Can implementing state policies help prevent the state-wide spread of COVID-19?

Cody Burker, Emily Fernandes, Margo Suryanaga

Contents

1 Which Covid-19 policies implemented by states have an effect on the statewide initial spread of Covid-19?

1.1 Introduction

The Covid-19 pandemic is the first pandemic seen in a century, affecting each individual worldwide. While many countries issued lockdowns, travel quarantines, and other Covid-19 restrictions on a federal level, the US federal government placed the power with the governors to regulate the implementation of these restrictions on a state-by-state basis.

There has been increased skepticism on the true effectiveness of restrictions in slowing down the spread of Covid-19, especially due to the amount of economic impact resulting from these procedures. Evaluating which Covid-19 restrictions helped prevent the initial spread of Covid-19 within that state helps citizens understand the importance of following these policies. Understanding the effectiveness of Covid-19 restrictions in slowing the initial spread of the disease also helps policy makers justify the implementation of these policies to their constituents. Additionally, this information might be of interest to public health officials who are looking for best practices to contain a future pandemic or infectious disease.

From this, we are focusing on the following main research question:
`Which Covid-19 policies implemented by states have an effect on the statewide initial spread of Covid-19?`

As part of our research we will specifically be looking at the following policies implemented by various states:
1. Stay at home/shelter in place order length
2. Face mask mandate specifically for businesses length
3. Interstate travel quarantine length
4. Restaurant shutdown length
5. Gyms shutdown length
6. Bars shutdown length
7. Declaration of a state of emergency

Understanding the nature of Covid-19 as being a highly infectious disease, we believe the first few months are the most crucial times in controlling an outbreak. Thus, we define an `initial spread of Covid-19` to be the number of cases at 05/31/2020. We believe this reflects the initial rise of cases from 0 to x within the first few crucial months of the pandemic.

We used the following data sources:
a) Covid-19 US State Policy Database A database of state policy responses to the pandemic, compiled by researchers at the Boston University School of Public Health.

We used this data source to get start and end dates to evaluate length in days of policies 1 through 6 listed above and the date a state of emergency was declared to evaluate the speed at which a state of emergency was declared within each state (policy 7). This source additionally provided us with population data by state and population density per square mile by state.

b) NY Times Covid-19 Data Repository A series of data files with cumulative counts of coronavirus cases in the United States, at the state and county level, over time.

We used this data source to get the number of cases at 05/31/2020 on a state level.

1.2 A Model Building Process

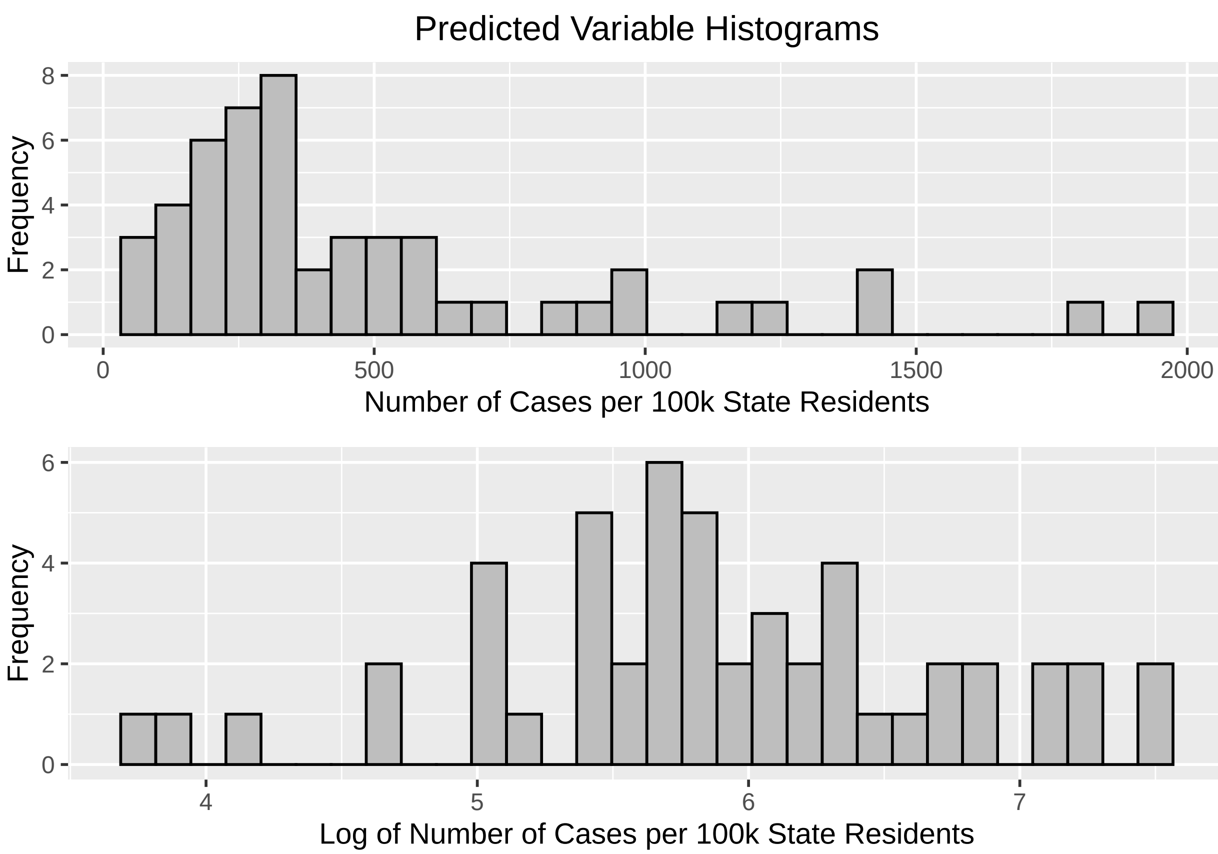
Our team is studying the factors that limited the initial spread of the Covid-19 virus, from the first case seen in the United States until May 31st. Our goal is to create a series of models to explain the relationship between the various state policies put in place to curtail the spread of the virus and the number of cases observed in each state. In the analysis we will use the following policies as input variables:

1. Business mask mandate (measured in days the policy was in place)
2. Stay at home/shelter in place order (measured in days the policy was in place)
3. Interstate travel quarantines (measured in days the policy was in place)
4. Restaurant/Bar/Gym Shutdowns (measured in days the policy was in place)
5. State of Emergency declared in relationship to 3-11-2020 (Date that the World Health Organization declared the Covid-19 Pandemic)

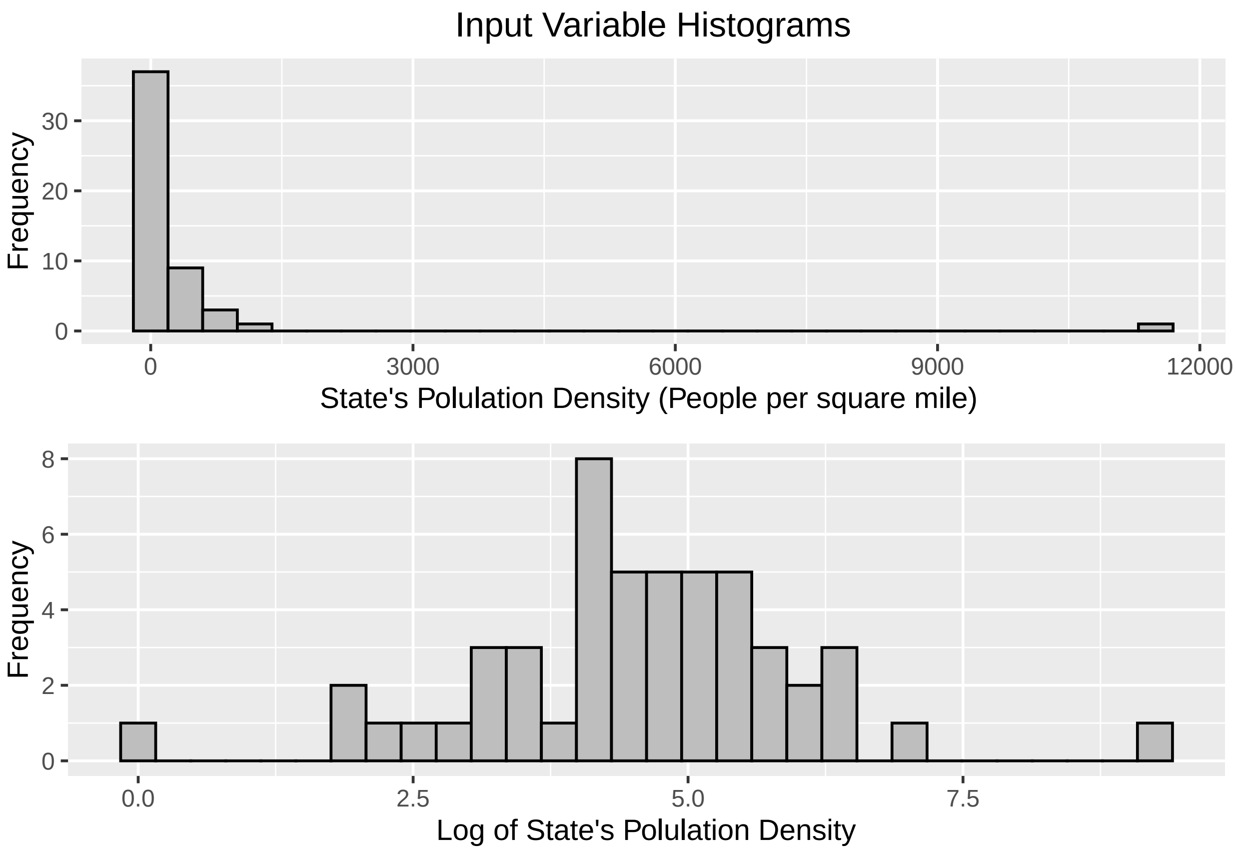
The spread of Covid-19 is not just influenced by state policies, but also by the relative closeness of people within each state. In an effort to capture this, we will also include the population density as an input variable. With these inputs, we hope to explain which state policies were effective at curtailing the initial spread of the Covid-19 virus.

1.3 Model 1

First, we began to explore the number of cases in each state per 100,000 residents. This allowed us to account for the number of cases while equalizing the predicted variable to account for the population of each state. For example, New York or California will have a much larger number of cases than Alaska simple due to the fact these states have many more residents. After viewing the histogram of the number of cases per state per 100k residents, our team chose to handle the skew of the data by completing a log transformation. The histogram of the log transformed variable can be seen below.



For our base model, we selected the one input variable that was a characteristic of the state and not a policy enacted by the state. We hypothesized the state's population density is the first building block to describe the number of Covid-19 cases on May 31st, 2020. This virus spreads thought close human contact and population density captures the relative closeness of residents in a state. Again, after viewing the initial highly skewed histogram we chose to complete a log transformation of this variable. The transformed histogram can be seen below.



Our base model will look at the log relationship of each state's cases on May 31st, 2020 per 100k residents with that state's population density.

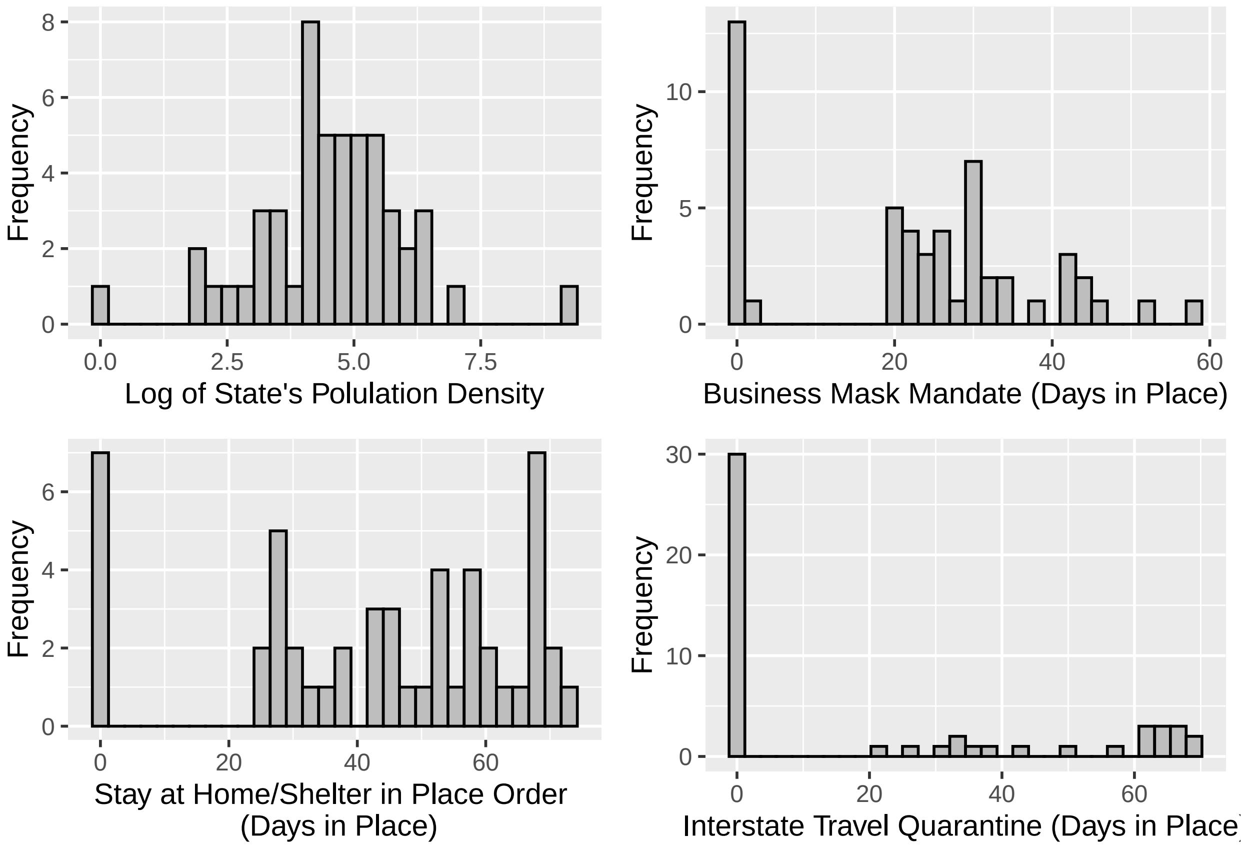
1.4 Model 2

For our next model, we built off the basic model and added several inputs to represent key state policies. This medium complexity model will account for the number of days each state had a:

1. Business mask mandate
2. Stay at home/shelter in place order
3. Interstate travel quarantine

The stay at home order was the first and strongest policy set by many states. This was a drastic policy enacted to limit face to face interactions in hopes that it would slow the spread of the virus. We expect a strong relationship between long stay at home orders and curtailed spread. The interstate travel quarantine was another attempt to limit face to face interactions with those who may not follow that state's policies. This variable captures each state's attempt to isolate themselves from other state policies. Finally, we included a variable to describe each state's mask mandate for businesses in order to capture the safety measures taken as states attempted to reopen their economies.

We first reviewed the histograms of our new input variables. Aside from the spike at 0 (which indicates states that did not wish to implement this specific policy), the variables look fairly normal.



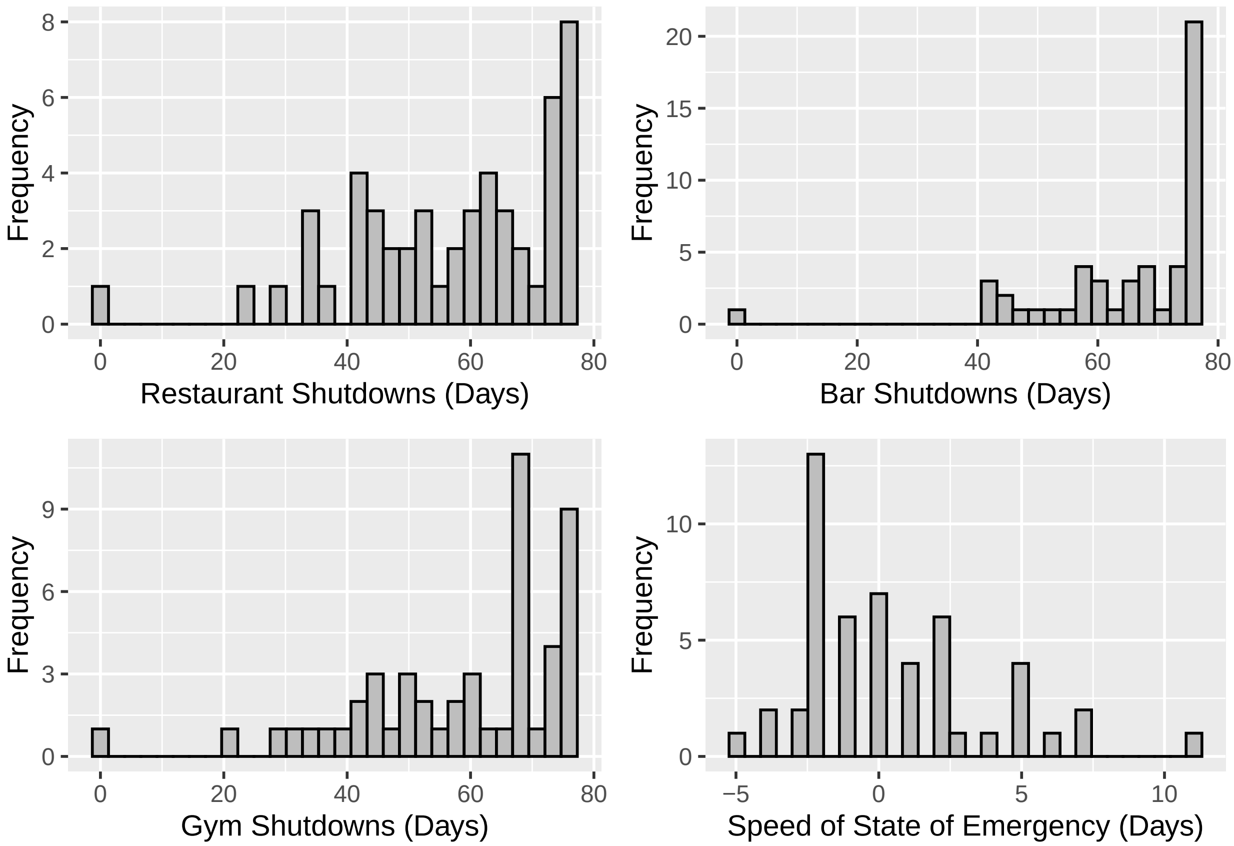
1.5 Model 3

And finally for our most complex model will also account for the number of days each state had:

1. Restaurant Shutdowns
2. Bar Shutdowns
3. Gym Shutdowns
4. State of Emergency Speed (measured in days from 3/11/2020)

In our final model we included data that supported the policies covered in our medium complexity model. For example, the shutting down of restaurants, bar, and gyms made it much easier for the public to comply with the stay at home mandate. If the businesses of the state are closed, where will people congregate? Unfortunately, this may lead to colinearily between the variables. This colinearily could add noise to our model if shutdowns occurred simultaneously or if shutdowns occurred at the same time as the stay at home order. We expect relationship between long closures and lower infection rates. Finally we attempted to add a variable to capture how quickly a state responded to this crisis. Preemptive policies will lead to fewer initial cases and therefore fewer initial vectors to spread the disease. This variable represents the days between when a state declared a state of emergency in relation to the day the World Health Organization (WHO) declared Covid-19 a pandemic. We expect negative or small number of days (States of Emergency called before or around 3/11/2020) to align with lower infection rates.

We reviewed the histograms of our new input variables. The speed of each state's State of Emergency declaration closely aligns to a normal distribution. However each of the shutdown variables displays a slight skew. While not ideal, this skew does not span over factors of tens and therefore we did not feel it appropriate to use a log transformation.



It is noteworthy that the variable describing the speed of State of Emergency Declaration may be negative. This means that the state in question declared a State of Emergency before the WHO declared Covid-19 as a pandemic.

1.6 Regression Table

Table 1 displays the regression of all 3 models. We are use robust standard error.

1.6.1 Model 1

We will begin by inspecting the first Model in Table 1. Here we see a significant relationship between Log (Population Density) and Log (Covid-19 Cases). Figure 1 displays a scatterplot of the two, and we do see some relationship. The coeffeficient is positive, indicating that a higher density is related to a larger number of Covid-19 cases per 100,000 people in a state. Because (using robust standard errors) this relationship is significant ($p < 0.001$) we reject the hypothesis that there is no relationship between Log (Population Density) and Log(Covid-19 Cases) in favor of the hypothesis that there is a relationship. The following model was used to test the relationship between the two variables:

$$\log(\text{Covid-19 Cases}) = \beta_0 + \beta_1 \log(\text{Population Density})$$

This model implies that every percent increase in Population Density (measured in people / square mile) causes a 0.36% increase in Covid-19 Cases per 100,000 people in a state, which has much practical significance. While it is likely not feasible to reduce Population Density, this number might inform policy makers as to which states are more vulnerable to future pandemics. Also, this model has an R^2 value of 0.39, it seems as if the transformed population density accounts for a surprising amount of the variance in Log (Covid-19 Cases).

Table 1: Regression Table

	Dependent Variable: Log (Covid-19 Cases)		
	Simple Model	Moderate Model	Complex Model
	(1)	(2)	(3)
Log (Population Density)	0.358***	0.329***	0.350***
	(0.063)	(0.084)	(0.105)
Business mask mandate (days)		0.016***	0.016**
		(0.006)	(0.007)
Stay at home/shelter in place order (days)		−0.010**	−0.010
		(0.005)	(0.007)
Interstate Travel Quarantine In Place Length (days)		−0.009**	−0.009*
		(0.004)	(0.005)
Restaurant Shutdown Length (days)			−0.002
			(0.015)
Bar Shutdown Length (days)			−0.008
			(0.018)
Gym Shutdown Length (days)			0.011
			(0.016)
State of Emergency Speed (days)			−0.027
			(0.033)
Constant	4.257***	4.628***	4.587***
	(0.306)	(0.379)	(0.995)
F-Statistic	32.4*** (df = 1; 49)	32.4*** (df = 5; 45)	4.15*** (df = 9; 41)
Observations	51	51	51
R^2	0.391	0.586	0.601
Adjusted R^2	0.378	0.550	0.525
Residual Std. Error	0.675 (df = 49)	0.574 (df = 46)	0.590 (df = 42)

Note: *p<0.1; **p<0.05; ***p<0.01

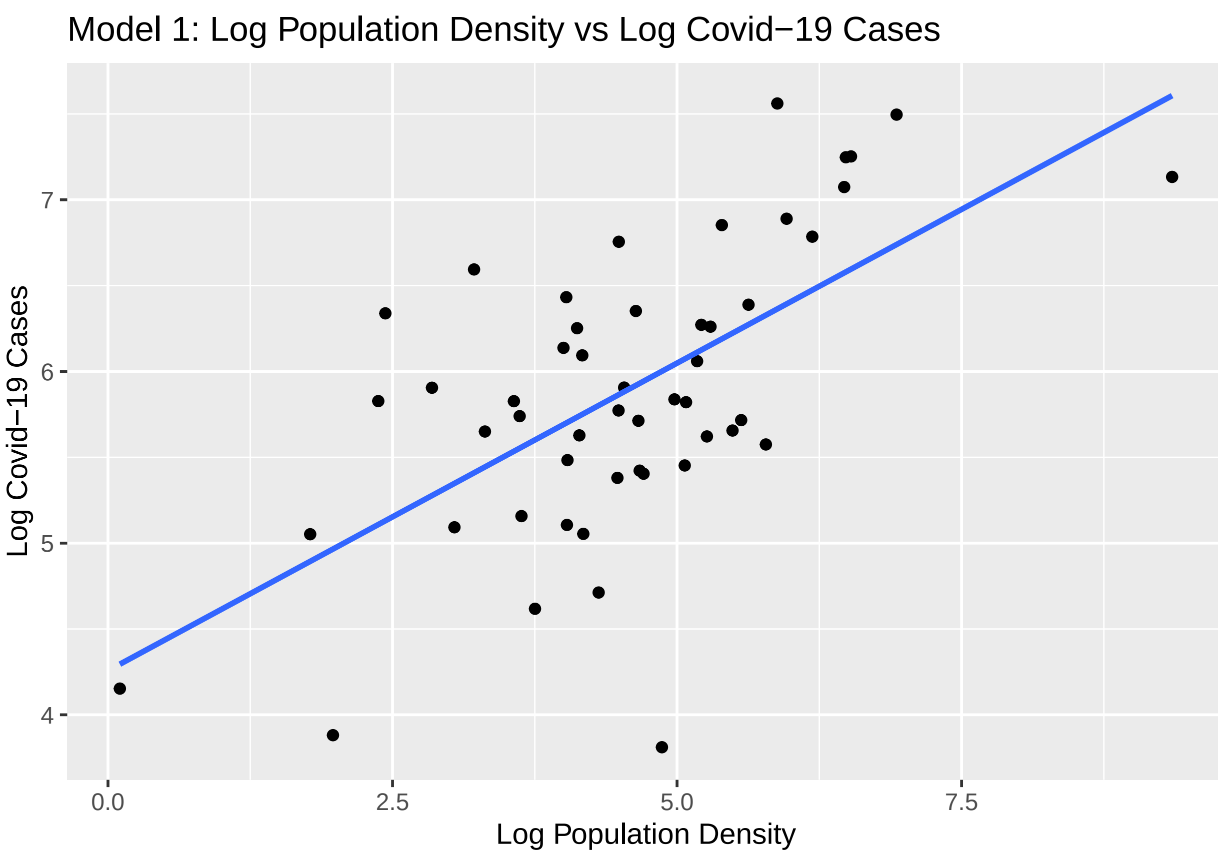


Figure 1: Scatterplot of Model1

1.6.2 Model 2

For our second model, we find evidence that there is a relationship between how long states took to enact business mask mandates after the pandemic was declared (Business mask Mandate), how long the stay at home order was in place (Stay at home/shelter in place order), and how long the interstate travel quarantine was in place (Interstate Travel Quarantine In Place Length) in addition to the Log (Population Density). The coefficient for Log (Population Density) decreases from 0.358 to 0.329 (but remains significant), indicating that some of its explanatory power is shared between the other variables we have included (and also perhaps there is a small amount of colinearity between this variable and one or more of the others).

When we inspect the Stay at home/shelter in place order coefficient, we see that it is both significant and negative This indicates that for every extra day that a state spend with a Stay at home/sheter in place order in effect if all else is held equal, we expect to see a 0.979% *decrease* in Covid-19 cases. This is a very large number, and because it has a small p-value we are sure that there is a relationship. This speaks strongly to the efficacy of such orders, and the low p-value (0.048) gives us reason to believe that there is a relatively low chance this relationship exists by chance.

Inspection of the Interstate Travel Quarantine In Place Length leads to a similiar result. We see that the p-value is low enough (0.041) that we should reject the idea that restricting travel does not affect Covid-19 cases. The effect size (1 more day of restricting interstate travel leads to 0.888% decrease in Covid-19 cases, all else held equal) is nearly as large as the Stay at home/shelter in place order, meaning that they have relatively similar effects on Covid-19 cases.

When we inspect the Business mask mandate variable, we find something surprising: it is significant, however it is also positive. This is counterintuitive: we would expect that the longer a mandate for business masks is in place, the lower the number of cases. One explanation for this might be reverse causality. Although we assumed that Business mask mandate influenced Log (Covid-19 Cases), it is possible that Log (Covid-19 Cases) also influenced Business mask mandate. For example, as the number of Covid-19 cases increased, states lengthened or reenacted their business mask mandates as a result/in reaction to rising cases. This would explain why, according to our model, every 1 days of increased business mask mandate is associated with a 1.65% increase in Covid-19 cases per 100,000 residents.

1.6.3 Model 3

For our last model, we add many extra variables. However, none of the additional variables were significant, and the Adjusted R^2 actually decreased (from 0.550 to 0.525) compared to the second model, indicating that this model is not efficient. We also seem some interesting effects on the variables that were already in model 2. Stay at home/shelter in place order and Interstate Travel Quarantine In Place Length actually are no longer significant. This might indicate that there exist one or more linear relationships between these two variables and the new variables added. Business mask mandate remains significant though. Log (Population Density) retains its significance (alluding to a very strong relationship). Overall this model does not seem to provide much useful information compared to the previous, simpler models.

1.7 Limitations of your Model

For each model we will evaluate the assumptions for a Classical Liner Model. These requirements are:

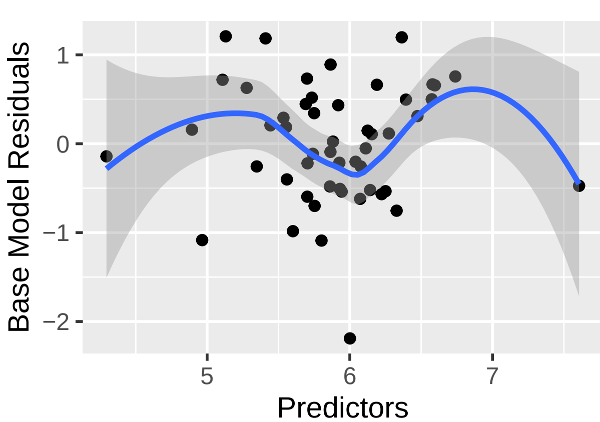
1. Independent and identically distributed data
2. Linear Conditional Expectation Exists
3. No Colinearity
4. No Homoskedastic Error
5. Normal Residuals

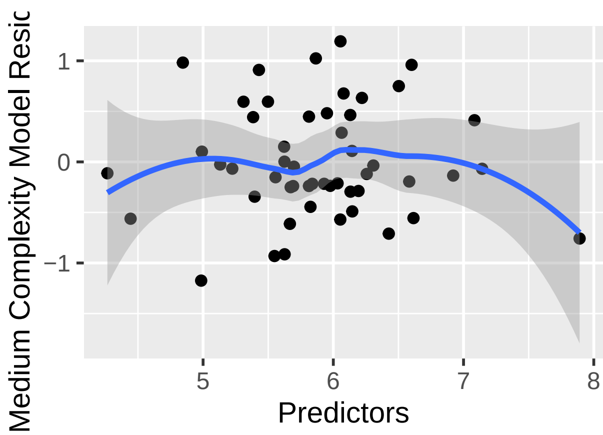
2 Independent and identically distributed data

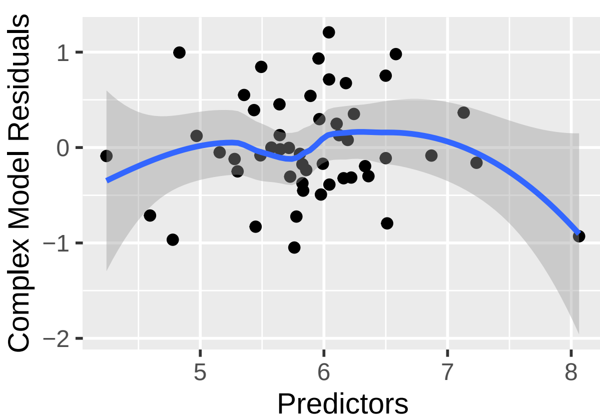
The data use in our models represents the number of Covid-19 cases and state level policy decisions for all 50 states of the United States and the District of Columbia. While we can assume the data from each state/district is identically distributed, there may be some difficulty making an argument that the data is independent. There is an argument to be made that the Covid-19 cases in one state may effect an adjacent state. Additionally the policies of one state, a collection of states, or the federal government may have an influencing effect of a state's policy. This lack of independence may explain the skew to the shutdown data. Large populated states such as California and New York began to shut businesses down early and many other smaller states followed suit shortly afterwards.

3 Linear Conditional Expectation Exists

To evaluate if the relationship between our input and output variables is indeed linear, we plotted the predicted values against their residuals.







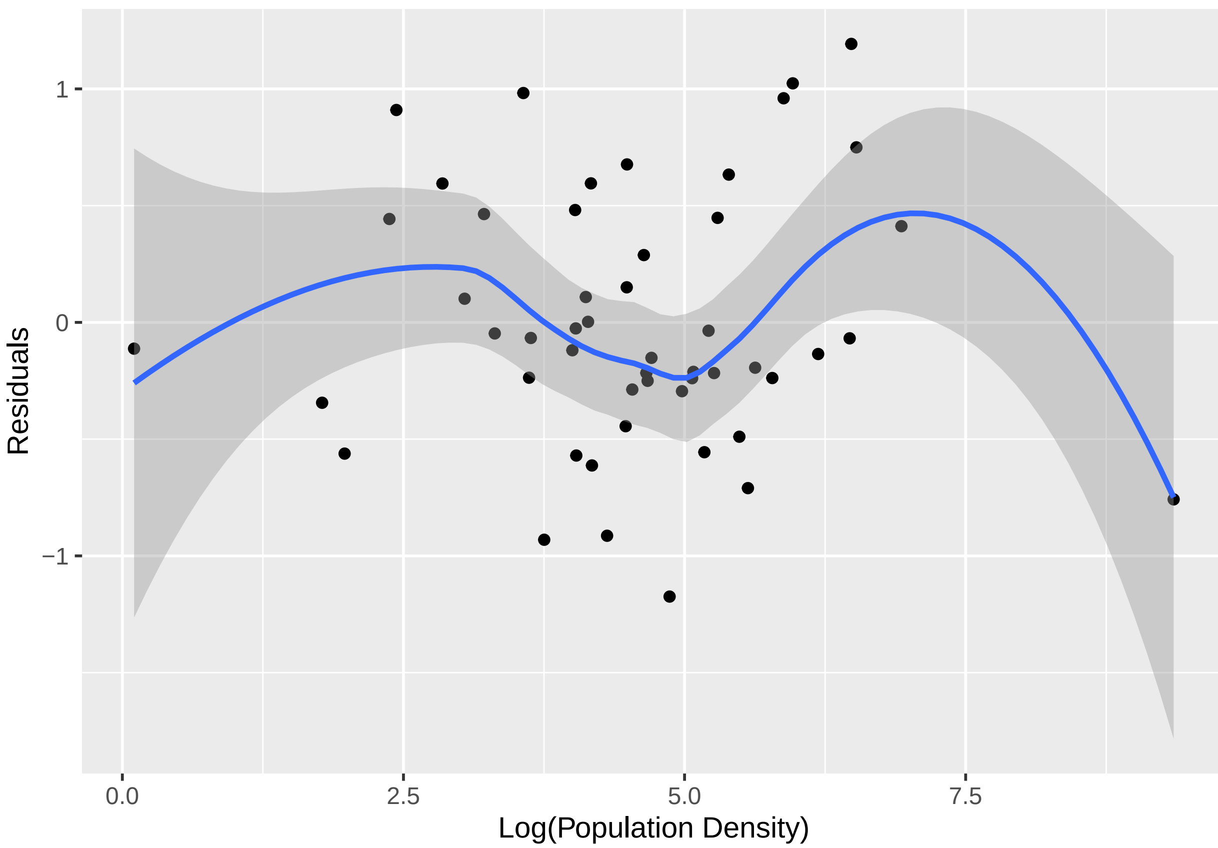
From the Residual/Predictors plot shown above, one can see that this basic model's predictors has a fairly linear relationship with residuals implying that a Linear Conditional Expectation exists for all models.

4 Colinearity

Our base mode cannot possibly have colinearity because it only has one input variable. We evaluated multicollinearity using variance inflation factors for our medium complexity model and complex model. Our medium complexity model's variance inflation factors are particularly small (less than 2) and imply that there is no concern for multicollinearity. However our complex model's variance inflation factors are quite large (several of the Shut Down variables are between 5-9). This implies that there may be multicollinearity. This judgment is subjective. If after reviewing the model coefficients we believe that this potential multicollinearity creates confusion in the model, the solution would be to limit our input variables to one variable to that we believe is representative of all the shut down inputs (for example, the model could include only the restaurant shutdowns instead of restaurant, bar, and gym shutdowns).

5 Homoskedastic Error

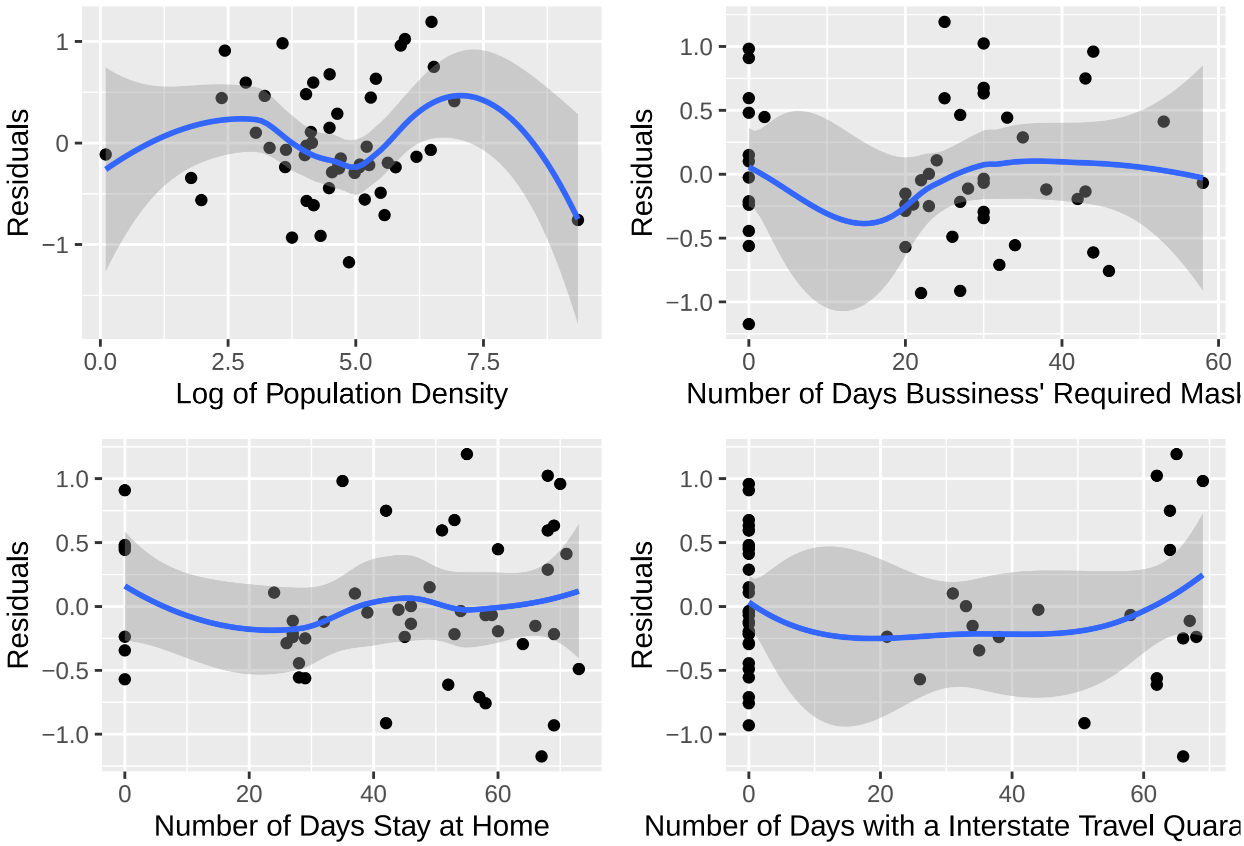
To evaluate the presence of homoskedastic error, we will plot the residuals in relationship to the model's inputs. For our base model, we will plot the base model residuals versus the log transformation of each state's population density.



The Residual/Input plot displays a flaring of the standard error at the extremes which implies the potential for Homoskedastic errors.

For our medium complexity model, we will plot the residuals in relationship to the model's four inputs. We will plot the base model residuals versus:

- the log transformation of each state's population density
- Business mask mandate (days in place)
- Stay at home/shelter in place order (days in place)
- Interstate travel quarantine (days in place)

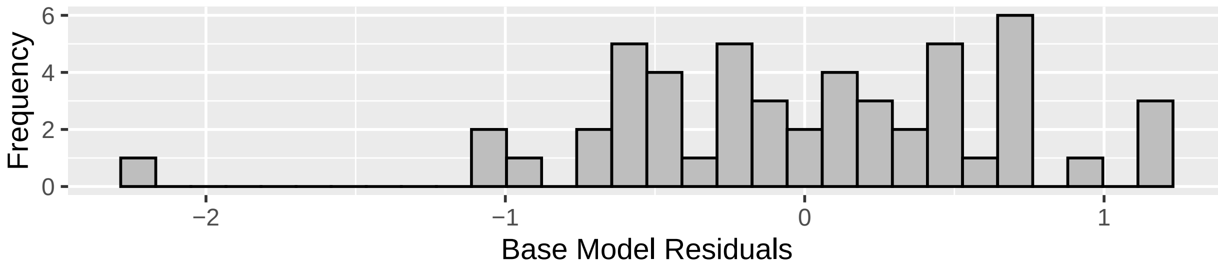


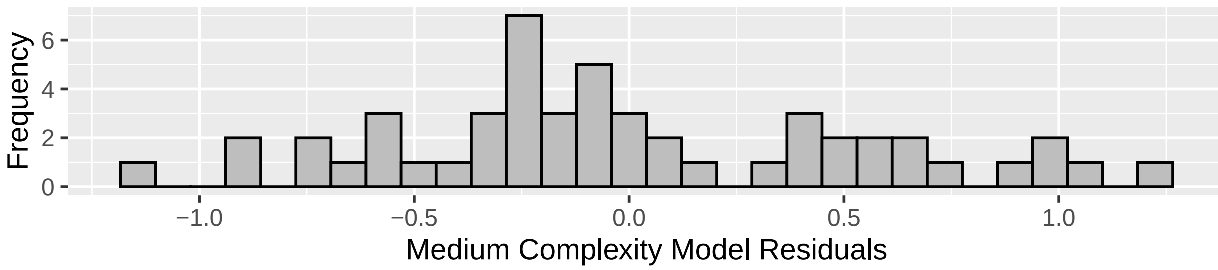
The Residual/Input plot for the log of the population density still displays a flaring of the standard error at the extremes. The business mask mandate and interstate travel quarantine also display inconsistent standard errors. These factors suggest the potential for Homoskedastic errors. To mitigate this, we will use robust standard error when reviewing the significance level of medium complexity model's coefficients.

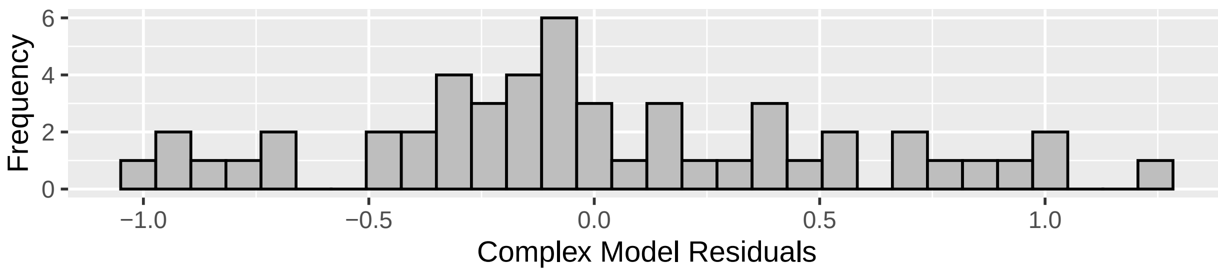
The log transformation of each state's population density also effects the complex model (suggesting Homoskedastic error), therefore we will use robust standard error when reviewing the significance level of all model's coefficients.

6 Normal Residuals

Lastly we have a histogram of each model's residuals to ensure they are normal. As you can see from the histograms below, the residual's are in fact normally distributed fro all models.







#Other Remarks

Finally it is important to not that our analysis has collapsed this time series data to merely look at one point in time. We have attempted to include variables that are measured in length of days in order to account for the importance of time on this analysis, but this does not counteract the fact that this analysis would be improved by looking at the timing of these policies being implemented.

6.1 Discussion of Omitted Variables

If the team has taken up an explanatory (i.e. causal) question to evaluate, then identify what you think are the 5 most important *omitted variables* that bias results you care about. For each variable, you should *reason about the direction of bias* caused by omitting this variable. If you can argue whether the bias is large or small, that is even better. State whether you have any variables available that may proxy (even imperfectly) for the omitted variable. Pay particular attention to whether each omitted variable bias is *towards zero or away from zero*. You will use this information to judge whether the effects you find are likely to be real, or whether they might be entirely an artifact of omitted variable bias.

6.2 Conclusion

Make sure that you end your report with a discussion that distills key insights from your estimates and addresses your research question.