How much did restricting mobility help in controlling the spread of COVID?

Satheesh Joseph, Paco Valdez, Yi Zhang

Introduction

Overview

The Covid-19 pandemic has taken more than 2.5 million lives worldwide. In the U.S. alone, more than half a million people died because of it. COVID-19 is thought to spread mainly through close contact from person to person, including among people who are physically near each other (within about 6 feet).[1]

As the virus spread, many cities and other areas have implemented a shutdown policy that significantly restricted the mobility of the population. Specifically, many companies have opted for a complete work-from-home policy.

As a team of Data Scientists, we're interested in assessing the effectiveness of restricting mobility in controlling the spread of Covid-19.

Research Question

Rather than investigating the effect of social distancing on a micro and personal interaction level, we are interested in its effect on a policy level. Specifically, the Research Question we're asking is:

`How much did restricting social mobility help in controlling the spread of COVID?`

The basic causal theory we're working under can be expressed by the diagram below:

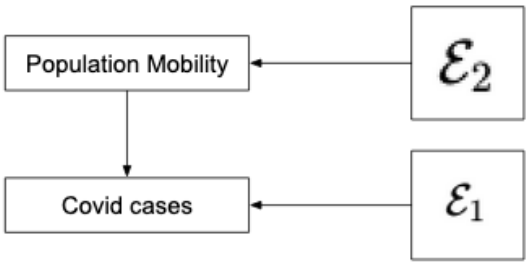

Figure 1: The Basic Causal Theory

Operationalization

To operationalize the research question, we propose the following plan for the causal models.

Firstly, we will use the New York Times[2] Covid-19 data for the number of Covid cases per State. And we will use the Covid-19 Community Mobility Report[3] for the population mobility scores. For the base model,

we will use the "residential mobility score" as a relatively stable representative of the general population's social mobility.

Secondly, mobility have drastically different effect on the spread of the virus depending on the population density. Thus, we believe that having the raw number of cases as the outcome variable can be misleading. Therefore, we will be using the number of cases per 100,000 residents as the outcome variable to normalize the population density. The population density data will come from the COVID-19 US State Policy Database[4].

More formally, our base model will assume a causal relationship of the form:

$$cases\_per\_100k\_residents = \beta_0 + \beta_1 residential\_mobility\_score$$

where each data point will represent a State in the U.S., and the outcome variable will be the total number of Covid-19 cases per 100,000 residents for that state since 2020-02-15 up until 2020-07-15, when the mask wearing survey[5] was taken.

And the independent variable will be the **average** residential mobility score of the state in the same time period.

Finally, for a more advanced model, we propose to use some control variables to assess the difference mobility score makes in relation to other factors, notably we'll include whether the "Stay at Home" order is in effect and whether the face mask policy is in effect and how long has been in place.

The mask wearing survey data will be operationalized by transforming it into probability of people wearing mask. The probability of people wears masks is calculated by assuming that survey respondents who answered 'Always' were wearing masks all of the time, those who answered 'Frequently' were wearing masks 80 percent of the time, those who answered 'Sometimes' were wearing masks 50 percent of the time, those who answered 'Rarely' were wearing masks 20 percent of the time and those who answered 'Never' were wearing masks none of the time.

For the most inclusive model, we will explore all the other mobility scores, including to workplaces, transit, grocery & pharmacy.

The models

Data Exploration

```
##     state           cases_per_100k      workplaces        transit
##  Length:50          Min.   :  89.48   Min.   :-36.76   Min.   :-49.822
##  Class :character   1st Qu.: 594.01   1st Qu.:-31.27   1st Qu.:-32.548
##  Mode  :character   Median : 865.29   Median :-28.46   Median :-22.882
##                     Mean   : 917.10   Mean   :-28.10   Mean   :-21.027
##                     3rd Qu.:1220.34   3rd Qu.:-24.32   3rd Qu.: -7.003
##                     Max.   :2091.42   Max.   :-20.57   Max.   :  9.651
##     grocery          residential       stay_home          fm
##  Min.   :-17.1842   Min.   : 5.658   Min.   :0.00   Min.   :0.00
##  1st Qu.: -4.8586   1st Qu.: 7.850   1st Qu.:1.00   1st Qu.:1.00
##  Median : -0.4868   Median : 9.441   Median :1.00   Median :1.00
##  Mean   : -0.6305   Mean   : 9.667   Mean   :0.78   Mean   :0.82
##  3rd Qu.:  3.0477   3rd Qu.:11.206   3rd Qu.:1.00   3rd Qu.:1.00
##  Max.   : 13.5066   Max.   :14.947   Max.   :1.00   Max.   :1.00
##    mask_prob        mask_days
##  Min.   :0.5645   Min.   :  0.00
##  1st Qu.:0.7095   1st Qu.: 29.75
##  Median :0.7736   Median : 69.50
##  Mean   :0.7761   Mean   : 55.92
##  3rd Qu.:0.8448   3rd Qu.: 75.00
##  Max.   :0.9327   Max.   :103.00
```

All the variables have their means and medians very close to each other, that suggests that there is no severe skewness.

By looking at the scatter plots we can notice that `workplaces` is the inverse of the `residential` mobility variable. Using both in a OLS regression probably will cause the coefficients to cancel each other. The `residential` variable shows promising results, it looks like it has a very linear relationship with `cases_per_100k`.

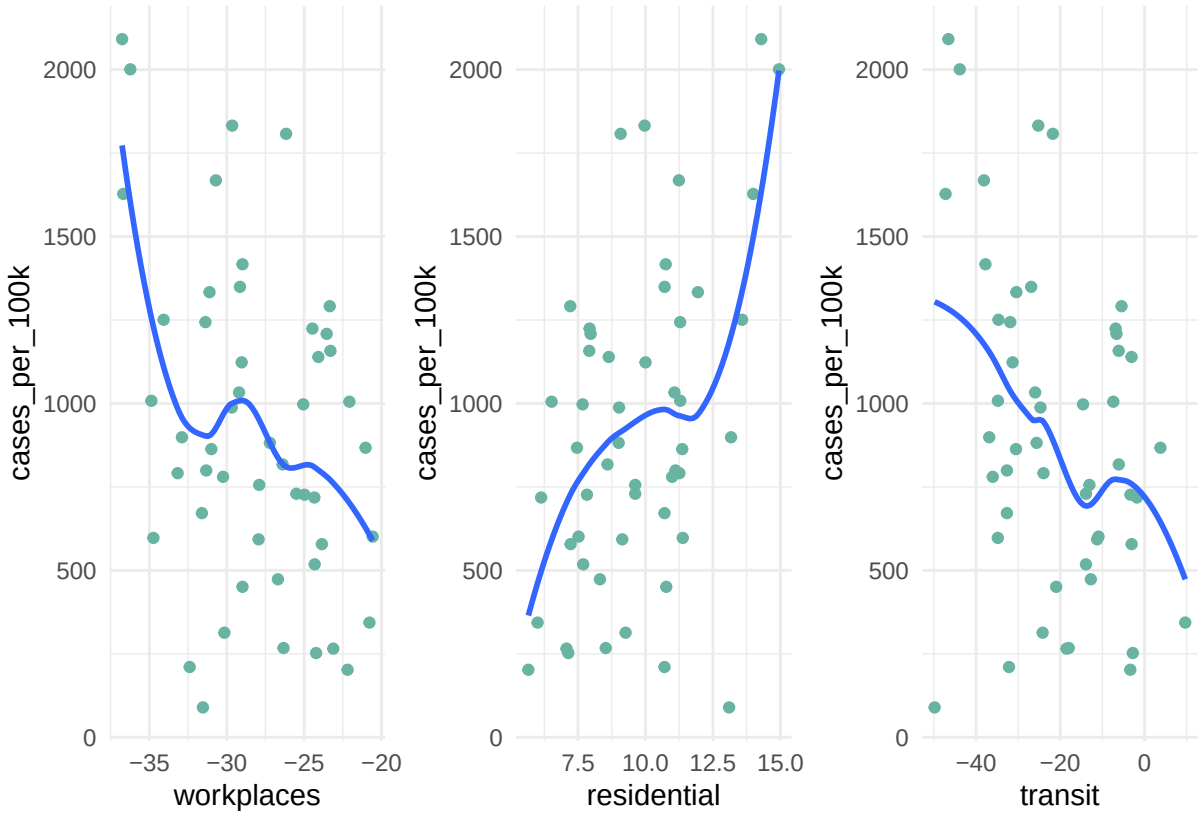

Figure 2: Scatter Plots for Cases per 100k and selected mobility variables

The Linear Models

Our intuition told us that the residential variable would be the best proxy to represent mobilty. At this point in time during the pandemic, the U.S. was registering its second wave of Covid-19 cases and non-essential business were still closed in most of the sates. One way to explain this relationship is that in the states where the residential mobility score increased, people were not sheltering at their homes. Instead people were going out with friends and family, thus spreading the virus.

Model One

In our initial model we will use the residential mobility score and cases per 100,000 residents:

$$cases\_per\_100k\_residents = \beta_0 + \beta_1 residential\_mobility\_score$$

```
##
## Call:
## lm(formula = cases_per_100k ~ residential, data = data)
##
## Residuals:
```

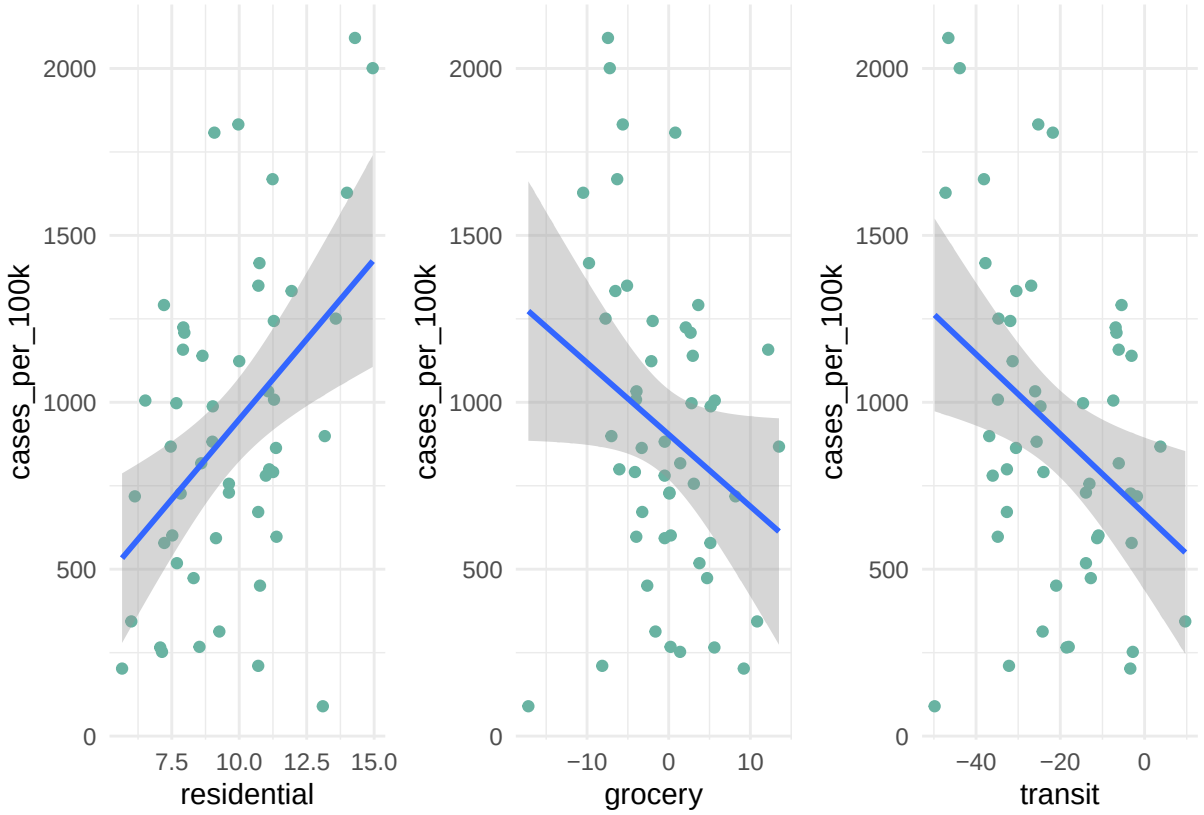

Figure 3: Scatter Plots for Cases per 100k and selected mobility variables with OLS regression line and standard errors

```
##      Min       1Q   Median       3Q      Max
## -1156.38  -277.06   -29.98   314.56   946.99
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -9.003    272.637  -0.033  0.97379
## residential   95.801     27.466   3.488  0.00105 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 437.7 on 48 degrees of freedom
## Multiple R-squared:  0.2022, Adjusted R-squared:  0.1856
## F-statistic: 12.17 on 1 and 48 DF,  p-value: 0.001052
```

We got a very significant p-value, which highlights the strong relationship between these two variables.

Model Two

In the second model we wanted to test if the probability of wearing a mask would have a significant impact on the amount of Covid-19 cases. Although by looking at the scatter plot of Cases per 100k and mask wearing probability there is no clear relationship, and the slope of the OLS line has the opposite direction to what our intuition tell us.

So, our second model contains the residential mobility score and the face mask wearing probability. The coefficient for Mask wearing probability (`mask_prob`), at least has the correct direction but fails to achieve

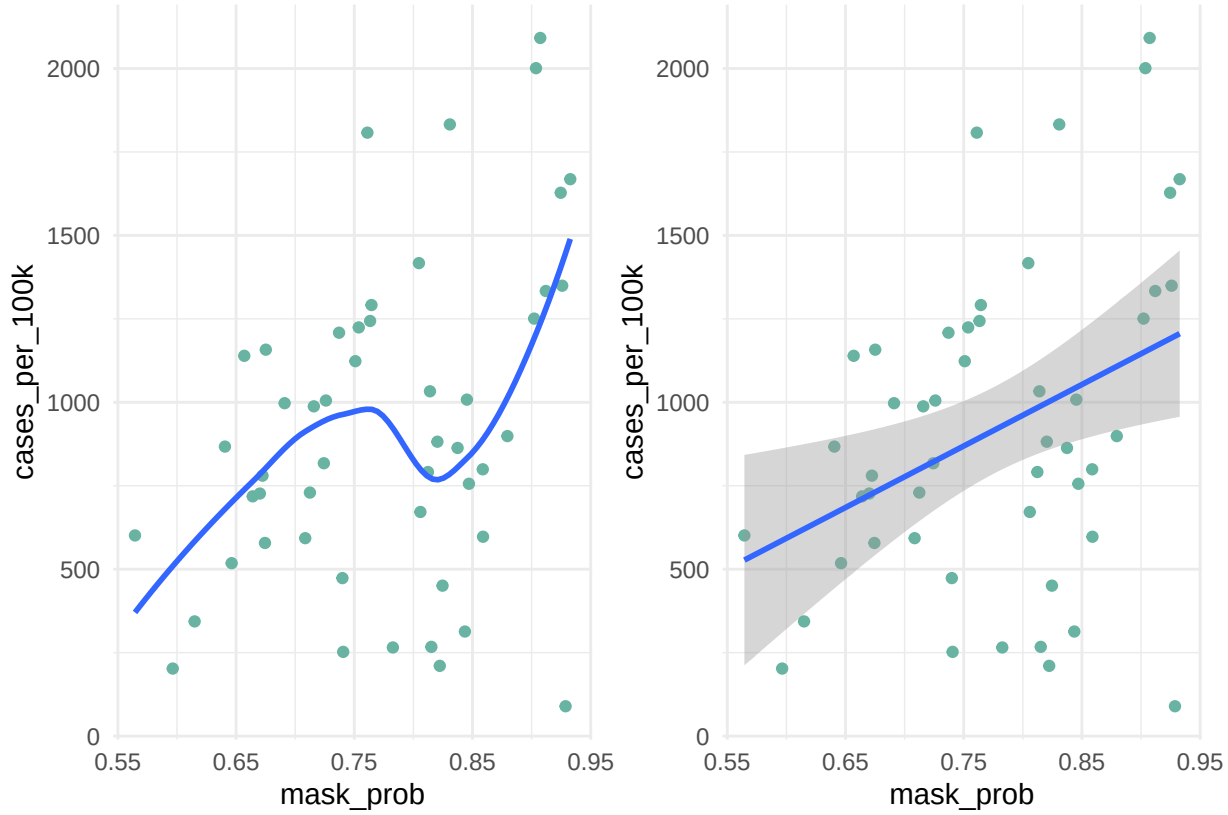

Figure 4: Scatter Plots for Cases per 100k and mask wearing probability

significance.

The residential mobility score coefficient is still significant, which give us more trust in that variable.

$$cases\_per\_100k\_residents = \beta_0 + \beta_1 residential\_mobility\_score + \beta_2 mask\_wearing\_probability$$

```
##
## Call:
## lm(formula = cases_per_100k ~ residential + mask_prob + fm, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1174.75  -271.92   -22.61   307.27   972.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -105.20     583.42  -0.180   0.8577
## residential    96.88      47.62   2.035   0.0477 *
## mask_prob     234.59    1160.85   0.202   0.8407
## fm           -117.45     190.97  -0.615   0.5416
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 445.3 on 46 degrees of freedom
## Multiple R-squared:  0.2087, Adjusted R-squared:  0.1571
## F-statistic: 4.045 on 3 and 46 DF,  p-value: 0.01237
```

Model Three

Our secondary question was if the other mobility variables could have a significant impact on the cases per 100,000 variable. In our data exploration, workplaces also looked promising but it looked that it acted in the opposite direction than residential. When we tested this variable alone we confirmed our intuition, the workplaces mobility score is also significant but acts in the opposite direction than residential.

When both are present residential still has a more significant effect. And the coefficient for workplaces mobility score is significantly lower and in the opposite direction than the previous model.

$$cases\_per\_100k\_residents = \beta_0 + \beta_1 residential\_mobility\_score + \beta_2 workplaces\_mobility\_score$$

```
##
## Call:
## lm(formula = cases_per_100k ~ residential + workplaces, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1281.37  -270.89   -45.09   295.65   932.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   485.43     492.39   0.986   0.3292
## residential   178.37      73.86   2.415   0.0197 *
## workplaces     46.01      38.23   1.203   0.2348
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 435.7 on 47 degrees of freedom
## Multiple R-squared:  0.2261, Adjusted R-squared:  0.1931
## F-statistic: 6.864 on 2 and 47 DF,  p-value: 0.002425
```

Our definitive third model is the comparison of all the mobility score variables, with the exception of the workplaces variable since it's the opposite of residential. The residential mobility score captures the information contained in the other mobility variables.

$$cases\_per\_100k\_residents = \beta_0 + \beta_1 residential\_mobility\_score + \beta_2 parks\_mobility\_score + \beta_3 grocery\_mobility\_score +$$

```
##
## Call:
## lm(formula = cases_per_100k ~ residential + workplaces + grocery +
##     transit + fm, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1029.12  -286.81   -77.41   299.90  1063.31
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -92.768    769.430  -0.121   0.9046
## residential  232.550     86.796   2.679   0.0103 *
## workplaces    46.183     41.588   1.110   0.2728
## grocery       32.569     22.584   1.442   0.1564
## transit       -3.475     11.824  -0.294   0.7702
```

```
## fm             8.395    192.443   0.044   0.9654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 438.5 on 44 degrees of freedom
## Multiple R-squared:  0.2662, Adjusted R-squared:  0.1828
## F-statistic: 3.192 on 5 and 44 DF,  p-value: 0.01521
```

The Regression Table

Model Limitations

TimeSeries:

The Research Question under study is well suited for a time series analysis. The initial discussion and work, were poised to well take off in including a weekly average and correlation with the cases offset by couple of weeks. As per the instruction and the limited machinery in place, the team has decided to take an average for the states from a time window where all the variables were produced and use it in the model.

Reverse Causality:

The Research Question under study finds relationship between Mobility and COVID cases. But, there is also the reverse path where more COVID cases creates fear among people. Thus, people limiting outdoor activities.

All Percent drops are not equal:

A 10% drop in mobility is really good for sparsely populated regions to go below a certain threshold in transmitting the virus. The same cannot be applied to densely populated cities. Thus, the non-linearity in the measures taken vs impact seen is omitted.

Classic Linear Model assumptions

1. IID Sampling None of the variables seem to violate IID.

2. Linear Conditional Expectation By looking at the predicted vs. residuals of the model it looks that relationship is very linear.

3. No Perfect Collinearity

```
## (Intercept) residential  workplaces     grocery     transit          fm
##  -92.767747  232.550240   46.182751   32.568578   -3.474629    8.395254
```

None of the coefficients was dropped, that means there is no Perfect Collinearity.

4. Homoskedastic Errors The scale-location plot is very close to a flat line, which lead us to believe that there is no major issues with heteroskedasticity.

5. Normally Distributed Errors The histogram looks does not show a significant deviation from a normal distribution and the qqplot also shows very low deviation from normality.

In general we can say that all the 5 CLM Assumptions are met.

Omitted Variables

Not following mandates:

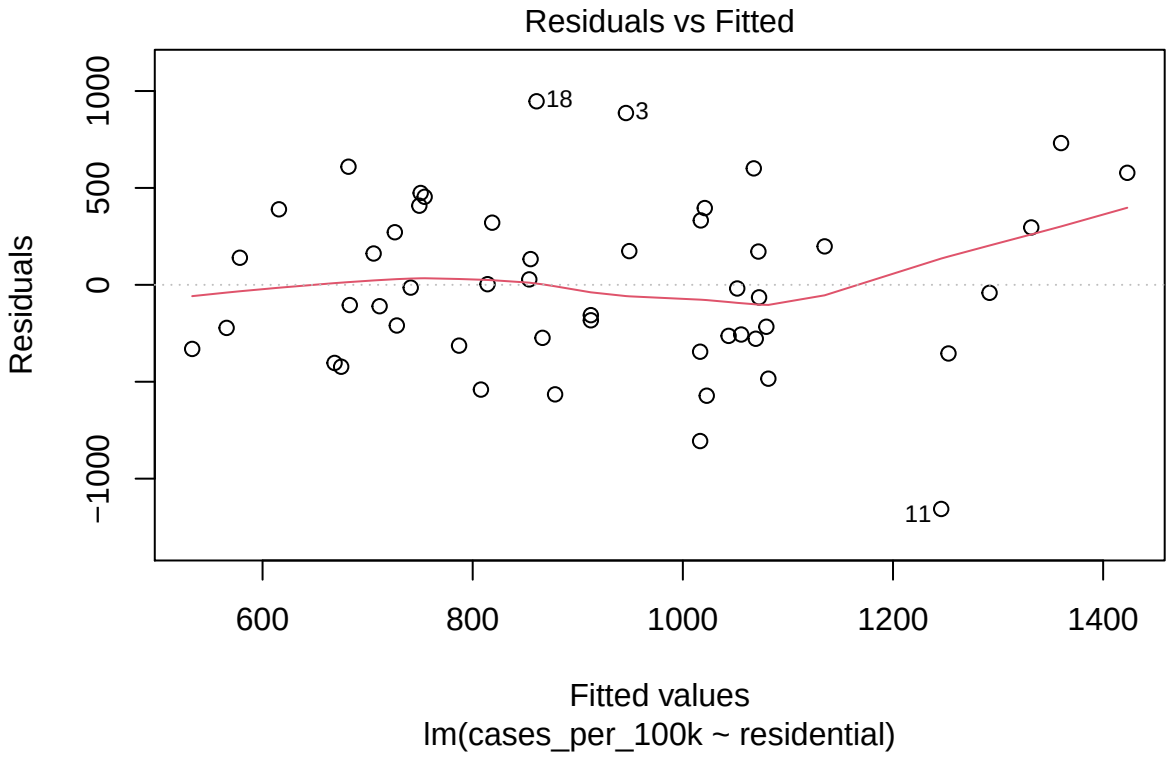

Figure 5: Residuals vs Fitted plot for Model One

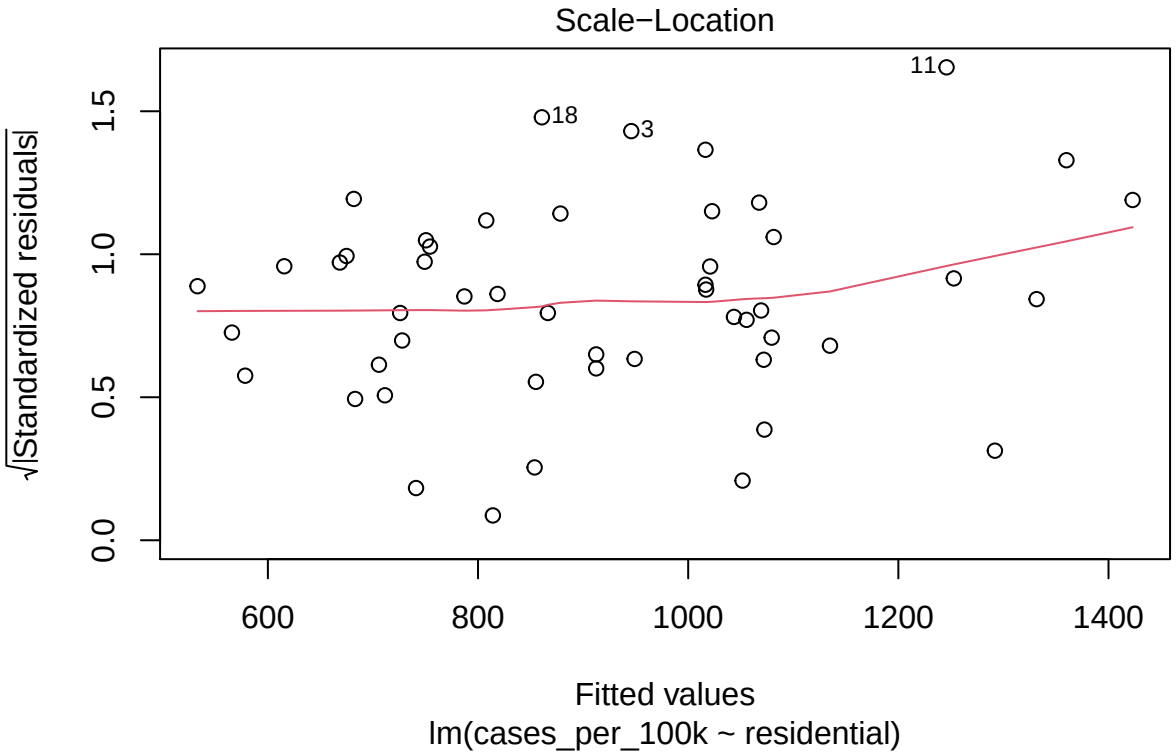

Figure 6: Scale-Location plot for Model One

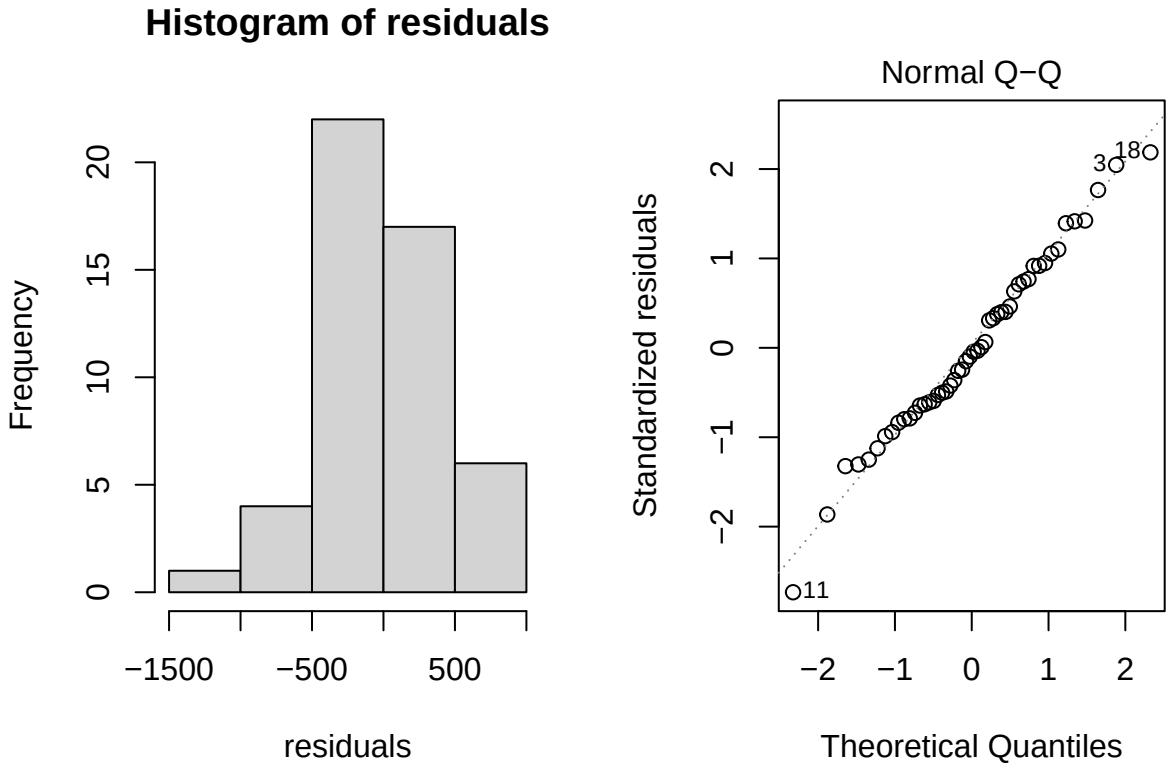

Figure 7: Histogram of residuals for model one and qq-plot for model one

Though, State and County put restrictions, including Mask mandate, Curfew, Social distance, Quarantine, park, school and restaurant closures, how much general public abide to these mandates is a very good [Omitted] indicator that influence the total cases. The omitted variable positively impacts outcome and positively related with measured variable, thus moving away from zero.

International/Non-Local Mobility:

As we all know Port of Entries like New York, San Francisco, Chicago and Seattle are the first affected places. State wide policies restricting mobility has varying effects on how close a location is to the airport or other transit points. Thus, closeness to transit stations and airport is an omitted variable. The omitted variable positively impacts outcome and positively related with measured variable, thus moving away from zero.

Conclusion

References:

[1]: https://www.cdc.gov/coronavirus/2019-ncov/faq.html
[2]: https://github.com/nytimes/covid-19-data
[3]: https://www.google.com/covid19/mobility
[4]: https://www.tinyurl.com/statepolicies
[5]: https://github.com/nytimes/covid-19-data/blob/master/mask-use