Final Project - Analysis of the early Covid-19 dataset by Elsevier

Liuyang Zheng - A15409562
lizheng@ucsd.edu

1. Introduction

The data we got are from The COVID-19 resource centre hosted by Elsevier. Elsevier has granted permission to make all its COVID-19 related research on the resource centre publicly available with acknowledgement of the original source. It consists of six statistical features of 16 countries in this pandemic: Total cases, active cases, recovery cases, week 4 deaths, CFR and week 5 deaths. Amongst them, India doesn't have week 5 deaths, and we will predict that in this report. We believe that this is an urgent health crisis, and that's why our report carries a lot of significance. We want to apply what we learned from this class to provide useful predictions for frontline workers fighting this disease.

The rest of this report will consist of four parts. In the first part, we fitted a linear model to the original dataset with all five numerical features. We also plotted the pairwise plot and the partial regression plot to check for the relationship between those features. In the second part, we removed Brazil and transformed the explanatory features with logarithm. Then we fitted a linear model on the transformed data. In both parts, we diagnosed the fitted model by examining their adjusted r-squared values. In the third part, we compared and evaluated the two models we fitted in the previous two parts. Last but not least, in the fourth part, we checked correlations between the explanatory variables and tried regularization to improve the model.

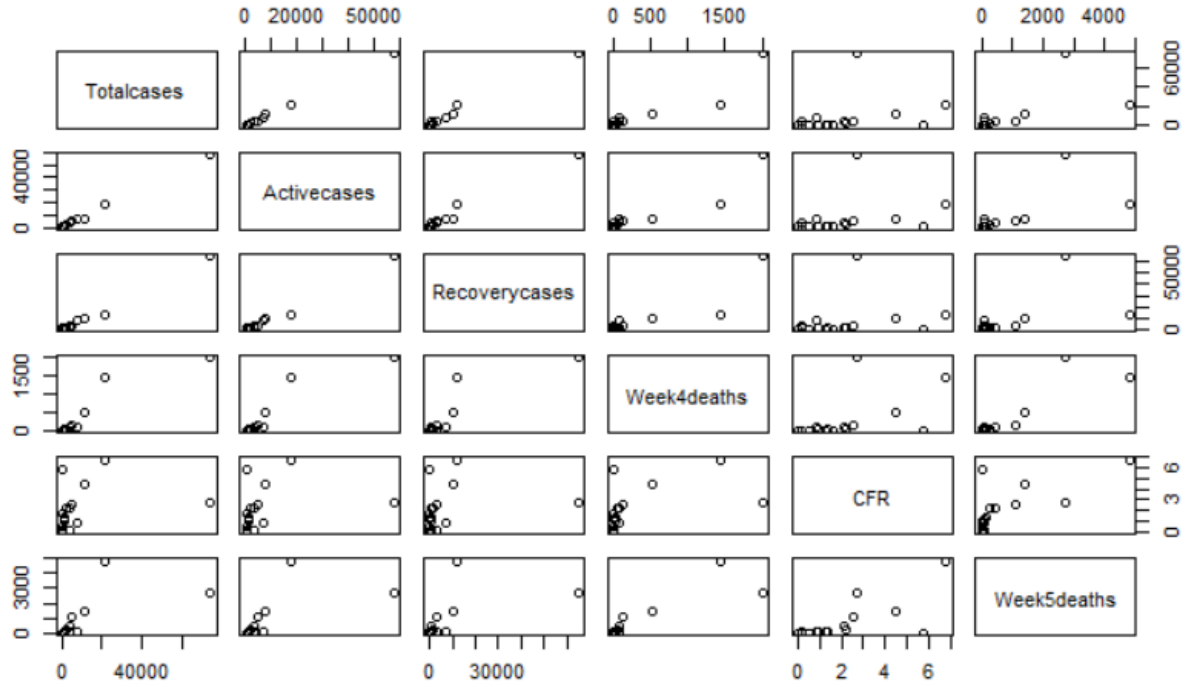
The conclusion part will conclude our questions and findings in a more augmented perspective. It also talks about the limitations of the data we have. For example, we need to keep in mind that the recovery number in the US was originally missing. Researches imputed the data by utilizing the fact that the total cases have a strong correlation with the recovery cases. The fitted model is not perfect and likely contains error, which might affect our prediction accuracy.

Additionally, in the appendix part, we will list more detailed descriptions of the code and graphs we used in our analysis. Graphs here will be referenced at the relevant parts of the report. The readers are expected to consult the appendix if they have any doubt on the procedures or results in the report. They will find a detailed description of each item regarding where they were mentioned in our analysis procedures.

2. Body

2.1 Fit on the Original

In this part, we first plotted the pairwise plot for all features to check their relationships, especially with the response variable, week 5 deaths.



From the graph, we can see that “week 5 deaths” actually has a linear relationship with all other features. But the relationships are seemingly logarithmic.

Next, we fitted a linear model on the dataset.[1] The fitted model has an adjusted R2 of 0.97 and a R2 of 0.98, which indicate a pretty good fitting. In addition, the followings are the intercept and coefficients:

The intercept is 84.42, which means if all explanatory variables are 0, the week 5 deaths number would be 84.42.

The coefficient for Total cases is -0.06999, which means that for every additional case, there would be 0.06999 decrease in the week 5 death number.

The coefficient for Active cases is 0.12155, which means that for every additional active case, there would be 0.12155 additional death cases in week 5.

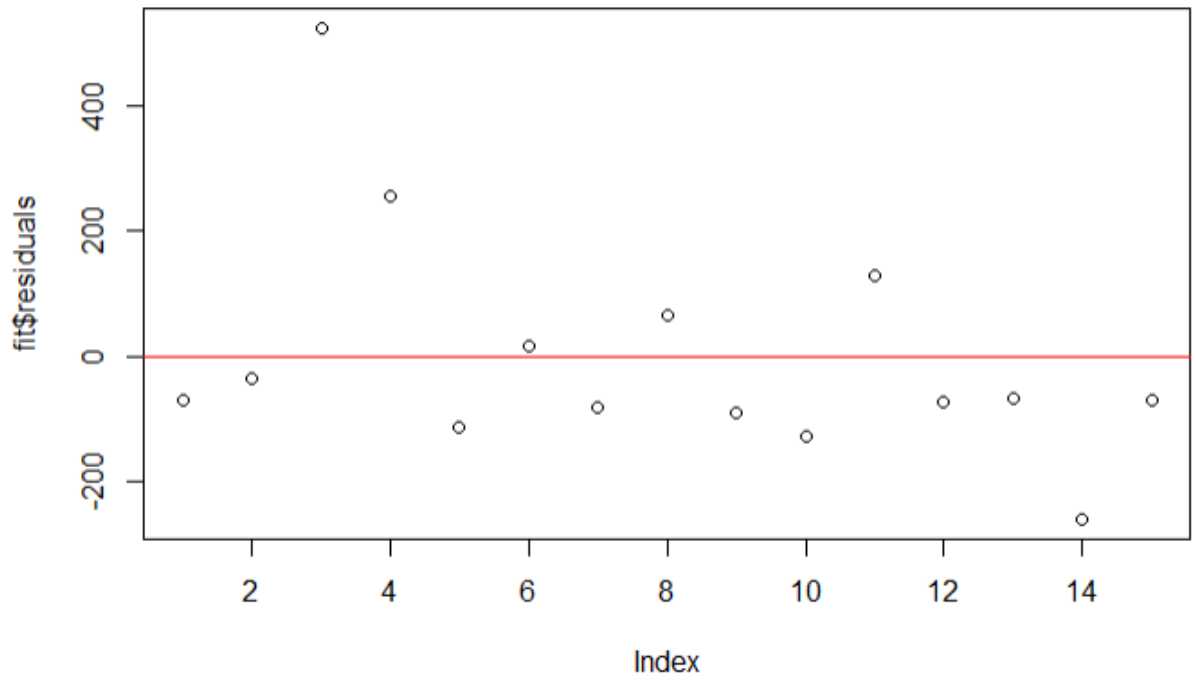
The coefficient for Recovery cases is -0.09571, which means that for every additional recovery case, there would be 0.09571 less death case number in week 5.

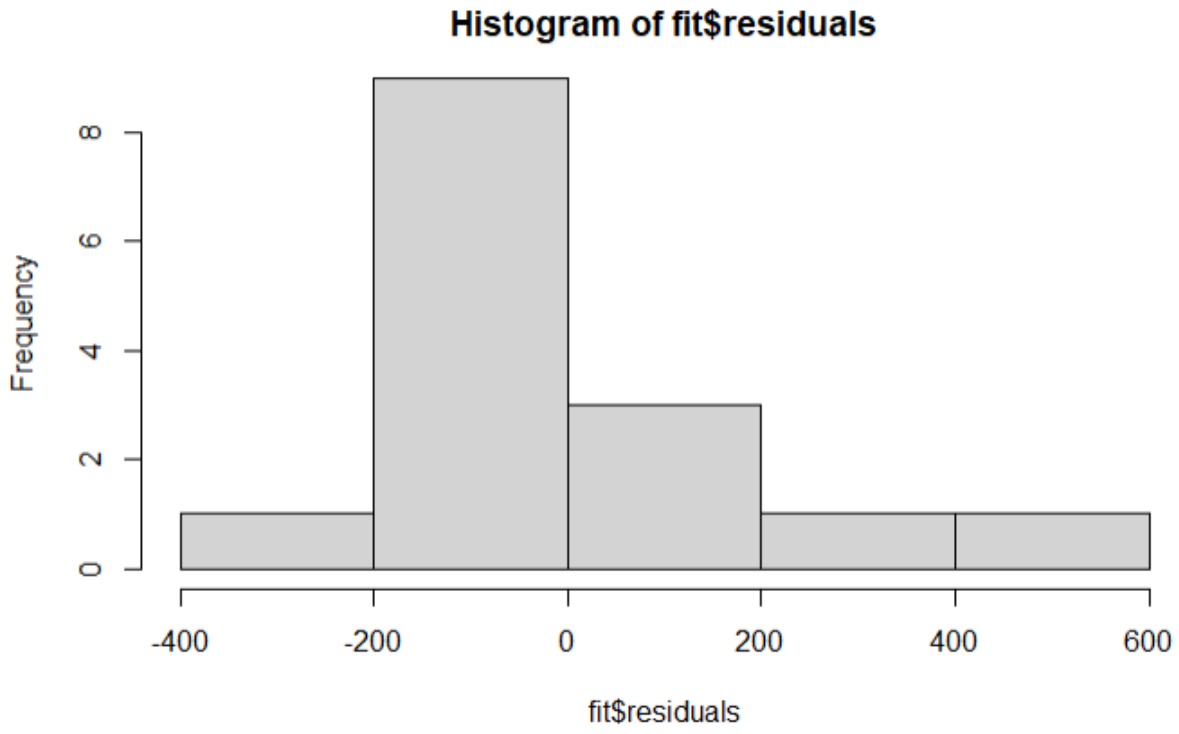
The coefficient for Week 4 deaths is 3.4975, which means that for every additional death case in week 4, there would be 3.4975 more deaths in week 5.

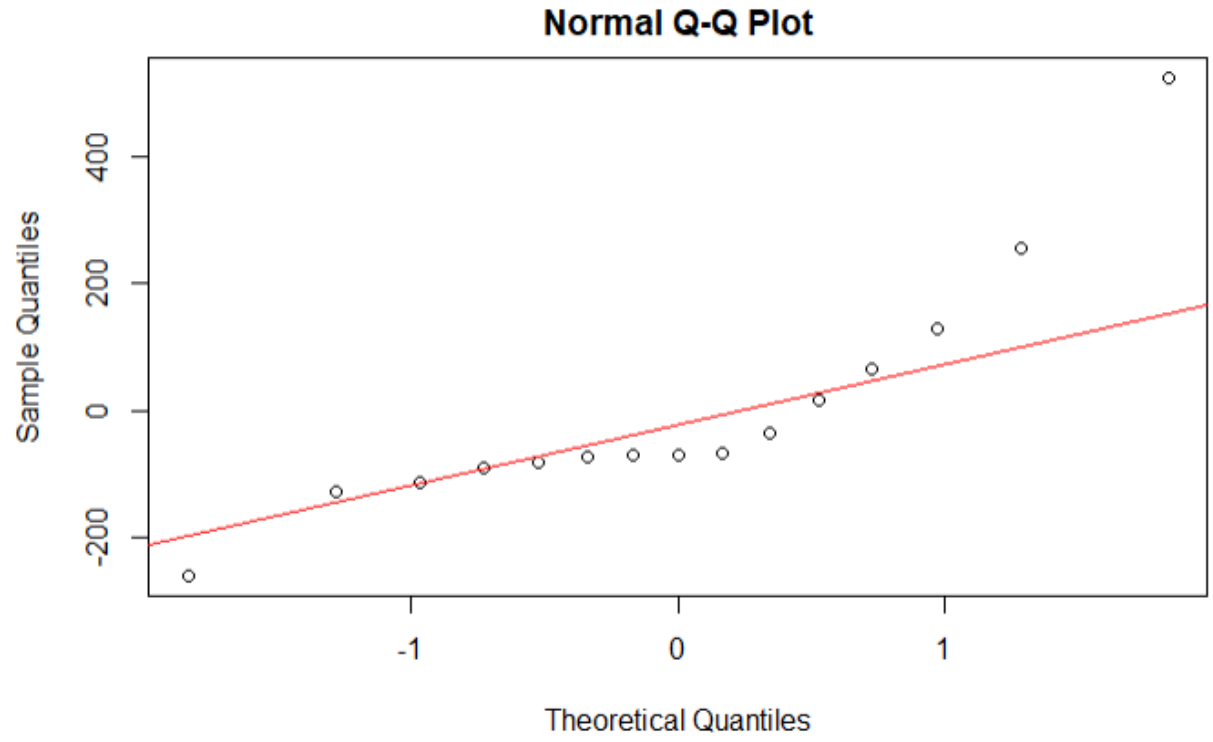
The coefficient for CFR is 33.51329, which means that if CFR increases from 0 to 1, the week 5 death number would increase by 33.51329.

Amongst all these variables, only week 4 deaths has a significant p value. This indicates that probably only "week 4 death" is really important in the prediction. The prediction interval we got for India's week 5 death is [-358.0397, 749.244], with a point estimate of 195.6021.

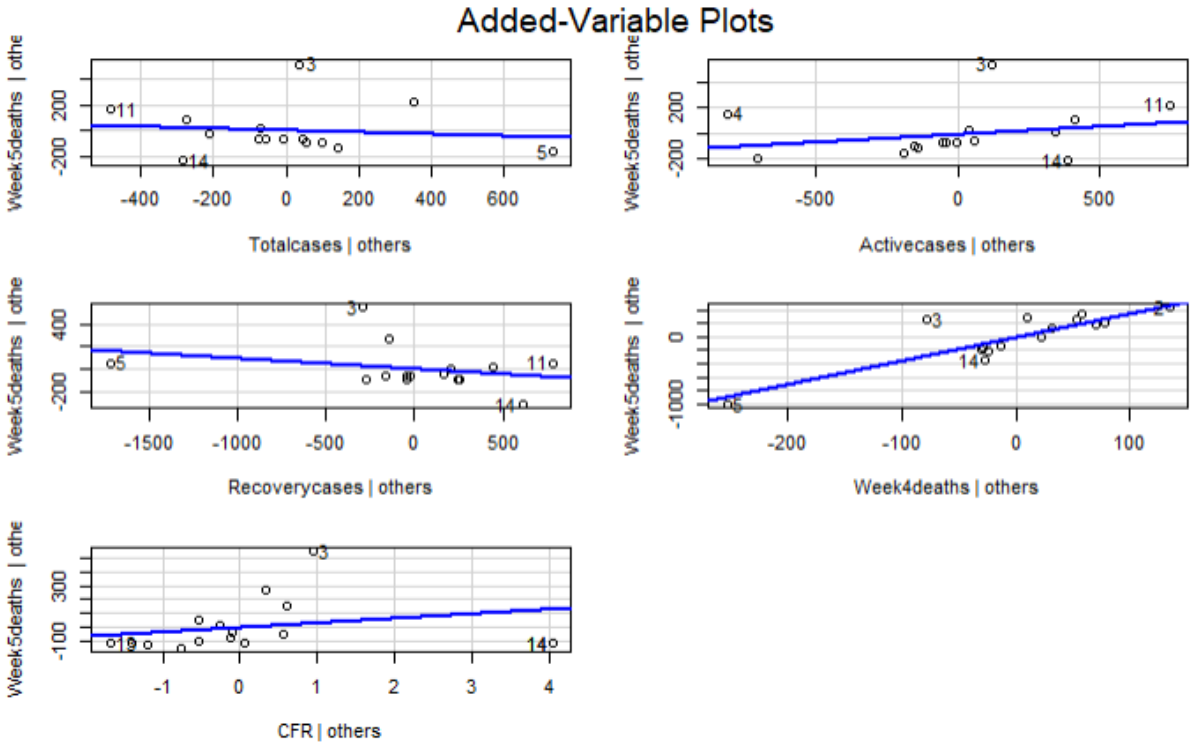
We also diagnosed the residuals to check the goodness of fitting. We found that there is a linearity in the data. And the residuals have constant variability. However, the residuals don't follow a normal distribution, which indicates that the model is not the optimal.




Histogram of fit$residuals
Frequency
0
2
4
6
8
-400
-200
0
200
400
600
fit$residuals


Normal Q-Q Plot
Sample Quantiles
-200
0
200
400
-1
0
1
Theoretical Quantiles

We also plotted the partial regression plot for the model to check the linearity between the response variable and individual explanatory variables.

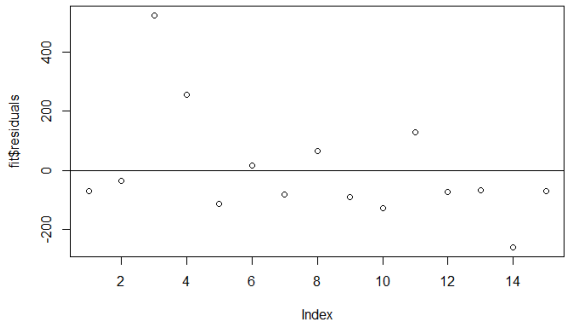


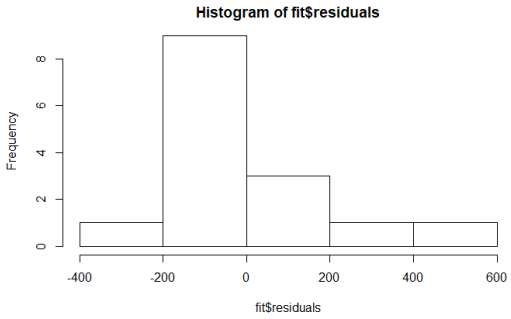
From the plot, it seems that “Week 5 Deaths” has a linear relationship with all five numeric features.

2.2 Fit on the Transformed

After we applied pairwise plots for all features in the first part, we found that “Week 5 deaths” actually has a linear relationship with all other features, but the relationships are seemingly logarithmic. Thus, for this part, we are going to apply a logarithmic transformation to the entire dataset and we may have to exclude Brazil in the process since there are 0s in the columns “Week 4 deaths” and “CFR” of Brazil data and we cannot apply logarithm to 0.

Also from the summary of fit in the first part we can see that only “Week 4 Deaths” has a significant p-value, and the p-values of other variables are all way much greater than the confidence degree. Moreover, we can plot the residuals of the fit in the first part to see whether they are normally distributed.





As we can see from the histogram of the residuals, obviously, they are not normally distributed, and most of them are below zero. Therefore, we may want to make an assumption that the dataset would fit well in log scale.

Then, we fit a linear model to the log scale of the dataset and find that it has an adjusted R^2 of 0.9038 and a R^2 of 0.9408, which indicate a pretty good fitting as well.[2]

And the following intercept and coefficients suggest that:

The intercept is -252.3691, which means that if all explanatory variables are 0, the week 5 deaths would be 10^-252.3691, which is very close to 0.

The coefficient for Total cases is 125.2276, which means that if the Total cases increase 10 times, the week 5 death number would increase 10^125.776 times.

The coefficient for Active cases is 1.6159, which means that if the Activecases increase 10 times, the week 5 death number would increase 10^1.6159 times.

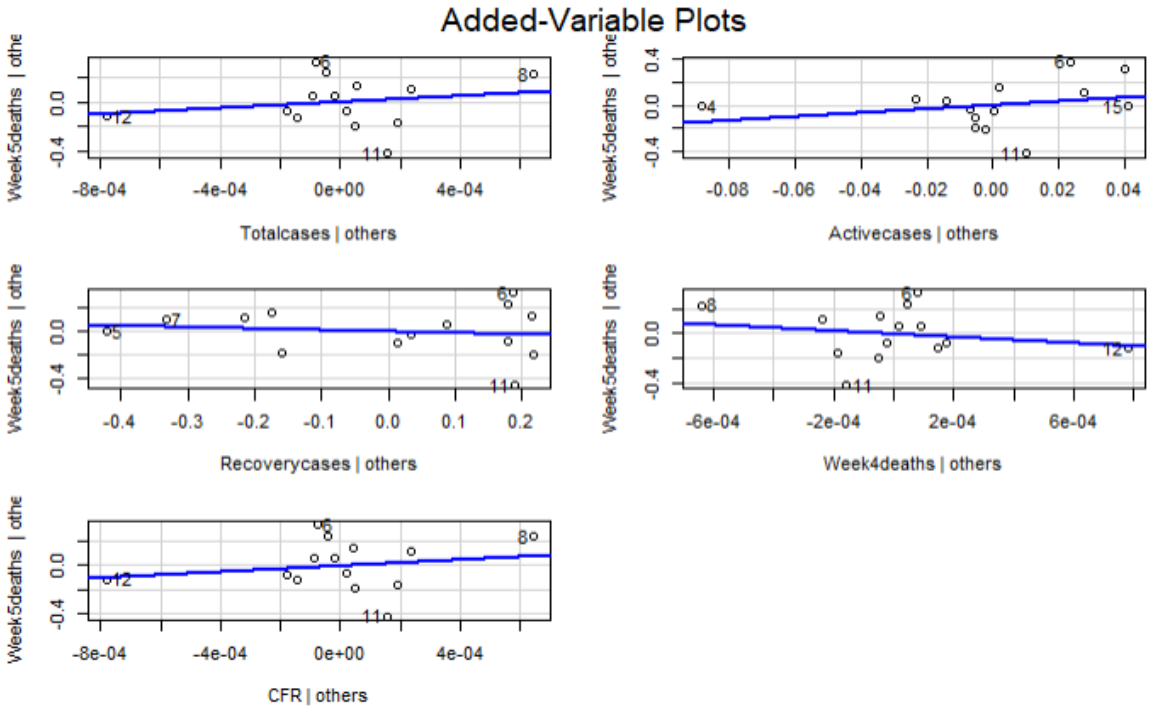
The coefficient for Recovery cases is -0.1072, which means that if the Recovery cases increase 10 times, the week 5 death number would decrease 10^0.1072 times.

The coefficient for Week 4 deaths is -125.9814, which means that if the Week 4 deaths increase 10 times, the week 5 death number would decrease 10^125.9814 times.

The coefficient for CFR is 127.0035, which means that if CFR increases from 0 to 10, the week 5 death number would increase by 10^127.0035.

Amongst all these variables, no variable has a significant p-value. This indicates that this model may be not suitable for the dataset.

Again, we plot the partial regression plot for the model to check the linearity between the response variable and individual explanatory variables. And the result shows:



From the above figures, we can see “Week 5 deaths” does not have a clear linear relationship with all five numeric features in log scale.

From the result, we can see the number of deaths in India in week 5 has a point estimate of 93.87909 with 95% confidence interval of [21.7043, 406.0616], and with 95% prediction interval of [12.76397, 690.4813].

```
[1] "Confidence Interval"
        fit      lwr      upr
16 93.87909 21.7043 406.0616
[1] "Prediction Interval"
        fit      lwr      upr
16 93.87909 12.76397 690.4813
```

2.3 Comparison

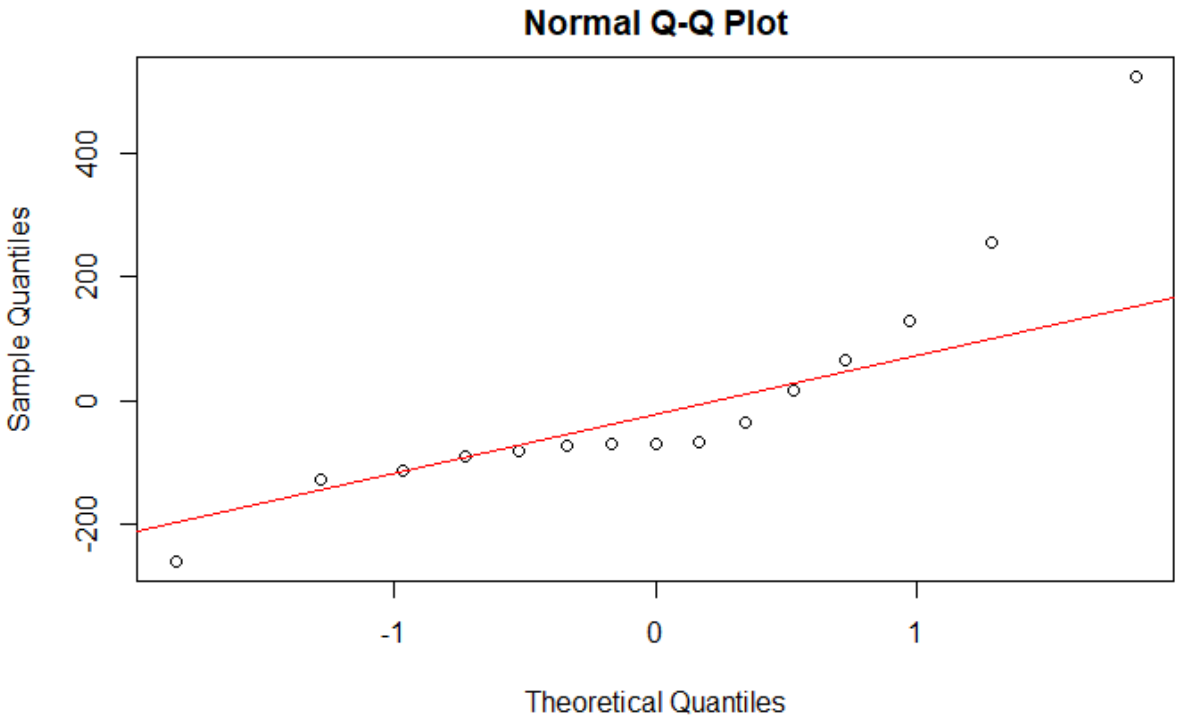
First, we need to compare two models and predictions from part1 and part2 from their predicted results. By the performance of the prediction, we could start to compare two models and find out the pros and cons of each model.

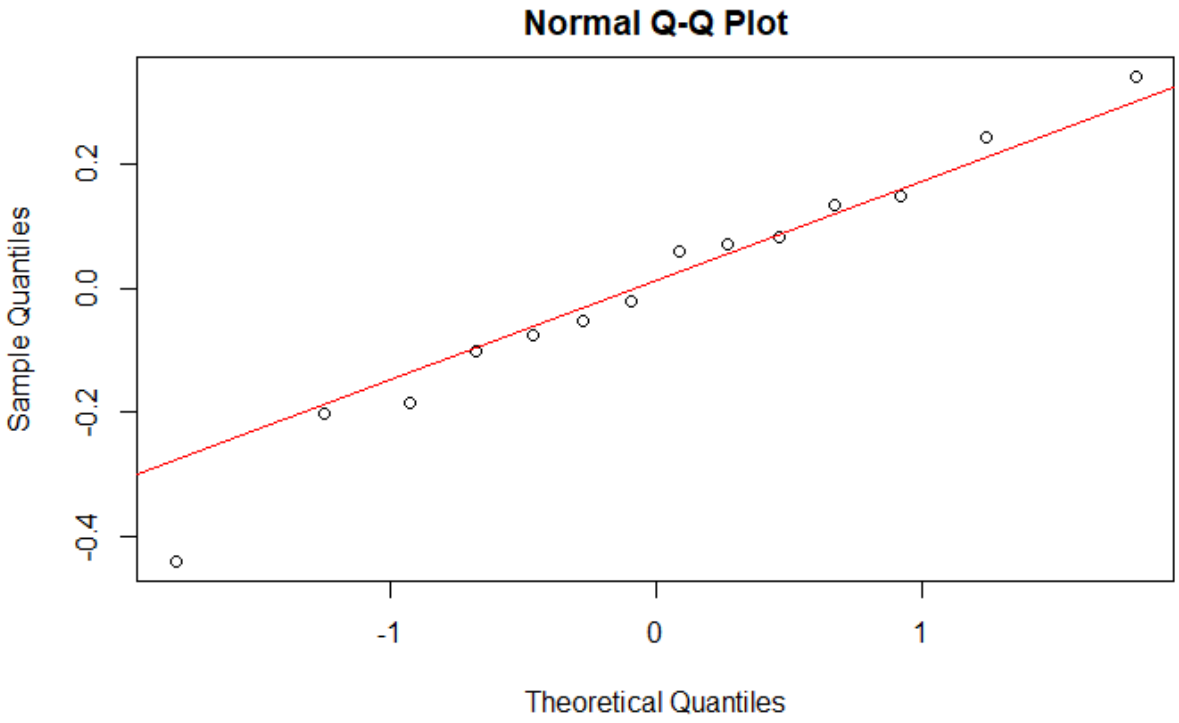
In the first part, the predicted result is that fit: 195.6021, lwr: 34.3962, upr: 356.808, which means that The point estimate is 195.6021. The 95% confidence interval is [34.3962, 356.808]. In the second part, the predicted result is shown in the graph below

```
        fit      lwr      upr
16 93.87909 21.7043 406.0616
```

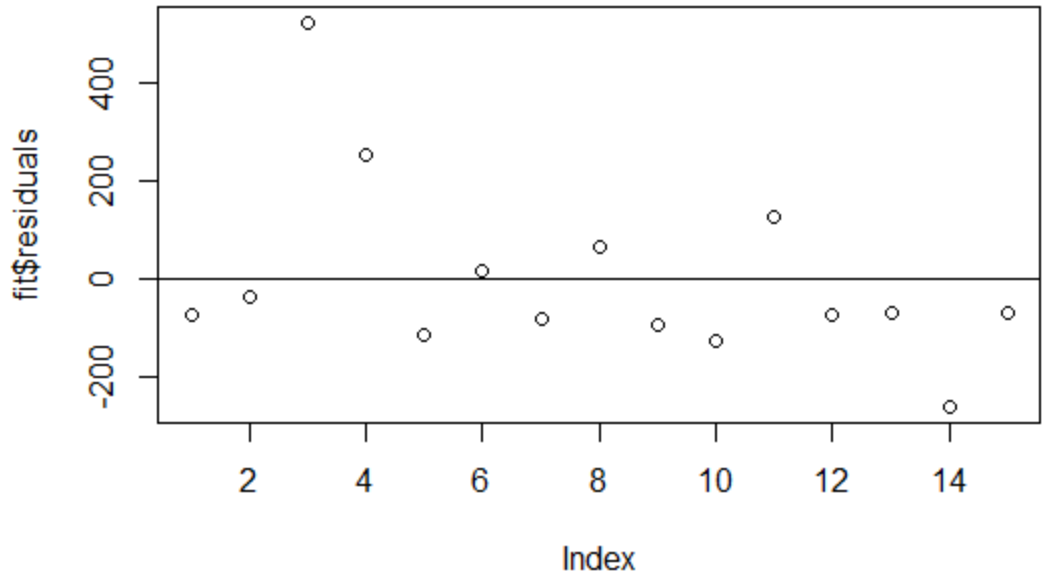
This means that the number of deaths in India in week 5 has a point estimate of 93.87909 with 95% confidence interval of [21.7043, 406.0616]. By comparing the point estimate of two parts, we could find out that 195.6021 no doubt is more accurate than 93.87909 since that the number from the project should be 211. Also, the confidence interval of the first part is more accurate than the second part. As a result, from the predicted value of two parts, we could find out that the first part is more accurate than the second part.

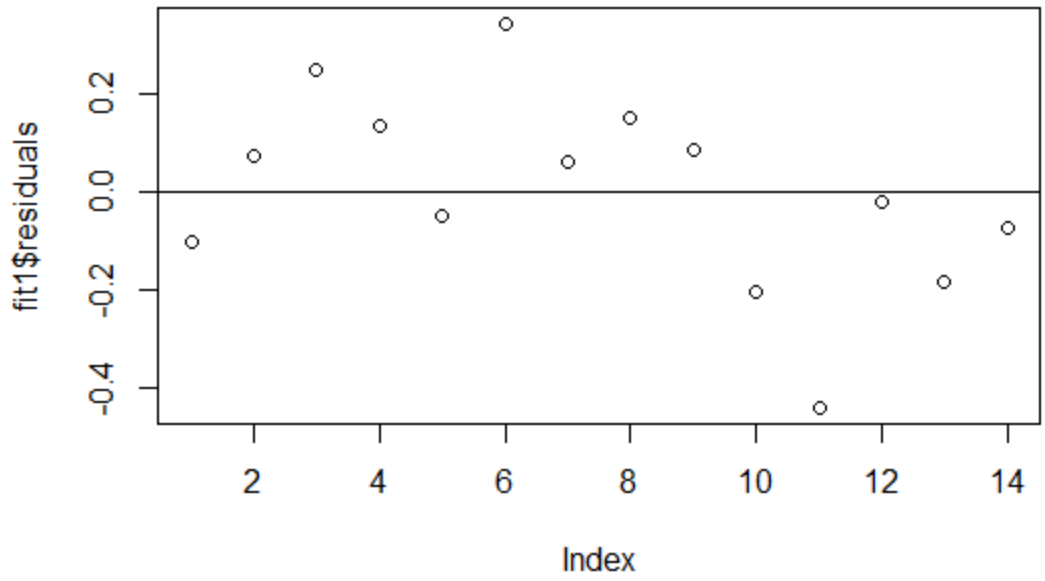
Second, after comparing the final point estimation and confidence interval, we should compare these two models from more aspects. We could compare these models in part 1 and part 2 by using residuals. There are two conditions we could check: residuals are nearly normal, residuals have constant variability. First, we could compare the Q-Q plot of two models. From the R code, we could get two graphs below.





The first graph is the Q-Q plot of model 1 and the second graph is the Q-Q plot of model 2. We could clearly recognize the difference between them. We could conclude that in part of residuals normality, the second model performs better than the first model.





The first residual plot above is the model in the first part and the second residual plot below is the model in the second part. We could find that the constant

variability is better in the second part. However, when we see the Residual standard error, we could find out that the Residual standard error of model 1 is 234.1, but the Residual standard error of model 2 is 0.2552, which means that model2 has better residual control.

Third, from R squared and Adjusted R-squared value, we could compare the performance of each model. In the first model in the first part, we find out that Multiple R-squared: 0.9807, Adjusted R-squared: 0.9701. In the second model in the second part, we find out that Multiple R-squared: 0.9408, Adjusted R-squared: 0.9038. From the data above, we could easily conclude that the first model in the first part has better fitting than the second model in the second part.
0.9807 > 0.9408; 0.9701 > 0.9038. We could say that the first model has a better fitting condition than the second model.

2.4

Regularization

In this part, we will utilize ridge regression to apply regularization in order to achieve a lesser variance with the tested data and restrict the influence of predictor variables over the output variable by compressing their coefficients.

3. Conclusion

From all the works we have done above, we may find it pretty hard to come up with a perfect model that can predict the number of deaths in India very well, but we do find that there are linear relationships lie between the week 5 deaths and all other features from the work in the first part and predict that the casualty should be around 196 in the fifth week. And applying logarithmic transformation to the dataset at part two might not help too much to predict the accurate number of deaths in week 5 in India since the fit gives the value of 94, which deviates too much from the number 211 that is predicted by the project.

Although using the linear regression model to predict the Week 5 deaths in India might help the government to develop more efficient methods to prevent the spread of the virus and better prepare for this disaster, we still have some data limitations in the project that might affect the accuracy of our prediction and therefore influence the whole outcome. For example, according to the original research, the recovery numbers from the US. was missing and the value was imputed using the strong correlation between the total cases and the recovery numbers, which might contain some errors and affect the accuracy. Apart from that, we may find that the data categories in our dataset are very limited since it does not include any information about the remedy that each country had already taken, and given the data of only 15 particular countries might also cause the data to be less representative and our prediction to be inaccurate. Thus, we

still have a lot of work to do when trying to create a perfect model to predict the number of deaths in the future.

4. Appendix

1. The model fitted on the original dataset.

```
Call:
lm(formula = Week5deaths ~ Totalcases + Activecases + Recoverycases +
    Week4deaths + CFR, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-259.52  -86.60  -69.72   41.31  524.50

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.42474  115.00886   0.734 0.481590
Totalcases   -0.06999    0.21816  -0.321 0.755657
Activecases   0.12155    0.15538   0.782 0.454134
Recoverycases -0.09571   0.10966  -0.873 0.405463
Week4deaths   3.49750    0.70392   4.969 0.000771 ***
CFR          33.51329   46.33829   0.723 0.487907
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 234.1 on 9 degrees of freedom
Multiple R-squared:  0.9807,    Adjusted R-squared:  0.9701
F-statistic:  91.7 on 5 and 9 DF,  p-value: 1.925e-07
```

2. The model fitted on the log scale dataset.

```
Call:
lm(formula = Week5deaths ~ Totalcases + Activecases + Recoverycases +
    Week4deaths + CFR, data = train1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44036 -0.09535  0.01911  0.12092  0.34237

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -252.3691   463.5709  -0.544    0.601
Totalcases    125.2276   231.8770   0.540    0.604
Activecases     1.6159     2.1969   0.736    0.483
Recoverycases  -0.1072     0.3223  -0.332    0.748
Week4deaths  -125.9814   231.8386  -0.543    0.602
CFR           127.0035   231.8375   0.548    0.599

Residual standard error: 0.2552 on 8 degrees of freedom
Multiple R-squared:  0.9408,    Adjusted R-squared:  0.9038
F-statistic: 25.43 on 5 and 8 DF,  p-value: 0.000103
```