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

Nathan Chiu, Mikayla Pugel, Joanie Weaver

10/28/2020

1. Introduction

The United States is approaching a total of 15 million confirmed cases of Covid-19 but individual states are contributing to this total at a variety of levels, even taking population into account. Although there are likely many significant factors that have influenced how many cases an individual state is now facing, we were most curious about how an early action that many states took to enact a stay at home / shelter in place order impacted case counts months later. The catch phrase as these policies initially rolled out across the US in March and April was to "flatten the curve", with this report we hope to answer "Did they?".

The shelter in place policy has attracted nationwide attention and drawn its share of proponents and detractors. Proponents argue that shelter in place policies enable residents to avoid unnecessary exposure to the virus while detractors argue that shelter in place is jeopardizing people's economic livelihoods or influencing them to put off medical or other necessary care. Understanding how shelter in place policy impacts the number of cases is all the more important since the number of daily COVID-19 cases has recently exceeded 100,000. Public policy ideally improves the health outcomes for American residents and our research is intended to evaluate the efficacy of shelter in place and other policies designed to curb the outbreak.

In the report that follows, we provide an answer to the research question of how does the (log) length of time elapsed with policy implementation in a given state relate to the coronavirus case rate per 100,000 people. In addition to the length of quarantine policy, we consider additional policies we believe would impact successful implementation of a shelter in place order such as quarantine for individuals entering from out of state, face mask requirements for employees at public facing businesses, and the number of tests run per 100,000 residents. Other dependent variables added to model 3 include children as a percentage of the population and median annual income to account for population demographics we believe are important to include based on the pandemic's contribution to widening the wealth gap and the known observation that fewer children have been sick with Covid-19 CDC. As we started this analysis in early November, we are using data that reflected the current state of Covid-19 in the US on 10/30/2020.

2. A Model Building Process

We are operationalizing the shelter in place policy by computing the number of days that shelter in place lasted for. We subtracted the start date from the end date of the shelter in place policy to find its duration. Length of shelter in place policy is the explanatory variable and the response variable is the Covid-19 case rate per 100,000 people. We also considered making the policies a dummy variable, but for the main model, it wouldn't make as much sense. We chose to use the case rate per 100,000 people because this figure controls for different population sizes of states. Our research operates under the assumption that citizens in a state with shelter in place policies comply with these policies.

The data as a whole offers a plethora of variables to examine and we want to be aware of cross-cutting the dataset. We realize that different variables may be related to one another and that we cannot accurately model the case rate based on one explanatory variable alone. This cross-cutting issue is mitigated by the fact that we are using three different models with increasing numbers of covariates to more properly assess and describe the relationship between shelter in place policies and case rate per 100,000 people. Overall, the data is appropriate for policy research given the sheer number of variables we can analyze. Another consideration

is that the dataset contains only 50 data points, one for each state, which means we have to use small sample linear regression, which has different assumptions.

As mentioned above, our primary research goal was to evaluate the length of time elapsed with the initial shelter in place policy that many states implemented. We plan to test how the length of this policy impacted the cumulative rates of COVID-19 per 100,000 people. Because states vary in the dempgraphics of their populations, initial number of Covid-19 cases, and vary extremely in their choice of policies (or not), we also plan to explore some additional variables to account for these differences and improve our model. Below we will describe how we evaluated this explanatory variable and other relevant covariates in a series of models.

```
## New names:
## * `White % of Total Population` -> `White % of Total Population...14`
## * `Black % of Total Population` -> `Black % of Total Population...16`
## * `Hispanic % of Total Population` -> `Hispanic % of Total Population...18`
## * `Other % of Total Population` -> `Other % of Total Population...20`
## * `White % of Total Population` -> `White % of Total Population...22`
## * ...
```

As described above, our primary goal for this study is to evaluate how the length of the initial shelter in place order that states enacted early in the pandemic impacted the cumulative rate of Covid-19 cases per 100,000 people.

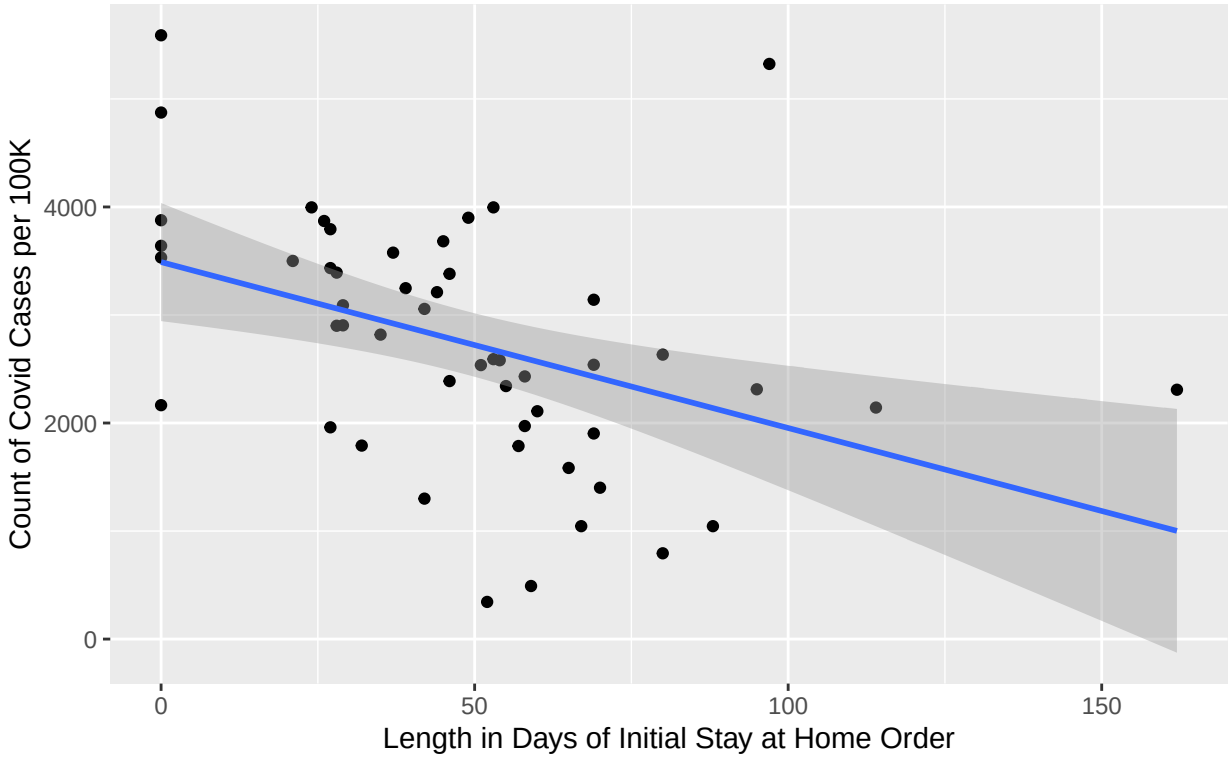
First, we created a new variable length to represent the length of time the shelter in place policy was in place in a given state based on the difference of the date of starting the shelter in place and the date of ending the shelter in place. Following are some assumptions and adjustments we made about the data: - There were 4 states (Connecticut, Kentucky, Oklahoma, and Texas) that issued a shelter in place policy but did not specifically restrict movement of the general public. In these four cases, we made the assumption that the public still treated the shelter in place policy the same way in these states as other states where the policy specifically restricted movement of the general public and calculated length based off of when this formal issue took place. The notes in the policies spreadsheet say that Kentucky's order expired on 6/29/2020, which we update in the end column of the spreadsheet. - The notes in the policies spreadsheet list Utah as an exception. Utah issued a stay at home order on 3/27/2020 saying 'stay at home as much as possible'. We are considering that the public treated this the same way as other states with a restrictive movement policy. Utah removed this on 4/17/2020. - California and New Mexico have not issued end dates for their stay at home orders according to the policies spreadsheet. California issued an updated Blueprint for a Safer Economy in the state with revised criteria for loosening and tightening restrictions on activities on 8/28/20. We take this to be the end of shelter in place for CA. (See: https://covid19.ca.gov/stay-home-except-for-essential-needs/). The policies spreadsheet notes that New Mexico updated their shelter in place policy on 7/16/20, we take this to be the end of the initial policy. - Arkansas, Iowa, Nebraska, North Dakota, South Dakota, and Wyoming never issued stay at home orders so we used length of 0 or a log length of 0 for these states.

Below we will walk you through some of our initial exploratory data analysis as we evaluated different variables (and transformations) to include in our models.

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```

```
## `geom_smooth()` using formula 'y ~ x'
```

Fig 1 − How did the length of the initial stay at home order impact cumulative rate of Covid−19 cases in the US



We can see in Fig 1 above that there appears to be a slight linear relationship between the length of the initial stay at home order and the cumulative rate of Covid-19 cases per 100,000 people in a given state. The longer this order was, the lower the rate. This is what we expected to see.

Let's see if we can improve this relationship with a log transformation. Taking log of the length of time will mean that we'll need to interpret the coefficients as a 1% increase in the length variable increases (or decreases) the covid cases variable by (coefficient/100) units.

```
## `geom_smooth()` using formula 'y ~ x'
```

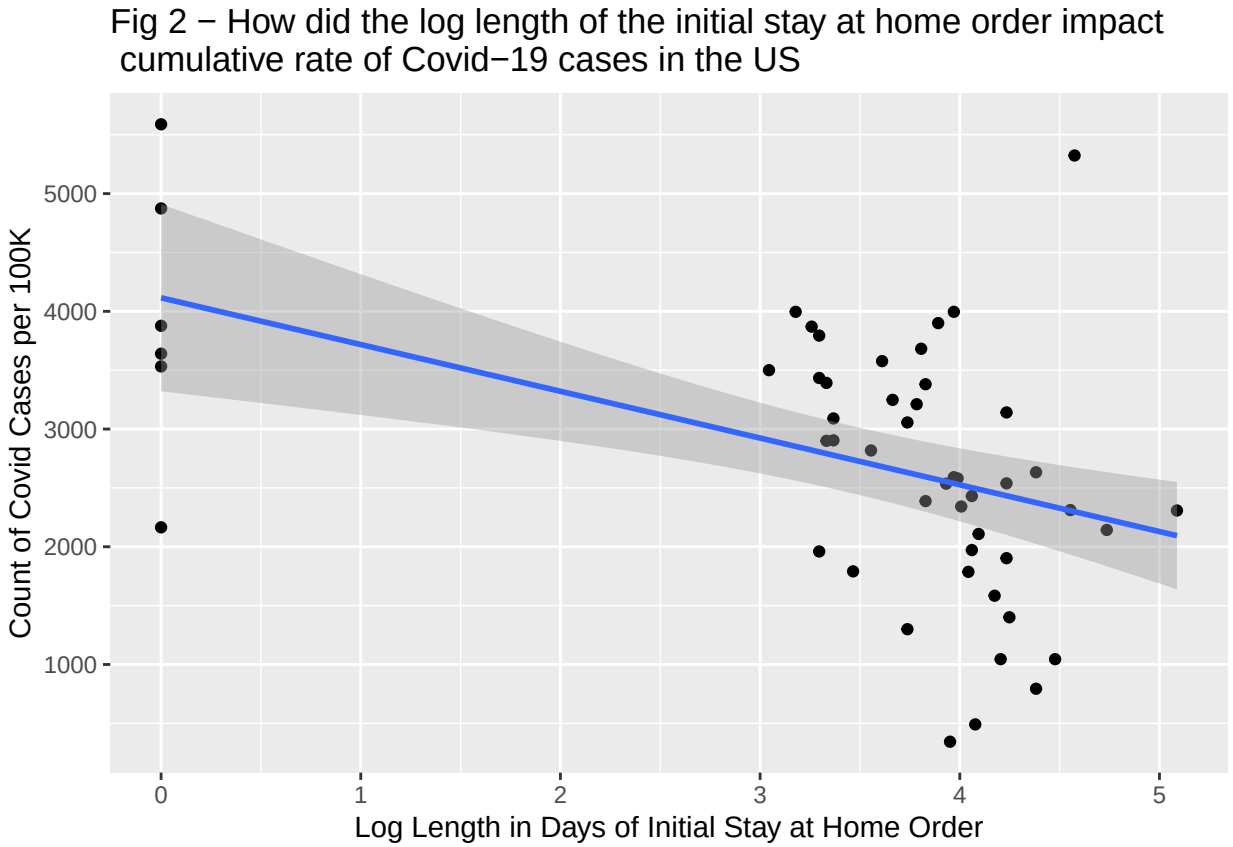


Fig 2 above also appears to have a slight linear relationship. Again, we see that the longer this order was, the lower the rate. Additionally, we can see that the data in Figure 2 is close to the line and less spread apart than in Figure 1, this likely would result in smaller errors. We can also see that the states that did not enact any shelter in place orders and had a length of 0 and log length of 0 are now farther apart from the other states. Two of these states have the highest Covid cases. Based on this analysis, we believe that log length will result in a better model so we will use the log of the initial shelter in place length in our models.

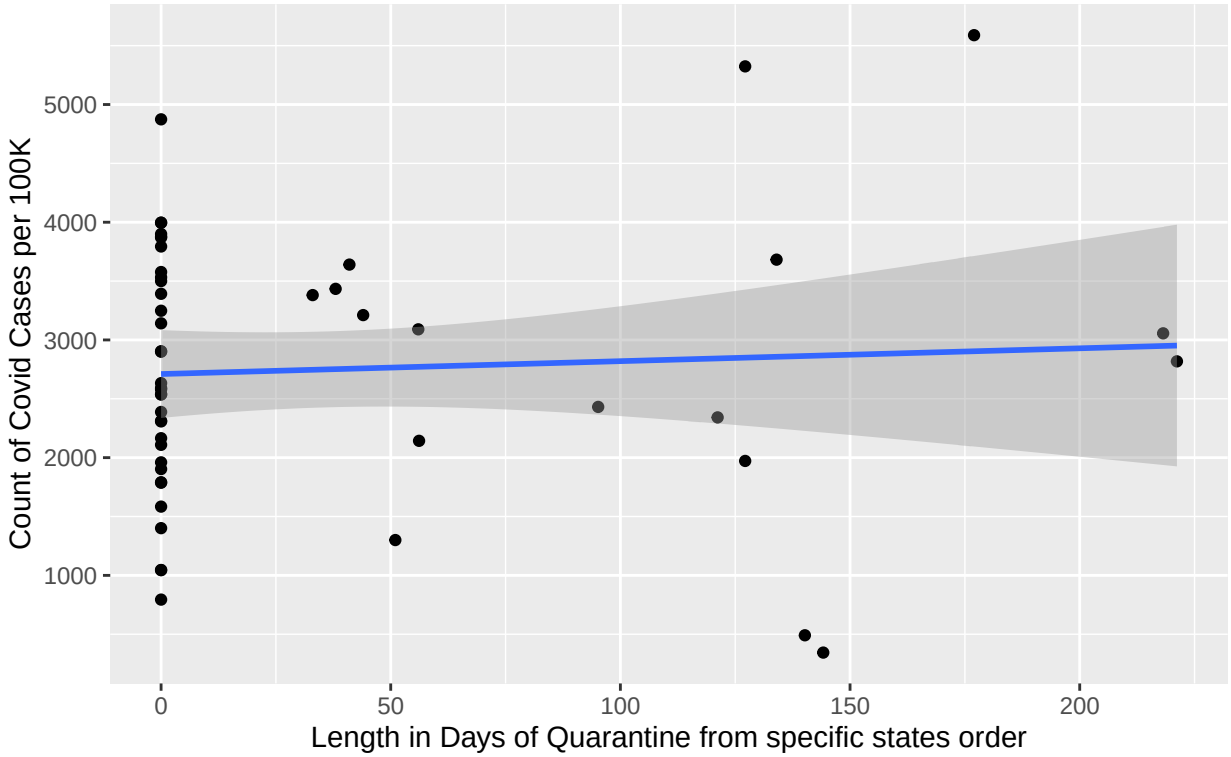
While an initial shelter in place may have improved covid rates in a given state, if individuals then would enter from other states, the state they are entering from may be less restrictive or no policies and thus they may be more likely to bring and spread covid. This is why we also wanted to explore the length of other restrictive policies such as quarantines for visitors or residents returning from out of state.

Below are some of our assumptions about the data for these fields as well as some additional exploratory data analysis about the lengths of these additional policies.

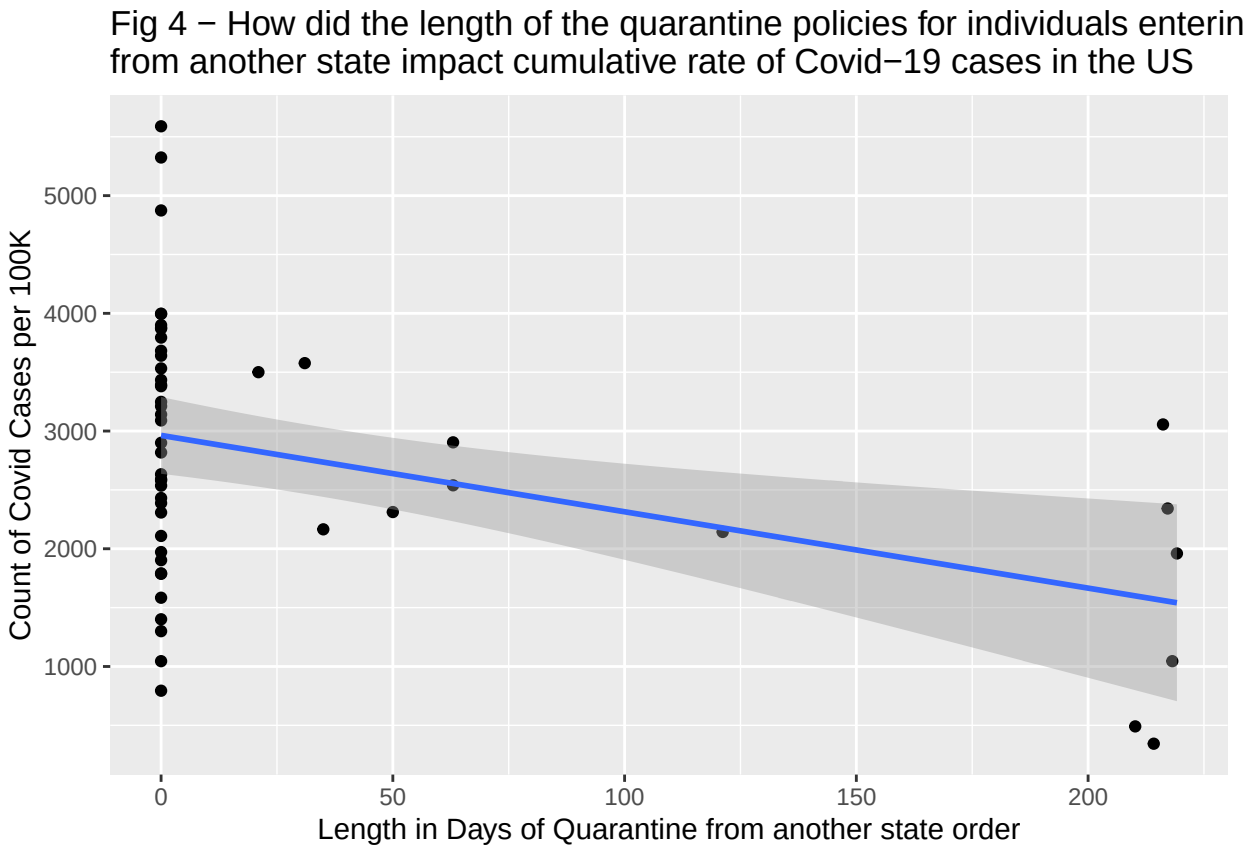
Some states enacted policies requiring quarantine for those entering the state from specific states and/or quarantine for all individuals entering the state from another state. - The policies spreadsheet noted that Connecticut, New Jersey, and New York all implemented a policy mandating quarantine from certain states. We updated the data to use the same date of starting this policy for New Jersey as Connecticut and New York even though review of New Jersey's policy indicated the quarantine was voluntary not mandatory, we are assuming that residents treat it the same as they would in a mandatory state. - The policies spreadsheet noted that Pennsylvania's Department of Health recommended that out-of-state visitors from states with widespread community spread self-quarantine on 4/13/2020 and again on 7/2/2020. We consider the first date when the policy started and make the assumption that the residents treated it the same way as if it were mandatory. - Any states that enacted a quarantine for individuals entering from other states and did not have a date all mandated quarantines ended, were considered to have ended the date of the data collection, 10/30/2020 - States that never enacted voluntary nor mandatory quarantines for individuals entering from out of state were considered to have a length of 0 for these policies.

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
## `geom_smooth()` using formula 'y ~ x'
```

Fig 3 − How did the length of the quarantine policies for individuals enterin specific state impact cumulative rate of Covid−19 cases in the US




```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
## `geom_smooth()` using formula 'y ~ x'
```

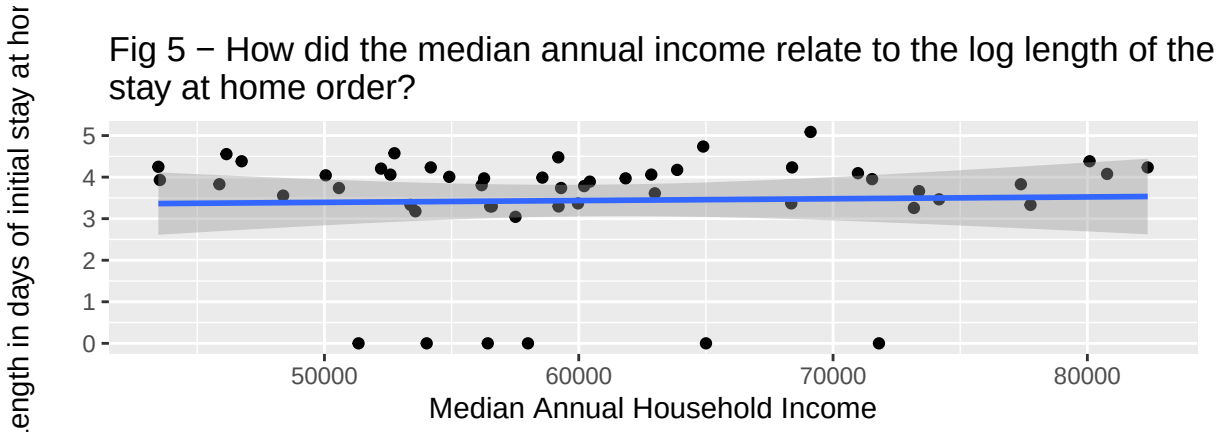


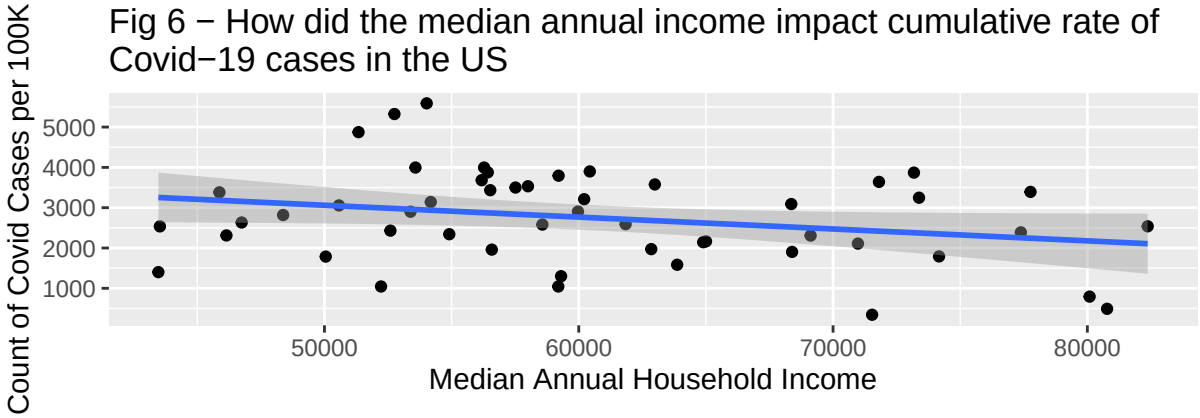
From Figures 3 and 4 above, we can see that most states did not enact any kind of quarantine policy, whether voluntary or mandatory, for individuals entering from out of state (either specific states or any other state). That said, we can see a strong linear relationship between the cumulative rate of Covid-19 cases per 100,000 people in a given state and the length of the quarantine policy for individuals entering from another state that's displayed in Figure 4. This is likely going to be a useful covariate in our model. Figure 3, which shows the length of policies for states who had specific states to quarantine from, does not appear to have as strong of a relationship so we plan to leave this out of our model.

We also wanted to understand how other factors might be relevant to include in the model that might have a relationship with the ability to stay at home and the rate of Covid cases such as income. For example, many people in lower income groups are not able to stay at home as their workplaces were considered essential or lost jobs and needed to go out to find other sources of income or food.

Note that the spreadsheet given to our section was missing a value in the median income column for Wyoming but since the information was referenced from the Kaiser Family Foundation, we updated to use the 2019 median income value for Wyoming found on their site here.

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



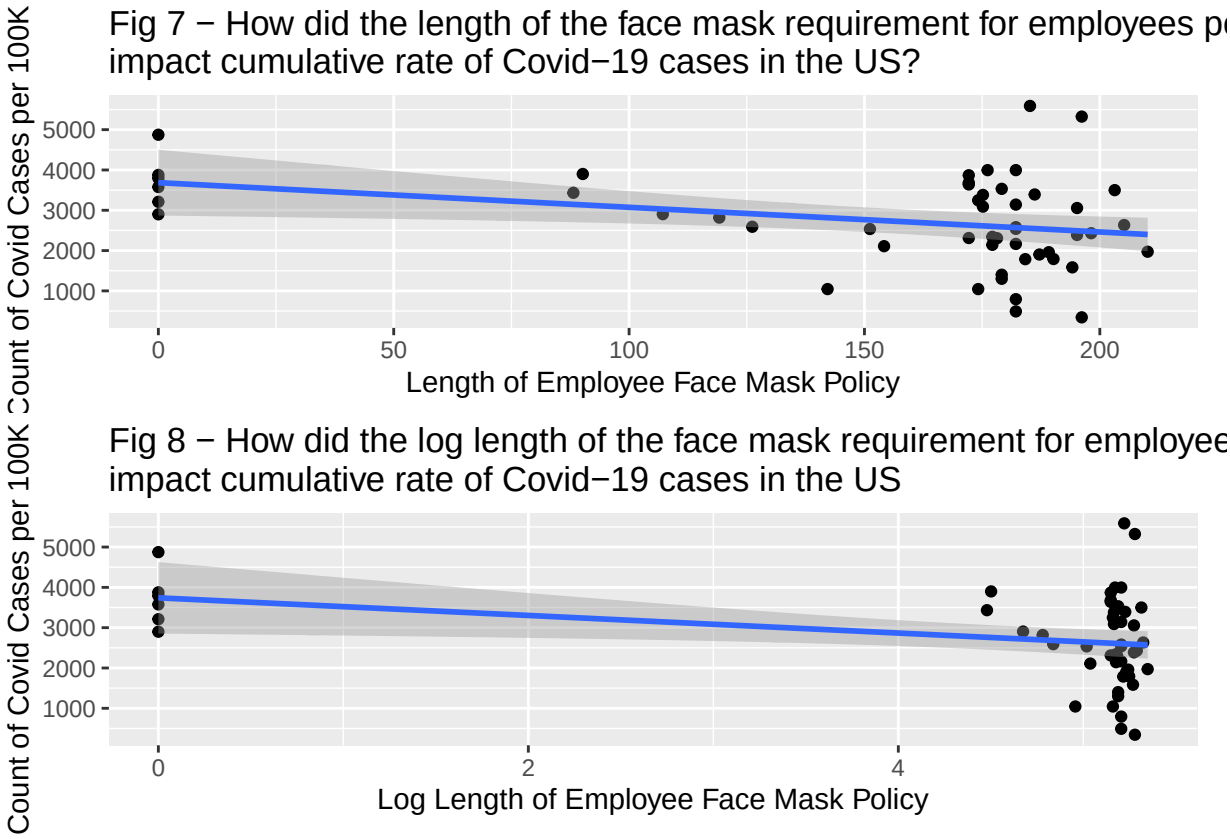


In Figures 5 and 6 above, we can see that there does appear to be a slight positive linear relationship between income and log length and a definite negative linear relationship with the covid rate. If we don't include this variable in the model, we might create a slight omitted variable bias away from zero.

Stay at home orders did not affect many employees who were considered essential who were not able to remain at home so we also wanted to look at the face mask requirement policy for employees. We know that many states also had individual mask policies but that many were not enforced for individuals. We believed that requiring employees to wear masks was more enforceable and thus more followed and would achieve its desired effect to reduce Covid-19 cases.

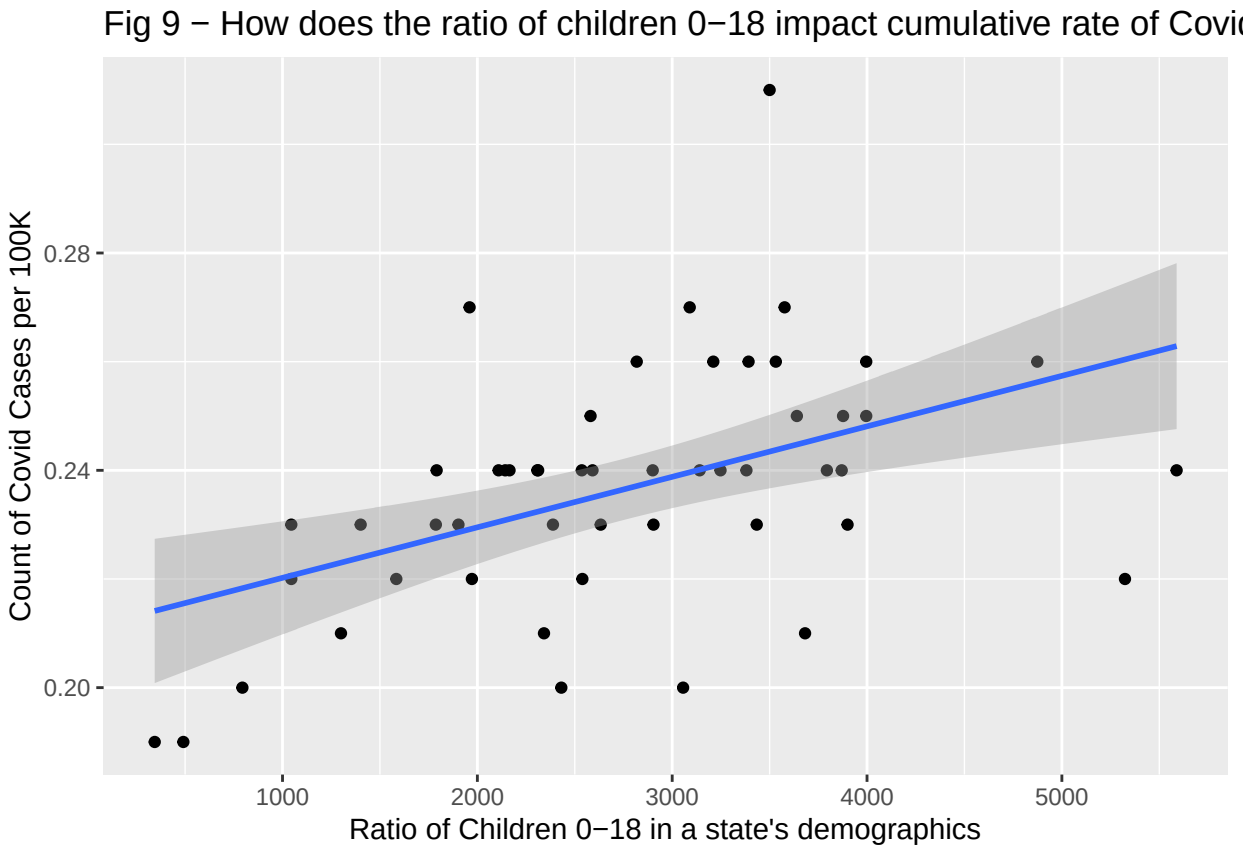
```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



From Figure 7, we can see that the length of the face mask mandate by employees in public-facing businesses has a slight linear relationship and looks to be a better fit than the log of this length because more of the points are closer to the best fit line. This makes sense that there's a relationship because covid cases and mandating employees to wear face masks because covid is airborne and most likely to transfer in indoor spaces which employees working in businesses will be in for many hours with other people.

```
## `geom_smooth()` using formula 'y ~ x'
```



Lastly, it's interesting to note that the ratio of children 0-18 in a state's population appears to have a positive relationship with the cumulative rate of Covid-19 cases per 100,000 people in a given state. Children have not been diagnosed with as many cases of Covid-19 so we initially believed children would have a negative relationship. This could indicate that children comprise many of the asymptomatic cases that don't get caught and thus they continue to spread to others without realizing.

Now that we've explored some likely covariates that include in our models, we will use them in 3 models we build below. Note that we are using the assumption of homoskedasticity in our data to use the classical standard errors instead of robust standard errors.

Model 1 Since our primary research goal is to understand how effective the length of the initial shelter in place policy was, we build our first model solely using the log of this length as the explanatory variable with the cumulative rate of Covid-19 cases per 100,000 people in a given state as the outcome. We are choosing to use only this log length as our sole dependent variable to understand how much of the variation of Covid-19 cases can be explained by that alone. Understanding this may be useful for policy makers in advocating for or against this specific policy for an entire state.

```
##
## Call:
## lm(formula = Cases ~ log_length, data = covid_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2201.74  -556.72    52.82   614.18  3025.88
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4115.0      395.3  10.410 5.22e-14 ***
```

```
## log_length    -397.2      107.3  -3.702 0.000542 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1015 on 49 degrees of freedom
## Multiple R-squared:  0.2186, Adjusted R-squared:  0.2026
## F-statistic: 13.71 on 1 and 49 DF,  p-value: 0.0005423
```

As we can see above the classical standard errors tests find that the log of the length to be statistically significant so that we can conclude that there is a statistically meaningful relationship between the log of length of the initial shelter in place/stay at home order and the cumulative rate of Covid-19 cases per 100,000 people in a given state as the outcome. We can assess the goodness of fit for this model by using the Adjusted R^2 value which is 0.20. That means only 20% of the variation of the Covid cases can be explained by the log length. Additionally, the standard error for log length in this model is 107.3 which represents a number of Covid-19 cases per 100,000 this estimate could be off by. This represents an error of 0.10% of the case rate per 100,000.

While the standard error is very small, the percentage of variation explained by log length is also small so we will explore adding more covariates to our model to help explain more of the variation of the covid rate that log length alone fails to explain.

Model 2 Next we wanted to include some additional covariates in addition to our key explanatory variable, log length of the initial shelter in place/stay at home order. The covariates we want to add to this model are the following: - Quar_Length_Another: The length of time of the policy mandating that individuals entering a state from another state must quarantine. In the first model, we saw a statistically significant negative relationship among the length of time of shelter in place vs. the cumulative rate of Covid-19 cases per 100,000 people. As the length of shelter in place increased, the cumulative rate of Covid-19 cases per 100,000 people decreased. Since states are not vacuums and residents may be moving between states that have different policies and different rates of Covid-19, including these policy variables helps us account for policy decisions that affected individuals entering states with tighter/looser policies from other states with tighter/looser policies. We are not including the length of time of policies mandating quarantine from specific states since our initial EDA did not show a strong relationship. - num_tests_perc: The total number of tests run per 100,000 people in a state has a relationship with the number of positives found. If there are difficulties getting tested, not all of the positives will be found as some people who are asymptomatic may never be tested or some who may not have been sick enough to get hospitalized will never be tested so we think this is important to include in the model to help explain some additional variation among states' covid case rates. - Face_mask_employee: The length of time of the policy mandating that employees in public facing businesses must wear masks. As many public facing businesses are considered essential and these employees are not able to stay at home, we believe adding this variable to our model helps account for people who were unable to stay at home but were protected at work from face masks they and their coworkers wore.

```
##
## Call:
## lm(formula = Cases ~ log_length + Quar_Length_Another + num_tests_perc +
##     Face_mask_employee, data = covid_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1445.03  -573.07    43.49   537.03  1758.23
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.435e+03  4.149e+02   8.279 1.15e-10 ***
## log_length          -2.439e+02  8.544e+01  -2.855  0.00644 **
## Quar_Length_Another -7.916e+00  1.642e+00  -4.822 1.59e-05 ***
```

```
## num_tests_perc       2.808e-02  5.619e-03   4.998 8.85e-06 ***
## Face_mask_employee  -6.119e+00  1.905e+00  -3.213  0.00240 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 761.1 on 46 degrees of freedom
## Multiple R-squared:  0.5874, Adjusted R-squared:  0.5515
## F-statistic: 16.37 on 4 and 46 DF,  p-value: 2.086e-08
```

As we can see above the classical standard errors tests find that length of initial shelter in place/stay at home policy, length of quarantine policy for individuals arriving from another state, the number of tests run per 100,000 people, and the length of the employee face mask mandate have a p value < 0.05 so that we can conclude that there is a statistically meaningful relationship between these variables and the cumulative rate of Covid-19 cases per 100,000 people in a given state as the outcome. We can also see that the Adjusted R^2 value has increased from 0.20 in the first model to 0.55 which means that 55% of the variance of the covid case rates can be explained by the variables in our second model. We can also see that the standard error of the log length has decreased to 85.44. Log length has also decreased in significance from the 0.001 level to 0.01 as it is now 0.006, this is still considered significant to us. We believe it likely decreased in significance as more of the variation is now being explained by other variables in the model.

For the reasons stated above, Model 2 does seem to be better than Model 1 at explaining the variation of the count of the Covid-19 rate but let's check using an F-test.

```
## Analysis of Variance Table
##
## Model 1: Cases ~ log_length
## Model 2: Cases ~ log_length + Quar_Length_Another + num_tests_perc + Face_mask_employee
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     49 50467182
## 2     46 26647808  3  23819373 13.706 1.615e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is really small which means that the additional variables we added to Model 2 from Model 1 do indeed produce a measurable improvement in the performance of the model.

For more details on the coefficients in this model, refer to the regression table section below.

Model 3 Lastly, we will include some covariates in addition to the ones used in the second model to err on the side of inclusion. In our initial EDA, we saw a linear relationship between children and covid cases as well as median annual household income. We want to include these in the model as children and median annual household income may affect the ability to catch cases, social distance or stay at home as discussed earlier in the paper. We still want to evaluate the impact of the log length of the initial stay at home order so we don't want to add too many other additional variables that may decrease the significance of this value.

```
##
## Call:
## lm(formula = Cases ~ log_length + Quar_Length_Another + Face_mask_employee +
##     num_tests_perc + Children + `Median Annual Household Income`,
##     data = covid_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1581.39  -427.65    -6.83   462.03  1412.74
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                         5.763e+01  1.639e+03   0.035  0.97211
## log_length                         -1.801e+02  8.352e+01  -2.156  0.03655 *
## Quar_Length_Another                -6.777e+00  1.578e+00  -4.295 9.48e-05 ***
## Face_mask_employee                 -5.120e+00  1.834e+00  -2.792  0.00772 **
## num_tests_perc                      2.849e-02  5.440e-03   5.238 4.38e-06 ***
## Children                            1.430e+04  5.135e+03   2.785  0.00786 **
## `Median Annual Household Income` -7.171e-03  1.046e-02  -0.686  0.49655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 709.6 on 44 degrees of freedom
## Multiple R-squared:  0.6569, Adjusted R-squared:  0.6101
## F-statistic: 14.04 on 6 and 44 DF,  p-value: 7.484e-09
```

Adding these new variables to our model, we see that log_length, Quar_Length_Another, Face_mask_employee, num_tests_perc, and Children all have p values < 0.05 so that we can conclude that there is a statistically meaningful relationship between these variables and the cumulative rate of Covid-19 cases per 100,000 people in a given state as the outcome. The Median Annual Household Income is not found to have a statistically meaningful relationship with the cumulative rate of Covid-19 cases per 100,000 people. This model has an Adjust R^2 of 0.61 which means that 61% of the variance of the covid case rates can be explained by the variables in our model. We can also see that the standard error of the log length has decreased to 84. Log length has also decreased in significance from the 0.01 level to 0.05 as it is now 0.03, this is still considered significant to us and we believe it likely decreased in significance as more of the variation is now being explained by other variables in the model. This might mean that including the `Median Annual Household Income` might be removing some significance of the log length variable in the model and could be a sign of collinearity. This makes sense as we did see a slight linear relationship earlier in the paper between log length and income. We can hypothesize that higher incomes make it easier to stay at home as many higher income jobs can be done remotely and with higher incomes you can afford to pay delivery fees for many of the goods you may need to leave your house for, making a stay at home policy more effective at reducing Covid case rates.

For the reasons stated above, Model 3 does seem to be little better than Model 2 at explaining the variation of the count of the Covid-19 rate but let's check using an F-test.

```
## Analysis of Variance Table
##
## Model 1: Cases ~ log_length + Quar_Length_Another + num_tests_perc + Face_mask_employee
## Model 2: Cases ~ log_length + Quar_Length_Another + Face_mask_employee +
##     num_tests_perc + Children + `Median Annual Household Income`
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)
## 1     46 26647808
## 2     44 22156460  2   4491348 4.4596 0.01724 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is less than 0.05 which means that the additional variables we added to Model 3 from Model 2 are producing a measurable improvement in the performance of the model, but this p value is 0.02 which is pretty close to 0.05 so it may not be found as significant in future replications.

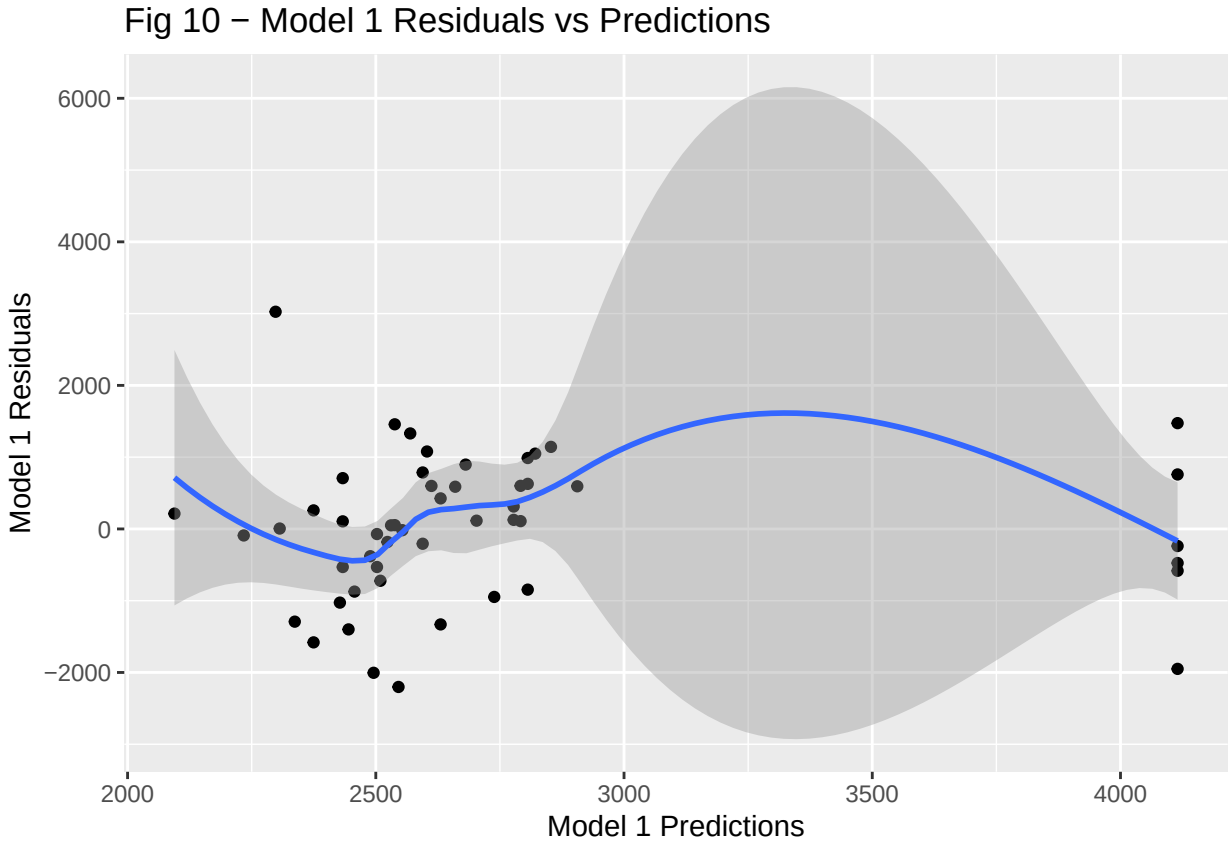
Although the F-test above finds that model 3 is able to explain more of the variance than model 2, it does so with a variable, income, that is not statistically significant to the model and reduces the significance of the log length, our primary explanatory variable in answering the question how does the length of the stay at home policy impact Covid cases. Since the third model only improves the adjusted R^2 from 55% in model 2 to 61% in model 3, we prefer model 2 because it still explains a majority of the variance, while keeping the log length significant and reducing the complexity of interpretation with less variables in the model.

3. Limitations of your Model

While we are not required to discuss in depth the limitations of our model, we did make some assumptions that others may find problematic that we will discuss below.

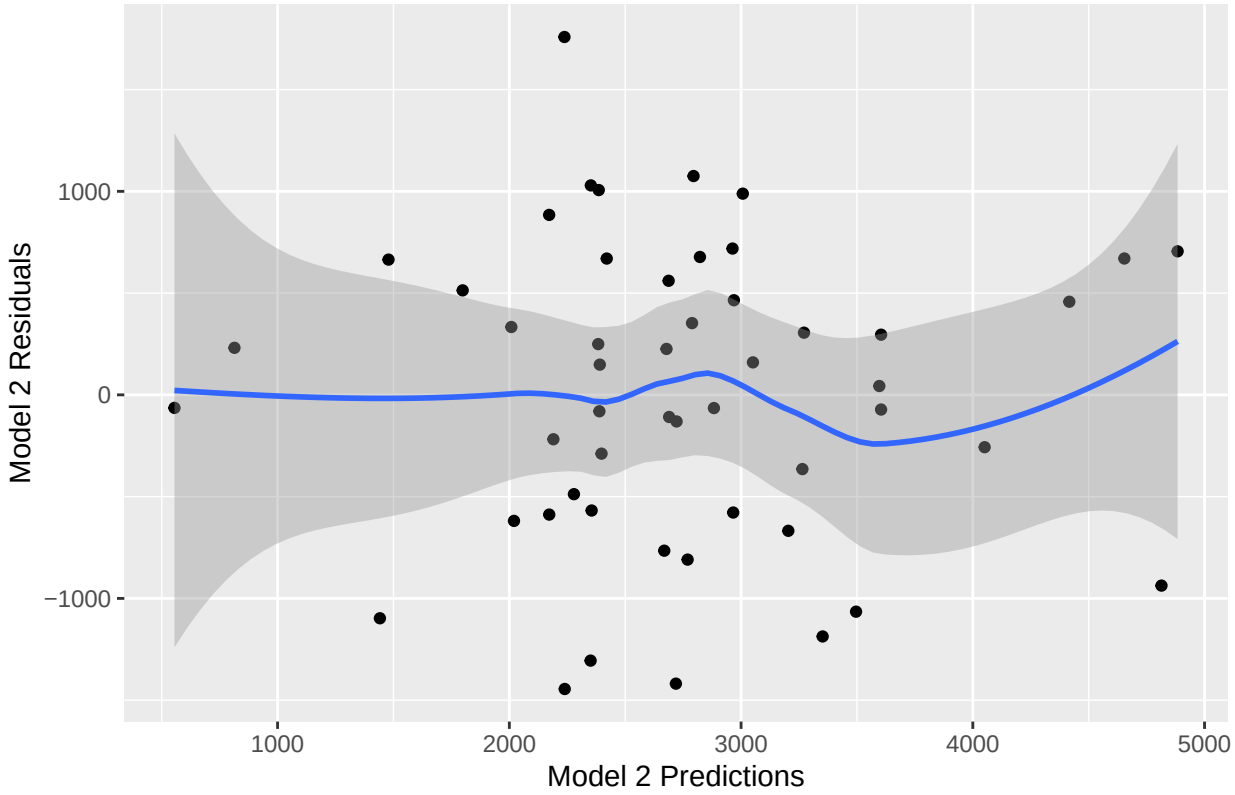
The first assumption is that our data was produced in an IID sampling process. This assumption is evaluated from assessing the design of how the data was collected. In this case, there are many issues with how the data was collected, mainly when it comes to the number of COVID-19 cases. The issue with the count of cases is the different ways that COVID-19 is tested for; some testing procedures are more accurate than others. Also, since the symptoms of COVID-19 vary in prominence in patients, there are many asymptomatic people who never get tested. However, we have assumed that even with these discrepancies, the data is not significantly different than it would be without. Mainly, due to the wide spread tracing and testing of people who come into contact with COVID-19 even if they do not have symptoms and that all of the tests being performed are FDA approved for a minimum amount of certainty. There could also be affects of clustering in the data, as all of the data is split up by state and this could impact the independence of the data. This is because there would be clusters of outbreaks from COVID-19 in certain areas in different cities. However, since we are not completing a statistical study of comparing two groups we can ignore the affects from clustering and assume that all states, since they are based in the United States, are generally similar.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

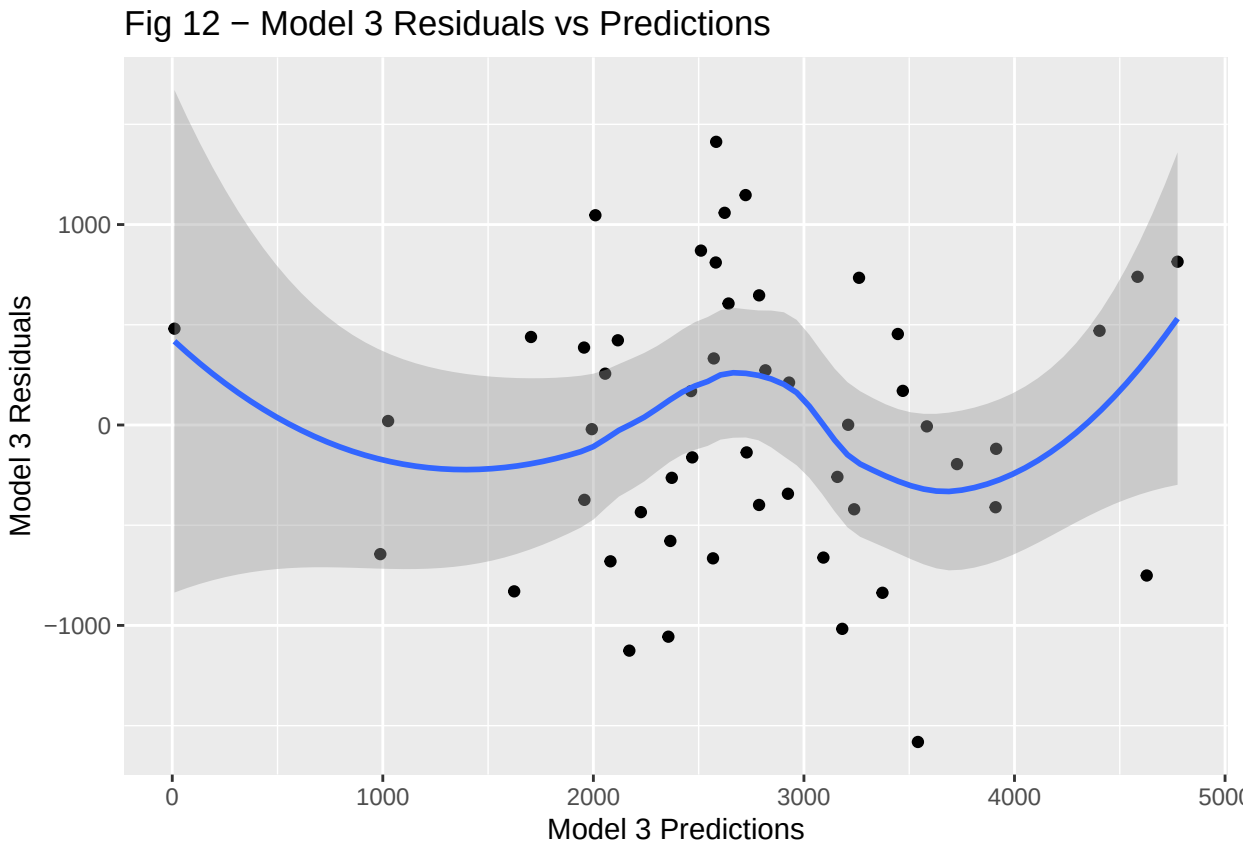



```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Fig 11 − Model 2 Residuals vs Predictions

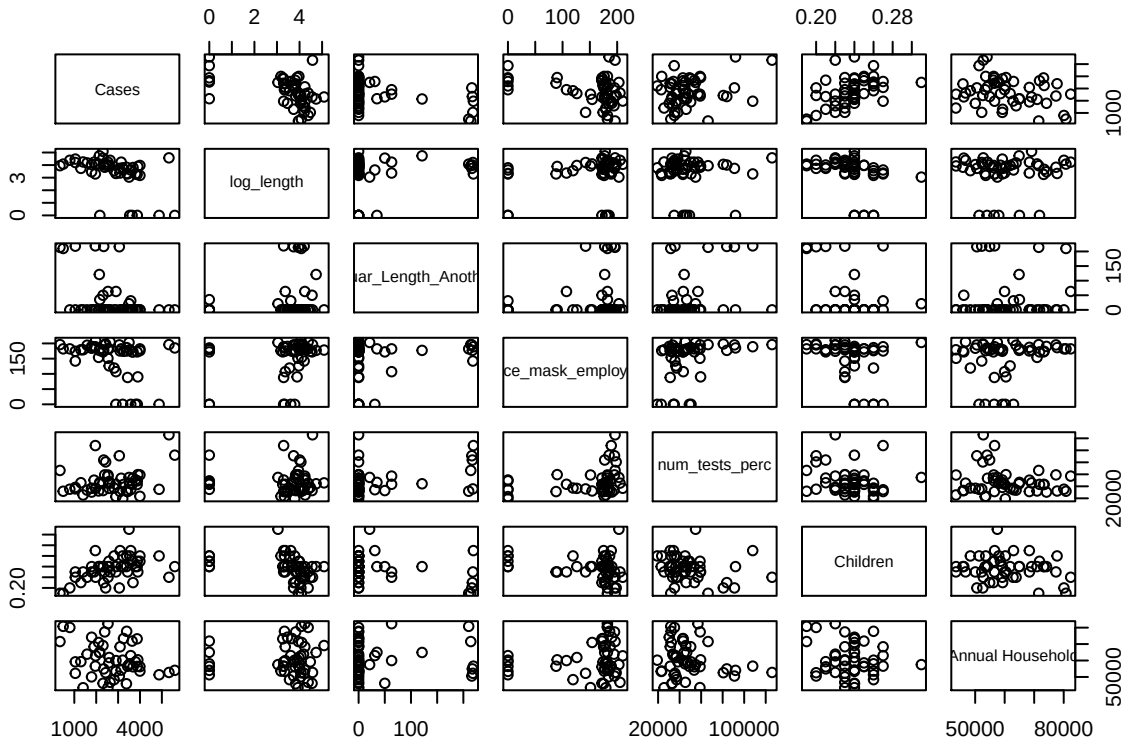



```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The second assumption for the Classical Linear Model to stand true, is the linear conditional expectation. This assumption checks to make sure that the data we are modeling correlate in a general linear fashion. This is important because the structure of the model we are creating is a linear combination of variables. If any variables do not meet the linear conditional expectation, then including them in the model would only pull the model away from the true model. The way to test for this is by comparing the models' residuals with the models' predicted values. A residual is the difference of the predicted value and the actual value for the model. From this graph, we want to see evidence of whether the data is linear. Figures 10-12 above display the results from the residuals from models 1, 2, and 3. Based on these figures, only model 2 satisfies this linear assumption. Model 1 and 3 do not have a general linear relationship and because of this we cannot assume that they are strong predictors. As model 2 is the only model to satisfy this assumption, this is another reason why Model 2 is our preferred model for understanding the relationship between the log length of the initial shelter in place policy and the Covid-19 case rate.

The third assumption is that the model cannot have perfect collinearity. Perfect collinearity is where a variable or a set of variables perfectly describe the output variable. If the model does have perfect collinearity, then the regression will not run or will drop a feature. If a model has nearly perfect collinearity, then the regressions will have large standard errors on collinear features. The graphs below show all the pairs of variables and how they correlate. From looking at the graphs it can be seen that none of the variables are perfect or near perfect collinear. This assumption is satisfied.



The fourth assumption is the model must have constant error variance, which also means that the data is not heteroskedastic. There are a couple things that cause heteroskedastic variance in the data. These include, changes in reporting over time, skewness in the data, changes in variance at different levels, and model misspecification. There are two ways in which we determined this, the ocular test and the statistical test. The ocular test graphs the model's predicted values with the model's residual values to see if the data fans out across the predicted values. The graphs for the ocular test can be seen above in the figures for assumption two. Based on these graphs, there is not an obvious fanning out. The statistical test is the Breusch-Pagan test which predicts a p-value. This test calculates whether the variance of the errors from a regression model is dependent on the values of the independent variables. If this holds true, then the data is heteroskedastic. If the p-value is significantly low, ($p < 0.05$) then the null hypothesis of homoskedasticity is rejected and we can assume the data to be heteroskedastic. In the test below, all the p-values for models 1, 2, and 3 are large, therefore we can assume there is not heteroskedasticity in our models. Since our data is not heteroskedastic, then we can assume that this assumption, constant error variance, is satisfied for all of our linear regression models. This also means that we will be able to use classical standard errors, as we did in our initial model evaluations above, for all of the models.

```
##
##  studentized Breusch-Pagan test
##
## data:  model_unrestricted
## BP = 0.51906, df = 1, p-value = 0.4712
```

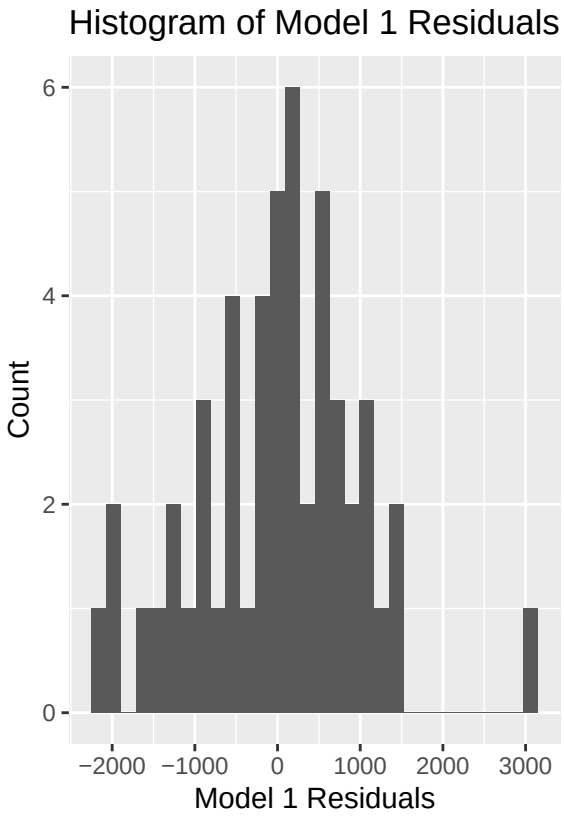
```
##
##  studentized Breusch-Pagan test
##
## data:  model_unrestricted_2
## BP = 5.0134, df = 4, p-value = 0.2859
```

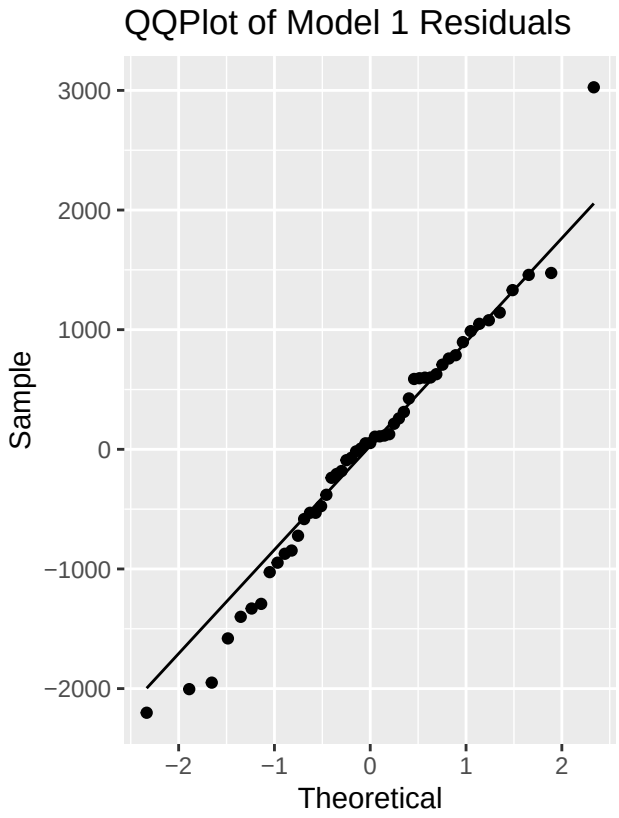
```
##
##  studentized Breusch-Pagan test
##
## data:  model_unrestricted_3
```

```
## BP = 8.8238, df = 6, p-value = 0.1837
```

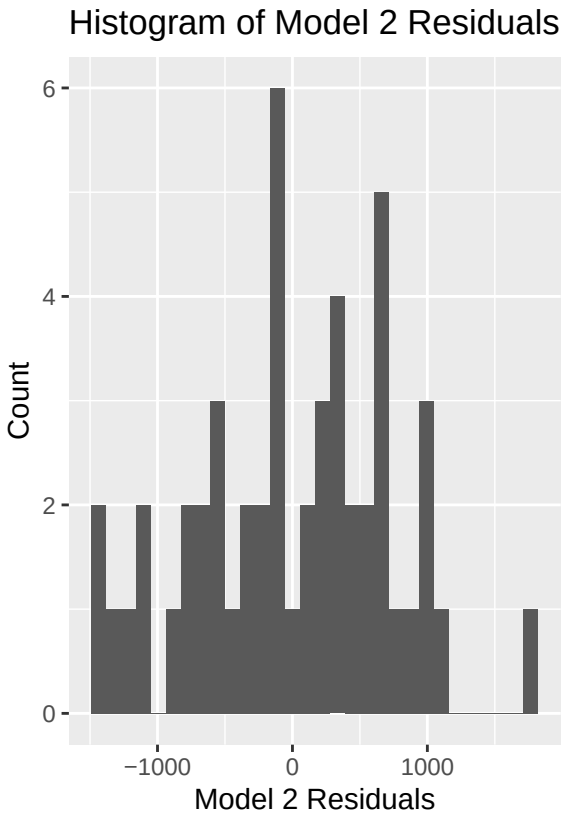
Based on the graphs below, the errors are generally normalized in all of the models. They all deviate slightly, however not enough to not assume normality. If a model's errors were not considered normally distributed, this does not create a problem for unbiasedness, and it is does not affect the standard errors. However, it would threaten the validity of our t-tests and confidence intervals.

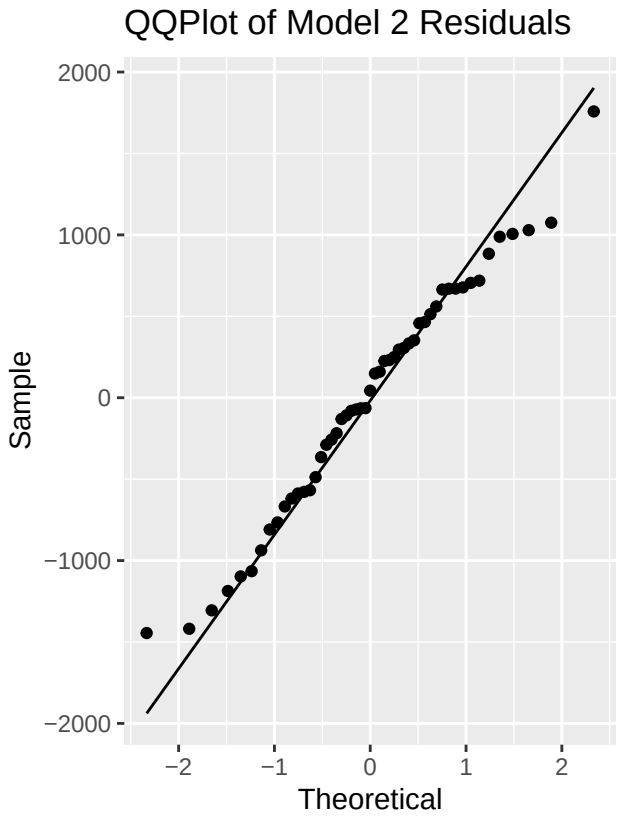
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



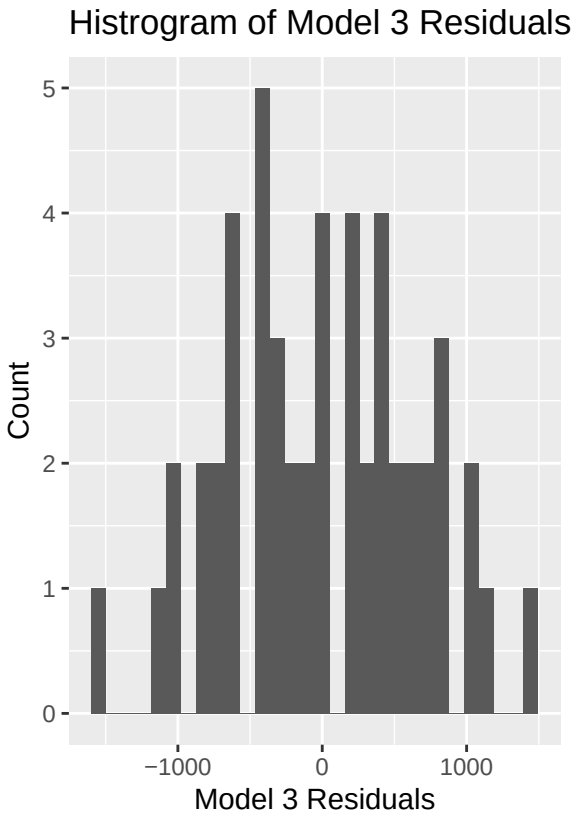


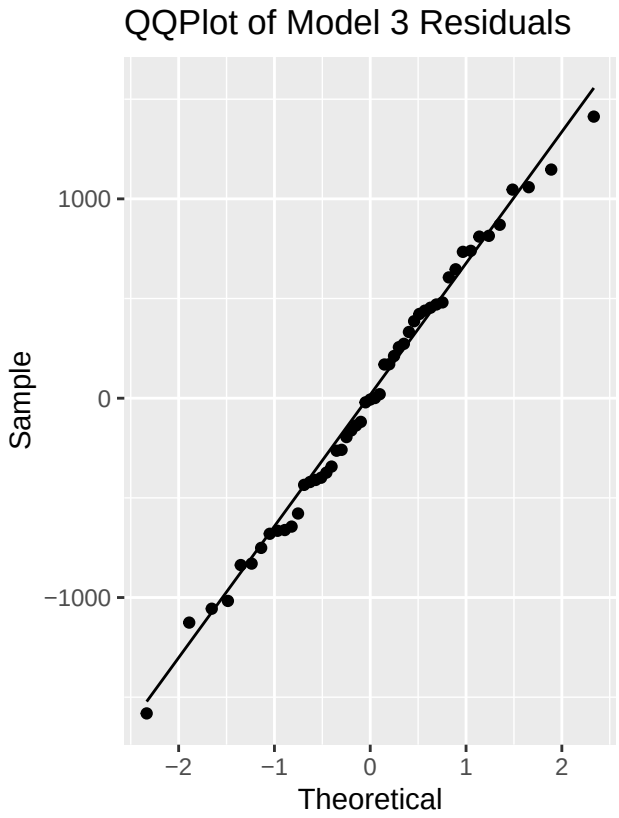

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```






```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





4. A Regression Table

```
##
## Table 1: The relationship between the length of the initial shelter in place/stay at home policy and
## ================================================================================================
##                                                  Dependent variable:
##                                     ---------------------------------------------------
##                                                          Cases
##                                           (1)                (2)                (3)
## ------------------------------------------------------------------------------------------------
## log_length                             -397.164***        -243.913**         -180.110*
##                                         (107.283)          (85.435)           (83.520)
##
## Quar_Length_Another                                        -7.916***          -6.777***
##                                                             (1.642)            (1.578)
##
## num_tests_perc                                             0.028***           0.028***
##                                                             (0.006)            (0.005)
##
## Children                                                                    14,302.200**
##                                                                             (5,134.884)
##
## `Median Annual Household Income`                                              -0.007
##                                                                               (0.010)
##
## Face_mask_employee                                         -6.119**           -5.120**
##                                                             (1.905)            (1.834)
##
## Constant                               4,115.032***       3,434.976***         57.629
##                                         (395.305)          (414.915)         (1,639.117)
##
## ------------------------------------------------------------------------------------------------
## Observations                                51                 51                 51
## R2                                         0.219              0.587              0.657
## Adjusted R2                                0.203              0.552              0.610
## Residual Std. Error                1,014.861 (df = 49) 761.118 (df = 46) 709.617 (df = 44)
## ================================================================================================
## Note:                                                  * p<0.05; ** p<0.01; *** p<0.001
```

Since we were able to accept the null hypothesis that the data is homoskedastic for all of the models in the assumptions section above, we are using classical standard errors in our regression table.

From the regression table above, we can see the following:

Model 1, which is solely based on the relationship of the log of the length of the initial shelter in place/stay at home, has an adjusted R2 of 0.20 which means only 20% of the of the variance of the covid cases is explained by the log length, which is not great. This model estimates the coefficient of log length to be -397.16 with a standard error of 83.52. This means that a 1% increase in the length of the initial shelter in place policy will result in a decrease of 397.16/100=3.97% of the number of cases in the cumulative Covid-19 rate per 100,000 people.

Model 2, which adds to Model 1 variables representing the length of the following policies: mandated quarantine for individuals arriving from another state and mandated quarantine for individuals arriving from specific states. This model has an adjusted R2 of 0.55, which means that 55% of the of the variance of the covid cases is explained by the variables in the model, which is much better than the earlier model. Using the classical standard errors, we can also see that: - an estimate of -243.913 with a standard error of

85.44 is associated with the log length of the shelter in place. This means that a 1% increase in the length of the initial shelter in place policy will result in a decrease of 244/100=2.44% of the number of cases in the cumulative Covid-19 rate per 100,000 people. This is a smaller change than in model 1, which is likely due to more of the variance being explained by other variables we have added to this model. The standard error for this variable is 85.44 which is greater than what we found in the first model. This could be an indication of collinearity. As Covid-19 is airborne can lead to superspreader events, an increase of 2.44% due to policies like this could be significant in battling this disease. - an estimate of -7.91 covid cases per 100,000 is associated with lengthening a quarantine if arriving from another state policy by a single day. This makes sense to us since we believe that setting policies mandating quarantine when entering from out of state help reduce visitors bringing and spreading Covid in the state. A decrease in 7 cases per 100,000 is very small but due to the exponential growth of Covid, this is likely a significant policy to evaluate and keep in the battle against Covid-19. - an estimate of 0.028 covid cases per 100,000 is associated with increasing the number of tests per 100,000 by one test. This makes sense as more tests available make it easier for positive cases to be detected. While 0.028 increases in detection per 1 test is less practically significant, this can also be thought of as an increase of 100 more tests will detect 2.8 more positives. This is significant to account for when evaluating positives in a state that may not have run as many tests. - An estimate of -6.12 covid cases per 100,000 is associated with a one day increase of the policy mandating employees in public facing businesses. A decrease in 6 cases per 100,000 is very small but due to the exponential growth of Covid, this is likely a significant policy to evaluate and keep in the battle against Covid-19.

Model 3, which adds to Model 2 variables representing the ratio of children in a state's age demographics and the median household income the state, has an adjusted R2 of 0.61, which means that 61% of the of the variance of the covid cases is explained by the variables in the model, which is pretty good. Using the classical standard errors, we can see that: - an estimate of -180 with a standard error of 85.44 is associated with the log length of the shelter in place. This means that a 1% increase in the length of the initial shelter in place policy will result in a decrease of 180/100=1.8% of the number of cases in the cumulative Covid-19 rate per 100,000 people. - an estimate of -6.78 covid cases per 100,000 is associated with lengthening a quarantine if arriving from another state policy by a single day. This is a smaller change than in model 2. - an estimate of 0.028 covid cases per 100,000 is associated with increasing the number of tests per 100,000 by one test. This is the same change as in model 2. - An estimate of -6.12 covid cases per 100,000 is associated with a one day increase of the policy mandating employees in public facing businesses. A decrease in 6 cases per 100,000 is very small but due to the exponential growth of Covid, this is likely a significant policy to evaluate and keep in the battle against Covid-19. - An estimate of 14,302.20 is associated with an increase in children by 1. Since children is a ratio of the number of people under 18 in the state's total age demographics, it is more practical to think of this as an increase of the children ratio from 0.20 to 0.21 can result in an increase of 143 more covid cases per 100,000 people with an error of 5. This is a large amount of cases and thus is significant for future investigation on why exactly this may be the case. - Median Annual Household Income is not found to be statistically significant in this model with the current data.

Some other observations we have made: - The coefficient for log_length and significance with each addition of other variables in the model, although it is still significant at the 0.05 level in model 3. This is likely due to the fact that many of the variables we have added to models 2 and 3 helped explain more of the variance of the covid cases and may have a relationship with the length of the stay at home or capability of people to stay at home. - The coefficient for the length of the policy mandating quarantine from another state also decreases its significance from model 2 to 3, indicating the additional variables added to the model 3 of children and income may affect how well a quarantine policy from another state works. - The coefficient for the number of tests run per 100,000 has the same coefficient in both models 2 and 3. This is likely due to the strong relationship this variable has with the number of positive cases found. - The coefficient for the constant significantly drops in the third model. It is also not statistically significant in this model. The constant is the most significant in the 2nd model when it also is found to have the lowest standard error.

As we've alluded to in a few sections above, since model 2 is the only model to meet all 5 of our CLM assumptions and is able to explain a majority of the variance of the Covid cases without creating large standard errors or losing statistical significance for our main explanatory variable, we would recommend using this model for future experimentation and evaluation. With less variables and the fact that all variables

are statistically significant also makes it easier to interpret when explaining to policy makers about potential actions to make that can reduce Covid cases in a state.

5. Discussion of Omitted Variables

We have chosen a descriptive model, but it is still worth mentioning the potential effects of omitted variables on our model as some of the variance from the omitted variables will get added to other variables standard errors in the model. Below is a list of potential omitted variables and the estimated direction of the bias they have on our current model coefficients:

- Percentage of people who follow each policy. We know that states make policies and even make them mandatory, but they are not enforced or followed by everyone and the proportion of people who do not follow greatly varies by state. It would be helpful to the model if each policy had an associated percentage of the people who follow out of all people in the state. The bias for each state variable would be different depending on the overall covid case count in the state. An increase in the ratio of people who follow each policy would have a negative relationship with the covid case rate and thus we'd see our coefficient estimates for the policies get closer to 0.
- Number of individuals travelling to the state from each other state. Not all states have individuals visit from out of state and some have more individuals visit from certain states than others. Since we include policy information on mandated quarantines for individuals visiting from out of state and we know not everyone follows policies, it would be useful to know how many people are visiting the state of interest from every other state. The bias for each state variable would be different depending on the overall covid case count in the state and thus it is difficult to estimate whether it would be positive or negative.
- Ratio of the population of a state living in cities with restrictive policies. While not all states implemented the policies we discuss above or implemented them for as long, some cities or counties within the states had their own more restrictive policies which would have impacted residents in those areas and possibly nearby areas to shelter in place or avoid certain businesses that were closed. Keeping this information out of our model results in coefficients that are towards zero as we are missing sets of relationships that would decrease the number of covid cases.
- Economic demographics such as % of people who are considered essential workers, % of people who are not able to work from home would help us understand more about how many people in a state are not able to follow a stay at home order and its benefits would not affect. These variables would have a positive impact with Covid cases indicating our current coefficient estimates for log_length would be farther away from zero.

6. Conclusion

We began this study to assess the effects of the (log) length of the shelter in place policy against the rate of coronarius cases per 100,000 in a given state. As such, our original question was: How does the log of the length of a state's shelter in place policy impact the number of cases per 100,000 people in that state? Model 1 describes the relationship between the number of cases per 100,000 people against the log of the length of the shelter in place policy. We found that the relationship between length of the shelter in place policy is a statistically significant factor at alpha = .01. While the R squared was lower than expected, we knew that the descriptive model would benefit from the addition of other variables.

Model 2 includes additional variables such as the length of time that a state has implemented a policy mandating that individuals entering from out of state must quarantine, face masks for employees, and number of tests per 100,000 residents. We think these factors are relevant and useful to include because essential businesses continue to operate and can contribute to the spread of the coronavirus. Mandating out of state travellers quarantine could also help curb the spread as with better diagnostics in the form of tests. All the dependent variables are statistically significant at the alpha = .05 level and the total R squared is .55. Given the balance between R squared and the number of terms involved, we believe model 2 is the ideal descriptive model to use.

Model 3 includes demographic variables like children and annual median household income. All variables are statistically significant at alpha = .05 except for annual median household income. The model has a high

R squared of .6569 especially compared to model 1, but there is still room for improvement. The results suggest that states with stricter travel and lockdown policies benefit from decreased case rates. At the same time, children and the number of cases bear a positive relationship with case rate. This has a number of implications. This could suggest that states with a higher proportion of children could have higher rates of infection due to children spreading Covid asymptomatically and undetected, potentially from going to school, day care, or hanging out with friends/neighbors unsupervised if schools are closed. Also the relationship between greater testing and higher case rates can be explained by the fact that states with more test kits can test more residents whereas states with less testing have to rely on hospital data to track coronavirus cases. Further analysis would be required to evaluate these assertions.

These results are relevant for policy makers who are now seeing the highest rates in the US and may be considering another shelter in place (like California has just announced) or other actions. The three models we have produced below show that the majority of the variance in the number of coronavirus cases can be explained by the aforementioned factors. In particular, mandating that long distance travellers quarantine has a strong impact on lowering coronavirus cases. While Americans may be anxious to get traveling again, people traveling across state boundaries should try to quarantine for two weeks. Restricting travel and exposure via open spaces limits the transmission of the virus, which our regressions support. Moreover, there is a positive relationship between COVID-19 cases and the proportion of children in a state. This may point policymakers into looking into school closures as a way of mitigating the spread of the virus or requiring more testing of children who may be carrying the virus asymptomatically. As winter approaches and the virus resurges, politicians and state officials should consider opting for all remote classes. Several states have already decided to transition from hybrid to fully online learning to hinder the reemergence of coronavirus. While the lockdown adversely affects business owners, the data points toward continued lockdowns to lower the number of coronavirus cases.