Final Project 632 Rough Draft

Nic James, Sri Chandu, Thomas Li

05/21/2020

Contents

Abstract

We looked at data collected for the COVID-19 pandemic combined with movement tracking collected by Google to see if we could determine which policy measure showed the most promise in effecting deaths from COVID-19 using multiple linear regression. We started by analyzing the data using only the COVID-19 data set but could not find a linear regression that worked. So we decided to use predictors solely from the Google Mobility data to see there were any linear correlations with death from COVID-19. We found that movement to residential, retail and recreational areas were correlated with an increase in deaths and that movement to grocery, pharmacy, parks, workplaces, and transit stations was correlated with a decrease in deaths.

Problem and Motivation

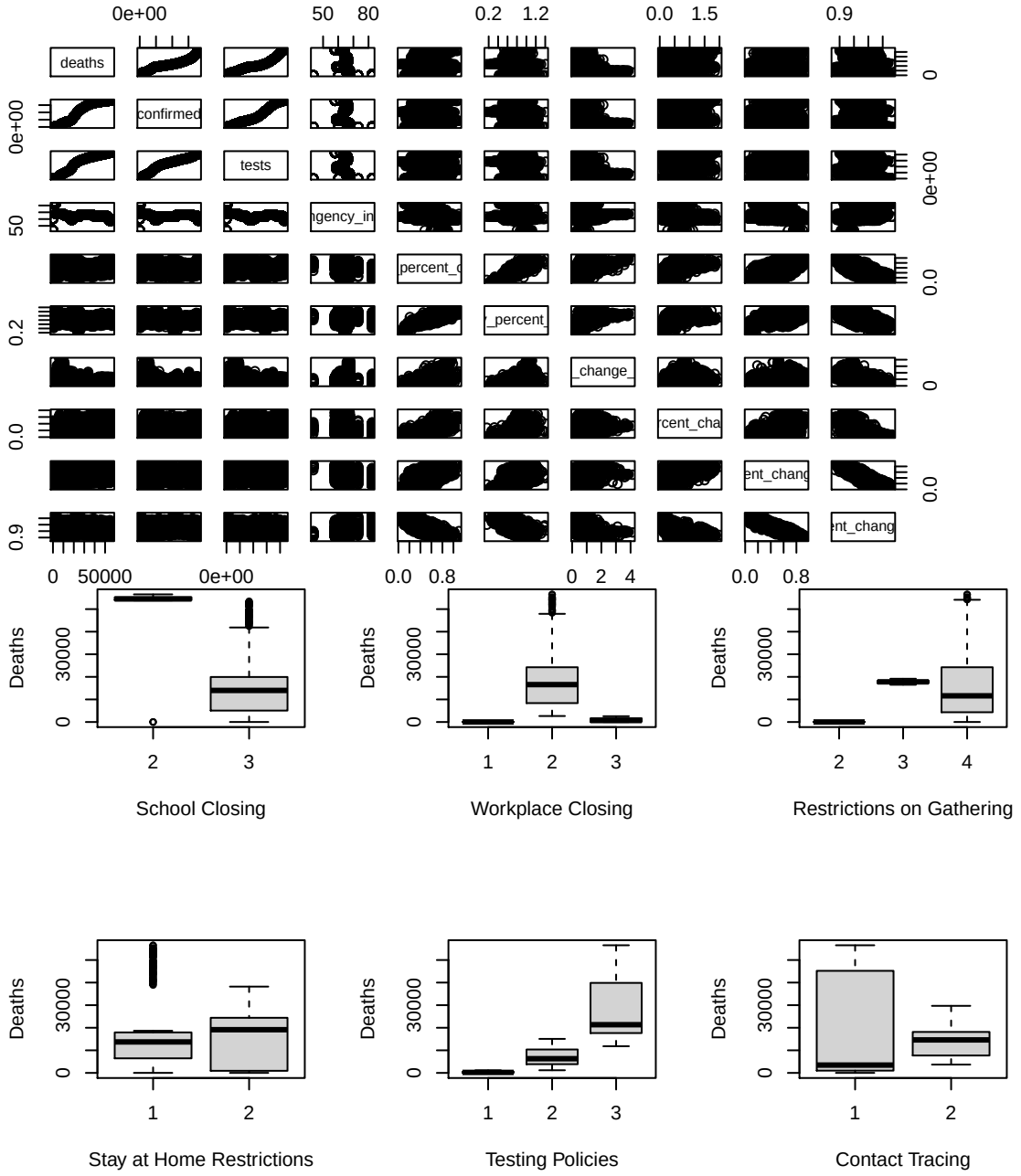
This past year our lives have changed drastically. We started school online instead of in person and have spent the majority of our year indoors, staying away from friends and family. We all know someone who has lost their lives to COVID-19 and we have experienced the effects of the virus daily. We wanted to use this data set because we were specifically interested in the Public Health measures like gathering restrictions, schools closing, workplace closing, and contact tracing during this pandemic. We wanted to see if we could find proof of how effective they were. We are hoping that by analyzing this data we will have a better understanding of which Public Health measures work the best. On a larger scale, analyses like this one will be extremely useful for future predictions and pandemics. Life styles and transportation have changed dramatically since the last worldwide pandemic and that was a hundred years ago. Having a recent event will help medical professionals to better understand the spread of this disease and hopefully make quicker and more effective decisions if this situation were to occur again in the future.

Data Description

This data set is a collection of governmental sources at national, regional, and city levels from 190 countries for COVID19. It includes time series of vaccines, test, cases, deaths, recovered, intensive therapy, and policy measures by Oxford COVID-19 Government Response Tracker. We will used the World Bank Google Mobility Reports as well. There are 16 variables in the base data set that we will be using for our regression. We will be limiting the location data strictly to California and using data from 3/15/2020 - 3/15/2021.

Our initial objective was to find out if running a linear regression of the Google Mobility data with the Covid-19 data had any significance in predicting the rate of deaths due to Covid-19. The Google mobility data recorded travel trends to categorized locations during the Covid-19 pandemic. This data is compared against a baseline reading; that is, the median value of each day of the week during a 5-week period (Jan 3 – Feb 6, 2020).

Variables used:
Original COVID-19 data: date, confirmed, tests, population, latitude, longitude, school_closing, workplace_closing, cancel_events, transport_closing, stay_home_restrictions, internal_movement_restrictions, international_movement_restrictions, information_campaigns, testing_policy, contact_tracing, stringency_index
World Bank data set: GDP per capita, GDP per capita growth, Poverty rate, Pollution in mcg
Google Mobility data set: retail_and_recreation_percent_change_from_baseline, grocery_and_pharmacy_percent_change_from_baseline, parks_percent_change_from_baseline, transit_stations_percent_change_from_baseline, workplaces_percent_change_from_baseline, residential_percent_change_from_baseline



Questions of Interest

1. What variables from the original data set have the most effect on deaths? We plan to use *deaths* as the response and *confirmed*, *tests*, *population*, *latitude*, *longitude*, *school_closing*, *workplace_closing*, *cancel_events*, *transport_closing*, *stay_home_restrictions*, *internal_movement_restrictions*, *international_movement_restrictions*, *information_campaigns*, *testing_policy*, *contact_tracing*, *stringency_index* from in the original COVID-19 data set as predictors to answer this question.

2. Are the trends in movement significantly correlated to deaths caused by COVID-19? We plan to use *deaths* as the response; *confirmed*, *tests*, *contact tracing*, and *stringency index* from in the original COVID-19 data set; and *retail and recreation percent change from baseline*, *grocery and pharmacy percent*

change from baseline, *parks percent change from baseline*, *transit stations percent change from baseline*, *workplaces percent change from baseline*, *residential percent change from baseline* from the Google Mobility data set to answer this question.

3. How does the economic profile of the country affect the mortality rate from COVID over the year 2020? What is the effect of air pollution (or exposure to air pollution) to the number of cases and the mortality rate from COVID? We plan to use *deaths* as the response; *confirmed*, *tests*, *contact tracing*, and *stringency index* from in the original COVID-19 data set; and *GDP*, *GDP per capita growth*, *Hospital beds per 1000*, *Poverty headcount Ratio*, *CO2 emissions per capita*, *Percentage of population exposed to high air pollution level*, *Annual percent Inflation*, *Total Reserves*, *Prevalence of Undernourishment*, and *Annual mean Air Pollution Exposure* from the World Bank data set to answer these questions.

Regression Analysis, Results and Interpretation

Important Details

Exploratory Analysis I:

We started by creating a data frame by filtering the data by California, and then by date. We ended with 365 rows of data for California. There is not any data for *vaccines*, *recovered*, *hosp*, *vent*, and *icu* so we removed these variables. We decided to use the following variables to create a new data frame date using the original COVID-19 data. We then turned the policy measures (categorical variables) into factors before we fitted a regression model. However factors need to have 2 or more levels in order to work so we also removed *cancel_events*, *international_movement*, and *transport_closing*. We also removed all variables that caused a singularity in the data. This left us with a base data set of *confirmed*, *tests*, *fschool_closing*, *fworkplace_closing*, *fgatherings_restrictions*, *fstay_home_restrictions*, *ftesting_policy*, *fcontact_tracing*, and *stringency_index* as the predictors to start looking for a linear regression model with.

We started by running a hypothesis test to see if we would prefer the null model against the full model. From Table 1 in Appendix 1, we can see that the p-value is $< 2.2e\text{-}16$ so we reject the null model as at least one predictor in the full model is significant.

Next we looked at the scatter plots for all of the variables. This was less useful since there were so many plots that it was difficult to see in detail (Plot 1, Appendix 1) so we looked at the scatter plots of the numerical data (Plot 2, Appendix 1) to see if there was anything that we could derive from the data. We assume that *confirmed* and *tests* would be a positive influence on the number of *deaths* because we see that there are positive linear relationships. The *stringency index* has a clear patterning to it that does not show any linear trends making it difficult to make any assumptions about it. We also should note that none of the variables are spread out. The data creates a line with the data points we will definitely need to transform this data to see if we can find a linear relationship. We then did an analysis of the categorical data using box plots (Plot 3, Appendix 1). From these we concluded that none of these variables have a constant variance and thus we will probably need to transform some if not all of the variables. We also looked at the added variable plots (Plot 4, Appendix 1) and summary to see if we should remove any variables. From the added variable plots we assumed that we will probably remove *testing policy*, *confirmed*, and *stringency index*. While *gathering restrictions 2* and *gathering restrictions 3* look like they should be removed, *gathering restrictions 4* looks to have some influence and therefore we chose to keep the *gathering restrictions*. However, the summary table shows us that *confirmed* and *gathering restrictions* will probably be removed. Which goes against what we previously thought when looking at only the scatter plots and the added variable plots. We will probably keep *testing policy* since only one of the dummy variables is not significant.

Next we did a variable selection using AIC and BIC stepwise selection (see Code 1-2, Appendix 1). We can see that the only variable that is not significant is a dummy variable for *testing policy* and we cannot remove it so we leave it in and we are left with a model of:

$deaths = \beta_0 + \beta_1 fstay.home.restrictions + \beta_2 tests + \beta_3 fworkplace.closing + \beta_4 fschool.closing + \beta_5 fcontact.tracing + \beta_6 ftesting.policy + \beta_7 stringency.index$

We looked at the residuals vs fitted and Q-Q plot (Plot 5, Appendix 1) to see if the linear assumptions were violated and to check to see if there were any outliers and/or leverage points. From the plots we can see that there is definite patterning in the residuals vs fitted plot and the Q-Q plot is heavy tailed showing violations of normality. We then run a powerTransform (Code 3, Appendix 1) on the numerical predictors and see that *tests* needs a cube root transformation and *stringency index* needs a logarithmic transformation. After we transform these predictors we checked to see if we needed to transform *deaths* and can see that the boxCox (Plot 6, Appendix 1) suggests a cube root transformation. This left us with a model of $\sqrt[3]{deaths} = \beta_0 + \beta_1 fstay.home.restrictions + \beta_2 \sqrt[3]{tests} + \beta_3 fworkplace.closing + \beta_4 fschool.closing + \beta_5 fcontact.tracing + \beta_6 ftesting.policy + \beta_7 log(stringency.index)$

After transformation we looked at the residual vs fitted and Q-Q plot again to check for linearity and check to see if there are still outliers. When we looked at the residuals vs fitted plot is still very patterned and the Q-Q plot is still heavy-tailed but less so (Plot 7, Appendix 1). We then used Cook's Distance plots (Plots 8-10, Appendix 1) and removed outliers (Code 5, Appendix 1)

Diagnostic Checks and Interpretation I:

After transformation and outlier/leverage points being removed we checked the summary (Table 4, Appendix 1) again and saw that *fworkplace_gathering_restrictions* was no longer significant and so we removed it and checked the linear assumptions again (Plot 11, Appendix 1). There was no difference seen. When we did a hypothesis test to see if we preferred the H_0 : mod.full6 or H_1 : mod.full5 (Code 6, Appendix 1), the p-value is 0.9087 showing us that we prefer the smaller model.

Our assumption that *confirmed* would have a strong effect on deaths was not correct. None of the data manipulation performed gave us a linear model as was seen by the clear patterning in the residuals vs. fitted plot. Because of this, we decided not to interpret this model.

Exploratory Analysis II:

a) Full Model (Base Covid + Google Mobility) Manipulating the Google mobility data into a workable data set resulted in the following predictors: *retail_and_recreation_percent_change_from_baseline*, *grocery_and_pharmacy_percent_change_from_baseline*, *parks_percent_change_from_baseline*, *transit_stations_percent_change_from_baseline*, *workplaces_percent_change_from_baseline*, and *residential_percent_change_from_baseline*. We then combined them with the full model from the base data set to create our new full model.

We ran an ANOVA comparing the full model with the null model (Table 1, Appendix 2a). The resulting p-value was <2.2e-16 so we reject the null model and conclude that there is at least one predictor in the full model that is significant.

Because the current model did not meet our assumptions of linearity, we decided to run a variable selection to choose a better model (Table 2, Appendix 2a). We compared eight models by looking at adjusted R-squared, CP values, and BIC values (Table 3, Appendix 2a). The model we chose had seven predictor variables. (Table 1-2, Plot 1-2, Appendix 2a)

Next, we ran transformations of all non-factor predictors. Resulting in a square root transformation for *confirmed* (Table 4, Appendix 2a). We also ran a transformation for the response variable which resulted in a square root transformation for *deaths* (Plot 3, Appendix 2a).

Next, we checked for outliers and high leverage points. We set $|r_i| > 2$ and plotted the data to identify bad leverage points (Plot 4, Appendix 2a). A plot to check for high Cook's distance values showed none were greater than 0.5 (Plot 5, Appendix 2a). As a result, we decided not to remove any data points.

Our final model does not include any Google mobility predictors. The response is $\sqrt{deaths}$ and the predictors are $\sqrt{tests}$, fschool_closing, fworkplace_closing, fstay_home_restrictions, ftesting_policy, and fcontact_tracing. Checking the final diagnostics plots, we see that the assumptions of linearity are not met (Plot 6, Appendix 2a).

For our next model, we decided to remove all base COVID-19 numerical data from the original data set. The model included categorical base COVID-19 data and Google mobility data. After exploring this model, we came to the same conclusion as the previous model, that assumptions of linearity were not met. Because of these results, we decided to remove all base COVID-19 predictors and work solely with Google mobility data.

We looked at adding a categorical variable to our model however the linear assumptions kept getting worse so we chose to leave the categorical variables out of our model and changed our questions to match this.

b) Full Model (Google Mobility Numerical Variables only) We did an analysis to see if we could find a linear regression of the Google Mobility variables when deaths was a response since the model when combined with the Covid-19 categorical variables removed all the Google Mobility variables. So, in our analysis we started off by doing an inital summary, our inital plots also showed that Linearity assumptions were not satisfied. It also showed that Workplace change was not significant, but we kept this and found that once we completed the transformations and removed Influential points, the variable did become significant. (Table 1, Appendix 2B) (Plot1, Appendix 2B).

Because the current model did not meet our assumptions of linearity, we decided to run a variable selection to help us narrow down significant predictors. We compared all models by looking at adjusted R-squared and CP values. (Table 2,Appendix 2B) The model we chose had all 6 predictor variables. Next, we decided to check if there were any necessary transformations for the predictors. We found that 5 out of 6 variables needed to be transformed.(Table 3, Appendix 2B).The assumptions were still not met (Plot 2, Appendix 2B). Then we needed to see if the Response variable needed to be transformed, and found that it required a square root transformation (Plot 3, Appendix 2B). When we checked linearity assumptions we found that the QQ plot was much better, even though the residualsvsFitted stayed approximately the same (Plot 4, Appendix 2B). We also found that once all transformations were done then the workplace variable that was initally not significant, stayed significant. (Table 4, Appendix 2B).

Lastly, we looked at outliers and found that there were some points that needed to be removed (Plot 5, Appendix 2B. Once we removed these, we checked our linear assumptions (Plot 6, Appendix 2b). We also checked to make sure that the removal of the outliers and high leverage points made our model better, and they did! (Plot 7, Appendix 2b).

Diagnostic Checks II:

And this plot shows us that we have a linear model and all of the assumptions are met as well as our data can.

Interpretation II:

$Y_{ijk} = \beta_1 var1 + \beta_2 var2$

Exploratory Analysis III:

For the World Bank Data set, we only had one annual data point for each country. This meant that we had the same data point for every day which caused a singularity when we tried to analyze the data and combine it with the original COVID data set. We also did not have state data which is what we wanted to analyze. At the end, we only had the country by country data and could only look at one day and found

that most of the variables were consistently lacking significant amounts of data points, which led to serious issues with the reliability of the data. Eventually we decided to only focus on the base data set and combine it with the Google Mobility tracking data.

Conclusions (200 words) - Thomas

The young man wanted a role model. He looked long and hard in his youth, but that role model never materialized. His only choice was to embrace all the people in his life he didn't want to be like. The young man wanted a role model. He looked long and hard in his youth, but that role model never materialized. His only choice was to embrace all the people in his life he didn't want to be like. The young man wanted a role model. He looked long and hard in his youth, but that role model never materialized. His only choice was to embrace all the people in his life he didn't want to be like.

The young man wanted a role model. He looked long and hard in his youth, but that role model never materialized. His only choice was to embrace all the people in his life he didn't want to be like. The young man wanted a role model. He looked long and hard in his youth, but that role model never materialized. His only choice was to embrace all the people in his life he didn't want to be like. The young man wanted a role model. He looked long and hard in his youth, but that role model never materialized. His only choice was to embrace all the people in his life he didn't want to be like.

Appendicies

Appendix 1: R Code for Original COVID-19 Data Set

Code 1: The AIC model kept fstay_home_restrictions, tests, fworkplace_closing, fschool_closing, fcontact_tracing, ftesting_policy, and stringency_index as the predictors in the ideal model. data

```
##
## Call:
## lm(formula = deaths ~ fstay_home_restrictions + tests + fworkplace_closing +
##     fschool_closing + fcontact_tracing + ftesting_policy + stringency_index,
##     data = fbase_data)
##
## Coefficients:
##              (Intercept)  fstay_home_restrictions2                     tests
##                -6.026e+03                -6.378e+03                 1.133e-03
##      fworkplace_closing2       fworkplace_closing3          fschool_closing3
##                 5.127e+03                 8.405e+03                -5.347e+03
##        fcontact_tracing2          ftesting_policy2          ftesting_policy3
##                -1.676e+03                 9.524e+02                -2.803e+01
##         stringency_index
##                 1.162e+02
```

Code 2: The BIC is the same as the AIC, we chose to use BIC.

```
##
## Call:
## lm(formula = deaths ~ fstay_home_restrictions + tests + fworkplace_closing +
##     fschool_closing + fcontact_tracing + ftesting_policy + stringency_index,
##     data = fbase_data)
##
## Coefficients:
##              (Intercept)  fstay_home_restrictions2                     tests
##                -6.026e+03                -6.378e+03                 1.133e-03
##      fworkplace_closing2       fworkplace_closing3          fschool_closing3
##                 5.127e+03                 8.405e+03                -5.347e+03
##        fcontact_tracing2          ftesting_policy2          ftesting_policy3
##                -1.676e+03                 9.524e+02                -2.803e+01
##         stringency_index
##                 1.162e+02
```

Code 3: The powerTransformation suggests we use a cube root transformation on *tests* and a logarithmic transformation on stringency_index.

```
## bcPower Transformations to Multinormality
##                   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## tests               0.2715        0.27       0.2178       0.3252
## stringency_index   -0.4417        0.00      -1.0026       0.1193
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                               LRT df       pval
```

```
## LR test, lambda = (0 0) 135.8532  2 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                               LRT df       pval
## LR test, lambda = (1 1) 420.7194  2 < 2.22e-16
```

Code 4: There are two leverage points and 3 outliers.

```
## 1 2
## 1 2
```

```
## 33 34 35
## 33 34 35
```

```
##   1   2   3   4   5   6   7   8   9  10  11  29  30  31  32  33  34  35 352 353
##   1   2   3   4   5   6   7   8   9  10  11  29  30  31  32  33  34  35 352 353
## 354 355
## 354 355
```

Code 5: Remove all leverage and outliers from fbase_data.

Code 6:

```
## Analysis of Variance Table
##
## Model 1: (deaths^{
##     1/3
## }) ~ fstay_home_restrictions + I(tests^{
##     1/3
## }) + fschool_closing + fcontact_tracing + ftesting_policy + log(stringency_index)
## Model 2: (deaths^{
##     1/3
## }) ~ fstay_home_restrictions + I(tests^{
##     1/3
## }) + fworkplace_closing + fschool_closing + fcontact_tracing +
##     ftesting_policy + log(stringency_index)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    352 103.02
## 2    351 103.02  1 0.0038684 0.0132 0.9087
```

Table 1: $H_0 : \beta_1 = \beta_2 = ... = \beta_9 = 0$
H_1 : At least one $\beta_i \neq 0$ for $i = 1, 2, ..., 9$

```
## Analysis of Variance Table
##
## Model 1: deaths ~ 1
## Model 2: deaths ~ confirmed + tests + fschool_closing + fworkplace_closing +
##     fgatherings_restrictions + fstay_home_restrictions + ftesting_policy +
##     fcontact_tracing + stringency_index
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
```

```
## 1    364 8.1640e+10
## 2    352 4.7587e+08 12 8.1164e+10 5003.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2: This summary table shows us that confirmed and gathering restrictions will probably be removed. We may keep testing policy since only one of the dummy variables is not significant.

```
##
## Call:
## lm(formula = deaths ~ confirmed + tests + fschool_closing + fworkplace_closing +
##     fgatherings_restrictions + fstay_home_restrictions + ftesting_policy +
##     fcontact_tracing + stringency_index, data = fbase_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3917.7  -536.0  -114.1   506.1  4360.5
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -6.059e+03  3.202e+03  -1.892  0.05927 .
## confirmed                  -3.993e-04  9.071e-04  -0.440  0.66009
## tests                       1.169e-03  8.586e-05  13.614  < 2e-16 ***
## fschool_closing3           -5.411e+03  5.241e+02 -10.325  < 2e-16 ***
## fworkplace_closing2         6.002e+03  1.330e+03   4.511 8.79e-06 ***
## fworkplace_closing3         9.029e+03  1.649e+03   5.476 8.29e-08 ***
## fgatherings_restrictions3 -2.429e+03  1.921e+03  -1.265  0.20685
## fgatherings_restrictions4 -2.317e+03  2.002e+03  -1.157  0.24806
## fstay_home_restrictions2  -6.471e+03  2.440e+02 -26.518  < 2e-16 ***
## ftesting_policy2           9.374e+02  3.378e+02   2.775  0.00582 **
## ftesting_policy3           1.705e+01  4.664e+02   0.037  0.97085
## fcontact_tracing2         -1.732e+03  3.069e+02  -5.643 3.44e-08 ***
## stringency_index           1.391e+02  6.855e+01   2.030  0.04314 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1163 on 352 degrees of freedom
## Multiple R-squared:  0.9942, Adjusted R-squared:  0.994
## F-statistic:  5003 on 12 and 352 DF,  p-value: < 2.2e-16
```

Table 3: This is a summary of the model found after BIC Stepwise Selection.

```
##
## Call:
## lm(formula = deaths ~ fstay_home_restrictions + tests + fworkplace_closing +
##     fschool_closing + fcontact_tracing + ftesting_policy + stringency_index,
##     data = fbase_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3925.3  -548.4  -118.5   559.9  4361.3
##
```

```
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -6.026e+03  1.972e+03  -3.056 0.002410 **
## fstay_home_restrictions2 -6.378e+03  2.051e+02 -31.101  < 2e-16 ***
## tests                     1.133e-03  1.147e-05  98.840  < 2e-16 ***
## fworkplace_closing2       5.127e+03  1.092e+03   4.695 3.82e-06 ***
## fworkplace_closing3       8.405e+03  1.217e+03   6.905 2.33e-11 ***
## fschool_closing3         -5.347e+03  4.619e+02 -11.576  < 2e-16 ***
## fcontact_tracing2        -1.676e+03  2.356e+02  -7.112 6.34e-12 ***
## ftesting_policy2          9.524e+02  3.350e+02   2.843 0.004723 **
## ftesting_policy3         -2.803e+01  4.298e+02  -0.065 0.948043
## stringency_index          1.162e+02  3.457e+01   3.360 0.000863 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1161 on 355 degrees of freedom
## Multiple R-squared:  0.9941, Adjusted R-squared:  0.994
## F-statistic:  6691 on 9 and 355 DF,  p-value: < 2.2e-16
```

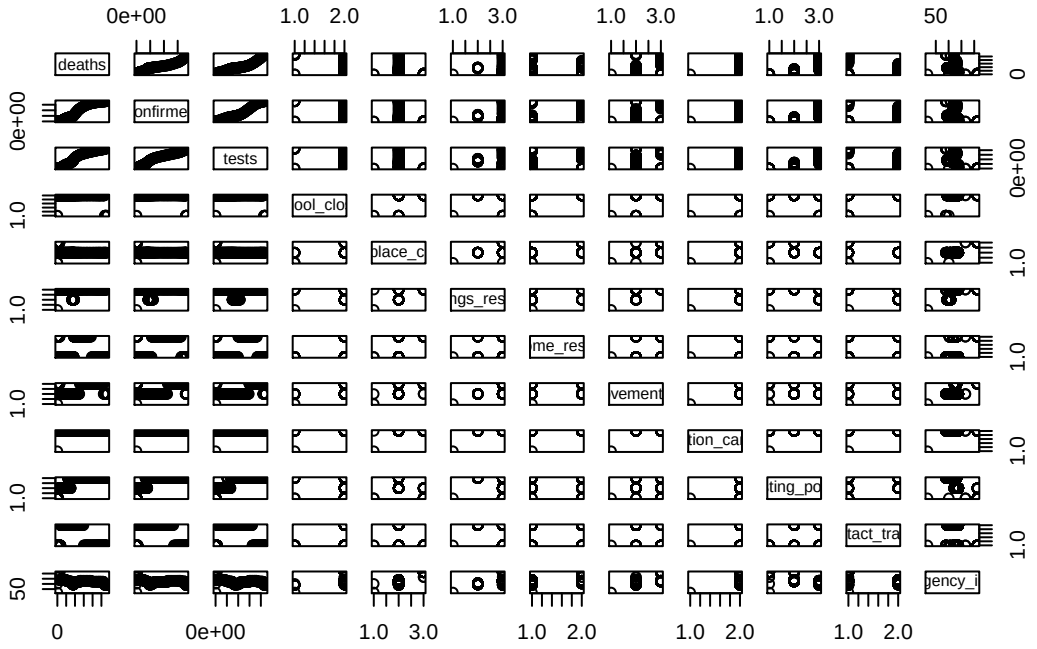
Table 4:

```
##
## Call:
## lm(formula = (deaths^{
##     1/3
## }) ~ fstay_home_restrictions + I(tests^{
##     1/3
## }) + fworkplace_closing + fschool_closing + fcontact_tracing +
##     ftesting_policy + log(stringency_index), data = fbase_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23640 -0.31719 -0.04829  0.21613  1.90884
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -21.143270   4.685643  -4.512 8.76e-06 ***
## fstay_home_restrictions2  -1.875099   0.095475 -19.640  < 2e-16 ***
## I(tests^{\n    1/3\n})     0.089391   0.000846 105.659  < 2e-16 ***
## fworkplace_closing3       -0.030021   0.261497  -0.115    0.909
## fschool_closing3          -2.000958   0.217052  -9.219  < 2e-16 ***
## fcontact_tracing2         -1.612712   0.092226 -17.487  < 2e-16 ***
## ftesting_policy2           3.274730   0.162020  20.212  < 2e-16 ***
## ftesting_policy3           3.402234   0.216109  15.743  < 2e-16 ***
## log(stringency_index)      6.000421   1.126961   5.324 1.81e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5418 on 351 degrees of freedom
## Multiple R-squared:  0.996,  Adjusted R-squared:  0.996
## F-statistic: 1.106e+04 on 8 and 351 DF,  p-value: < 2.2e-16
```

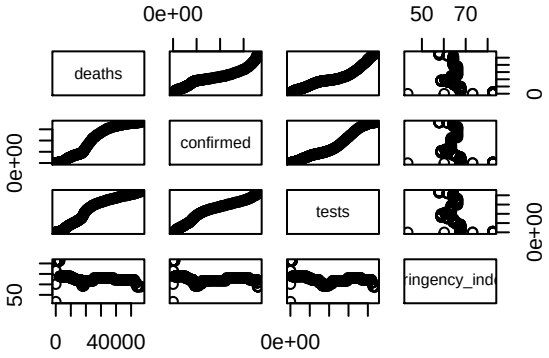
Table 5: This is our final model.

```
##
## Call:
## lm(formula = (deaths^{
##     1/3
## }) ~ fstay_home_restrictions + I(tests^{
##     1/3
## }) + fschool_closing + fcontact_tracing + ftesting_policy + log(stringency_index),
##     data = fbase_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23530 -0.31336 -0.04861  0.21695  1.90872
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2.085e+01  3.928e+00  -5.308 1.97e-07 ***
## fstay_home_restrictions2   -1.877e+00  9.370e-02 -20.034  < 2e-16 ***
## I(tests^{\n    1/3\n})      8.943e-02  7.853e-04 113.871  < 2e-16 ***
## fschool_closing3           -1.992e+00  2.034e-01  -9.796  < 2e-16 ***
## fcontact_tracing2          -1.611e+00  9.136e-02 -17.638  < 2e-16 ***
## ftesting_policy2            3.281e+00  1.527e-01  21.481  < 2e-16 ***
## ftesting_policy3            3.402e+00  2.158e-01  15.765  < 2e-16 ***
## log(stringency_index)       5.925e+00  9.165e-01   6.465 3.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.541 on 352 degrees of freedom
## Multiple R-squared:  0.996,  Adjusted R-squared:  0.996
## F-statistic: 1.267e+04 on 7 and 352 DF,  p-value: < 2.2e-16
```

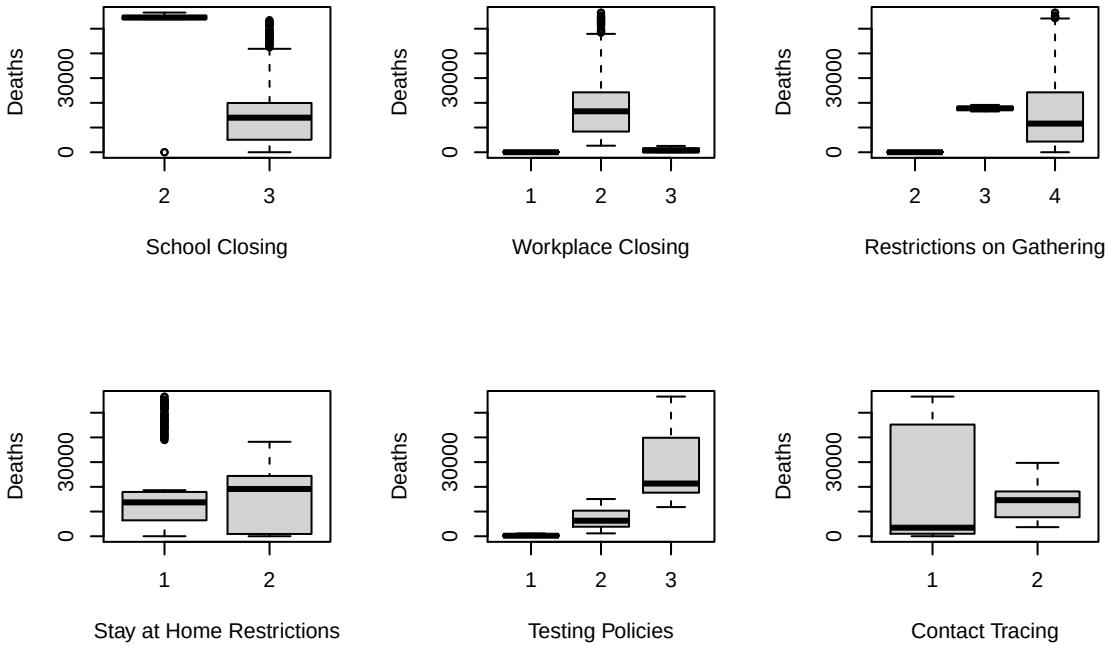
Plot 1: Each of these are really small and it is hard to derive anything useful from them.



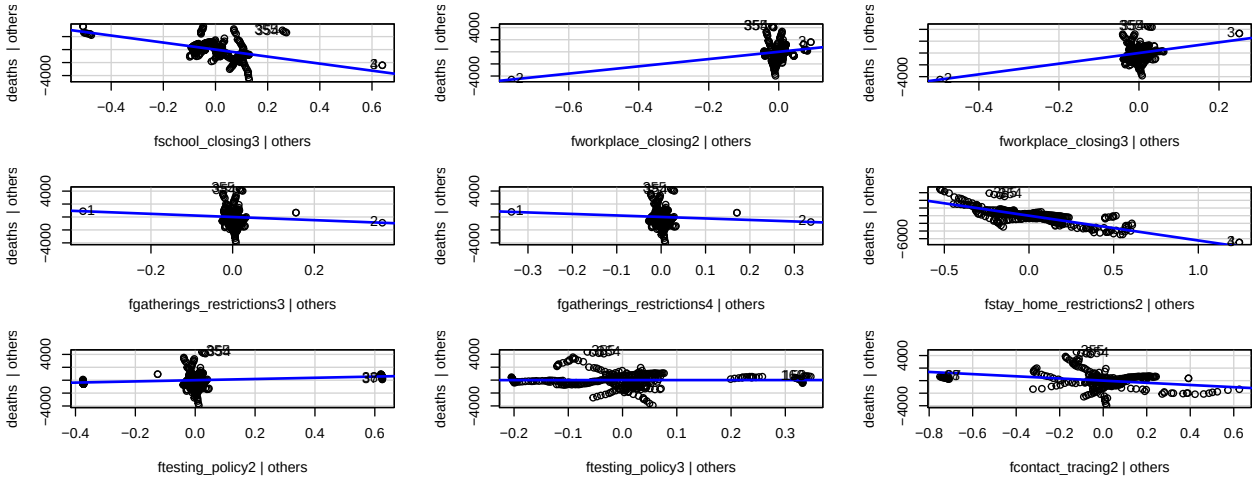
Plot 2: Scatterplots of the numerical variables in the original data set.

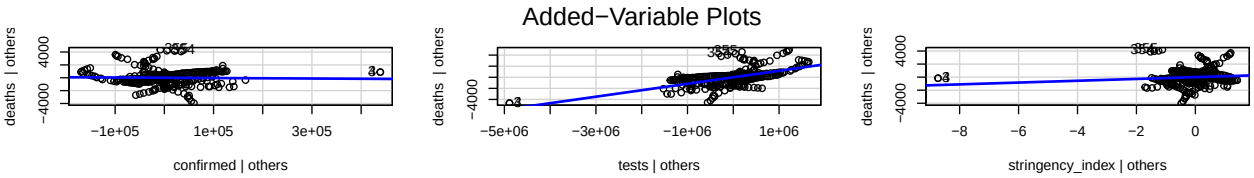


Plot 3: Box plots of the categorical variables in the original data set.

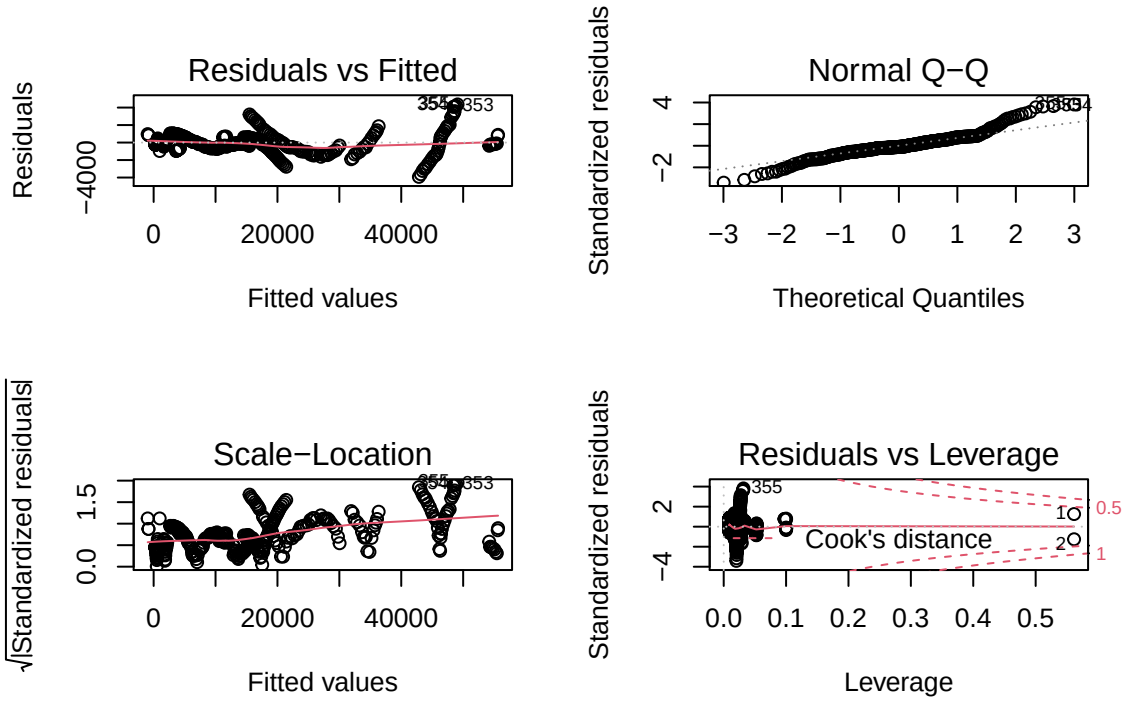


Plot 4: Added variable plots for both the categorical variables.

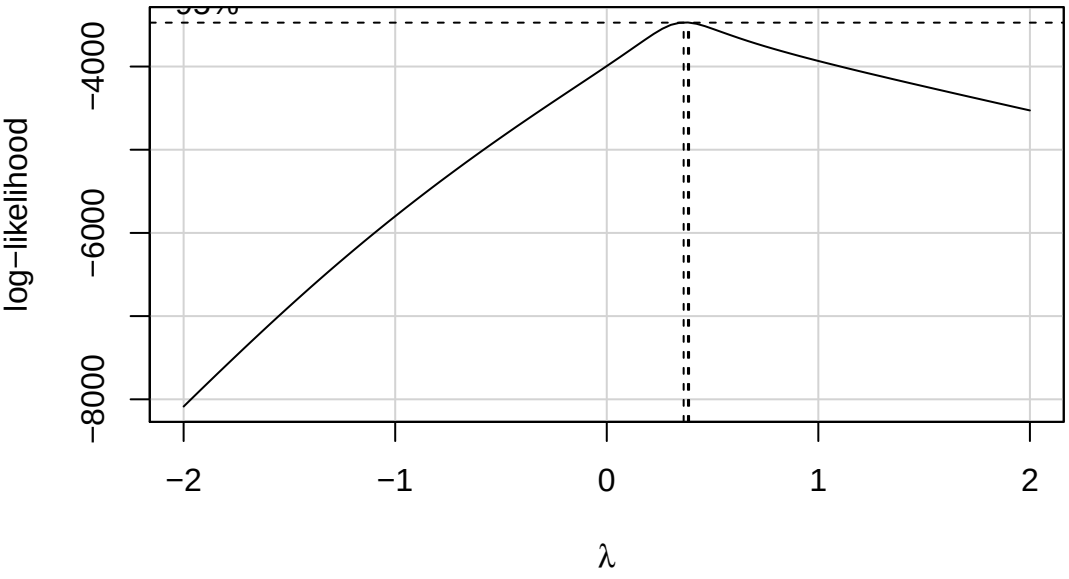




Plot 5: Checking normality prior to transformation

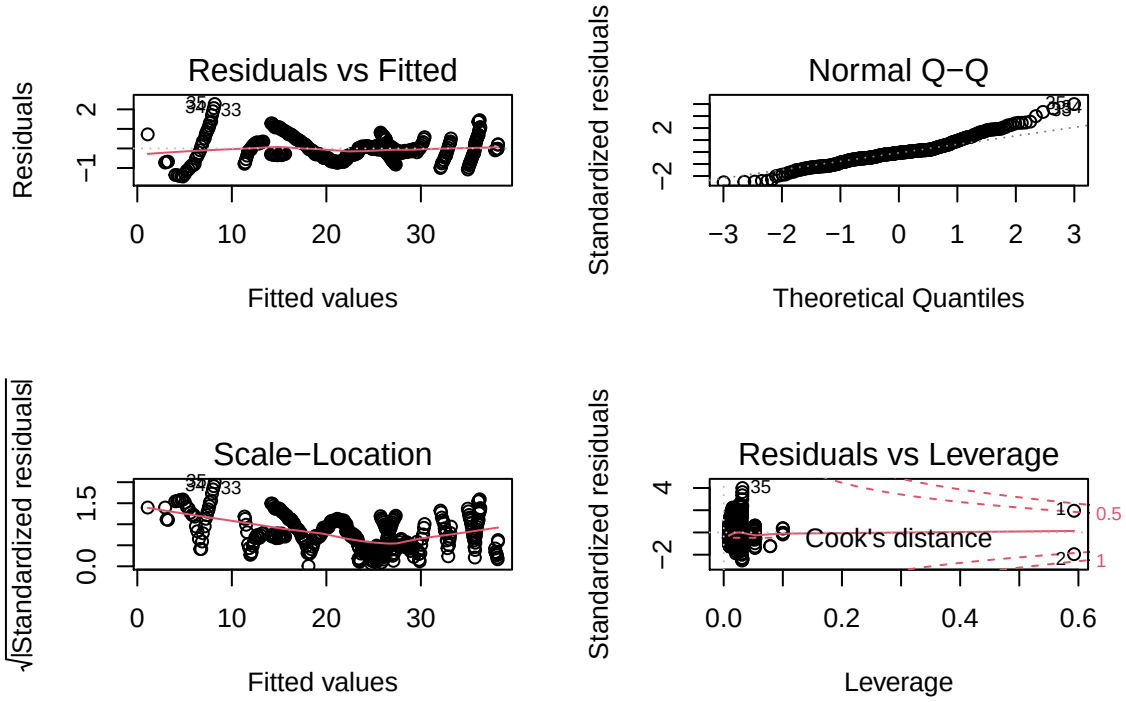


Plot 6: According to boxCox we should do a cube root transformation on the response *deaths*.

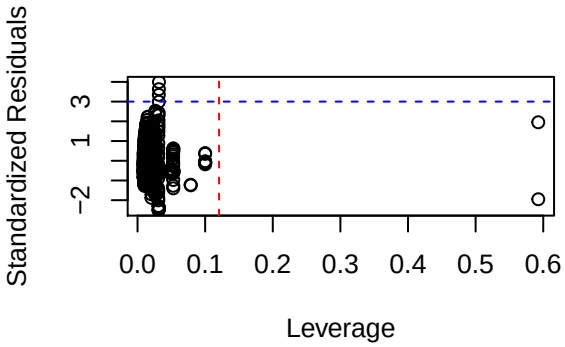


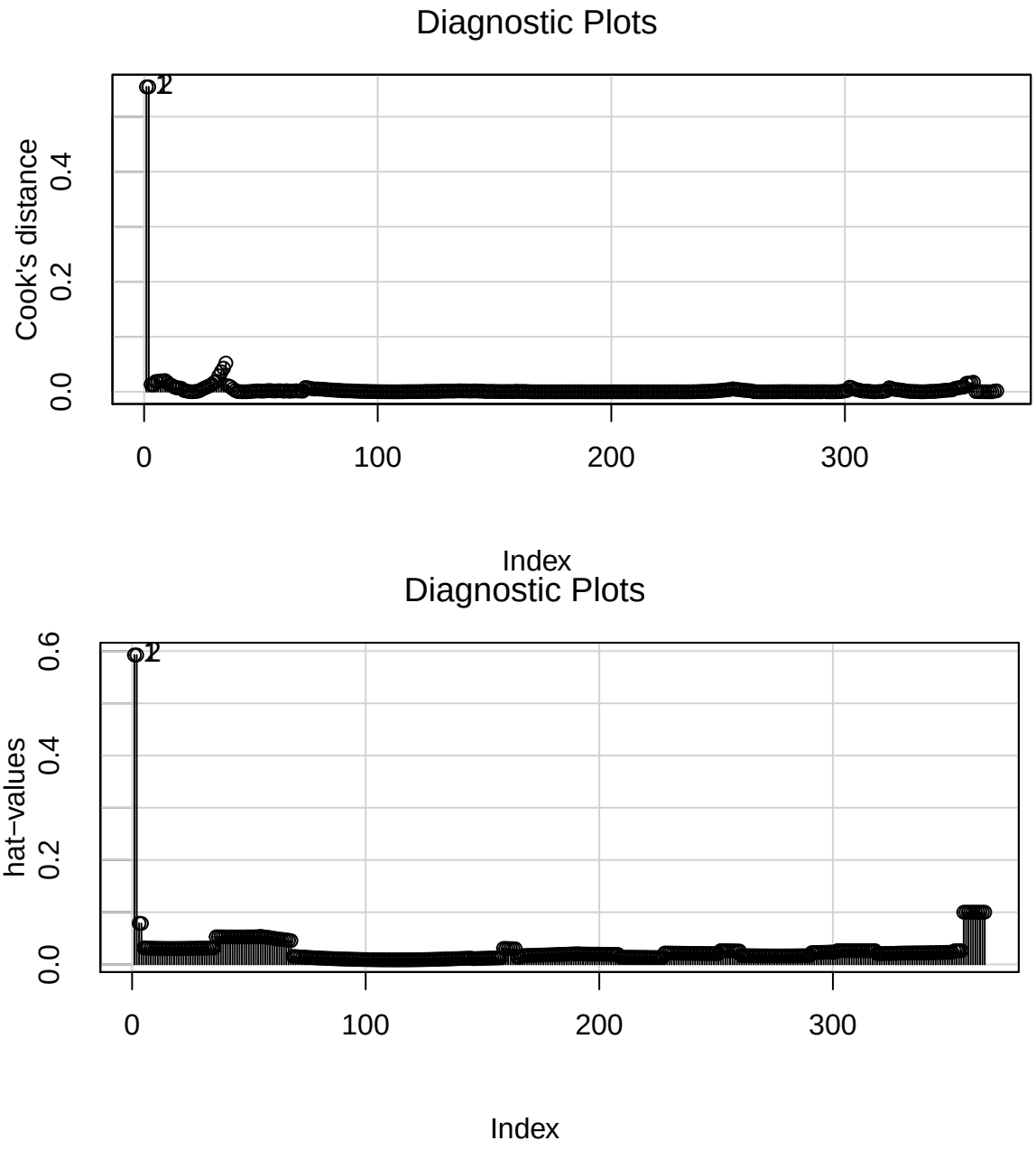

```
## [1] 0.3838384
```

Plot 7: Post transformation linearity check. The residuals vs fitted plot is still very patterned and the Q-Q plot is still heavy-tailed but less so.

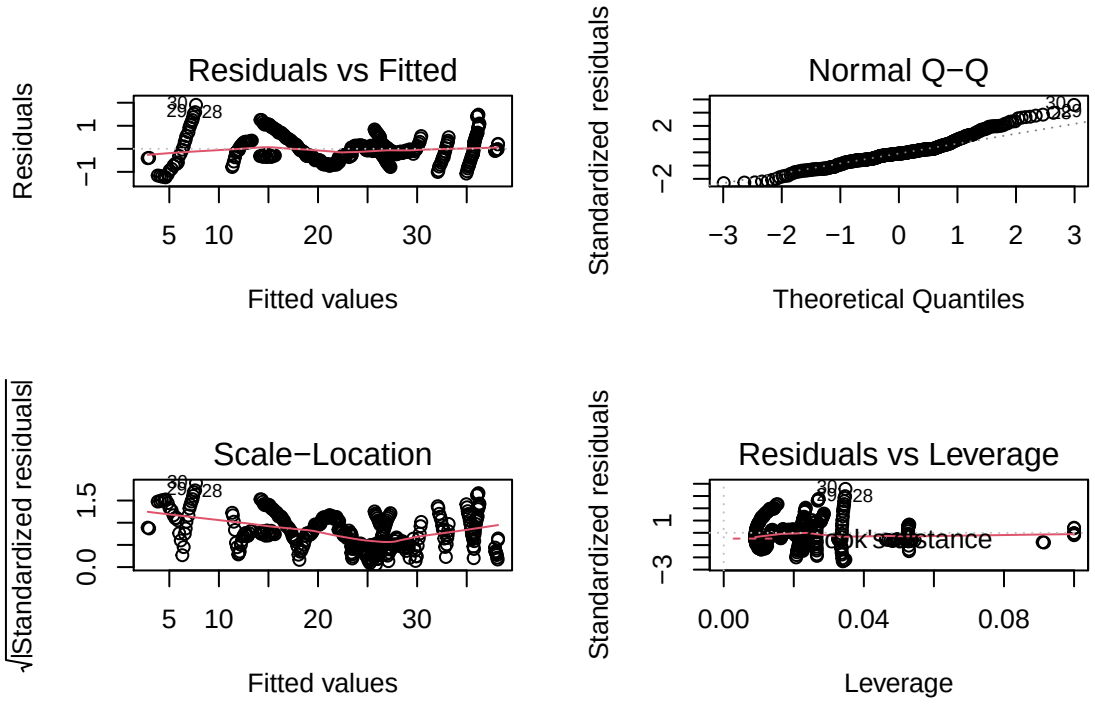


Plot 8: From the Cook's InfuleceIndexPlot and hat-values influenceIndexPlot we can see that we should definitely look at points 1 and 2.

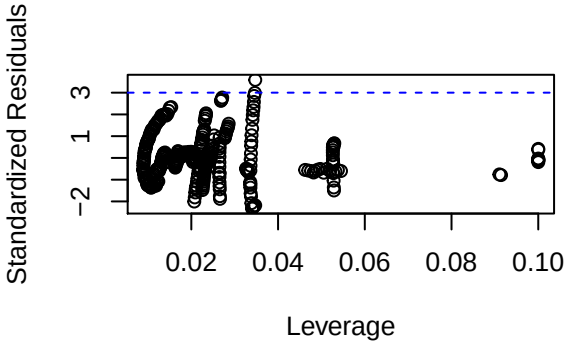


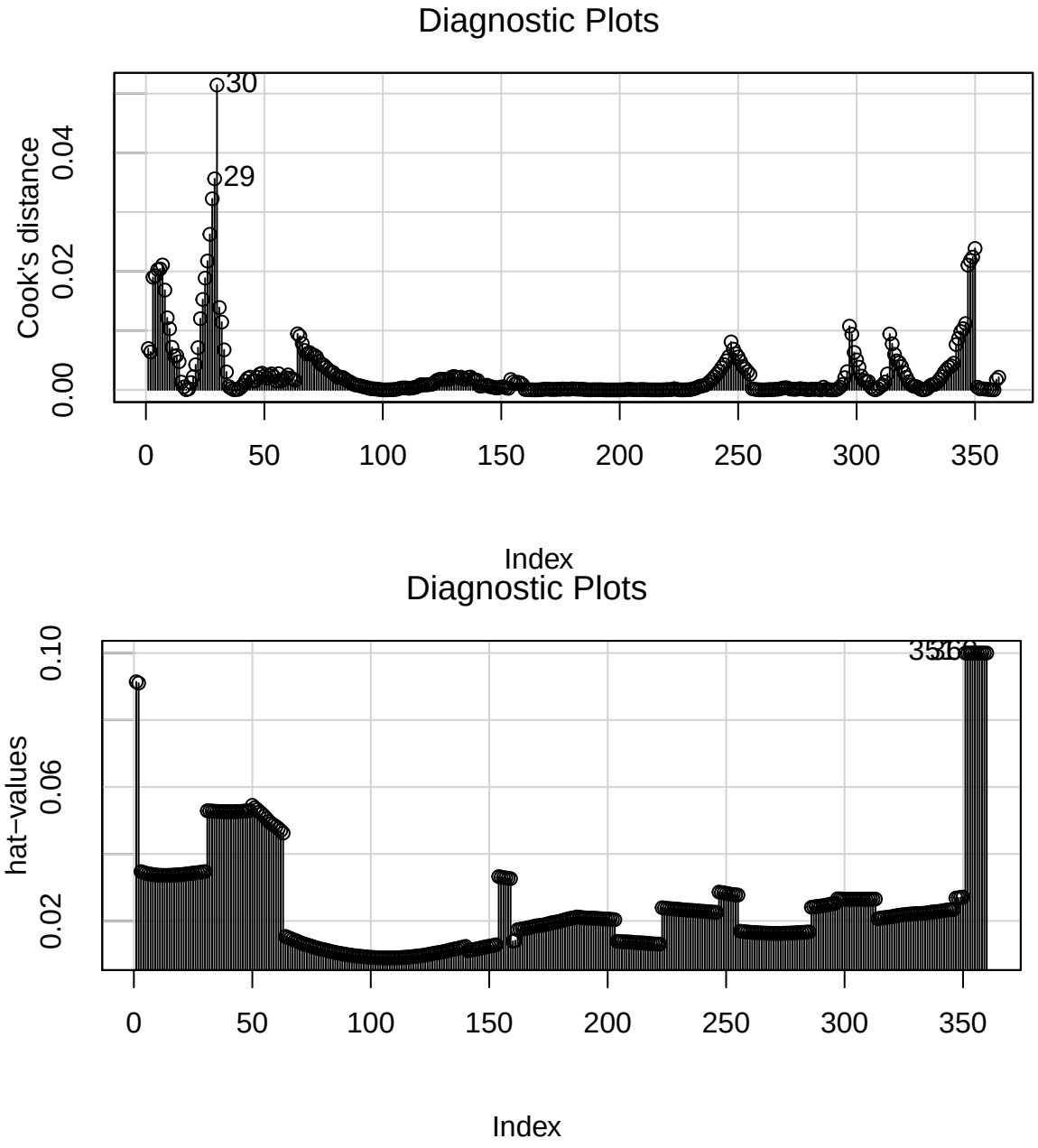


Plot 9: Checking the linearity after the removal of rows 1, 2, 33, 34, and 35. Looks a little better.

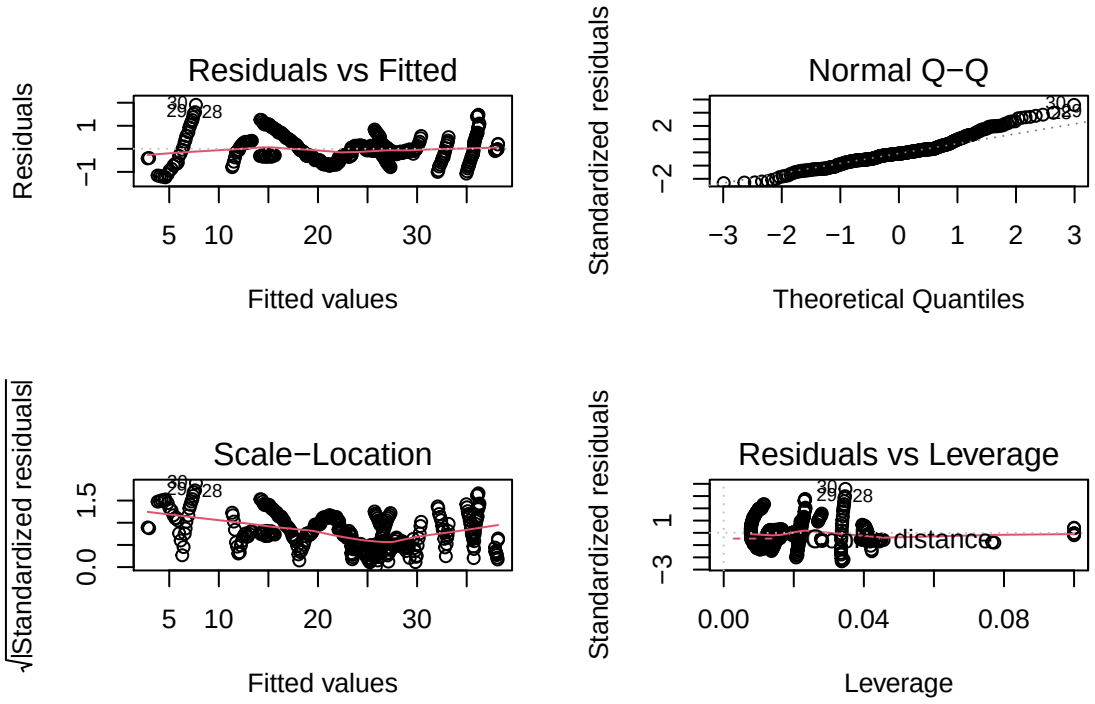


Plot 10: Checking Cook's plot again after the removal of rows 1, 2, 33, 34, and 35. Looks better.





Plot 11: After removing the outliers work_place_restrictions is no longer significant so we removed it and the plots look roughly the same so we remove it from our model.



Appendix 2a: R Code for Google Mobility + Categorical Variables

Code 1:

Table 1:

```
## Analysis of Variance Table
##
## Model 1: deaths ~ 1
## Model 2: deaths ~ retail_and_recreation_percent_change_from_baseline +
##     grocery_and_pharmacy_percent_change_from_baseline + parks_percent_change_from_baseline +
##     transit_stations_percent_change_from_baseline + workplaces_percent_change_from_baseline +
##     residential_percent_change_from_baseline + date + confirmed +
##     tests + fschool_closing + fworkplace_closing + fgatherings_restrictions +
##     fstay_home_restrictions + ftesting_policy + fcontact_tracing +
##     stringency_index
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  12972 3.0456e+12
## 2  12953 1.6726e+10 19 3.0288e+12 123455 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2:

```
## Subset selection object
## Call: regsubsets.formula(deaths ~ retail_and_recreation_percent_change_from_baseline +
##     grocery_and_pharmacy_percent_change_from_baseline + parks_percent_change_from_baseline +
##     transit_stations_percent_change_from_baseline + workplaces_percent_change_from_baseline +
##     residential_percent_change_from_baseline + date + confirmed +
```

```
##     tests + fschool_closing + fworkplace_closing + fgatherings_restrictions +
##     fstay_home_restrictions + ftesting_policy + fcontact_tracing +
##     stringency_index, data = cacovid_mobility)
## 19 Variables  (and intercept)
##                                                  Forced in Forced out
## retail_and_recreation_percent_change_from_baseline     FALSE      FALSE
## grocery_and_pharmacy_percent_change_from_baseline      FALSE      FALSE
## parks_percent_change_from_baseline                     FALSE      FALSE
## transit_stations_percent_change_from_baseline          FALSE      FALSE
## workplaces_percent_change_from_baseline                FALSE      FALSE
## residential_percent_change_from_baseline               FALSE      FALSE
## date                                                   FALSE      FALSE
## confirmed                                              FALSE      FALSE
## tests                                                  FALSE      FALSE
## fschool_closing3                                       FALSE      FALSE
## fworkplace_closing2                                    FALSE      FALSE
## fworkplace_closing3                                    FALSE      FALSE
## fgatherings_restrictions3                              FALSE      FALSE
## fgatherings_restrictions4                              FALSE      FALSE
## fstay_home_restrictions2                               FALSE      FALSE
## ftesting_policy2                                       FALSE      FALSE
## ftesting_policy3                                       FALSE      FALSE
## fcontact_tracing2                                      FALSE      FALSE
## stringency_index                                       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          retail_and_recreation_percent_change_from_baseline
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) " "
## 8  ( 1 ) " "
##          grocery_and_pharmacy_percent_change_from_baseline
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) " "
## 8  ( 1 ) " "
##          parks_percent_change_from_baseline
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) " "
## 8  ( 1 ) " "
##          transit_stations_percent_change_from_baseline
```

```
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) " "
## 8  ( 1 ) " "
##          workplaces_percent_change_from_baseline
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) " "
## 8  ( 1 ) " "
##          residential_percent_change_from_baseline date confirmed tests
## 1  ( 1 ) " "                                      " "  "*"       " "
## 2  ( 1 ) " "                                      " "  "*"       " "
## 3  ( 1 ) " "                                      " "  " "       "*"
## 4  ( 1 ) " "                                      " "  " "       "*"
## 5  ( 1 ) " "                                      " "  " "       "*"
## 6  ( 1 ) " "                                      " "  " "       "*"
## 7  ( 1 ) " "                                      " "  " "       "*"
## 8  ( 1 ) " "                                      "*"  " "       "*"
##          fschool_closing3 fworkplace_closing2 fworkplace_closing3
## 1  ( 1 ) " "              " "                 " "
## 2  ( 1 ) " "              " "                 " "
## 3  ( 1 ) " "              " "                 "*"
## 4  ( 1 ) "*"              " "                 " "
## 5  ( 1 ) "*"              " "                 " "
## 6  ( 1 ) "*"              " "                 "*"
## 7  ( 1 ) "*"              "*"                 "*"
## 8  ( 1 ) "*"              "*"                 "*"
##          fgatherings_restrictions3 fgatherings_restrictions4
## 1  ( 1 ) " "                       " "
## 2  ( 1 ) " "                       " "
## 3  ( 1 ) " "                       " "
## 4  ( 1 ) " "                       " "
## 5  ( 1 ) " "                       " "
## 6  ( 1 ) " "                       " "
## 7  ( 1 ) " "                       " "
## 8  ( 1 ) " "                       " "
##          fstay_home_restrictions2 ftesting_policy2 ftesting_policy3
## 1  ( 1 ) " "                      " "              " "
## 2  ( 1 ) "*"                      " "              " "
## 3  ( 1 ) "*"                      " "              " "
## 4  ( 1 ) "*"                      " "              " "
## 5  ( 1 ) "*"                      " "              " "
## 6  ( 1 ) "*"                      "*"              " "
## 7  ( 1 ) "*"                      "*"              " "
## 8  ( 1 ) "*"                      "*"              " "
##          fcontact_tracing2 stringency_index
```

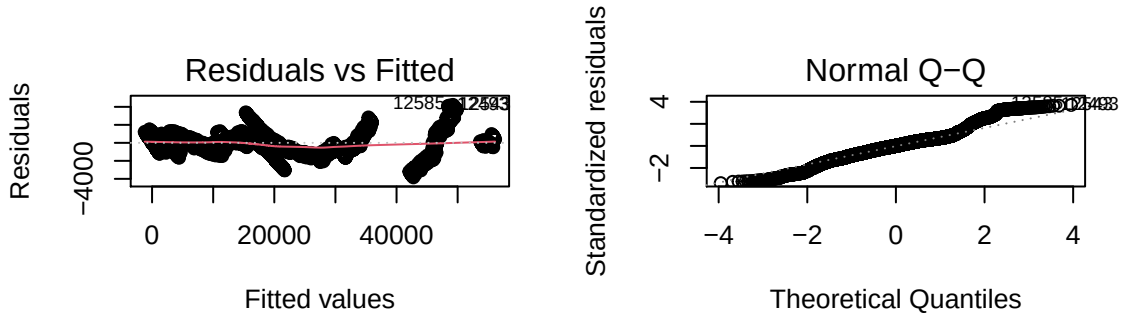
```
## 1  ( 1 ) " "               " "
## 2  ( 1 ) " "               " "
## 3  ( 1 ) " "               " "
## 4  ( 1 ) " "               "*"
## 5  ( 1 ) "*"               "*"
## 6  ( 1 ) "*"               " "
## 7  ( 1 ) "*"               " "
## 8  ( 1 ) "*"               " "
```

Table 3:

```
##   subset.summary.adjr2 subset.summary.cp subset.summary.bic
## 1            0.9605253        80128.9929          -41912.02
## 2            0.9780352        38831.2128          -49508.58
## 3            0.9881143        15062.2614          -57466.84
## 4            0.9922731         5256.1681          -63044.89
## 5            0.9930667         3385.6083          -64442.39
## 6            0.9933658         2681.1650          -65006.04
## 7            0.9939441         1318.7697          -66180.73
## 8            0.9941868          747.6231          -66702.84
```

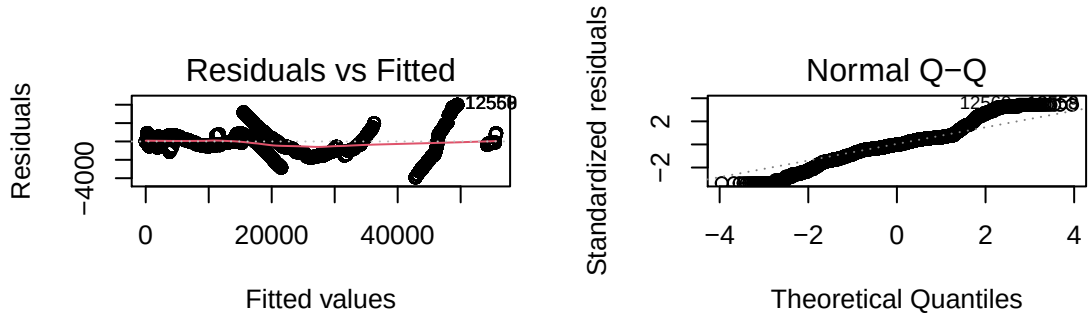
Table 4:

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.3642        0.36       0.3529       0.3754
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 5177.776  1 < 2.22e-16
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 8574.034  1 < 2.22e-16
```

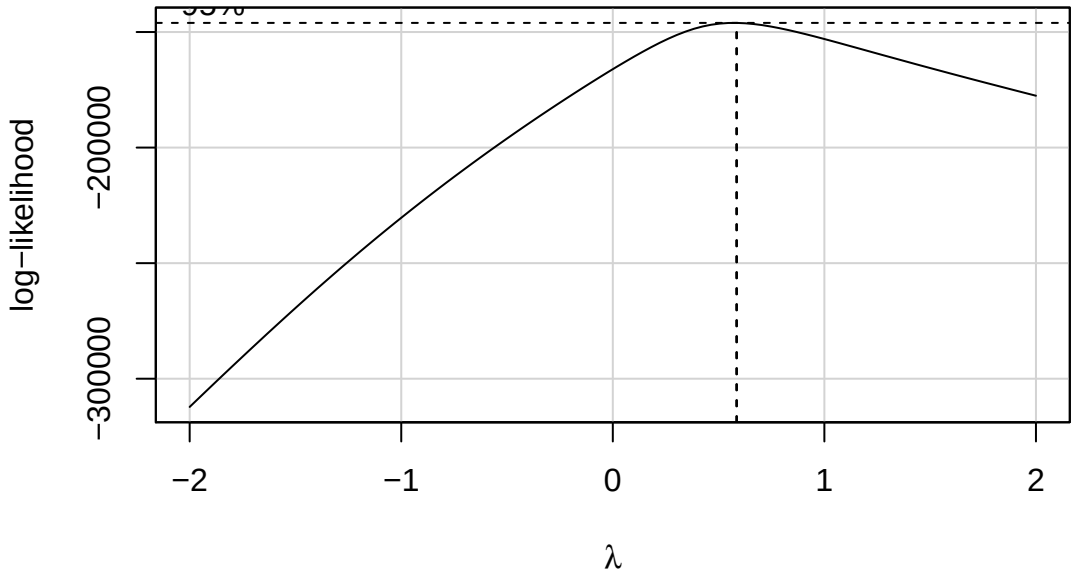


Plot 1:

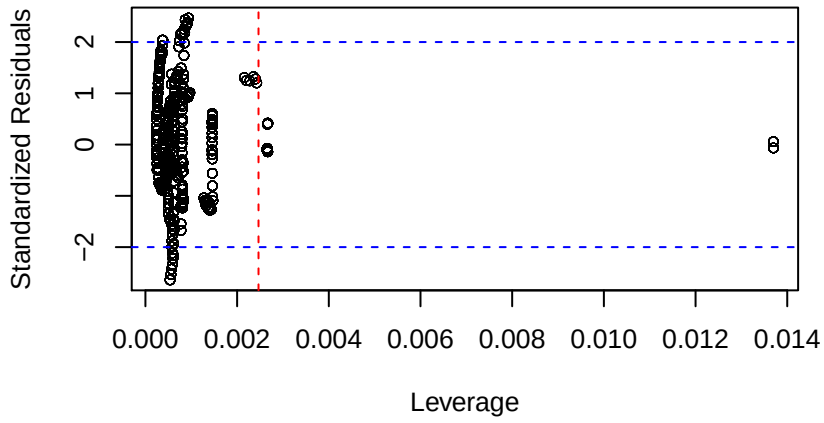
Plot 2:

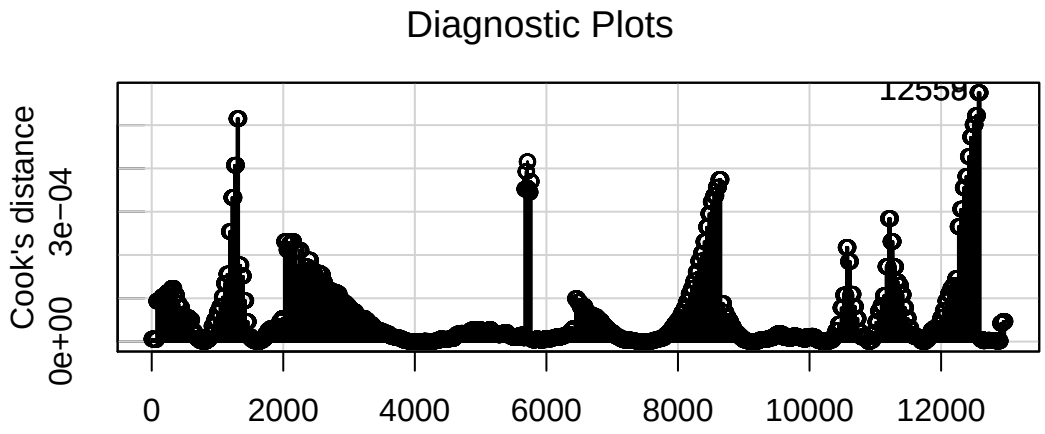


Plot 3:

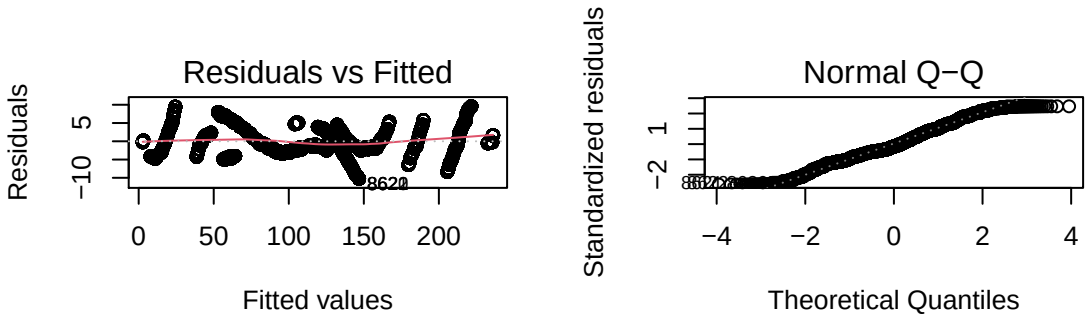


Plot 4:





Plot 5:



Plot 6:

Appendix 2b: R Code for Google Mobility Only

Table 1:

```
##
## Call:
## lm(formula = deaths ~ retail_and_recreation_percent_change_from_baseline +
##     grocery_and_pharmacy_percent_change_from_baseline + parks_percent_change_from_baseline +
##     transit_stations_percent_change_from_baseline + workplaces_percent_change_from_baseline +
##     residential_percent_change_from_baseline, data = cacovid_mobility)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28185  -8374  -3152   5292  44220
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                       73831.6     6714.1  10.997
## retail_and_recreation_percent_change_from_baseline  78601.3     1302.4  60.352
## grocery_and_pharmacy_percent_change_from_baseline  -70671.7     1483.6 -47.636
## parks_percent_change_from_baseline                  -6796.8      337.1 -20.163
## transit_stations_percent_change_from_baseline      -11553.1      576.5 -20.040
## workplaces_percent_change_from_baseline             -1849.6     1847.5  -1.001
## residential_percent_change_from_baseline           -32250.9     5220.5  -6.178
```

```
##                                                 Pr(>|t|)
## (Intercept)                                      < 2e-16 ***
## retail_and_recreation_percent_change_from_baseline  < 2e-16 ***
## grocery_and_pharmacy_percent_change_from_baseline   < 2e-16 ***
## parks_percent_change_from_baseline                  < 2e-16 ***
## transit_stations_percent_change_from_baseline       < 2e-16 ***
## workplaces_percent_change_from_baseline              0.317
## residential_percent_change_from_baseline           6.7e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12420 on 12966 degrees of freedom
## Multiple R-squared:  0.3436, Adjusted R-squared:  0.3433
## F-statistic:  1131 on 6 and 12966 DF,  p-value: < 2.2e-16
```

Table 2:

Table 3:

```
## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    1.5059        1.51       1.4613       1.5505
## Y2    0.6780        0.68       0.6239       0.7320
## Y3    0.3274        0.33       0.3044       0.3503
## Y4    0.2022        0.20       0.1793       0.2250
## Y5    1.2205        1.22       1.1899       1.2511
## Y6   -2.0512       -2.00      -2.2008      -1.9015
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                       LRT df       pval
## LR test, lambda = (0 0 0 0 0 0) 16049.92  6 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                       LRT df       pval
## LR test, lambda = (1 1 1 1 1 1) 9967.929  6 < 2.22e-16
```

```
## Subset selection object
## Call: regsubsets.formula(deaths ~ retail_and_recreation_percent_change_from_baseline +
##     grocery_and_pharmacy_percent_change_from_baseline + parks_percent_change_from_baseline +
##     transit_stations_percent_change_from_baseline + workplaces_percent_change_from_baseline +
##     residential_percent_change_from_baseline, data = cacovid_mobility)
## 6 Variables  (and intercept)
##                                                 Forced in Forced out
## retail_and_recreation_percent_change_from_baseline     FALSE      FALSE
## grocery_and_pharmacy_percent_change_from_baseline      FALSE      FALSE
## parks_percent_change_from_baseline                     FALSE      FALSE
## transit_stations_percent_change_from_baseline          FALSE      FALSE
## workplaces_percent_change_from_baseline                FALSE      FALSE
## residential_percent_change_from_baseline               FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
```

```
##          retail_and_recreation_percent_change_from_baseline
## 1  ( 1 ) "*"
## 2  ( 1 ) "*"
## 3  ( 1 ) "*"
## 4  ( 1 ) "*"
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
##          grocery_and_pharmacy_percent_change_from_baseline
## 1  ( 1 ) " "
## 2  ( 1 ) "*"
## 3  ( 1 ) "*"
## 4  ( 1 ) "*"
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
##          parks_percent_change_from_baseline
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) "*"
## 4  ( 1 ) "*"
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
##          transit_stations_percent_change_from_baseline
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) "*"
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
##          workplaces_percent_change_from_baseline
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) "*"
##          residential_percent_change_from_baseline
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
```

Table 4:

```
##
## Call:
## lm(formula = sqrt(deaths) ~ (I(retail_and_recreation_percent_change_from_baseline)^2) +
##     sqrt(grocery_and_pharmacy_percent_change_from_baseline) +
##     log(parks_percent_change_from_baseline) + log(transit_stations_percent_change_from_baseline) +
##     workplaces_percent_change_from_baseline + (I(1/residential_percent_change_from_baseline)^2),
##     data = cacovid_mobility)
##
```

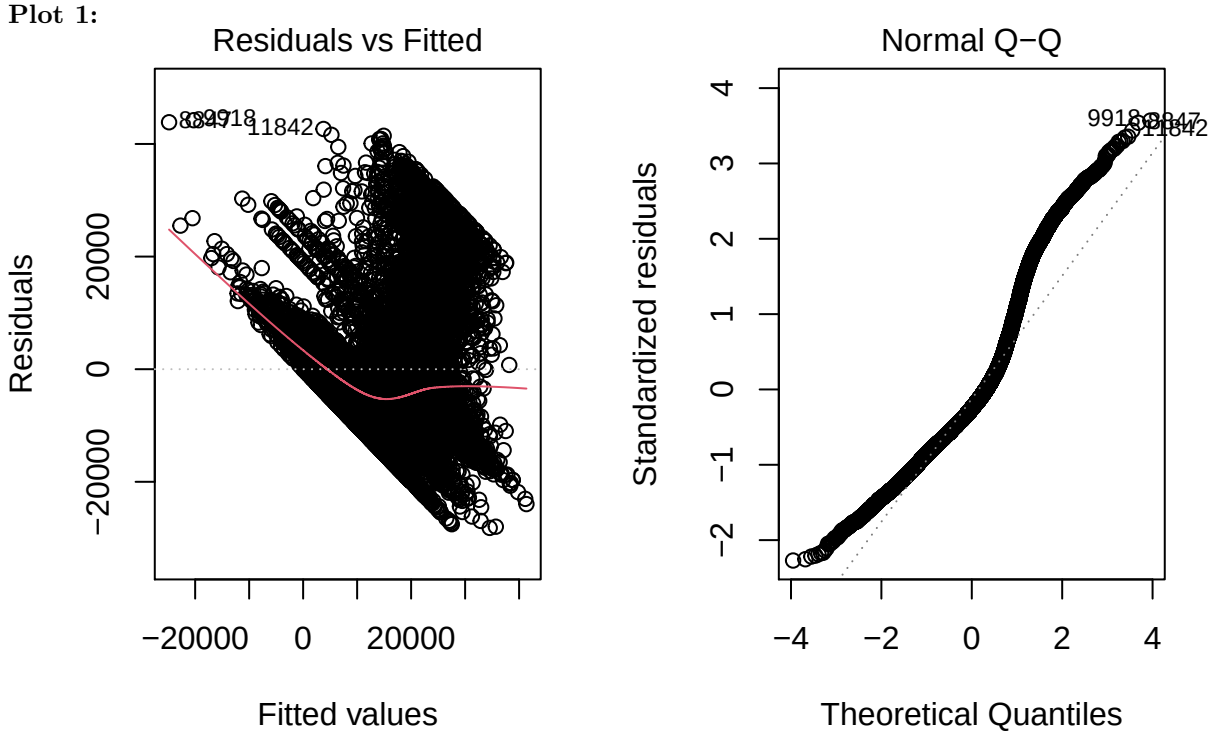
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -176.126  -27.167   -4.463   27.244  196.108
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                           0.285     19.775
## I(retail_and_recreation_percent_change_from_baseline)    347.544      4.857
## sqrt(grocery_and_pharmacy_percent_change_from_baseline) -536.451     10.121
## log(parks_percent_change_from_baseline)                -15.918      1.160
## log(transit_stations_percent_change_from_baseline)     -44.958      1.272
## workplaces_percent_change_from_baseline                -55.107      6.961
## I(1/residential_percent_change_from_baseline)          392.199     21.562
##                                                    t value Pr(>|t|)
## (Intercept)                                          0.014    0.988
## I(retail_and_recreation_percent_change_from_baseline)    71.555  < 2e-16 ***
## sqrt(grocery_and_pharmacy_percent_change_from_baseline) -53.005  < 2e-16 ***
## log(parks_percent_change_from_baseline)                -13.721  < 2e-16 ***
## log(transit_stations_percent_change_from_baseline)     -35.358  < 2e-16 ***
## workplaces_percent_change_from_baseline                -7.917 2.63e-15 ***
## I(1/residential_percent_change_from_baseline)          18.189  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.15 on 12966 degrees of freedom
## Multiple R-squared:  0.4313, Adjusted R-squared:  0.4311
## F-statistic:  1639 on 6 and 12966 DF,  p-value: < 2.2e-16
```

Table 5:

```
##
## Call:
## lm(formula = sqrt(deaths) ~ (I(retail_and_recreation_percent_change_from_baseline)^2) +
##     sqrt(grocery_and_pharmacy_percent_change_from_baseline) +
##     log(parks_percent_change_from_baseline) + log(transit_stations_percent_change_from_baseline) +
##     workplaces_percent_change_from_baseline + (I(1/residential_percent_change_from_baseline)^2),
##     data = cacovid_mobilityGO)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -176.073  -27.154   -4.421   27.246  168.148
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                           4.020     19.781
## I(retail_and_recreation_percent_change_from_baseline)    348.571      4.860
## sqrt(grocery_and_pharmacy_percent_change_from_baseline) -537.789     10.119
## log(parks_percent_change_from_baseline)                -15.904      1.159
## log(transit_stations_percent_change_from_baseline)     -44.946      1.271
## workplaces_percent_change_from_baseline                -53.700      6.964
## I(1/residential_percent_change_from_baseline)          388.194     21.568
##                                                    t value Pr(>|t|)
## (Intercept)                                          0.203    0.839
```

```
## I(retail_and_recreation_percent_change_from_baseline)    71.726  < 2e-16 ***
## sqrt(grocery_and_pharmacy_percent_change_from_baseline) -53.147  < 2e-16 ***
## log(parks_percent_change_from_baseline)                 -13.718  < 2e-16 ***
## log(transit_stations_percent_change_from_baseline)      -35.372  < 2e-16 ***
## workplaces_percent_change_from_baseline                  -7.711 1.34e-14 ***
## I(1/residential_percent_change_from_baseline)            17.998  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.12 on 12965 degrees of freedom
## Multiple R-squared:  0.4321, Adjusted R-squared:  0.4318
## F-statistic:  1644 on 6 and 12965 DF,  p-value: < 2.2e-16
```

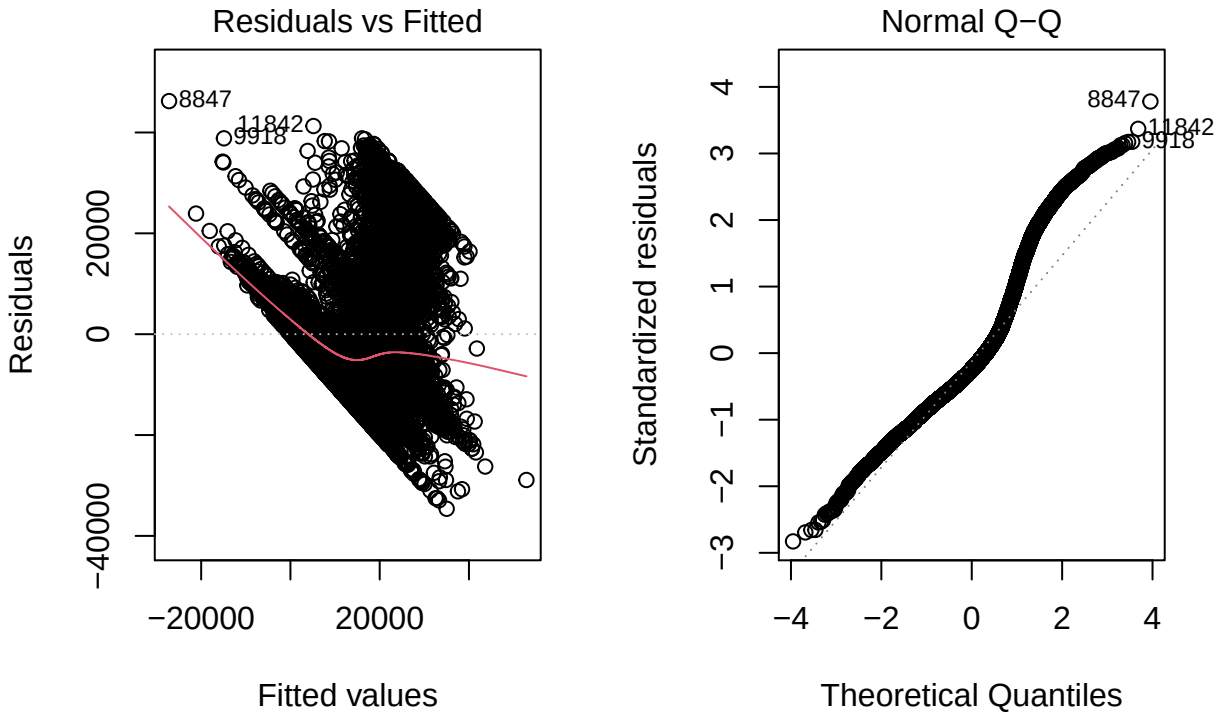
Plot 1:



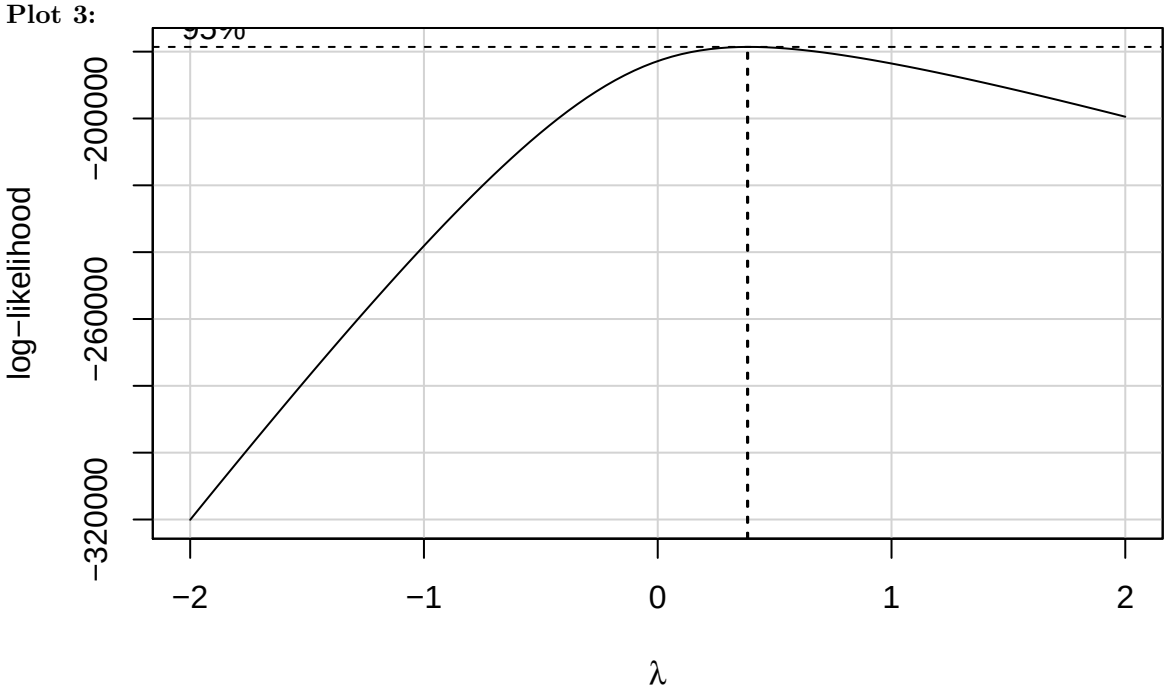
Plot 2:

```
##
## Call:
## lm(formula = deaths ~ (I(retail_and_recreation_percent_change_from_baseline)^2) +
##     sqrt(grocery_and_pharmacy_percent_change_from_baseline) +
##     log(parks_percent_change_from_baseline) + log(transit_stations_percent_change_from_baseline) +
##     workplaces_percent_change_from_baseline + (I(1/residential_percent_change_from_baseline)^2),
##     data = cacovid_mobility)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34621  -8102  -3024   5052  46223
##
## Coefficients:
```

```
##                                                    Estimate Std. Error
## (Intercept)                                         22644.7     5249.5
## I(retail_and_recreation_percent_change_from_baseline)     82092.9     1289.4
## sqrt(grocery_and_pharmacy_percent_change_from_baseline) -122943.4     2686.7
## log(parks_percent_change_from_baseline)              -5349.0      308.0
## log(transit_stations_percent_change_from_baseline)       -10124.3      337.5
## workplaces_percent_change_from_baseline                -221.4     1847.8
## I(1/residential_percent_change_from_baseline)         48677.4     5723.9
##                                                    t value Pr(>|t|)
## (Intercept)                                          4.314 1.62e-05 ***
## I(retail_and_recreation_percent_change_from_baseline)    63.670  < 2e-16 ***
## sqrt(grocery_and_pharmacy_percent_change_from_baseline) -45.760  < 2e-16 ***
## log(parks_percent_change_from_baseline)                 -17.368  < 2e-16 ***
## log(transit_stations_percent_change_from_baseline)      -29.995  < 2e-16 ***
## workplaces_percent_change_from_baseline                  -0.120    0.905
## I(1/residential_percent_change_from_baseline)             8.504  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12250 on 12966 degrees of freedom
## Multiple R-squared:  0.3609, Adjusted R-squared:  0.3606
## F-statistic:  1221 on 6 and 12966 DF,  p-value: < 2.2e-16
```



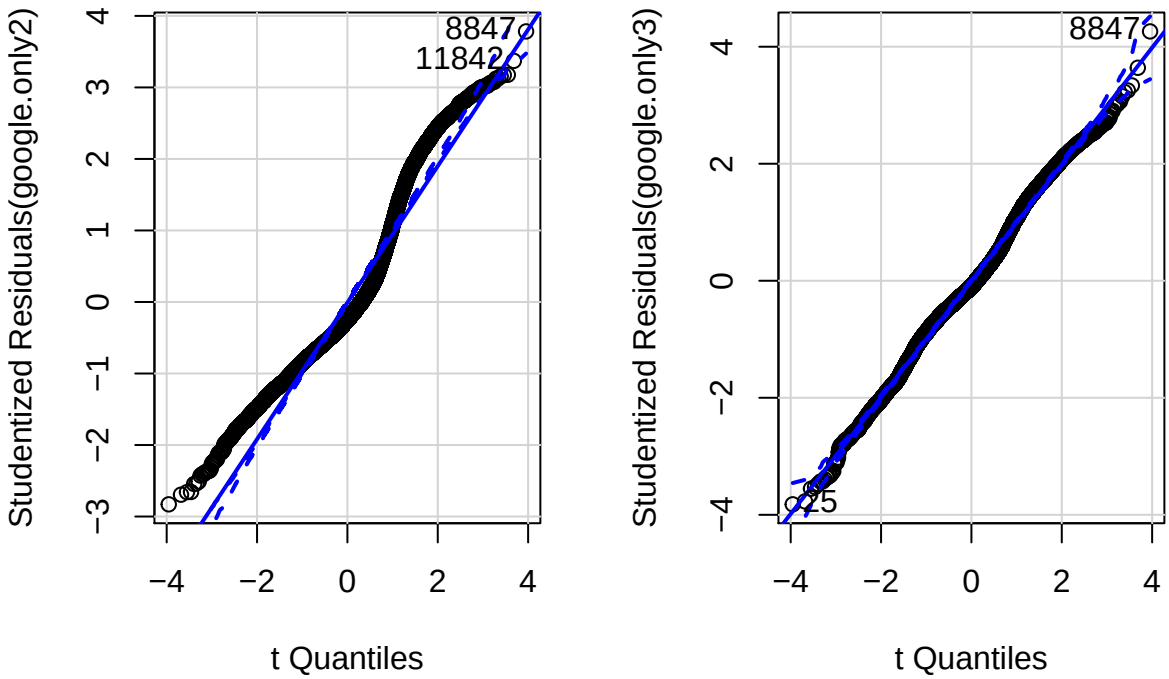
Plot 3:




```
## [1] 0.3838384
```

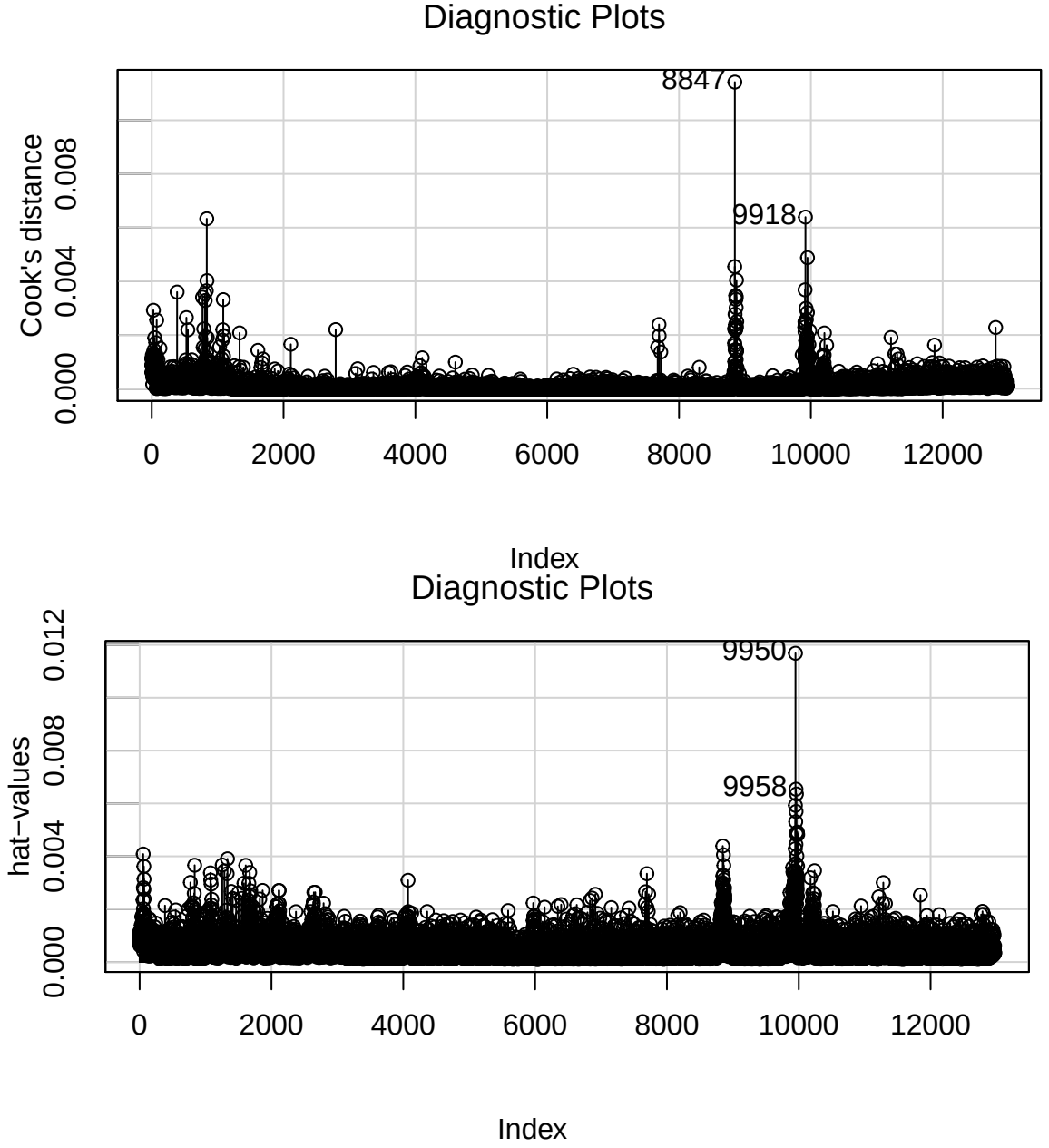
Plot 4:

```
## [1]  8847 11842
```




```
## [1]   25 8847
```

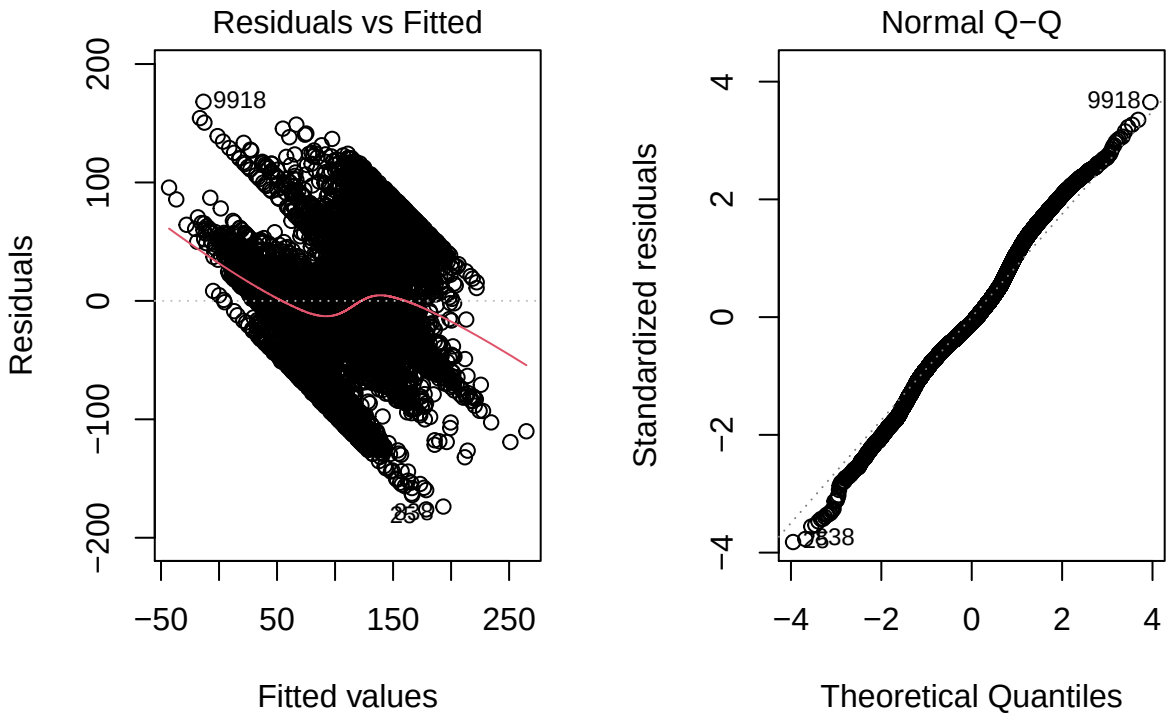
Plot 5:



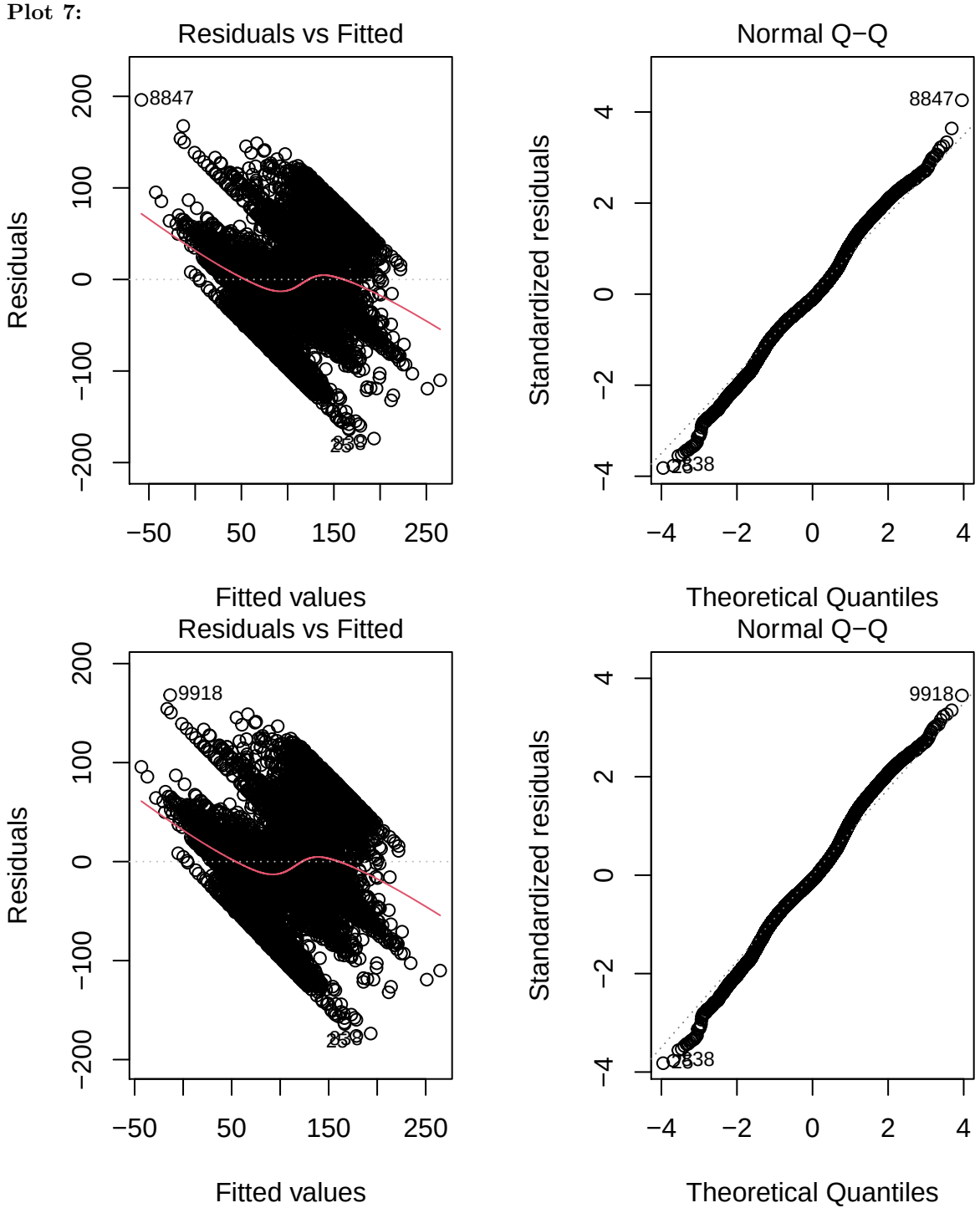
Plot 6:

```
##
## Call:
## lm(formula = sqrt(deaths) ~ (I(retail_and_recreation_percent_change_from_baseline)^2) +
##     sqrt(grocery_and_pharmacy_percent_change_from_baseline) +
##     log(parks_percent_change_from_baseline) + log(transit_stations_percent_change_from_baseline) +
##     workplaces_percent_change_from_baseline + (I(1/residential_percent_change_from_baseline)^2),
##     data = cacovid_mobilityGO)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -176.073  -27.154   -4.421   27.246  168.148
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                           4.020     19.781
## I(retail_and_recreation_percent_change_from_baseline)    348.571      4.860
## sqrt(grocery_and_pharmacy_percent_change_from_baseline) -537.789     10.119
## log(parks_percent_change_from_baseline)               -15.904      1.159
## log(transit_stations_percent_change_from_baseline)    -44.946      1.271
## workplaces_percent_change_from_baseline               -53.700      6.964
## I(1/residential_percent_change_from_baseline)         388.194     21.568
##                                                    t value Pr(>|t|)
## (Intercept)                                          0.203    0.839
## I(retail_and_recreation_percent_change_from_baseline)    71.726  < 2e-16 ***
## sqrt(grocery_and_pharmacy_percent_change_from_baseline) -53.147  < 2e-16 ***
## log(parks_percent_change_from_baseline)              -13.718  < 2e-16 ***
## log(transit_stations_percent_change_from_baseline)   -35.372  < 2e-16 ***
## workplaces_percent_change_from_baseline              -7.711 1.34e-14 ***
## I(1/residential_percent_change_from_baseline)        17.998  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.12 on 12965 degrees of freedom
## Multiple R-squared:  0.4321, Adjusted R-squared:  0.4318
## F-statistic:  1644 on 6 and 12965 DF,  p-value: < 2.2e-16
```



Plot 7:



Appendix 3: R Code for World Bank

Base Data with World Bank imported, GDP, GDP growth, Hospital beds/1,000 ppl, Poverty= Poverty head count ratio at 1.90 a day(% of pop), CO2em: CO2 emissions (metric tons/capita), Air pollution: % Pop exposed to levels exceeding WHO guidelines,

```
x <- covid19()
```

```
## We have invested a lot of time and effort in creating COVID-19 Data Hub, please cite the following wh
##
##   Guidotti, E., Ardia, D., (2020), "COVID-19 Data Hub", Journal of Open
##   Source Software 5(51):2376, doi: 10.21105/joss.02376.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {COVID-19 Data Hub},
##     year = {2020},
##     doi = {10.21105/joss.02376},
##     author = {Emanuele Guidotti and David Ardia},
##     journal = {Journal of Open Source Software},
##     volume = {5},
##     number = {51},
##     pages = {2376},
##   }
##
## To retrieve citation and metadata of the data sources see ?covid19cite. To hide this message use 'ver
```

```
covid19<- covid19(level = 1, start = "2020-03-15", end = "2021-03-15",)
wb <- c("gdp" = "NY.GDP.MKTP.CD", "hosp_beds" = "SH.MED.BEDS.ZS","gdp_grow" = "NY.GDP.MKTP.KD.ZG","pover
wbdcovid  <- covid19(wb = wb)
```

```
## We have invested a lot of time and effort in creating COVID-19 Data Hub, please cite the following wh
##
##   Guidotti, E., Ardia, D., (2020), "COVID-19 Data Hub", Journal of Open
##   Source Software 5(51):2376, doi: 10.21105/joss.02376.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {COVID-19 Data Hub},
##     year = {2020},
##     doi = {10.21105/joss.02376},
##     author = {Emanuele Guidotti and David Ardia},
##     journal = {Journal of Open Source Software},
##     volume = {5},
##     number = {51},
##     pages = {2376},
##   }
##
## To retrieve citation and metadata of the data sources see ?covid19cite. To hide this message use 'ver
```

```
wbdcovid
```

```
## # A tibble: 94,748 x 42
## # Groups:   id [199]
##    iso_alpha_3 id    date       vaccines tests confirmed recovered deaths  hosp
##    <chr>       <chr> <date>        <dbl> <int>     <int>     <int>  <int> <dbl>
```

```
##  1 AFG         AFG   2020-01-22       NA    NA        NA        NA     NA    NA
##  2 AFG         AFG   2020-01-23       NA    NA        NA        NA     NA    NA
##  3 AFG         AFG   2020-01-24       NA    NA        NA        NA     NA    NA
##  4 AFG         AFG   2020-01-25       NA    NA        NA        NA     NA    NA
##  5 AFG         AFG   2020-01-26       NA    NA        NA        NA     NA    NA
##  6 AFG         AFG   2020-01-27       NA    NA        NA        NA     NA    NA
##  7 AFG         AFG   2020-01-28       NA    NA        NA        NA     NA    NA
##  8 AFG         AFG   2020-01-29       NA    NA        NA        NA     NA    NA
##  9 AFG         AFG   2020-01-30       NA    NA        NA        NA     NA    NA
## 10 AFG         AFG   2020-01-31       NA    NA        NA        NA     NA    NA
## # ... with 94,738 more rows, and 33 more variables: vent <int>, icu <int>,
## #   population <int>, school_closing <int>, workplace_closing <int>,
## #   cancel_events <int>, gatherings_restrictions <int>,
## #   transport_closing <int>, stay_home_restrictions <int>,
## #   internal_movement_restrictions <int>,
## #   international_movement_restrictions <int>, information_campaigns <int>,
## #   testing_policy <int>, contact_tracing <int>, stringency_index <dbl>,
## #   iso_alpha_2 <chr>, iso_numeric <int>, currency <chr>,
## #   administrative_area_level <int>, administrative_area_level_1 <chr>,
## #   administrative_area_level_2 <lgl>, administrative_area_level_3 <lgl>,
## #   latitude <dbl>, longitude <dbl>, key <lgl>, key_apple_mobility <chr>,
## #   key_google_mobility <chr>, gdp <dbl>, hosp_beds <dbl>, gdp_grow <dbl>,
## #   poverty <dbl>, co2em <dbl>, pollution <dbl>
```

Only variables from WB and Confirmed and Death

```
wbcovdata<- subset(wbdcovid, select = c("date", "confirmed", "deaths", "iso_alpha_3", "administrative_a
```

```
fgpd<- as.integer(wbcovdata$gdp)
```

```
## Warning: NAs introduced by coercion to integer range
```

```
fgdp_grow<- as.integer(wbcovdata$gdp_grow)
fhosp_beds<- as.integer(wbcovdata$hosp_beds)
fpoverty<- as.integer(wbcovdata$poverty)
fco2em<- as.integer(wbcovdata$co2em)
fpollution<- as.integer(wbcovdata$pollution)

#wbcovdata$gdp %>% replace_na(0)
#wbcovdata$gdp_grow %>% replace_na(0)
#wbcovdata$hosp_beds %>% replace_na(0)
#wbcovdata$poverty %>% replace_na(0)
#wbcovdata$co2em %>% replace_na(0)
#wbcovdata$pollution %>% replace_na(0)


cleandata <- na.omit(wbcovdata)
cleandata
```

```
## # A tibble: 51,375 x 11
##    date       confirmed deaths iso_alpha_3 administrative_are~      gdp gdp_grow
##    <date>         <int>  <int> <chr>       <chr>                  <dbl>    <dbl>
```

```
##  1 2020-03-11        12      1 ALB         Albania              1.53e10     2.24
##  2 2020-03-12        23      1 ALB         Albania              1.53e10     2.24
##  3 2020-03-13        33      1 ALB         Albania              1.53e10     2.24
##  4 2020-03-14        38      1 ALB         Albania              1.53e10     2.24
##  5 2020-03-15        42      1 ALB         Albania              1.53e10     2.24
##  6 2020-03-16        51      1 ALB         Albania              1.53e10     2.24
##  7 2020-03-17        55      1 ALB         Albania              1.53e10     2.24
##  8 2020-03-18        59      2 ALB         Albania              1.53e10     2.24
##  9 2020-03-19        64      2 ALB         Albania              1.53e10     2.24
## 10 2020-03-20        70      2 ALB         Albania              1.53e10     2.24
## # ... with 51,365 more rows, and 4 more variables: hosp_beds <dbl>,
## #   poverty <dbl>, co2em <dbl>, pollution <dbl>
```

Graphs to check for Normality and variance

```
economic<- lm(deaths ~ confirmed, gdp, gdp_grow, poverty, data= cleandata) # Economic
airqual<- lm(deaths ~ co2em, pollution, data= wbcovdata) #Air Quality

summary(economic)
summary(airqual)
```

Appendix 4: Exploratory analysis not used in final paper

Github Link: https://github.com/oboechick/STAT632FinalProject

Appendix 5: Data Variable Description

- **date** - Observation date
- **confirmed** - Cumulative number of confirmed cases
- **tests** - Cumulative number of tests
- **population** - Total population
- **latitude** - Latitude (Check to see if more than 1 since we are only using CA)
- **longitude** - Longitude (Check to see if more than 1 since we are only using CA)
- **school_closing** - 0: No measures - 1: Recommend closing - 2: Require closing (only some levels or categories, eg just high school, or just public schools - 3: Require closing all levels
- **workplace_closing** - 0: No measures - 1: Recommend closing (or work from home) - 2: require closing for some sectors or categories of workers - 3: require closing (or work from home) all-but-essential workplaces (eg grocery stores, doctors).
- **cancel_events** - 0: No measures - 1: Recommend canceling - 2: Require canceling gatherings_restrictions 0: No restrictions - 1: Restrictions on very large gatherings (the limit is above 1000 people) - 2: Restrictions on gatherings between 100-1000 people - 3: Restrictions on gatherings between 10-100 people - 4: Restrictions on gatherings of less than 10 people.
- **gatherings_restrictions** - 0: No restrictions - 1: Restrictions on very large gatherings (the limit is above 1000 people) - 2: Restrictions on gatherings between 100-1000 people - 3: Restrictions on gatherings between 10-100 people - 4: Restrictions on gatherings of less than 10 people.
- **transport_closing** - 0: No measures - 1: Recommend closing (or significantly reduce volume/route/means of transport available) - 2: Require closing (or prohibit most citizens from using it).
- **stay_home_restrictions** - 0: No measures - 1: recommend not leaving house - 2: require not leaving house with exceptions for daily exercise, grocery shopping, and "essential" trips - 3: Require not leaving house with minimal exceptions (e.g. allowed to leave only once every few days, or only one person can leave at a time, etc.).

- **internal_movement_restrictions** - 0: No measures - 1: Recommend closing (or significantly reduce volume/route/means of transport) - 2: Require closing (or prohibit most people from using it).
- **international_movement_restrictions** - 0: No measures - 1: Screening - 2: Quarantine arrivals from high-risk regions - 3: Ban on high-risk regions - 4: Total border closure.
- **information_campaigns** - 0: No COVID-19 public information campaign - 1: public officials urging caution about COVID-19 - 2: coordinated public information campaign (e.g. across traditional and social media).
- **testing_policy** - 0: No testing policy - 1: Only those who both (a) have symptoms AND (b) meet specific criteria (eg key workers, admitted to hospital, came into contact with a known case, returned from overseas) - 2: testing of anyone showing COVID-19 symptoms - 3: open public testing (eg "drive through" testing available to asymptomatic people).
- **contact_tracing** - 0: No contact tracing - 1: Limited contact tracing, not done for all cases - 2: Comprehensive contact tracing, done for all cases.
- **stringency_index** - Stringency of governmental responses.
- **retail_and_recreation_percent_change_from_baseline** - comparison of pre-Covid-19 pandemic to Covid-19 pandemic travel trends to destinations classified as retail and recreation
- **grocery_and_pharmacy_percent_change_from_baseline** - comparison of pre-Covid-19 pandemic to Covid-19 pandemic travel trends to destinations classified as grocery stores and pharmacies
- **parks_percent_change_from_baseline** - comparison of pre-Covid-19 pandemic to Covid-19 pandemic travel trends to destinations classified as outdoor parks
- **transit_stations_percent_change_from_baseline** - comparison of pre-Covid-19 pandemic to Covid-19 pandemic travel trends to destinations classified as transit stations
- **workplaces_percent_change_from_baseline** - comparison of pre-Covid-19 pandemic to Covid-19 pandemic travel trends to destinations classified as work places
- **residential_percent_change_from_baseline** - comparison of pre-Covid-19 pandemic to Covid-19 pandemic travel trends to destinations classified as residential

wb - World Bank Data

Source

<URL: https://covid19datahub.io>

References

Guidotti, E., Ardia, D., (2020), "COVID-19 Data Hub", Journal of
Open Source Software 5(51):2376, doi: 10.21105/joss.02376 (URL:
https://doi.org/10.21105/joss.02376).