Analysis of Covid 19 Virus Data in the World

Final Project Report

Surbhi Rathore

MS CS

Table of Contents

1. Abstract

This project aims to analyze the on Corona Virus Disease 2019 (COVID-19), also known as the coronavirus or COVID, Coronavirus disease is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first known case was identified in Wuhan, China, in December 2019. The disease has since spread all over the world, leading to an ongoing pandemic. Thus, it became more important to analyze the disease and understand the statistics to spread awareness among people. Today when there is no immunity against the disease, we can probably understand the statistics of how precarious the disease is, how much deaths it has caused or can cause in future and act accordingly. The Data source is Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. It contains raw data with confirmed cases recorded along with death reported and recovered cases. The main purpose of analysis is to understand and visualize the effects throughout the world and apply models to predict for better results. The virus has still not been discovered overall, the effects, the cure, has no concrete information to be followed. In that case, analyzing it with various perceptions can help in tackling the disease better, it is socially disturbing for people across the globe but results of analysis may give an idea to make a few decisions for betterment.

2. Questions of Interest

The main objective of the project is to address the following questions:

i) Deaths due to Covid 19 is more than recovered cases? If yes, then what are the factors responsible for spreading the disease? What are the current conditions of major countries? and if

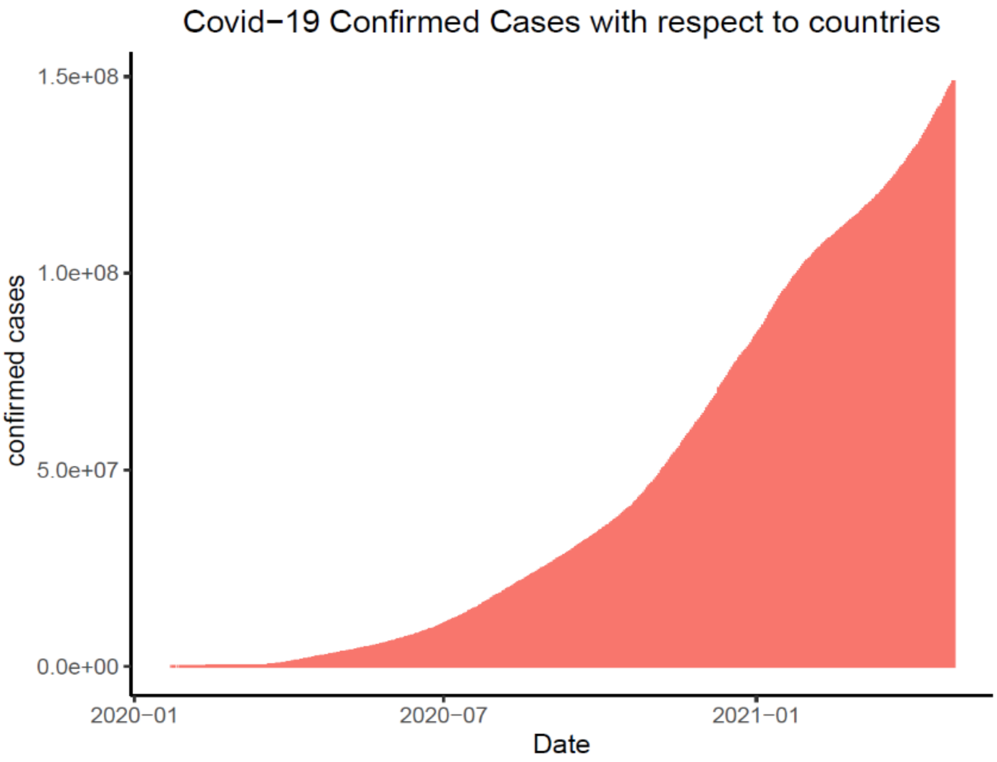
ii) How much effect do predictors like economy, estimated population density, estimated population have on the spread of Covid 19 virus?

These questions are part of my own analysis over the data as the major issues of concern is the spread of the disease and outcomes from this analysis may help to identify the factors and the result can be utilized to curb such problems to an extent.

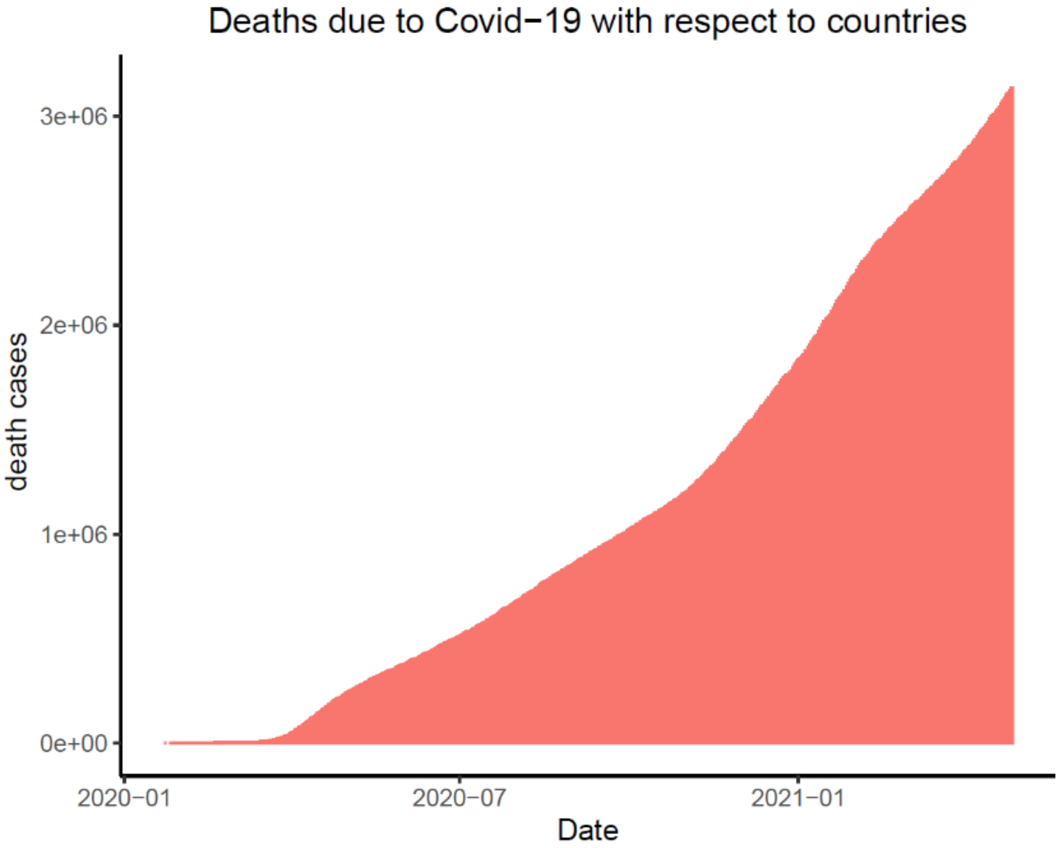
3. Data Cleaning

To create country level, I aggregated at country level, further refined country level confirmed cases and aligning it from wide to long, refined country level death cases and aligned it wide to long along with refining country level recovered cases and aligned it wide to long. Post refining, I joined all the three refined datasets obtained from raw data, using full join. Date variable required a fix hence, converted it from character to date, which gave me a country table with all the required information's. Later according to requirements, I further days column to keep a track of cases increasing in days.
I also created a consolidated world level data frame and extracted data for specific countries along with removing na values.

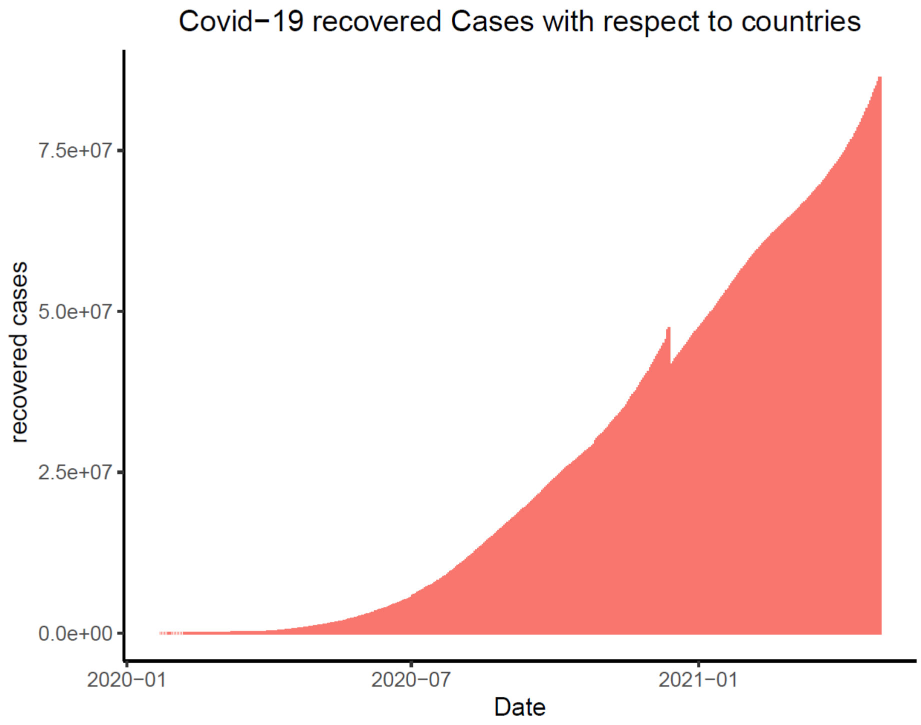
4. Observations on the confirmed, death and recovered case numbers



Plot shows the COVID 19 confirmed cases, the spread took over more that 140 million people across countries from January 2020 - January 2021. The wave stated in the month of December 2019 and led to several symptoms that existed for 14 or more days followed by testing results.

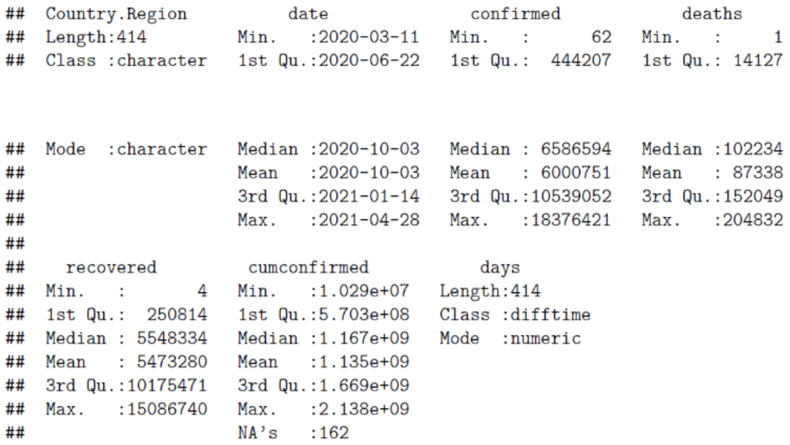


The plot shows the number of death cases due to covid 19 virus. the numbers started ranging from zero
to 100, 1000 and reached up to a million by January 2021. currently number of deaths have reached up to 3 million.



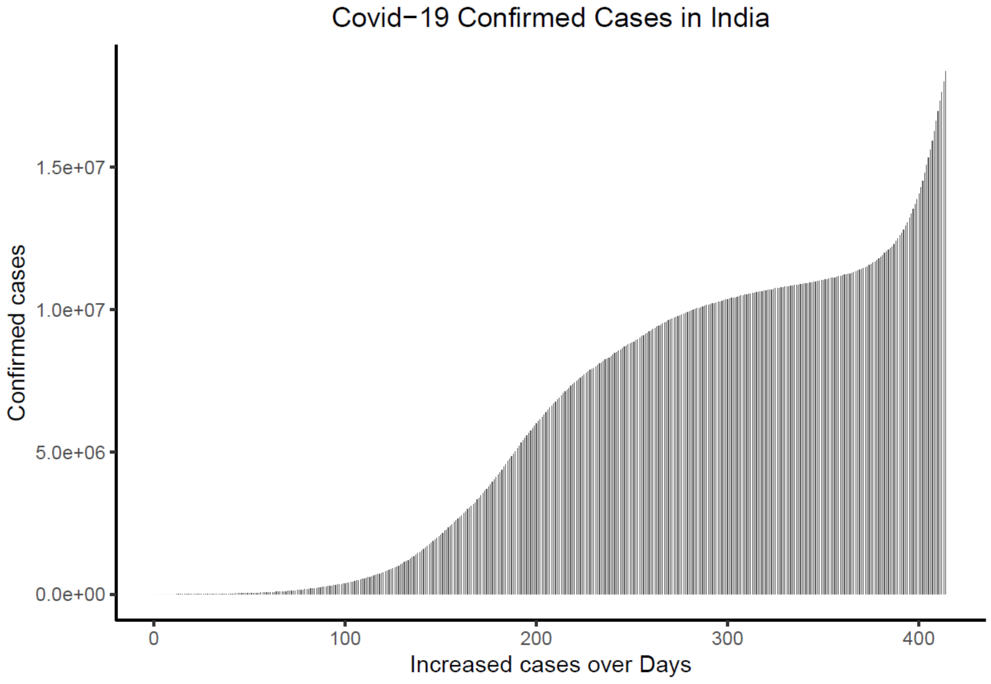
The plot shows the record of recovered cases. It accounts for people who were infected but recovered from the deadly virus effects. Current number ranges up to 75 million.

5. Observation statistical measures for country India



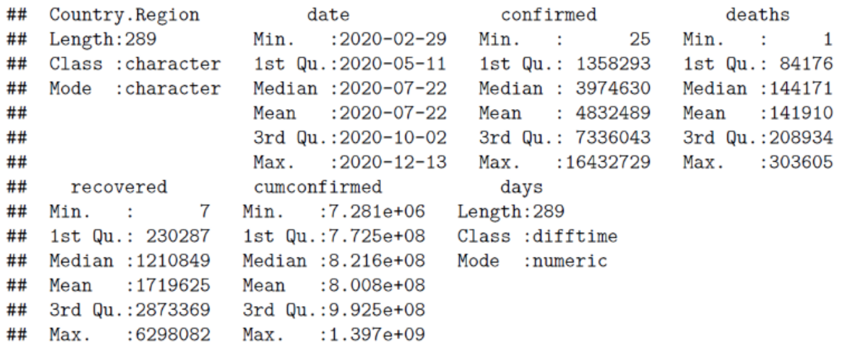
```
##  Country.Region          date              confirmed           deaths
##  Length:414         Min.   :2020-03-11   Min.   :      62   Min.   :     1
##  Class :character   1st Qu.:2020-06-22   1st Qu.:  444207   1st Qu.: 14127


##  Mode  :character   Median :2020-10-03   Median : 6586594   Median :102234
##                     Mean   :2020-10-03   Mean   : 6000751   Mean   : 87338
##                     3rd Qu.:2021-01-14   3rd Qu.:10539052   3rd Qu.:152049
##                     Max.   :2021-04-28   Max.   :18376421   Max.   :204832
##
##    recovered          cumconfirmed          days
##  Min.   :       4   Min.   :1.029e+07   Length:414
##  1st Qu.:  250814   1st Qu.:5.703e+08   Class :difftime
##  Median : 5548334   Median :1.167e+09   Mode  :numeric
##  Mean   : 5473280   Mean   :1.135e+09
##  3rd Qu.:10175471   3rd Qu.:1.669e+09
##  Max.   :15086740   Max.   :2.138e+09
##                     NA's   :162
```

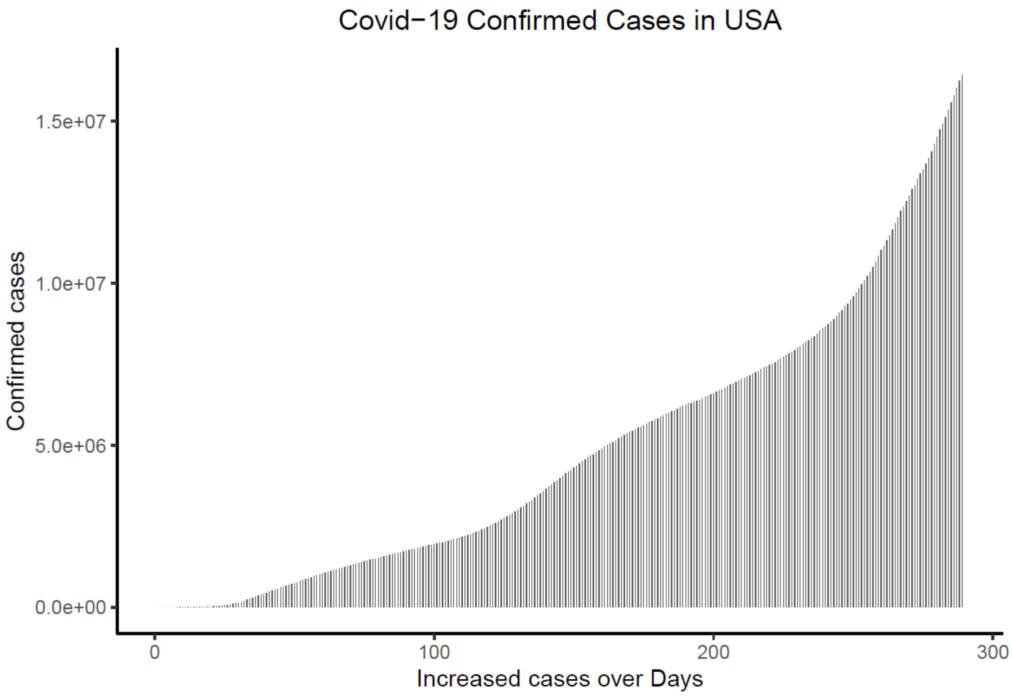


Summary of India Mean of the confirmed cases for India was recorded as 6000751 by September 2020 while the mean of death cases was recorded as 87338 followed by recovered cases around 5473280.

6. Observation statistical measures for country USA



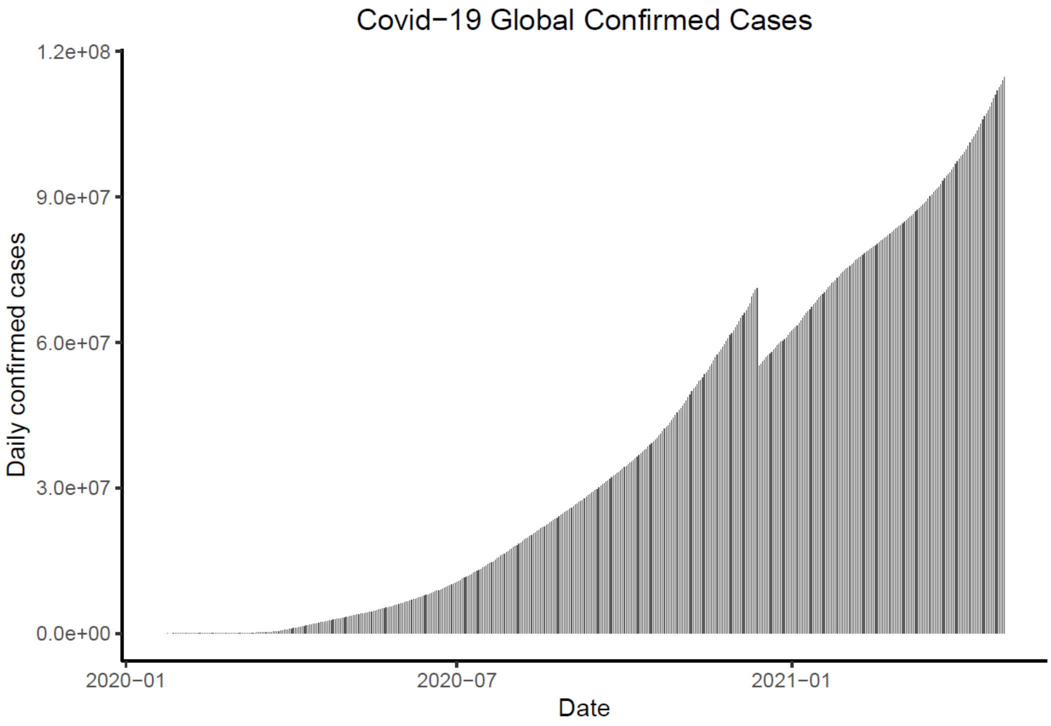
```
##  Country.Region          date                confirmed            deaths
##  Length:289         Min.   :2020-02-29   Min.   :       25   Min.   :     1
##  Class :character   1st Qu.:2020-05-11   1st Qu.: 1358293   1st Qu.: 84176
##  Mode  :character   Median :2020-07-22   Median : 3974630   Median :144171
##                     Mean   :2020-07-22   Mean   : 4832489   Mean   :141910
##                     3rd Qu.:2020-10-02   3rd Qu.: 7336043   3rd Qu.:208934
##                     Max.   :2020-12-13   Max.   :16432729   Max.   :303605
##    recovered         cumconfirmed           days
##  Min.   :      7   Min.   :7.281e+06   Length:289
##  1st Qu.: 230287   1st Qu.:7.725e+08   Class :difftime
##  Median :1210849   Median :8.216e+08   Mode  :numeric
##  Mean   :1719625   Mean   :8.008e+08
##  3rd Qu.:2873369   3rd Qu.:9.925e+08
##  Max.   :6298082   Max.   :1.397e+09
```

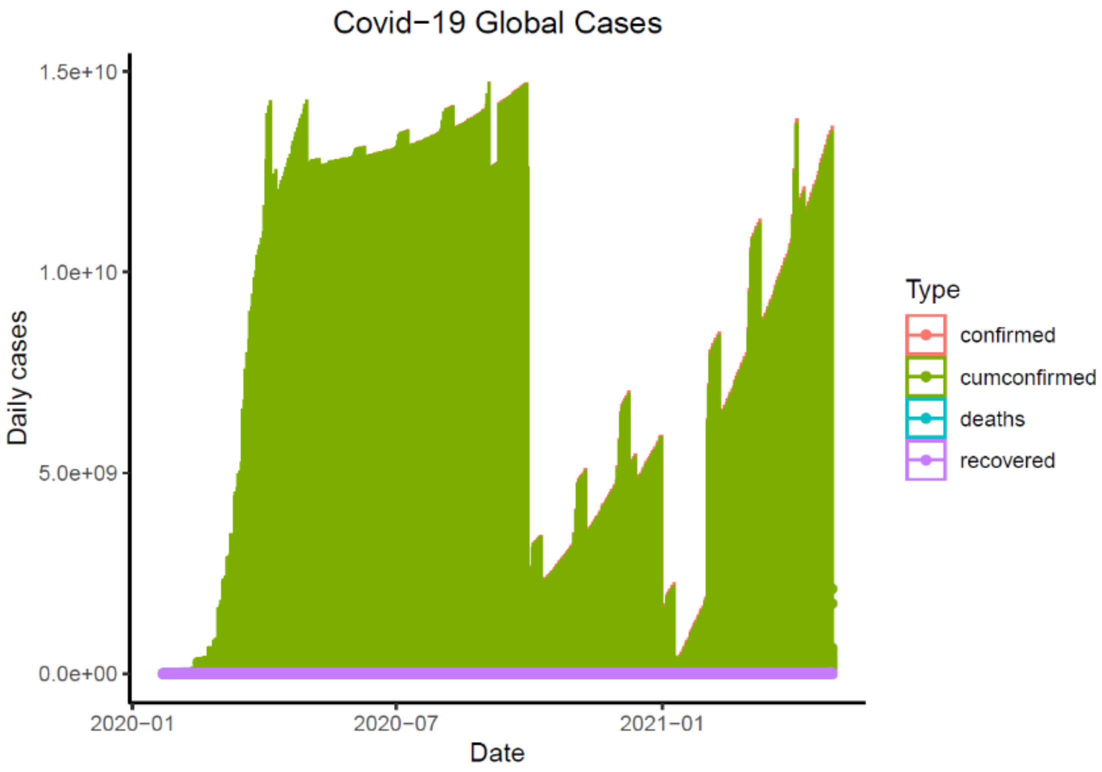


Summary of USA Mean of the confirmed cases for USA was recorded as 4832489 by September 2020 while the mean of death cases was recorded as 141910 followed by recovered cases around 1719625.

Indian population and American population have huge difference, no conclusion can be drawn with the analysis
but we can get an idea of the current numbers.

7. Analysis of Global Confirmed cases



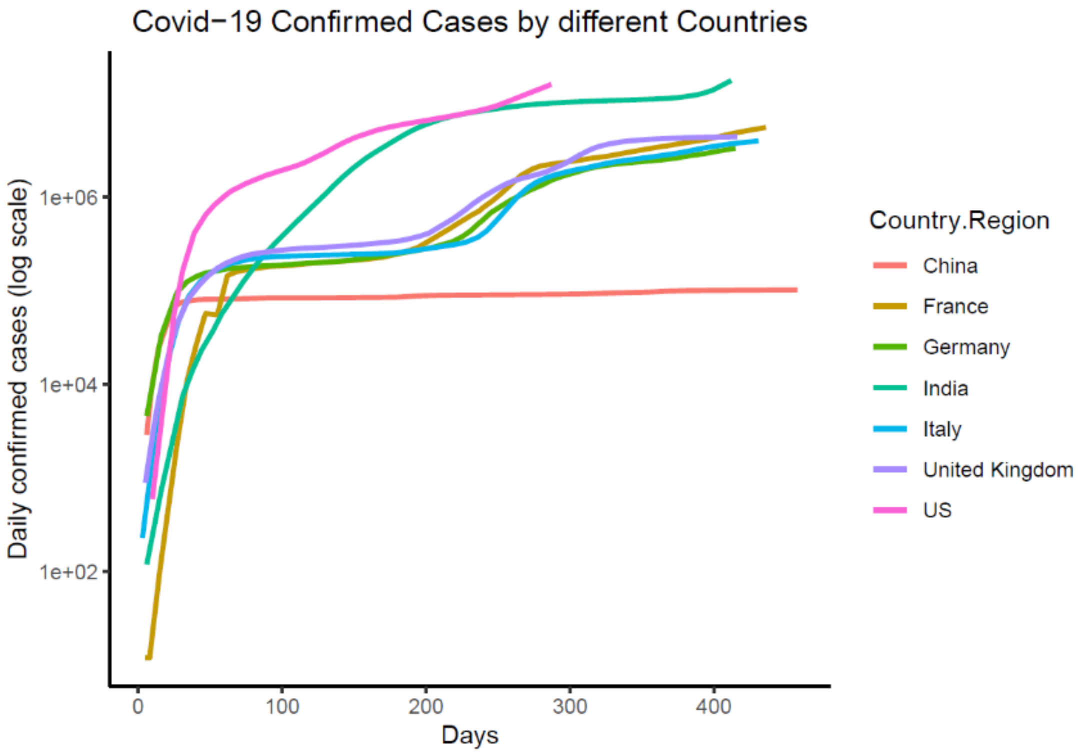


Observations:

Looking at the plots for data set globally, we can make following relevant observations:

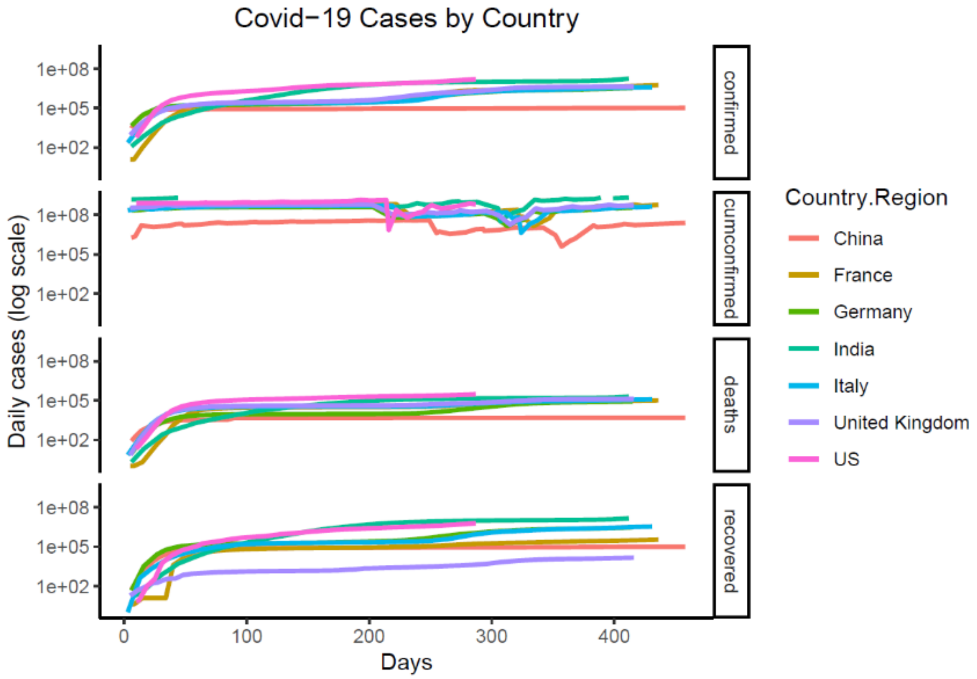
- The sum of data is growing over time. This lets us view the total contribution so far.
- Though we have recovered case yet it does not rule out the death numbers and demands people to follow safety measures to cope up with the pandemic.

8. Analysis of Cases by selected countries



Observations:

- The disease is wide spreading, though it started from china, but china is stabilizing the numbers by taking strict measures.
- Other countries like India and US still have no control over the disease and is infecting people thus increasing the numbers.
- France, Italy UK, Germany showing increasing number but little less than numbers of India and US.

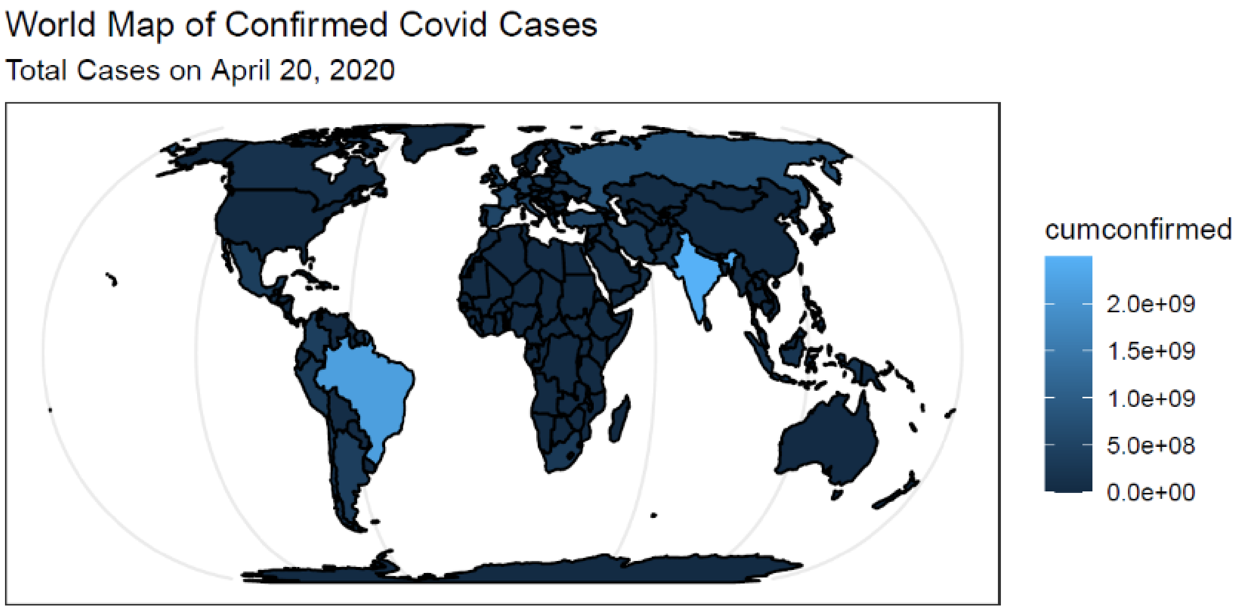


Graph shows us collaborative reports for confirmed, deaths, recovered and cumulative confirmed cases for countries specified for graph 1.

Observations:

- The most promising results in the graph for all the 3 cases is for China, even though the country led to spread of this deadly virus, it played better role in controlling the infection and bending down the death rates as well.
- Another promising result is the recovered cases from India, even though the cumulative confirmed cases are higher but somehow the virus is leading to less deaths compared to other countries. Reason may be better immunity or health care infrastructure (It’s just speculation). Hence India shows highest number of recovered cases.
- In addition, US shows least number in the recovered cases, whereas the confirmed cases are increasing by each passing day, even the death cases have higher number compared to other countries.

9. Cumulative Cases all over the world, World Map.

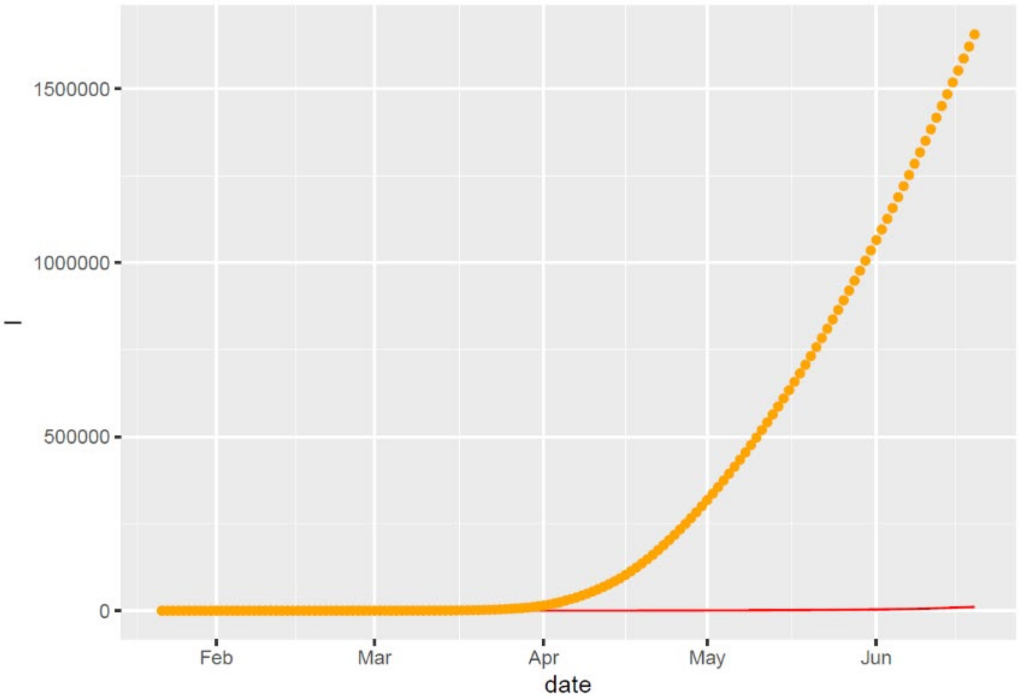

World Map of Confirmed Covid Cases
Total Cases on April 20, 2020
cumconfirmed
2.0e+09
1.5e+09
1.0e+09
5.0e+08
0.0e+00

Graph is a World map which shows cumulative confirmed cases all over the world.

Observations:

- It shows the numbers in shades of blue and has recognized least to increasing number of cases.
- Currently India shows the highest number of cases as its covered in lightest shade of blue.
- Brazil stands second with in terms of cumulative confirmed cases.

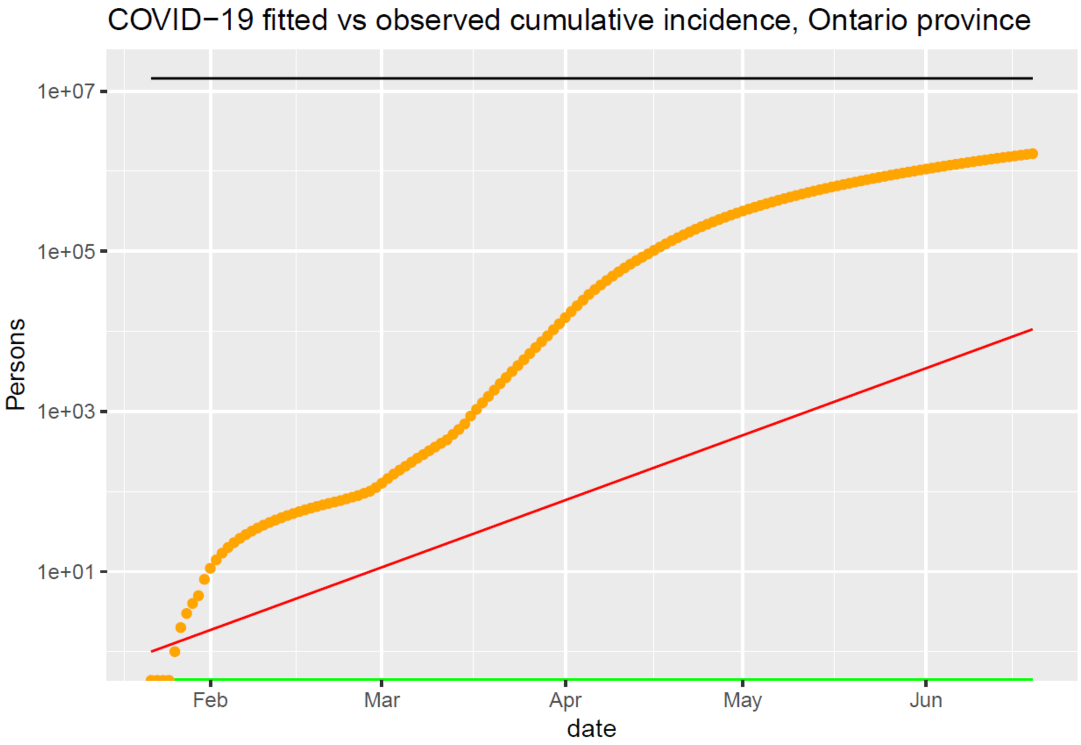
10. SIR Model Implementation (Susceptible, Infected, Recovered - Prediction)



Graph is the plot of infected (red line) along with cumulative confirmed cases (orange line).

Observation:

- In general, increase in infection is proportional to confirmed cases hence the number should have gone higher for infected as well, but my model does not perform better in fetching results for infected numbers (red line).
- I plan on improving results with better approach.



Graph shows the SIR model along with cumulative confirmed numbers.

Observation:

- The graph does show rising infection numbers but shows flat reading for both susceptible and recovered, which mean the infection is rising and the whole population is susceptible but there is no recovery rate, this output is not acceptable,
- Considering the current scenario, the prediction model is basically not accurate as it should show figures for all three measures.
- I plan to work on my model for accurate predictions.

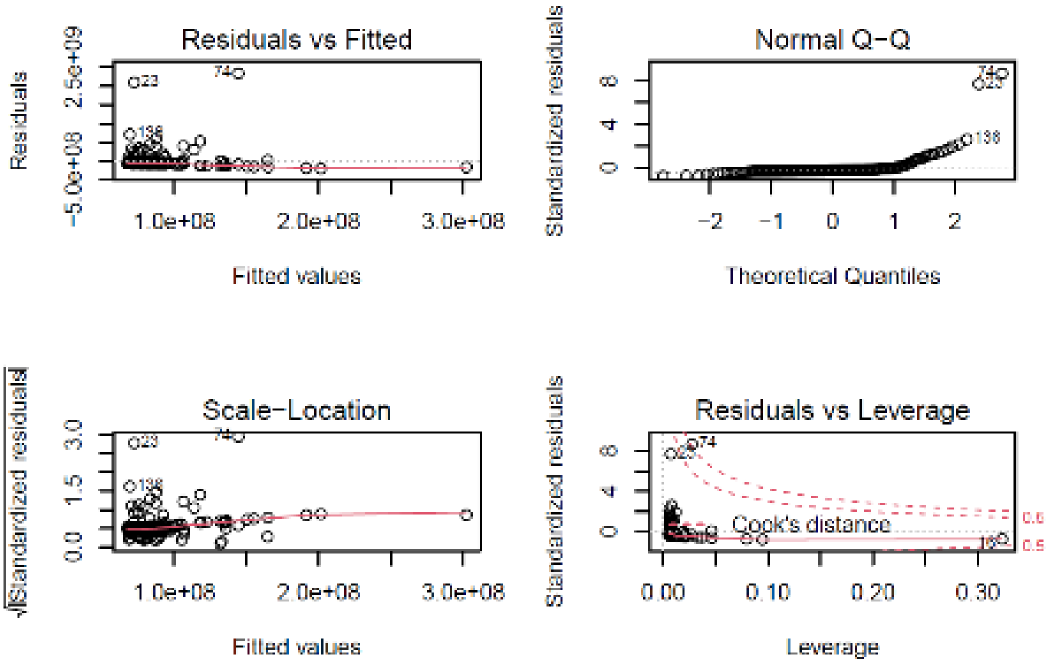
An approach I followed here is discussed below:

The SIR Model for Spread of Disease - The Differential Equation Model

The SIR model provides us with insights and predictions of the spread of the virus in communities that the recorded data alone cannot. It aims to predict the number of individuals who are susceptible to infection, are actively infected, or have recovered from infection at any given time.

- An **SIR** model is an epidemiological model that computes the theoretical number of people infected with a contagious illness in a closed population over time.
- The name of this class of models derives from the fact that they involve coupled equations relating the number of susceptible people $S(t)$, number of people infected $I(t)$, and number of people who have recovered $R(t)$.

11. Implementation of Linear Model



Plot: Linear Regression Model

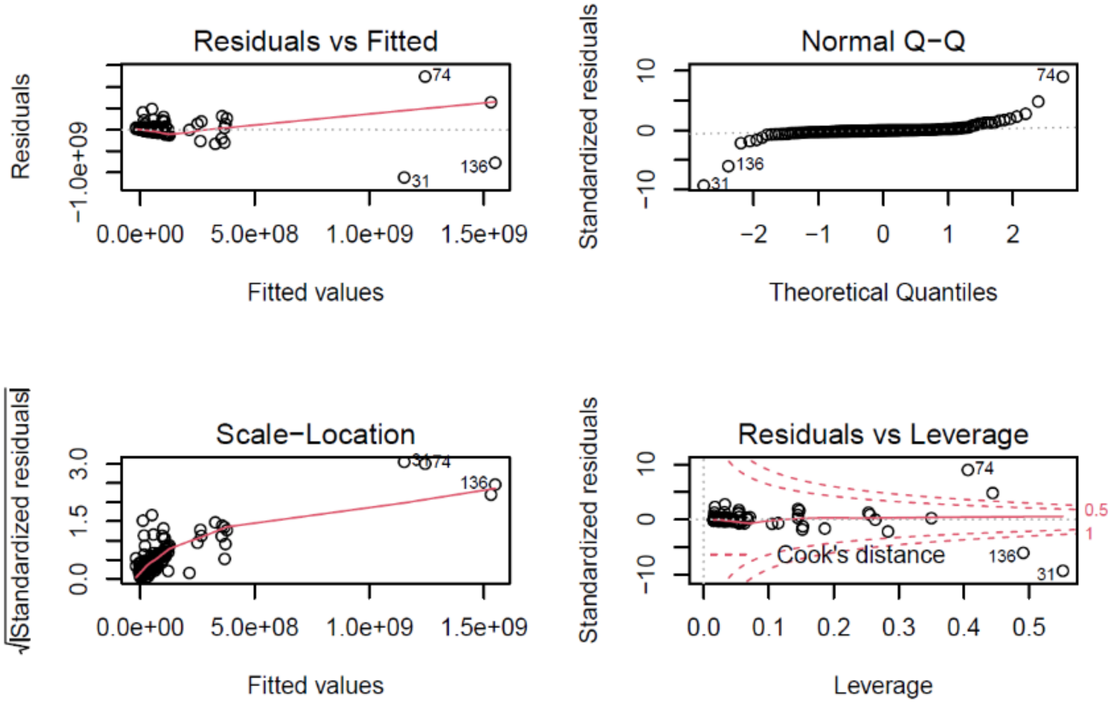
Summary Observation:

Linear Model, the Pr(>t) acronym found in the model output relates to the probability of observing any 15 value equal or larger than t. A small p-value indicates that it is unlikely we will observe a relationship
between the predictor (pop_est_dens) and response (cumconfirmed) variables due to chance. Typically, a p-value of 5% or less is a good cut-off point. In our model, the p-values are very close to zero. Three stars (or asterisks) represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between pop_est_dens and cumconfirmed cases.

Plot Observations:

- The results show a strong evidence that estimated population density is the major influencing cumulative confirmed cases.
- Shown residual plots also adds strength for the same. There are few outliers found while performing the linear fit.
- Approximately 90+% of variation in value variable can be explained by this model with these two independent variables (pop_est_dens and cumconfirmed).
- Very low P-value also strengthens the assumption.
- The residual standard error also shows there is not much distance between our observed value from the predicted value.
- We can see from the plots that there exist leverage points but, they are not much influential.

12. Implementation of Multiple Linear Regression Model



Plot: Multiple Linear Regression Model

Summary Observation:

Multiple Linear Regression Model In multiple linear regression, the R2 represents the correlation coefficient between the observed values of the cumconfirmed and the fitted i.e.pop_est_dens + pop_est + economy values of cumconfirmed, for this reason, the value of R will always be positive and will range from zero to one. A problem with the R2, is that it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response.

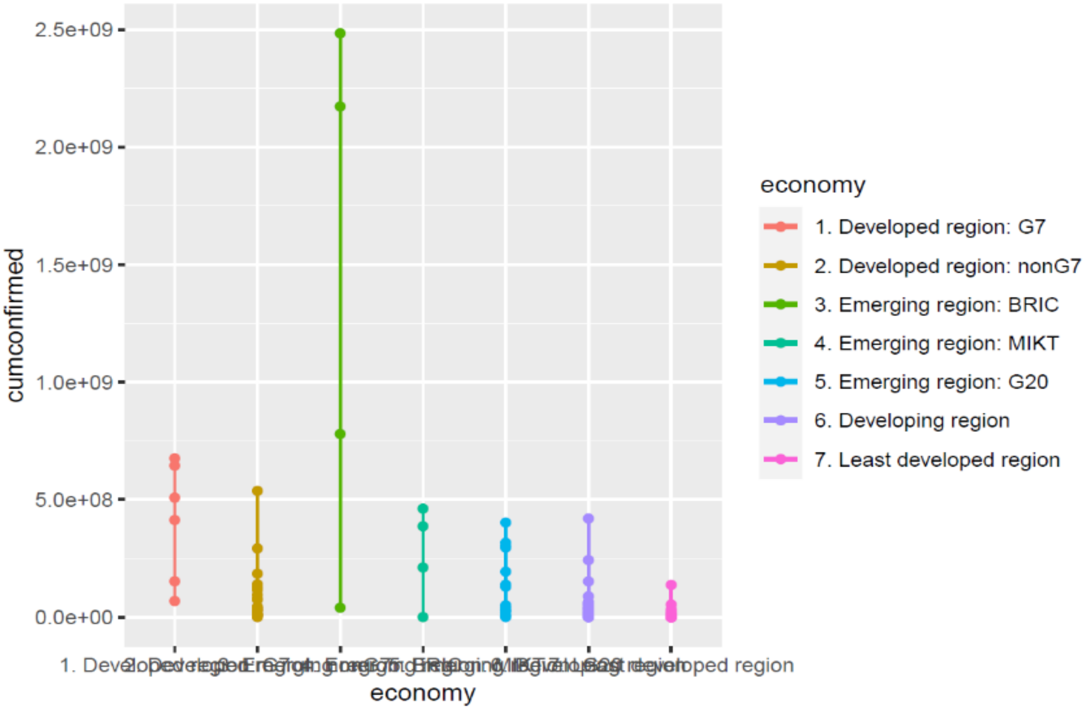
A solution is to adjust the R2 by considering the number of predictor variables. The adjustment in the “Adjusted R Square” value in the summary output is a correction for the number of predictor variables included in the prediction model. The lower the RSE, the more accurate the model, In our case RSE is 0.4902.

A small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between pop_est_dens + pop_est + economy and cumconfirmed cases.

Plot Observations:

The results show a strong evidence that estimated population density, total population estimate and
economy collectively is the majorly influencing cumulative confirmed cases.

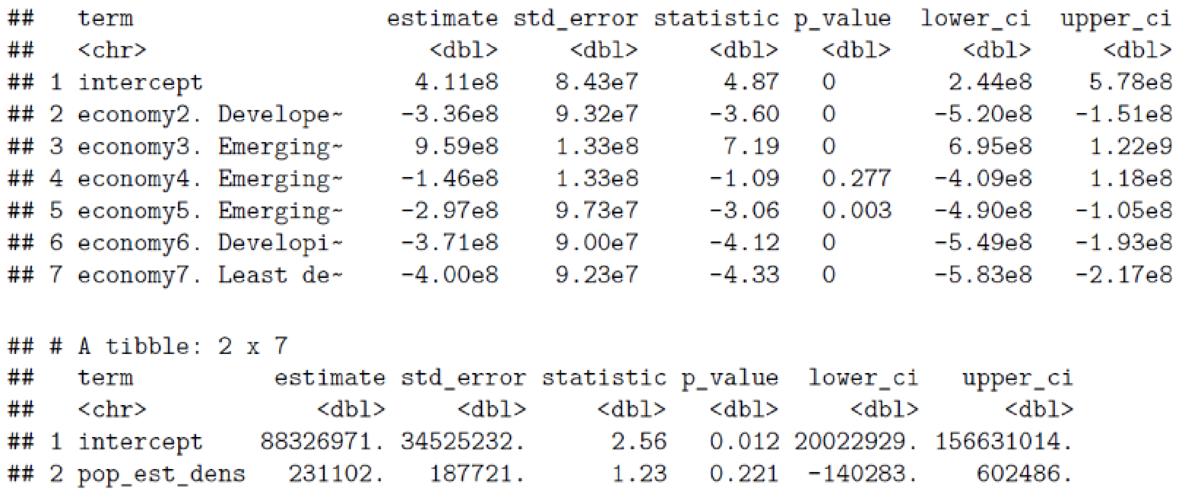
13. Analyzing the role of economy behind increasing case numbers.



Observations:

- Well, the plot does not clearly explain the relation between the spread of virus among people belonging to developing or emerging economy. It shows variable results.
- We would agree with confirmed numbers shown for emerging region: MIKT for probable reasons like, they may not have better access to infrastructures or may have to meet people in order to get their work done which would have led to transmission of virus or many more.
- But if we consider, least developed region, they have the least number of cumulative confirmed cases, which is great considering their economic conditions.
- In contrast to above note, even developed region has second highest number of cases.

14. Linear Model, Predictor- Economy and Linear Model, Predictor- Population Density Estimate



```
##   term                  estimate std_error statistic p_value  lower_ci  upper_ci
##   <chr>                    <dbl>     <dbl>     <dbl>   <dbl>     <dbl>     <dbl>
## 1 intercept              4.11e8    8.43e7      4.87   0          2.44e8    5.78e8
## 2 economy2. Develope~   -3.36e8    9.32e7     -3.60   0         -5.20e8   -1.51e8
## 3 economy3. Emerging~    9.59e8    1.33e8      7.19   0          6.95e8    1.22e9
## 4 economy4. Emerging~   -1.46e8    1.33e8     -1.09   0.277     -4.09e8    1.18e8
## 5 economy5. Emerging~   -2.97e8    9.73e7     -3.06   0.003     -4.90e8   -1.05e8
## 6 economy6. Developi~   -3.71e8    9.00e7     -4.12   0         -5.49e8   -1.93e8
## 7 economy7. Least de~   -4.00e8    9.23e7     -4.33   0         -5.83e8   -2.17e8

## # A tibble: 2 x 7
##   term          estimate std_error statistic p_value  lower_ci   upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>     <dbl>      <dbl>
## 1 intercept    88326971. 34525232.      2.56   0.012 20022929. 156631014.
## 2 pop_est_dens   231102.   187721.      1.23   0.221  -140283.    602486.
```

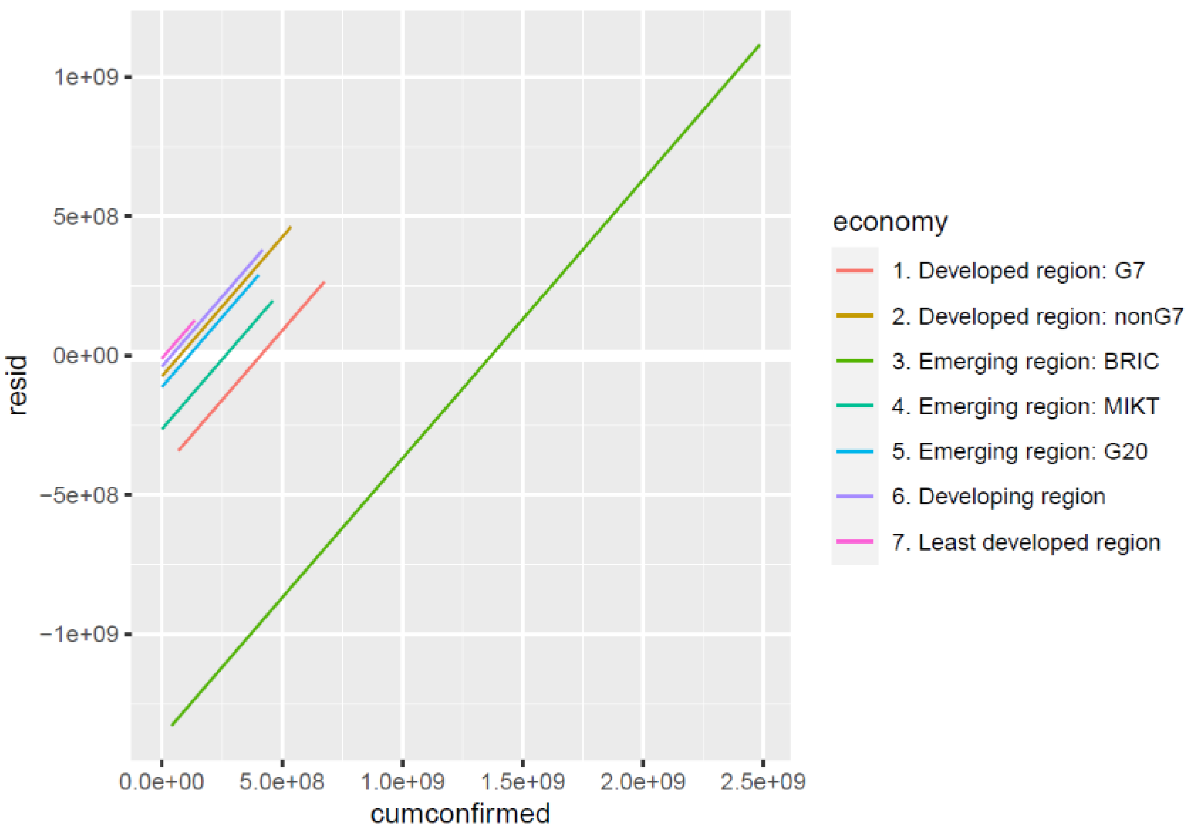
Linear Model, Predictor- Economy

To understand the relationship better, I did explore the linear model with economy as predictor and
cumconfirmed as target, and to our surprise the model does work better. the P value and R squared both are close to 0, which explains the assumption.

Linear Model, Predictor- Population Density Estimate

I had an inclination towards this particular model, as I believed the spread will increase if people assemble in closed area or are around others (Just an idea which is far away from the concept of population density but does explains the concept of density in a closed area). And this model does help in understanding it, it gives us p-value: 0.2883 and Adjusted R-squared: 0.001011. It does explain that our model fits with the data provided and spread of the disease can be related to higher population density estimated.

15. Residuals Plot in terms of economy division

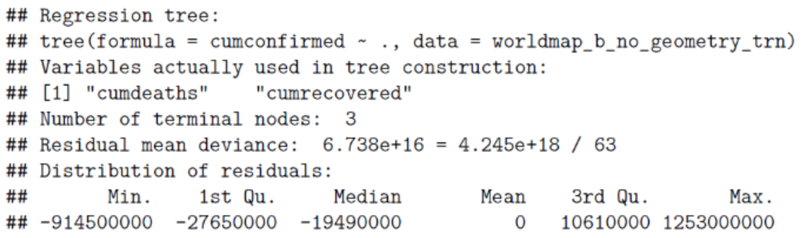


Residuals: The basic idea behind finding residuals is that we can get the difference between predicted value and the measured values and find the line that fits our data set.

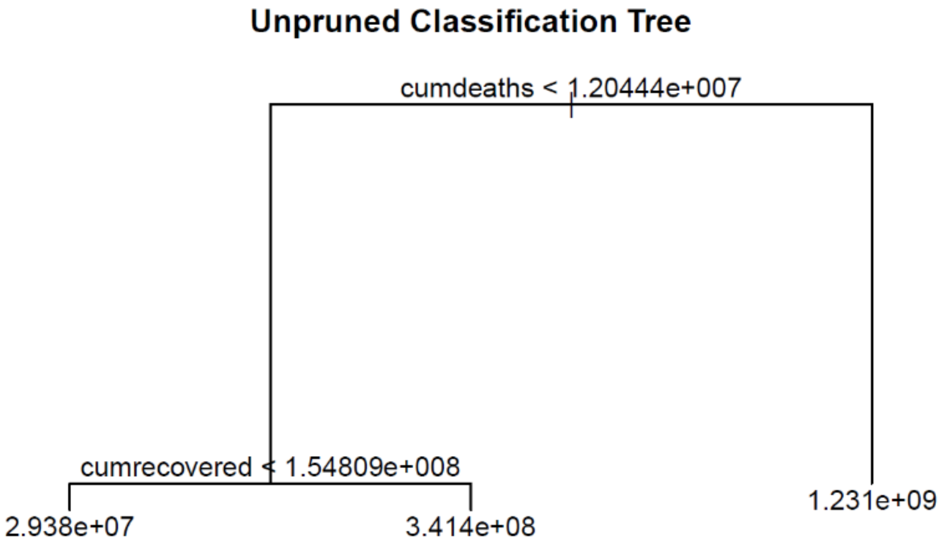
Conclusion:

16. Regression Tree Analysis:

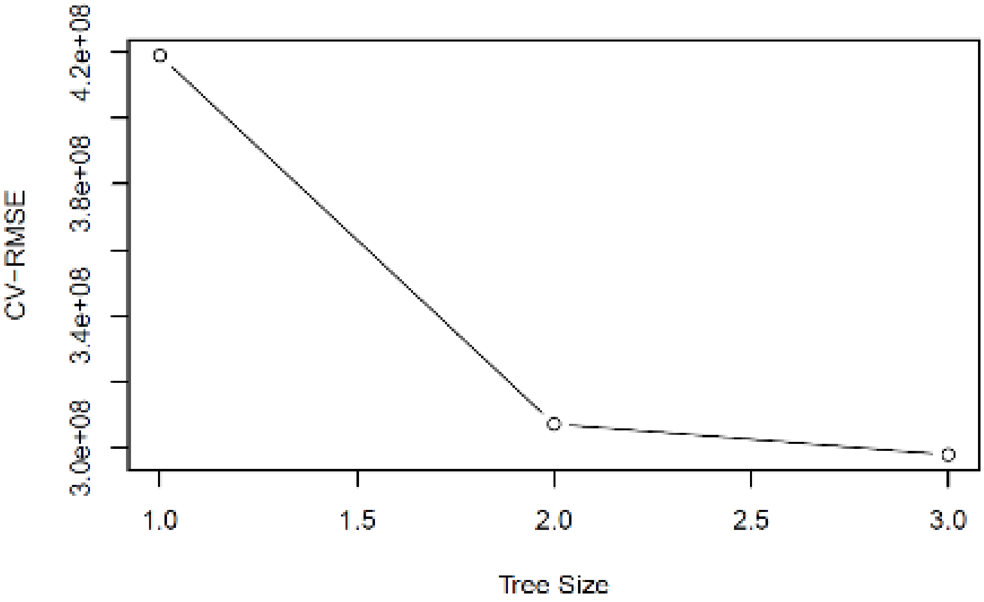
Performed another type of statistical model called Regression Trees, because it has a good graphical Representation. Basic Regression trees Covid 19 data set. The regression tree building methodology allows input variables to be a mixture of continuous and categorical variables. A decision tree is generated when each decision node in the tree contains a test on some input variable’s value. The terminal nodes of the tree contain the predicted output variable values. Basic Regression trees Covid 19 data set. The regression tree building methodology allows input variables to be a mixture of continuous and categorical variables. A decision tree is generated when each decision node in the tree contains a test on some input variable’s value. The terminal nodes of the tree contain the predicted output variable values.



```
## Regression tree:
## tree(formula = cumconfirmed ~ ., data = worldmap_b_no_geometry_trn)
## Variables actually used in tree construction:
## [1] "cumdeaths"    "cumrecovered"
## Number of terminal nodes:  3
## Residual mean deviance:  6.738e+16 = 4.245e+18 / 63
## Distribution of residuals:
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -914500000  -27650000  -19490000          0   10610000 1253000000
```

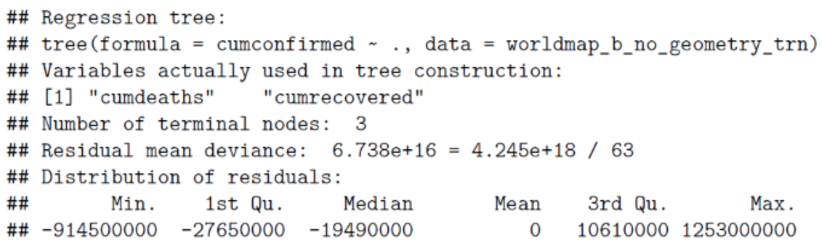


As with classification trees, we can use cross-validation to select a good pruning of the tree.

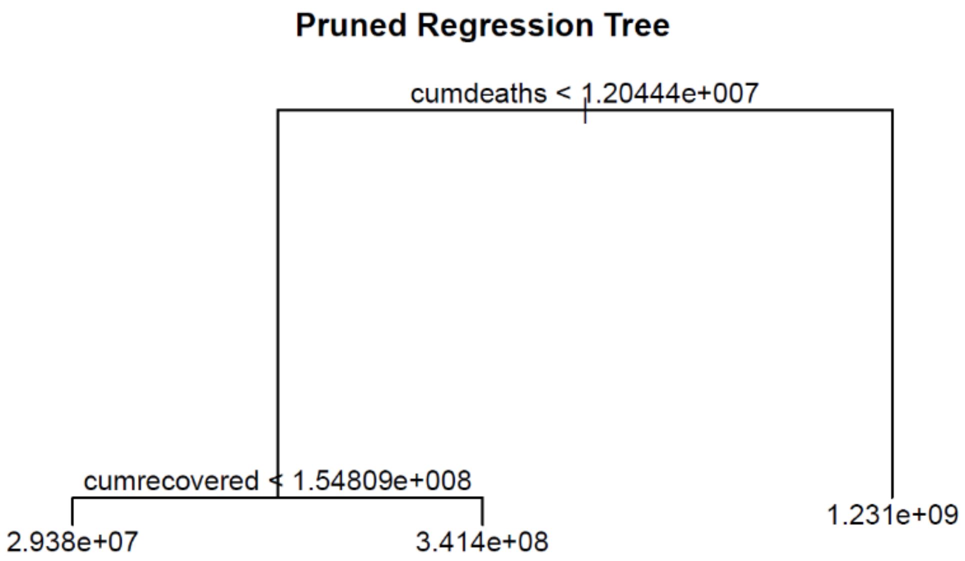


Cross-validation refers to a method for measuring the performance of a given predictive model on new test
data sets.
The basic idea, behind cross-validation techniques, consists of dividing the data into two sets:
1.The training set, used to train (i.e. build) the model;
2. The testing set (or validation set), used to test (i.e. validate) the model by estimating the prediction error.

While the tree of size 380000000 does have the lowest RMSE, we'll prune to perform just as well. The pruned
tree is, as expected, smaller and easier to interpret.

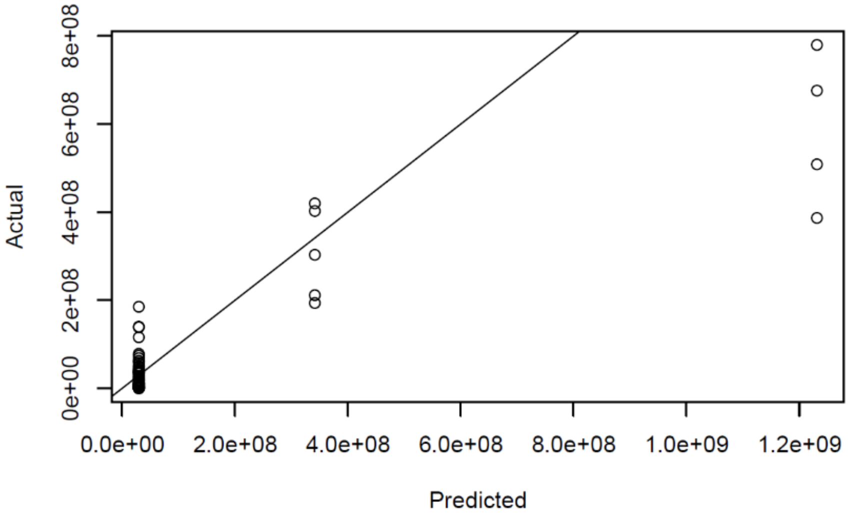


```
## Regression tree:
## tree(formula = cumconfirmed ~ ., data = worldmap_b_no_geometry_trn)
## Variables actually used in tree construction:
## [1] "cumdeaths"    "cumrecovered"
## Number of terminal nodes:  3
## Residual mean deviance:  6.738e+16 = 4.245e+18 / 63
## Distribution of residuals:
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -914500000  -27650000  -19490000          0   10610000 1253000000
```

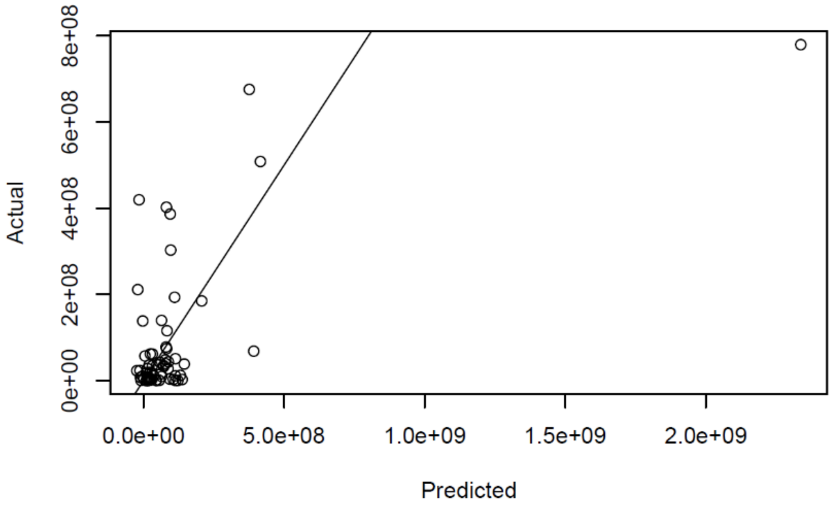


Regressing tree does not provide us better model view to understand the distribution of data, let’s compare
this regression tree to an additive linear model and use RMSE as our metric. We obtain predictions on the train and test sets from the pruned tree. We also plot actual vs predicted.

This plot may look odd. We’ll compare it to a plot for linear regression below.

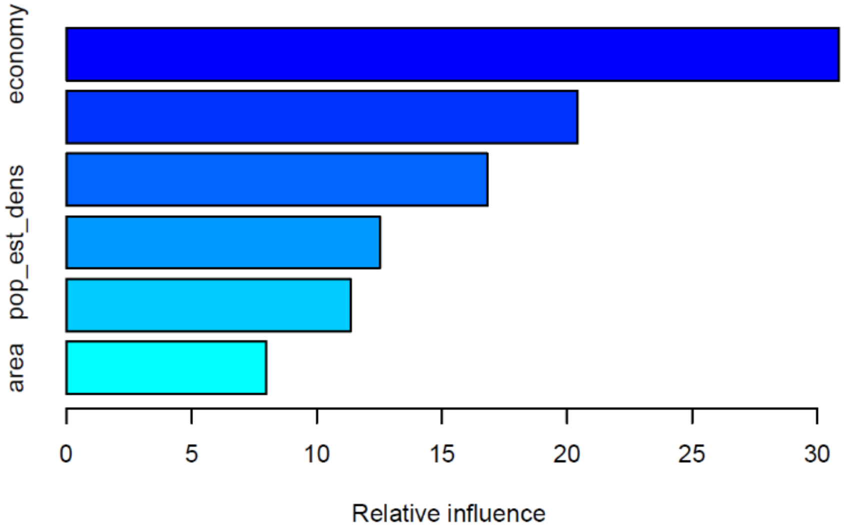


Here, using an additive linear regression the actual vs predicted looks much more like what we are used to.

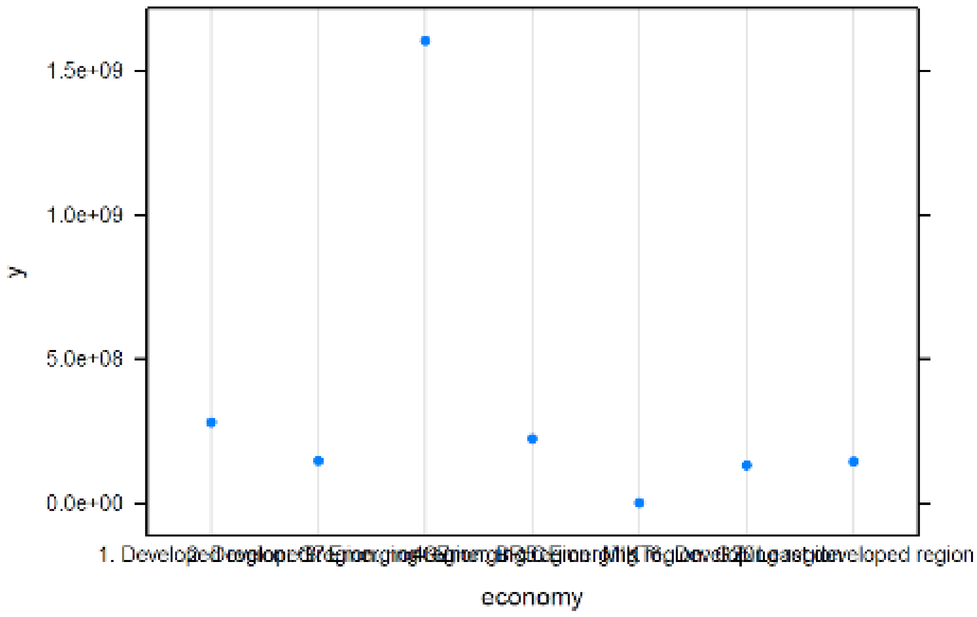


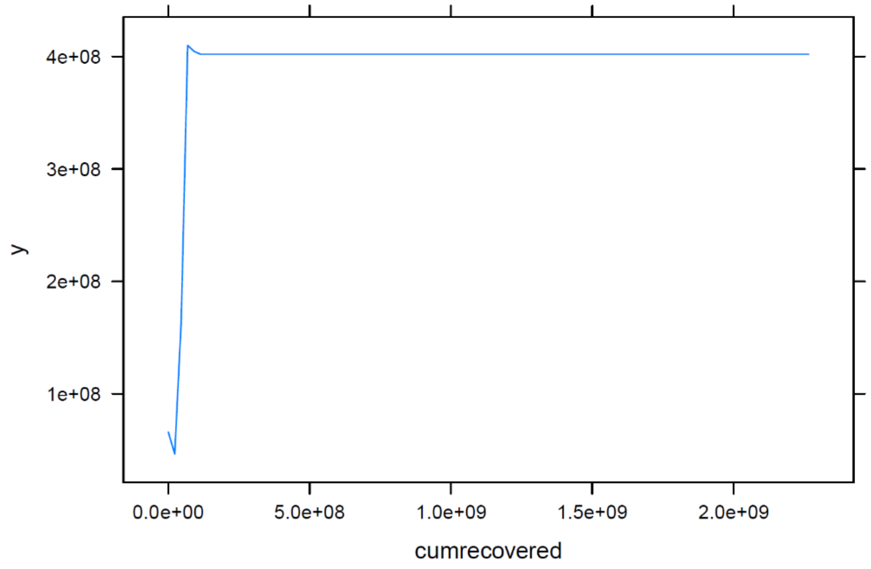
The tree after pruning seems to be easier to interpret and has a better graphical representation

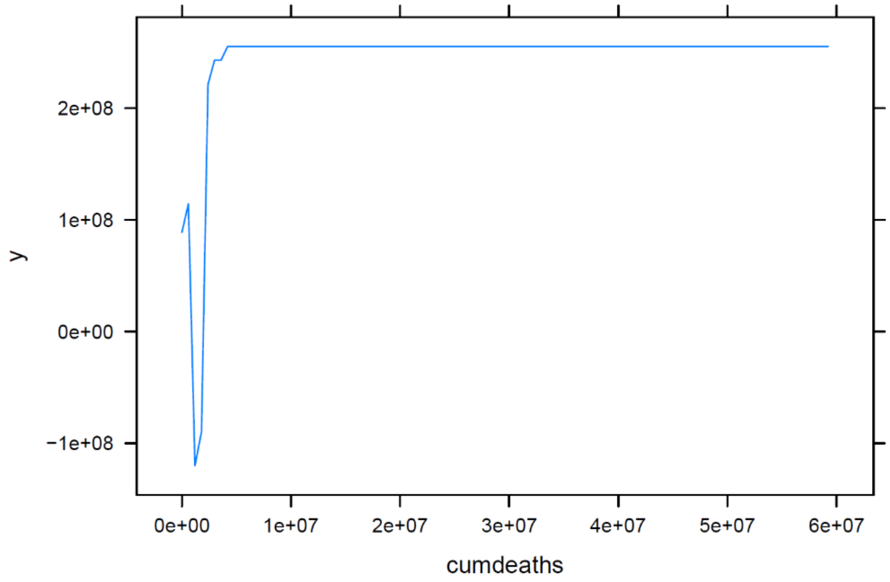
Performed Boosting analysis to get better fit and clear plots:



```
##                       var   rel.inf
## economy           economy 30.876848
## cumrecovered cumrecovered 20.420271
## cumdeaths       cumdeaths 16.824129
## pop_est_dens pop_est_dens 12.524883
## pop_est           pop_est 11.358863
## area                 area  7.995007
```


1.5e+09
1.0e+09
5.0e+08
0.0e+00
y
1. Develo2e. Develope3r7egEmoenrgin#g4Eg..mioenrgBeRdCeiognionmg MIIST6glDoenvel3l0Lpaisntgdeveloped region
economy


4e+08
3e+08
2e+08
1e+08
y
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
cumrecovered



- The best predictor seems to be economy.
- Cumrecovered estimate is the next best predictor followed by cumdeaths, pop_est_dens, area and pop_est.
- The tree() function has used cumdeaths and cumrecovered for building the Regression tree.

17. Overall Conclusion:

Overall, we can conclude that cumulative deaths and cumulative recovered is directional proportional to the cumulative confirmed, the relationship can be established using an independent variable Population estimation of the particular country. Also, cumulative deaths and cumulative recovered are the most important predictors that influence the values. Economy is another factor, that can be used as predictor. Economy in every sense is related to people of particular country, the better the economy the better the infrastructure and health care systems and opposite in case of emerging economy or least developed nation.

18. Future Research:

Analysis using more predictors.
Analyze the problem with SIR model and get it working as it is one of the most preferred method for predictions.
Perform analysis on genomic forms
Analyze the importance of social rules over Covid spread, like Locked down, Night curfews etc.
These could be a good cases for future research.

19. References:

- https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model
- https://www.rdocumentation.org/
- https://coronavirus.jhu.edu/