Case Study On Covid19

~ Samriddh Gupta

Introduction

The outbreak of Covid-19 is a hot topic of discussion all around the world. Covid-19 is now spread in 210 countries and is threatening the lives of millions. People all around the world are trying to help with whatever they can. they are also showing great acts of bravery and kindness to the poor and those in need. I wish to give my respect to those who have died and there to family members for there losses. In this project, I try to display the data I got from the internet about the virus in certain countries and try to find any relationship based on data and make models from it and prediction based on it. As we all know that Doctors all around the world are trying their best to keep the death toll to a minimum. The question that I wish to answer here is whether the data we got could be of help to anyone and the visualization that I made could be of use in any research. The two cases which I study were India and Italy. I chose India as it is my home country and has the second-highest population in the world, but we still have a lot fewer cases in terms of China which has the highest population. I chose Italy as my second case study because it was gaining a lot of attention from news media because of its higher number of dead people there.

Get Data

As I said earlier the data was collected from different websites, I got the Data for India from the following link:

`https://www.kaggle.com/imdevskp/covid19-corona-virus-india-dataset#complete.csv`

The data is a detailed explanation of the number of total of cases along with the recovered cases and the total number of people dead. The data also contains the states where the cases were reported and the date at which they are reported. The data also contain the longitude and latitude of that place. The data started from 30th January 2020 till 5th April 2020. The states contain data from 34 States and Union Territories in India. The data was almost ready to use but there were few things like formating the date etc which needed to be done. The data was also either group by the States or by Dates to get more meaning full results.

The date for Italy was taken from the kaggle as well and was similar to the Data I got for India. It was taken from the following link:

`https://www.kaggle.com/sudalairajkumar/covid19-in-italy`

The Data Contains a lot of pieces of information, but the ones that I used were the Total number of cases, Total People recovered and total people who have died. The data also comes with the dates for the data mention earlier and in which region they were reported. Another part of the data which I used was longitude and latitude of the data which helped me with the shiny app part of the project. The Data was ready to use but had to be selected and grouped down first. It was done similar to the Indian data. The data's date ranges from 24th February 2020 to 2nd April 2020.

The only problem with both of the data sets was that they gave us information till the 1st week of April, so It was a little bummer, but I was still able to get some good results from the data.

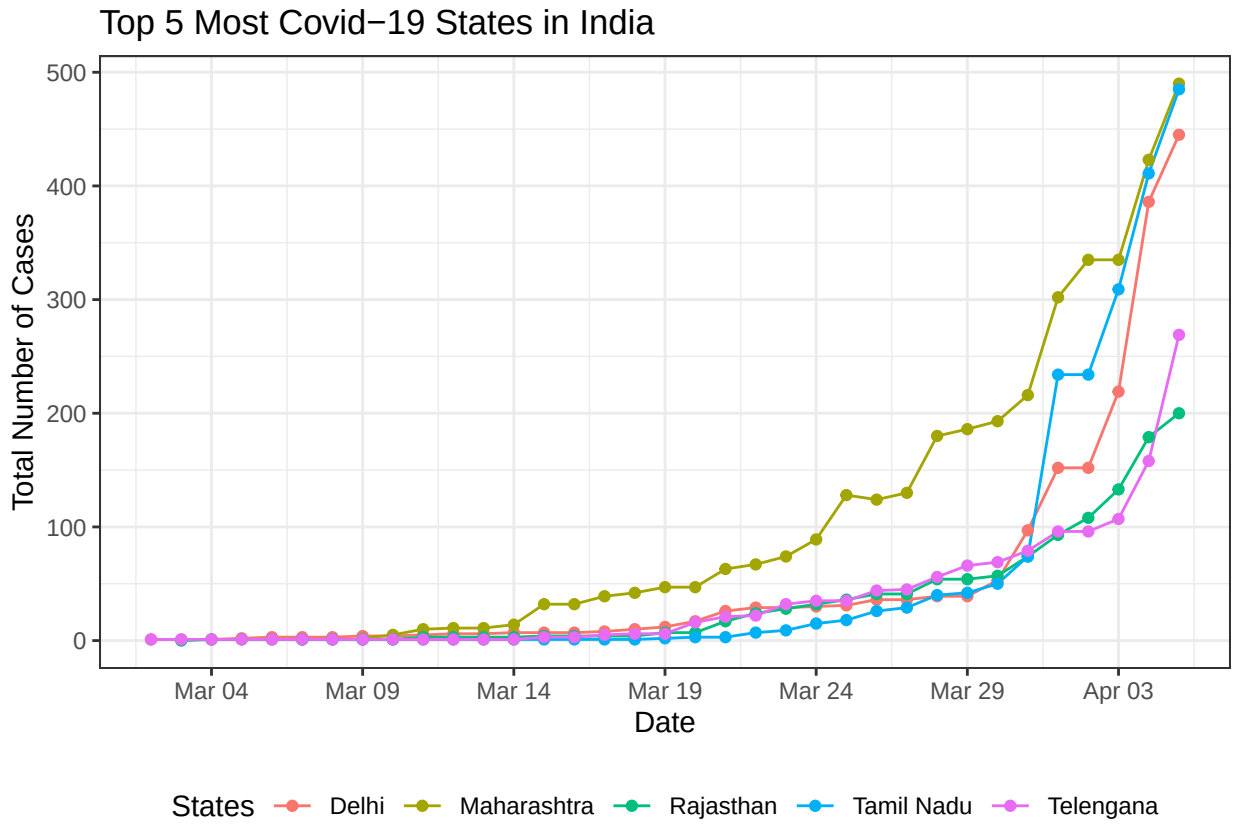
Analysis and Visualization

Most of the work I did here involves the concepts of Regression, Data Mining, and Data Analysis. The research question I look at was how does the Covid-19 Spread in different countries based on the number of days it been since the report of First cases there.

part1 Case Study: India

Selection By State

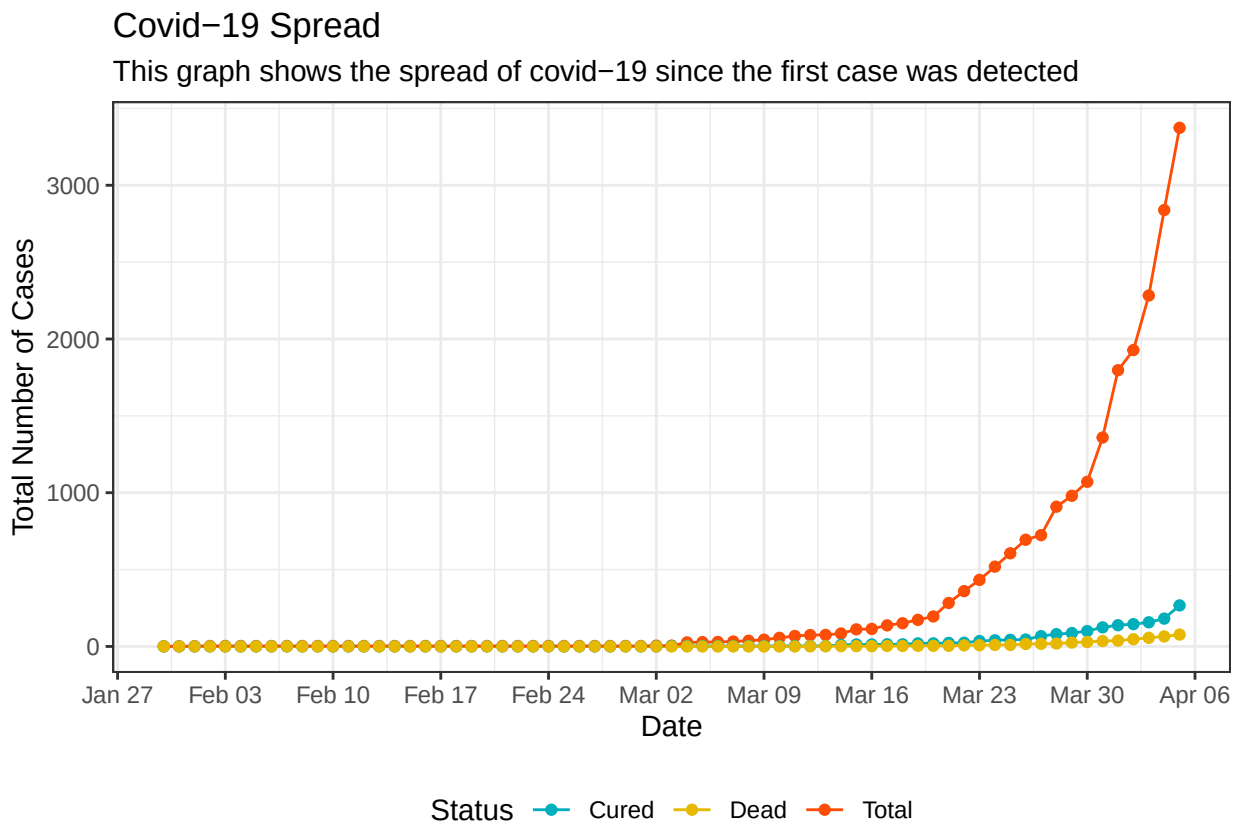
Firstly, I showed the top 5 states where Covid-19 is more active in India.



It gives the information about the states in India where the spread of the virus is high and which states you could choose to be safe. It also tells whether your states is one where Covid-19 is more active or not.

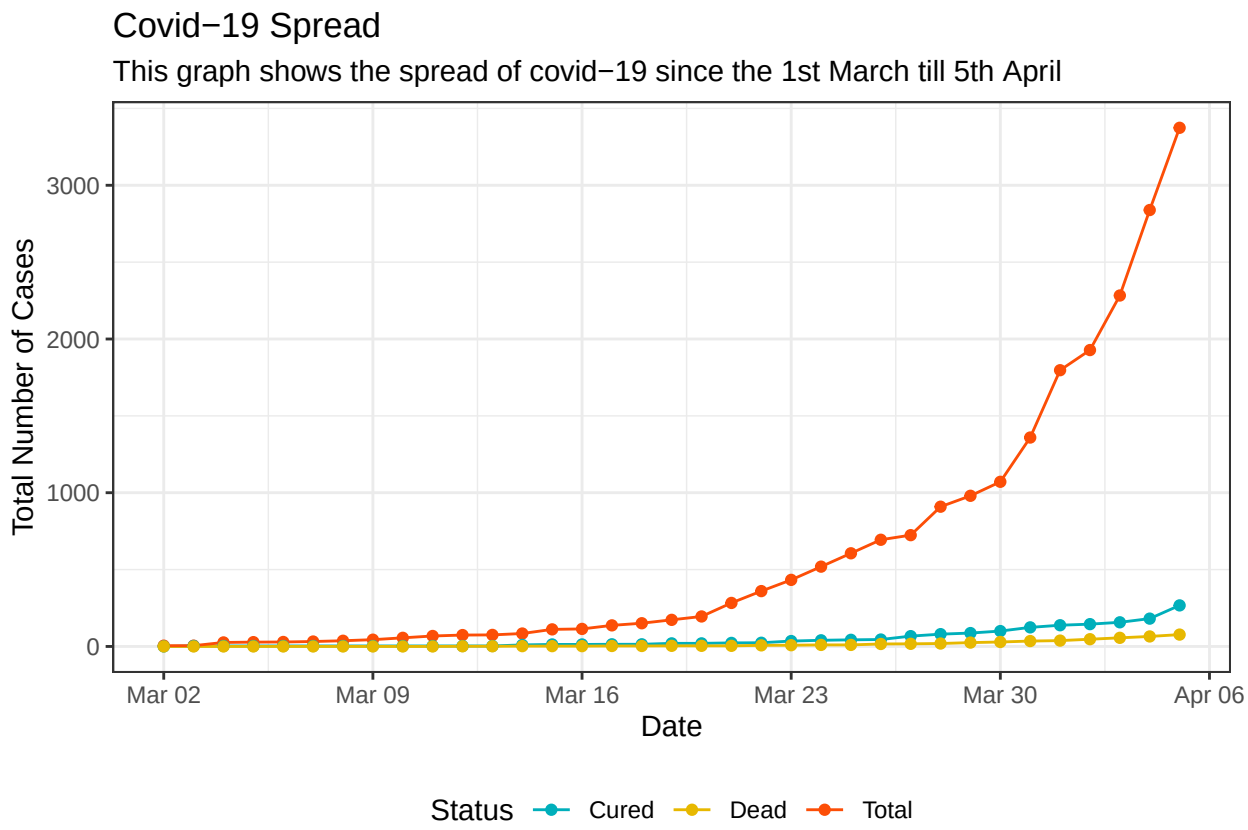
Total Number Of Cases

Now, the following graph shows the spread of covid-19 in India over the past two months since the first official cases were conformed.



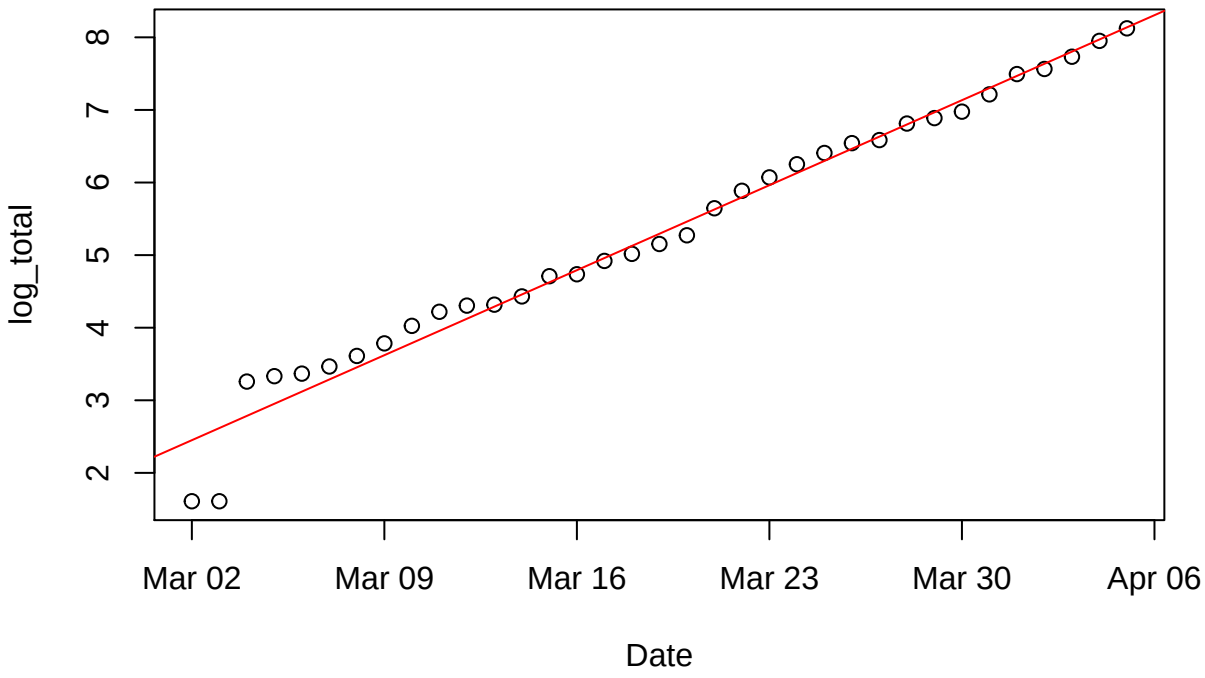
The covid-19 was not spread in India till 2nd march, where we can see more reported cases started coming. There are some of the dates where I can examine and say that those where the people because of which virus spread. For example, Indian singer Kanika Kapoor Come from a flight from London on 9th March. She did not went through security like everyone else and went on to attend a couple of parties and seminars and was tested positive on March 20. The spike which we see in around 26 march is because of her.

Since we are not seeing any changes in the number of covid-cases from February to March, we will remove that from the graph and plot it once again from march till 5th of April.



Model

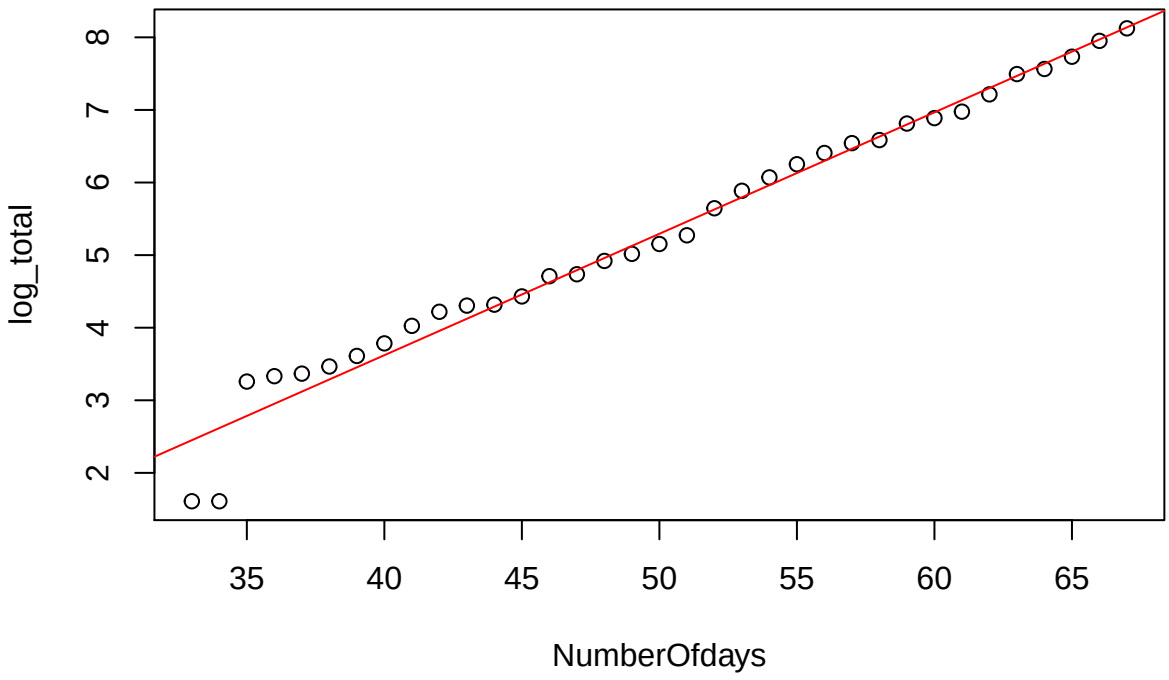
Now since we have prepared the data to our needs and displayed them above, let's make a linear model for the total case and try to find some relation there. I tried the number of equations like square, exponential and square root but I found good relation with the log and I will be using that over here.



The linear model we made is near perfect fit and most of the points lies on the line. The equation we get is as follows:

```
y(Total cases) = -3066.1624 + 0.1675 (Date of that day)
```

now entering the date here feels odd and does not make any sense for us , so I made a new change here where rather than using the date, we will use number of days it had been since the spread of outbreak.



now the date issue is fix, we get the following new equation:

```
y(Total Number of Cases)= -3.0785 + 0.1675 * x(Where x is number of days it been since the outbreak)
```

Tests and Results

Now let us see the 98% confidence interval for B1 and B0 for the equation

```
##                       1 %       99 %
## (Intercept)  -3.6536819 -2.4863742
## NumberOfdays  0.1558382  0.1787222
```

Since we have negative Intercept(at any percentage of confidence interval), we can say that the log_total will be negative unless certain days are when by. I think that means that virus will not spread rapidly unless someday went by.(Which make sense as we need more people to spread it more quickly)

now, Lets check if Number of days have a positive association on the active cases

H0: B1=0 H1: B1>0

```
##
## Call:
## lm(formula = log_total ~ NumberOfdays, data = by_date1)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.00806 -0.07080  0.01691  0.13927  0.47332
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.07003    0.23873  -12.86 2.11e-14 ***
## NumberOfdays  0.16728    0.00468   35.74  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2796 on 33 degrees of freedom
## Multiple R-squared:  0.9748, Adjusted R-squared:  0.9741
## F-statistic:  1278 on 1 and 33 DF,  p-value: < 2.2e-16
```

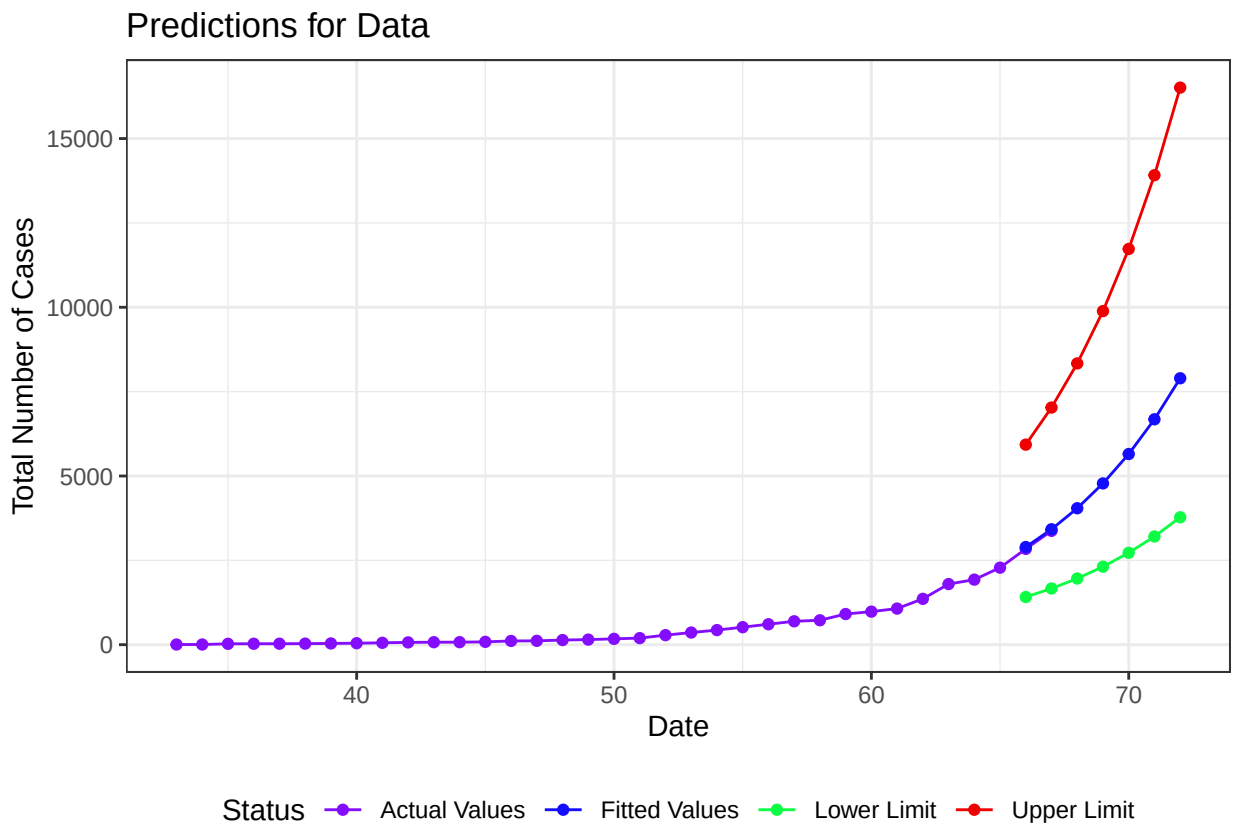
p-value: $<$ 2.2e-16 since the p-value is less than alpha(for any value of alpha) we can say that there is a positive association between the total cases and Number of days it been since the start of the virus in India.

Predictions

Lets make predictions of what will happen from a week from now if the Virus is spread according to the model. The table we find below gives us prediction for a given day and alosi gives us lower limit and upper limit for the graph.

```
##   NumberOfdays lower_limit fitted_values upper_limit
## 1           66        1413          2894        5929
## 2           67        1665          3421        7029
## 3           68        1962          4044        8336
## 4           69        2312          4780        9887
## 5           70        2723          5651       11728
## 6           71        3207          6680       13914
## 7           72        3776          7896       16511
```

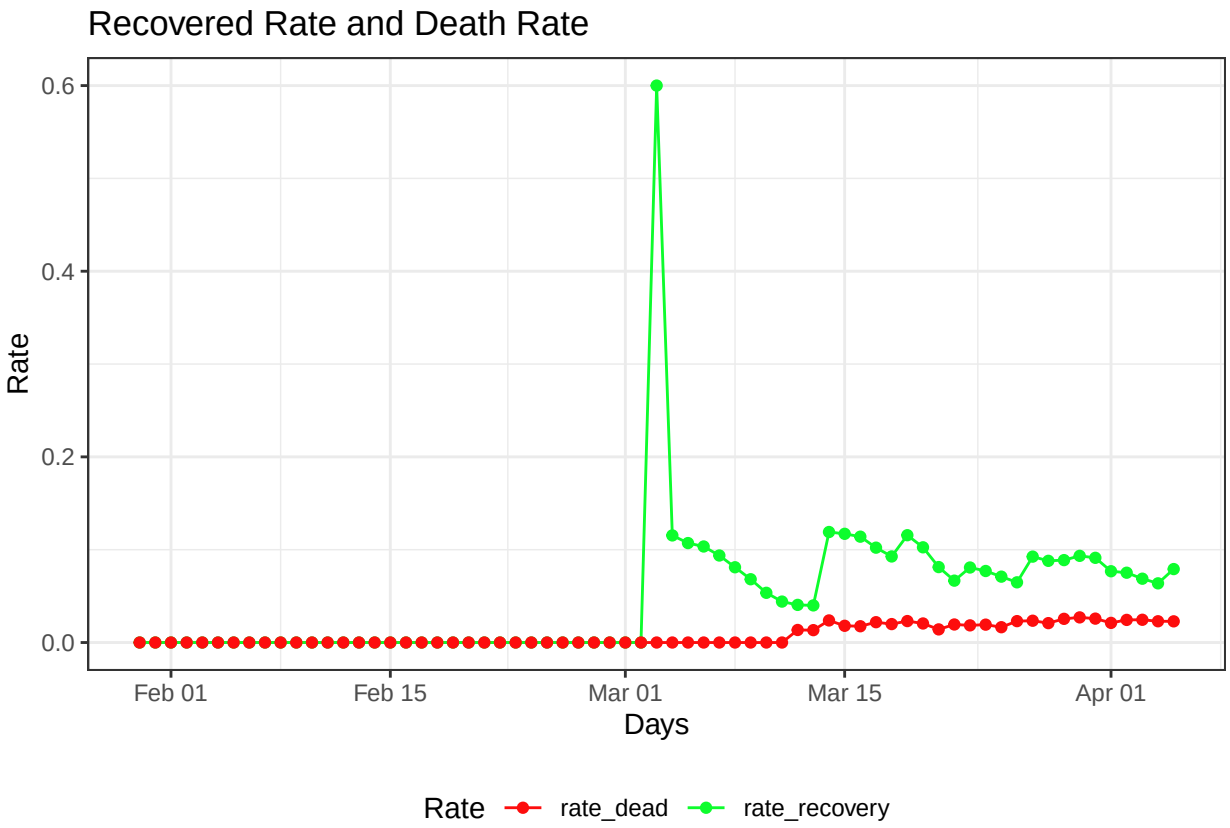
This table shows how many people will be affected by the end of that day. Now you will notice that we already have actual data for 66 and 67th day. I did this so that we can see if the model can give us the correct results. Now here we are representing that data



Now over here we can see how the data goes from actual value and how fitted value added here along with Lower limit and upper limit. We can also see that the two points which are connected there are almost overlapping the fitted value. Hence I can say the model will be really close fit.

Recovered Rates VS Death Rates

Now lets check the trends recovery rate and death rate

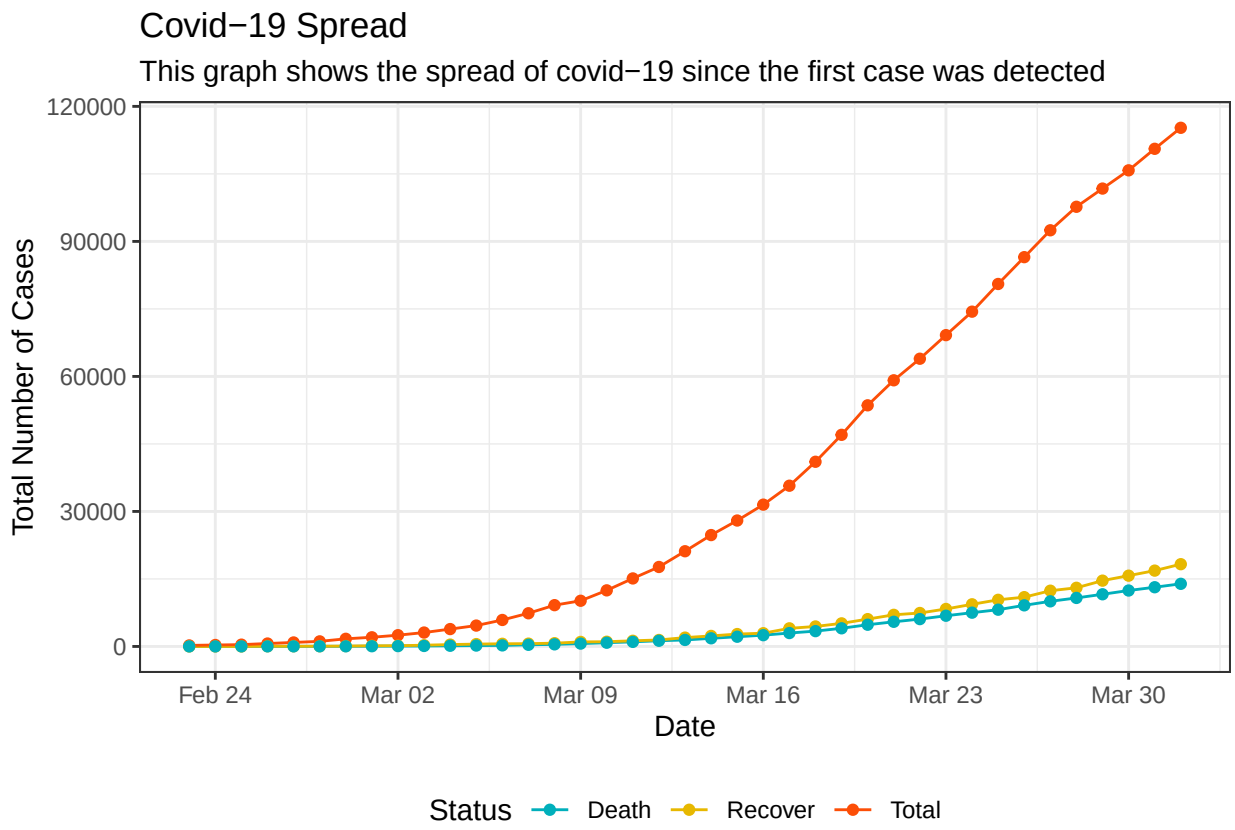


As we can see here till March first, then we see a huge increase around 2nd March as the number of Recovered increase around that time but there was a huge drop around 3 march as the number of case increase around that time and then it goes up and down as the number of cases increases or decreases accordingly. The death rate is more or less is same as the number of death are increasing but not as fast as in any other country.

part2 Case Study: Italy

Total Number Of Cases

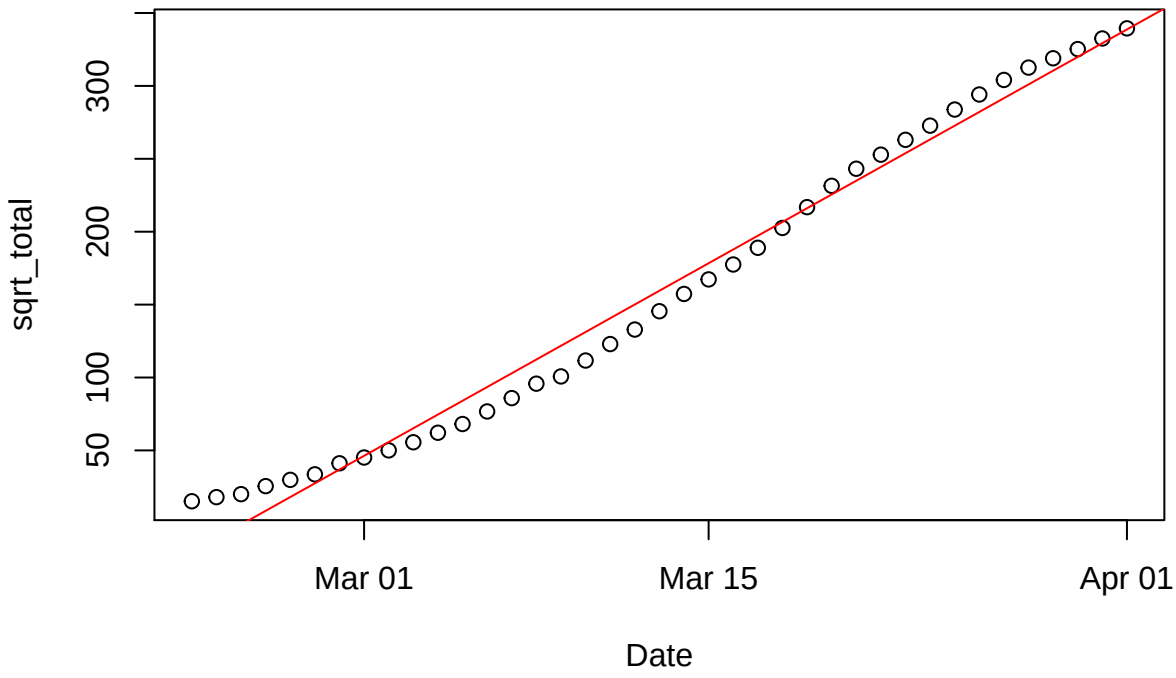
Now as for Italy, let see the graph.



This graph tells us how to fast the cases in Italy spike and give up other useful information like people recovered and people that died.

Model

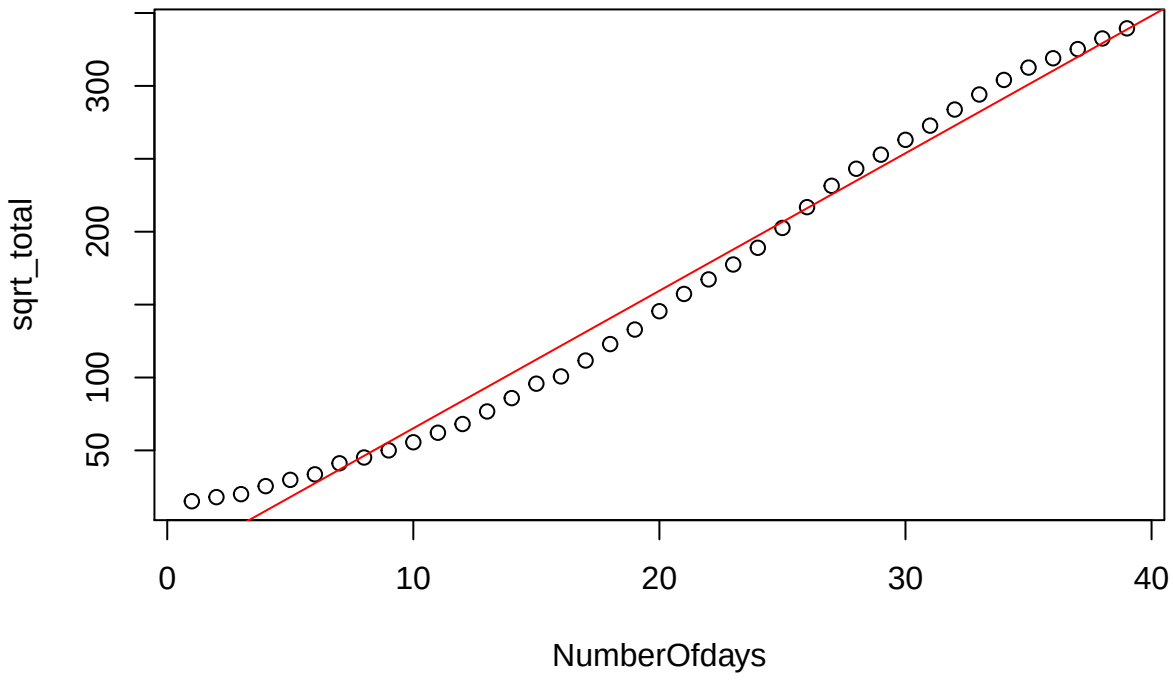
Now, I tried to make a model based on data and tried many functions but finally, the function which worked was square root function which gives us near-perfect fit.



like I said, I tried many methods before finally getting a little better result with the square root function. The function lies mostly on the line. the function comes out to be as follows:

```
Y= -172800+9433*X
where X= Date and Y are number of cases
```

But again , it feels odd to enter date in X value for me so I adjusted the value with number of days it been since outbreak.



now the equation we get from the model is:

```
Y=-29.159 +9.433*X
where X is number of days and Y is number of cases
```

Tests and Results

now lets check for 98% confidence interval for B1 and B0 for the equation we got

```
##                       1 %       99 %
## (Intercept)  -40.393972 -17.924081
## NumberOfdays   8.943428   9.922542
```

Since we have negative Intercept(at any percentage of confidence interval), we can say that the sqrt_log will be negative unless certain days are when by. I think that means that virus will not spread rapidly unless someday went by. (Which make sense as we need more people to spread it more quickly). Also, the slope over here and slope earlier have a lot of difference as well. people in Italy are getting infected very quickly than people in India. This can be due to the reason that India has to prepare for the whole country lockdown around march when the number of cases in India started increasing while in Italy, the Government was a little late to responded and failed to take such measures. Furthermore, This is something that I think, that infected people who are traveling will travel more towards western Side by either having transit flight in any of European Union Countries eg Italy, France, but people from china have less possibility to coming to India for transit flight and I think that is the reason why the virus does not spread that much in India

now, Let's check if Number of days have a positive association on the active cases

H0: B1=0 H1: B1>0

```
##
## Call:
## lm(formula = sqrt_total ~ NumberOfdays, data = by_date_Italy1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.026 -12.056   0.746   9.329  34.859
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -29.1590     4.6207  -6.311 2.39e-07 ***
## NumberOfdays   9.4330     0.2013  46.850  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.15 on 37 degrees of freedom
## Multiple R-squared:  0.9834, Adjusted R-squared:  0.983
## F-statistic:  2195 on 1 and 37 DF,  p-value: < 2.2e-16
```

p-value: $<$ 2.2e-16 since the p-value is less than alpha(for any value of alpha) we can say that there is a positive association between the total cases and Number of days it been since the start of the virus in Italy.

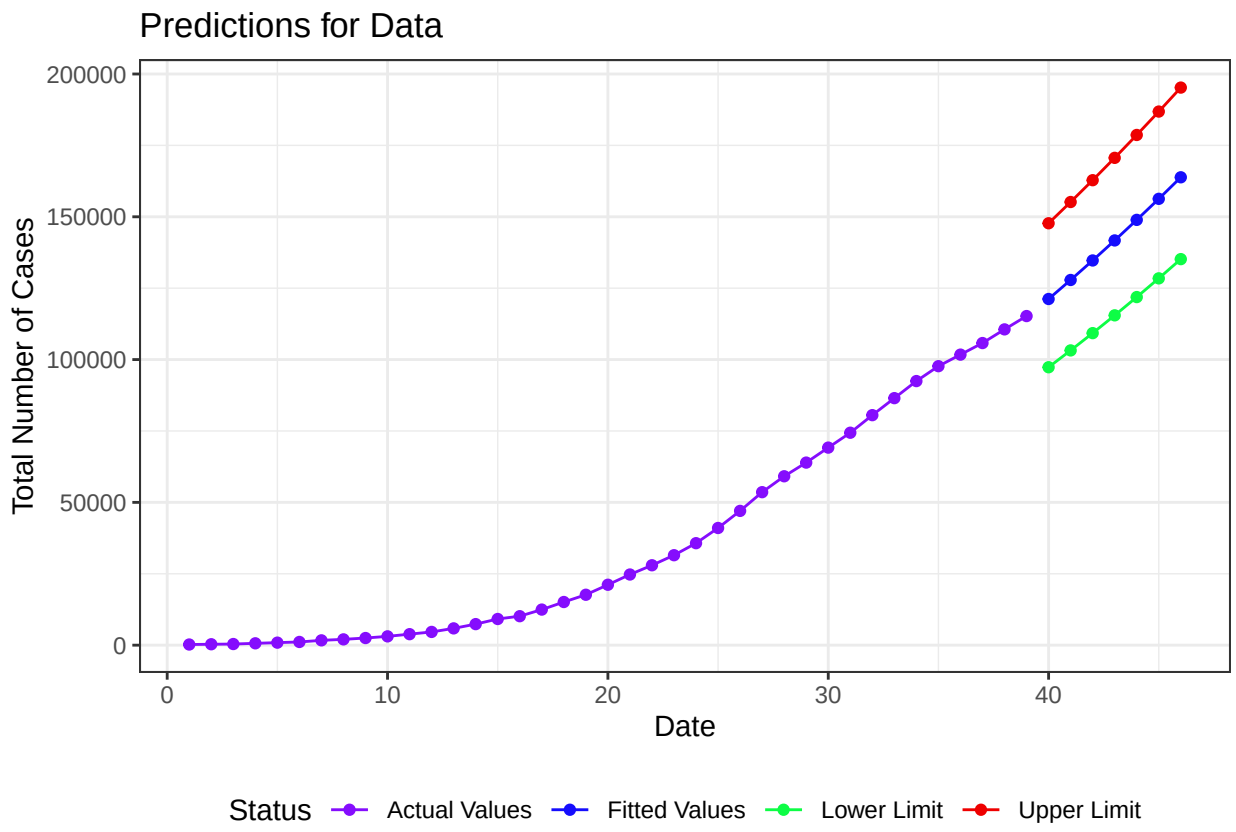
Predictions

Let's make predictions of what will have from a week from now if the Virus is spread according to the model. The table shown below shows us the fitted value we will get along with the upper limit and the lower limit of the data.

```
##   NumberOfdays lower_limit fitted_values upper_limit
## 1           40       97322        121216      147730
## 2           41      103209        127873      155177
## 3           42      109265        134708      162812
## 4           43      115490        141722      170635
## 5           44      121883        148913      178647
## 6           45      128445        156282      186848
## 7           46      135173        163829      195238
```

This table shows how many people will be affected by the end of that day.

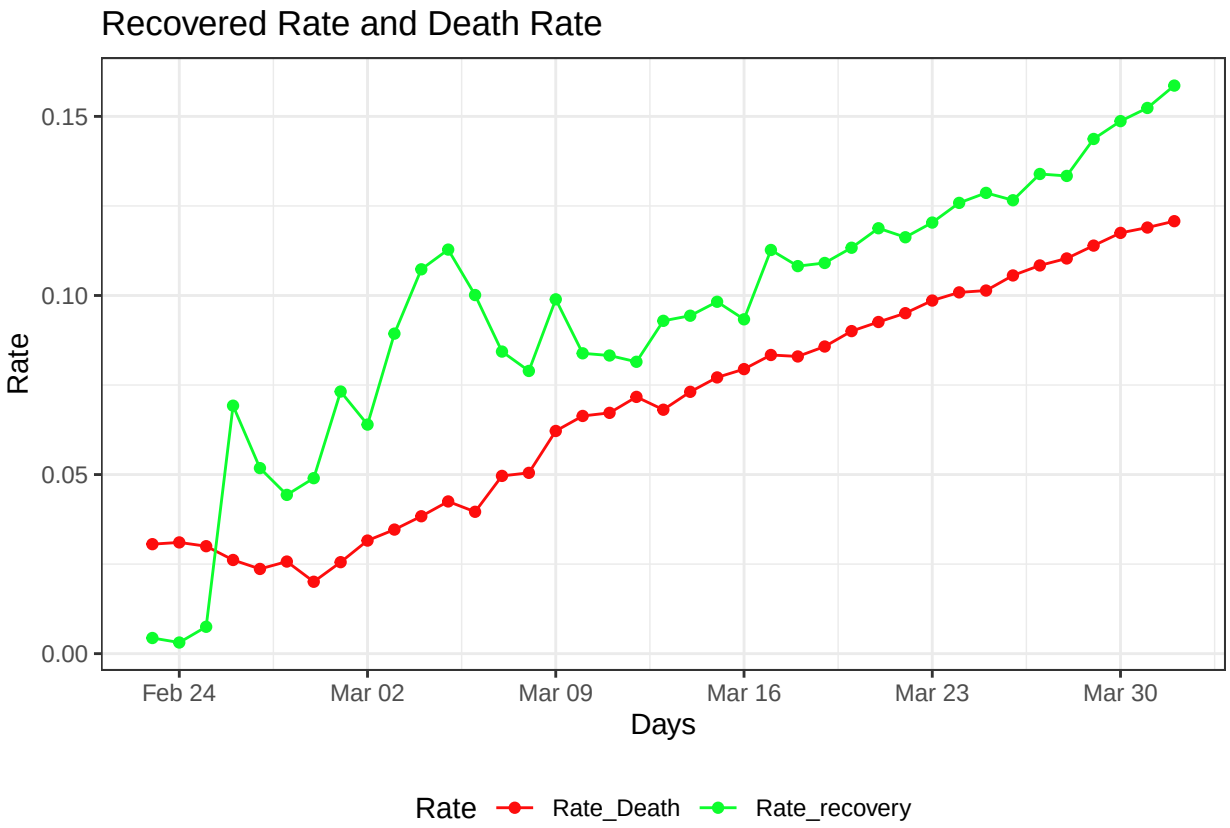
Now, after getting those values lets plot them and see how does the graph looks.



Now over here we can clearly see how data will fit into the model since it was increasing periodically. Also we can see how bad it can get and how much better it get be for people in Italy.

Recovered Rates VS Death Rates

Now lets check the trends recovery rate and death rate



As we can see here the recovery rates were lower than the death rates in the starting but then it more or less follow the liniear patter and same goes for the death rates as well. The only think which I can say is that the number of recovered patients as well number of death patients are increasing conituiously and that is the reson we see the positive trends.

Conclusion and Biased

In the End, We have multiple graphs here which tell us a lot about the virus and the most important thing to do right now is to avoid social events and public places so that neither you get the virus nor you become a carrier for it. This virus threat is real and people who think that going to work or going out will help the country are just trying to make excuses. It might help you and the country in the short run, but this could have a disastrous effect in the long run as the virus can improve and kill many more people like the swine flu. The data we got here does not have any bias in it and I know that there are other factors which may or may not effects the spread of the virus and further study is definitely needed here but I can say for sure that the number of cases will not slow down if people do not understand the importance of social distancing or problems related to leaving homes in quarantine.

Shiny APP

The Shiny app can be seen on this link

`https://samriddh202.shinyapps.io/Case_Study_On_Covid19/`