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:

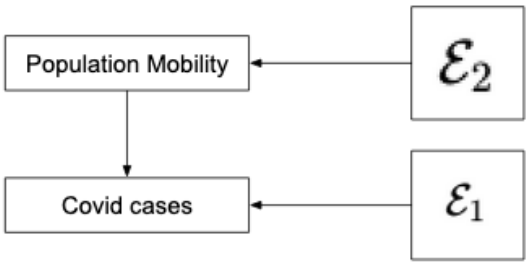

Figure 1: The Basic Causal Theory

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 two factors about facial masks. One is the probability of someone wearing a mask, gathered from New York Time county level data, and the other is whether the mask wearing mandate has been in place for the state.

The probability of people wearing 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, parks, and grocery & pharmacy.

The models

Data Exploration

First of all, we gather and transform all the data identified from above to State level aggregated in the time range of 2020-02-15 to 2020-07-15.

The table below shows a summary of all the variables we intend to use.

	variables (N = 50)
cases_per_100k	
minimum	89.48
median (IQR)	865.29 (594.01, 1,220.34)
mean (sd)	917.10 ± 485.03
maximum	2,091.42
workplaces	
minimum	-36.76
median (IQR)	-28.46 (-31.27, -24.32)
mean (sd)	-28.10 ± 4.40
maximum	-20.57
transit	
minimum	-49.82
median (IQR)	-22.88 (-32.55, -7.00)
mean (sd)	-21.03 ± 14.65

	variables (N = 50)
maximum	9.65
grocery	
minimum	-17.18
median (IQR)	-0.49 (-4.86, 3.05)
mean (sd)	-0.63 ± 6.15
maximum	13.51
parks	
minimum	-36.60
median (IQR)	53.34 (31.33, 83.90)
mean (sd)	54.19 ± 39.92
maximum	141.70
residential	
minimum	5.66
median (IQR)	9.44 (7.85, 11.21)
mean (sd)	9.67 ± 2.28
maximum	14.95
mask_policy	
minimum	0.00
median (IQR)	1.00 (1.00, 1.00)
mean (sd)	0.82 ± 0.39
maximum	1.00
mask_prob	
minimum	0.56
median (IQR)	0.77 (0.71, 0.84)
mean (sd)	0.78 ± 0.10
maximum	0.93

As can be seen, 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 of the various mobility scores, we notice that `workplaces` is the inverse of the `residential` mobility variable. And we see in Figure 2 that they have almost perfect inverse correlation, having a linear correlation coefficient of -0.93. Therefore we would only use one out of these two.

In Figure 3, plotting the conditional expected value of `cases_per_100k` against `residential` and from the data shows a reasonably linear relationship, hence we decide to use it as the representative of the mobility scores. Other scores also show nice linear relationships. We will use them in the most inclusive model.

The Linear Models

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)
```

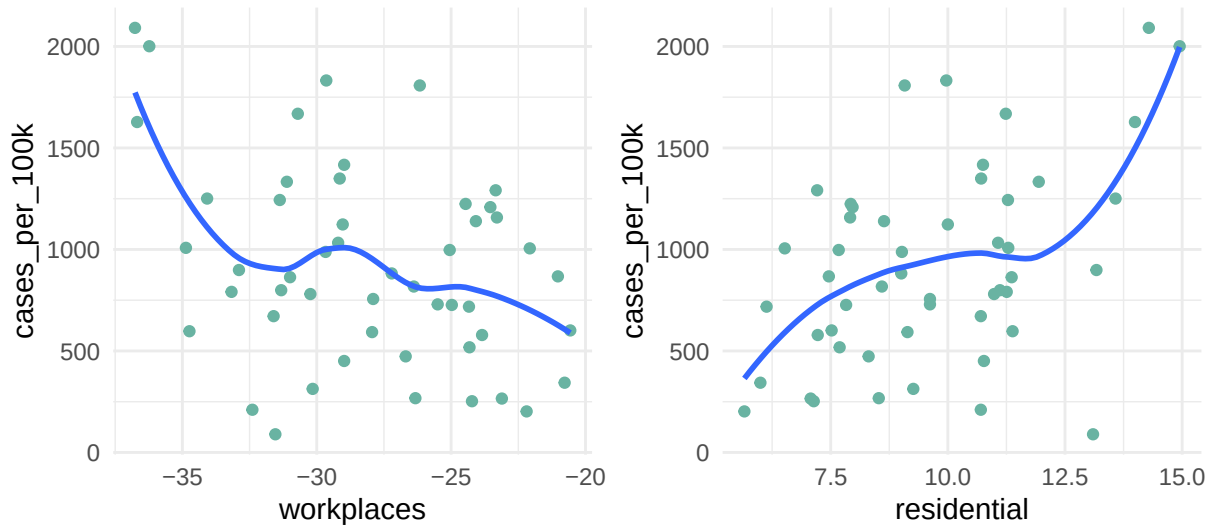

Figure 2: Scatter Plots for Cases per 100k and selected mobility variables

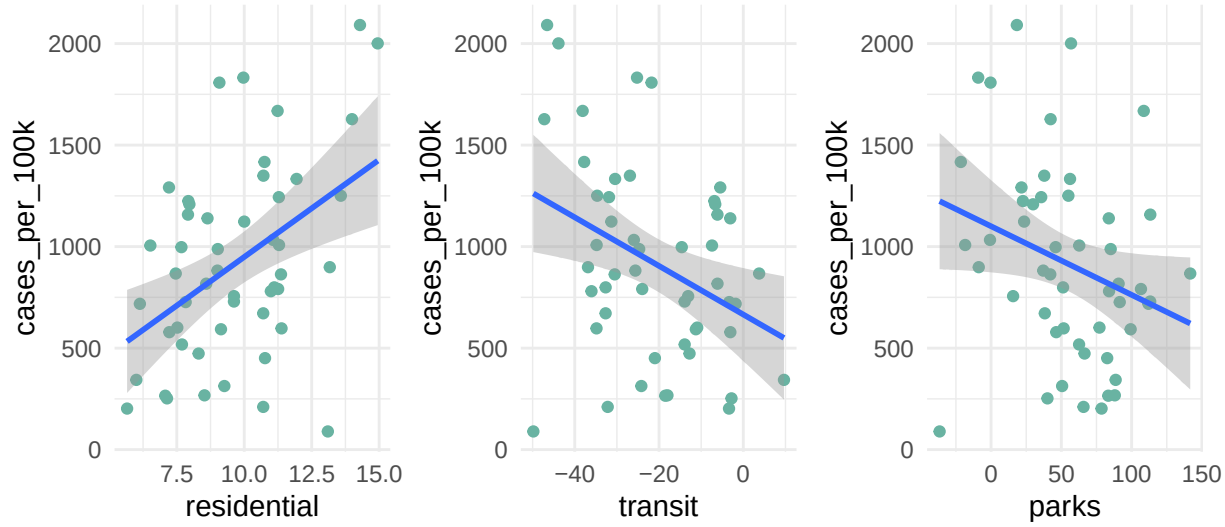

Figure 3: Scatter Plots for Cases per 100k and selected mobility variables (regression line and standard errors)

```
##
## Residuals:
##      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
```

Here we have a significant p-value withe residential mobility score. So it seems that one point increase in the residential mobility score leads to, on average, almost 96 more Covid cases.

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.

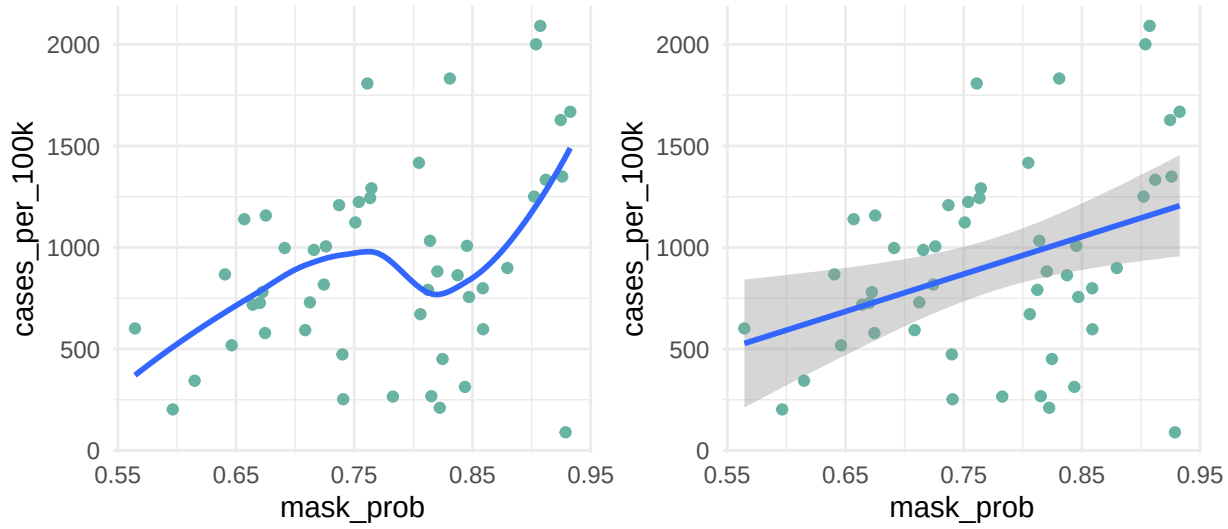

Figure 4: Scatter Plots for Cases per 100k and mask wearing probability

So, our second model contains the residential mobility score, the face mask wearing probability, as well as the control variable of whether the mask policy has been in place.

$$cases\_per\_100k\_residents = \beta_0 \tag{1}$$
$$+ \beta_1 residential\_mobility\_score \tag{2}$$
$$+ \beta_2 mask\_wearing\_probability \tag{3}$$
$$+ \beta_3 mask\_policy\_in\_effect \tag{4}$$

```
##
## Call:
## lm(formula = cases_per_100k ~ residential + mask_policy + mask_prob,
##     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_policy  -117.45     190.97  -0.615   0.5416
```

```
## mask_prob     234.59    1160.85   0.202   0.8407
## ---
## 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
```

The interesting thing here is that it shows, the probability of wearing a mask (`mask_prob`) seems to cause more cases, however the mask policy being in place cases fewer cases.

However neither variables are statistically significant, which give us more confidence that the mobility score is still the most important.

Model Three

Our secondary question was if the other mobility variables could have a significant impact on the outcome variable. As previously mentioned, the workplace mobility score is too close to perfect linearity with residential, so we'll leave it out, but we can run the model against all other mobility scores.

$$cases\_per\_100k\_residents = \beta_0 \tag{5}$$
$$+ \beta_1 residential\_mobility\_score \tag{6}$$
$$+ \beta_2 mask\_wearing\_probability \tag{7}$$
$$+ \beta_3 mask\_policy\_in\_effect \tag{8}$$
$$+ \beta_4 parks\_mobility\_score \tag{9}$$
$$+ \beta_5 grocery\_mobility\_score \tag{10}$$
$$+ \beta_6 transit\_mobility\_score \tag{11}$$

```
##
## Call:
## lm(formula = cases_per_100k ~ residential + mask_policy + mask_prob +
##     parks + grocery + transit, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -967.28 -336.03  -57.95  294.34  918.84
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -986.617    870.485  -1.133  0.26332
## residential  195.329     71.840   2.719  0.00941 **
## mask_policy  -47.341    191.031  -0.248  0.80545
## mask_prob    494.244   1192.953   0.414  0.68071
## parks         -3.513      1.929  -1.821  0.07557 .
## grocery       46.291     23.749   1.949  0.05782 .
## transit        5.218     11.058   0.472  0.63943
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 431.1 on 43 degrees of freedom
## Multiple R-squared:  0.3067, Adjusted R-squared:  0.2099
## F-statistic:  3.17 on 6 and 43 DF,  p-value: 0.01155
```

Perhaps somewhat surprisingly, all the prior variables retained their previous directions under earlier models. Residential mobility remains very significant. At the same time, higher mobility in grocery shopping and public transit seam to cause Covid cases to increase, whereas going to parks seem to cause fewer cases, although none of these variables are statistically significant to prove very useful.

The Regression Table

The models above are summarized in the following regression table.

Table 2: Comparisoin of Regression Models

	Dependent variable:		
	cases_per_100k		
	(1)	(2)	(3)
residential	95.801***	96.881**	195.329***
	(27.466)	(47.615)	(71.840)
mask_policy		−117.449	−47.341
		(190.972)	(191.031)
mask_prob		234.585	494.244
		(1,160.854)	(1,192.953)
parks			−3.513*
			(1.929)
grocery			46.291*
			(23.749)
transit			5.218
			(11.058)
Constant	−9.003	−105.201	−986.617
	(272.637)	(583.424)	(870.485)
Observations	50	50	50
R^2	0.202	0.209	0.307
Adjusted R^2	0.186	0.157	0.210
Residual Std. Error	437.718 (df = 48)	445.297 (df = 46)	431.130 (df = 43)
F Statistic	12.166*** (df = 1; 48)	4.045** (df = 3; 46)	3.170** (df = 6; 43)
Note:			*p<0.1; **p<0.05; ***p<0.01

Model Limitations

Time Series

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 Given that each data point is aggregated from a U.S. state, the number of cases and mobility are largely independent. However the mask wearing policy as well as the populations willingness to carry them out do tend to cluster somewhat based on geography and political leanings, so they can be argued to be less independent of each other.

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

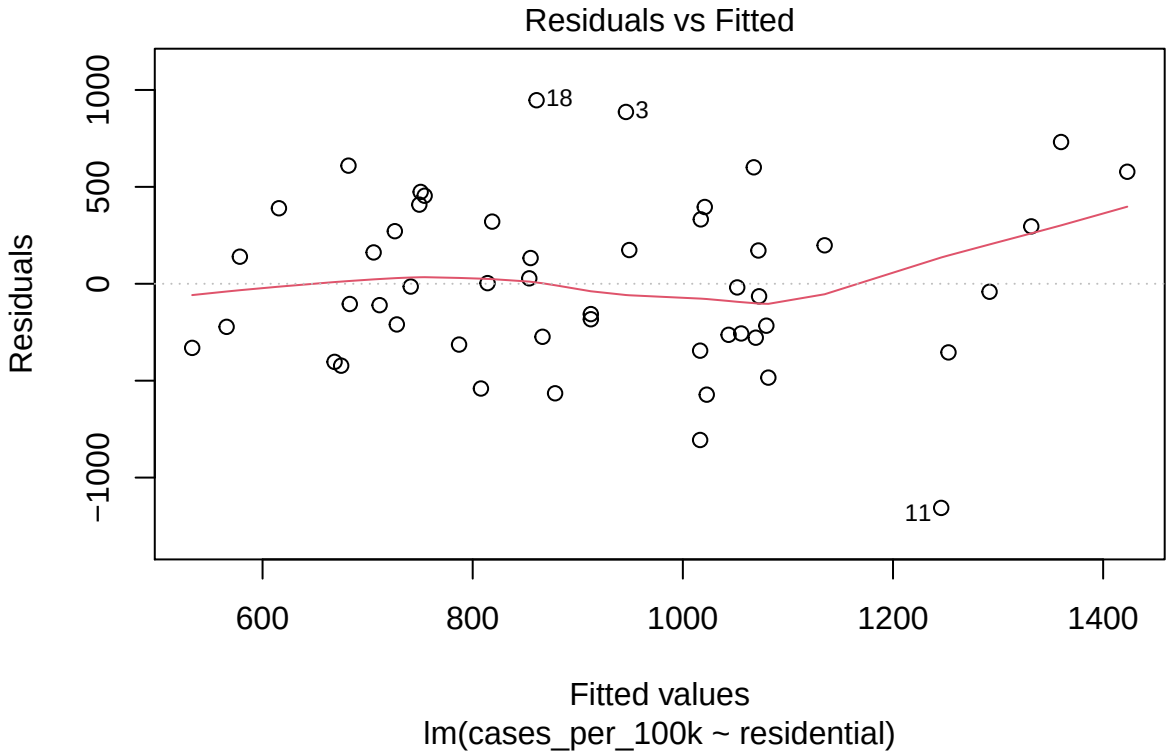

Figure 5: Residuals vs Fitted plot for Model One

3. No Perfect Collinearity

```
## (Intercept) residential mask_policy   mask_prob       parks     grocery
## -986.616791  195.328970  -47.340986  494.244101   -3.513109   46.291300
##     transit
##    5.217702
```

For even the most inclusive model, none of the coefficients of the variables were dropped, meaning that there is no perfect colinearity within independent variables.

However we did notice some high colinearity between some mobility scores and decided to use a subset, notably between workplaces and residential scores.

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.

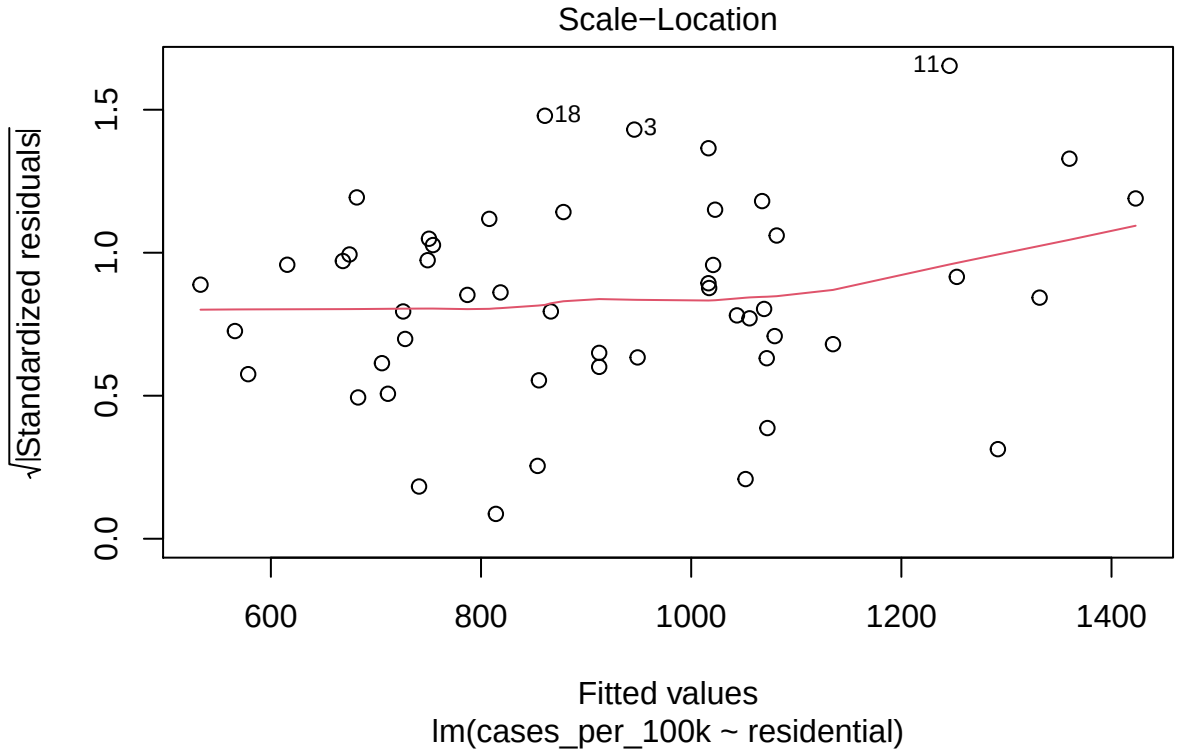

Figure 6: Scale-Location plot for Model One

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

In general we can say that all the 5 CLM Assumptions are met to a reasonable degree.

Omitted Variables

Not following mandates

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 such as 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

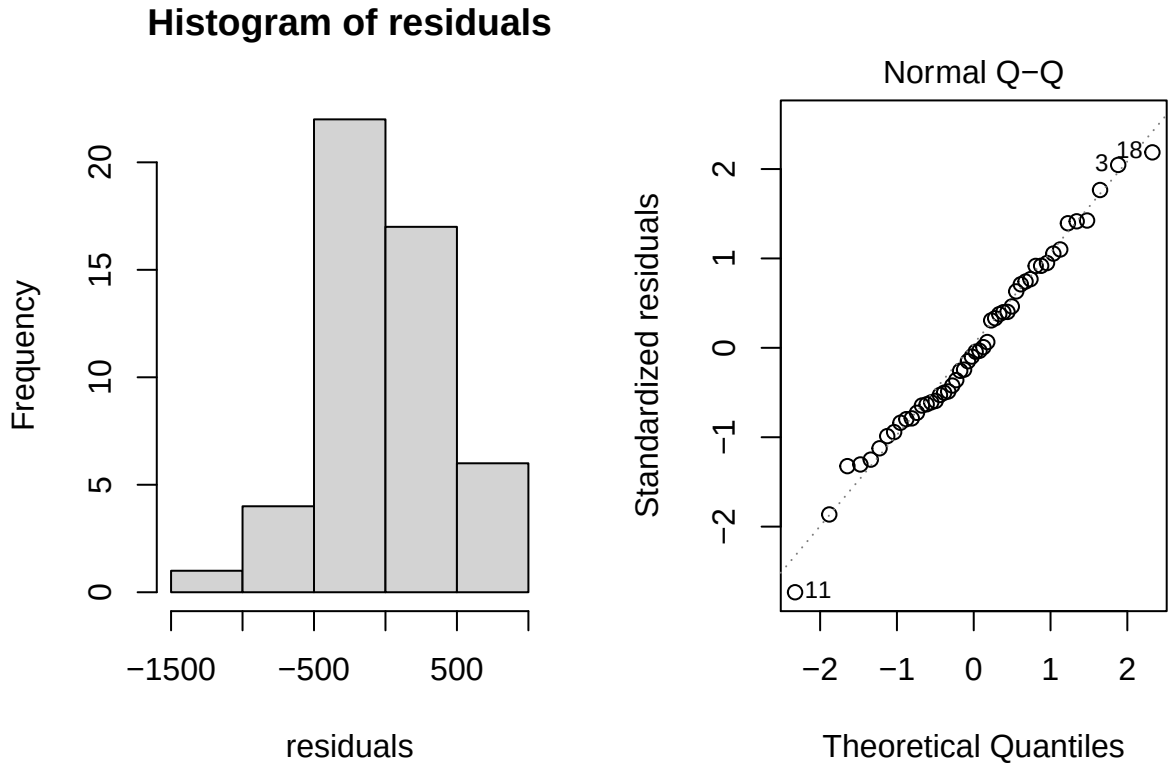

Figure 7: Histogram of residuals for model one and qq-plot for model one

variable positively impacts outcome and positively related with measured variable, thus moving away from zero.

Population and Population Density

Even as we captured the population in the outcome variable by considering cases per 100k residents, we're not taking into account other non-linear relationship between population density and the spread of Covid. For example in a high density metropolitan area, the spread of Covid might be super-linearly prominent.

Conclusion

Based on the very limited data points that we have and our analysis, it does seem that people moving around is indeed a significant (statistically and practically) factor in the spread of Covid-19. Higher mobility scores in enclosed areas does seem to have a causal effect on higher number of cases per resident in a State.

We've found this to be consistent even with a number of control variables such as the mask policy and the proportion of population actually wearing masks.

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