STA 141: Analysis of Covid Case Growth Rates Across Different Countries

Group 12

3/2/2021

Abstract

Our main question of interest for this exploratory data analysis is whether or not there is a difference in the 200 day cumulative new case growth rate amongst WHO (World Health Organization) member countries that had implemented a mask mandate, testing, lockdown policy, and/or restaurant closure mandate, versus those that didn't implement all or part of these policies during that timeframe. The 200 day period that will be examined for each country will start from the day of the 1st confirmed new case of COVID-19. Furthermore, only those countries that recorded at least 800 new cases in the first 100 days of data collection were included as part of this data analysis. This restriction was implemented due to the fact that countries falling outside this range may have too few cases to be able to accurately judge if any relationship is present between the aforementioned enacted policies and new case growth.

Introduction

The results of this data analysis could better inform us on the effectiveness of mask, testing, and other public policies in stemming the spread of COVID-19. In turn, this information could have significant public policy implications as countries decide whether or not to implement mask mandates, introduce more widespread testing, mandate restaurant closures, and/or enact lockdowns for their citizens to combat the COVID-19 pandemic. The results of this analysis can be used to build political and public support either for or against these measures based on its results. Additionally, if such policies are shown to be effective in stemming the spread of COVID-19, this analysis would allow researchers to better forecast the impact of COVID-19 on countries based on whether or not they had implemented the mentioned policies.

We will be examining two datasets in this report. One was produced by the World Health Organization (WHO) regarding the COVID-19 pandemic. This dataset features the following variables of interest: number of new COVID cases, deaths attributable to COVID-19, and a cumulative measure of both data points since January 3rd, 2020 until the present. Data has been compiled from all countries that are members of the WHO. The second dataset was published in the journal Nature and includes data from the European Center for Disease Prevention and Control, Oxford, and other sources. This particular dataset examines the response of governments to COVID-19. Some of the key variables from this dataset that will be relied on for this report include: obligation to wear masks (binary value tracking whether or not mask mandate is in place), testing policy (binary value tracking whether or not a public testing policy is in place), number of new cases reported each day for each respective country, etc.

At the onset, we hypothesize that because masks prevent air from flowing between persons, they will be effective in reducing the transmission of COVID-19 amongst a population. Therefore, we expect to see that countries that implemented the mask mandate during the 200 day timeframe will have a lower cumulative new case growth rate versus countries that did not implement such policies until after the mentioned timeframe. Furthermore, having a public testing policy would allow COVID-19 cases to be caught early on so that people can go into isolation and minimize their likelihood of spreading the virus to their peers. Instituting a lockdown and/or laws requiring restaurant closures would be expected to have a similar impact due to the fact that they also restrict people from interacting with one another and encourage isolation. So it can also be hypothesized that countries with public testing policies, a lockdown in place, and/or laws mandating restaurant closures may have a lower cumulative new case growth rate over the 200 days in question.

Background

As previously mentioned, two datasets were used in the creation of this analysis. The first such dataset was obtained from the WHO and featured variables pertaining to new cases of COVID-19 infections as well as deaths attributable to the virus. As aforementioned, this data had been harvested since January 3rd 2020 up to the present. The WHO went about obtaining the information needed for this dataset by collecting data via official communications, which were conducted per International Health Regulations, and supplemented this with data reported by countries' governmental health agencies (this includes any information disclosed by these agencies on their websites and social media channels). During this process, the number of new cases and deaths catalogued were only recorded if these cases were laboratory-confirmed and fell within the WHO's definitions for confirmed cases of COVID-19. Specifically, in order to be classified as a new case of COVID-19, a person would have to have a positive Nucleic Acid Amplification Test result and/or be identified as being positive for the SARS-CoV-2 Antigen. Additional data apart from that reported by member countries was obtained from the European Center for Disease Prevention and Control. Possible sources of bias in the WHO dataset may include: bias on the part of health ministries in WHO member countries when reporting their data, lack of accurate reporting from WHO member countries with underdeveloped health systems and a lack of significant healthcare infrastructure, bias in the AI web-scraping algorithm used to harvest data from public press releases and social media posts made by governmental health agencies, etc.

The second dataset was obtained from the University of Oxford's COVID-19 Government response tracker and features several variables documenting the governmental response of countries to the COVID-19 pandemic. In total, 13 public health measures and 7 measures of economic policy were catalogued in this dataset. Each of these measures were assessed with a binary datapoint (1 or 0) depending on whether or not the policy indicated had been implemented by the country in question. In a few cases, a measure of "0.5" was used to denote partial implementation of a particular policy. Those of particular interest for this analysis included: obligation to wear masks in public settings, whether or not a public testing policy was in place, whether or not a domestic lockdown had been initiated, and whether restaurant closures had been mandated. Data for each of the public health measures catalogued by the dataset were obtained from the Assessment Capacities Project (ACAPs), the International Monetary Fund (IMF), and from information made publicly available by countries' governmental health agencies. Possible sources of bias in this dataset may include: bias on the part of governmental health agencies when publicly disclosing information about their policy responses to COVID-19, lack of coordinated policy measures implemented by countries with weakly structured governments, etc.

Citation: Hale, Thomas, Noam Angrist, Emily Cameron-Blake, Laura Hallas, Beatriz Kira, Saptarshi Majumdar, Anna Petherick, Toby Phillips, Helen Tatlow, Samuel Webster (2020). Oxford COVID-19 Government Response Tracker, Blavatnik School of Government.

Existing research published in the American Journal of Tropical Medicine and Hygiene as well as by the National Academy of Sciences in the United States has shown that COVID-19 transmission rates are forecasted to be 7.5 times higher in countries without a mask mandate as opposed to countries with one in place. And as expected, countries with a mask mandate have a lower COVID-19 related mortality rate than those that have not implemented such policies. Additionally, it was observed that the duration of time spent in lockdown was inversely correlated with COVID-19 related mortality rate, which is also in line with expectations. On the other hand, increased public testing for COVID-19 was shown to not have a statistically significant relationship with COVID-19 related mortality rate.

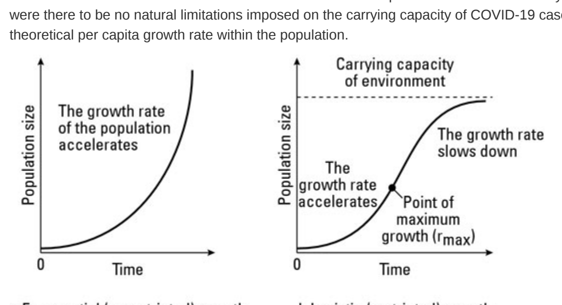
Citation: Leffler, Christoper T, et al. "Association of Country-Wide Coronavirus Mortality with Demographics, Testing, Lockdowns, and Public Wearing of Masks." The American Journal of Tropical Medicine and Hygiene, vol. 103, no. 6, 13 Aug. 2020, pp. 2400–2411., doi:https://doi.org/10.4269/ajtmh.20-1015.

Decriptive Analysis

A good way of measuring the spread of COVID-19 across a range of time is through logistic growth modeling (insert citation of study). Logistic growth is commonly used in fields like biology and ecology to model the growth of populations in an environment. The logistic growth equation used here (from the R package "growthcurver") takes the form of:

$$N_t = \frac{K}{1 + \left(\frac{K - N_0}{N_0}\right) e^{-rt}}$$

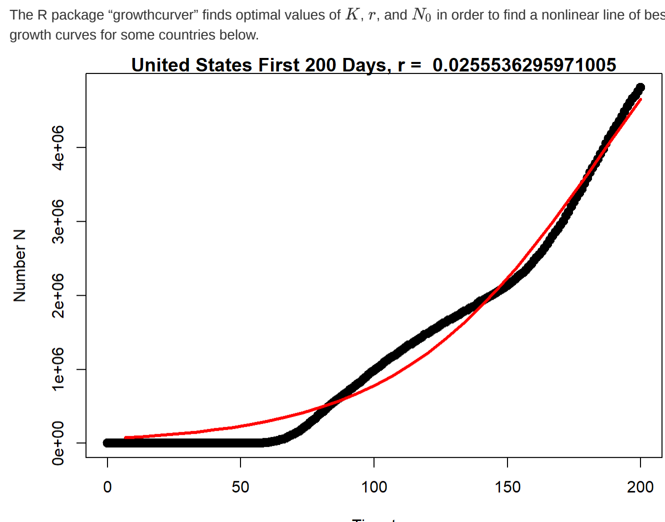
Where N_t describes the number of cumulative cases at time t. The parameter K is the theoretical maximum possible population size, which in this context describes the theoretical carrying capacity of COVID-19 cases in a given country. The parameter N_0 is the theoretical number of cumulative cases on day 0 for the fitted growth curve. Lastly, the parameter r is known as the intrinsic growth rate. It can be thought of as the maximum theoretical rate of increase of the number of cases per individual. That is to say, r represents the growth rate which would be observed were there to be no natural limitations imposed on the carrying capacity of COVID-19 cases in a given country. It represents the maximum theoretical per capita growth rate within the population.

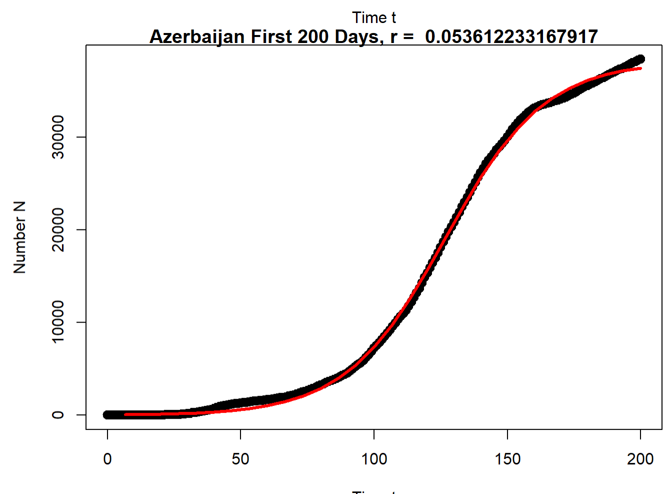


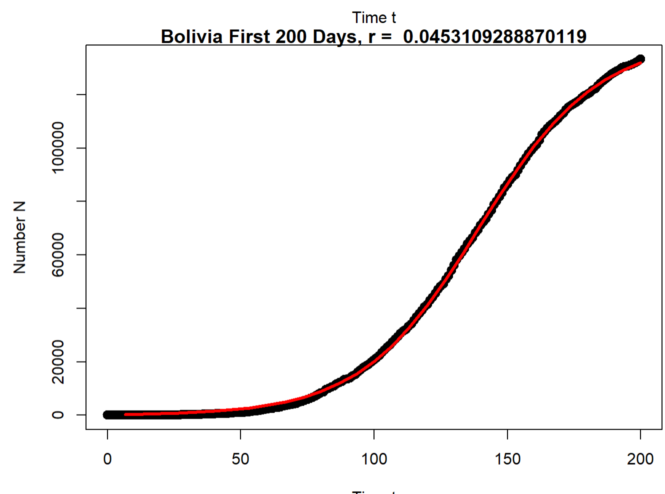
a Exponential (unrestricted) growth b Logistic (restricted) growth

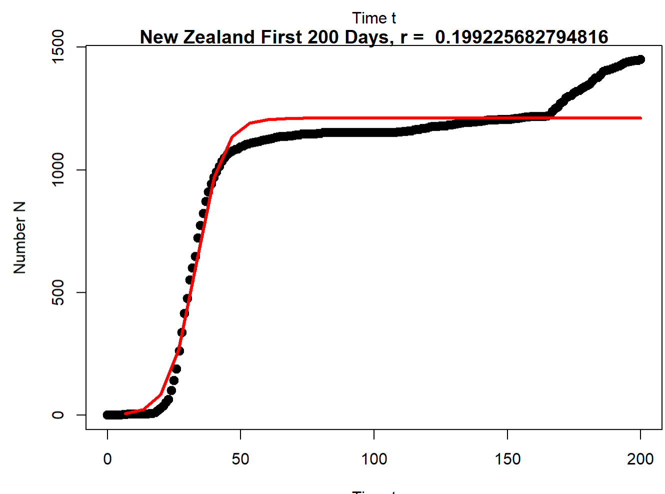
This image explains the reasoning behind using r as our outcome variable. You can see it controls the maximum slope, or maximum rate of growth, of the fitted logistic growth curve.

The R package "growthcurver" finds optimal values of K, r, and N_0 in order to find a nonlinear line of best fit. We can see some fitted logistic growth curves for some countries below.





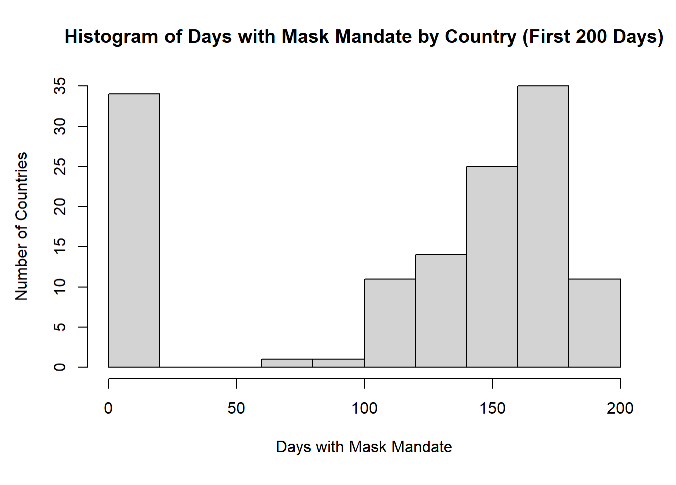




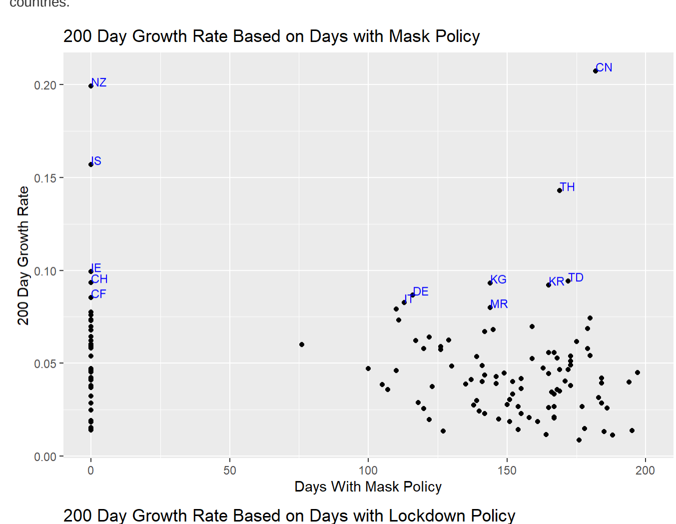
We can also observe the parameters which the growthcurver package fits to each country. For Denmark, for example, we obtain the parameters shown in the output below. Note that r, intrinsic growth rate, is reported as 0.087. This implies that the theoretical maximum rate of spread of COVID-19 in Denmark over the initial 200 day period was 0.087. Each person in the country would be expected to spread the disease to 0.087 other people during the height of contagion.

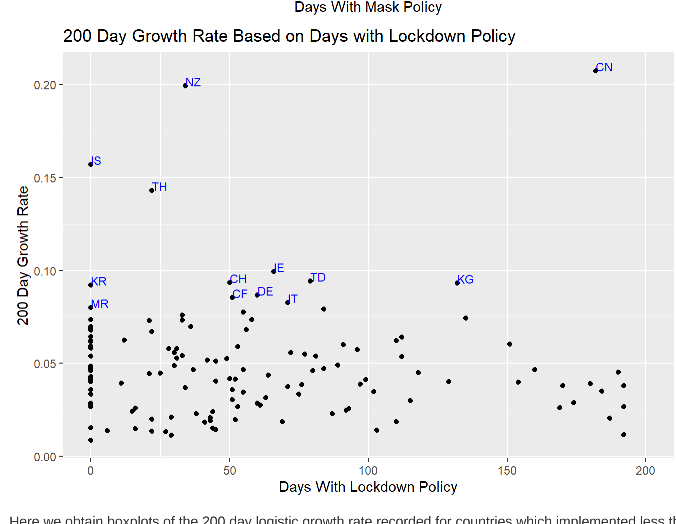
```
## Fit data to K / (1 + ((K - N0) / N0) * exp(-r * t)):
##      K   N0  r
##   val:   194735.012  359.292 0.087
##   Residual standard error: 10165.39 on 198 degrees of freedom
##
## Other useful metrics:
##   DT 1 / DT  auc_l   auc_e
##   8  1.2e-01 24792895.84 24533550.5
```

Visualizing a histogram of the number of days in which a mask mandate was in place may also shed insight into how different countries chose to respond to the initial onset of the virus. From the histogram, we observe that although the majority of countries implemented mask mandates for more than 100 days of the initial 200 day period, there were still a handful of countries which implemented very few days of mask mandates and/or none at all.

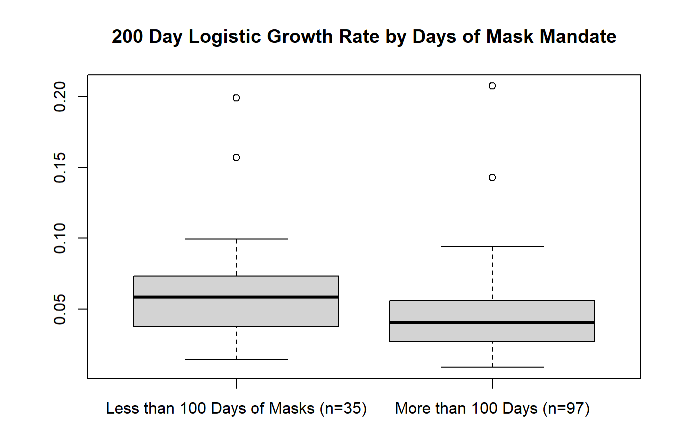


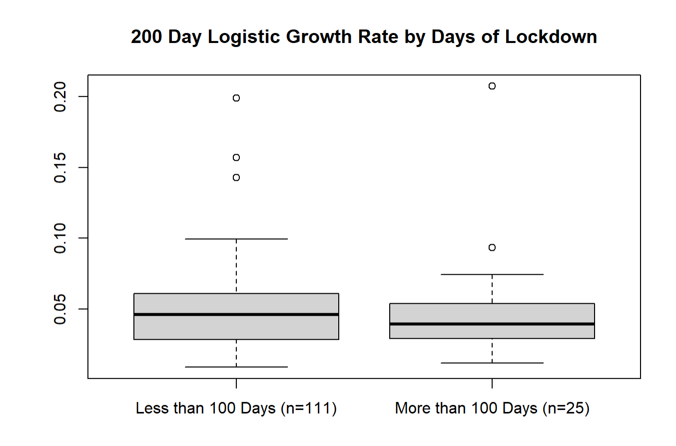
Here we can visualize 200 day growthrates based on the number of days mask mandates and lockdown policies were implemented in given countries.





Here we obtain boxplots of the 200 day logistic growth rate recorded for countries which implemented less than 100 days of certain policy measures and more than 100 days of certain policy measures within the 200 day time period.






```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.008744 0.028168 0.046033 0.049249 0.060924 0.199226
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01160 0.02892 0.03913 0.04773 0.05361 0.20742
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01401 0.03734 0.05829 0.05913 0.07324 0.19923
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.008744 0.026712 0.040257 0.045350 0.055703 0.207422
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Adding case mortality column

Inferential Analysis

Our primary question of interest is whether certain policies enacted at the national and sub-national level may be associated with different outcomes in the 200 day growth rate of cumulative new cases of COVID-19 as measured by a fitted logistic growth curve. The policies we have examined include mask wearing obligations, testing implementation, lockdown enforcement, and mandatory restaurant closures. In the Oxford dataset, which tracks daily COVID-19 cases as well as daily policy implementations through the use of indicator variables, we obtained the sum of the total number of days in which policy measures were in place over the 200 day period. Each policy indicator variable took a value of 1 if the given policy was implemented at the national or sub-national level on that day, and 0 if the policy was not enacted at both the national and sub-national level. The 200 day period is defined here with day 1 being the date on which the first case was recorded in the given country. This was done so as to ensure that measurement of growth rate was consistent for all countries in question. Although we should note that this approach fails to take into account other relevant factors, such as seasonal climate and population density.

The dataset which is implemented in the model contains 127 observations. In order to increase the reliability of the model, we decided to only take into account countries which recorded at least 800 cases by the 100th day after initial infection. We found that countries below this threshold tended to report numbers which were inconsistent enough (i.e. large jumps or gaps in reported data) to cause the fitted logistic growth curves to be unrepresentative of the actual spread of the disease. After doing this, we were left with 138 observations (countries) in total.

It is important to emphasize that the goal of this model is not to assess causality, but to determine what policy measures, if any, are associated with different outcomes in terms of the country's 200 day logistic growth rate, r.

A secondary question of interest then entails which of these different policies, if any, are associated with COVID-19 logistic growth rates measured over the 200 day interval.

For the purposes of answering these questions, we employed regression analysis. The outcome variable Y_i is defined as the 200 day logistic growth rate r as determined by a fitted logistic growth curve with the R package 'growthcurver'. The package determines the optimal parameters K, r, and N_0 (see the descriptive analysis section for explanation of these parameters) based on the logistic growth equation:

$$N_t = \frac{K}{1 + \left(\frac{K - N_0}{N_0}\right) e^{-rt}}$$

This equation is commonly used to model population growth of organisms in different environments.

We first employ a linear regression model of the form:

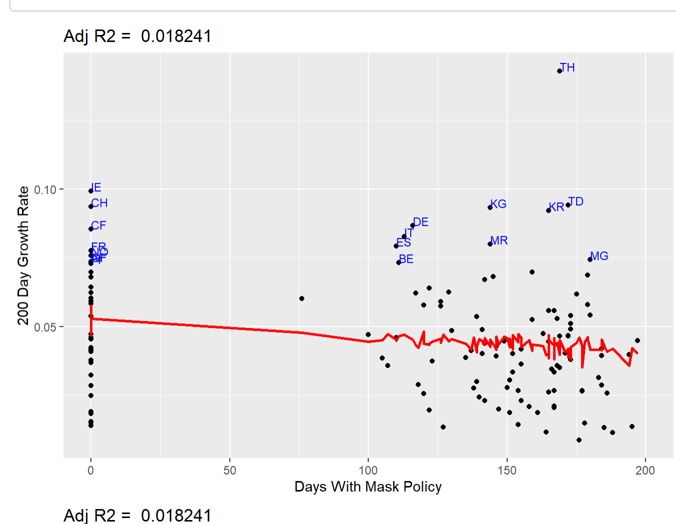
$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} \qquad i = 1, \ldots, 127$$

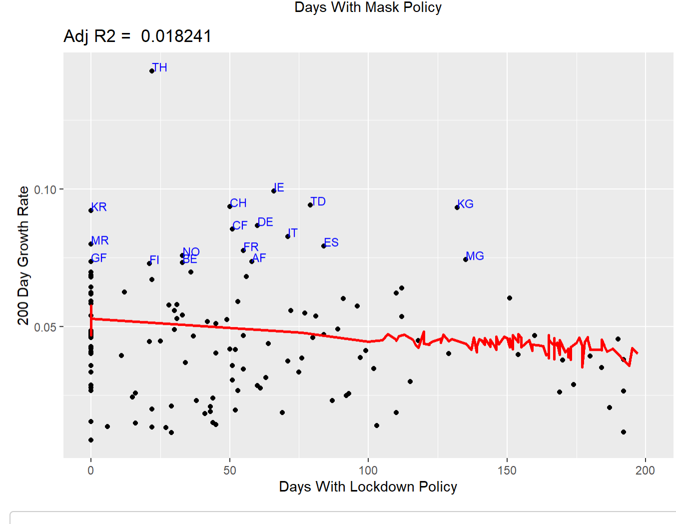
Where each vector $x_{1,i}, \ldots, x_{4,i}$ is the number of days for which the mask, lockdown, testing, and restaurant closure policies were implemented for the i-th country over the 200 day time period. The outcome variable is the parameter r for the i-th country as determined by that country's fitted logistic growth curve over the same 200 day time period.

The linear regression model relies on several key assumptions.

1. Linearity: The relationship between the response variable and the predictor variables is linear in parameters. This means that Y can be expressed as a linear combination of the parameters present in the model.
2. Homoskedasticity: Variance across the error terms is constant for every value of the response variable Y
3. Errors are independent and identically distributed: In addition to the assumption of constant variance across residuals, residuals must follow the distribution $\epsilon \sim N(0, \sigma_\epsilon^2)$
4. Variables are measured or observed reliably and without reporting error.

```
##
## Call:
## lm(formula = unlist(gr) ~ days_mask + days_testing + days_restaurant_close +
##     days_lockdown, data = test.data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.040691 -0.015811 -0.000893  0.012342  0.097998
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.604e-02  5.080e-03  11.031   <2e-16 ***
## days_mask             -6.067e-05  3.082e-05  -1.969   0.0513 .
## days_testing           9.163e-06  2.644e-05   0.346   0.7296
## days_restaurant_close -1.803e-05  3.179e-05  -0.567   0.5716
## days_lockdown         -3.474e-05  3.820e-05  -0.909   0.3649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02285 on 122 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.04941,    Adjusted R-squared:  0.01824
## F-statistic: 1.585 on 4 and 122 DF,  p-value: 0.1825
```

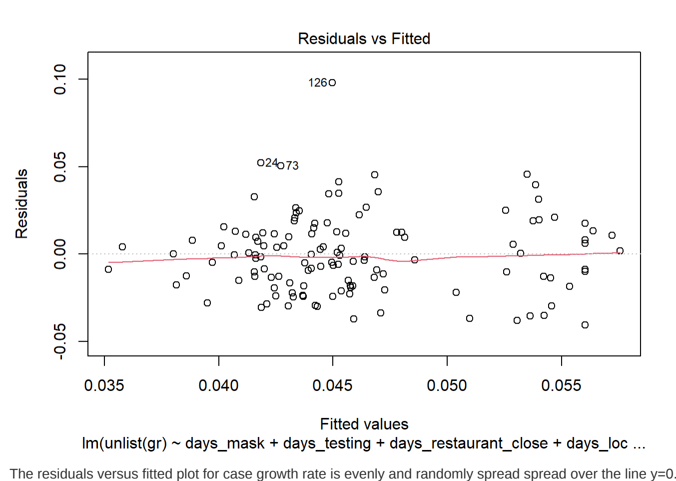





```
##                   2.5 %       97.5 %
## days_mask -0.0001216855 3.396577e-07
```

We find that the model which measures 200 day case growth rate finds mask policies to be significantly associated with the outcome at the $\alpha = 0.05$ significance level. In addition, the mask coefficient is negative, suggesting that implementation of mask mandates is associated with a lower 200 day growth rate.

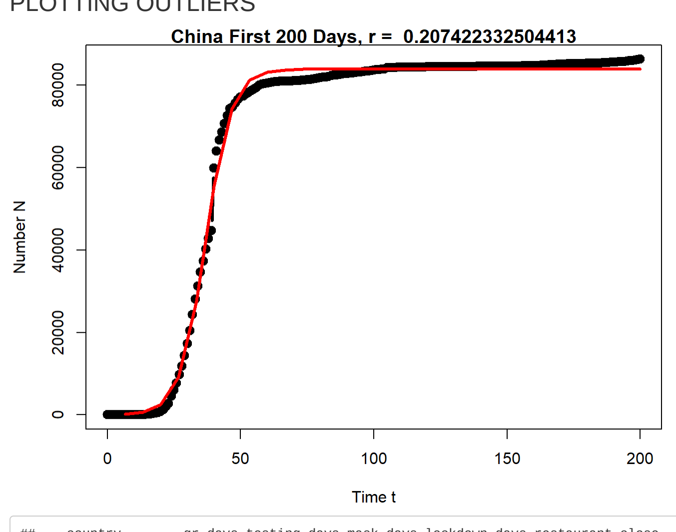
Sensitivity Analysis



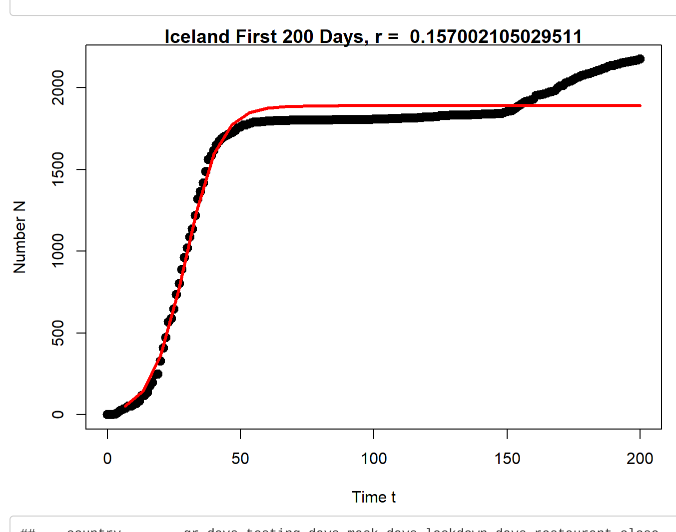
The residuals versus fitted plot for case growth rate is evenly and randomly spread spread over the line y=0. The spread widen slightly. And we have note the 3 outlier observations 64, 73 and 107. Overall based on this plot the model satisfies linearity and equal varaince of errors assumptions.

```
##    country        gr days_testing days_mask days_lockdown days_restaurant_close
## 26      CN 0.2074223          118       182           182                     0
```

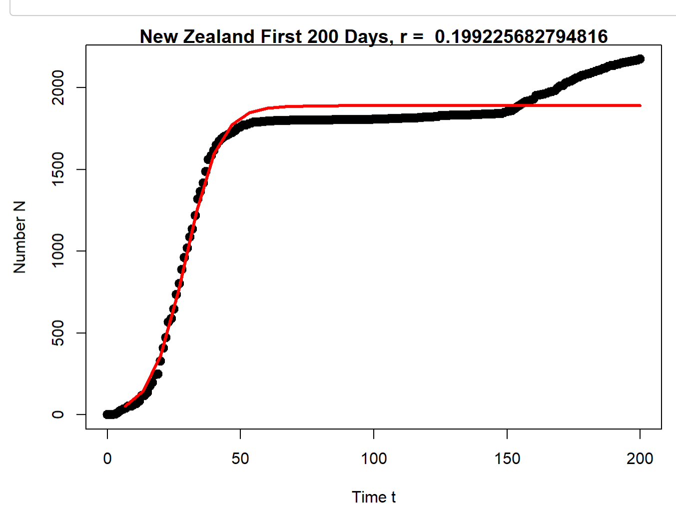
PLOTTING OUTLIERS

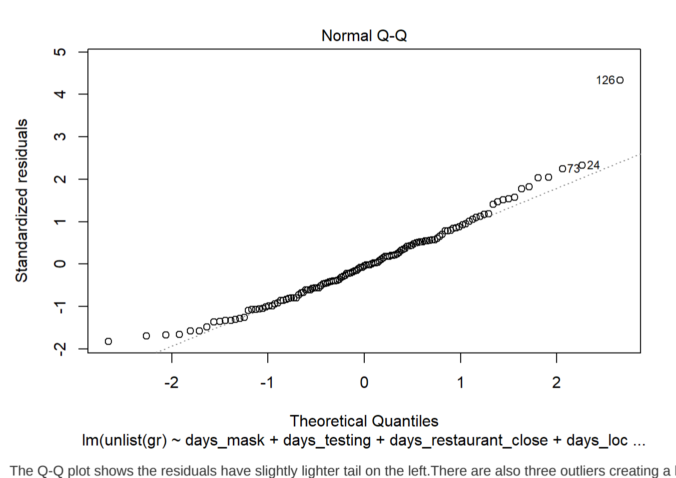



```
##    country        gr days_testing days_mask days_lockdown days_restaurant_close
## 59      IS 0.1570021          188         0             0                   177
```

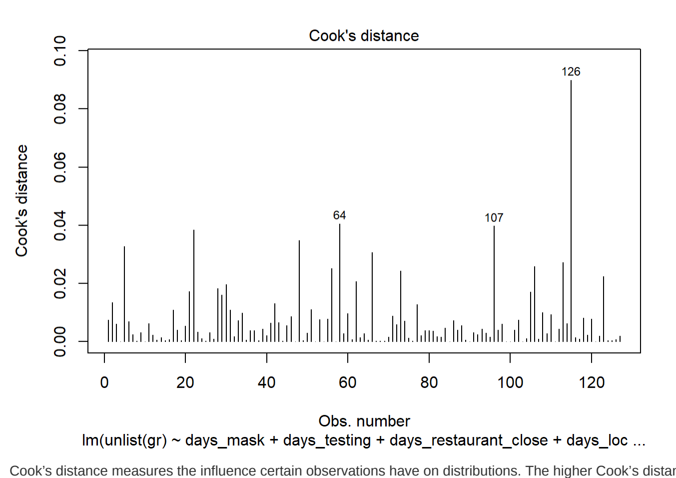



```
##    country        gr days_testing days_mask days_lockdown days_restaurant_close
## 91      NZ 0.1992257          168         0            34                    51
```





The Q-Q plot shows the residuals have slightly lighter tail on the left.There are also three outliers creating a heavier tail on the right. They are the same observations we saw in fitted versus residual plot. Two out of three outliers are very close to the line as are the rest of the observations. So the errors are normally distributed.



Cook's distance measures the influence certain observations have on distributions. The higher Cook's distance is the more influential the observation is. There are three observations that we noted in the above plots that also have a high Cook's distance. They are observations 64,73 and 107. The Cook's distance associated with these three was much lower than the three previous outliers, (China, Iceland, New Zealand) that were already removed from data.

Discussion & Conclusion

In this report we performed an analysis of two data sets obtained from the World Health Organization and the University of Oxford's COVID-19 Government response tracker. Our question of interest is surrounded around the 200 day cumulative COVID-19 case growth rate for several WHO member countries. Specifically, we want to study the differences in this growth rate amongst countries that implemented mask mandate, testing, lockdown policy, and/or restaurant closures during the 200 day timeframe versus those that didn't. The key question is if there is a statistically significant difference between the new case growth rates between the two groups.

To answer this question we first presented appropriate plots in the report. Among these plots were box plots that showed that when testing, masks, lockdown procedures, and/or restaurant closure mandates were implemented, the cumulative case growth over the 200 days being studied were much lower than when these policies were not put into place during this timeframe. To test for statistical significance we constructed a multiple variable general linear model to fit our data set. We set the new case proportion growth rate as the response variable, and we set whether masks mandates, testing, lockdown policies and restaurant closures were implemented as the independent variables. This allowed us to statistically assess whether or not the difference in cumulative new case growth rates were significant. We found that the number of days a country implemented mask mandates for were statistically significant at the $\alpha = 0.05$ significance level. However, it is important to note that this analysis finds association between mask mandates and lower disease spread, but that this relationship is not necessarily causal.

For future research and policy making, one can create a model where vaccination and immunity data is also taken into consideration and applied into the model. This model would produce more current up to date results as vaccinations have just begun to ramp up around the world, which means that one can hypothesize that new case growth would go down as distribution of vaccines increases.