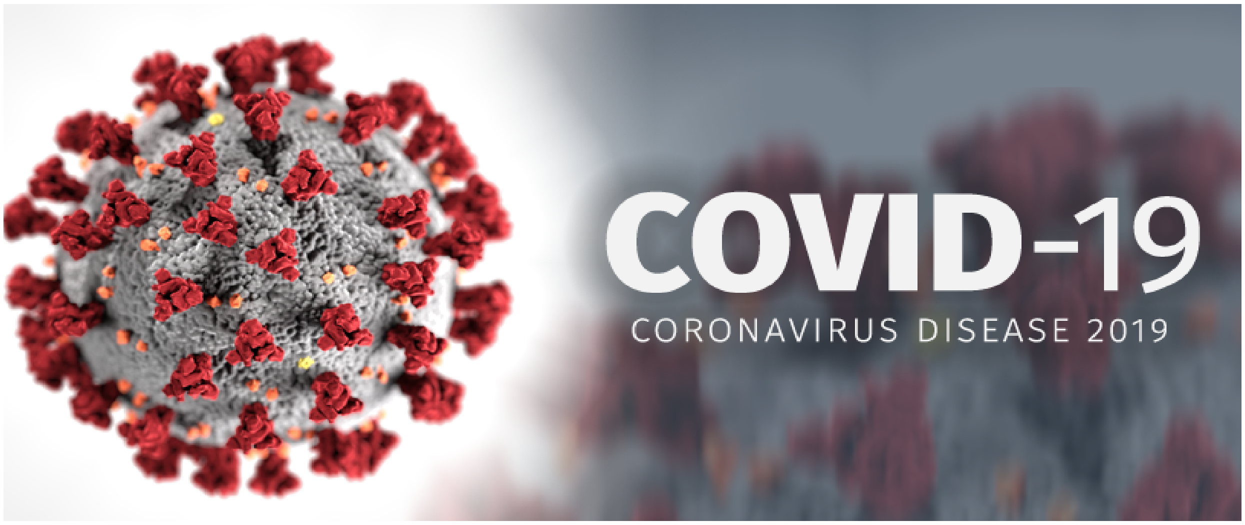
Influential Factors of Covid-19 Death in the U.S.



By Siyuan Ding, Xinyuan Wang, Yiqing Xin, Qixuan Ye, Mutong Yin

Go to the slides for Brief Introduction and Outline: https://docs.google.com/presentation/d/1anXSIesgsFGKTul9sPFnoEfHOap_rVXFOqX-VPjN0hI/edit?usp=sharing

1. Contributions: (In alphabetical order)

Siyuan Ding (siding@ucdavis.edu)	● Questions creation, data aggregation ● PCA (Principal Component Analysis), analysis ● Linear Regression, Report design
Xinyuan Wang (xwwwang@ucdavis.edu)	● MSPR (Mean Square Prediction Error) ● Statistic analysis (AIC, BIC and interaction) ● Data transformation and interpretation
Yiqing Xin (yqxin@ucdavis.edu)	● Multiple linear regression, data visualization, Outlier ● Statistic Analysis ● Report Revision
Qixuan Ye (qye@ucdavis.edu)	● Linear regression, analysis ● Final report design
Mutong Yin (mtyin@ucdavis.edu)	● Datasets collection, factor comparison ● Report design, PPT

2. Background

A new type of coronavirus named COVID19 has been widespread around the world, including the U.S., since the end of 2019. The virus comes without a clear cause, and with the lack of an effective vaccine, a large population of death has been induced.

The goal of our project is to find factors that affect the large death population caused by COVID19.

We use the dataset "COVID19_state" as our project data. It contains 26 variables regarding the potential causes of the large deaths in the U.S. by state. Twenty-four of them are numerical variables, and two of them are categorical variables. We modified the dataset by changing the 20th variable "Med-Large Airports" from numerical data to categorical "With/Without" and removing the "infected population" and "tested population" variables since their relationship to the death population are too close so that they also could be regarded as a response variable. We also remove the "school closure time" since it has empty data. The datasets have both "population" and "population density" variables. Since "population" may affect the "death population" too directly, we choose to use "population density" in our analysis and delete the "population" variable. So, our analysis's final dataset contains 22 variables, with 2 of them as categorical variables and 20 of them as numerical variables. Then, we build linear and multiple regression models, analyze which potential cause is the most significant, and how the linear regression model fits the data.

3. Key Question

1. How effective is a simple linear regression model explaining the relationship between the potential causes and the death population?

2. Among the influential factors, which one is the most significant?

3. What will happen if we construct multiple regression with the top two/three/four/five potential causes from question 1? Analyze and conclude the results.

4. Analysis Plan

Since we are interested in the factors that influence the severity of this Covid outbreak, we have the option to choose between "Deaths" and "Infected." We finally decided to use "Death" as our response variable because we are interested in spreading the virus and how states respond to it.

Besides, a simple look at the dataset tells the diversity of factor levels in the variables. So above all, some simplifications could be conducted. For instance, we turned the "Med.Large.Airports", which indicates the number of medium and large-sized airports in a state, from numerical to non-numeric of "with/without," i.e., a dummy variable. The change from quantitative variable to qualitative variable reduces the factor levels to two.

After making some adjustments, we need to split the data of different states into two groups, using training data to build the models and test data to show the fitness. We set aside a portion of observations for each state as test data and use the remaining observations as training data. "set.seed" function is used for random selection.

5. Data exploration and transformation

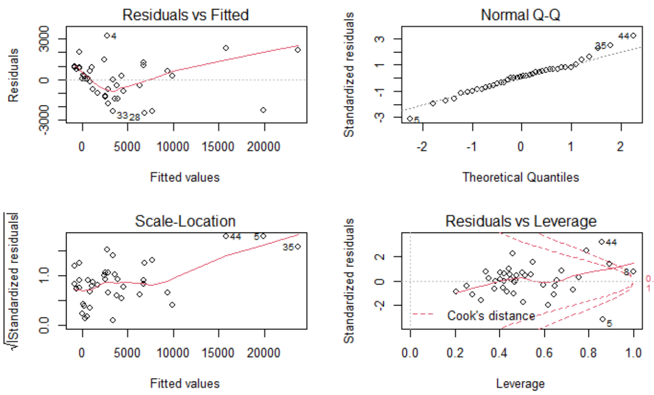


Fig. 1. Diagnostic Plot of the Training Data Set Without Transformation

By observing the training data set's diagnostic plot without transformation, we realize that the data does not meet many assumptions (see fig. 1). From the Residual vs. Fitted plot, we see that the linearity is under question, for the data are not locating around the 0 horizontal-line. The Normal Q-Q plot shows that the data is heavy-tailed. The Scale-location graph shows that the homoscedasticity assumption does not necessarily hold, and we also have some high leverage points according to the residual vs. fitted plot.

Thus, we need to find the right type of data transformation before proceeding to the next analysis steps. By using the boxcox() function, we see that the lambda value of the full first-order model is around 0.25, indicating that this model needs a power of 0.25 transformation on the response variable "Deaths" so that the distribution of this non-normal response variable could become normal (see fig. 2).

We also have one data entry in our training data set with high leverage (see fig. 3). By observing this particular entry, we see that this data represents the District of Columbia. Since we are interested in the relationship between the potential causes and the death population in each state, we decide to rule out the District of Columbia since its demography contains significant differences compared with other states. With the current data set, we may be unable to analyze DC into the model considerations.

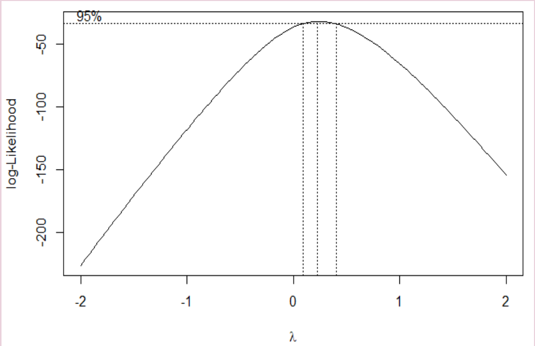


Fig. 2. Boxcox Graph of the Full First Order Model

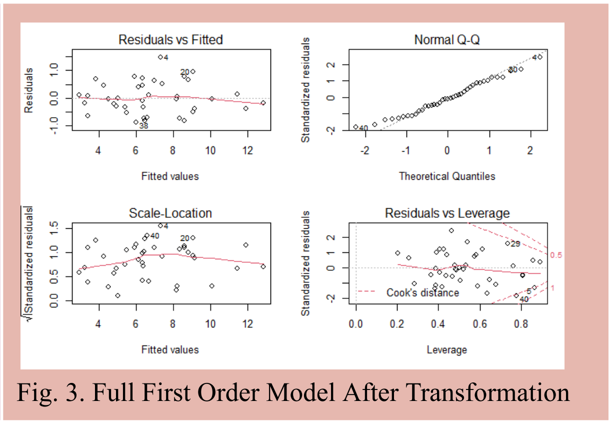
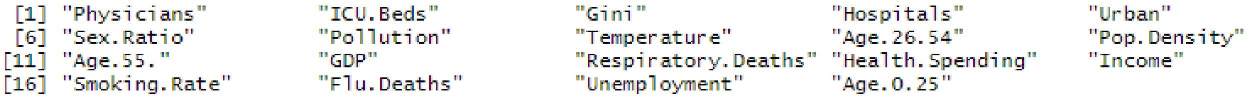


Fig. 3. Full First Order Model After Transformation

6. Analysis and Results

6.1 Linear Regression

After the transformation, we applied linear regression to our transformed response variable (short for variable). The goal for this step is to choose the five top most statistically significant variables. First of all, we removed two categorical variables, which are “state” and “Med.Large.Airports,” and kept the rest numerical variables to see the linearity between “deaths” and other variables. Then, we used the summary() function to see the f value of each variable and get the descending order with order () and “decreasing = true” functions inside the order() functions. Lastly, with the names() function, we can get the variables in a significant descending order (see fig. 4). We could choose the five most significant variables now, but we need to confirm that the variables we got are not highly correlated. Therefore, we need to do Principal Component Analysis (PCA) to see the correlation between variables.



```
 [1] "Physicians"    "ICU.Beds"      "Gini"                "Hospitals"         "Urban"
 [6] "Sex.Ratio"     "Pollution"     "Temperature"         "Age.26.54"         "Pop.Density"
[11] "Age.55."       "GDP"           "Respiratory.Deaths"  "Health.Spending"   "Income"
[16] "Smoking.Rate"  "Flu.Deaths"    "Unemployment"        "Age.0.25"
```

Fig. 4. Names of Highest Correlation Factors in Order

6.2 Principal Component Analysis (PCA)

The PCA test estimates the correlation between two variables. A higher correlation between two variables means these two variables are dependent on each other, which implies we only need to choose one of them as the parameter in a model. The reason is that we want to keep the variables independent so that there is no interaction between them acted on the model. By using the prcomp() and biplot() functions, we tested the top five variables derived from linear regression that are “Physicians,” “ICU.Beds,” “Gini,” “Hospitals,” and “Urban.” Based on figure 5, while “Urban” and “Gini” variable vectors form a right angle that shows a low correlation, “physicians,” “ICU.Beds,” and “Hospitals” are positively correlated since vectors of them overlap.

Hence, we keep “Gini” and “Urban,” then choose “Physicians” to represent “Physicians,” “ICU.Beds,” and “Hospitals,” since “Physicians” has the highest significance among these three. The next top two variables are “Sex.Ratio” and “Pollution.” In figure 6, although the vectors of “Physicians” and “Pollution” do not parallel, two variable vectors are close to each other. Thus, for the same reason, we utilize “Physicians” to represent “Physicians” and “Pollution.” The variable after “Pollution” is the “Temperature. ” And the biplot funded by “Physicians,” “Gini,” “Urban,” “Sex.Ratio,” and “Temperature” shows these five variables are independent because there are no parallel or overlaps (see fig. 7). Therefore, the best top five variables chosen to construct multiple regression are “Physicians,” “Gini,” “Urban,” “Sex.Ratio,” and “Temperature.” We will take particular care of the variable Physicians” and denote this simple linear regression model as sel1.

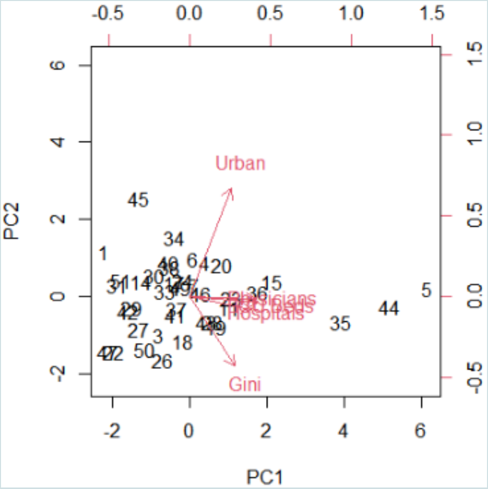


Fig. 5. Plots with Five Variable Vectors: “Physicians,” “ICU.Beds,” “Gini,” “Hospitals,” and “Urban.”

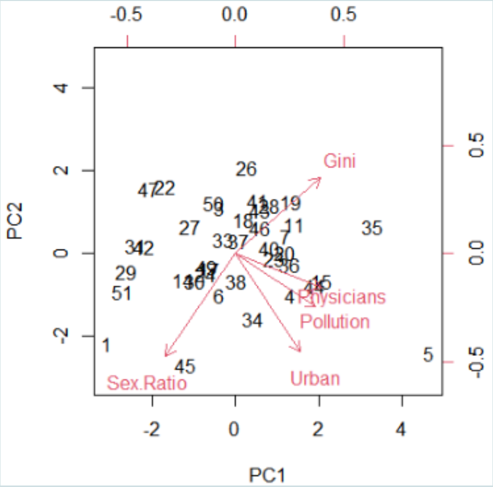


Fig. 6. Plots with Five Variable Vectors: “Physicians,” “Gini,” “Urban,” “Sex.Ratio,” and “Pollution.”

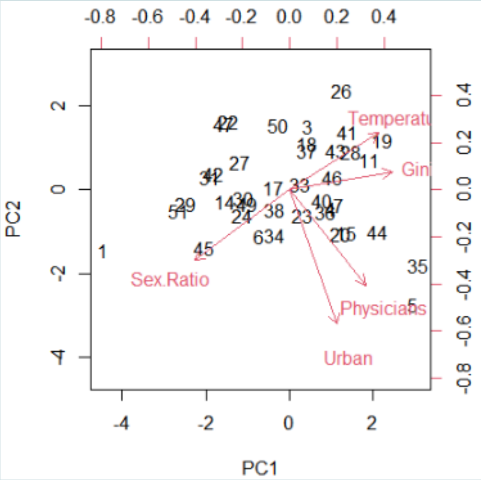


Fig. 7. Plots with Five Variable Vectors: “Physicians,” “Gini,” “Urban,” “Sex.Ratio,” and “Temperature.”

6.3 Multiple Linear Regression

In this part, we used the top five fittest and independent variables obtained from linear regression and PCA sessions to construct four multiple linear regression. By plotting graphs of multiple linear regression and multiple logistic regression with the top two variables, “Physicians” and “Gini,” we found no difference between diagnostic plots of linear one and logistic one. The only difference is that the red line in the logistic function is more curved or smoothing than the red line in the linear function (see fig. 8 & fig. 9). Therefore, to keep the linear consistency above, we choose to use multiple linear regression to represent the multiple regression part.

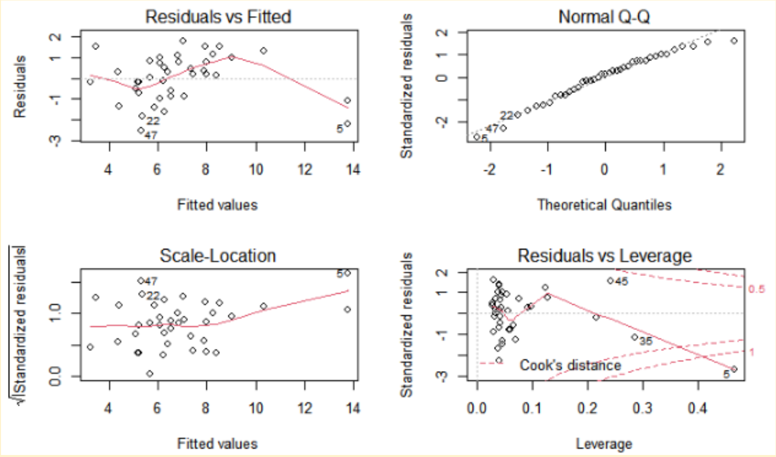


Fig. 8. Multiple Linear Regression: Deaths V.s. Physicians+Gini

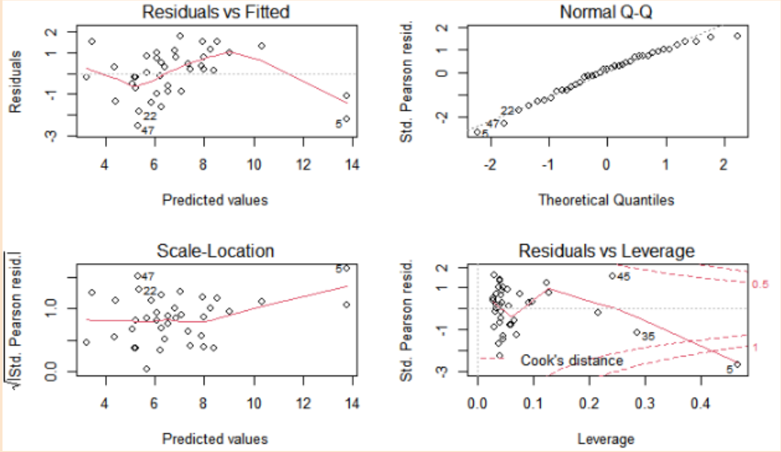


Fig. 9. Multiple Logistic Regression: Deaths V.s. Physicians+Gini

These five variables are “Physicians,” “Gini,” “Urban,” “Sex.Ratio,” and “Temperature.” And the four multiple linear regressions are four models constructed by top two variables, top three variables, top four variables, and top five variables, respectively. In order to find the best model among these four models, we calculated and compared the adjusted r-squared values and Akaike information criterion (AIC) values. While the adjusted r-squares are modified r-squares based on added variables, AIC estimates the relative amount of information lost by a given model. Less information a model loses, the higher quality a model has. Thus, the model with the highest adjusted r-squared value and the smallest AIC value will be the best-fitted model. Through the calculation, the best multiple regression is the model with all five variables, which has adjusted r-squared value = 0.8851063 and AIC = -8.7.

On the other hand, we analyze models from diagnostic graphs of those four multiple linear regressions. Compared with the "Residuals vs Fitted" graphs, we checked the red lines inside and found the models with the top two variables and top three variables have larger residuals than the other two models, which shows a less fit (see fig. 10 & fig. 11). Then, compare the models with the top four variables and top five variables, we realized that the "Normal Q-Q" plot of the model with four variables has lesser points close to the dashed line than the model with five variables (see fig. 12 & fig. 13). This implies that the top five variables model is a better model than the model with the top four variables. To conclude, the multiple linear regression with all top five variables is the best model among those four models in this session, and we will denote this model as sel2.

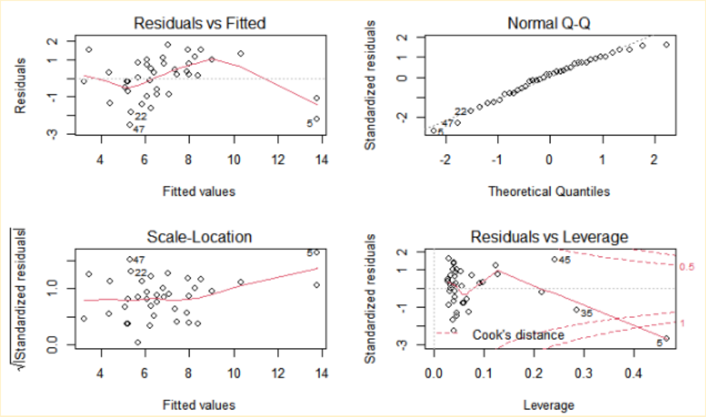


Fig. 10. Deaths V.s. Physicians+Gini

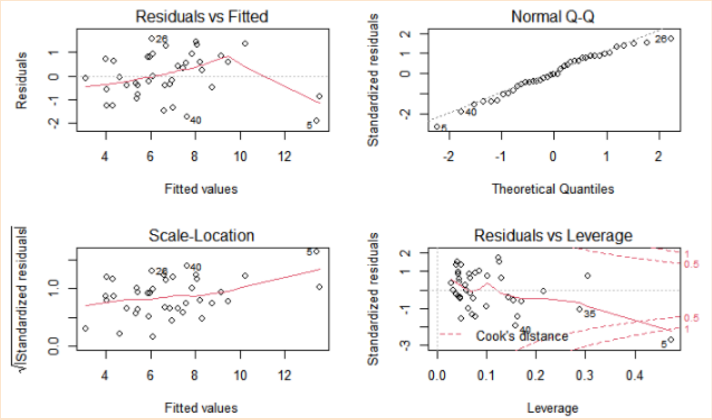


Fig. 11. Deaths V.s. Physicians+Gini+Urban

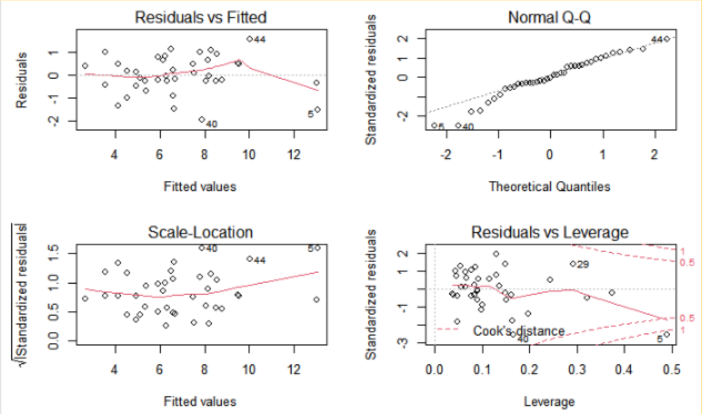


Fig. 12. Deaths V.s. Physicians+Gini+Urban+Sex.Ratio

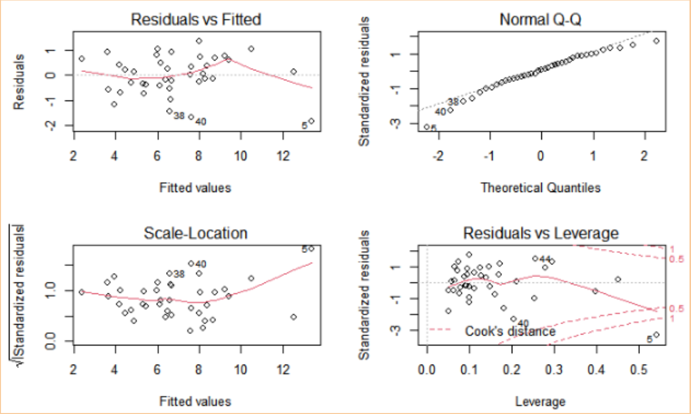


Fig. 13. Deaths V.s. Physicians+Gini+Urban+Sex.Ratio+Temperature

6.4 Multiple Regression Model with AIC, BIC Model Selection Criteria

As we have explored linear regression and used its result to construct some multiple regression models, we would also like to consider first-order models that are purely based on model selection criteria AIC and BIC (Bayesian information criterion). To do so, we set up a model0 and a modelF corresponding to the lower bound and the upper bound of our model selection range. Model0 is the model with the only intercept, whereas modelF is the full first-order model. Then we use stepwise algorithms to find the optimal model with the minimum AIC and BIC value starting from model0 and modelF, respectively.

Using AIC, we have obtained two distinct models by starting from model0 and modelF. We denote them as sel3 (see fig. 14) and sel4 (see fig. 15). Using BIC, we have also obtained two distinct models, and we will denote them as sel5 (see fig. 16) and sel6 (see fig. 17). From their diagnostic plot, both models seem to hold their assumptions reasonably, so we would have to decide which one is the fittest after our model validation process.

Fig. 14. Deaths V.s. Physicians + Sex.Ratio + Urban + Age.0.25 + Gini + Hospitals + Pollution + Smoking.Rate

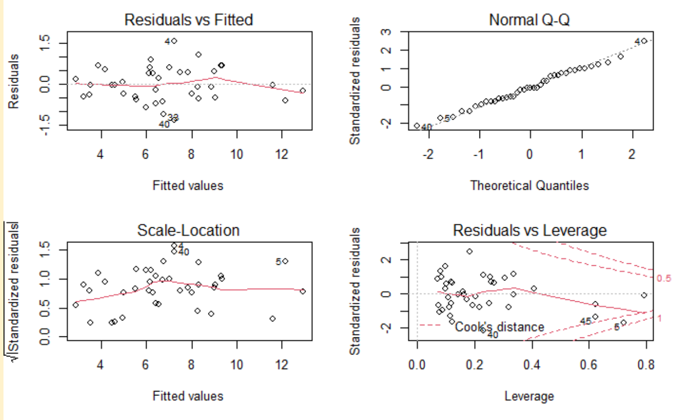
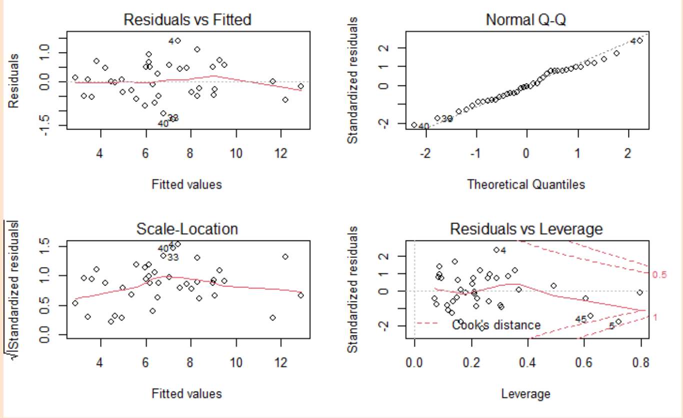


Fig. 15. Deaths V.s. Gini + Sex.Ratio + Smoking.Rate + Physicians + Hospitals + Pollution + Urban + Age.26.54 + Age.55



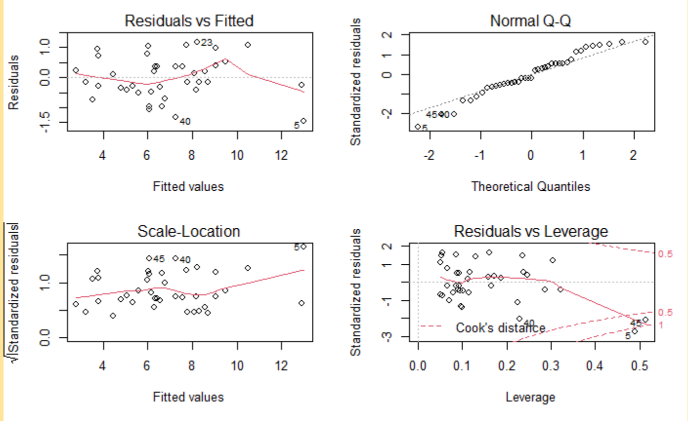


Fig. 16. Deaths V.s. Physicians + Sex.Ratio + Urban + Age.0.25 + Gini

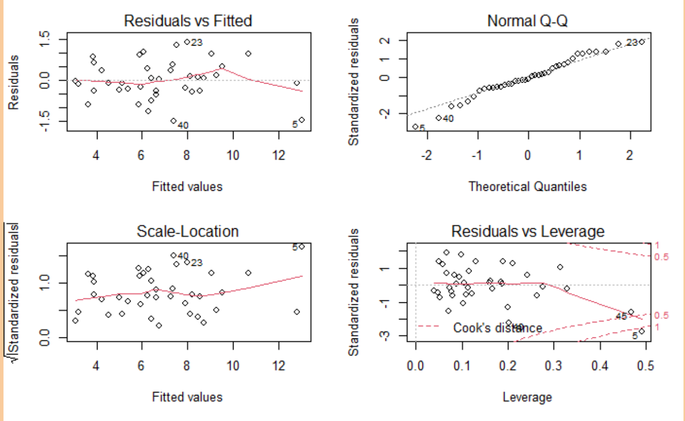


Fig. 17. Gini + Sex.Ratio + Physicians + Urban + Age.55.

6.5. Multiple Regression Model with Interaction Effects

As we have gone over different model selection methods regarding the first order multiple regression model, it is also necessary to check multiple regression models with interaction effects. There might be some other interesting patterns to be studied. Building on the spirit of that, we again use AIC and BIC to find the best multiple regression model with interaction. We find that our selection method agrees on the following model, denoted as sel7 (see fig. 18). This model is particularly interesting, as it demonstrated that the interaction effect between "Physician" and "Urban" might decrease the number of deaths, and the interaction terms between "Age.0.25" with "Gini" would increase the number of deaths.

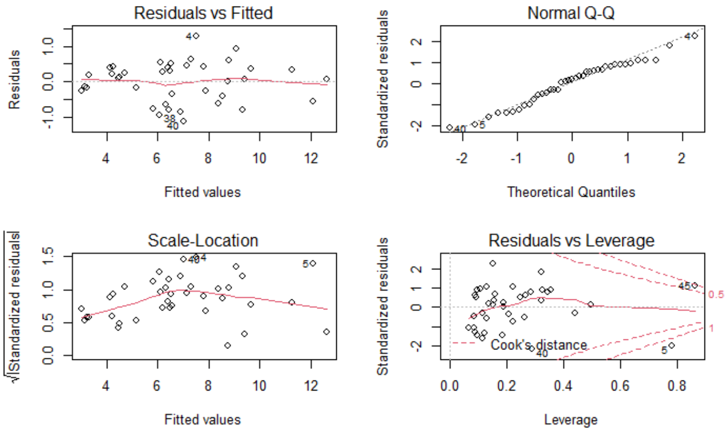
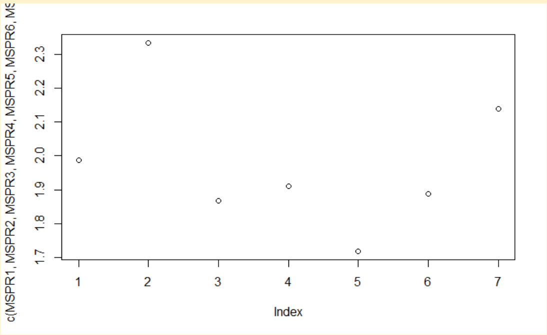


Fig. 18. Deaths V.s. Physicians + Sex.Ratio + Urban + Age.0.25 + Gini + Health.Spending+Physicians:Urban + Age.0.25:Gini

6.6 The Fittest Models and Mean Square Prediction Error (MSPR)

Mean square prediction error evaluates the ability of the model to predict accurately. The smaller MSPR value would suggest that the model is closer to the truth. From the previous model selection methods, we have seven models that could be considered, denoted as sel1 to sel7. As we see in the plot of MSPR value (Fig. 19) of all final model candidates, we see that sel5 has the lowest MSPR value, indicating that this model seems to be the most accurate one among all other models. As we see the "Boxcox" graph for all the models, we realized that two of our models need some model transformations (Fig. 20). Thus, we added the additional model transformation to sel1 and sel7 and obtained the result that sel1 has a large prediction error despite its proved to be the most significant simple linear regression model. Sel7, on the other hand, has demonstrated the best prediction ability with the lowest MSPR value after the model transformation. We have that the corresponding MSPR value of each model are sel1=2906.966, sel2= 2.334768 sel3= 1.867212 sel4= 1.91085 sel5=1.718682 sel6=1.887055 sel7=0.6453922. The best first-order model is sel5, which contains variables "Physicians," "Sex.Ratio," "Urban," "Age.0.25," and "Gini." Based on our model selection criteria, the overall best model is the interaction model, which has the lowest MSPR value and the highest adjusted R-squared value of 0.9357. Sel2, our hand-picked model, appears to be the least accurate model based on MSPR value and significant factors, but still holding a high Adjusted R squared value of 0.92.

Fig. 19. MSPR Values for Final Model Candidate



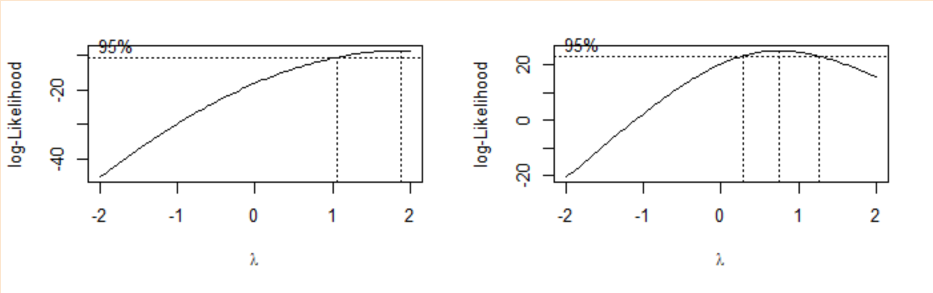


Fig. 20. Boxcox Plot for sel1 and sel7

7. Conclusion

According to the calculation above, we reconstructed selected data, split data, and used data transformation to explore the relationship between the death population and correlated variables in the data frame. Through applying methods of linear regression, principal component analysis, multiple linear regression with top five variables, multiple regression with step() function, and multiple regression with interaction effect, we compared f statistics, adjusted r-squares, AIC, BIC, and MSPR to figure out the fittest model of estimating the death population.

The model that we constructed, though it might be informative, yet has some potential problems. The first one is the amount of data that it contains is not as much as we would like since only 50 states (and a special district that we rule out) are in the U.S.. Thus, it is not easy to construct accurate models.

According to the linear regression models, the "Physicians" is the most vitally correlated variable to the death population. However, the number of physicians is a correlation instead of causation since more physicians in a state mean more death cases will be recorded in hospitals. Therefore, linear regression is not sufficient and not effective to explore the causes of the death population due to Covid-19. The same should also apply to ICU.beds and hospitals. The first variable that seems to be a causation factor is the variable "Gini," representing economics's inequality. As we see positive coefficients to be significant, we may conclude that equality within economics would likely result in more deaths. Similarly, we have "Urban" as another variable that causes more death cases, which makes sense since a more urbanized region would likely make the virus easier to spread and thus generate more cases of deaths.

Then, through utilizing the PCA method, we figured out five independent variables that positively correlate to the death population; these five variables formed sel2 - Deaths V.s. Physicians+Gini+Urban+Sex.Ratio+Temperature. Compared to the other three models with the top two, top three, and top four variables, respectively, sel2 is the best model to explain the death population. Nevertheless, the PCA method provided limited qualified variables and compositions to form a suitable multiple regression. Thus, step() function, AIC, BIC, and MSPR are used to verify the correctness of sel2 and find out whether a better model exists.

We see that, based on adjusted R-squared value, AIC, BIC, and MSPR, sel2 is not the best multiple regression model. Even so, sel2 is a much better model than the predictability of the simple linear regression model. It further proved that simple linear regression on this data set might not be effective.

The best first-order multiple regression model that we found was sel5, which contains Physicians + Sex.Ratio + Urban + Age.0.25 + Gini. We are quite surprised that "Sex.Ratio" appeared to be in every multiple regression that we found, and it is negatively related to deaths. Recall that the "Sex.Ratio" factor is calculated based on the population of males over females in each state. Thus the "Sex.Ratio" factor would indicate that a state with more females would be more likely to have more deaths due to Covid compare with a state with more males. This could be justified by a study from Georgia University, "Sex-dependent regulation of social reward by oxytocin receptors in the ventral tegmental area," Johnathan M. Borland and his colleagues have illustrated that females find social interactions to be more rewarding than males; therefore, females are more likely to be engaged with people, which explains this phenomenon. Also, the fact that "Age.0.25" included in the model could be interpreted as young people would have higher mobility and more risk-seeking than a more aged population, hence contributing more to the death population during this pandemic.

The multiple regression model with interaction effects is more interesting, and it is also the model that best describes the data set. It contains the following factors: Physicians + Sex.Ratio + Urban + Age.0.25 + Gini + Health.Spending + Physicians:Urban + Age.0.25:Gini. The "Health.spending" factor is once again more of a correlation factor than a causation factor. The first interaction term Physicians:Urban contributes negatively towards the death population, indicating that if a state is highly urbanized and has enough physicians, then the number of deaths will reduce. This is explainable because highly urbanized states generally have better medical conditions that could reduce death due to viruses. For example, better doctors, better equipment, and more resources. The other interaction term, Age.0.25:Gini, positively affects the death population, which means that if a state has more young people, and the inequality of the economy in that state is significant, we would expect a higher death population.

8. Reference (MLA)

Brownlee, Jason. “Probabilistic Model Selection with AIC, BIC, and MDL”.October 30, 2019 https://machinelearningmastery.com/probabilistic-model-selection-measures/

Prabhakaran, Selva. “Outlier Treatment With R | Multivariate Outliers.” *r-Statistics.co*, 2016, r-statistics.co/Outlier-Treatment-With-R.html.

Ranger, Night. “COVID-19 State Data.” *Kaggle*, 3 Nov. 2020, www.kaggle.com/nightranger77/covid19-state-data.

Shedden, Kerby. “Prediction”, November 9, 2018. http://dept.stat.lsa.umich.edu/~kshedden/Courses/Regression_Notes/prediction.pdf