Lab 2: Regression to Study the Spread of Covid-19

Charis Chan, Joyce Ching, Inderpal Kaur

10 December, 2020

1. Introduction

Covid-19 has drastically changed our lives in terms of our plans for the future, the workplace environment, meeting new people, and even simply where we go day to day. Because Covid-19 is an infectious disease easily transmittable through close contact with others, it is important to enact policies to encourage people to stay at home as much as they can to protect both themselves and others around them.

But the longer the pandemic continues, pandemic fatigue becomes an increasingly important issue. Pandemic fatigue as defined by the Word Health Organization is "feeling demotivated about following recommended behaviours to protect themselves and others from the virus." This often results in being less vigilant about safety regulations like stay at home orders. We wanted to know if pandemic fatigue is reflected in the data. We ourselves feel pandemic fatigue but do we see this phenomenon play out for the rest of the country? Do people continue to abide by stay at home policies the longer they are in effect?

Research Question: In states that enacted stay at home orders, did people follow the order less strictly over time? Is there a relationship between the order's duration and the change in people's mobility from the start of the order to the end?

Data Sources: The data that we used to investigate our question contains state-level information about demographics, mobility, Covid-19 case and death statistics, and policy information about stay at home orders and other closures/mandates.

Mobility: The mobility information comes from Google's COVID-19 Community Mobility Report which shows trends in visits and lengths of stay in six location categories (residential, grocery and pharmacy, parks, workplaces, transit stations, and retail and recreation). The data measures the daily percent change in mobility in comparison to a baseline value for that corresponding day of the week. This baseline is set as the median value for that day of week from the pre-C OVID period of Jan. 3 - Feb. 6, 2020. To measure the "strictness" with which people follow stay at home orders, we used this data to see if mobility increased (larger percent change compared to baseline) or decreased (smaller percent change compared to baseline) over the course of the policy.

State-Level Policy: The COVID-19 US State Policy Database (CUSP) by the Boston University School of Public Health documents the dates of health and social policies enacted by individual states. These policies include stay at home order start and end dates as well as the start and end dates for other closures/reopenings and health mandates. To find the duration of these policies, we subtracted the policy start date from the end date to calculate the number of days the order had been in effect.

State-Level Characteristics: We obtained information about state-level demographics from a Covid-19 specific dataset that had combined sources from the Kaiser Family Foundation and CUSP. Variables included state governor, population density, percent of population at risk for serious illness, median annual household income, age demographics, etc.

Covid-19: We downloaded data from the Center for Disease Control and Prevention (CDC) about United States COVID-19 Cases and Deaths by State over Time. This data provides daily Covid-19 case numbers by state, and we transformed it to include the number of daily new cases and deaths by state per 100,000 people.

Data Limitations:

Mobility: There may be some bias in terms of the sampling procedure for the mobility data because it only records the movements of people with smartphones and the appropriate privacy settings. In addition, Google notes that there is variation in the categorization of "places" between locations (ex. categorization of residential areas or grocery and pharmacy may vary based on location). The dataset already excludes data where the information was not statistically significant or there was not enough privacy to protect people's information. Some other limitations of the residential mobility information that we have access to includes the possibility for time-dependent trends to influence residential mobility. For example, because this data is compared to a January baseline, there may be a natural trend for people to stay at home more when the weather is colder and go outside more often when the weather becomes warmer. In addition, our residential mobility data shows when people are visiting residential areas/staying in residential areas for longer, however increases in residential mobility may reflect both people staying at home more often or people visiting other people's houses more often (which may not be according to stay at home policy).

State-Level Policy and Characteristics: There may be variation in state policies because of different levels of strictness and enforcement at the local level. Additionally, different counties within the same state may implement different supplementary policies based on the current local Covid case numbers. These variations indicate that stay at home orders may vary by state and county, which might impact our measurement of the results.

Covid-19: The CDC notes that there may be differences in how or how often states report data about Covid cases and deaths, which could have an impact on analysis of daily trends.

2. A Model Building Process

To answer our research question, we want to use a descriptive linear model to examine the relationship between the duration of a stay at home order and changes in public response to the policy.

To track when stay at home orders were in effect, we used information from the Covid-19 US State Policy Database about the start dates and end dates for stay at home orders in every US state. This data represents statewide orders to stay at home, but there may be discrepancies between counties in the same state based on the strictness of county-specific policies and enforcement, and we note this as a limitation of our data. We excluded states that either did not enact a stay at home order or enacted a policy that did not require people to stay at home by specifically restricting the movement of the general public (i.e. Arkansas, Connecticut, Iowa, Kentucky, Nebraska, North Dakota, Oklahoma, South Dakota, Texas, Utah, Wyoming). Additionally, we chose to exclude states that had not yet ended their stay at home orders (i.e. California and New Mexico) because we compared changes in mobility from the order start date to the end date.

We created an order length variable that measures the number of days that the stay at home order was in place (i.e. the number of days from the order start date to the end). We investigated changes in mobility based on this duration variable rather than the calendar start and end dates. As a result, different states with the same order duration may have implemented their orders at different times. We chose to use this method because we wanted to understand the relationship between the length of the order and mobility, but we note that depending on when states chose to enact their stay at home orders, factors such as current national sentiment, weather, and other time-related variables may have played a role in people's mobility patterns.

To understand how patterns in people's mobility change over time, we used data from Google's Covid-19 Community Mobility Report to track the percent change from baseline in visits to specific locations. We chose to focus on the percent change in residential mobility as our measure for how closely people were following stay at home orders under the assumption that staying at home would be reflected in the data as an increase in stays in residential areas. We also limited our focus to one mobility measure because changes in residential mobility were fairly correlated with the other five measures. (See EDA for a plot that displays the correlations between each mobility measure.) We selected state-level mobility data instead of the more granular county-level mobility data because our policy data was also given at the state level.

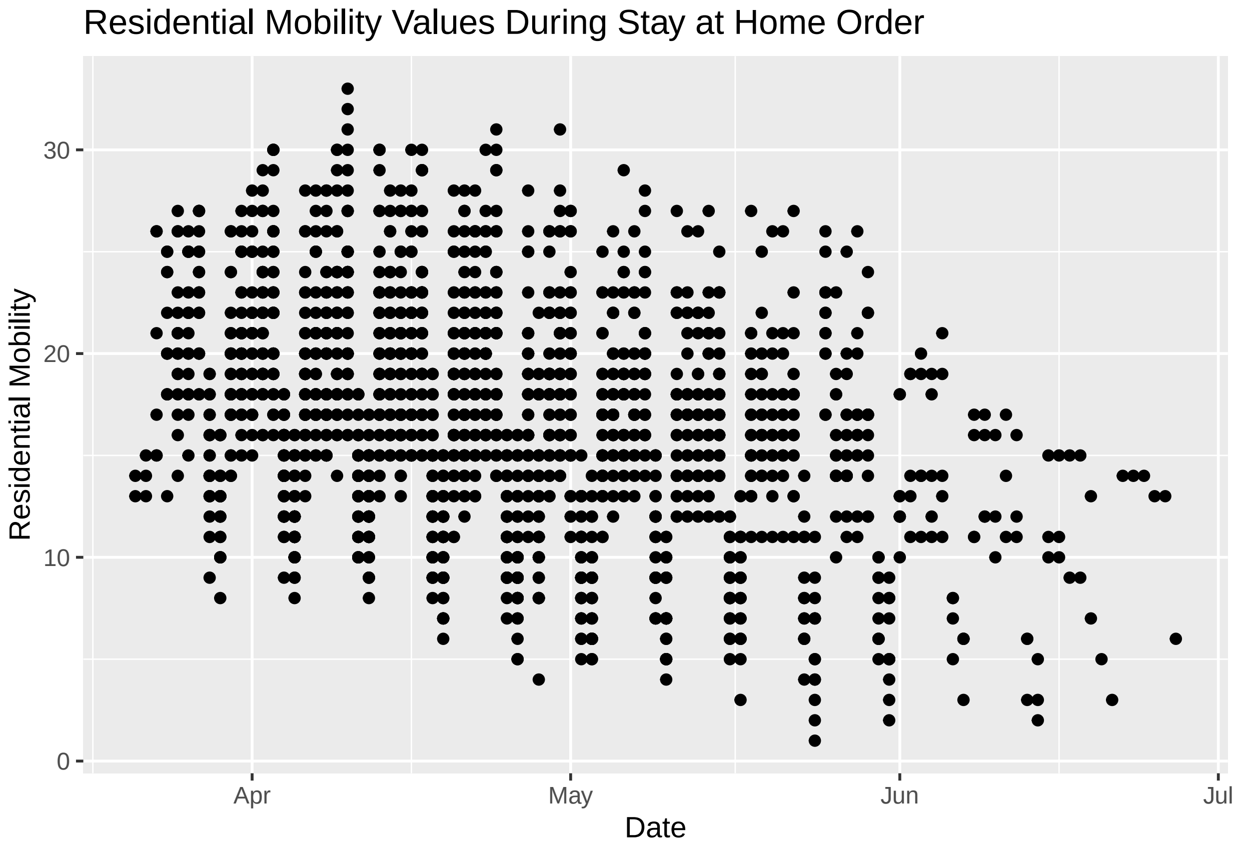
We created a residential mobility difference variable which took the difference between the average residential mobility during the first week the order was in effect and the average during the last week it was in effect. We chose to take averages of the first and last week because the mobility data uses the corresponding day of the week to set the baseline and we did not want our analysis to be influenced by the day of week that an order happened to start or end. For example, if the average for the first week was 15% and the average for the last week was 10%, our measure would tell us that the average change compared to baseline in residential mobility dropped by 5 percentage points over the course of the stay at home order. Positive residential mobility differences indicate that at the end of the order, people did not stay at home as much as they did at the beginning (the average change compared to baseline decreased over time). Negative differences indicate the opposite.

Using the residential mobility data for each state, our modeling goal was to describe the relationship between changes in mobility and the duration of stay at home orders to see if people began following the order less strictly over time, and if so, how quickly. If this pattern exists, we expect that it would be reflected in the data by a positive trend between order duration and residential mobility difference. (i.e. the trend shows that states with longer stay at home orders have larger residential mobility differences).

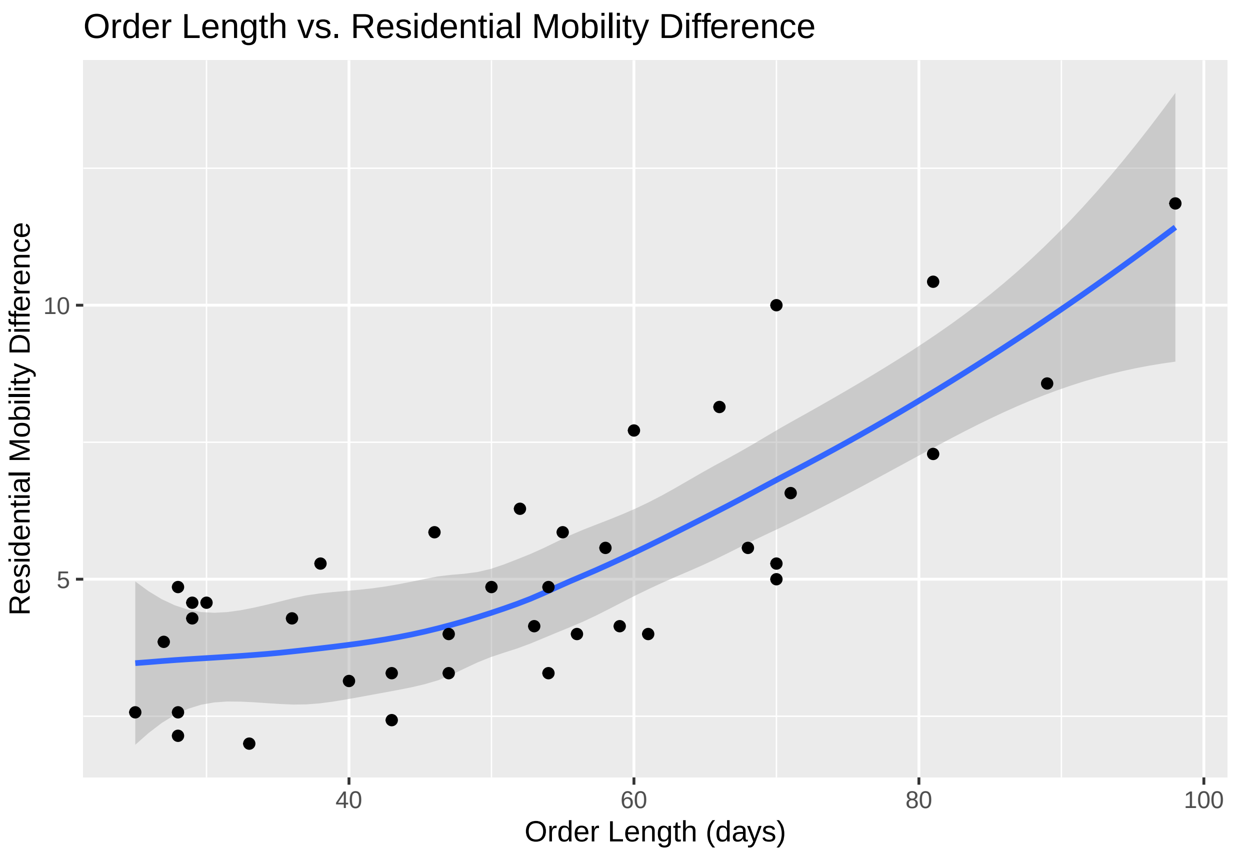
From the data we have, some covariates that could help us describe the relationship between mobility differences and order duration include state-level information about the number of known Covid-19 cases and deaths at the time. We believe that the choices people make about following the order could be influenced by the local case numbers (i.e. people in areas with proportionally higher cases and deaths may be more aware of the spread of the virus and choose to stay home for that reason).

Another set of covariates includes state-level policies for reopening, population demographics, and political leaning. These characteristics may also help us describe the relationship between order duration and mobility difference.

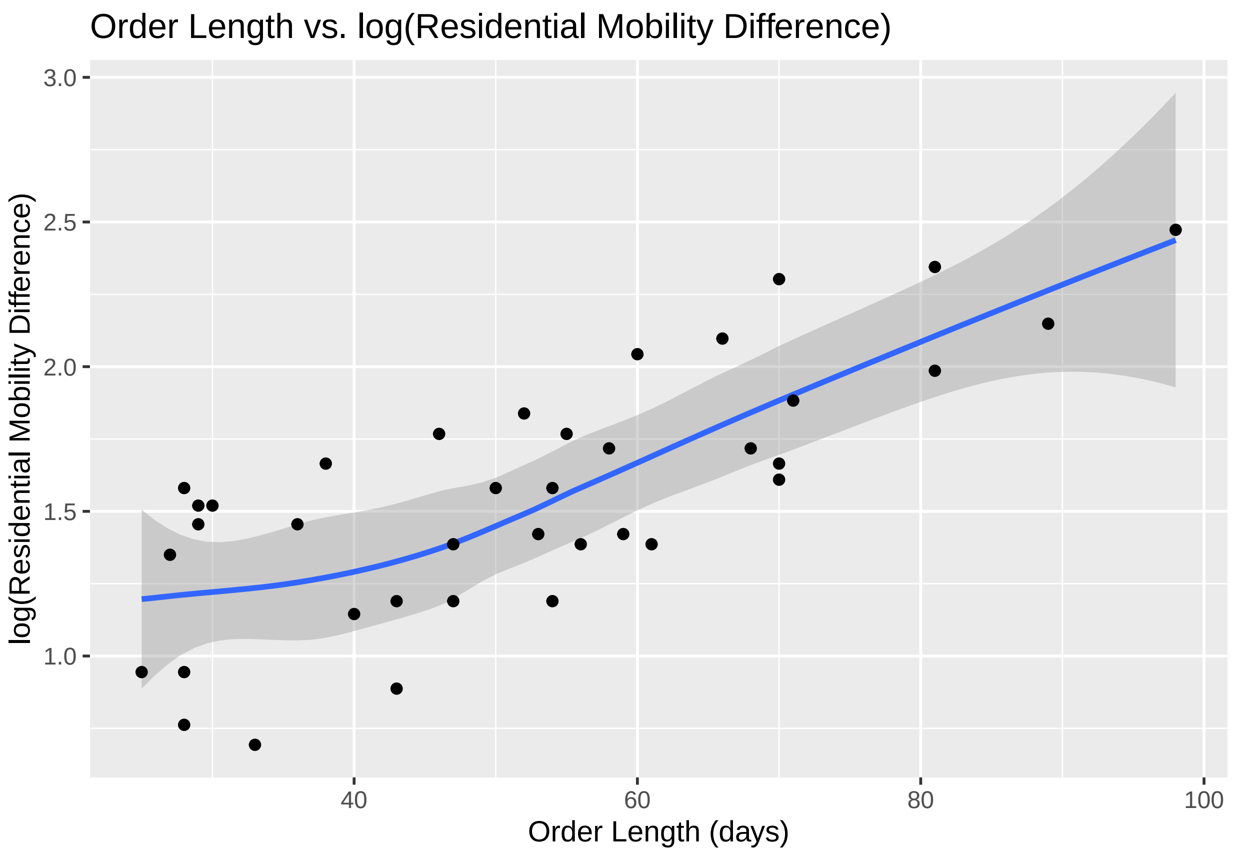
To get a sense for some of the overall trends in residential mobility, we plotted the mobility data over time for each state that implemented and enforced a stay at home order with a set start and end date. We noticed that the percent change from baseline in residential mobility seemed to decrease over time, indicating that people may have been staying at home less at the end of the order than they were at the beginning. The plot below also reveals a seasonality reflected in the days of the week, and it seems that the weekends have lower rates of residential mobility which implies that people are staying home less on the weekends. It may look like there is less data provided for the weekends than the weekdays, but the residential mobility varies less across the states for the weekends. We jittered the plot but it didn't look much different so we took it out but the plot accounts for all the state's weekend data.



To further investigate and model the relationships between time, mobility, and the order, we took the average residential mobility difference from the first and last week of the stay at home order and plotted it with the order length. At a glance, it seems like longer stay at home orders see a larger mobility difference (there's a bigger difference between how people behaved at the beginning of the order vs. the end). Another thing to note is that there are fewer data points for the longer orders than shorter so the longer orders may have more of a influence on how the model will come out than the shorter ones, this is part of the reason why we took out California and New Mexico from this model (both of which have not ended their stay at home order).



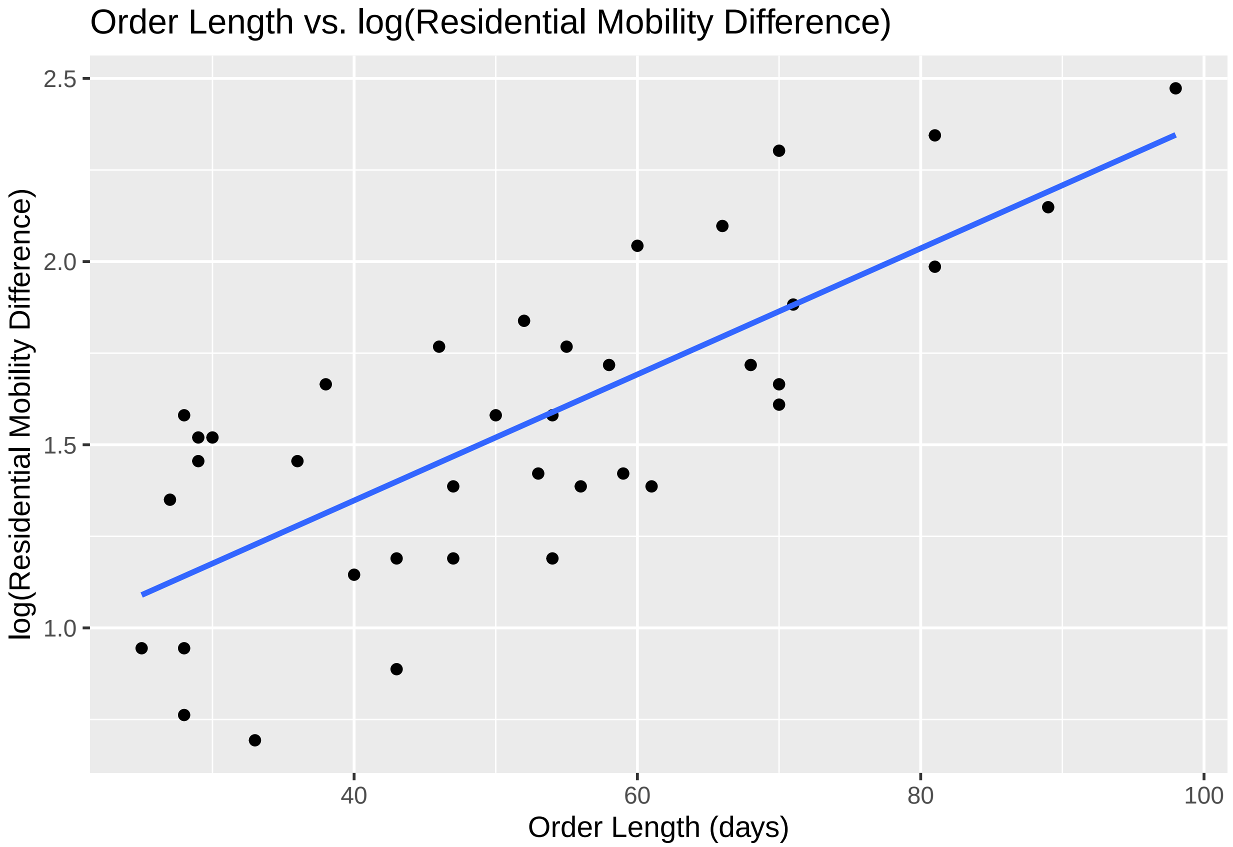
We noticed that the data has some clusters around the lower mobility difference values, so we decided to log the residential mobility difference to transform the data to have a more linear relationship for modeling.



Baseline Model

For our baseline model, we only included information about the two key variables (order length and log residential mobility difference) to describe the relationship between them. The equation is as follows:

$$log(residential\ diff) = \beta_0 + (order\ length)\beta_1$$

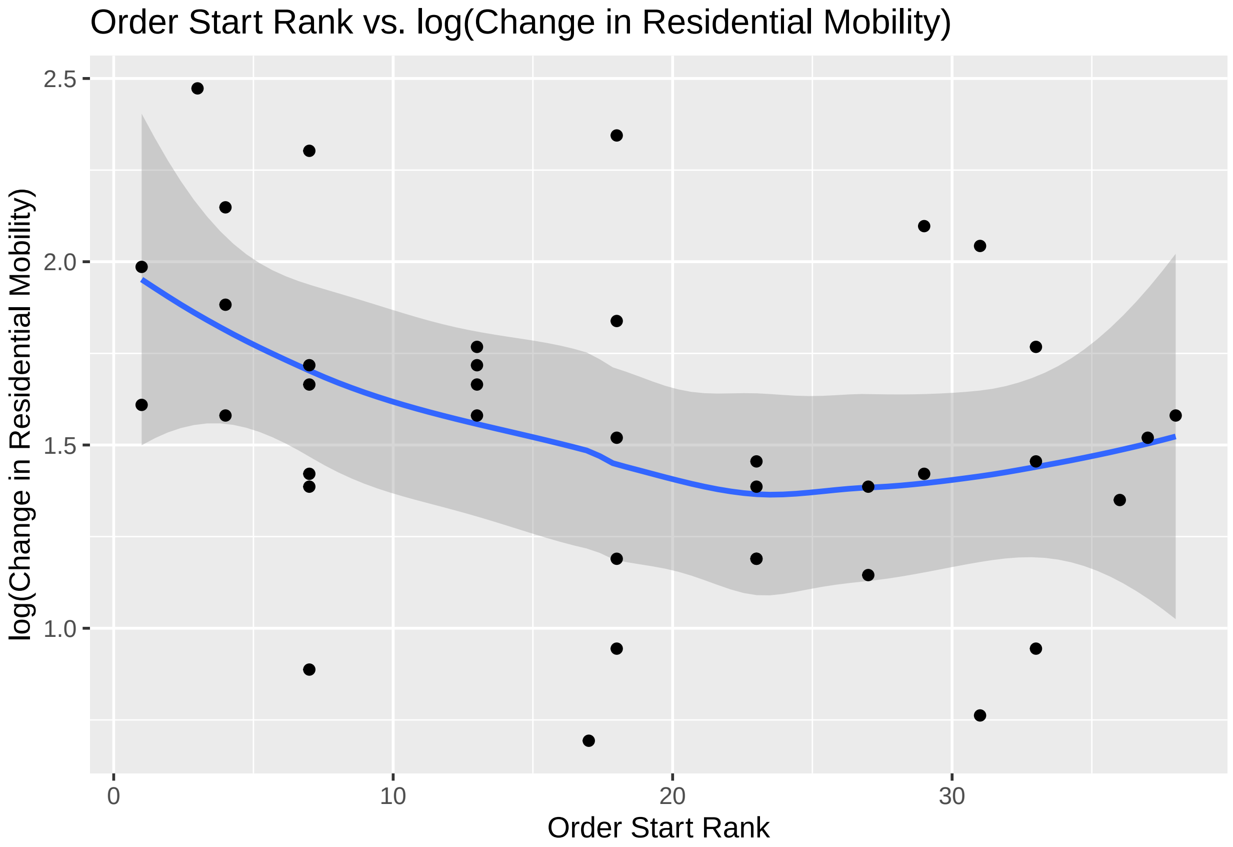


A visual inspection of the fitted values shows that our baseline model fits the data reasonably well and captures the general trend between the two variables.

Model 1

For our first improved model, we added a feature that indicates the start rank of each state's stay at home order. To calculate order start rank, we sorted the states by the start date of their order so that the state that instituted their order first would be rank 1, and so on. In the case of ties (i.e. states that began their stay at home orders on the same day), states were given the same rank. Because our measure of order length does not distinguish between states that implemented their stay at home orders for the same number of days but at different times, we decided to include order start rank to account for differences between states that did not implement orders concurrently. This variable allows us to incorporate information into our model about how the novelty of the order at the national level may have influenced public compliance.

The plot below suggests that states that have a higher order start rank may see smaller changes in residential mobility over the course of the stay at home order.



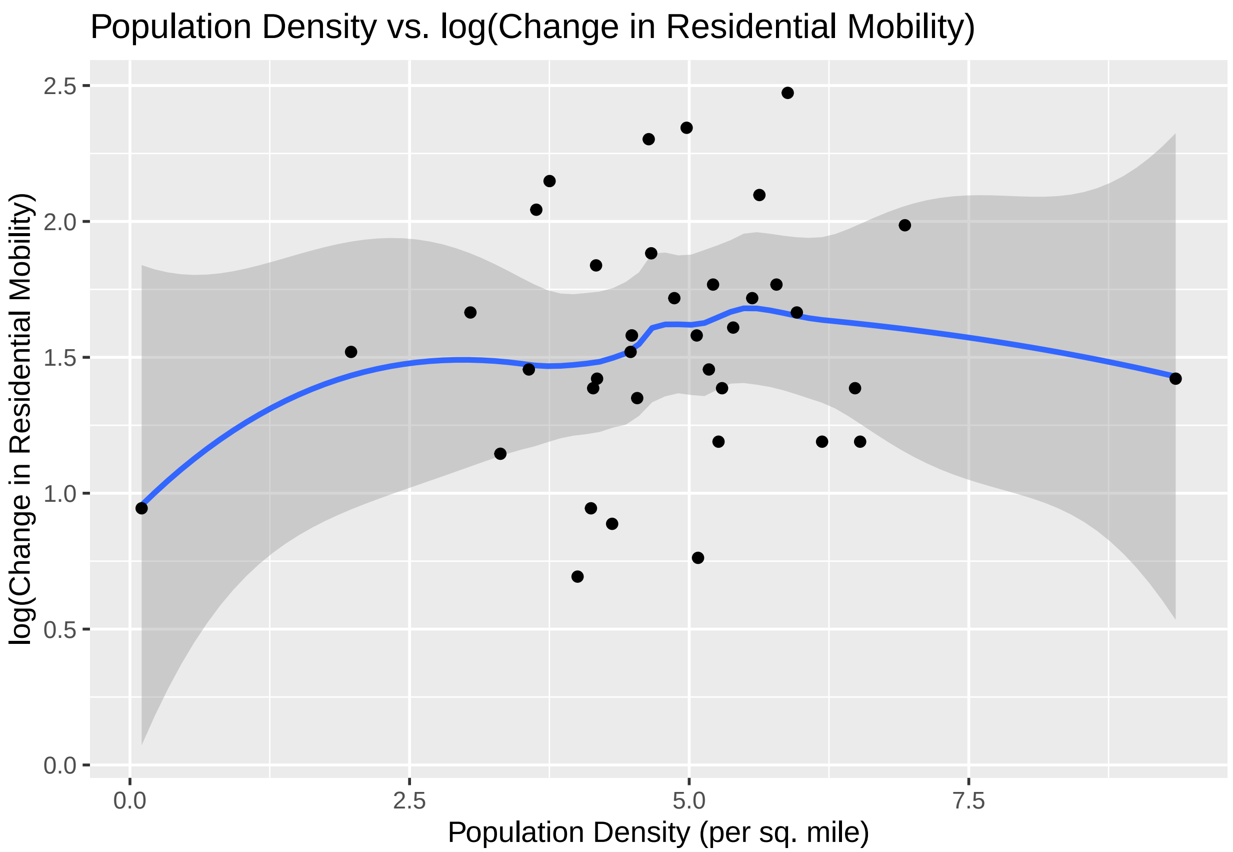
The equation for Model 1 is as follows:

$$log(residential\ diff) = \beta_0 + (order\ length)\beta_1 + (order\ start\ rank)\beta_2$$

Model 2

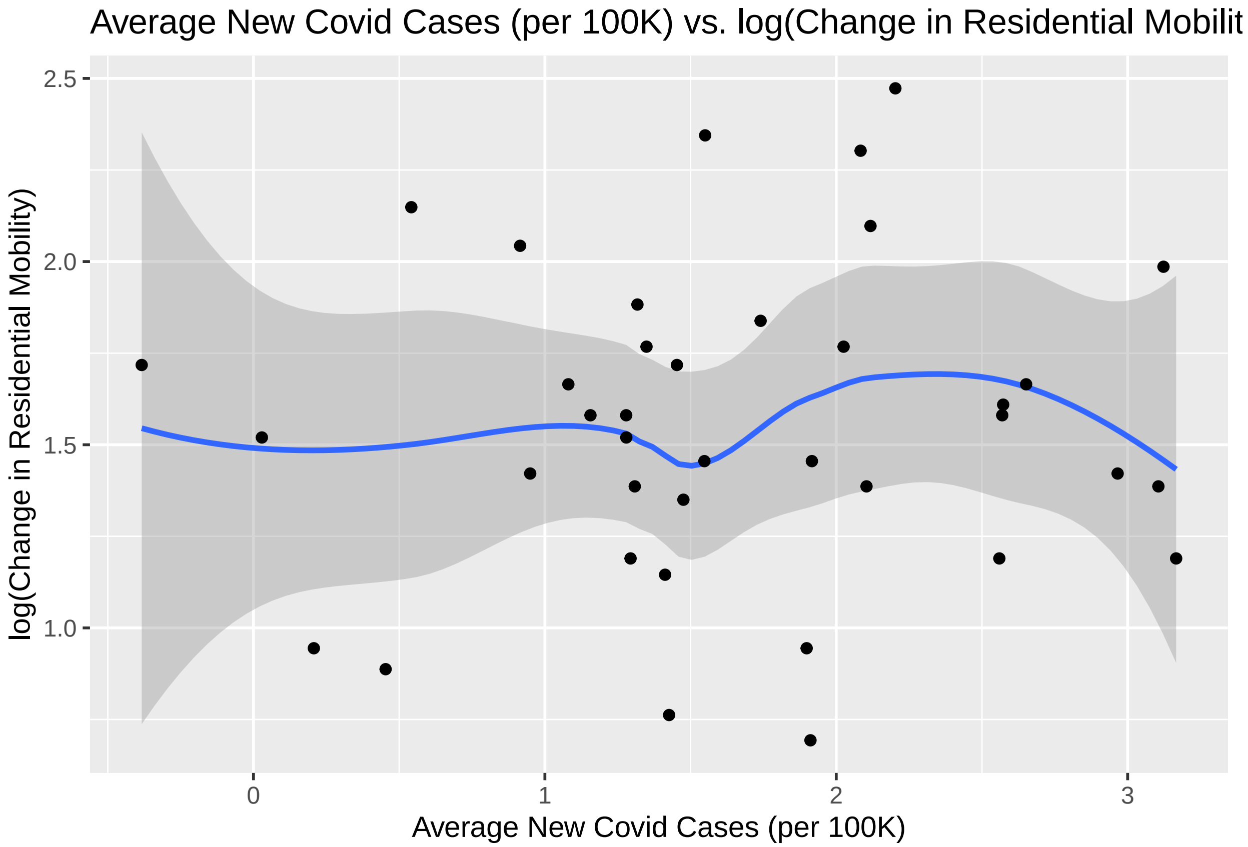
For our second improved model, we chose to add covariates that we thought might help us describe the relationship between order length and residential mobility difference and explain some additional variance in the data. Based on our exploratory data analysis, we selected the following three variables from the data:

Population density per square mile: To account for differences between states that have largely urban or largely rural areas, we included a variable that represents the total population density of the state. We hypothesized that population density could impact the types of jobs in the state (blue-collar vs. white-collar, essential vs. non-essential) as well as how much people would have to change their everyday habits to isolate themselves from other people. These differences could affect people's mobility patterns and how they respond to stay at home orders. Because this variable had a long right tail, we log transformed it before including it in the model.

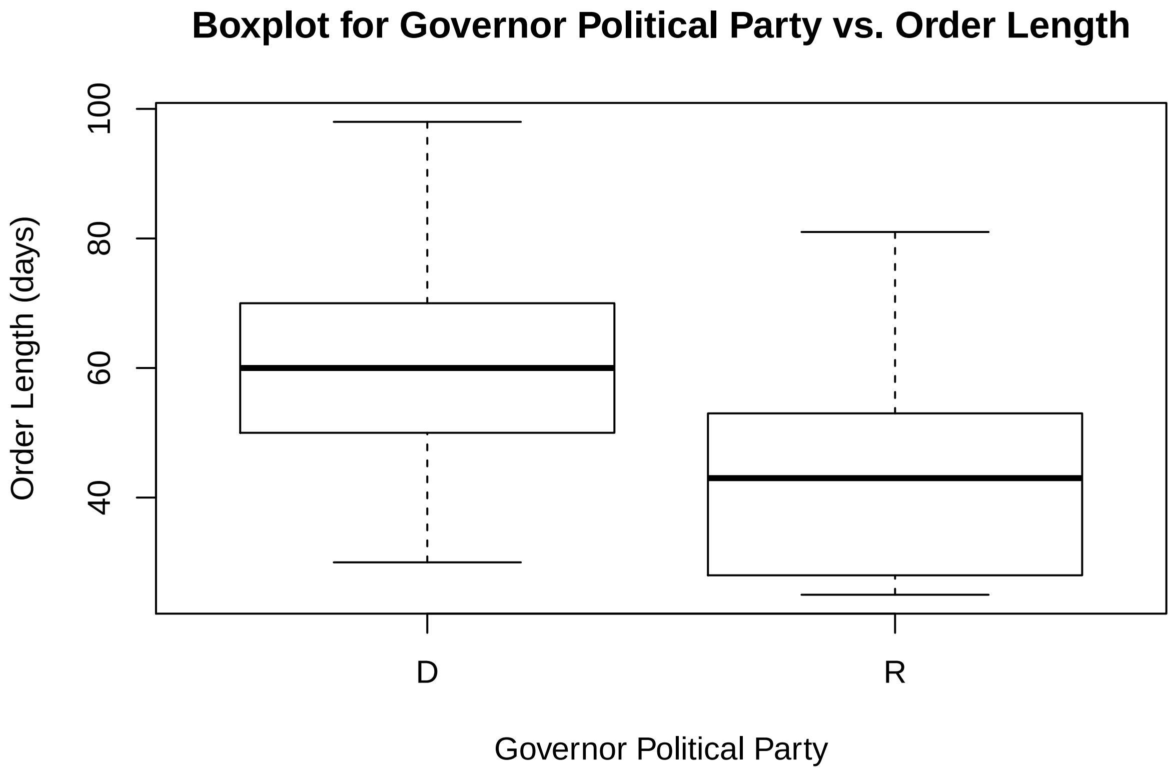


Average new cases per 100K: To calculate the average new Covid-19 case numbers per 100K, we took the average of the daily new case numbers that were reported during the state's stay at home order and normalized those averages by state population. (As we note in our Data Limitations section above, this measurement is slightly limited by differences in reporting and testing across states.)

We included this covariate under the assumption that the choices people make about following stay at home orders could be influenced by the local case numbers. For example, in our EDA we noticed that states with the highest average case numbers during the stay at home order tended to be a few outliers (ex. New York) because most orders were implemented between March and May. Therefore, we chose to include this variable to account for local differences in Covid conditions and their effect on order strictness as well as individual mobility. Like population density, we log transformed this variable because of a long right tail.



Governor political party: The issue of closures and health mandates has become somewhat of a political issue in the U.S. We included the political party of the state governor to investigate the possibility that politics played a role in the state's response to the pandemic and the public's acceptance. From our EDA, we noted that states with Republican governors (17 of 38 states in our dataset) tended to have much shorter stay at home policies than states with Democratic governors.



The equation for Model 2 is as follows:

$$log(residential\ diff) = \beta_0 + (order\ length)\beta_1 + (order\ start\ rank)\beta_2 + log(population\ density)\beta_3 +$$
$$log(avg\ new\ cases\ per\ 100K)\beta_4 + (governor\ political\ party)\beta_5$$

3. Limitations of the Model

Below we list the CLM assumptions for which we made adjustments or noted any limitations. For a full list of all 5 CLM assumptions and the corresponding information/tests see the CLM Assumptions notebook.

1. IID Sampling

Reference population: US states that implemented stay at home orders.

Limitations:

a) The mobility data doesn't include the entire state population (nonrandom sample of users).

- Proximity in terms of social distance: Only users who have a mobile device and a Google Account are represented. Users also need to have opted-in to Location History for their Google Account.
- Proximity in terms of physical distance: People may not have data or live in places with connectivity issues. Privacy thresholds also may not be met if somewhere isn't busy enough to ensure anonymity.

b) States might be systematically different from one another in various ways. For example, states that are closer together geographically may be influenced by similar weather patterns or economic conditions, while states that are farther away geographically may not be as similar.

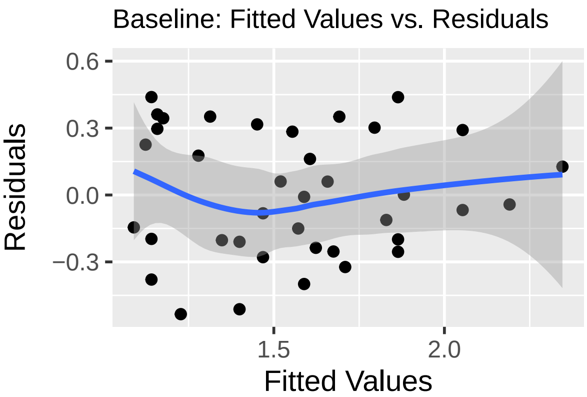
These limitations mean that our data is not guaranteed to be IID for the entire population. Given the granularity of the data we had available to us, we attempted to incorporate covariates in our models that might account some of the differences between states, but we acknowledge that this approach may not be sufficient.

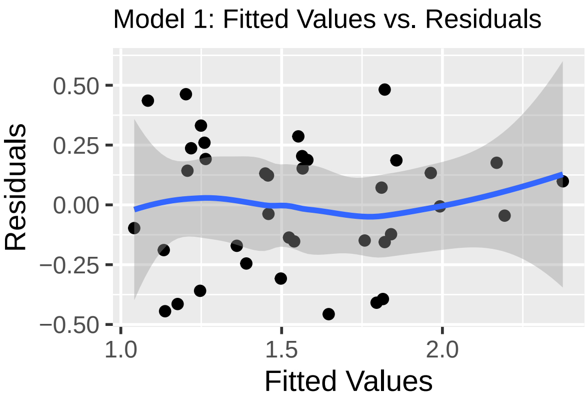
Statistical consequences: Having non-independent samples prevents us from being able to provide guarantees about the entirety of our reference population.

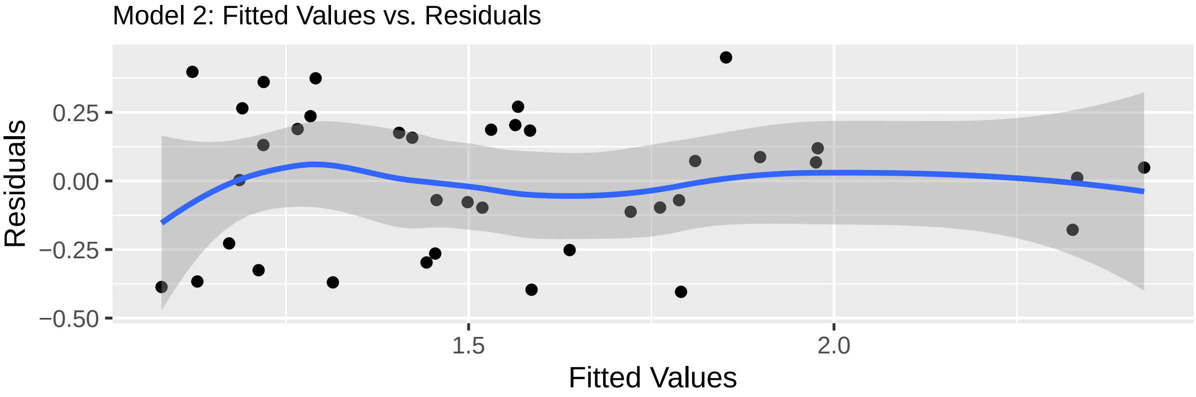
Mitigating the consequences: We need to narrow down our research question to only include individuals who fit the population of those who are sampled and to adjust our measures of uncertainty to reflect the clustered nature of the data generating process.

4. Homoskedastic Conditional Variance

Ocular test: A visual inspection of our graphs of the fitted values vs. the residuals shows that there is relatively little "fanning out" effect. The residuals are roughly arranged in a flat line at 0, which means that the ocular test does not show strong evidence of heteroskedasticity.







Breusch-Pagan test: Although the test says that we should not reject the null hypothesis that there is homoskedastic conditional variance at the 95% confidence level, the p-values are small enough that we decided to use robust standard errors in our final model to account for the observed variation.

```
##
##  studentized Breusch-Pagan test
##
## data:  baseline
## BP = 4.3904, df = 1, p-value = 0.03614
```

```
##
##  studentized Breusch-Pagan test
##
## data:  version_1
## BP = 3.4058, df = 2, p-value = 0.1822
```

```
##
##  studentized Breusch-Pagan test
##
## data:  version_2
## BP = 10.827, df = 5, p-value = 0.05492
```

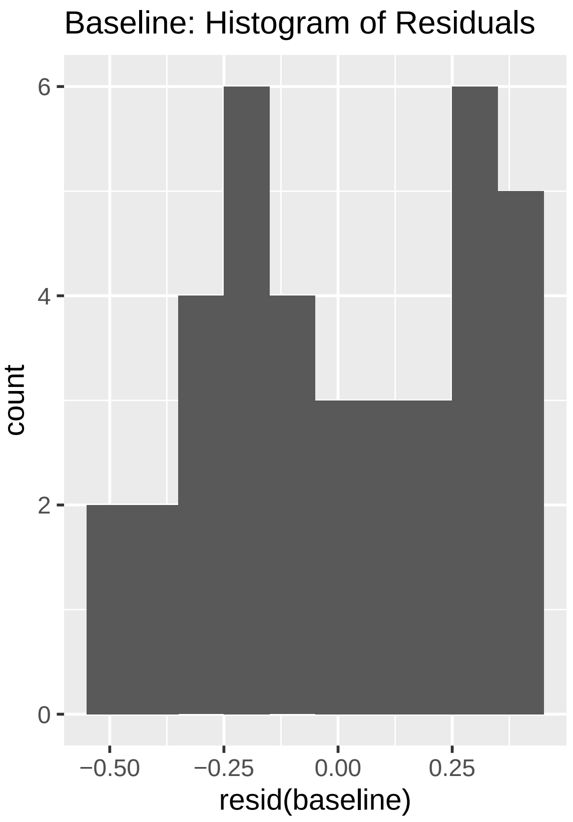
As we see in our regression table (below), the larger robust standard errors do not affect the statistical significance of our findings. That is, the relationship where a 1 day increase in order length is correlated with a 2.59% increase in mobility difference remains significant at the 99.99% confidence level.

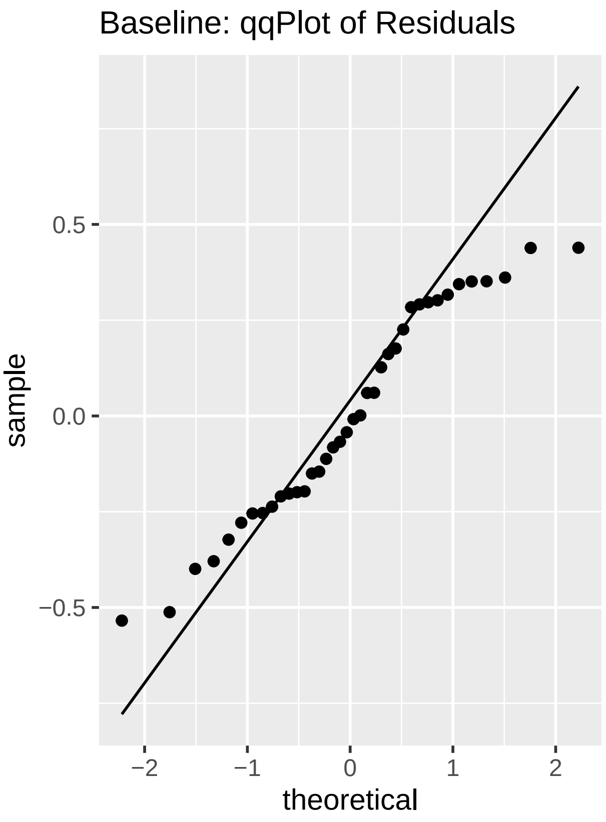
```
##
## t test of coefficients:
##
```

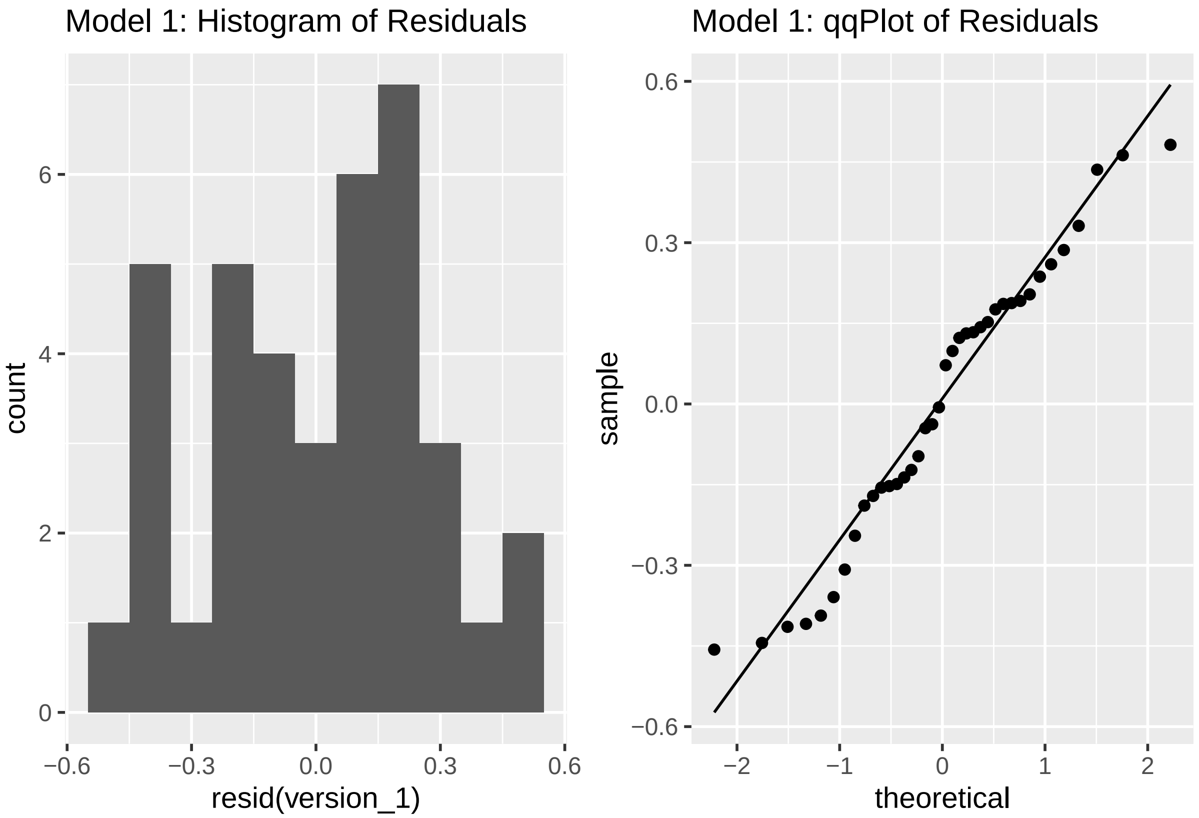
```
##                                              Estimate Std. Error t value
## (Intercept)                                 0.2518667  0.2642428  0.9532
## order_length                                0.0259331  0.0030336  8.5486
## order_start_rank                            0.0140486  0.0044092  3.1862
## log(`Population density per square miles`) -0.0817488  0.0445082 -1.8367
## log(avg_new_cases_per_100K)                 0.0316349  0.0621858  0.5087
## governor_political_partyR                   0.0853574  0.0901689  0.9466
##                                            Pr(>|t|)
## (Intercept)                                 0.34765
## order_length                               9.07e-10 ***
## order_start_rank                            0.00321 **
## log(`Population density per square miles`)  0.07556 .
## log(avg_new_cases_per_100K)                 0.61444
## governor_political_partyR                   0.35092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

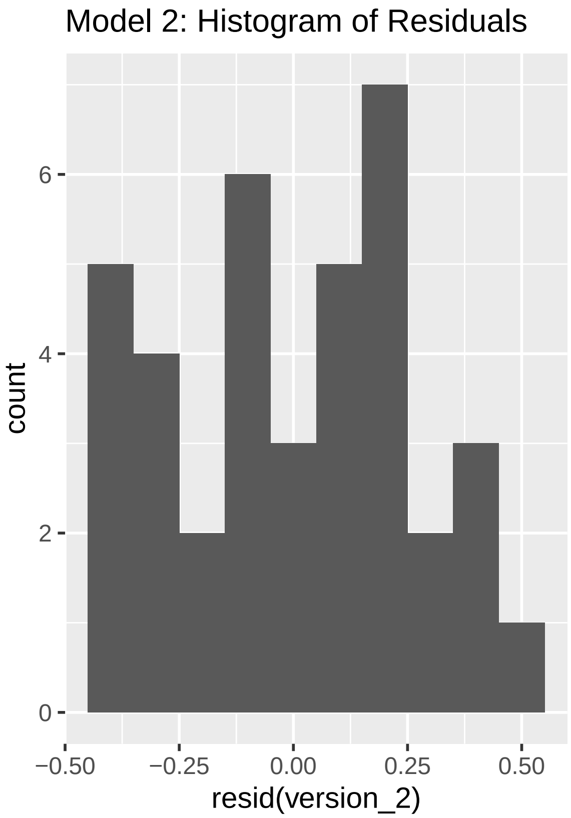
5. Normally Distributed Errors:

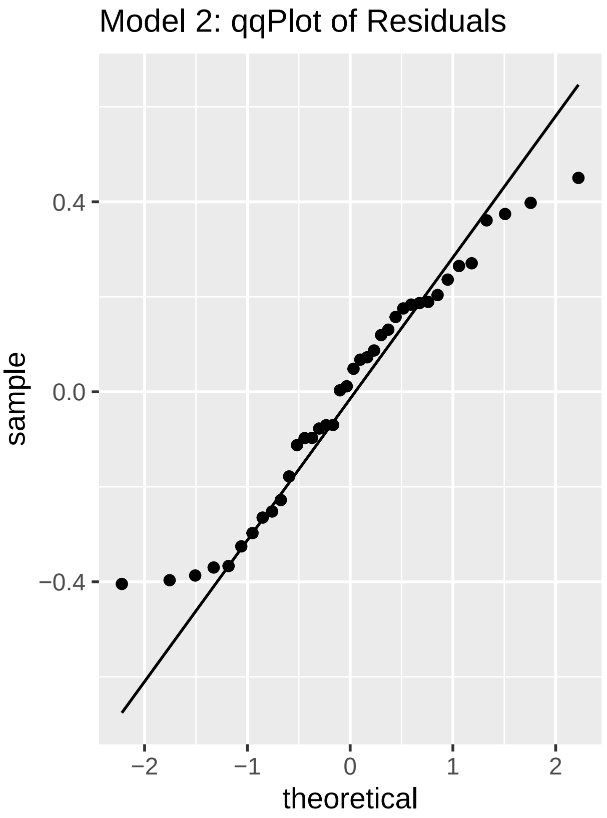
The histograms of the residuals and the qqplots show the errors to have some deviations from normality. This problem threatens the validity of our significance tests and confidence intervals. We tried to fix this problem with various logarithmic and polynomial variable transformations such as but were unsuccessful in finding a transformation that would give us approximately normally distributed errors.


Baseline: Histogram of Residuals


Baseline: qqPlot of Residuals


Model 1: Histogram of Residuals
count
0
2
4
6
−0.6
−0.3
0.0
0.3
0.6
resid(version_1)
Model 1: qqPlot of Residuals
sample
0.6
0.3
0.0
−0.3
−0.6
−2
−1
0
1
2
theoretical


Model 2: Histogram of Residuals
count
0
2
4
6
−0.50
−0.25
0.00
0.25
0.50
resid(version_2)


Model 2: qqPlot of Residuals
sample
0.4
0.0
−0.4
−2
−1
0
1
2
theoretical

4. A Regression Table

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Dec 10, 2020 - 1:28:41 PM

Table 1:

	Dependent variable:		
	log(residential_diff)		
	order	order + rank	order + rank + density + cases + politics
	(1)	(2)	(3)
order_length	0.017***	0.021***	0.026***
	(0.002)	(0.003)	(0.004)
order_start_rank		0.010*	0.014***
		(0.005)	(0.005)
log('Population density per square miles')			−0.082
			(0.060)
log(avg_new_cases_per_100K)			0.032
			(0.079)
governor_political_partyR			0.085
			(0.109)
Constant	0.660***	0.259	0.252
	(0.148)	(0.268)	(0.339)
Observations	38	38	38
R^2	0.571	0.615	0.663
Adjusted R^2	0.559	0.593	0.610
Residual Std. Error	0.284 (df = 36)	0.273 (df = 35)	0.267 (df = 32)
F Statistic	47.826*** (df = 1; 36)	28.005*** (df = 2; 35)	12.578*** (df = 5; 32)
Note:			*p<0.1; **p<0.05; ***p<0.01

Baseline: The table above shows that the regression line from our baseline model fits the data fairly well. According to the regression output, order length is significant at the 99.99% confidence level with a coefficient of 0.017. This means that a 1 day increase in order length is correlated with a 1.7% increase in residential mobility difference, supporting the argument that longer stay at home orders are correlated with fewer people adhering to the order. This effect is practically significant as the range of mobility difference throughout the states is between 2 and 11, so a 1.7% increase for each additional day the order is in place quickly adds up to significant changes in mobility difference for states with longer stay at home orders.

Model 1: With the addition of order start rank as control variable, order length remains significant at the 99.99% confidence level and the coefficient has increased to 0.021. Our Model 1 now indicates a 2.1% change in mobility difference for every 1 day increase in order length. Order start rank is statistically significant at the 90% confidence level with a positive coefficient of 0.01 which supports the relationship of higher start ranks (later start dates) being correlated with fewer people adhering to the order. Looking at the adjusted R-squared value of 0.593, Model 1 captures more of the variance compared to our Baseline which had an adjusted R-squared of 0.559. This increase supports our decision in adding order start rank as a control variable as it does have an effect on our dependent variable. We ran an additional F-test comparing our Baseline model and Model 1 to see if we could find evidence supporting our inclusion of order start rank in our model. The results show statistical significance at the 90% confidence level allowing us to reject the null hypothesis of insignificant improvements in explaining variance.

```
## Analysis of Variance Table
##
## Model 1: log(residential_diff) ~ order_length + order_start_rank
## Model 2: log(residential_diff) ~ order_length
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
## 1     35 2.6042                                
## 2     36 2.9082 -1  -0.30395 4.0851 0.05096 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 2: With the further addition of log(population density per square mile), log(average number of new COVID cases per 100k people), and the state governor's political party as control variables, order length continues to remain significant at the 99.9% confidence level and the coefficient has increased now to 0.026. Our Model 2 now indicates a 2.6% change in mobility difference for every 1 day increase in order length. Out of our control variables, order start rank is the only one statistically significant at the 90% confidence level and has a positive coefficient of 0.014. Looking at the adjusted R-squared value of 0.61, Model 2 captures more of the variance compared to Model 1's adjusted R-squared of 0.593. However, with such a small increase in R-squared despite the addition of multiple control variables, our F-test comparing Model 1 and Model 2 was unable to support significant improvements in our model's ability to explain variance from the inclusion of these control variables.

```
## Analysis of Variance Table
## 
## Model 1: log(residential_diff) ~ order_length + order_start_rank + log(`Population density per square
##     log(avg_new_cases_per_100K) + governor_political_party
## Model 2: log(residential_diff) ~ order_length + order_start_rank
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     32 2.2837                            
## 2     35 2.6042 -3  -0.32055 1.4972  0.234
```

5. Discussion of Omitted Variables

We are answering a descriptive question, not explanatory

Although we are trying to answer a descriptive question, we wanted to discuss omitted variables that could potentially bias or affect the significance of our results.

1. Changes in weather as the United States as a whole moved from spring to summer could result in people having greater motivation to go outside due to the warmer weather. Although we would expect higher temperatures to correspond with increases in mobility differences, it's unclear how temperature differences would affect the lengths of stay at home orders and we are therefore unable to determine the direction of this bias.
2. Instead of ending their stay at home orders, some states (ex. CA) chose instead to amend their orders to allow certain reopenings to occur. Such policies would result in both longer order lengths and greater mobility differences, leading us to expect the omission of this variable to have a positive bias that pushes the observed effect away from zero.
3. There is a lot of county-level variation within a state that we are not accounting for with our model. Whether this involves differences in population density, demographics or the restrictiveness of orders issued by individual counties, the full effects of population density and the average number of new Covid-19 cases per 100k people may not be accurately represented in our model results. Additionally, there may be relationships between variables that would appear on the county level but not state level that we would want to include in our model if we had operated on that level of granularity.
4. The ratio of blue-collar versus white-collar workers within a state would indicate the proportion of people who are motivated to commute to work as much as possible to earn income compared to those who are more likely to be able to work remotely for longer periods of time. Although we do not know whether or not this characterization of constituents would actually influence policy-makers' decisions, we would expect states with higher ratios to face more pressure to shorten their stay at home order lengths. Along with the greater urgency with which people would try to go back to work that would increase mobility differences, we would expect the omission of this variable to have a negative bias that pushes the observed effect towards zero.
5. Racial demographics have been historically correlated with income and recent reports have also shown that Covid-19 disproportionately impacts specific groups. As a result, the racial makeup of a state may be correlated with our variables of interest in ways that are not fully captured when observing the effects of median income and the number of Covid-19 cases on their own.

6. Conclusion

Our models have shown the lengths of stay at home orders do have a statistically significant direct correlation with changes in residential mobility at the 99.99% confidence level. As we added additional variables to our model to control for possible confounding factors, the coefficient for order length increased to our final Model 2 result which indicates a 2.6% change in mobility difference for every 1 day increase in order length.

Regarding our primary research question, we did observe that people follow stay at home orders less strictly over time through the significant rise in the decrease of residential mobility when comparing the average mobility in the first week of the stay in place order to the average mobility in the last week of the stay at home order. In other words, the longer a stay at home order is in effect, the less strictly people follow it from the beginning to the end of the order.

In the future, we would like to distinguish between the lack of mobility difference from states that never really started following the order in the first place and states that continued to follow the order just as strictly as they did at the beginning. As mentioned in our Discussion of Omitted Variables, we would also like to conduct research using more granular data at the county-level to isolate variations in population density, demographics, and county-level restrictions that are not captured when only evaluating variables at the state-level. Additionally, we would like to conduct a time-series analysis to evaluate trends in residential mobility over time instead of just snapshotting the first and last weeks of stay at home orders.