**Factors influenced the numbers of COVID-19 infected people in different states in the United States.**

Yuting Zhu(179009638), Jianze Wang(180002524) and Heyuan

Huang(172008062)

Rutgers University-New Brunswick

**ABSTRACT**

In the research, we tried to figure out the factors that had a significant influence on the number of COVID-19 infected people in different states based on the data collected on April 10th and built the finest regression model.

We utilized stepwise and cross-validation methods to fit the finest model. By stepwise, we filtered the major elements which influence the diagnosed population and eliminate the not essential factors. Then to examine which one of two models generated by stepwise model selection is the best, we need to apply the k-fold cross-validation method by the metric of RMSE.

Based on the result of model selection method and k-fold cross validation, the final model is Y = -6213.48 - 0.176 X1 + 179.502X2 + 20.732X3 + error, where X1 stands for the Population of the Homeless, X2 stands for the Population of Residents, and X3 stands for the Population Density.

The outbreak of coronavirus is too sophisticated to be completely studied by a simple linear regression model, our primary intention is to determine if the number of people diagnosed can be roughly estimated in terms of the population related factors, such as number of residents, population density, etc, keeping political things aside. However, we speculate that we may miss some of the important factors so that the model is strongly suffering from the omitting variable bias. Probably, different states are undergoing different stages of outbreak and to avoid such kind of time bias, it would be better to use the final number of people diagnosed after the end of the outbreak, but it is far from over. Furthermore, some of the factors are hard to quantify.

**INTRODUCTION**

The COVID-19 is one of the most serious public health problems in the world. Whats more, COVID-19 is rapidly becoming Americas leading cause of death. All human beings care about the situation in American. Whats the reason that COVID-19 is spreading so fast in American? Why the percentages of affected people in states divergent? Why do New York State, New Jersey State and Michigan State have the most people who are affected by COVID-19?

In order to figure out the questions, we tried to find the factors that influenced the numbers of COVID-19 infected people in different states in the United States and find a finest model to figure out how much each factor influenced the number of people who were infected in different states. According to the results, people could have a deeper understanding of the spread of the epidemic. The public could know more about the disease and the policy maker could design the propaganda plan of the precautionary measures of the COVID-19 more precisely.

In the research data, we chose the numbers of people who were confirmed positive for COVID-19, resident population(M), population density(/mile^2), 2019 Q3 GDP(B), rate of people who have bachelor degree and higher, employment rate, homeless population of all states in United States except Washington DC. We did not include Washington DC in the experiment because the population density in this area is much higher than other states in the United States and we thought that including Washington DC would make the fitting model biased.

First of all, to clarify the impact factors of the six elements we mentioned above, we leverage Stepwise to go through different models by adding or cutting elements one by one, to filter the fitted model and get rid of the insignificant elements.

Next, we wanted to select the best model among the resulting models from the model selection method. We tested the models using k-fold cross validation to check which one is consistent with the rule of parsimony. K-fold CV divided the data set into k folds and tested every fold of data using the selected model, and then generated overall metrics, like RMSE, Rsq, etc. Since if the model was overfitting, then every single metric from one fold would be large so that the overall metrics would be even larger. Hence, we chose the best model by the smallest overall metric.

METHODS AND PROCESS

**-----------------------------------Stepwise----------------------------------**

**Utilize stepwise to filter the major elements which influence the diagnosed population by the metric of p-value.**

**CODE:**

**lmod <- lm(P\_diagnosed ~ P\_resident + P\_Density + Q3GDP + Rate\_Bechelor + Rate\_Employment + P\_Homeless, data=Project)**

**ols\_step\_all\_possible(lmod)**

**ols\_step\_forward\_p(lmod,details=TRUE)**

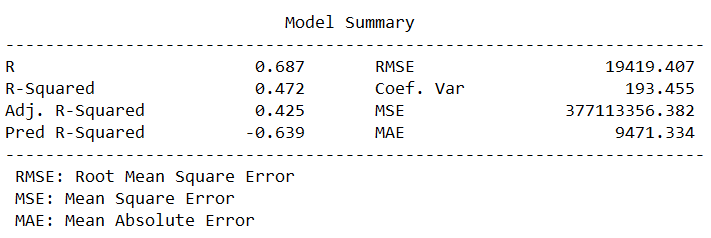
**ols\_step\_backward\_p(lmod,details=TRUE)**

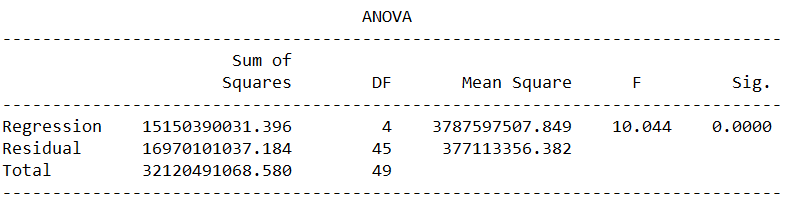
**ols\_step\_both\_p(lmod,details=TRUE)**

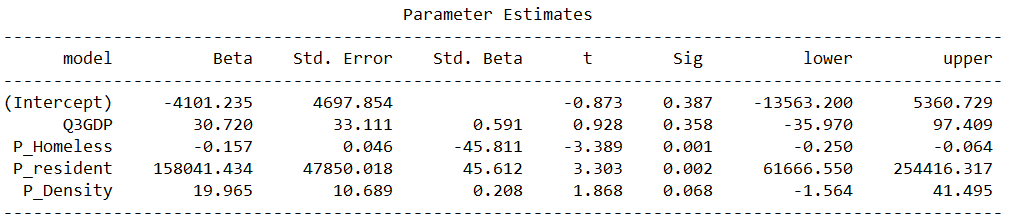
**\*\*\*\*\*\*\*\*\*\*\*only partial of the output of above code will be showed following\*\*\*\*\*\*\*\*\*\*\*\***

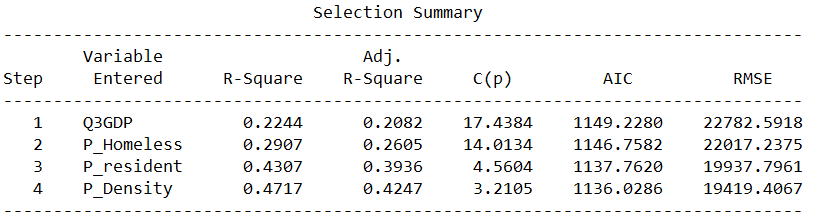
**Forward Selection Method:**

**Final Model Output**

****

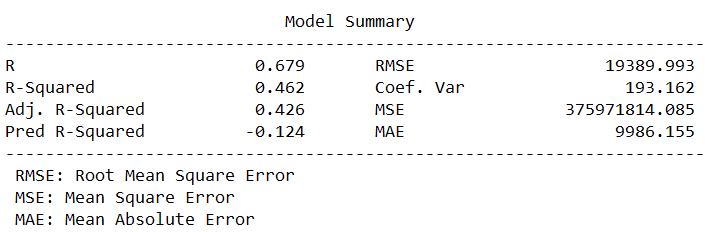
****

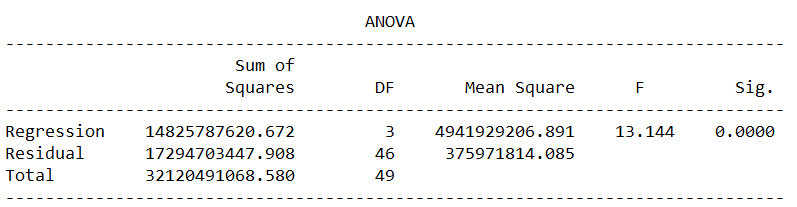
****

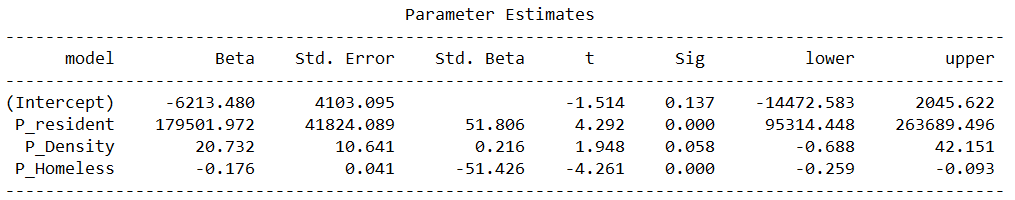
****

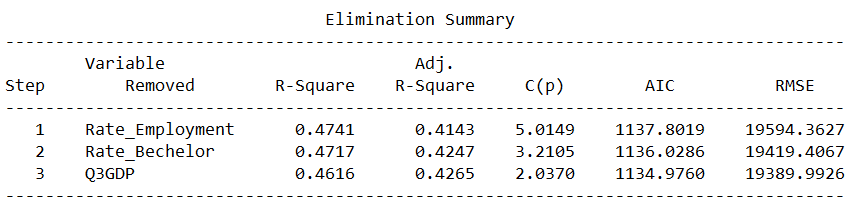
**Backward Elimination Method**

**Final Model Output**

****

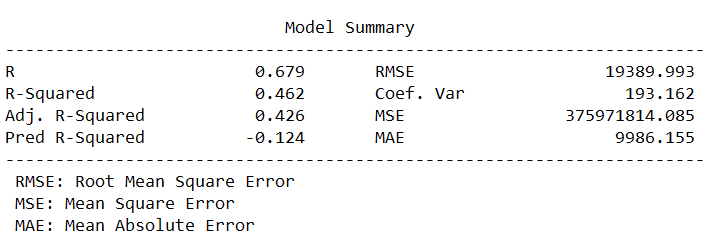
****

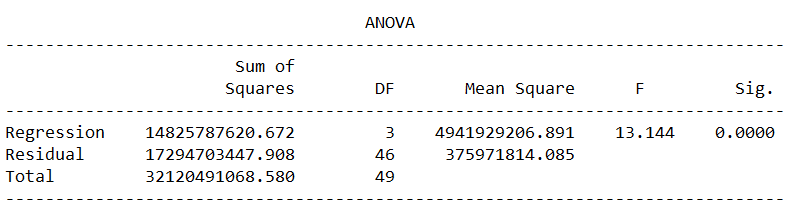
****

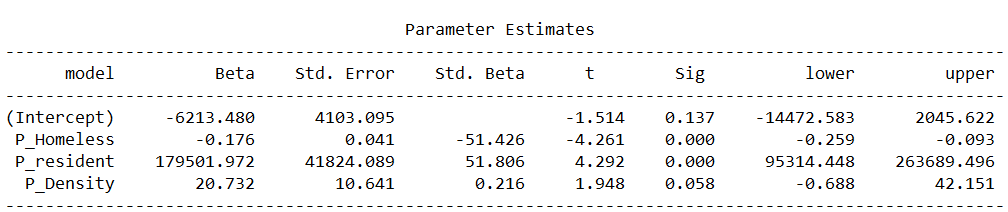
****

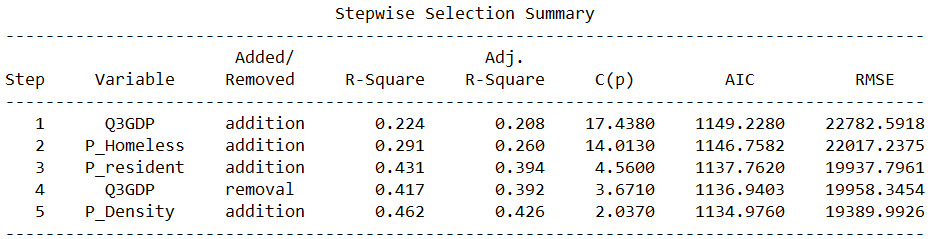
**Stepwise Selection Method**

**Final Model Output**

****

****

****

****

The final model from the Forward Selection Method is P\_diagnosed = P\_resident + P\_Density + Q3GDP + P\_Homeless. The final model from the Backward Elimination Method and Stepwise Selection Method is P\_diagnosed = P\_resident + P\_Density + P\_Homeless.

To examine which one of two models generated by stepwise model selection is the best, we need to apply the k-fold cross-validation method by the metric of RMSE. Based on the situation that we only have 50 data points, which is small, it would be practically more better to test which one is overfitting using 5 fold CV than using 10 fold CV.

**-------------------------------Cross Validation------------------------------**

**Code:**

#model generated by forward

model1 <- lm(P\_diagnosed ~ P\_resident + P\_Density + Q3GDP + P\_Homeless,data = covid19)

#model generated by backward and stepwise

model2 <- lm(P\_diagnosed ~ P\_resident + P\_Density + P\_Homeless,data = covid19)

#CV for model1

set.seed(32145)

train.control <- trainControl(method = "cv", number = 5)

model\_forward <- train(P\_diagnosed ~ P\_resident + P\_Density + Q3GDP + P\_Homeless,data = covid19, method = "lm", trControl = train.control)

#CV for model2

set.seed(12345)

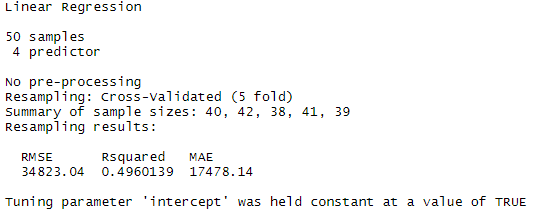
train.control <- trainControl(method = "cv", number = 5)model\_backward <- train(P\_diagnosed ~ P\_resident + P\_Density + P\_Homeless,data = covid19, method = "lm", trControl = train.control)

# Summarize the results

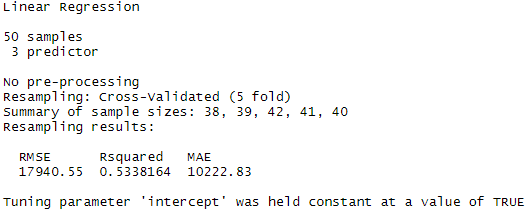
print(model\_forward)

print(model\_backward)

**Test result of model 1:**

****

**Test result of model 2:**



As we are using RMSE as the metric of k-fold CV test, we want to pick the model with the smaller RMSE. The RMSE of the model 1 is 34823.04 and the one of model 2 is 17940.55. Hence, we choose the model 2 as the final model to explore the relationship between the number of people who are diagnosed with COVID-19 and different factors.

**RESULTS**

Based on the result of model selection method and k-fold cross validation, the final model is

Y = -6213.48 - 0.176 X1 + 179.502X2 + 20.732X3 + error,

where X1 stands for the Population of the Homeless, X2 stands for the Population of Residents, and X3 stands for the Population Density.

To make the model more explicit, we let the unit of X2 to be thousands of people instead of millions of people which would make the coefficient too large.

**DISCUSSION**

While the real world issue on the outbreak of coronavirus is too sophisticated to be completely studied by a simple linear regression model, our primary intention is to determine if the number of people diagnosed can be roughly estimated in terms of the population related factors, such as number of residents, population density, etc, keeping political things aside.

Intuitively, as we wished, the number of people diagnosed should be positively related with the number of the homeless. Because homeless people are more likely to be exposed to the virus and they do not have enough medical support. However, the model that we finally have has a negative coefficient for that, which means that states having less homeless people are more likely to have a higher number of people diagnosed positive for the virus. We speculate that we may miss some of the important factors so that the model is strongly suffering from the omitting variable bias. Moreover, probably, different states are undergoing different stages of outbreak, which means that some of them may not reach the apex but some of them have reached, causing different acceleration of the growth of the number of diagnoses in each state on the date we picked for building the model. Therefore, the 50 data collected may not be the most suitable for estimation. To avoid such kind of time bias, it would be better to use the final number of people diagnosed after the end of the outbreak, but it is far from over.

Furthermore, some of the factors are hard to quantify. For instance, the fact that NYC has the highest number of diagnosed, besides it has a tremendous volume of people, is partially because of the extremely high population mobility, which is really difficult to find relevant data or be expressed numerically. The geographical factor is also crucial but impossible to be qualified as well. It is almost certain that the fact that NJ has the second highest number has something to do with the fact that NJ is contiguous to NYC. However, those geographical variables are impossible to be added to the regression model.

**Data Scource**

1 population density for states in United States <https://state.1keydata.com/state-population-density.php>

2 age average by state in the United States <https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_educational_attainment>

3 educational attainment by state in the United States <https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_educational_attainment>

4 unemployment rates for states, seasonally adjusted <https://www.bls.gov/web/laus/laumstrk.htm>

5 resident population of the U.S. in 2019, by state <https://www.statista.com/statistics/183497/population-in-the-federal-states-of-the-us/>

6 2019 Q3 GDP by state in the United States <https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_GDP>

7 homeless population by state 2020 <https://worldpopulationreview.com/states/homeless-population-by-state/>

8 number of people who were confirmed affected by COVID-19 by state until 4.10 2020 <https://coronavirus.1point3acres.com/>