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

```
## 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
```

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 $< 2.2e\text{-}16$ 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. 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

Finally we looked the residuals vs fitted and Q-Q plot to see if the linear assumptions were violated.

Since having 9 dummy variables creates a very complicated model we decided to only use 1 categorical variable in the final model, we kept contact tracing.

Diagnostic Checks I:

Exploratory Analysis II:

Insert Google Mobility Analysis here

Diagnostic Checks II:

Insert Google Mobility Diagnostics here

Exploratory Analysis III:

Insert World Data Analysis here

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

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

Code 1:

```
##
## 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
```

The BIC is the same as the AIC, so 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
```

Table 1: H_0 : The null mo

```
## 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
```

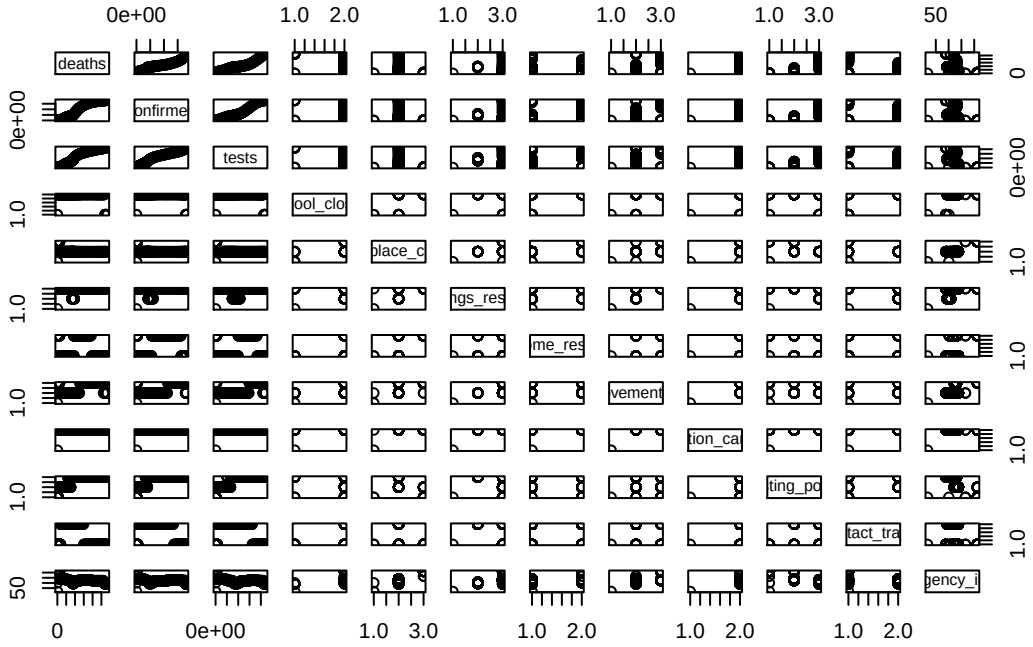
```
summary(step_bic)
```

Table 3:

```
##
## 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
```

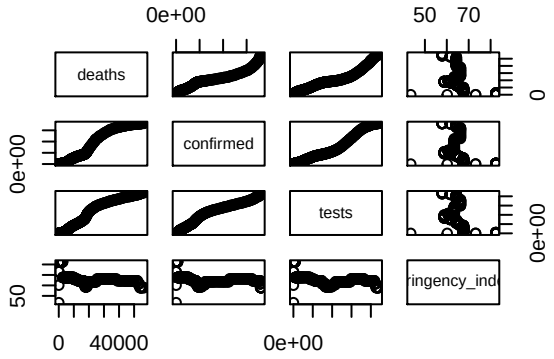
Plot 1: Each of these are really small and it is hard to derive anything useful from them.

```
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)
```



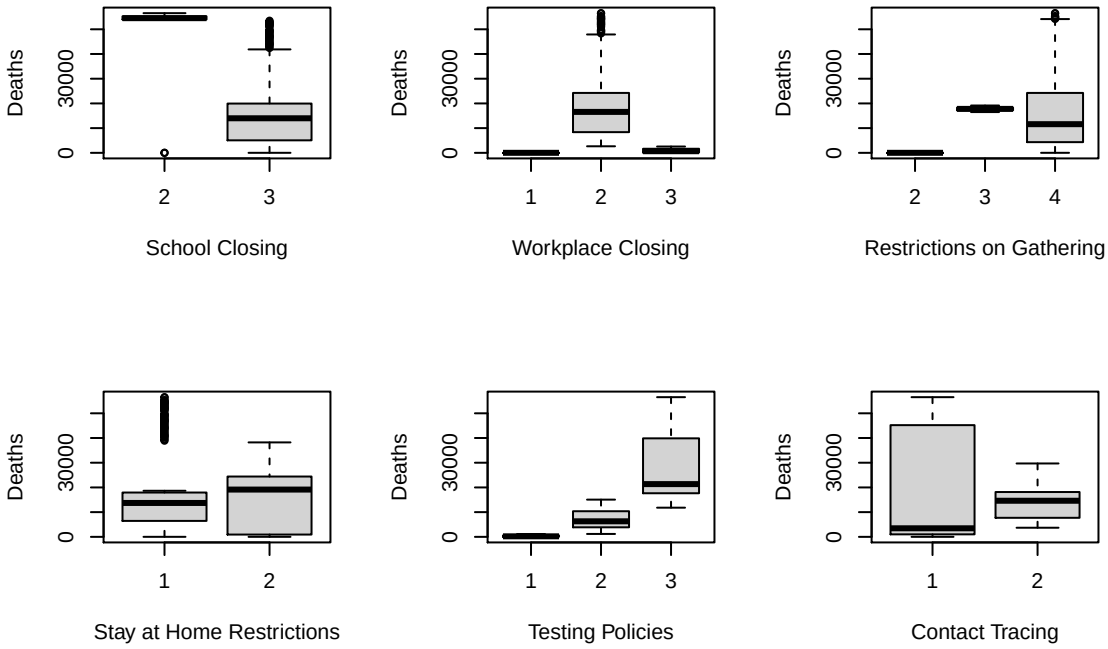
Plot 2: Scatterplots of the numerical variables in the original data set.

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



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

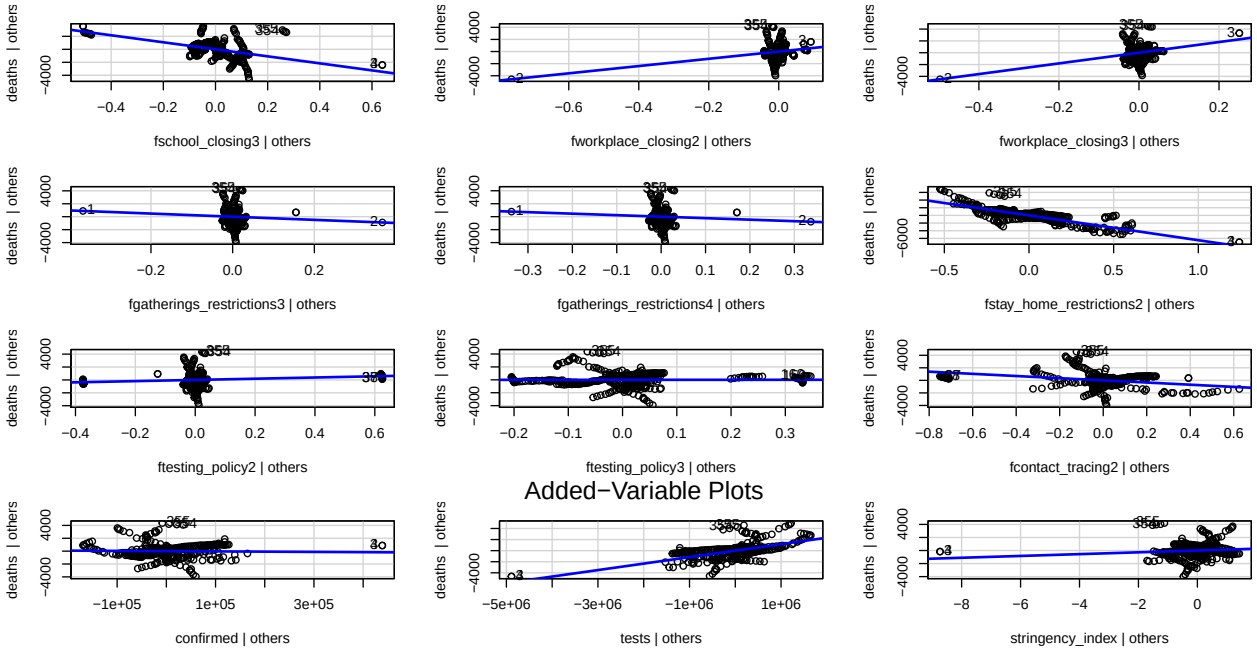
```
par(mfrow=c(2,3))
boxplot(deaths ~ fschool_closing, data = fbase_data, ylab = "Deaths", xlab = "School Closing")
boxplot(deaths ~ fworkplace_closing, data = fbase_data, ylab = "Deaths", xlab = "Workplace Closing")
boxplot(deaths ~  fgatherings_restrictions, data = fbase_data, ylab = "Deaths", xlab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



Appendix 2: Exploratory analysis not used in final paper

Base Data:

Google Mobility Data

World Bank Data Story

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).