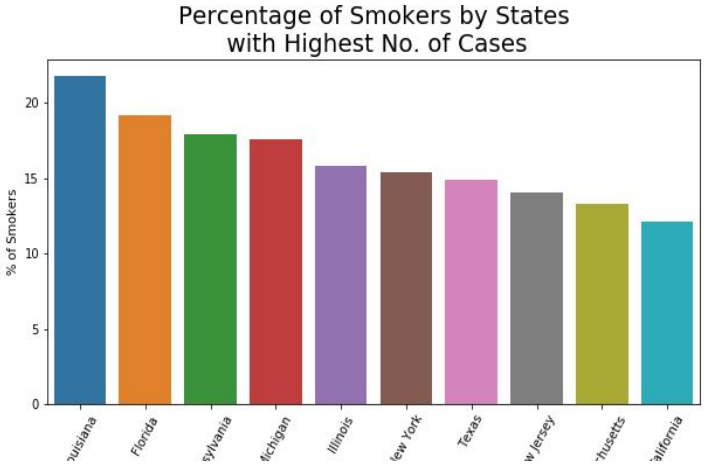

Figure 3. The scatterplot shows the Democrat to Republican ratio and the total cases of coronavirus in the state.

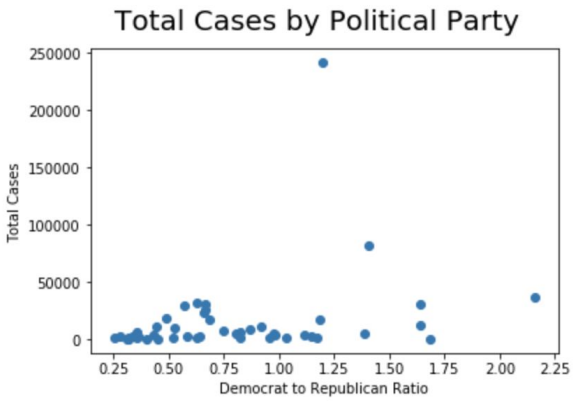

Figure 4. In order to best predict states with high coronavirus cases using pre-existing conditions, we conducted principal components analysis. We created a scree plot as pictured below where 95% of the variance is described by the first principal component, accounting for a large amount of variance.

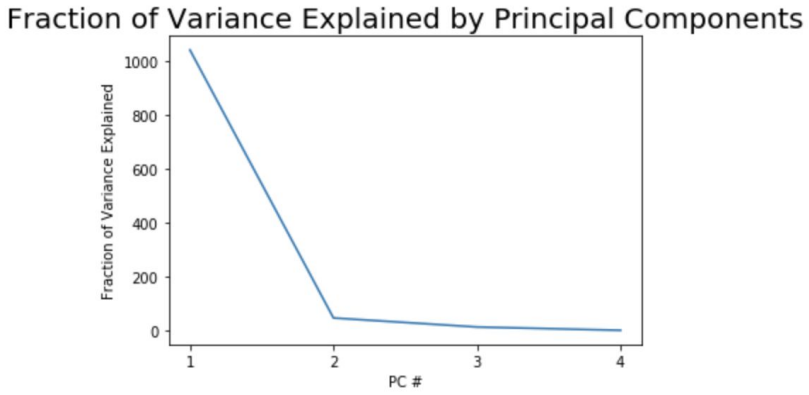

Figure 5. We conducted scatterplots of total cases of the top ten states versus hospitals/ICU beds and total cases and the number of intensive care unit beds by top ten states. New York was an outlier in both plots.

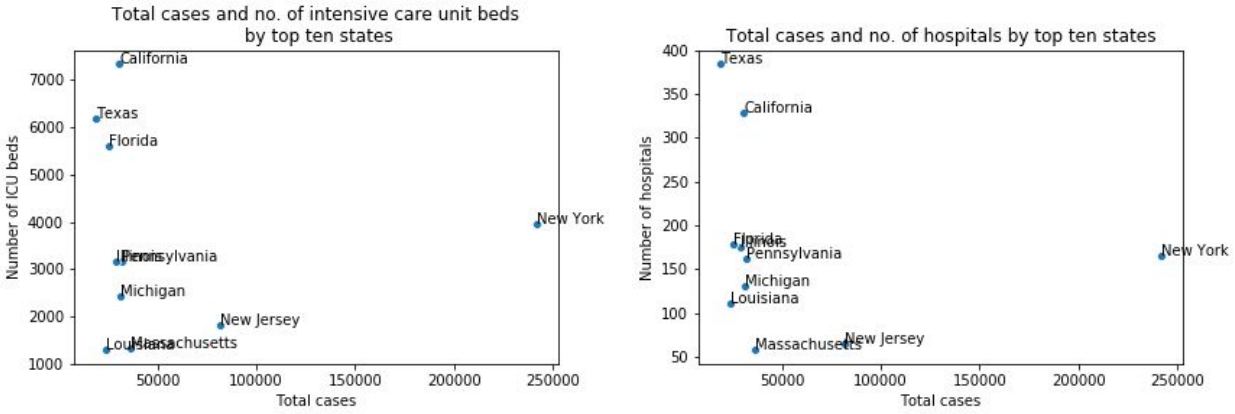

Figure 6. We conducted a plot for the residuals versus actual total cases on test data and we saw a line that was not completely horizontal. Another way to look at the test data is a scatterplot of the true (x-axis) and predicted (y-axis) number of cases; a model with low bias would have points that fell around the red line (where $Y = \hat{Y}$).

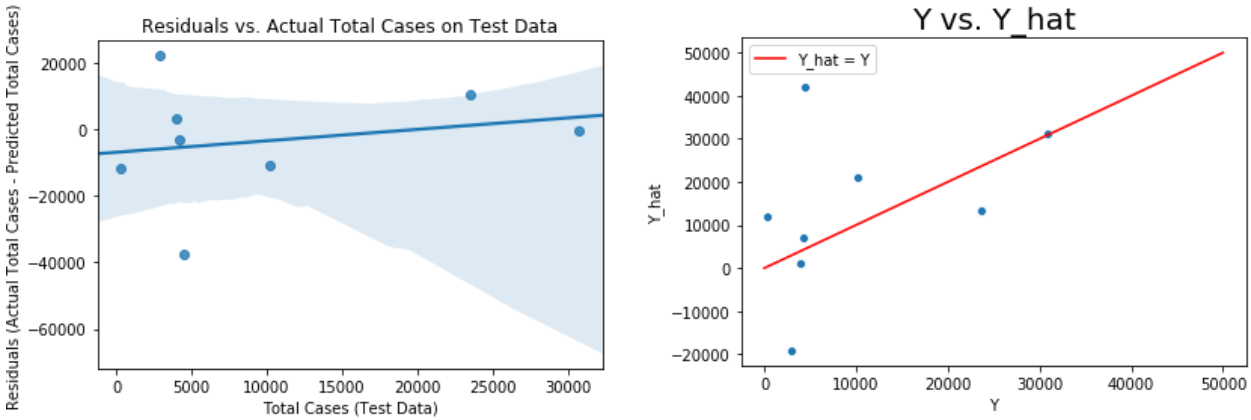

Figure 7. When we performed linear regression on the ten states with highest counts of positive cases using the county population/sex and age categories, we received a perfect line, meaning that the predictions were accurately predicting the total coronavirus case values, $R^2 = 1.0$. When looking at the data for all states using only population age and sex categories, the predictions were not as high, but they followed the same trend, $R^2 = 0.84$

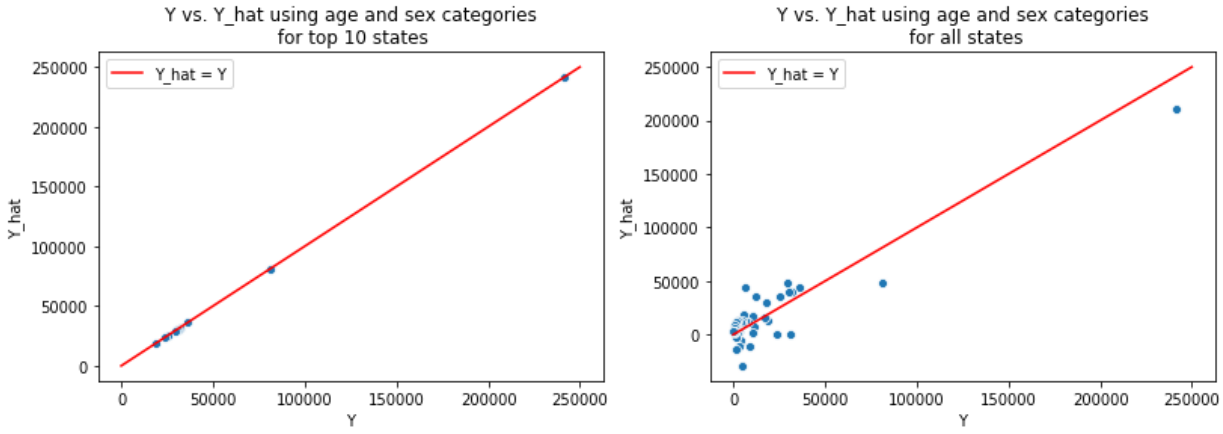

Figure 8. When using the same predictors as in *Figure 7* we see that they are not as accurate on the test data – ideally we should see the points fall on the $\hat{Y} = Y$ red line.