Math189 Final Project

Ruixuan Zhang, Zhao Jin, Yiwei Yang

1.Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is the acronym of "coronavirus disease 2019". In the past twenty years, two additional coronavirus epidemics have occurred. SARS-CoV provoked a large-scale epidemic beginning in China and involving two dozen countries with approximately 8000 cases and 800 deaths, and the MERS-CoV that began in Saudi Arabia and has approximately 2,500 cases and 800 deaths and still causes sporadic cases.

Common symptoms include fever, cough, fatigue, shortness of breath, and loss of smell and taste.[While the majority of cases result in mild symptoms, some progress to acute respiratory distress syndrome (ARDS) likely precipitated by a cytokine storm multi-organ failure, septic shock, and blood clots. The time from exposure to onset of symptoms is typically around five days but may range from two to fourteen days.The virus is primarily spread between people during close contact,most often via small droplets produced by coughing, sneezing, and talking.The droplets usually fall to the ground or onto surfaces rather than travelling through air over long distances. Less commonly, people may become infected by touching a contaminated surface and then touching their face. It is most contagious during the first three days after the onset of symptoms, although spread is possible before symptoms appear, and from people who do not show symptoms.

Due to COVID-19's super high infectiousness, it has quickly spread globally. The potential for these viruses to grow to become a pandemic worldwide seems to be a serious public health risk. In a meeting on January 30, 2020, per the International Health Regulations (IHR, 2005), the outbreak was declared by the WHO a Public Health Emergency of International Concern (PHEIC) as it had spread to 18 countries with four countries reporting human-to-human transmission.Concerning COVID-19, the WHO raised the threat to the CoV epidemic to the "very high" level, on February 28, 2020.On March 11, as the number of COVID-19 cases outside China has increased 13 times and the number of countries involved has tripled with more than 118,000 cases in 114 countries and over 4,000 deaths, WHO declared the COVID-19 a pandemic.

In this report, our main goal is to find the best model that balances explanatory power and prediction ability that fits the dataset and to make predictions according to our model.

2. Data

The dataset we used comes from Samit Ghosal’s research article. This dataset contains 6 variables (Totalcases, Activecases, Recoverycases, Week4deaths, CFR, Week5deaths) for 16 countries, while the ‘Week5deaths’ of India is missing (Since India just enter its 5th Week when the data was collected)
We also introduce two datasets from the World Bank which contain the GDP and elder percentage of all 264 countries and areas in the world from 1960 to 2018. We only adopt the data from 2018 to make better predictions.

3. Background

According to the World Health Organization (WHO), viral diseases continue to emerge and represent a serious issue to public health. In the last twenty years, several viral epidemics such as the severe acute respiratory syndrome coronavirus (SARS-CoV) in 2002 to 2003, and H1N1 influenza in 2009, have been recorded. Most recently, the Middle East respiratory syndrome coronavirus (MERS-CoV) was first identified in Saudi Arabia in 2012.

The CoVs have become the major pathogens of emerging respiratory disease outbreaks. They are a large family of single-stranded RNA viruses (+ssRNA) that can be isolated in different animal species. For reasons yet to be explained, these viruses can cross species barriers and can cause, in humans, illness ranging from the common cold to more severe diseases such as MERS and SARS. Interestingly, these latter viruses have probably originated from bats and then moved into other mammalian hosts — the Himalayan palm civet for SARS-CoV, and the dromedary camel for MERS-CoV — before jumping to humans. The dynamics of SARS-Cov-2 are currently unknown, but there is speculation that it also has an animal origin.

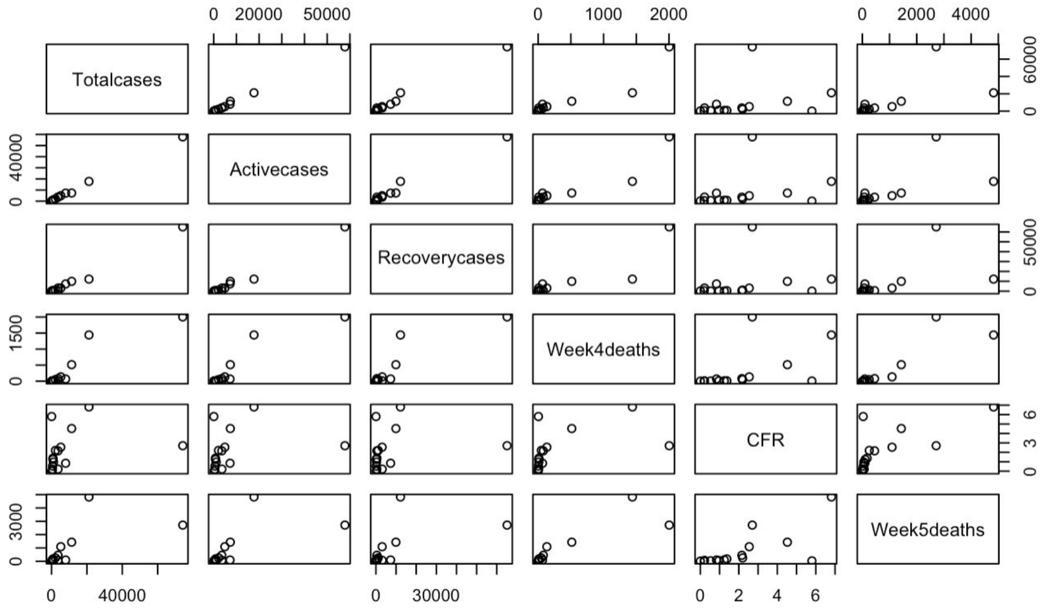
The pandemic of COVID-19 (Coronavirus disease 2019) caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) has created a havoc on the human civilization.What makes it more scary is the novel strain of the virus and the unknowns associated with it . The present strategy has been to prevent its spread by social isolation and a scientific overdrive to manufacture newer rapid diagnostic kits as well as medications . Coronavirus belongs to a family of RNA viruses within the virus family Coronaviridae, order Nidovirales . Coronaviruses are divided into three groups depending on the antigenic spikes produced by different protein structures of the virus (spike, membrane & nucleocapsid). The SARS coronavirus falls under group 2.

4. Analysis

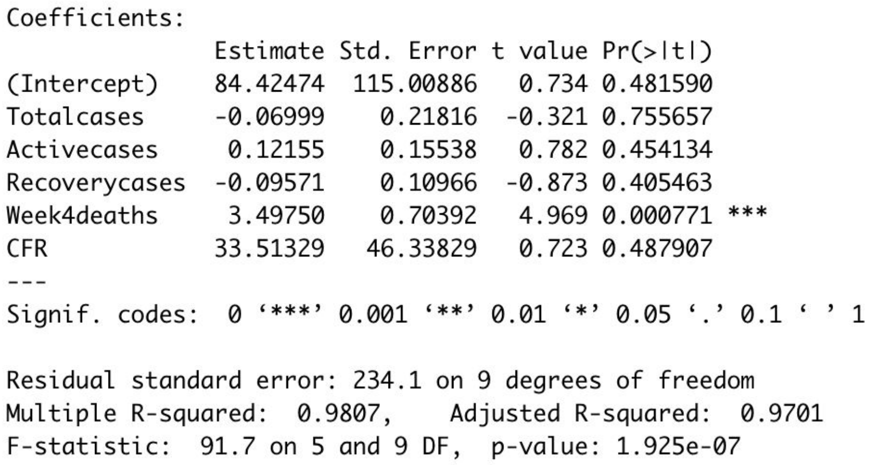
4.1 Fitting and prediction(Model1 & Model2)

Model 1(Original Data):

We want to predict week 5 death using our data. First, we examine the pairplot of our data. Although outliers presented, we can see that week 5 death is slightly positively associated with total case, active case and week 4 death, also week 5 death has a stronger positive association with CFC. Thus, we want to perform a multivariable linear regression towards our data in predicting the week 5 death.



Our first predicting model gives the following statistic.

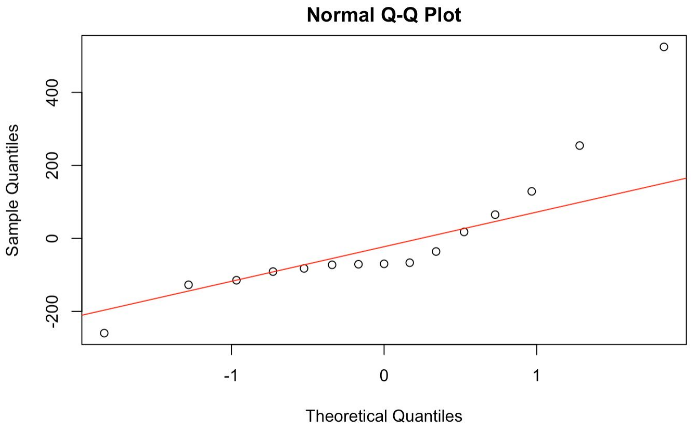


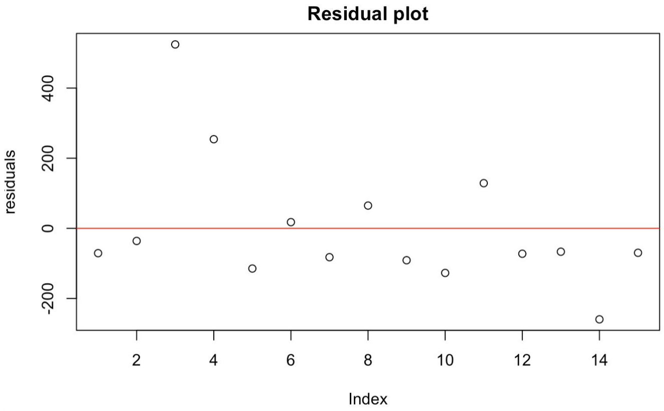
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    84.42474  115.00886   0.734 0.481590    
Totalcases     -0.06999    0.21816  -0.321 0.755657    
Activecases     0.12155    0.15538   0.782 0.454134    
Recoverycases  -0.09571    0.10966  -0.873 0.405463    
Week4deaths     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
```

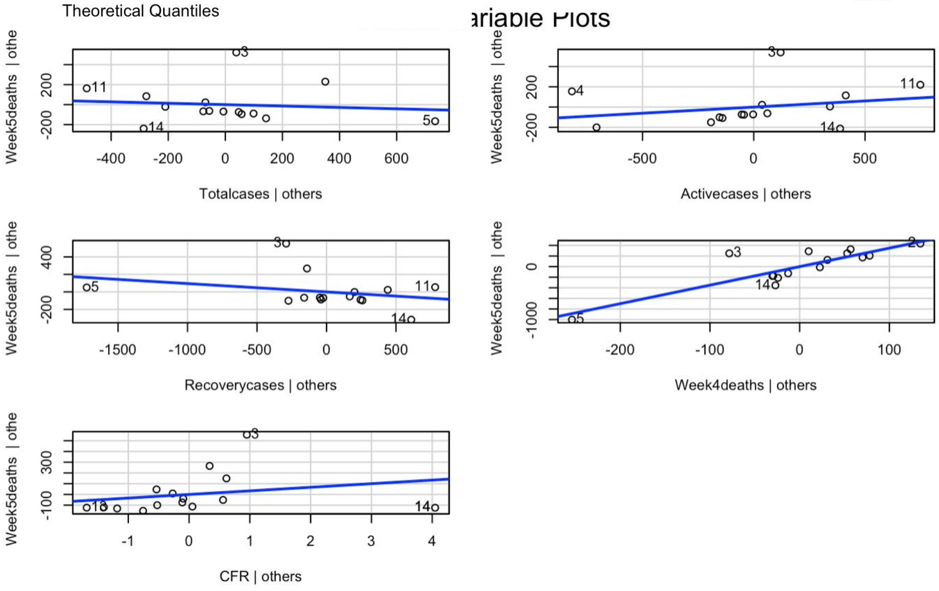
Our regression model is generally good with r-square of 0.98 and adjusted r-square of 0.97.

By examining the coefficient, we found out that the week 5 death is primarily determined by CFR and week 4 death. CFR is the Total deaths and case fatality rates which combines the info from infected cases, active cases and recovery numbers. Thus, it’s reasonable that CFR and week 4 death are important in determining the week 5 death. Examining closely on the coefficient of the variables, we found out the slope for CFR is larger compared to other variables. This may be because the value in CFR is much smaller than the value in other columns. Also, we found it interesting that the recovery cases are negatively related to the week 5 death, which is pretty reasonable since we expected less death for the country that more people have been recovered. Moreover, we found out the total case resulted in a negative association with the week 5 death. We doubt this result since the p value for this variable is large and we tend to expect more death for the country with more total cases.

Next, we want to check our qq plot, the residual plot and the leverage plot. We can see except the outliers, the model fits the data in an acceptable way.

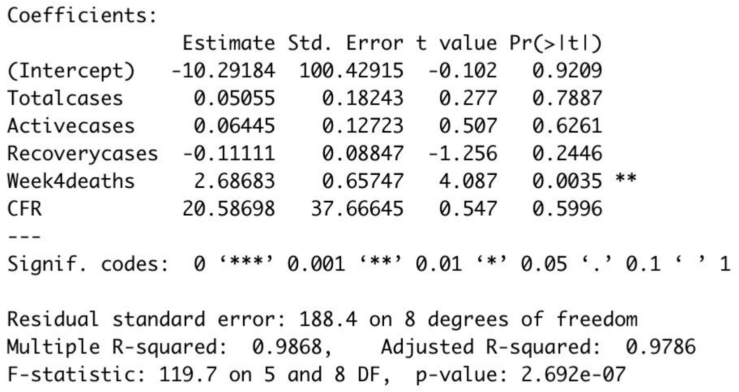




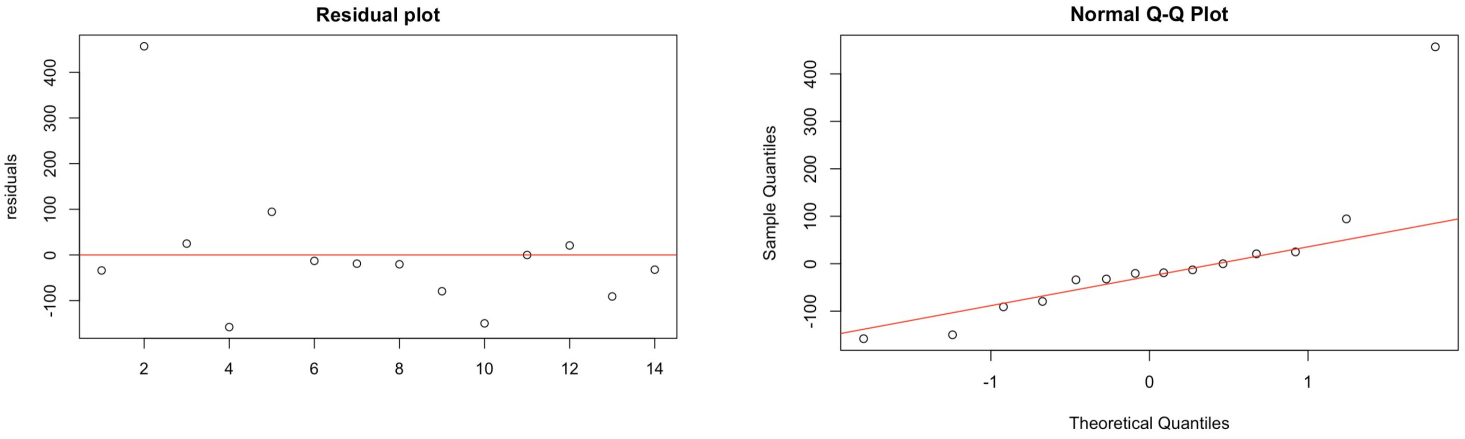


As indicated above, this model may be influenced a lot by the present of outliers. We want to take a closer look about the outliers. By investigating our dataset and the plot, we find out China and Italy served as two outliers in different ways. China has really high total cases, active cases, recovery cases and week 4 cases, whereas Italy has a regular value for those variables, but a super high week 5 death number. We want to analyze our model by dropping those outliers as well as dropping them separately.

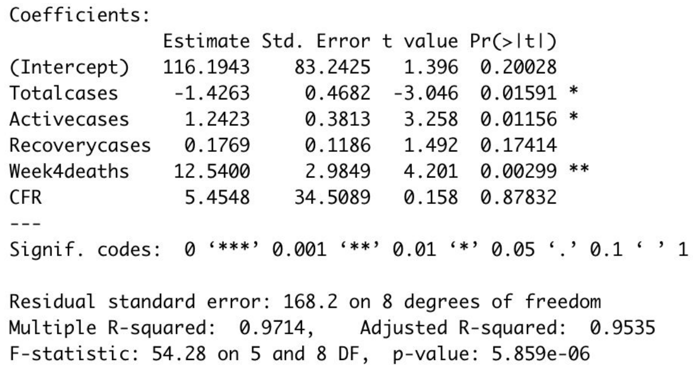
By dropping data from China, we get a better fitting model with a slight increase of R square, whereas outlier is still present in residual plot and qq plot.



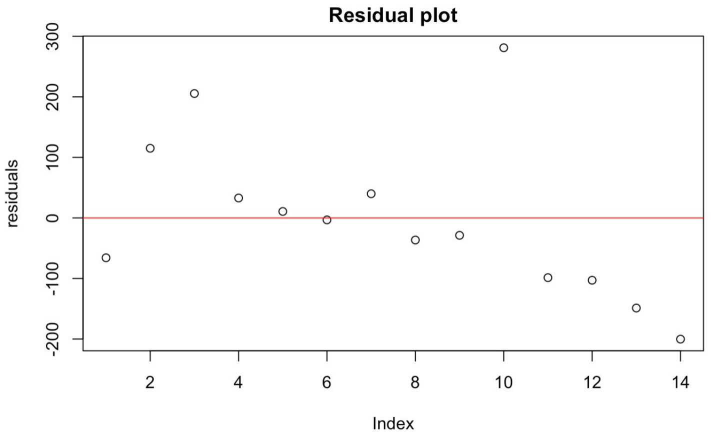
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -10.29184  100.42915  -0.102   0.9209
Totalcases      0.05055    0.18243   0.277   0.7887
Activecases     0.06445    0.12723   0.507   0.6261
Recoverycases  -0.11111    0.08847  -1.256   0.2446
Week4deaths     2.68683    0.65747   4.087   0.0035 **
CFR            20.58698   37.66645   0.547   0.5996
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 188.4 on 8 degrees of freedom
Multiple R-squared:  0.9868,    Adjusted R-squared:  0.9786
F-statistic: 119.7 on 5 and 8 DF,  p-value: 2.692e-07
```

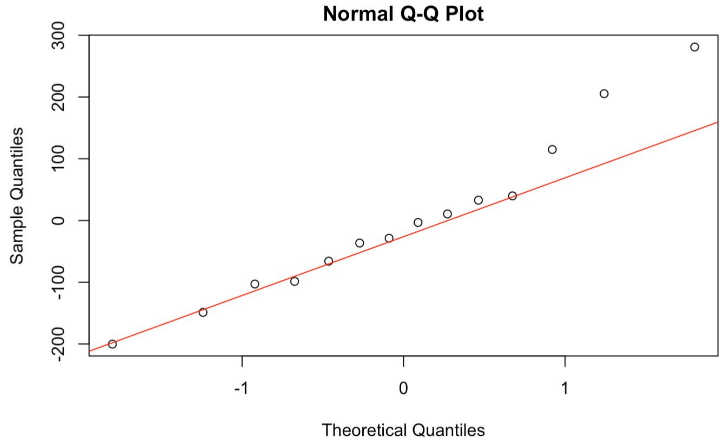


By dropping data from Italy, our fitting model has a less r square value, but a significant improvement in p values for most variables (except CFR) and a better performance in qq plot. However, if we examine the residual plot, we find out a pattern that the residual moves from positive to negative in our prediction model.

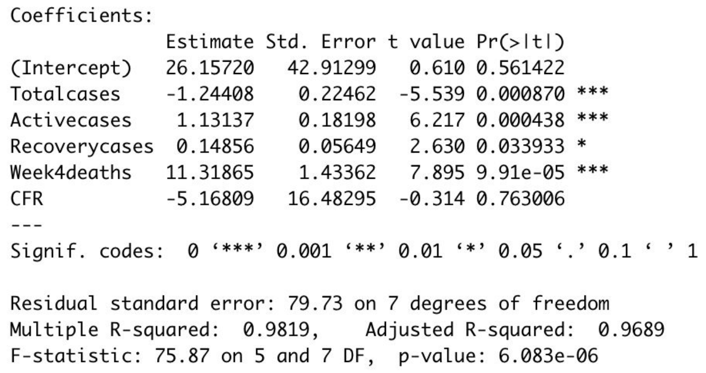


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   116.1943    83.2425   1.396  0.20028
Totalcases     -1.4263     0.4682  -3.046  0.01591 *
Activecases     1.2423     0.3813   3.258  0.01156 *
Recoverycases   0.1769     0.1186   1.492  0.17414
Week4deaths    12.5400     2.9849   4.201  0.00299 **
CFR             5.4548    34.5089   0.158  0.87832
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 168.2 on 8 degrees of freedom
Multiple R-squared:  0.9714,    Adjusted R-squared:  0.9535
F-statistic: 54.28 on 5 and 8 DF,  p-value: 5.859e-06
```

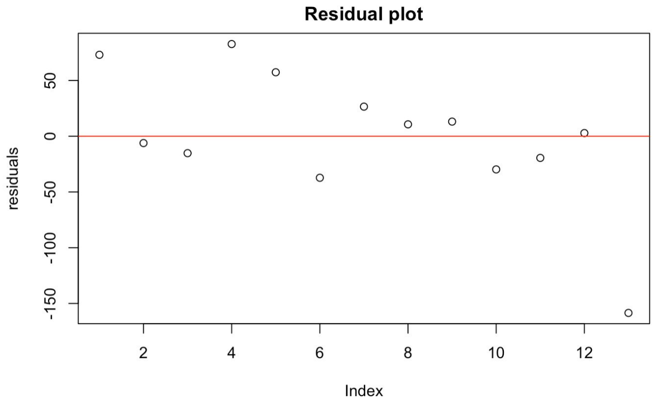


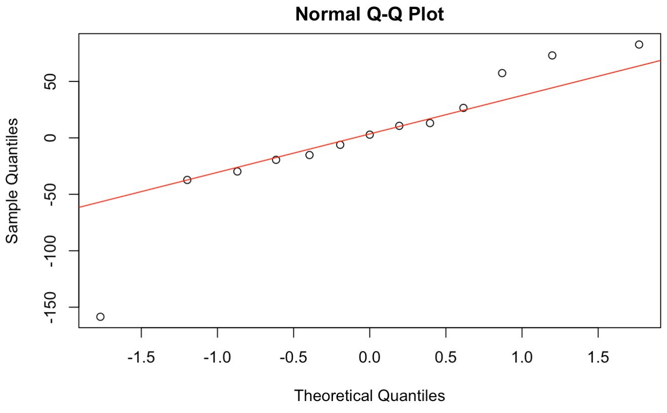


By dropping both two outliers, we get a better fitting model with a small increase in R square value, a significant improvement in p value for all variables and a better performance in residual plot and qq plot. Also, we noticed that the slope for CFR changed from positive 30 to -5, which means those outliers, especially the data from Italy, influenced significantly in our fitting model.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   26.15720   42.91299   0.610 0.561422    
Totalcases    -1.24408    0.22462  -5.539 0.000870 ***
Activecases    1.13137    0.18198   6.217 0.000438 ***
Recoverycases  0.14856    0.05649   2.630 0.033933 *  
Week4deaths   11.31865    1.43362   7.895 9.91e-05 ***
CFR           -5.16809   16.48295  -0.314 0.763006    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.73 on 7 degrees of freedom
Multiple R-squared:  0.9819,	Adjusted R-squared:  0.9689 
F-statistic: 75.87 on 5 and 7 DF,  p-value: 6.083e-06
```

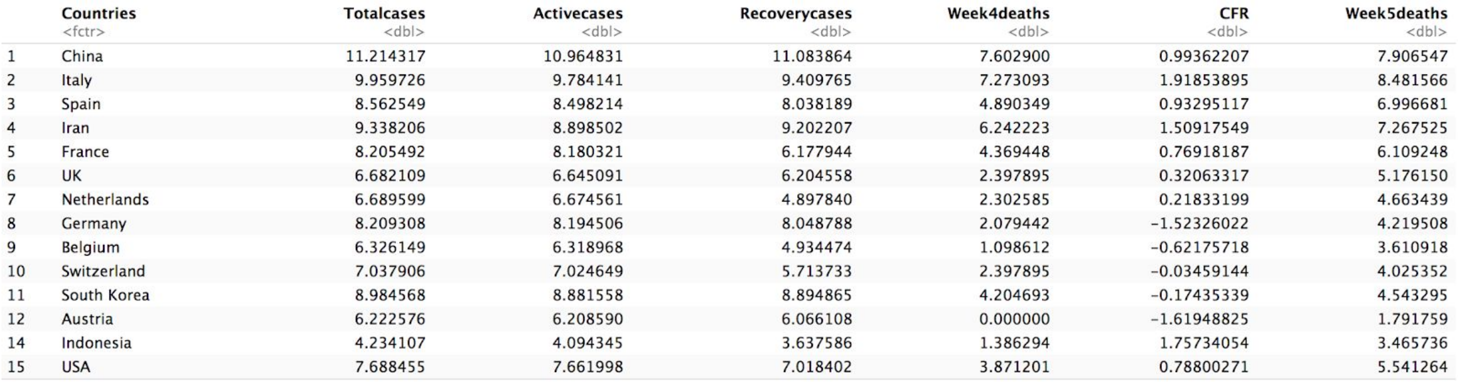




Still, we want a prediction for India’s week 5 death from our original multivariable linear regression model.
We predict the week 5 death for India is 195.6021.
95% confidence interval for mean of predictions is [34.3962,356.808].
95% prediction interval is [-358.039,749.244].

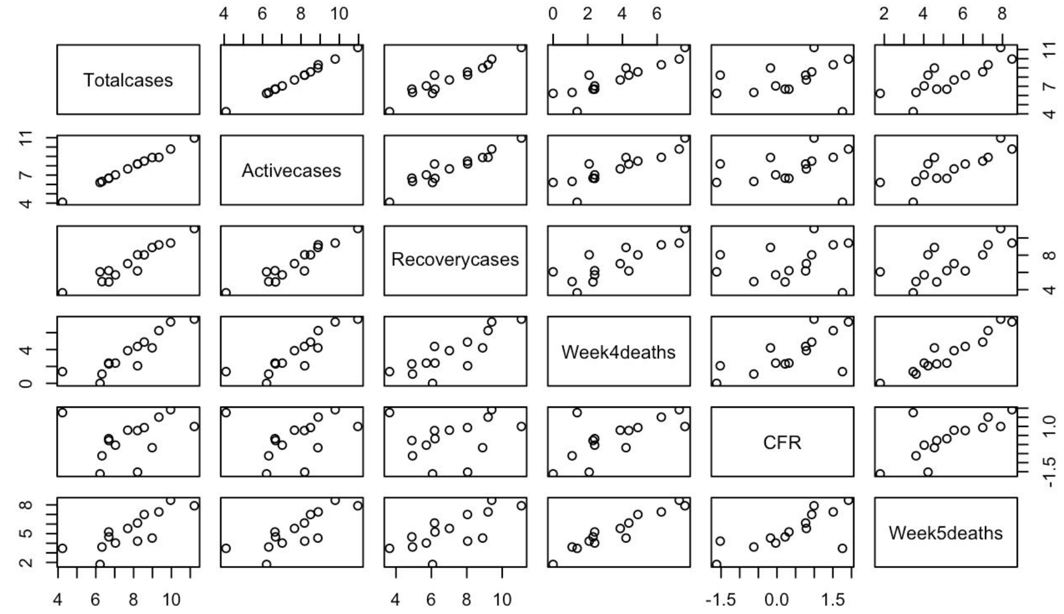
Model 2(logarithmic form of data):

As we can see in the pair graph in the first part, the patterns between each pair of variables are not strictly linear. Also, the residual plot shows extreme outliers, which means the linear model is very sensitive to outliers. These extreme outliers are influential and may cause errors to our linear model. Therefore, we revise the model by taking a logarithmic transformation of the entire dataset to reduce the effect of some influential skewed data(outliers). But before we apply the logarithmic transformation of the dataset, we need first eliminate Brazil's data. That is because 'Week4deaths' and 'CFR' are both 0. If we take a log of them, the entries will become '-inf' and cause errors.

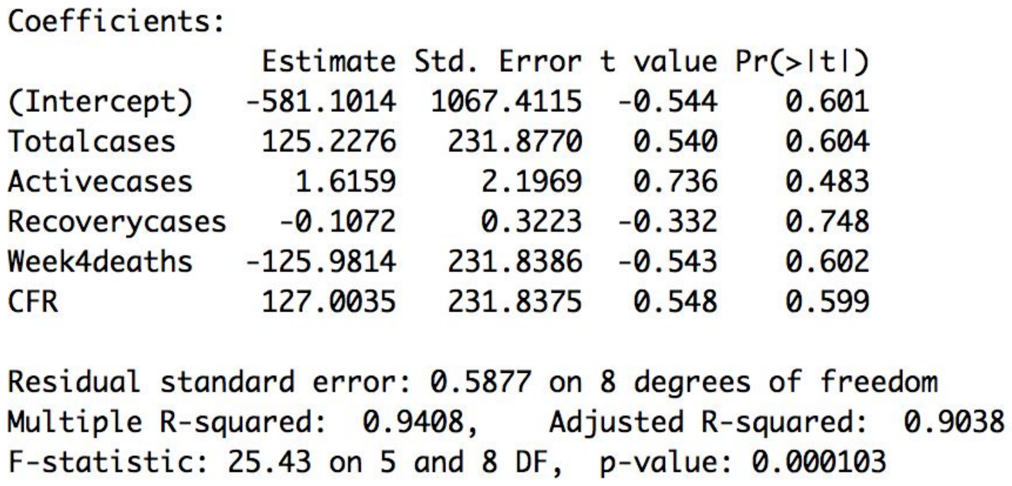


	Countries <fctr>	Totalcases <dbl>	Activecases <dbl>	Recoverycases <dbl>	Week4deaths <dbl>	CFR <dbl>	Week5deaths <dbl>
1	China	11.214317	10.964831	11.083864	7.602900	0.99362207	7.906547
2	Italy	9.959726	9.784141	9.409765	7.273093	1.91853895	8.481566
3	Spain	8.562549	8.498214	8.038189	4.890349	0.93295117	6.996681
4	Iran	9.338206	8.898502	9.202207	6.242223	1.50917549	7.267525
5	France	8.205492	8.180321	6.177944	4.369448	0.76918187	6.109248
6	UK	6.682109	6.645091	6.204558	2.397895	0.32063317	5.176150
7	Netherlands	6.689599	6.674561	4.897840	2.302585	0.21833199	4.663439
8	Germany	8.209308	8.194506	8.048788	2.079442	−1.52326022	4.219508
9	Belgium	6.326149	6.318968	4.934474	1.098612	−0.62175718	3.610918
10	Switzerland	7.037906	7.024649	5.713733	2.397895	−0.03459144	4.025352
11	South Korea	8.984568	8.881558	8.894865	4.204693	−0.17435339	4.543295
12	Austria	6.222576	6.208590	6.066108	0.000000	−1.61948825	1.791759
14	Indonesia	4.234107	4.094345	3.637586	1.386294	1.75734054	3.465736
15	USA	7.688455	7.661998	7.018402	3.871201	0.78800271	5.541264

We draw the pair graph to see the pattern between each pair variables. Compared to the original pair graph, our new graph shows a noticeable linear relationship between each pair of variables, also implying that there might be collinearities between variables.

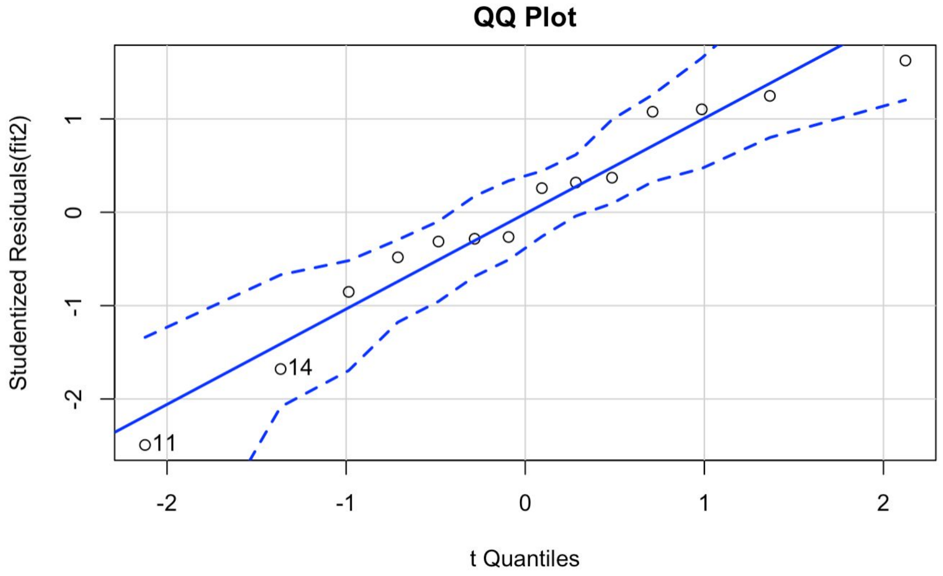


We use our new dataset to fit a new linear model. The coefficients are larger than those of the first model, which means that the response variables will be very sensitive to just a little change in the explanatory variables. Moreover, their p-values are very large, meaning that it is highly possible that we fail to reject the hypothesis that their coefficients are equal to 0. Also, it is possible that the collinearity between variables might influence the coefficients.

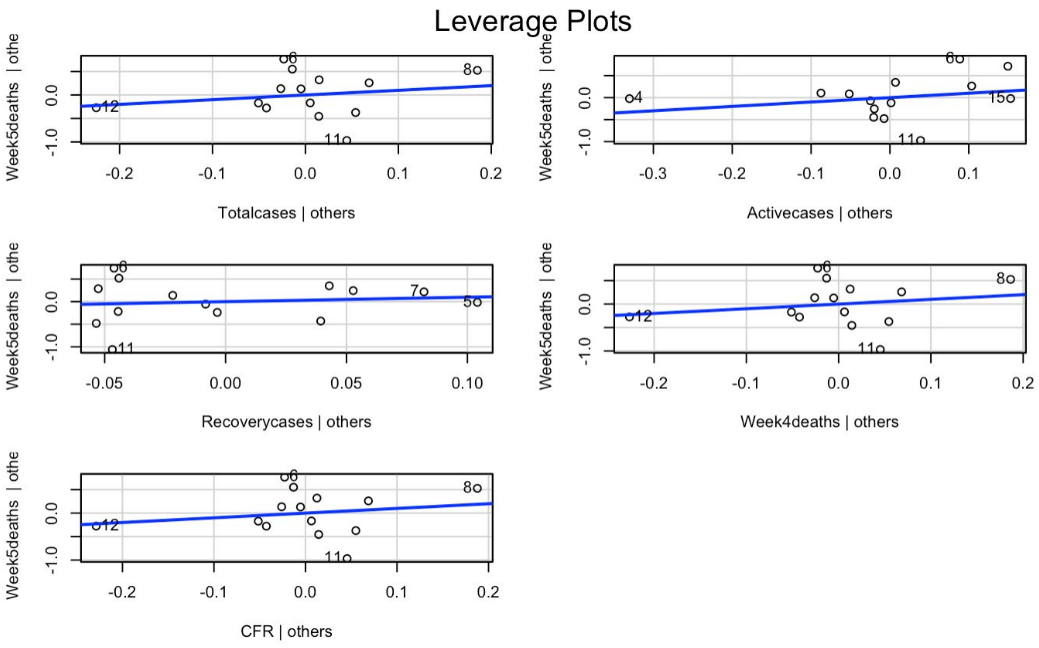


```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -581.1014  1067.4115  -0.544    0.601
Totalcases     125.2276   231.8770   0.540    0.604
Activecases      1.6159     2.1969   0.736    0.483
Recoverycases   -0.1072     0.3223  -0.332    0.748
Week4deaths   -125.9814   231.8386  -0.543    0.602
CFR            127.0035   231.8375   0.548    0.599

Residual standard error: 0.5877 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
```

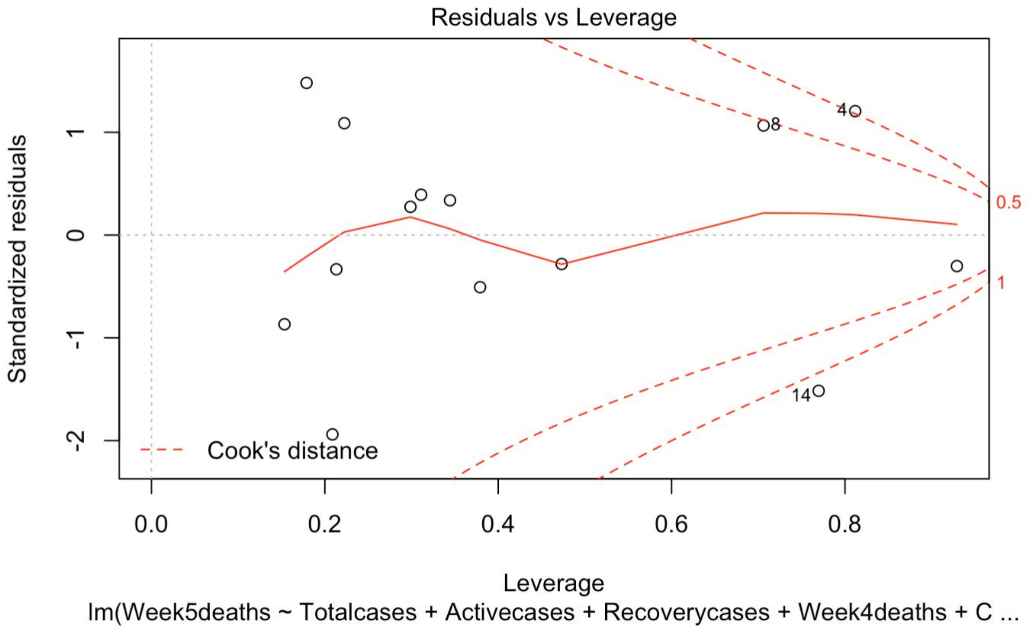
Evaluate the new model:
we can see our new model's residuals lie approximately along the line which indicates a better fit.



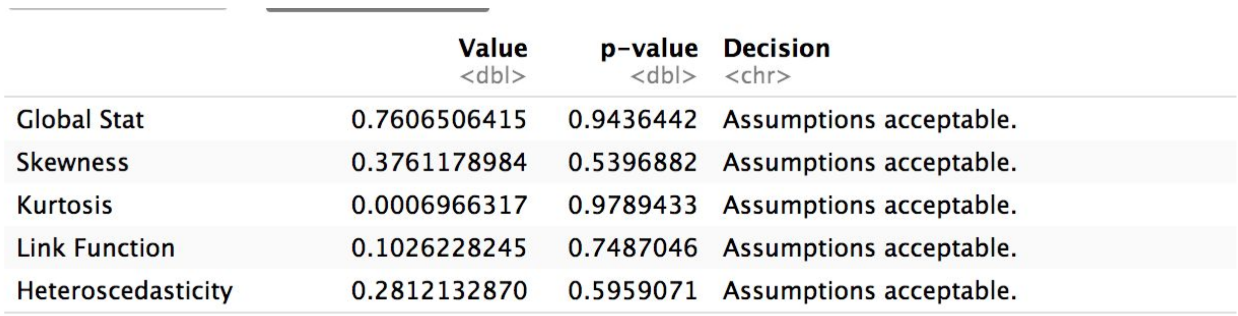
Our new leverage plot also shows that our new model is less affected by outliers.



Moreover, we have less residuals that are outside of Cook's distance, which indicate that we have less influential outliers.

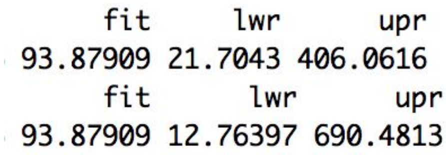


Moreover, I installed package gvlma to perform a global validation of linear model assumptions as well separate evaluations of skewness, kurtosis, and heteroscedasticity. We can see that our new model satisfies all of the assumptions.



	Value <dbl>	p-value <dbl>	Decision <chr>
Global Stat	0.7606506415	0.9436442	Assumptions acceptable.
Skewness	0.3761178984	0.5396882	Assumptions acceptable.
Kurtosis	0.0006966317	0.9789433	Assumptions acceptable.
Link Function	0.1026228245	0.7487046	Assumptions acceptable.
Heteroscedasticity	0.2812132870	0.5959071	Assumptions acceptable.

Now, we use our new model to give a point estimate and a prediction interval for the number of deaths in India in week 5.
We estimate the week5 death to be 93.87909 with
95% confidence interval to be [21.7043,406.0616]
95% prediction interval to be [12.76397,690.4813]



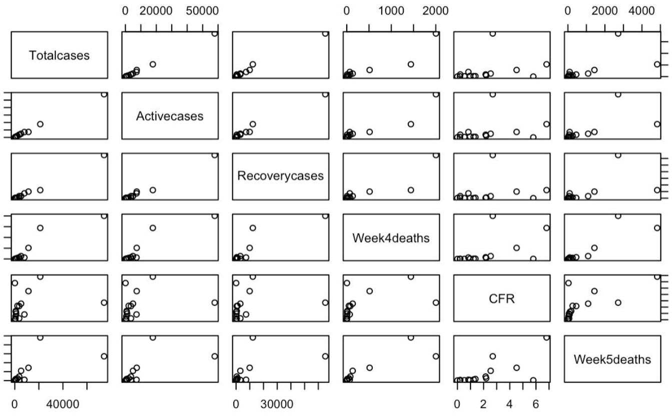
```
      fit     lwr      upr
 93.87909 21.7043 406.0616
      fit      lwr      upr
 93.87909 12.76397 690.4813
```

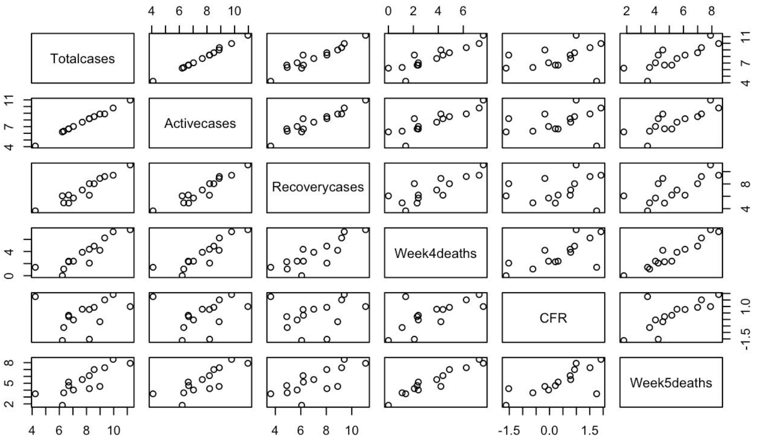
Comparison between two models:

Data: in our first model, we take all 15 countries' data into account. However, in our second model, Brazil's data was discarded as we apply logarithmic transformation of the dataset. The second model lacks one row of data.

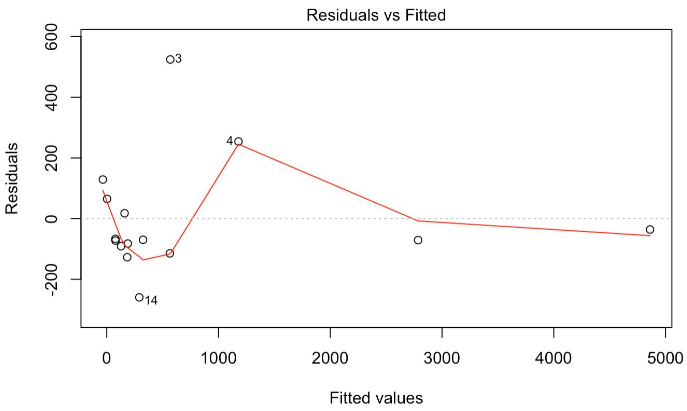
Coefficients and p values: the first model has coefficients with small p-values indicating that the coefficient of certain variables cannot be 0 and they are related to the response variables. However, the second model has coefficients with very large p-values implying that these variables may not fit in the regression model. Another way to explain large p-values is that there might be strong collinearity between variables.

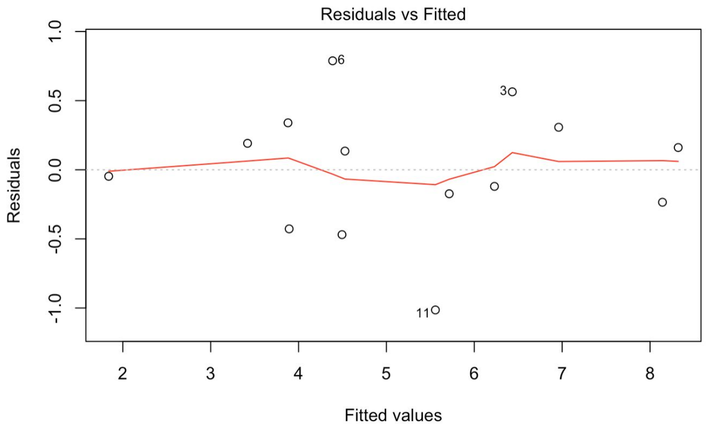
Regression Diagnostics:
By comparing two pair graphs, we can clearly see that the second one shows linear patterns between each pair of variables, may indicating collinearity between them.



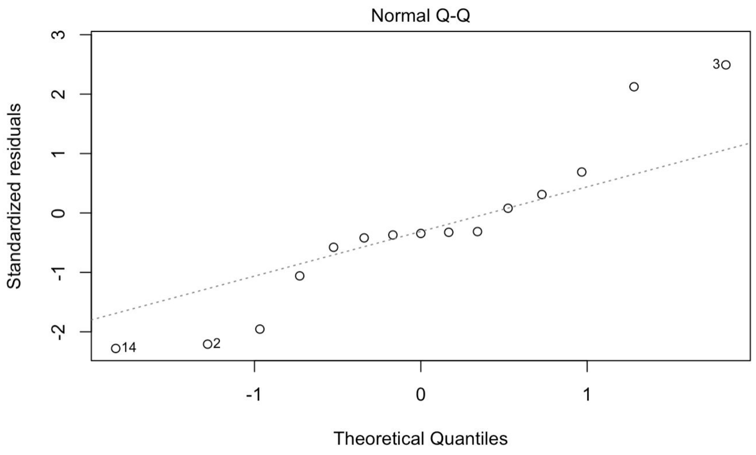


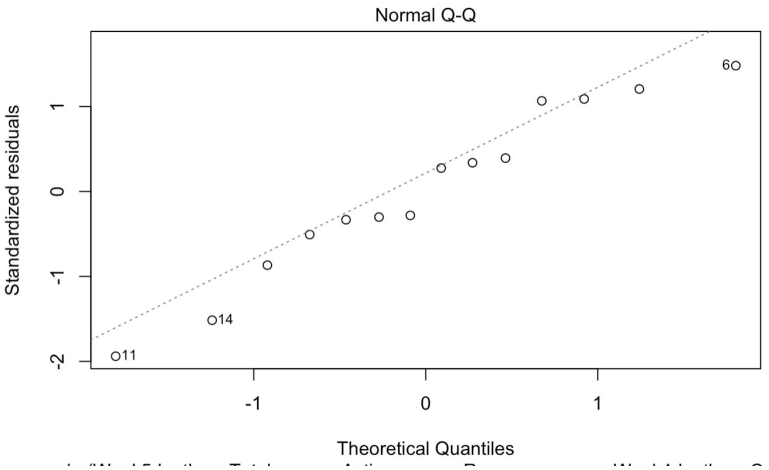
The residuals of the first plot do not spread randomly. There are extreme outliers. The second model shows a much better fit as the residuals spread randomly.



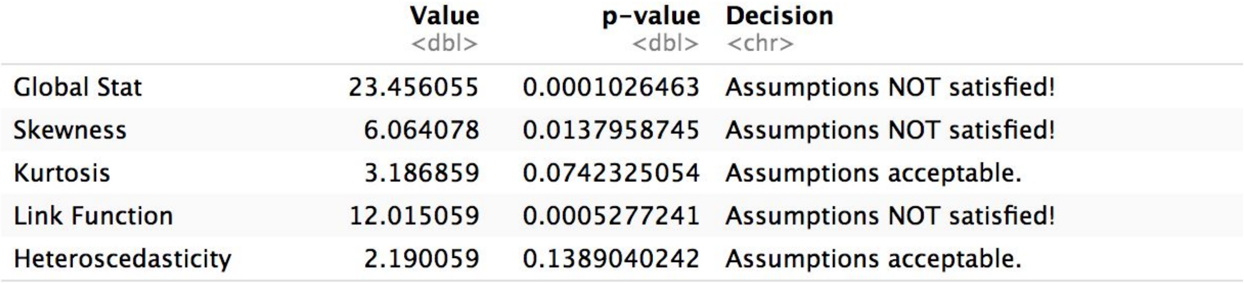


However, the first model shows better normality of the residuals then the second model.



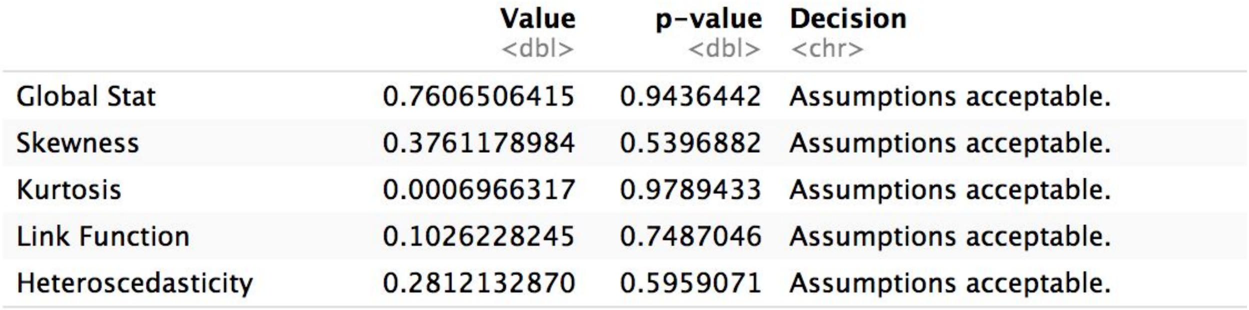


Moreover, we use gvlma to perform a global validation of linear model assumptions as well separate evaluations of skewness, kurtosis, and heteroscedasticity. The first model fails to satisfy all the evaluations but the second model does.
First:



	Value	p-value	Decision
	<dbl>	<dbl>	<chr>
Global Stat	23.456055	0.0001026463	Assumptions NOT satisfied!
Skewness	6.064078	0.0137958745	Assumptions NOT satisfied!
Kurtosis	3.186859	0.0742325054	Assumptions acceptable.
Link Function	12.015059	0.0005277241	Assumptions NOT satisfied!
Heteroscedasticity	2.190059	0.1389040242	Assumptions acceptable.

Second:



	Value	p-value	Decision
	<dbl>	<dbl>	<chr>
Global Stat	0.7606506415	0.9436442	Assumptions acceptable.
Skewness	0.3761178984	0.5396882	Assumptions acceptable.
Kurtosis	0.0006966317	0.9789433	Assumptions acceptable.
Link Function	0.1026228245	0.7487046	Assumptions acceptable.
Heteroscedasticity	0.2812132870	0.5959071	Assumptions acceptable.

Coefficients and r-square:
The second model’s coefficients are much larger than the first one. This indicates that any small change in explanatory variables may lead to large changes in our prediction. If there are any variables that are not related to our prediction. It may lead to errors in predictions.
The first model’s r-square value is larger than the second.

	Pros	Cons
The First Model	More data, better normality of residuals, small p-values, higher r-square value	Influential extreme outliers, relationships between each variable are not clear
The Second Model	patterns of residuals are spread randomly, fewer extreme outliers	Lack of one row of data, large p-values, strong collinearity between variables, lower r-square, the residuals are not spread normally, very sensitive to changes in explanatory variables(even unrelated variables)

4.2 Model evaluation & selection

We adopt model selection in this part to see whether the variable in our dataset is related.
Model1:

Step	Variables included	R-square(adjusted)
Full	Totalcases + Activecases +Recoverycases + Week4deaths + CFR	0.9701
Step1	Activecases +Recoverycases + Week4deaths + CFR	0.9727
	Totalcases + Recoverycases + Week4deaths + CFR	0.9712
	Totalcases + Activecases+Week4deaths + CFR	0.9708
	Totalcases + Activecases +Recoverycases + CFR	0.8991
	Totalcases + Activecases +Recoverycases + Week4deaths	0.9715

Step2	Recoverycases + Week4deaths + CFR	0.9701
	Activecases + Week4deaths + CFR	0.945
	Activecases +Recoverycases + CFR	0.8389
	Activecases +Recoverycases + Week4deaths	0.974
Step3	Recoverycases + Week4deaths	0.9723
	Activecases + Week4deaths	0.9491
	Activecases +Recoverycases	0.7353

Best Model: Activecases +Recoverycases + Week4deaths

Model2:

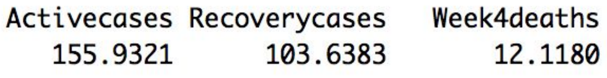
Step	Variables included	R-square(adjusted)
Full	Totalcases + Activecases +Recoverycases + Week4deaths + CFR	0.9038
Step1	Activecases +Recoverycases + Week4deaths + CFR	0.9114
	Totalcases +Recoverycases + Week4deaths + CFR	0.9078
	Totalcases + Activecases + Week4deaths + CFR	0.9133
	Totalcases + Activecases +Recoverycases + CFR	0.9113
	Totalcases + Activecases +Recoverycases + Week4death	0.9113
Step2	Activecases + Week4deaths + CFR	0.9184
	Totalcases + Week4deaths + CFR	0.9112
	Totalcases + Activecases + CFR	0.9183
	Totalcases + Activecases + Week4deaths	0.9182
Step3	Week4deaths + CFR	0.9132
	Activecases + CFR	0.9201

	Activecases + Week4deaths	0.9107
Step4	CFR	0.4717
	Activecases	0.5951

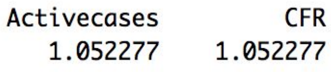
Best Model: Activecases + CFR

Check collinearity between variables

We have different model selections for our two models. The highest r-square value for the first model is 0.974, and for the second model is 0.9201. The first model has much better quality of linear regression. However, by comparing two models' variation inflation factor(vif), all of the variables in the first model show very large vif values, meaning there are collinearities between variables. The second model has appropriate vif values, smaller than 2.5, for their variables.

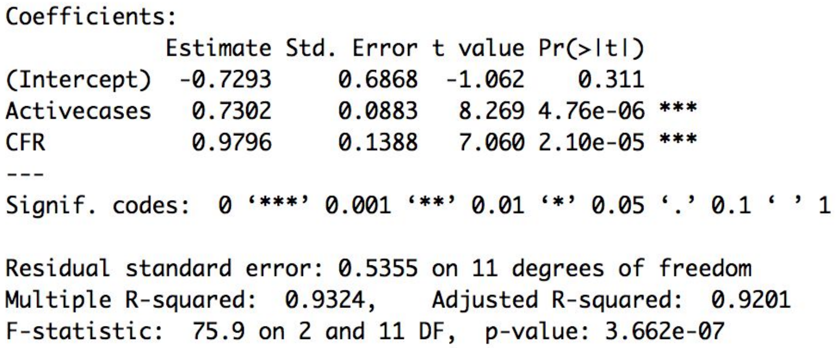


```
Activecases Recoverycases   Week4deaths
   155.9321      103.6383       12.1180
```



```
Activecases         CFR
   1.052277    1.052277
```

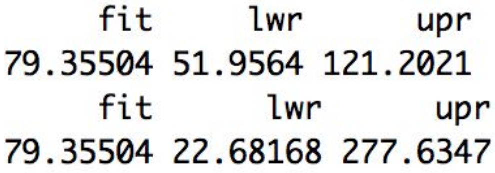
Moreover, in comparing pros and cons of these two models, we find that the second model is a better model as it decreases the effects of extreme outliers and has a randomly spread residual plot.
So our chosen model is:
log(Week5deaths) = -0.7293+0.7302*Activecases+0.9796*CFR
(or Week5deaths = exp(-0.7293+0.7302*Activecases+0.9796*CFR))



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7293     0.6868  -1.062    0.311
Activecases   0.7302     0.0883   8.269 4.76e-06 ***
CFR           0.9796     0.1388   7.060 2.10e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5355 on 11 degrees of freedom
Multiple R-squared:  0.9324,    Adjusted R-squared:  0.9201
F-statistic:  75.9 on 2 and 11 DF,  p-value: 3.662e-07
```

Although the adjusted r-square value is lower than the first model, we are sure the explanatory variables are related to the response variable ,the p-value is statistically significant, and there is no collinearity between variables.

Again, we use the new model to predict Week5deaths and its intervals. We predict the week 5 death for India to be 79.



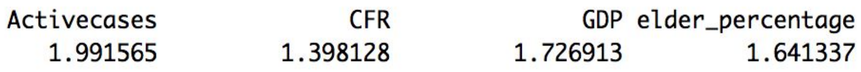
```
     fit     lwr      upr
79.35504 51.9564 121.2021
     fit      lwr      upr
79.35504 22.68168 277.6347
```

95% confidence interval for mean of predictions is [51,122].
95% confidence interval for mean of predictions is [22,278].

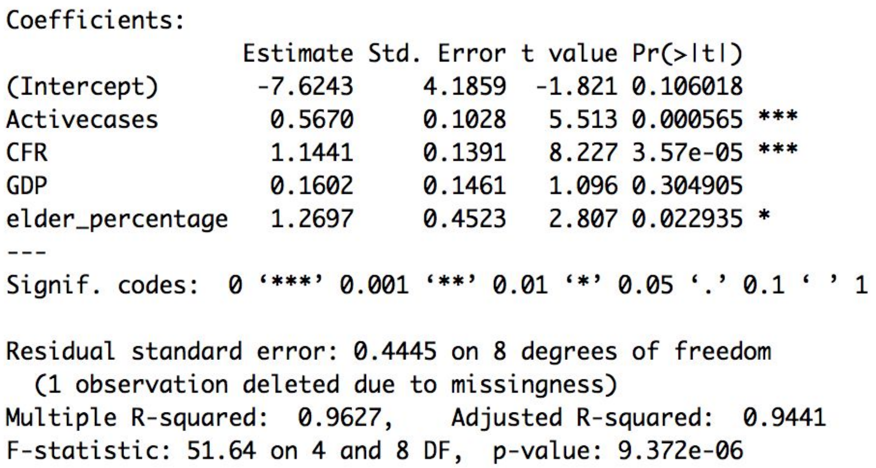
4.3 Advanced Analysis

Consider other possible response variables.

According to CDC, older adults and people who have serious underlying medical conditions might be at higher risk for severe illness from COVID-19. The reason is that COVID-19 will lead to serious complications such as respiratory problems and acute liver injury and elder people are quite sensitive due to their reduced immune function. Thus, we adopt a dataset from The World Bank which includes the percentage of the elder people (aged above 65) of each country to investigate whether the model can be improved.
Also, we make a guess that for countries which have lower GDP, they may have a higher death rate. We upload GDP and elder people's percentage into our data set and make a new model. The vif values for variables are smaller than 2.5, indicating there are no collinearities between variables.

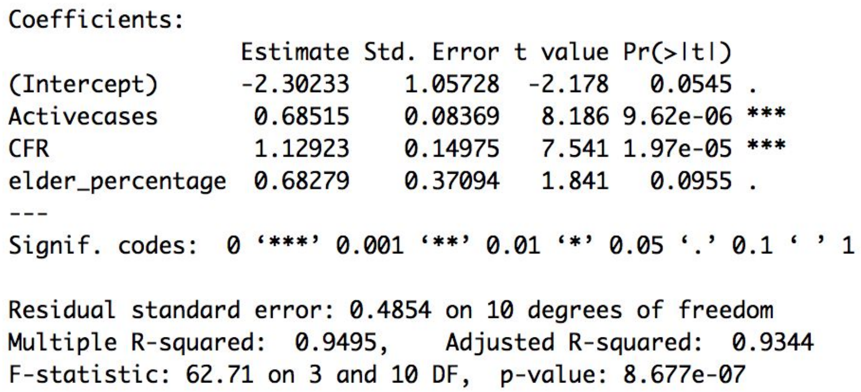


```
 Activecases              CFR              GDP elder_percentage 
    1.991565         1.398128         1.726913         1.641337 
```



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -7.6243     4.1859  -1.821 0.106018    
Activecases        0.5670     0.1028   5.513 0.000565 ***
CFR                1.1441     0.1391   8.227 3.57e-05 ***
GDP                0.1602     0.1461   1.096 0.304905    
elder_percentage   1.2697     0.4523   2.807 0.022935 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4445 on 8 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9627,	Adjusted R-squared:  0.9441 
F-statistic: 51.64 on 4 and 8 DF,  p-value: 9.372e-06
```

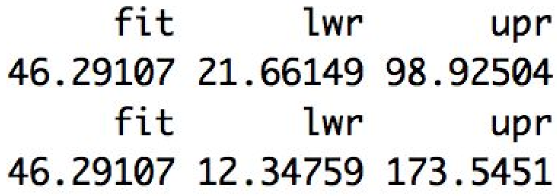
From the coefficients table above, we can infer that elder_percentage is also statistically. But p-value for GDP is high, so we delete that, and make a new model based on Activecases, CFR, and elder_percentage. All of the variables are statistically significant as they are smaller than 0.1. Moreover, adjusted R-square value improves to 0.9344 indicating a good quality of fit.



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -2.30233    1.05728  -2.178   0.0545 .  
Activecases       0.68515    0.08369   8.186 9.62e-06 ***
CFR               1.12923    0.14975   7.541 1.97e-05 ***
elder_percentage  0.68279    0.37094   1.841   0.0955 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4854 on 10 degrees of freedom
Multiple R-squared:  0.9495,	Adjusted R-squared:  0.9344 
F-statistic: 62.71 on 3 and 10 DF,  p-value: 8.677e-07
```

So our final model is :
log(week5deaths) =
2.30233+0.68515*Activecases+1.12923*CFR+0.68279*elder_percentage

We predict week 5 deaths to be 46



```
     fit      lwr      upr 
46.29107 21.66149 98.92504 
     fit      lwr      upr 
46.29107 12.34759 173.5451 
```

95% confidence interval for mean of predictions is [21,99].
95% confidence interval for mean of predictions is [12,173].

5.Discussion & Conclusion

From Section 4.1, we have developed a linear regression model from the original dataset and another linear regression model from logarithmic form of data. Our data fits both models well, which means linear regression is a good model to choose. Then we compare two models and then find that the model from logarithmic form of data is more accurate and reasonable. By applying model selection and regularization to both models in Section 4.2, we removed high-related variables in both models. Then for the log model we choose, only Active Cases and CFR becomes the predictor.

Then in Section 4.3, we introduce additional datasets to test whether GDP and elder adults percentage affect our linear prediction model. After fitting the model, we find that though there are no collinearities between two variables and the adjusted r square improved, the p-value of GDP is quite high. Thus, we only keep elder people percentage as an effective predictor.
Thus, we conclude that the week 5 death can be predict through a linear regression model and our final model is :
log(week5deaths) = 2.30233+0.68515*Activecases+1.12923*CFR+0.68279*elder_percentage
Our prediction of India's week 5 deaths is to be 46
We also want to discuss the data limitation. The dataset we used only contains the most impacted countries, and there are only 15 valid observations.. Moreover, the accuracy of the dataset we used is doubted since there is a difference between the value in this dataset and others dataset with the same columns. As a result, the accuracy of our model and prediction is highly affected by these aspects.

6. Contribution

We first brainstorm the project together. The first model is contributed to Zhao Jin.The final model and model selection is contributed to Yiwei Yang. Then the remaining part is finished by Ruixuan Zhang.

7.Appendix

7.1 Work Cited

1. Samit Ghosal, Sumit Sengupta, Milan Majumder, Binayak Sinhad. Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020). Diabetes & Metabolic Syndrome: Clinical Research & Reviews, Volume 14, Issue 4, July–August 2020, Pages 311-315https://www.sciencedirect.com/science/article/pii/S1871402120300576
2. Marco Cascella; Michael Rajnik; Arturo Cuomo; Scott C. Dulebohn; Raffaela Di Napoli. Features, Evaluation and Treatment Coronavirus (COVID-19) https://www.ncbi.nlm.nih.gov/books/NBK554776/
3. Quick R: Multiple regression https://www.statmethods.net/stats/regression.html
4. The World Bankhttps://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS