Lab 2 (Draft)

Charis Chan, Joyce Ching, Inderpal Kaur

Last Knit: 08 December, 2020

Introduction

Covid-19 has drastically changed our lives, in terms of: plans for the future, workplace environment, meeting new people, and even simply where we go day to day. With Covid-19 being an infectious disease that is easily transmittable through close contact with others, it is important to put policies to force people to stay at home as much as they can to protect both themselves and others around them. But policies can be enacted and then ignored. How can we figure out if people are actually abiding by these stay at home orders or not? And would people still abide by the policy the longer the order goes on? Looking at this problem from a policymaker's perspective, we would want to take into account whether or not the public will abide by the policy in order to understand the impact of the policy. This information could be used to gauge public opinion or attitudes toward the policy. We can then use the results of this research combined with recommendations from public health experts to design timelines for shutdowns that will be more effective or less frustrating for the general public. Answering the following questions would be key to figuring out an effective policy to help lower the impact of Covid-19 in our states.

Research Question: Do people follow stay at home orders less strictly over time? Is there a relationship between people's mobility and the number of days that a stay at home/shelter-in-place order is enacted?

Data: The data that we used to investigate our question contains information about county-level mobility, Covid-19 case and death statistics, as well as state-level 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 the percentage change in visits and lengths of stay at different places. This data is compared to the baseline median value for the corresponding day of the week from the pre-COVID 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 before the order ended.

Stay at home/shelter-in-place order policies:

State-level: 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.

County-level: We downloaded data from Healthdata.gov which includes a standardized dataset on COVID-19 State and County Policy Orders.

To find the duration of the stay at home policies, we subtracted the policy start date from the current date to calculate the number of days the order had been enacted.

Covid-19 Information: We downloaded county-level populations, cases, deaths from USAFacts, which collects the daily county-level cumulative totals from state public health websites. We transformed this data to also include the number of new cases and deaths per day proportional to the population of the county.

Appropriateness of our data:

Mobility data: 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 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 over time. For example, since 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).

Policy data: There may be variation in state and county policies because of different levels of strictness and enforcement at the local level.

A Model Building Process

To answer our research question, we want to understand how patterns in people's mobility change over time while a stay at home order is in effect. To measure changes in mobility, we used data from Google's Covid-19 Community Mobility Report to track the percent change from baseline in visits and lengths of stay at specific locations. We chose to focus on the percent change from baseline in residential mobility as our measure for how closely people were adhering to stay at home orders under the assumption that staying at home would be reflected in the mobility data as an increase in visits/lengths of stay in residential areas. We chose to use county-level mobility data based on the Mobility Report's recommendation that we not compare changes between regions with different characteristics due to differences in location accuracy and place categorization.

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 only used data from states/counties that enacted a stay at home order for some period of time and had residential mobility data. We excluded states that enacted stay at home policies that did not specifically restrict the movement of the general public or require people to stay at home.

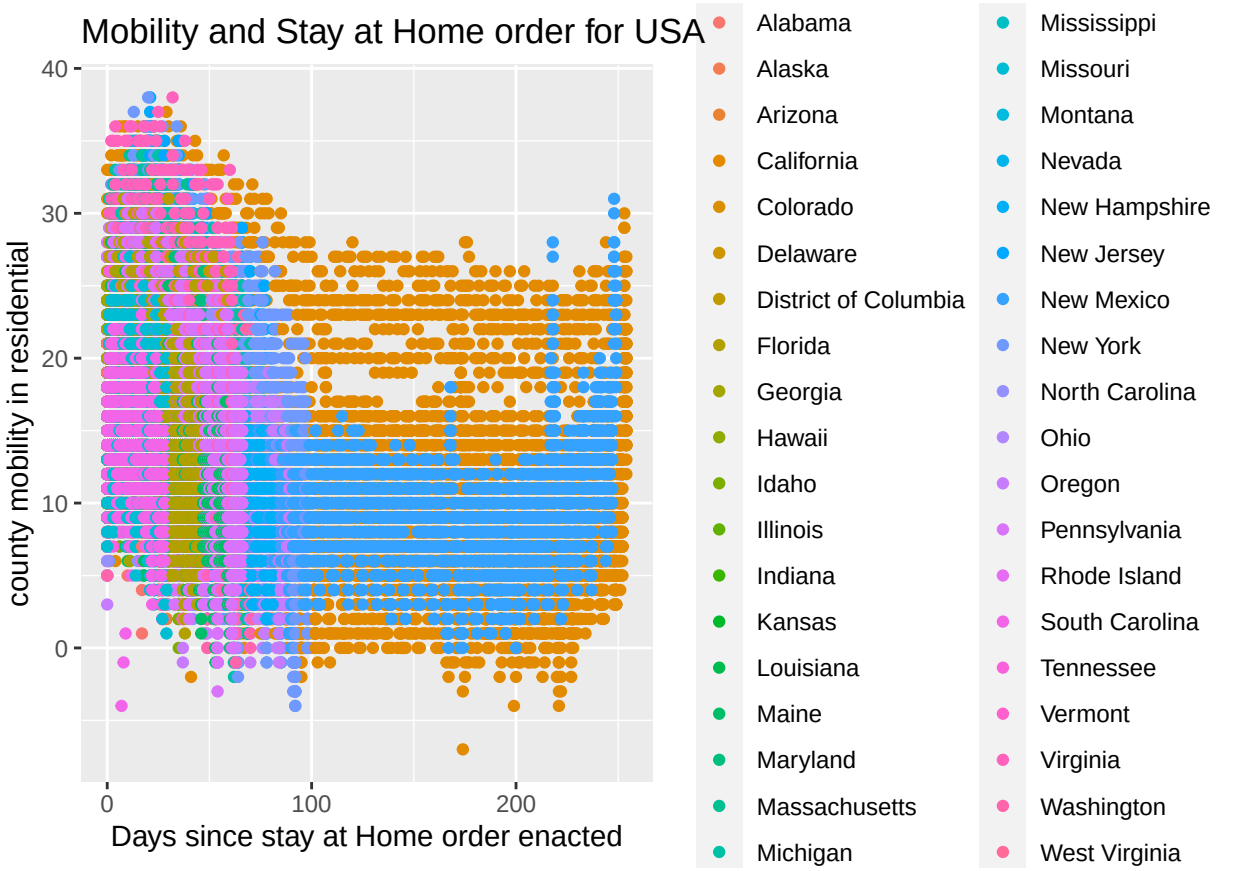
To reflect changes in mobility over time for the duration of the stay at home order, we created a order duration variable that measures the number of days since the stay at home order took effect (i.e. on the start date, the value of the variable would be 1, and on the end date the value of the variable would be the total number of days the order was in effect). We investigated trends in mobility based on this duration variable, which means that we tracked trends from "Day 1" rather than starting on the same calendar date for all states. As a result, counties in different states may have "Day 1" data from different calendar dates. 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.

Using the daily residential mobility data for each county, 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 over time between order duration and percent change from baseline in residential mobility (i.e. the trend shows that people are not staying at home as much as they were when the order was first enacted).

Some covariates that could help us describe the relationship between mobility changes and order duration include county-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 more cases and deaths as a proportion of the population 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 services and businesses (restaurants, bars, gyms, theaters, retail, etc.). Some states amended stay at home orders to gradually reopen certain businesses over time, and so people may be more likely to leave their homes to visit these businesses once they are open.

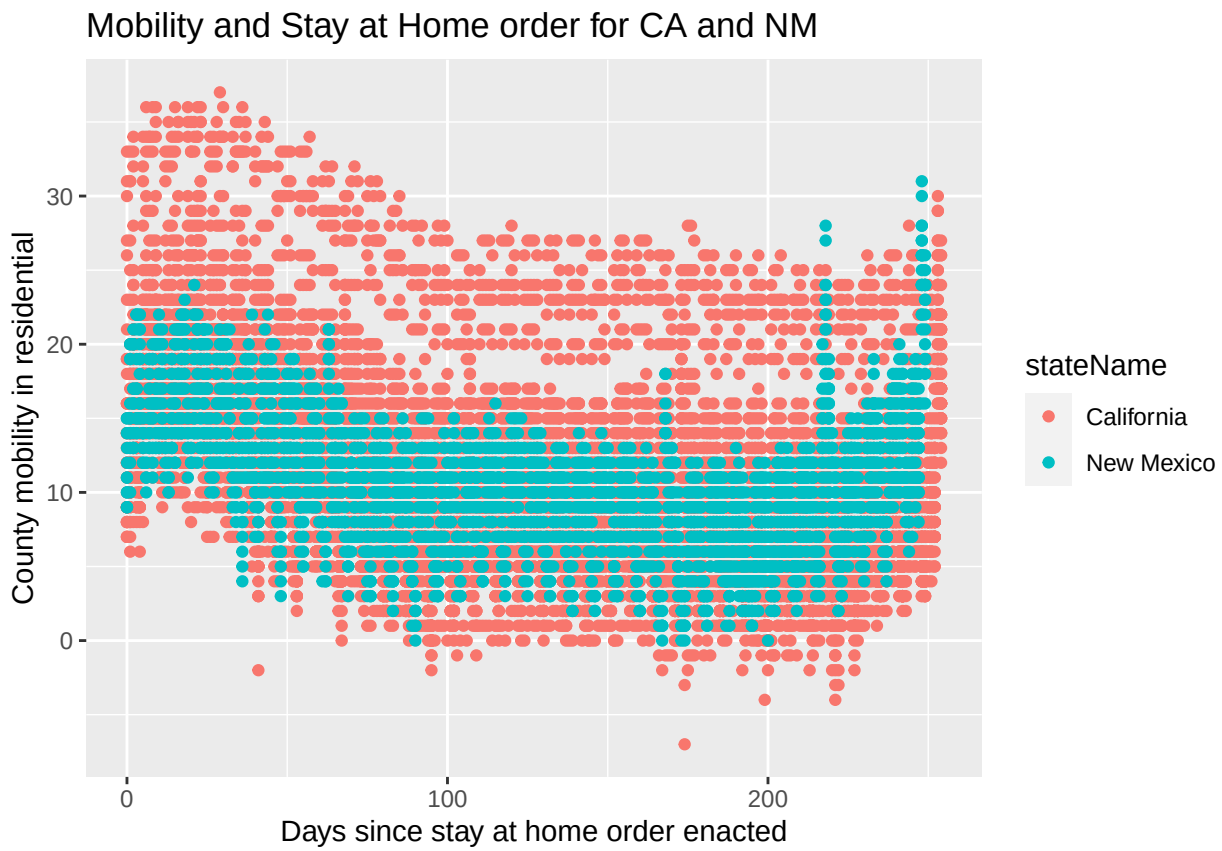
```
## Warning: Removed 54428 rows containing missing values (geom_point).
```



To get a sense for some of the overall national trends in mobility vs. order duration, we plotted these variables for every county for which we had the appropriate data. Although there is a lot of variation across states and counties, the graph seems to show a gradual negative trend between duration and residential mobility on a national level. In other words, as the stay at home order continues, the percent change from baseline in visits/lengths of stay at residential locations decreases (i.e. people seem to be spending less time at home).

Though there may be some national trend, the specific patterns get obscured by variation between states. Therefore, we chose to focus our analysis on a few particular states to narrow the scope of our modeling. Of the states represented, California and New Mexico are the only states that did not end their stay at home orders which have lasted for over 200 days while most other states ended their orders after about 100 days. We ultimately chose to focus on data from these two states because they have longer ranges of data to examine. We note that reducing the scope of our data may reduce variation but ultimately limits the generalizability of our findings to be specific to the states of California and New Mexico.

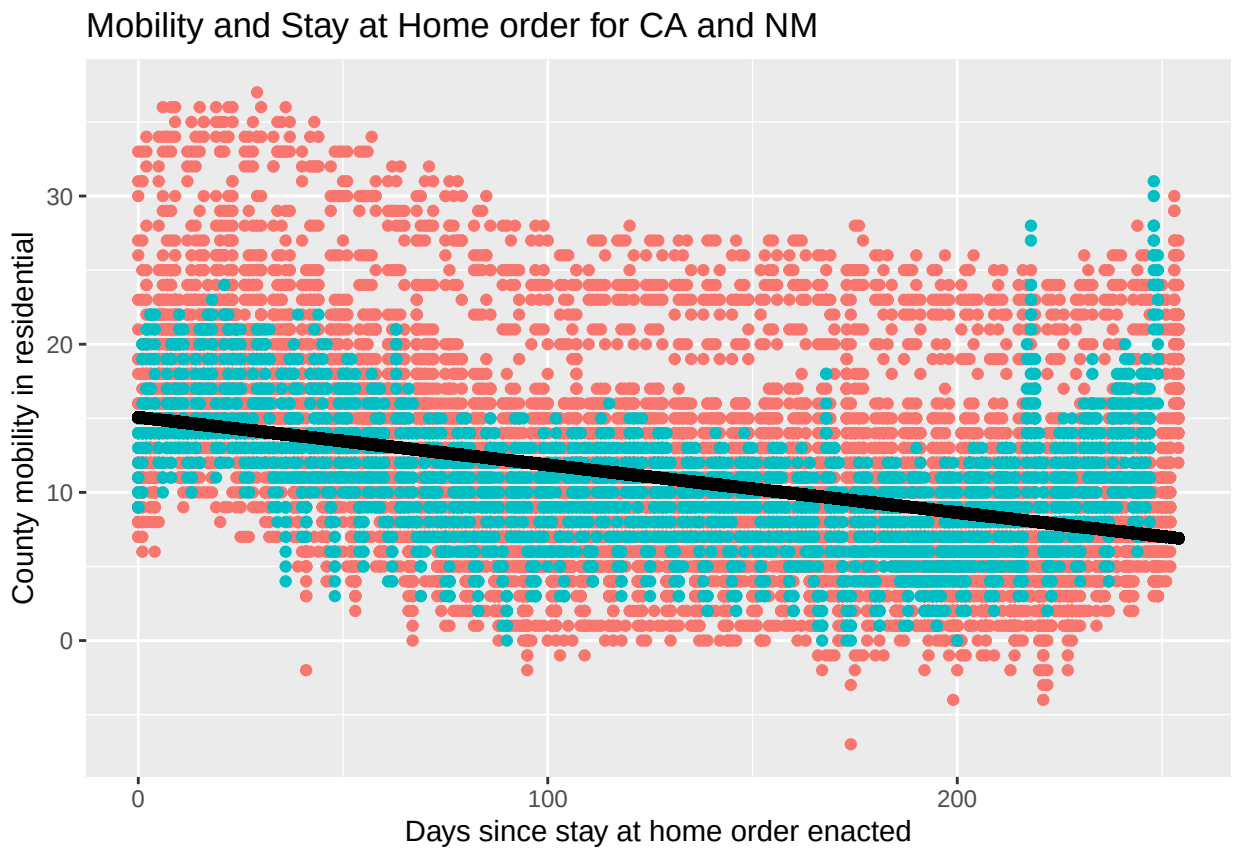
```
## Warning: Removed 5895 rows containing missing values (geom_point).
```



For our baseline model, we only included information about the two key variables (stay at home order duration and residential mobility) for counties in California and New Mexico. The visualization above shows that the trend between these variables for the two states is roughly similar to the trend that we saw at the national level. California counties in particular seem to have more variation than New Mexico counties, which may be a result of the difference in size and population between the two states.

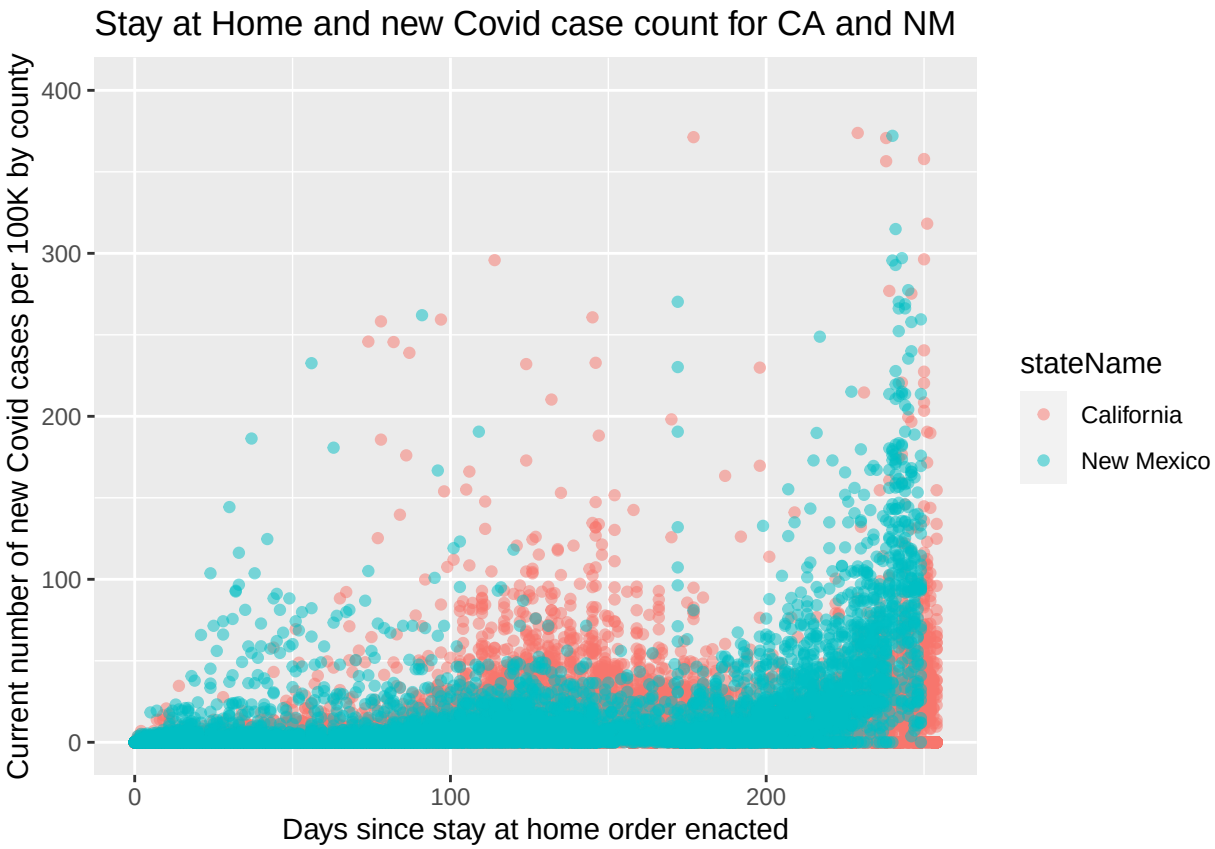
```
##
## Call:
## lm(formula = county_residential ~ order_start_days, data = only_CA_NM)
##
## Coefficients:
##      (Intercept)  order_start_days
##         15.06960          -0.03216
```

```
## Warning: Removed 5895 rows containing missing values (geom_point).
```



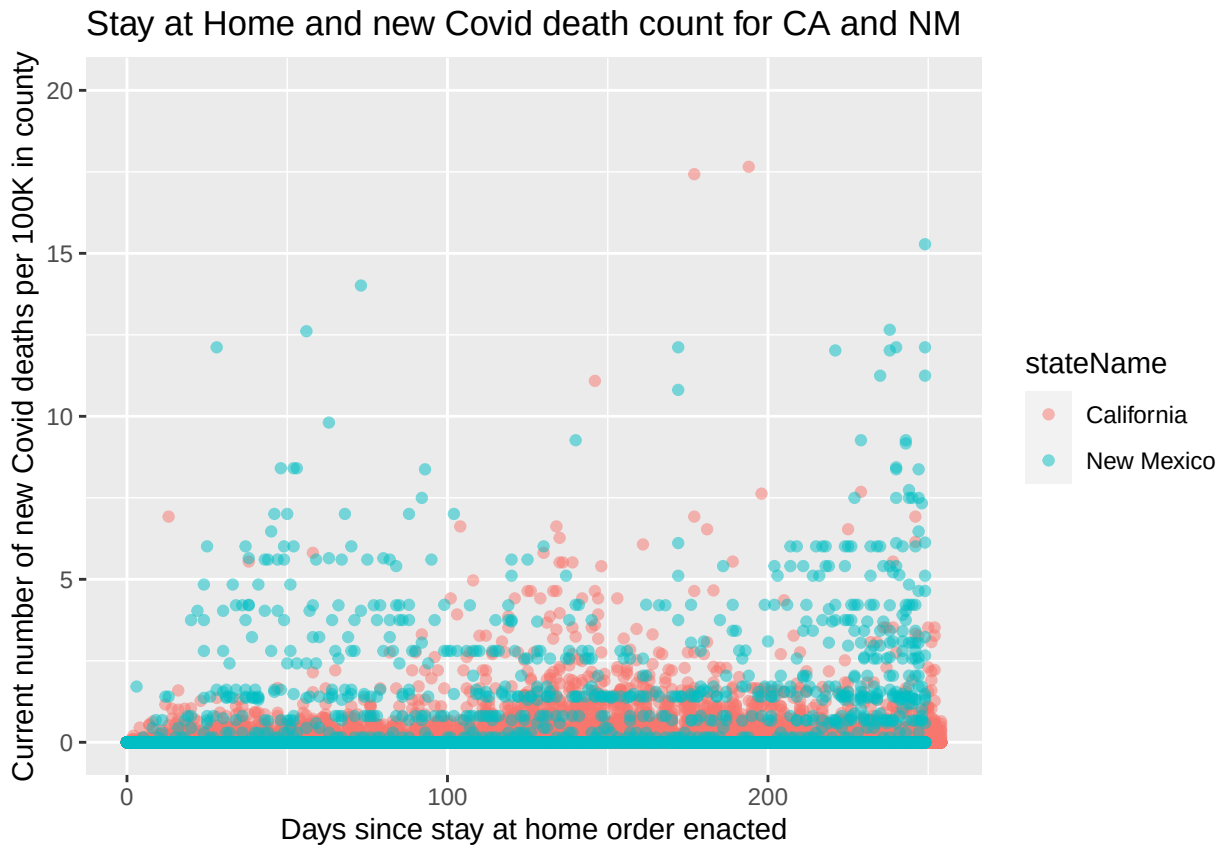
When comparing the trend of new Covid case numbers (per 100K population) vs. residential mobility over the duration of the stay at home order, we see a spike in new case counts after about 100 days since the order was enacted. This time period is preceded by what appears to be a gradual decline in residential mobility (i.e. people were staying at home less than they were at the start of the order) and this matches our assumptions because the time between exposure and symptom onset for Covid-19 can be up to 14 days. Looking at the 2 weeks prior to the spike in new cases, the timeline roughly corresponds to the decline in residential mobility.

```
## Warning: Removed 279 rows containing missing values (geom_point).
```



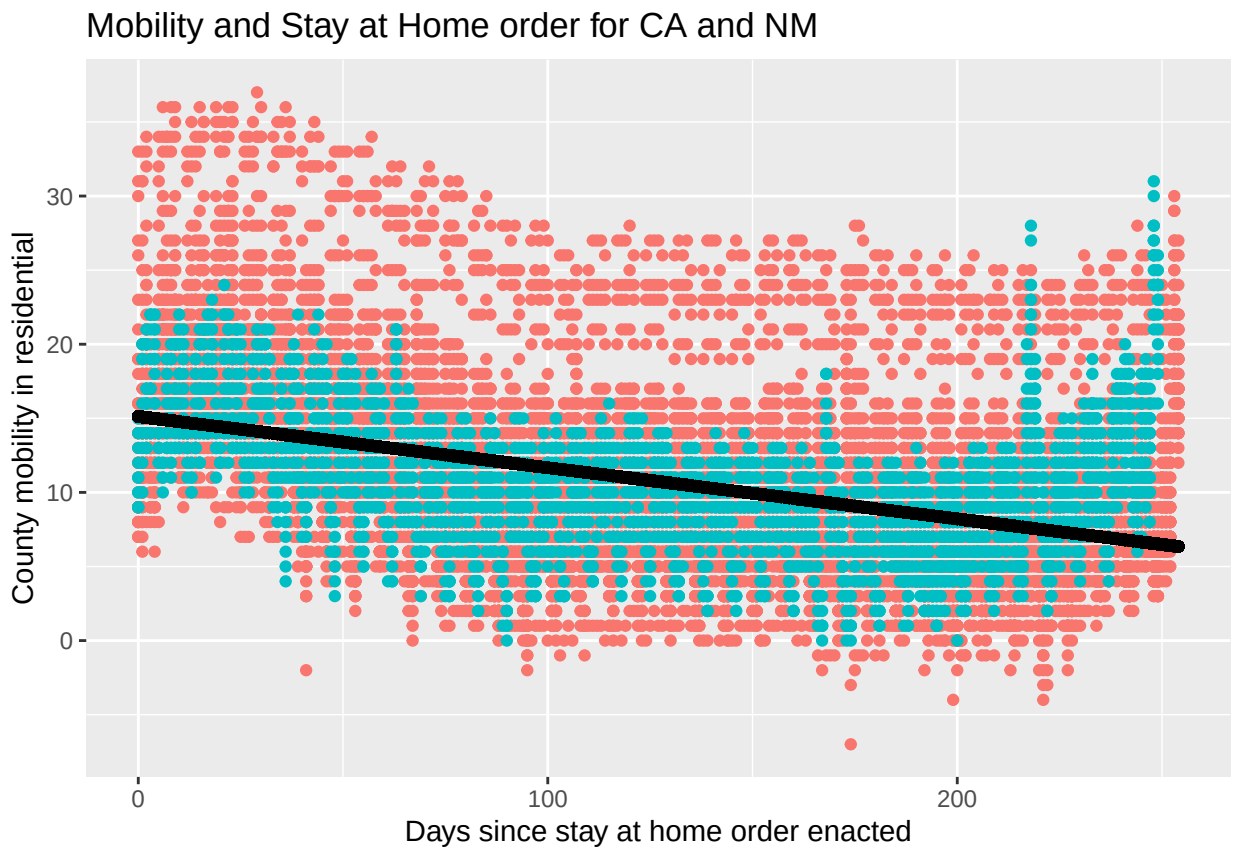
We also examined the trend of new Covid deaths (per 100K population) during the stay at home order but we did not see a strong pattern with respect to mobility over time. Since the number of Covid deaths reflects incidents of exposure from several weeks prior and may be complicated by existing health conditions, we believe these factors may contribute to the lack of apparent pattern in the graph below. Additionally, we were concerned about the possibility for collinearity between Covid case numbers and Covid deaths, so we chose to exclude this feature from our model.

```
## Warning: Removed 268 rows containing missing values (geom_point).
```




```
##
## Call:
## lm(formula = county_residential ~ order_start_days + new_cases_per_100K,
##     data = only_CA_NM)
##
## Coefficients:
##        (Intercept)    order_start_days  new_cases_per_100K
##           15.12274            -0.03455             0.01925

## Warning: Removed 5895 rows containing missing values (geom_point).
```



For our second improved model, we included covariates that represent policies for closing services and businesses. Amendments to stay at home orders may have made people more likely to follow them while businesses were closed and less likely to follow them while businesses were reopening. In particular, we chose to include the length of restaurant and retail closures as well as mask mandate policies. Our other options in the data included bar and gym closures, but we chose not to include the first out of concern for collinearity with restaurant closures and the second because it did not produce significant results.

```
##
## Call:
## lm(formula = county_residential ~ order_start_days + new_cases_per_100K +
##     restaurants_closed + retail_closed + mask_mandate, data = only_CA_NM)
##
## Coefficients:
##         (Intercept)    order_start_days  new_cases_per_100K
##            15.84604            -0.06825             0.01399
## restaurants_closed       retail_closed        mask_mandate
##             0.01793             0.05337             0.05437
```

Limitations of the Model

Below we list the 5 CLM assumptions and the corresponding information/tests that we use to check these assumptions:

1. IID Sampling

Reference population: the entire US population

Time series issues (not independent variables): We have multiple points from the same county on different

days, and each day likely affects the previous day.

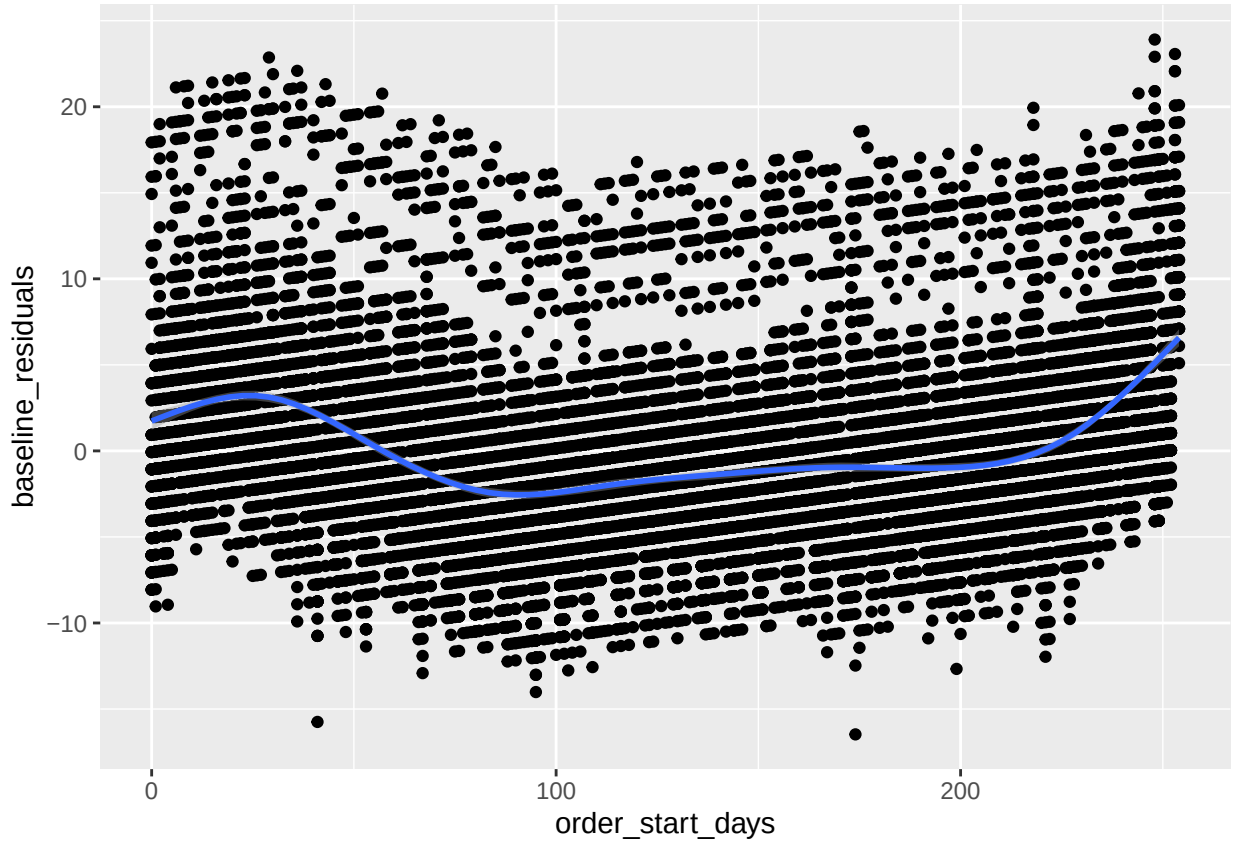
Mobility data doesn't include the entire population (nonrandom sample of users) - Close in terms of social distance: Only users who have a mobile device and a Google Account. Users also need to have opted-in to Location History for their Google Account. - Close in terms of physical distance: People may not have data or live in places with sparse connection (connectivity issues). Privacy thresholds also may not be met if somewhere isn't busy enough to ensure anonymity.

Statistical consequences - Non-independent: unable to provide guarantees about the population (only informative of clusters). - Non-identical: Most statistics are *fundamentally* changed.

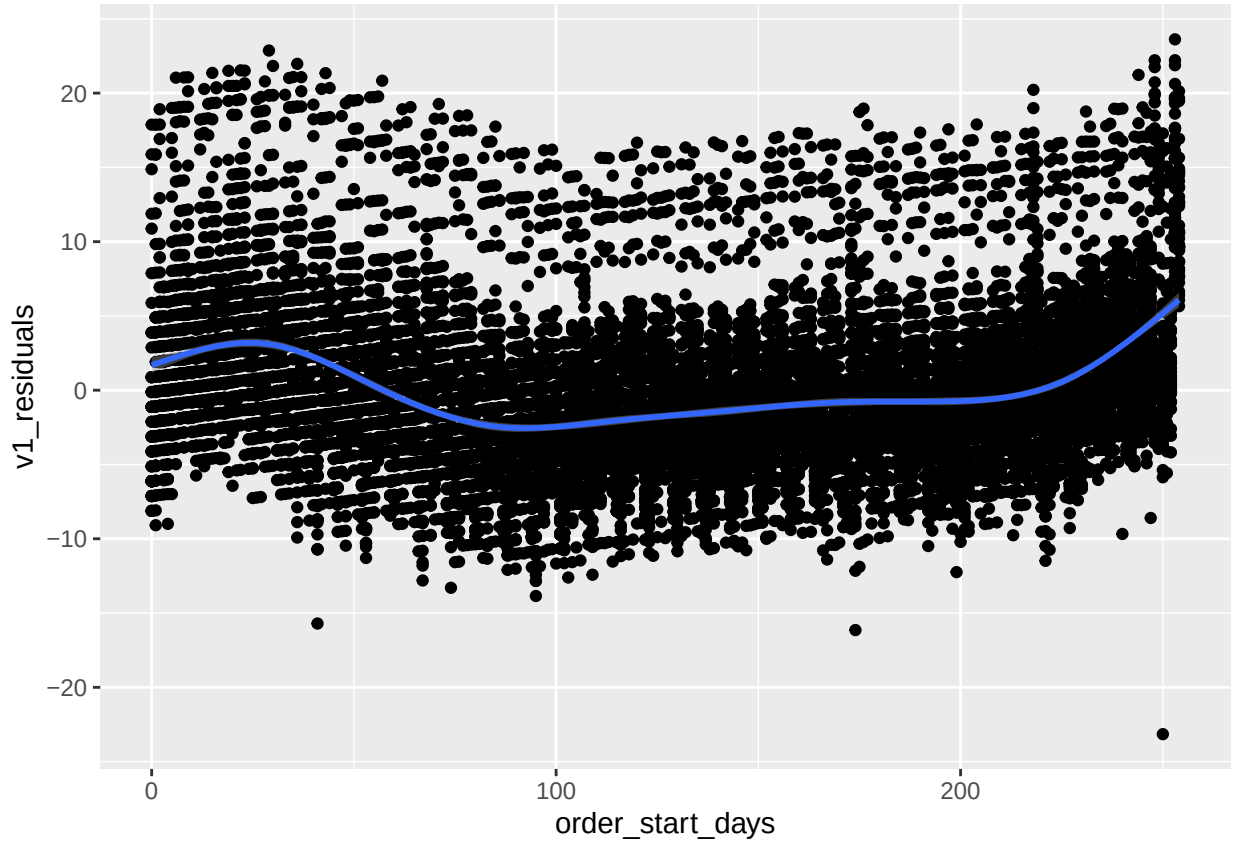
Mitigating the consequences: We need to narrow down our research question to only include individuals who fit the population of those who are sampled. We also need to adjust measures of uncertainty to reflect the clustered nature of the data generating process.

2. Linear Conditional Expectation: The data fits a linear conditional expectation (refer to graph of model residuals).

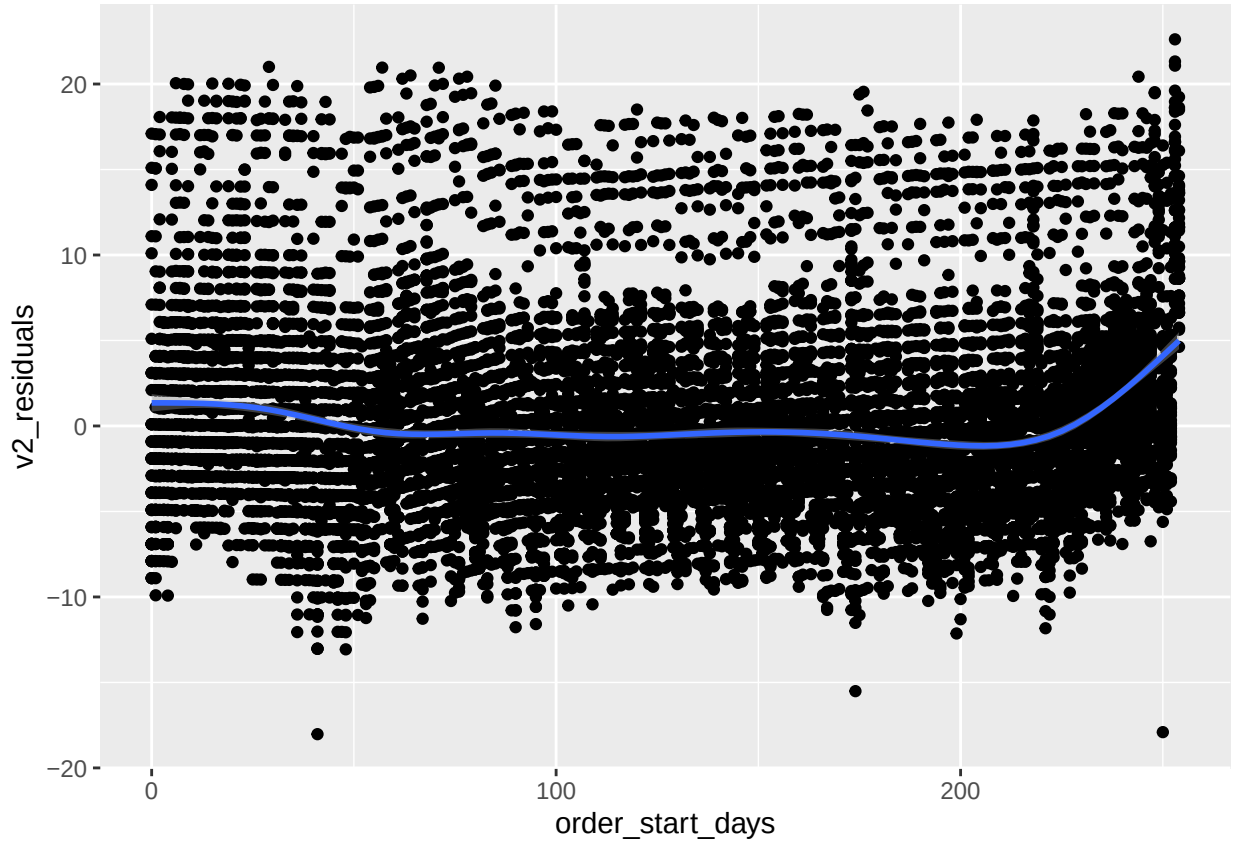
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

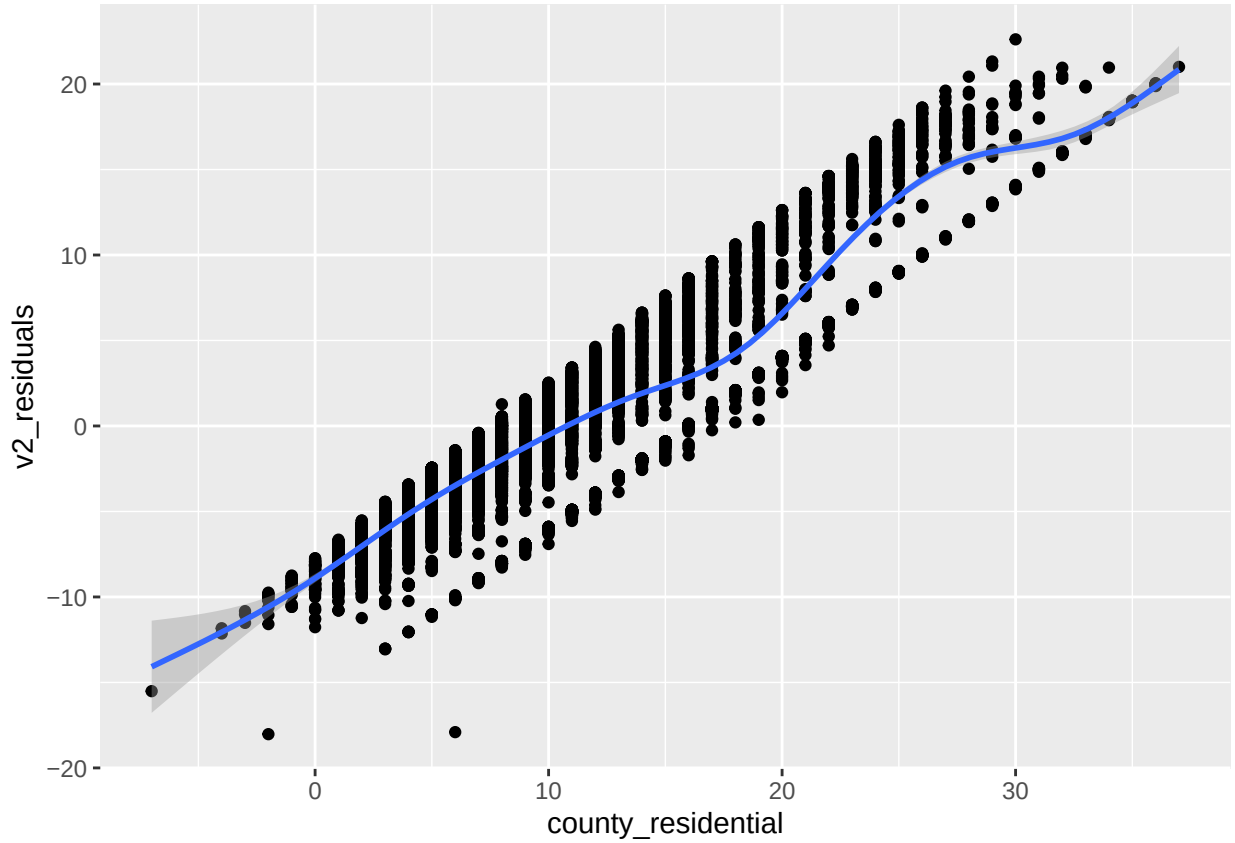


3. No Perfect Collinearity: One data series cannot be produced *exactly* through a simple transformation of other data series. Even though we created new variables from existing ones, we did not include the original variables in our models. We do not have "near perfect" collinearity based on background knowledge on the nature of our variables

4. Homoskedastic Conditional Variance

- Ocular: our graphs of the residuals show that there is no "fanning out" effect
- Breusch-Pagan test: says that we should reject the null hypothesis that there is homoskedastic conditional variance (we'll look into this in future drafts).

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
##
##  studentized Breusch-Pagan test
##
## data:  baseline
## BP = 14.601, df = 1, p-value = 0.0001328
```

```
##
##  studentized Breusch-Pagan test
##
## data:  version_1
## BP = 19.641, df = 2, p-value = 5.433e-05
```

```
##
##  studentized Breusch-Pagan test
##
## data:  version_2
## BP = 54.105, df = 5, p-value = 1.995e-10
```

5. Normally Distributed Errors: Errors appear to be normally distributed.

[Note: CLM assumptions are primarily used for smaller datasets, but our dataset contains ~13,000 data points]

A Regression Table

```
stargazer(
  baseline,
  version_1,
```

```
  version_2, type = "latex", se = list(sqrt(diag(vcovHC(baseline))),
                                       sqrt(diag(vcovHC(version_1))),
                                       sqrt(diag(vcovHC(version_2)))),
  column.labels = c("order", "order + cases", "order + cases + closures"))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Dec 08, 2020 - 11:50:52 AM

Table 1:

	Dependent variable:		
	county_residential		
	order	order + cases	order + cases + closures
	(1)	(2)	(3)
order_start_days	−0.032***	−0.035***	−0.068***
	(0.001)	(0.001)	(0.002)
new_cases_per_100K		0.019***	0.014***
		(0.003)	(0.002)
restaurants_closed			0.018***
			(0.004)
retail_closed			0.053***
			(0.006)
mask_mandate			0.054***
			(0.002)
Constant	15.070***	15.123***	15.846***
	(0.103)	(0.104)	(0.171)
Observations	14,539	14,289	14,289
R^2	0.151	0.160	0.217
Adjusted R^2	0.151	0.160	0.216
Residual Std. Error	5.622 (df = 14537)	5.621 (df = 14286)	5.429 (df = 14283)
F Statistic	2,592.878*** (df = 1; 14537)	1,360.226*** (df = 2; 14286)	789.808*** (df = 5; 14283)
Note:			*p<0.1; **p<0.05; ***p<0.01

Looking at the coefficients for the models above, negative coefficients represent a decrease in percent change from baseline in residential mobility (i.e. people are staying at home less than they were before) as the associated feature increases. The coefficient for stay at home order duration is negative, which shows a negative relationship between duration and mobility. Adding in other covariates for new cases and business closure policies produces positive coefficients, which indicates a positive relationship between duration and mobility (i.e. people stay at home more the longer that businesses are closed/cases are rising). However, most of these coefficients are very small in terms of practical meaning, which may be due to some of the limitations of our model.

The coefficients that we see in the regression table above are significant, but due to the size of our data and the limitations of our model above, we will likely revise our framework to take into account the effect of time dependencies in a more systematic way (ex. time series linear models). We also notice that most of the relationships are absorbed by the constant term in our models, which may be a result of the lack of variation

in county-level information. In future drafts, we plan to either include policy information at a more granular level or pull back on the specificity of our data to only track state-wide changes in mobility.

Discussion of Omitted Variables

We are answering a descriptive question, not explanatory

Here are some omitted variables that we care about:

- Including the tier/phase details for county and state policies would allow us to control for the specific levels of limited mobility the policy expects from the county/state constituents.
- Including the political party associated with the government officials or voting results for each county would help control for the ideological differences regarding COVID-19 that are likely to be correlated with the demographics of each party.
- Including the age demographic of each county population controls for the generalization that younger people are more likely to feel that they are at lower risk from the more severe effects of COVID-19. As a result, they are less likely to take the restrictions as seriously and may be more mobile despite shelter-in-place policies.
- Including the average income of the county population would control for the fact that counties with more people working blue-collar jobs will have people going back in-person to work sooner than those with white-collar workers who are more likely to work remotely.

[If we have time, we will look into the direction and size of the biases caused by omitting these variables]

Conclusion

Although our results appear to be significant in our regression table, we have noted several limitations and un-met assumptions that we would like to address in future drafts before taking those results as conclusive descriptions of the relationship between residential mobility and the duration of stay at home orders. We are currently working on improving our model to account for more states aside from CA and NM or include specific phase/tier policy details. We believe that these steps would allow us to increase variation in features that have limited information and decrease overall variation to see a clearer trend.