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

Cody Burker, Emily Fernandes, Margo Suryanaga

Contents

1 Introduction **2**

2 Model Building **2**
2.1 Model 1 . . . 3
2.2 Model 2 . . . 4
2.3 Model 3 . . . 5

3 Regression Table **6**
3.1 Model 1 . . . 7
3.2 Model 2 . . . 8
3.3 Model 3 . . . 9

4 Limitations **9**
4.1 Independent and identically distributed data . . . 9
4.2 Linear Conditional Expectation Exists . . . 9
4.3 Perfect Collinearity . . . 10
4.4 Homoskedastic Error . . . 10
4.5 Normal Residuals . . . 12
4.6 Other Remarks . . . 12

5 Discussion of Omitted Variables **12**
5.1 Omitted variable 1: Number of tourists . . . 12
5.2 Omitted variable 3: Amount of public sanitation stations . . . 13
5.3 Omitted variable 4: Days with cold or freezing temperatures . . . 13
5.4 Omitted variable 5: Number of schools shut down . . . 14

6 Conclusion **14**

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.

2 Model Building

To best answer our research question, we created 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. Throughout our analysis we will address the 7 policies we list in `Introduction` as the following independent 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 Shutdowns (measured in days the policy was in place)
5. Bar Shutdowns (measured in days the policy was in place)
6. Gym Shutdowns (measured in days the policy was in place)
7. 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 through person-to-person contact within each state. In an effort to capture this, we will also include the population density as an independent

variable. With these variables, we hope to explain which state policies were effective at curtailing the initial spread of the Covid-19 virus.

```
# Variable Transformations
Model_Data=total_data
Model_Data$case.per.100k = (Model_Data$cases/Model_Data$population)*100000
Model_Data$case.log = log(Model_Data$case.per.100k)
Model_Data$population.density.log = log(Model_Data$population.density)
```

2.1 Model 1

First, we began to explore the number of cases in each state per 100,000 residents, which is our dependent variable. This allowed us to account for the number of cases while accounting for the population of each state. For example, we expect New York and California to have a much larger number of cases than Alaska simply due to these states having many more residents. After viewing the histogram of the number of cases per state per 100,000 residents, our team chose to reduce the skew of the data by applying a log transformation to this variable. The histogram of the log transformed variable can be seen in Figure 1.

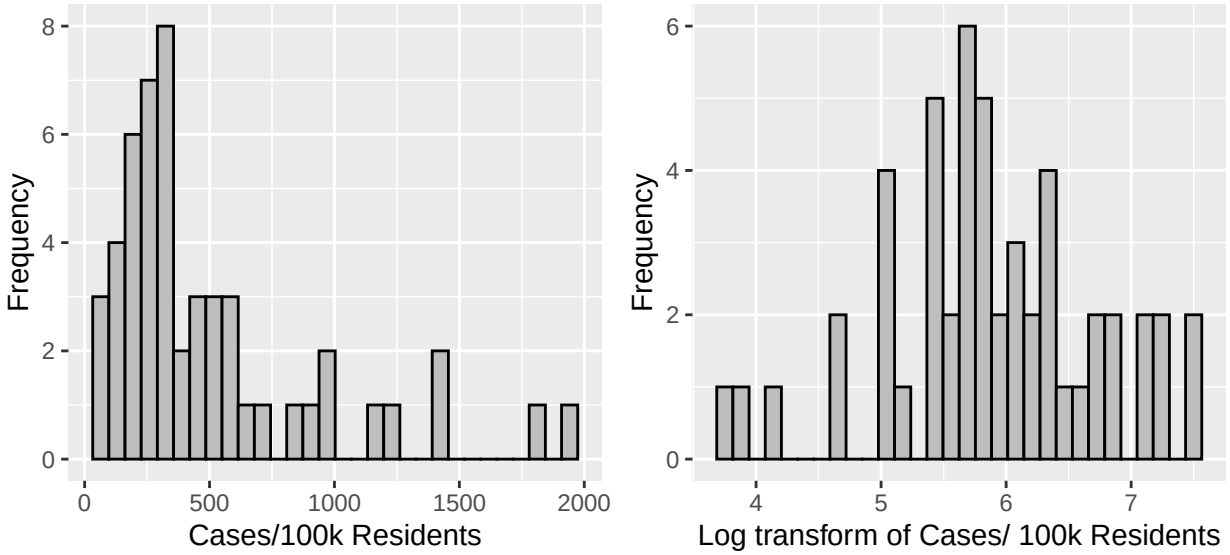

Figure 1: Log Transform of Dependent Variable

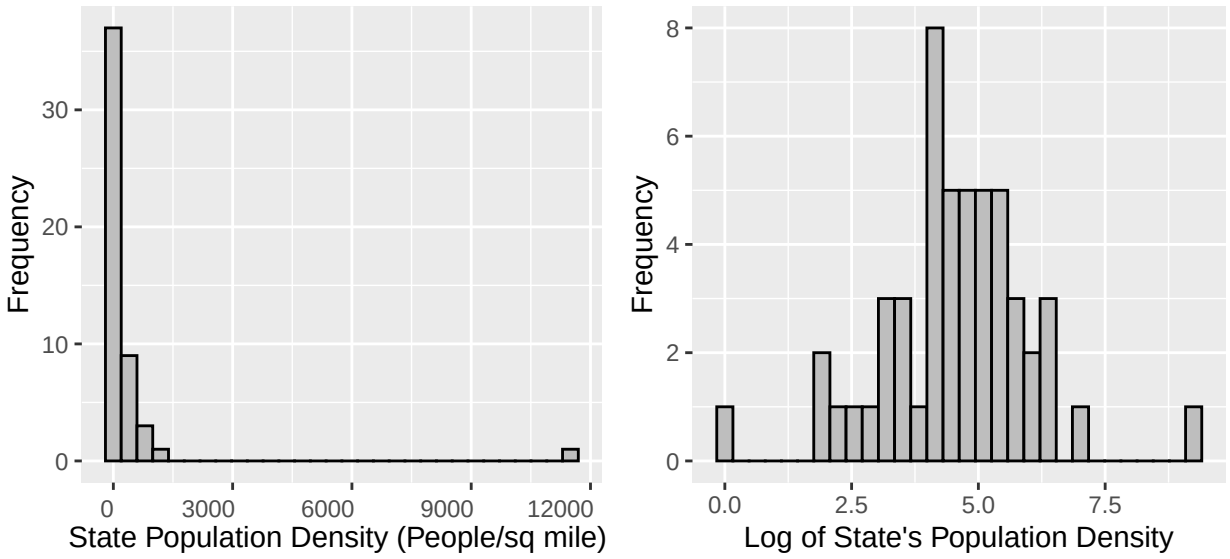

Figure 2: Log Transform of Population Density

For our base model, we selected the one independent variable that was a characteristic of the state and not a policy enacted by the state. We believe a state's population density can be used to represent the largest transmitter of Covid-19, direct person-to-person contact. Therefore, we hypothesize the state's population density is the first building block in describing the number of Covid-19 cases on 05/31/2020. Again, after viewing the initial highly skewed histogram we chose to perform a log transformation on this variable. The transformed histogram can be seen in Figure 2.

Our base model looks at the relationship of the natural log of each state's cases on 05/31/2020 per 100,000 residents with the natural log of state's population density.

```
#Base Model
model1 = lm(case.log~population.density.log, data = Model_Data)
```

2.2 Model 2

For our next model, we added three independent variables representing key state policies to our base model. This medium complexity model will account for the number of days each state had a:

- Stay at home/shelter in place order
- Interstate travel quarantine
- Business mask mandate

The stay at home order was the first and strongest policy set by many states. This was a drastic policy enacted to limit direct contact between people in the hopes that it would slow the spread of Covid-19. We expect a strong relationship between lengthier stay at home orders and curtailed spread. The interstate travel quarantine was another attempt to limit direct contact with visitors from other states (especially those who may have been subject to less restrictive policies). Our final additional independent variable describes the length of each state's face mask mandate for businesses in order to capture the safety measures taken as states attempted to reopen their economies.

The histograms of our three new input variables show distributions which are in line with our expectations, as can be seen in Figure 3. We did notice a spike within all three distributions at zero but we attribute this to states that did not wish to implement this specific policy. Hence our team did not opt to implement any type of variable transformation.

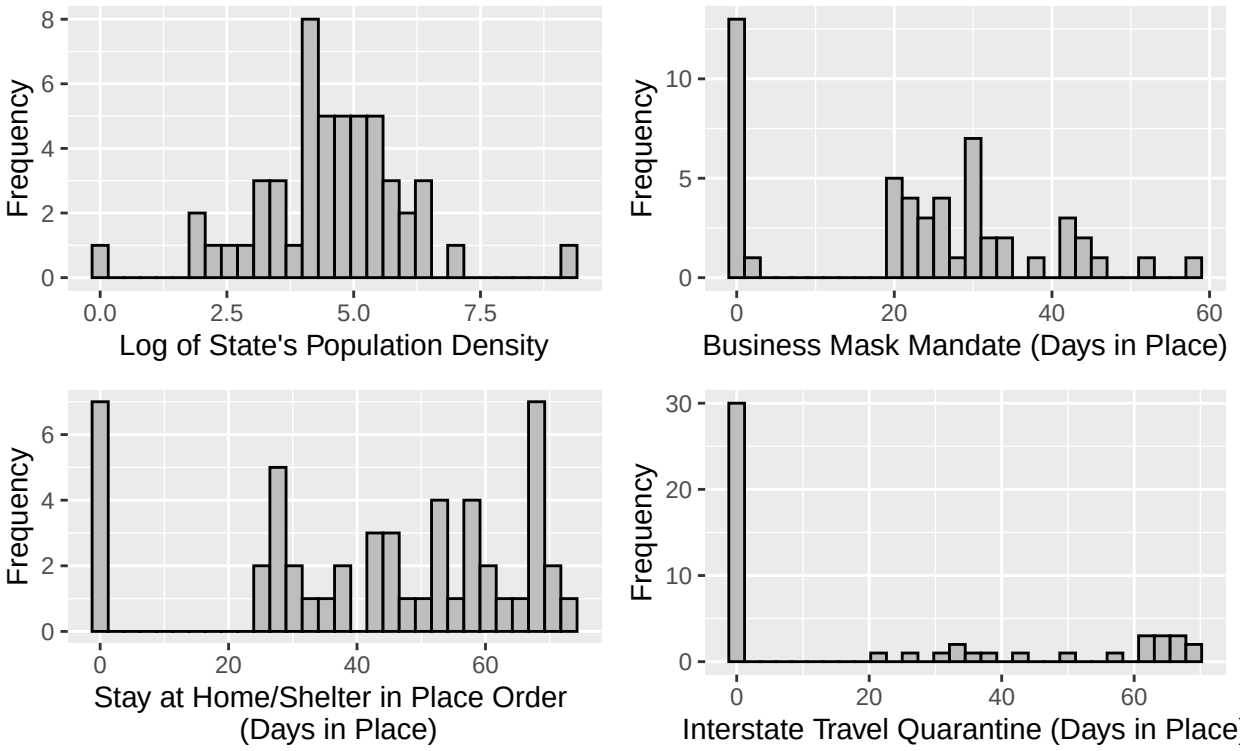

Figure 3: Model 2 Independent Variables

```
#Medium Model
model2 = lm(case.log~population.density.log+
                      maskB.total+
                      stay.home.total+
                      travel.total, data = Model_Data)
```

2.3 Model 3

And finally for our most complex model, we additionally account for the number of days each state had:

- Restaurant Shutdowns
- Bar Shutdowns
- Gym Shutdowns

We have also included a variable to account for the speed each state declared a State of Emergency. This variable is the number of days between the declared State of Emergency and the day the World Health Organization (WHO) declared Covid-19 a pandemic (3/11/2020).

In our final model we included data that further supported the policies covered in our medium complexity model. For example, the shutting down of restaurants, bars, and gyms made it much easier for the public to comply with the stay at home mandate. After all, if all the public spaces within the state are closed, where will people congregate? Unfortunately, this may also lead to multicolinearity between the variables, adding noise to our model if shutdowns occurred simultaneously or if shutdowns occurred at the same time as the stay at home order. We expect a relationship between longer shutdowns and lower infection rates.

We also added an independent 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 number of days between when a state declared a state of emergency and the day the WHO declared Covid-19 a pandemic. 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 officially declared Covid-19 a pandemic. We expect a negative or small number of days (States of Emergency called before or around 3/11/2020) will align with lower infection rates.

We reviewed the histograms of our new input variables, as seen in Figure 4. 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.

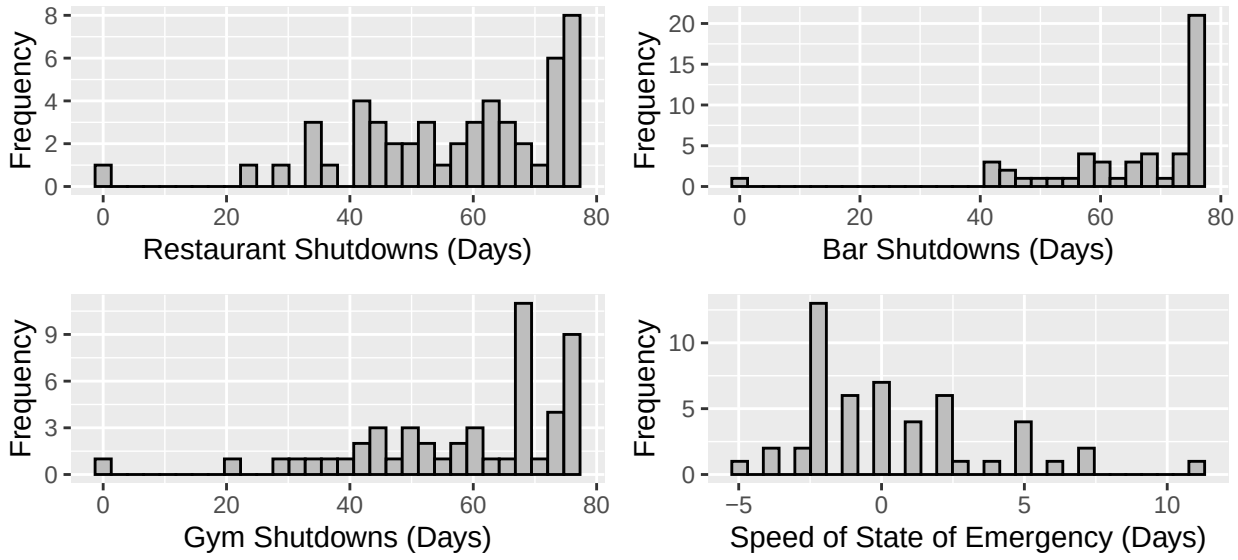

Figure 4: Histograms for Additional Variables in Model 3

```
#Complex Model
model3 = lm(case.log~population.density.log+
                      maskB.total+
                      stay.home.total+
                      travel.total+
                      res1.total+
                      bar1.total+
                      gym1.total+
                      state.of.emergency.speed, data = Model_Data)
```

3 Regression Table

Table 1 displays the regression of all 3 models. We used robust standard error in our calculations. What follows is an analysis of the results of our models.

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 (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

3.1 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 5 displays a scatterplot of the two, and we do see some relationship exists. The coefficient is positive, indicating that a greater population density is related to a larger number of Covid-19 cases per 100,000 people within 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. This translates to have great practical significance when we look at a state-by-state comparison. 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. Furthermore, this model has an R^2 value of 0.39, indicating that the transformed population density may account for a surprising amount of the variance in Log (Covid-19 Cases).

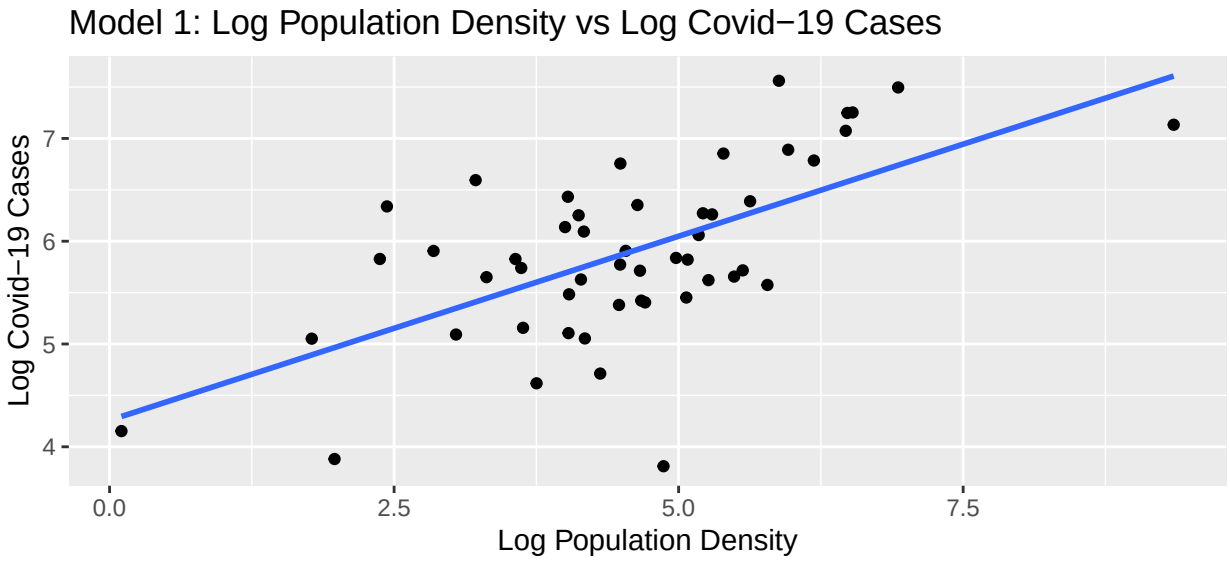

Figure 5: Scatterplot of Model 1

3.2 Model 2

For our second model, we find evidence that there is a relationship between our dependent variable and how long the business face mask mandate was in place (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 variable Log (Population Density). The coefficient for Log (Population Density) decreases slightly 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 has a Stay at home/shelter 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 a relationship exists. This speaks strongly to the efficacy of such orders, and the low p-value of 0.048 gives us reason to believe that there is a relatively low probability this relationship exists by chance.

Inspection of the Interstate Travel Quarantine In Place Length leads to a similar 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 instituted or extended 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. As will be further explained within the `Omitted Variables` section, we believe this may also be attributed to an omitted variable bias.

A one-way analysis of variance (ANOVA) test reveals that the addition of these three variables to the base model results in a statistically significant benefit.

3.3 Model 3

For our last model, we added an additional four extra variables. However, we found that none of the additional variables were statistically significant and that 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 see 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 exists one or more colinear relationships between these two variables and the new variables added. Business mask mandate remains significant and 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.

A one-way analysis of variance (ANOVA) test reveals that the addition of these four variables to the medium complexity model does not result in a statistically significant benefit.

4 Limitations

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

4.1 Independent and identically distributed data

The data used in our models represents the number of Covid-19 cases and state level policy decisions for all 50 states of the United States along with the District of Columbia. We can assume the data sampled from each state/district is identically distributed and independent from one another.

We do acknowledge some concern may exist as to the true independence of each state in 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.

4.2 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.

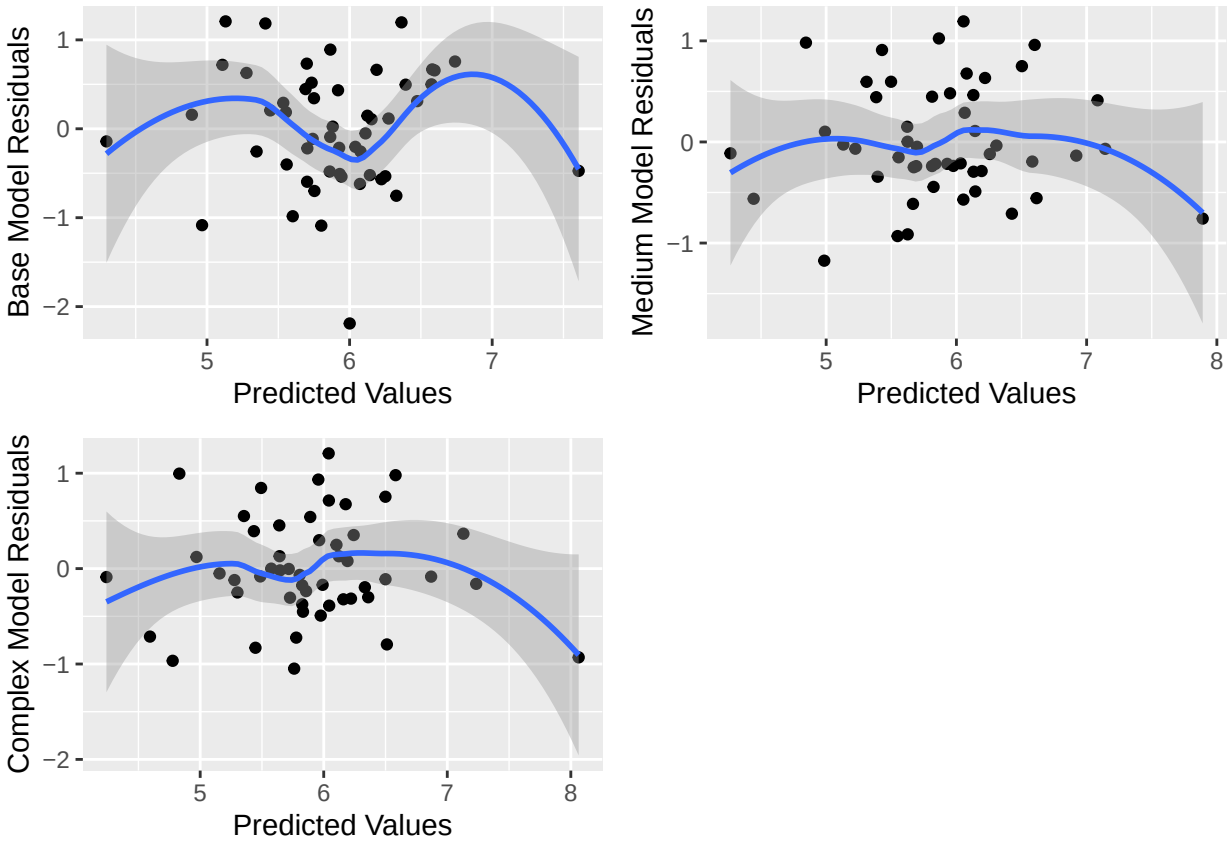

Figure 6: Independent Variable vs Residuals for Models 1,2,3

From the Residual/Predictors plot shown in Figure 6, 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.

4.3 Perfect Collinearity

Our base model cannot possibly have perfect 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).

4.4 Homoskedastic Error

To evaluate the presence of heteroskedastic errors, we will plot the residuals in relationship to the model's independent variables. For our base model, we will plot the base model residuals versus the log transformation of each state's population density. This is shown in Figure 7.

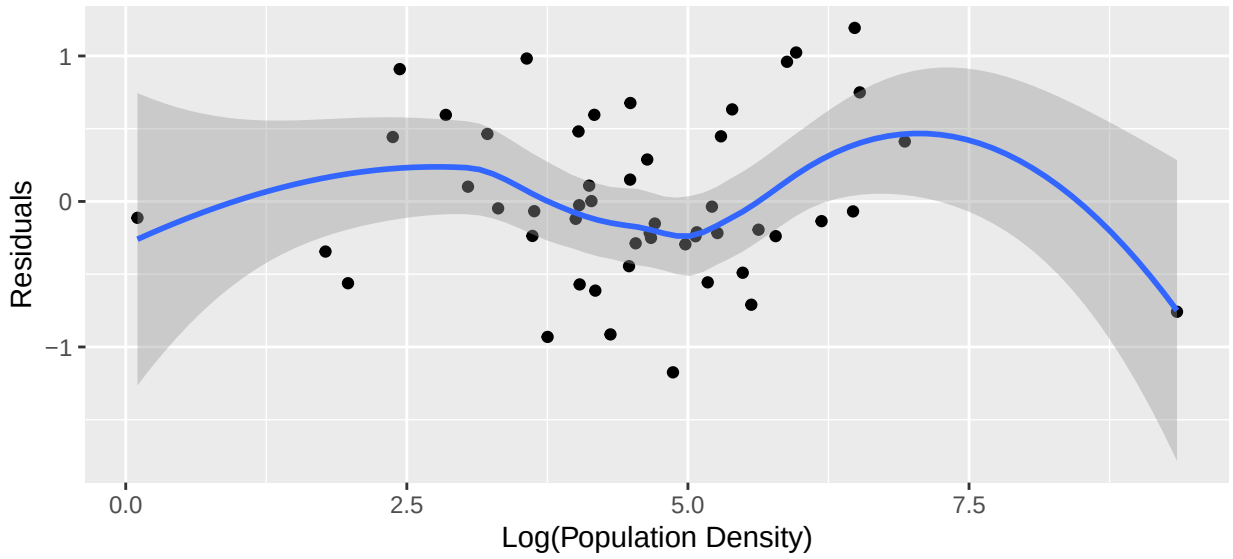

Figure 7: Homoskedastic Error of Model 2

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

For our medium complexity model, we will plot the residuals in relationship to the model's four dependent variables. 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)

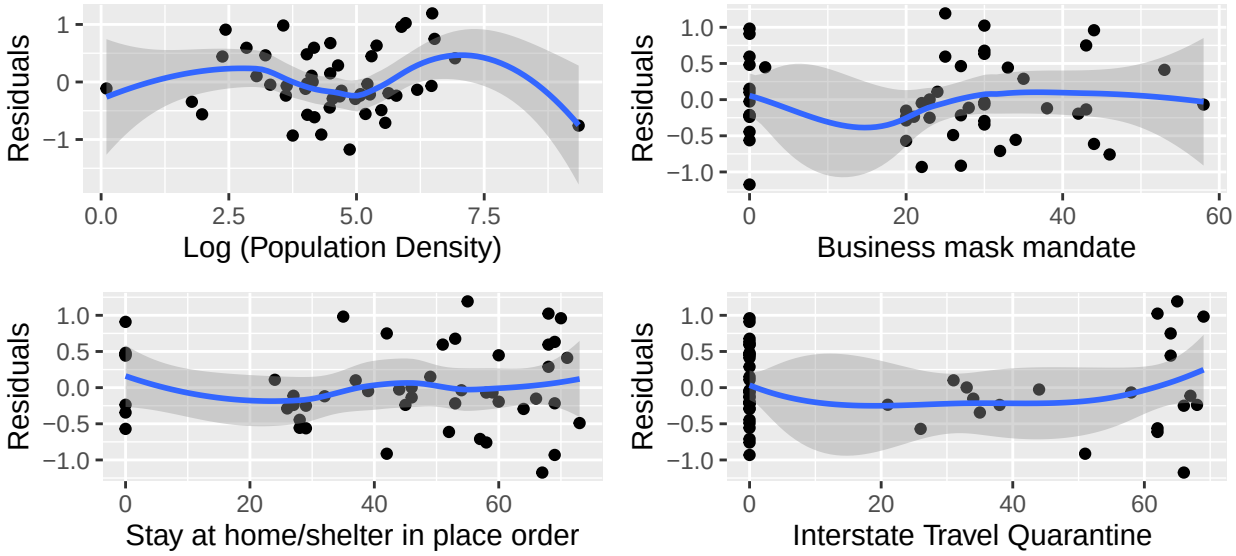

Figure 8: Homoskedastic Plots for Medium Complex Model

The Residual/Input plot for the log of the population density still displays a flaring of the standard error of the residuals at the extremes, as shown in Figure 8. The business mask mandate and interstate travel

quarantine also display inconsistent standard errors. These factors suggest the potential for heteroskedastic 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 affects the complex model (suggesting heteroskedastic error), therefore we will use robust standard error when reviewing the significance level of all model's coefficients.

4.5 Normal Residuals

Lastly we have a histogram of each model's residuals to ensure they are normal. As you can see from the Figure 9, the residuals are in fact normally distributed for all models.

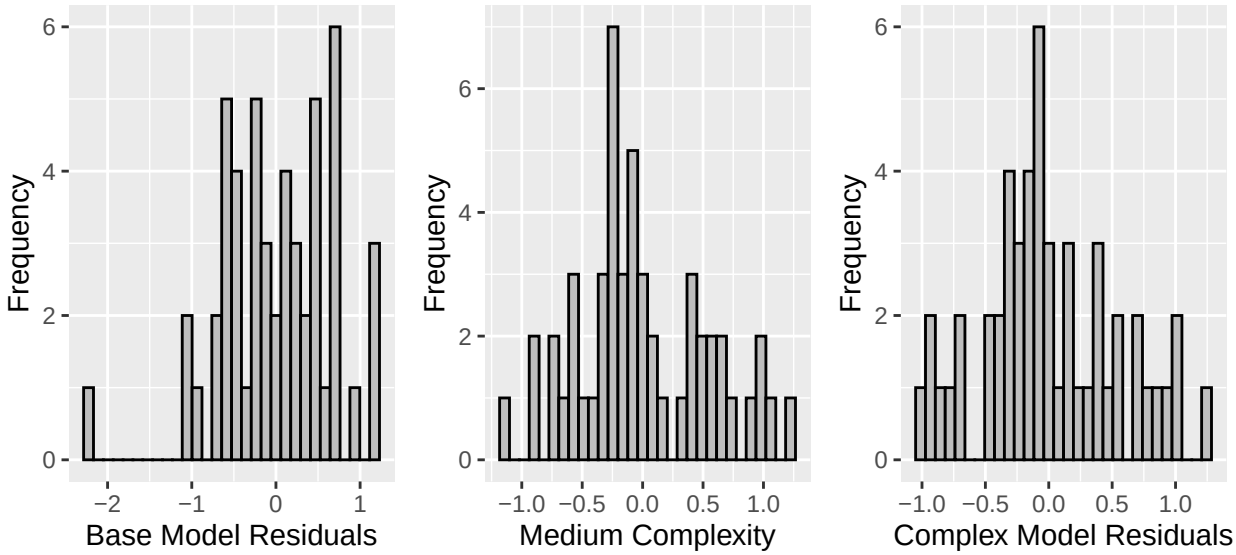

Figure 9: Distribution of Residuals of All 3 Models

4.6 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.

Additionally, our variables capture the length of time a mandate was in place, but not necessarily how effectively the mandate was enforced. For example, it is possible a state implemented an interstate travel quarantine but did not have the required infrastructure to enforce it. This means that we may not be capturing the effects of an interstate travel quarantine, but instead the effect of a state declaring an interstate travel quarantine.

5 Discussion of Omitted Variables

5.1 Omitted variable 1: Number of tourists

The first omitted variable we consider is number of tourists. Through background knowledge, our expectation is that the greater the number of tourists the quicker the spread of Covid-19. States with a greater number of tourists may also be more inclined to have a interstate travel quarantine in place for a longer period of time in order to help contain the spread of Covid-19. While we believe this variable may have an impact

on our model, we did not have access to it as part of the two data sources mentioned in our `Introduction` section and thus have omitted it.

We believe there would be a positive correlation between the number of tourists and our variable `Interstate Travel Quarantine in Place Length`. As mentioned above, we believe the greater the number of tourists a state has the longer it would have an interstate travel quarantine in place. Additionally, we believe the number of tourists would have a positive effect on the model output as a larger number of tourists increases the likelihood of a quicker spread of COVID-19. Since the coefficient for our variable `Interstate Travel Quarantine in Place Length` is negative, we believe the direction of the bias is towards zero.

5.1.1 Omitted variable 2: Number of essential businesses

The second omitted variable we consider are the number of essential businesses within a state. Essential businesses are exempt from statewide stay at home or shelter in place orders and can opt to still have their employees come into offices. Therefore, we believe states with a greater number of essential businesses have a greater potential of a lengthier business face mask mandate. Additionally, due to the fact that there would be a greater number of people who would be unable to stay at home, we expect everyone else who can would be expected to stay at home for a longer period of time. We did not have access to this data as part of the two data sources mentioned in our `Introduction` section and thus have omitted it from our model.

We believe there would be a positive correlation between the number of essential businesses and our variable `Business mask mandates` due to our previous reasoning. Additionally, we believe the number of essential businesses would have a positive effect on the model output as a larger number of essential businesses mean a larger number of the population who cannot stay at or work from home. Since the coefficient of our variable `Business mask mandates` is positive, we believe the direction of the bias is away from zero.

We believe there would also be a positive correlation between the number of essential businesses and our variable `Stay at home/shelter in place order` due to our reasoning above. As mentioned earlier, we also believe that the number of essential businesses would have a positive effect on the model output. Therefore, since the coefficient of our variable `Stay at home/shelter in place order` is negative, we believe the direction of the bias is towards zero.

5.2 Omitted variable 3: Amount of public sanitation stations

Another omitted variable of interest is the amount of public sanitation stations that exist within a state. One of the recommended ways to prevent the spread of COVID-19 is through sanitizing which could be greatly improved through a greater amount of public sanitation stations. We would additionally expect there to be a greater number of public sanitation stations with a more densely populated state.

We believe there would be a positive correlation between the amount of public sanitation stations and our variable `Log(Population Density)` due to our reasoning above. Additionally, we believe the amount of public sanitation stations would have a negative effect on the model output as we believe increased sanitation decreases the rate at which COVID-19 spreads. Therefore, since the coefficient of variable `Log(Population Density)` is positive, we believe the direction of the bias is towards zero.

5.3 Omitted variable 4: Days with cold or freezing temperatures

Our fourth omitted variable we consider are the number of days with cold or freezing temperatures. We believe this omitted variable may have an effect on the bias of the stay at home length. From prior knowledge, we believe that people are more likely to obey the stay at home order on days with cold or freezing temperatures and are more likely to leave their homes on days with warm temperatures. Therefore, we would expect state officials to be more likely to implement or extend stay at home or shelter in place orders should they expect cold or freezing days (for example, a blizzard on a day).

We believe there would be a positive correlation between the number of days with cold or freezing temperatures and our `Stay at home/shelter in place order` variable due to the reasoning mentioned above. Additionally, days with cold or freezing temperatures would force more people to stay indoors and therefore

we believe this omitted variable would have a positive effect on the model itself. Since the coefficient of our `Stay at home/shelter in place order` variable is negative, we believe the direction of our bias is towards zero.

5.4 Omitted variable 5: Number of schools shut down

The final omitted variable we consider are the number of schools shut down. From prior knowledge we know that some schools partially shut down, some completely shut down and went virtual, and some went on a hybrid model where students came in for a partial week. Therefore, though we recognize that we could retrieve the number of schools within a state we chose to omit this variable due to us not being able to confirm the true number of schools fully shut down.

We believe there would be a positive correlation between the number of schools shut down to our variable `Log(Population Density)`. Additionally, we believe this omitted variable would have a negative coefficient within our model. Therefore, we believe the direction of the bias is towards zero.

6 Conclusion

Our research has confirmed our suspicion that population density has an effect on Covid-19 spread ($R^2 = 0.39$). Furthermore, we were able to detect a relationship between key actions taken at the state-level, and Covid-19 spread. We found that Stay at home/shelter in place orders and Interstate Travel Quarantines both led to lower Covid-19 rates (-0.979% per day and -0.888% per day, respectively). However we found that longer Business Mask Mandates were associated with *higher* Covid-19 rates (1.65% increase in cases per 100k people per day). This could mean that longer mask mandates led to higher Covid-19 rates, but it could also be due to reverse-causality bias (states put longer policies in place as a response to higher Covid-19 rates). Moreover, we reason that due to an omitted variable (*Number of essential businesses*) the Business Mask Mandates is likely biased away from zero. Overall, we found interesting results that demonstrate the effect that public health policy had on the transmission of Covid-19.