Final Project 632 Rough Draft

Nic James, Sri Chandu, Thomas Li

05/21/2020

Contents

Abstract (100 words) - Nic

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.

Problem and Motivation (200 words) - Sri

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.

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 in the original COVID-19 Data Hub data set:
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

Variables used on top of base data set:
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 model using the contact tracing is the best predictor of deaths? We plan to use *deaths* as the response and *confirmed*, *tests*, *contact tracing*, and *stringency index* from in the original COVID-19 data set as predictors to answer this question.

2. How does the economic profile of the country affect the mortality rate from COVID over the year 2020? 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 per capita*, *GDP per capita growth*, and *Poverty rate* from the World Bank data set to answer this question.

3. 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 per capita*, *GDP per capita growth*, and *Pollution in mcg* from the World Bank data set to answer this question.

4. Using the Google Mobility Data, are policy measures that are non-restrictive with movement significant in preventing spread of 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 questions.

5. Using the Google Mobility Data, are policy measures that are restrictive with movement more significant than non-restrictive measures in preventing the spread of COVID-19 - Response: deaths - Predictors: looking at both movement restrictive and non-restrictive and comparing their significance

Regression Analysis, Results and Interpretation

Important Details

We did a hypothesis test to determine whether or not we had any significant variables in our full model to start. We used $H_0 : \beta_1 = \beta_2 = ... = \beta_9 = 0$ and $H_1 :$ 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
```

```
# full fitted model w/ removed predictor variables
google.full <- lm(deaths ~ retail_and_recreation_percent_change_from_baseline + grocery_and_pharmacy_pe

google.null <- lm(deaths ~ 1, data = cacovid_mobility)
```

```
anova(google.null, google.full)
```

```
## 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
```

Exploratory Analysis I:

We started by creating a data frame by filtering the data first by United States of America, secondly by California, and finally 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, *tests*, *confirmed*, *recovered*, *deaths*, *hosp*, *vent*, *icu*, *latitude*, *longitude*, *population*, *vaccines*, *school_closing*, *workplace_closing*, *cancel_events*, *gatherings_restrictions*, *transport_closing*, *stay_home_restrictions*, *internal_movement_restrictions*, *international_movement_restrictions*, *information_campaigns*, *testing_policy*, *contact_tracing*, and *stringency_index*. Since we don't have data for *vaccines*, *recovered*, *hosp*, *vent*, and *icu* we removed these variables. All variables now have data in every row. 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 *internal_movement_restrictions*, *information_campaigns*, *population*, *longitude*, and *latitude* as they have the same data for every row causing 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-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. From this we can see that confirmed and tests both have a positive linear relationship with deaths. This would lead us to assume that confirmed and tests would be a positive influence on the number of deaths. 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 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 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. We may 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 and we cannot remove it without removing a significant variable 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 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.

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 look we see that 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). However we can still see that there is probably an outlier from the Cook's Distance plot (Plot 8, Appendix 1) so next we looked for outliers and leverage points (Code 4, Appendix 1). We removed the outliers and leverage points one at a time and stopped after removing rows 1, 2, 33, 34, and 35 (Code 5, Appendix 1). When we checked the diagnostics the Q-Q plot (Plot 9, Appendix 1) improved a little and the Cook's Distance plot again (Plot 10, Appendix 1) we were much happier with this result.

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 did not appear to be a difference between them. 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.

Diagnostic Checks I:

Interpretation I:

Exploratory Analysis II:

Full Model (Base Covid + Google Mobility) After working solely with the base COVID-19 data, we decided to add in the Google mobility data. First, we read in and subset the Google mobility data. The data only included reports in CA and ranged from Mar 13, 2020 to Mar 14, 2021. We also took out 4 columns of data that were identifiers and not relevant for our data analysis. Lastly, we removed all rows with at least one NA and converted all data from percentages to decimals. After changing the Google mobility data, we merged the modified base COVID-19 data and Google mobility data into one data frame. The Google mobility variables added as predictors are *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*. The modified base COVID-19 data set included all variables with values in every row.

We started off by running a linear model summary of the the full model (Table A, Appendix 1). We saw that the there were variables that were singularities (i.e. *longitude*, *latitude*, *population*, *finternal_movement_restrictions2*, and *finformation_campaigns2*) so we removed these from the model as well. Next, we ran an ANOVA comparing the modified full model with the null model (Table B, Appendix 1). The resulting p-value was <2.2e-16 so we reject the null model and conclude that there is at least one predictor variable in the full model that is significant. The next step we took was to check the QQ plot and residuals vs. fitted plot (Plot A, Appendix 1). Visually, we saw that it did not meet the assumptions of linearity. The QQ plot did not follow a linear trend and the residuals vs. fitted plot showed obvious patterning.

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 (Table C, Appendix 1). We compared all eight models by looking at adjusted R-squared, CP values, and BIC values (Table D, Appendix 1). The model we chose had seven predictor variables, but still did not show linearity (Plot B, Appendix 1). Thus, we decided to check if there were any necessary transformations for the predictors.

We ran transformations of all non-factor predictors. This resulted in a square root transformation for *confirmed* (Table 5, Appendix A). We also ran transformations for the response variable which resulted in a square root transformation for *deaths* (Plot C, Appendix 1).

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.

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.

Interpretation II:

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

Exploratory Analysis III:

Due to the base data not being linear we decided to add but…

Interpretation

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

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

```
(step_aic <- step(mod.0, scope = list(lower = mod.0, upper = mod.full), trace = 0))
```

```
##
## 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.

```
(step_bic <- step(mod.0, scope = list(lower = mod.0, upper = mod.full), trace = 0))
```

```
##
## 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.

```
pt <- powerTransform(cbind(tests, stringency_index) ~ 1, data = fbase_data)
summary(pt)
```

```
## 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.

```
mod.full4 <- lm((deaths^{1/3}) ~ fstay_home_restrictions + I(tests^{1/3}) + fworkplace_closing +
    fschool_closing + fcontact_tracing + ftesting_policy + log(stringency_index),
    data = fbase_data)
# leverage point calculations
p <- 10
n <- nrow(fbase_data)
mod.full4_hat <- hatvalues(mod.full4)

leverage <- which(mod.full4_hat > 4*(p+1)/n)

# Find outliers
mod.full4_out <- rstandard(mod.full4)
outliers<- which(abs(mod.full4_out) > 3)

# Cook's points and hat values
mod.full4.cooks <- cooks.distance(mod.full4)
cooks <- which(mod.full4.cooks > 4/(n-p-1))

leverage
```

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

```
outliers
```

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

```
cooks
```

```
##   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.

```
fbase_data1<-fbase_data[-c(1,2,33,34,35), ]
```

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.

```
summary(mod.full)
```

```
##
## 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.

```
summary(step_bic)
```

```
##
## 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
```

```
summary(mod.full5)
```

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.

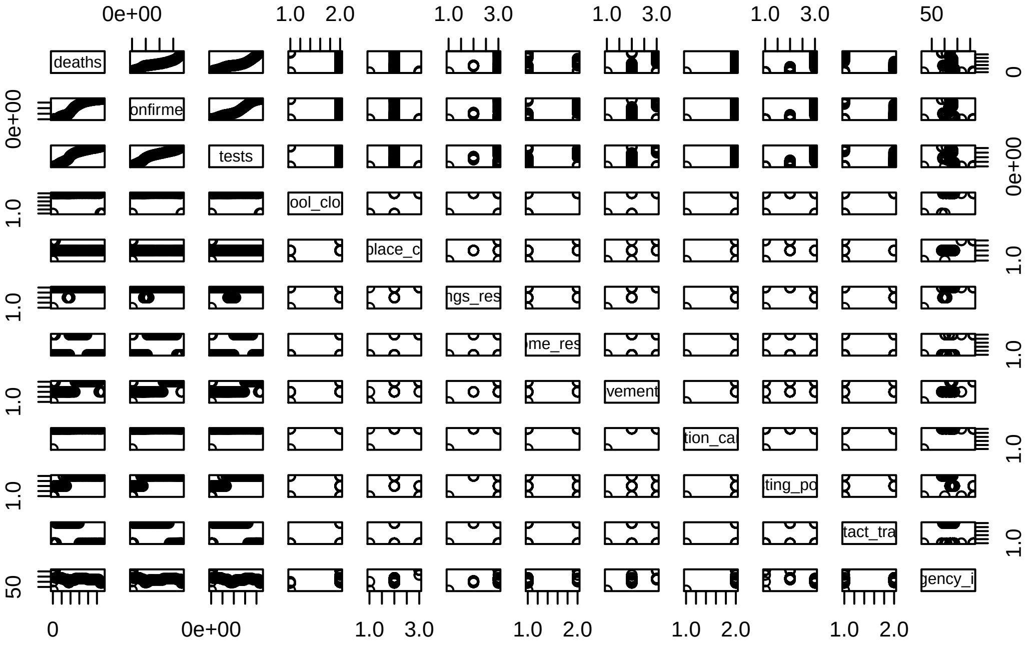
```
mod.full6 <- lm((deaths^{1/3}) ~ fstay_home_restrictions + I(tests^{1/3}) +
    fschool_closing + fcontact_tracing + ftesting_policy + log(stringency_index),
    data = fbase_data1)
summary(mod.full6)
```

```
##
## 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
```

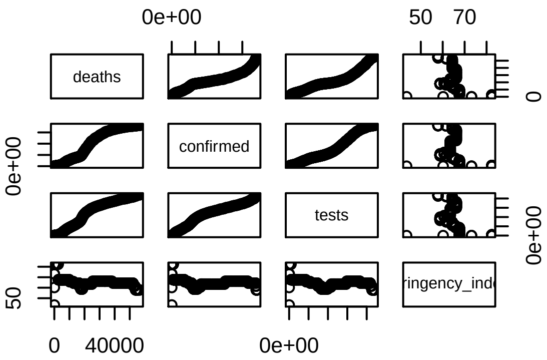
Plot 1: Each of these are really small and it is hard to derive anything useful from them.

```
pairs(deaths ~ confirmed + tests + fschool_closing +
                 fworkplace_closing + fgatherings_restrictions + fstay_home_restrictions +
                 finternal_movement_restrictions  +finformation_campaigns + ftesting_policy +
                 fcontact_tracing + stringency_index, data = fbase_data)
```



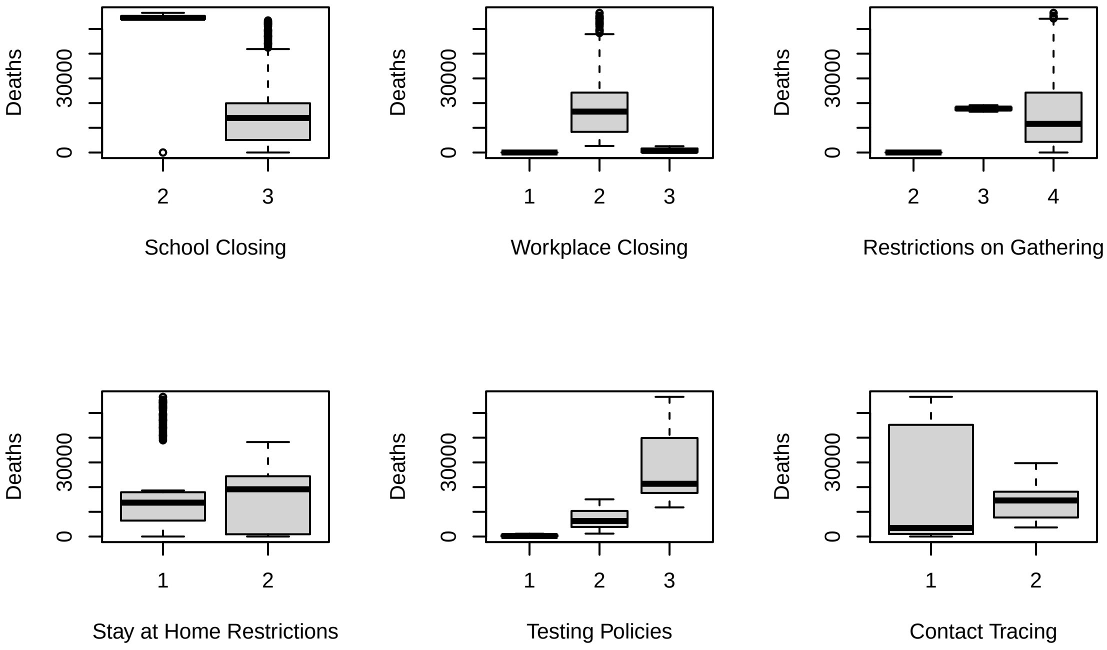
Plot 2: Scatterplots of the numerical variables in the original data set.

```
pairs(deaths ~ confirmed + tests +  stringency_index, data = fbase_data)
```



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

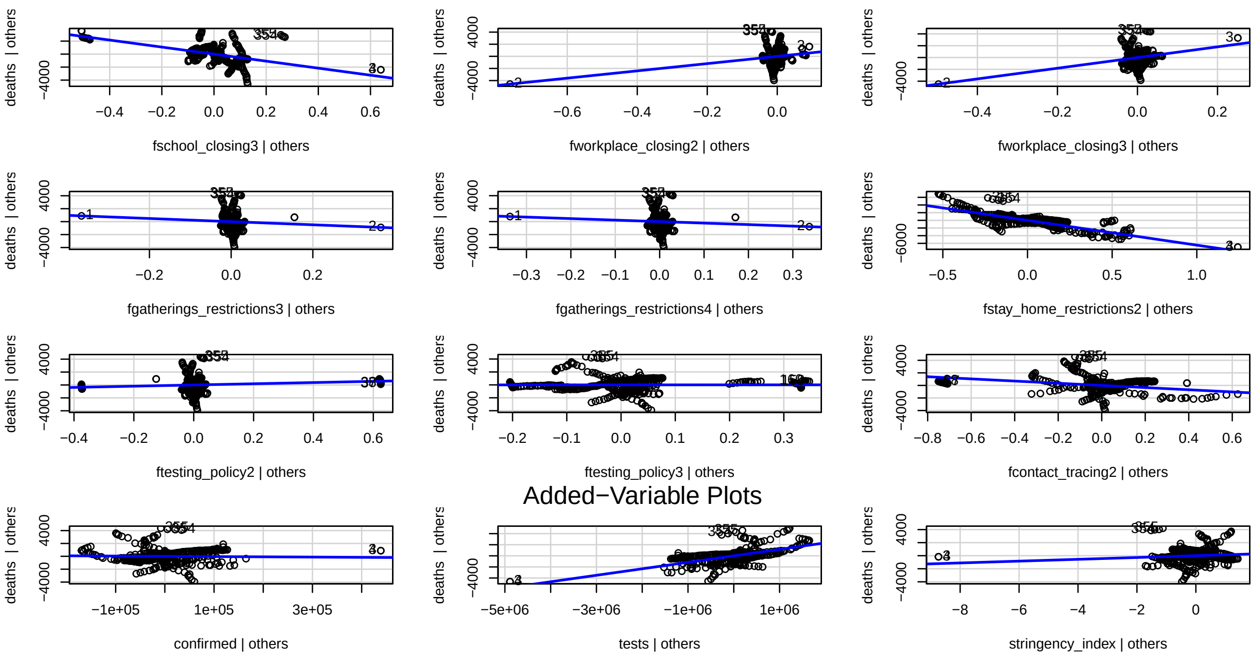
```
par(mfrow=c(2,3))
boxplot(deaths ~ fschool_closing, data = fbase_data, ylab = "Deaths", xlab = "School Closing")
boxplot(deaths ~ fworkplace_closing, data = fbase_data, ylab = "Deaths", xlab = "Workplace Closing")
boxplot(deaths ~  fgatherings_restrictions, data = fbase_data, ylab = "Deaths", xlab = "Restrictions on
boxplot(deaths ~  + fstay_home_restrictions, data = fbase_data, ylab = "Deaths", xlab = "Stay at Home Re
boxplot(deaths ~  + ftesting_policy, data = fbase_data, ylab = "Deaths", xlab = "Testing Policies")
boxplot(deaths ~  + fcontact_tracing, data = fbase_data, ylab = "Deaths", xlab = "Contact Tracing")
```



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

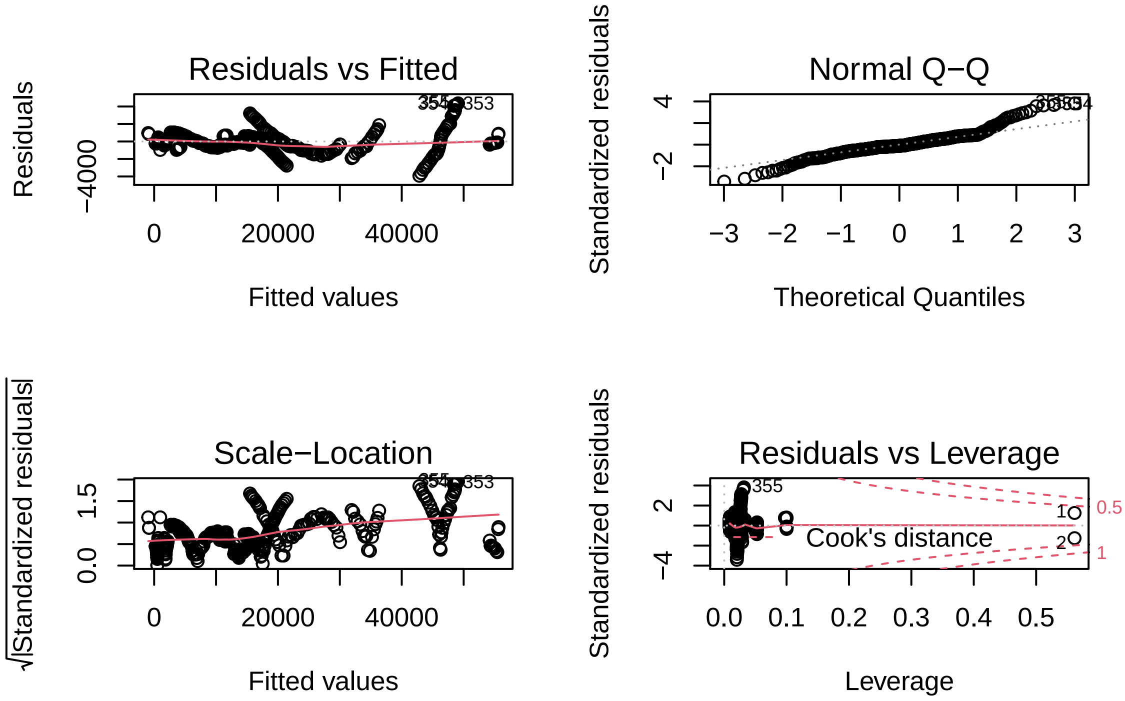
```
mod.full1 <- lm(deaths ~ fschool_closing +
                 fworkplace_closing + fgatherings_restrictions + fstay_home_restrictions +
                        ftesting_policy + fcontact_tracing + confirmed + tests + stringency_index, data = fbas

avPlots(mod.full1)
```



Plot 5: Checking normality prior to transformation

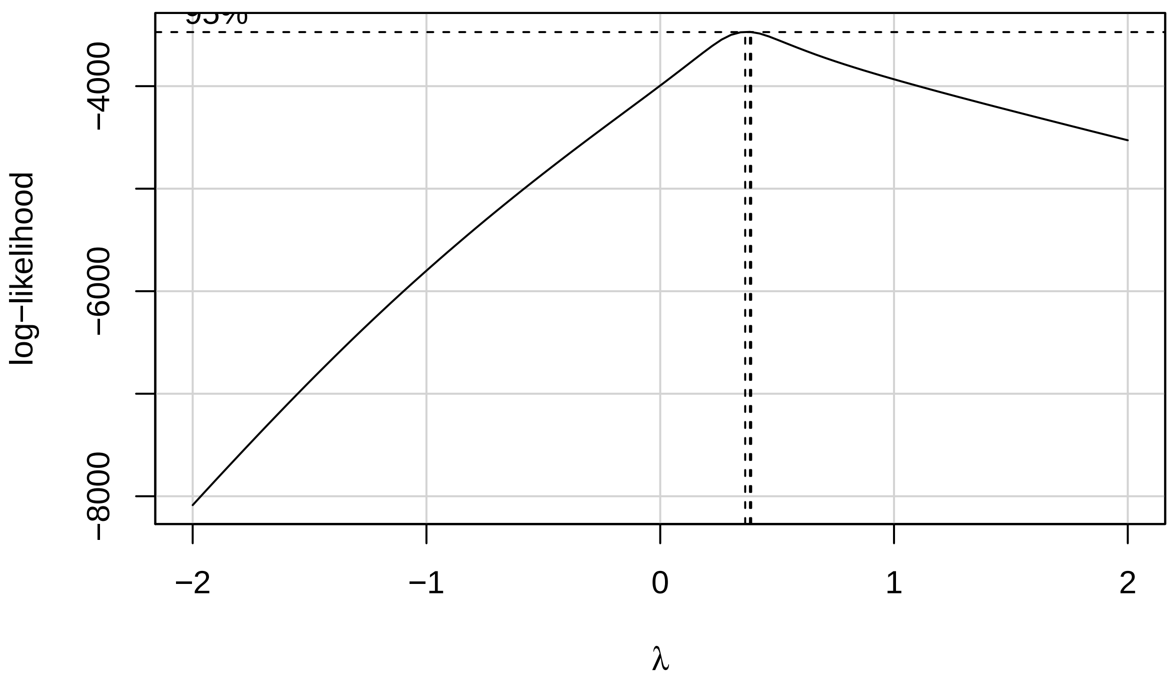
```
mod.full2 <- lm(deaths ~ fstay_home_restrictions + tests + fworkplace_closing +
    fschool_closing + fcontact_tracing + ftesting_policy + stringency_index,
    data = fbase_data)
par(mfrow=c(2,2))
plot(mod.full2)
```



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

```
mod.full3 <- lm(deaths ~ fstay_home_restrictions + I(tests^{1/3}) + fworkplace_closing +
    fschool_closing + fcontact_tracing + ftesting_policy + log(stringency_index),
    data = fbase_data)

bcTrans <- boxCox(mod.full3)
```

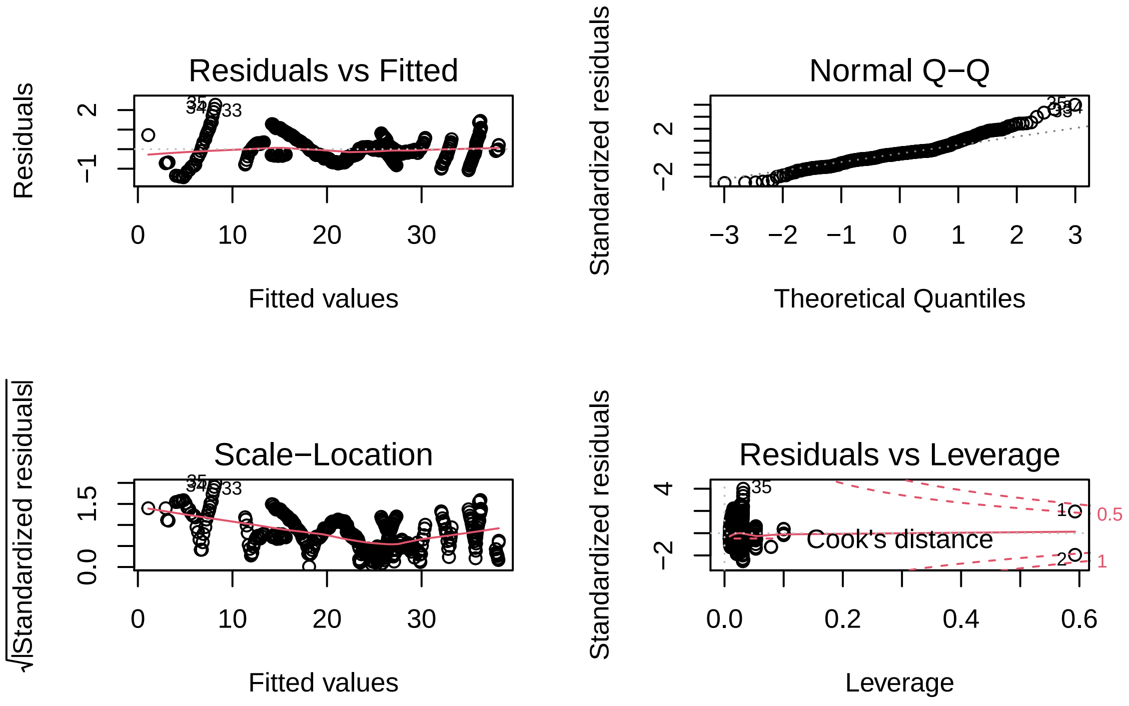


```
opt.lambda <- bcTrans$x[which.max(bcTrans$y)]
opt.lambda
```

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

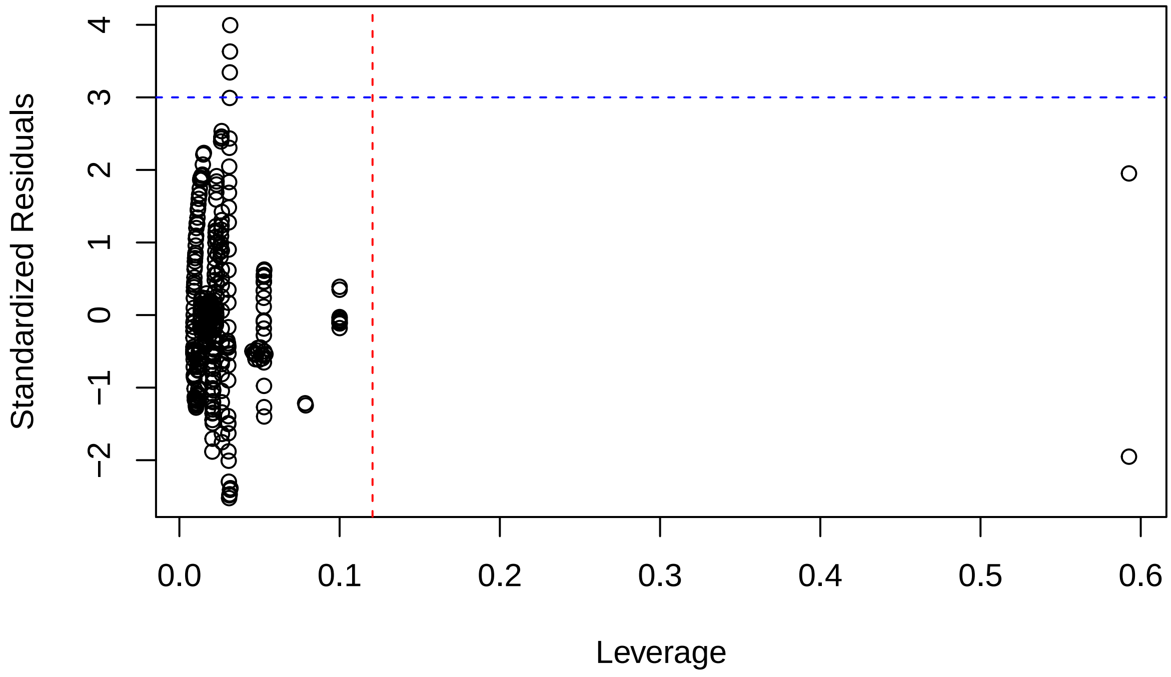
Plot 7: Post transformation linearity check. The residuals vs fitted plot is still very patterned and the Q-Q plot is still heavy-tailed but less so.

```
mod.full4 <- lm((deaths^{1/3}) ~ fstay_home_restrictions + I(tests^{1/3}) + fworkplace_closing +
    fschool_closing + fcontact_tracing + ftesting_policy + log(stringency_index),
    data = fbase_data)
par(mfrow=c(2,2))
plot(mod.full4)
```

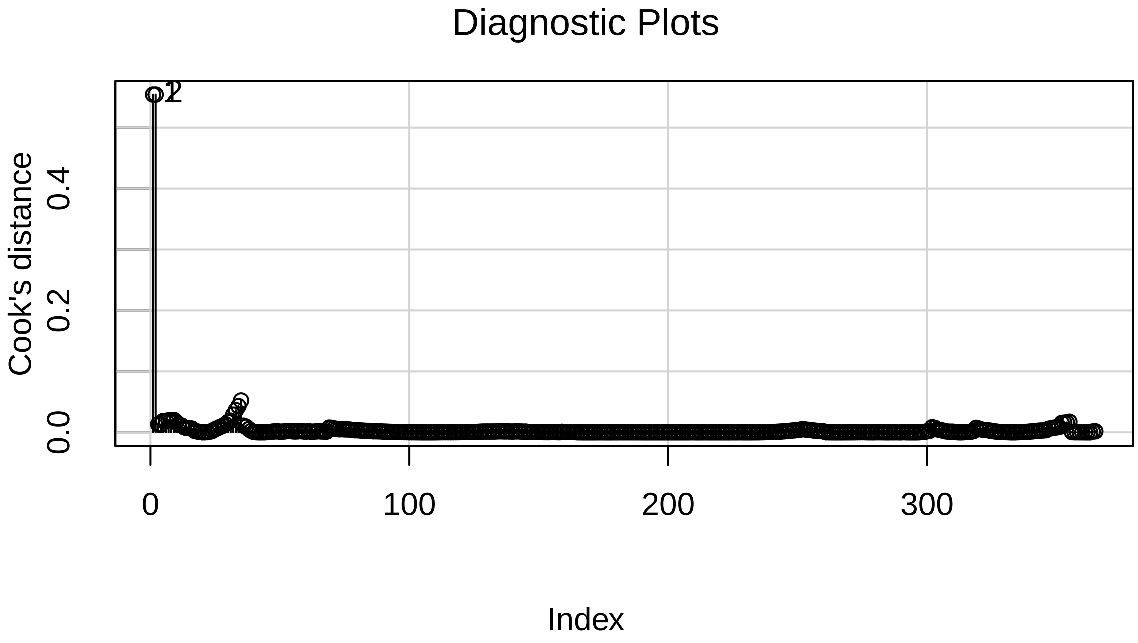


Plot 8: From the Cook's InfuleceIndexPlot and hat-values influenceIndexPlot we can see that we should definitely look at points 1 and 2.

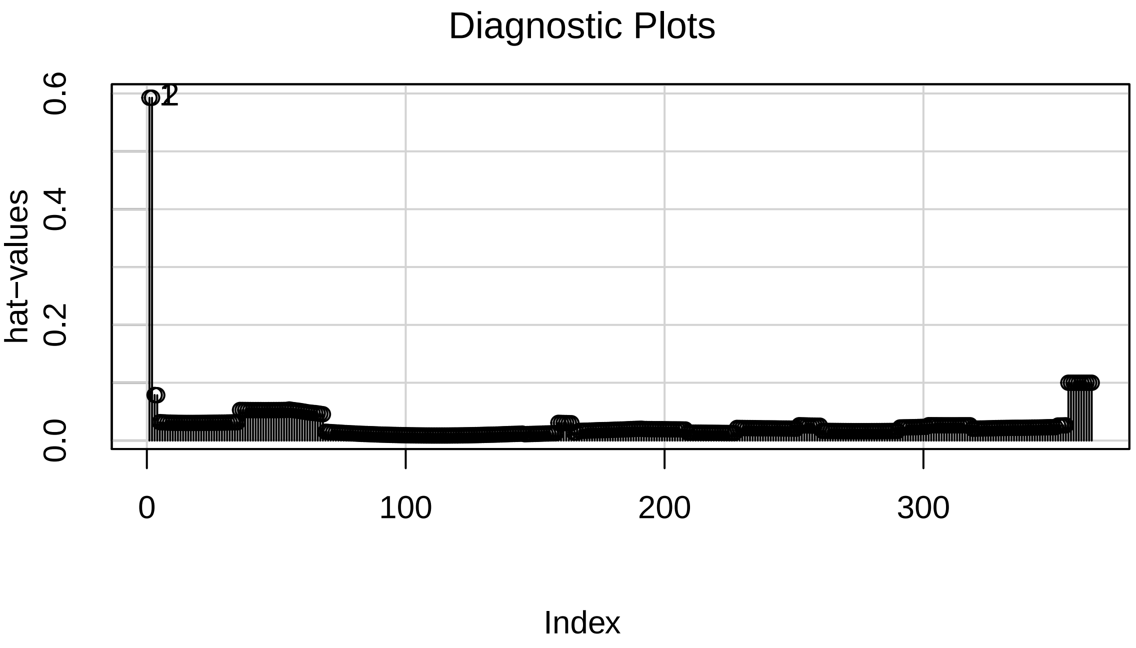
```
# plot of high leverage points and outliers
plot(hatvalues(mod.full4), rstandard(mod.full4), xlab = "Leverage",
     ylab = "Standardized Residuals")
abline(v = 4*(p+1)/n, col = "red", lty = 2)
abline(h = c(-3,3), col = "blue", lty =2)
```



```
# Cook's points and hat values
mod.full4.cooks <- cooks.distance(mod.full4)
influenceIndexPlot(mod.full4, vars = "Cook")
```

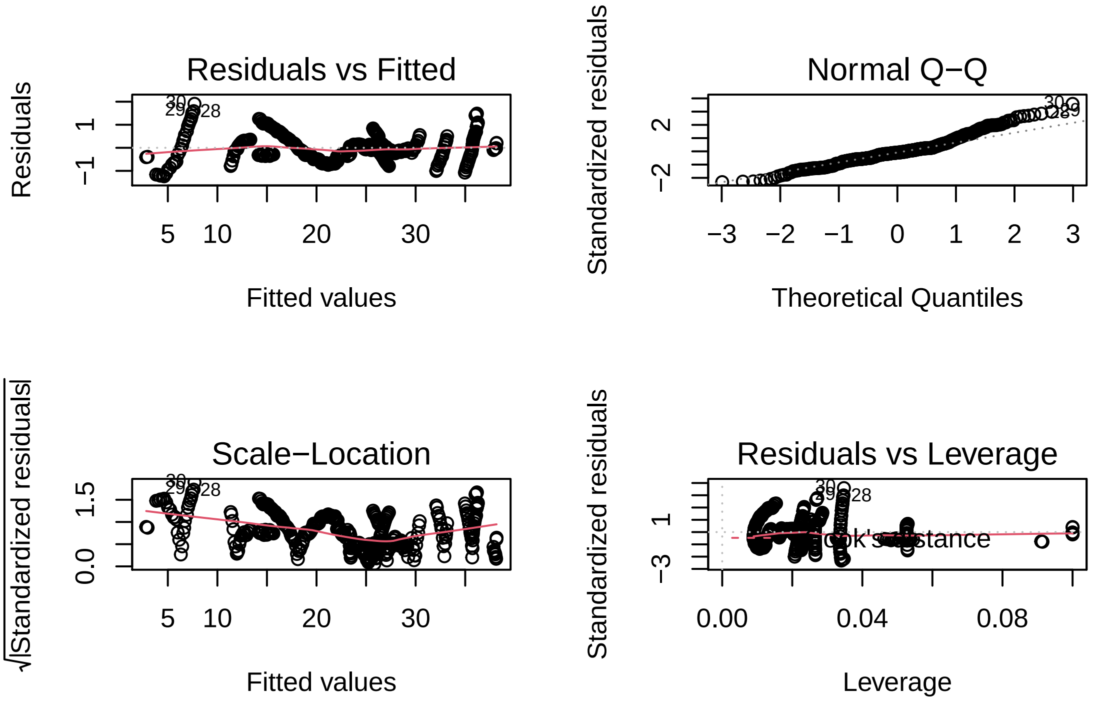


```
influenceIndexPlot(mod.full4, vars = "hat")
```



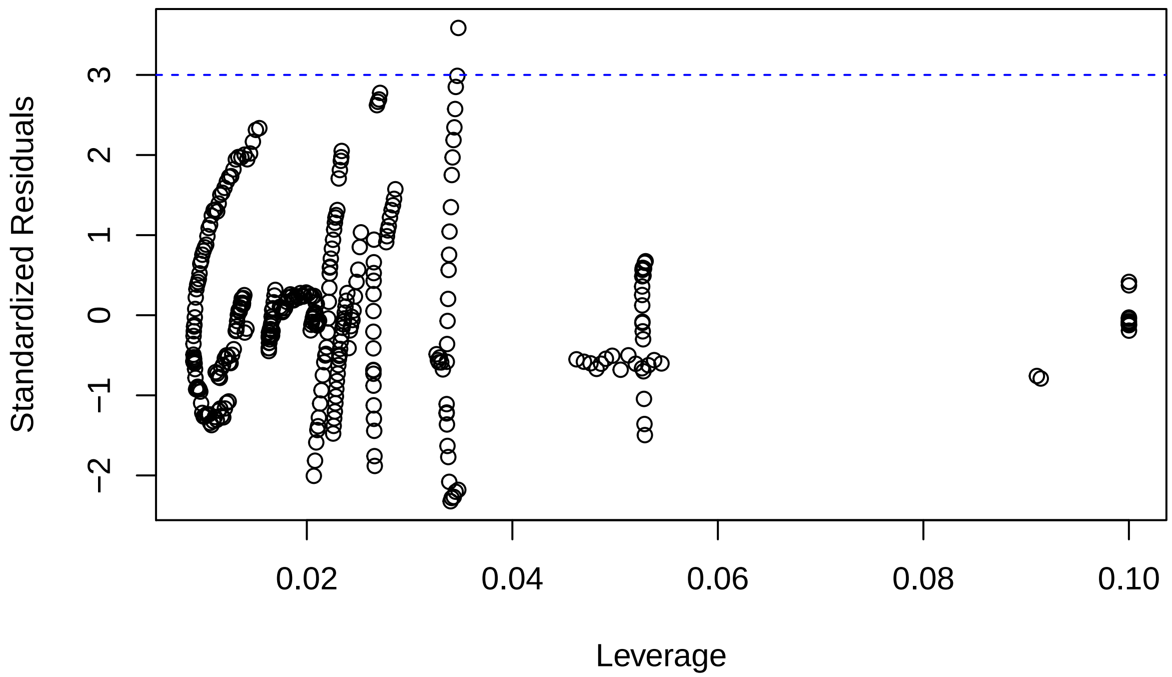
Plot 9: Checking the linearity after the removal of rows 1, 2, 33, 34, and 35. Looks a little better.

```r
mod.full5 <- lm((deaths^{1/3}) ~ fstay_home_restrictions + I(tests^{1/3}) + fworkplace_closing +
    fschool_closing + fcontact_tracing + ftesting_policy + log(stringency_index),
    data = fbase_data1)
par(mfrow=c(2,2))
plot(mod.full5)
```

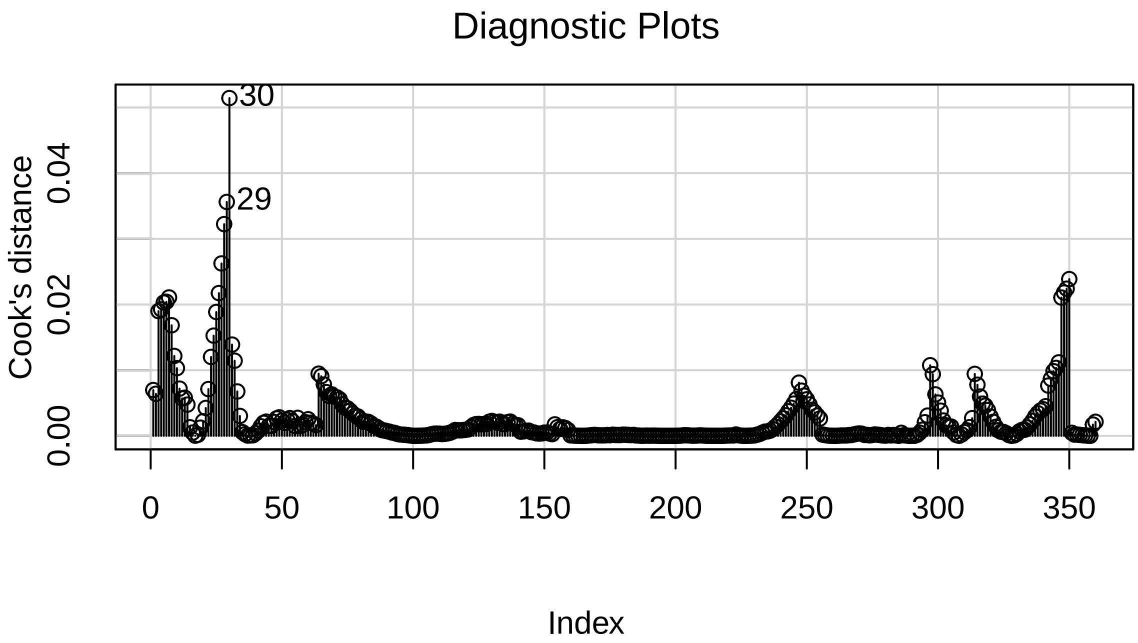


Plot 10: Checking Cook's plot again after the removal of rows 1, 2, 33, 34, and 35. Looks better.

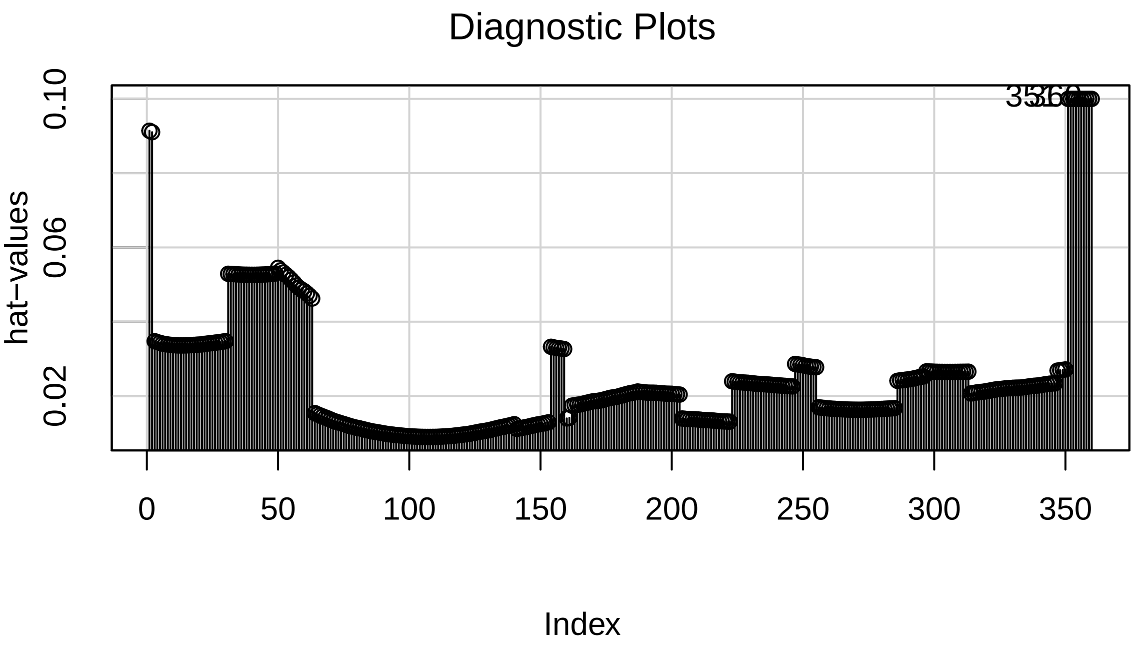
```
# plot of high leverage points and outliers
plot(hatvalues(mod.full5), rstandard(mod.full5), xlab = "Leverage",
     ylab = "Standardized Residuals")
abline(v = 4*(p+1)/n, col = "red", lty = 2)
abline(h = c(-3,3), col = "blue", lty =2)
```



```
# Cook's points and hat values
mod.full4.cooks <- cooks.distance(mod.full5)
influenceIndexPlot(mod.full5, vars = "Cook")
```

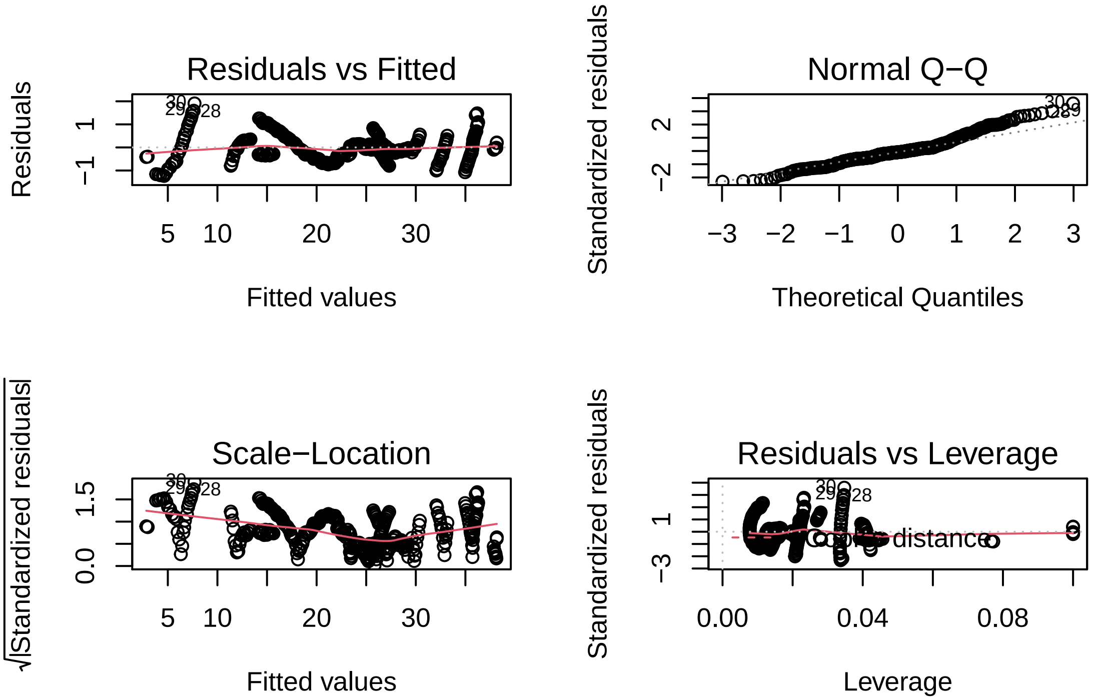


```
influenceIndexPlot(mod.full5, vars = "hat")
```



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.

```
par(mfrow=c(2,2))
mod.full6 <- lm((deaths^{1/3}) ~ fstay_home_restrictions + I(tests^{1/3}) +
    fschool_closing + fcontact_tracing + ftesting_policy + log(stringency_index),
    data = fbase_data1)
plot(mod.full6)
```



Appendix 2: Exploratory analysis not used in final paper

Github Link:

Appendix 3: 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).