A Regression Analysis on COVID-19 and States Political Affiliation

W203: Statistics for Data Science

By: Sean Lo

Introduction

Beginning around 2019, the world faced a pandemic, from the coronavirus (Covid-19), that would claim the lives of millions of people. For many, this was the first time experiencing such a devastating event. This virus was highly infectious and spread at rapid rates. Governments were then looked upon to figure out how to control the spread, to save as many lives as possible. In the United States, this came in the form of both Federal and State level mandates, like the implementation of shelter in place, business closures, and mask requirements. However, people had many different opinions on the best way of controlling Covid-19, especially between political parties, leading to states implementing different guidelines at various times. In the United States, two political parties, the Democratic Party and the Republican party, usually control state governments. The party in control of a particular state's legislature dictated the guidelines to enact. In this study, we analyzed whether a state's controlling party had any impact on the spread of the virus, by looking at the state's number of infections per capita. By being able to draw conclusions on these effects, we can better prepare ourselves against future pandemics and potentially save lives.

Research question: Does the party affiliation of the state legislature affect the Covid-19 infection rate of a state?

H_0: A state's party affiliation does not have a significant effect on the Covid-19 infection rate per capita.

H_a: A state's party affiliation has a significant effect on the Covid-19 infection rate per capita.

Data Exploration

The data set used in this analysis includes the following sources.

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. The following information was applicable to the question which is the focus of this paper:

- **State characteristics**: population density and population of each state
- **State of emergency**: how many days the emergency was declared in each state[1].
- **Stay at home**: how many days the stay at home order was declared in each state[2].
- **Face masks**: how many days the face mask mandate was in place in each state[3].

[1] March 17,2021 was assumed to be the last date of data refresh and chosen as the cut off point for states that did not have an end date the declaration

[2] Same as above. In addition CT, KY, OK and TX did not have any stay at home orders.

[3] March 16,2021 was assumed to be the last date of data refresh and chosen as the cut off point for states that did not have an end date for the face mask mandate. Montana did not have a face mask mandate (only implemented in a few counties).

- **Closures and Reopenings**: data tracking the duration of closure for restaurants, movie theaters and gyms was tracked. There were at most 3 waves of closures of facilities. In this case the data for each wave was added up and used in the models. The following data for closures and reopenings was not used:
    - Daycares, schools and nursing homes were not tracked as they were closed for every state. Data did not provide any differentiators among the states.
    - Bars, hair salons and casino closures lacked completeness across states and hence were not used.

New York Times Covid Tracking Data

Data tracking covid cases across every state in USA was obtained from their github[4]. This data set provided the number of covid cases by state (prediction variable) for our regression analysis.

State Legislature Political Affiliation

The political affiliation for each state was manually coded into a data set. The data was obtained from the NCLS website[5].

The data for Nebraska, Minnesota and Wyoming was dropped as these states have either a split legislature or are non-partisan. These states do not have a party affiliation whereas we are specifically looking for data that has an association between a state legislature's party affiliation and the covid infection rate in the state.

State Population

The population of each state was obtained from the World Population Review website[6]

The data provides state population metrics which enable us to calculate the output variable of infection rate per capita by state.

Data Analysis

All the data sources were combined using the state code which was available in the source data. The data that underwent calculations was calculated as follows:

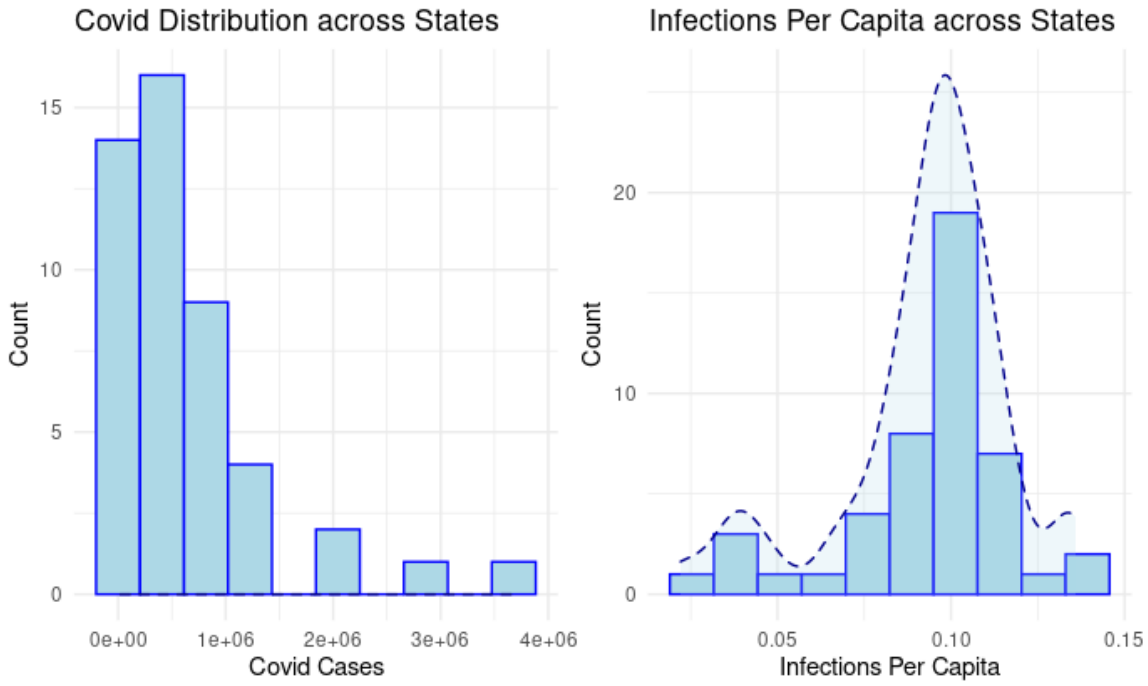
Combined Data	**Source Components of Data**
Duration of Restaurant Closures	Closed restaurants & Reopened restaurants +

[4] https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv
[5] https://www.ncsl.org/research/about-state-legislatures/partisan-composition.aspx
[6] https://worldpopulationreview.com/states

	Closed restaurants x2 & Reopened restaurants x2 + Closed restaurants x3 & Reopened restaurants x3
Duration of Movie Theater Closures	Closed movie theaters & Reopened movie theaters + Closed movie theaters x2 & Reopened movie theaters x2 + Closed movie theaters x3 & Reopened movie theaters x3
Duration of Stay at Home Mandate	Stay at home/shelter in place & End stay at home/shelter in place
Duration of Emergency Declaration	State of emergency issued & State of emergency expired
Duration of Face Mask Mandate	Public face mask mandate & End face mask mandate
Infection per Capita	Number of Covid Cases & State Population

Outcome Variable Distribution

Understanding how viruses transmit the disease, it's not surprising that positive infections are highly skewed (left graph). This is because COVID is highly transmittal and increases exponentially as more people get infected. With this in mind, we decided to calculate our outcome variable as a rate to help normalize this data set. By taking the total infections and dividing that by the state's population, we get infections per capita. As shown in the right graph, we were able to transform it into a roughly normal histogram.



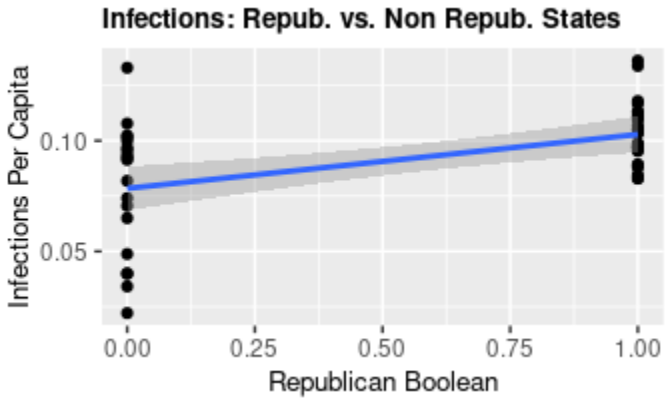
Correlations

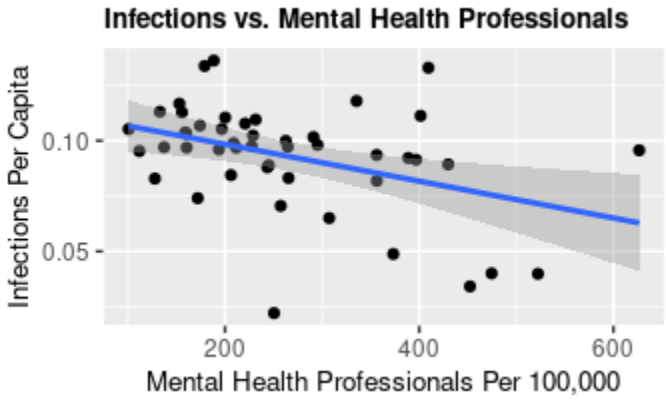
The correlations between variables and the outcome can be observed with the scatter plots below.

Republican States: Our data has categorized each state as either republican with a value of 1, or non-republican with a value of 0. The graph below shows a correlation between the "Is Republican" variable

and the outcome (“Infections per Capita”). Therefore the decision was made to include this as a variable in the regression.

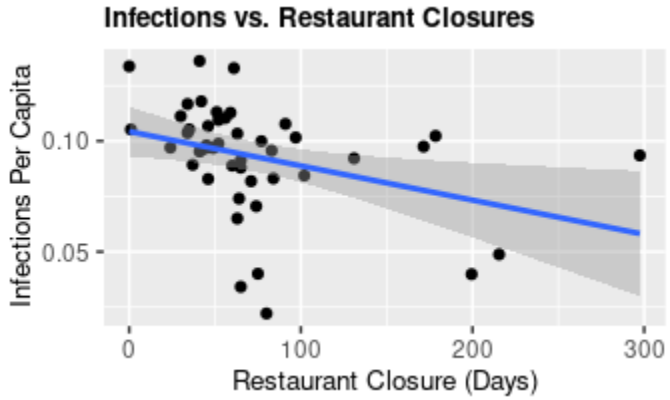
Mental Health Professionals: This variable was also included in the regression as it shows a correlation with the outcome (“Infections per Capita”). The reasoning behind why infections per capita would drop (as indicated in the graph) with increased presence of mental health professionals is not clear. This will be further reviewed in the regression models.

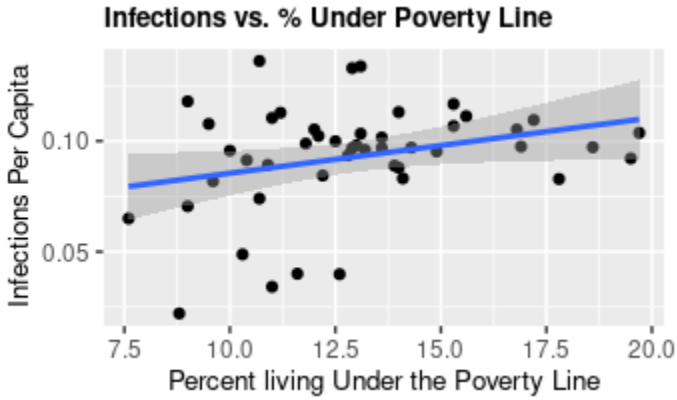




Restaurant Closures:The total number of days that restaurants were closed in a state is captured by this variable. The correlation between the infection per capita and restaurant closures is understandably negative.

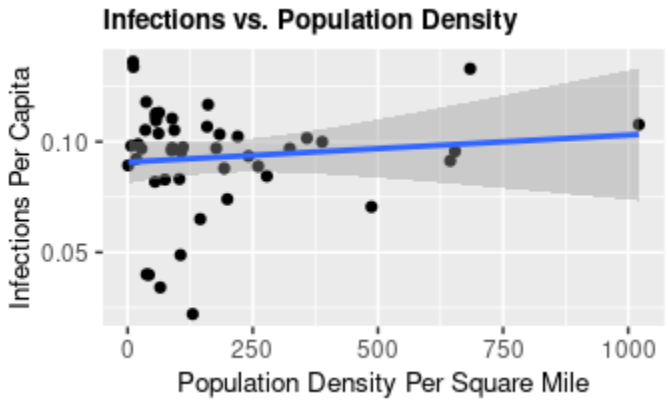
Percent of Population Poverty Line: The higher the poverty level in a state the fewer purchases of personal protective equipment and isolation guidelines can be followed by the population. The data shows that infection rates do rise with the level of poverty in the graph below.

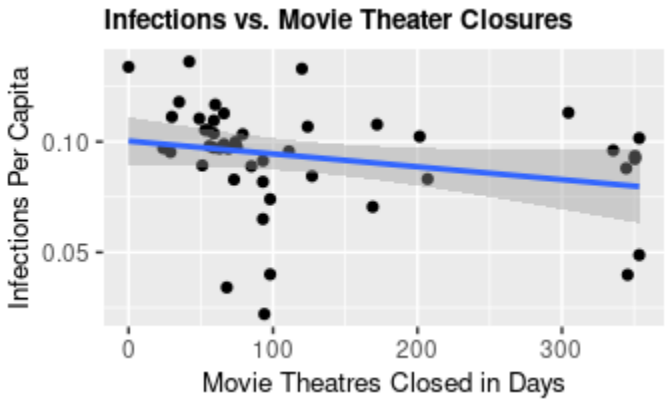




Population Density: The population density was expected to be a highly meaningful variable with a strong correlation as higher density means less social distancing however people are likely still able to isolate themselves during Covid even though the population density of a state is high. The correlation between population density and infections per capita is present but is weak. Using a log for population density did not provide any benefits or better linearity with the outcome.

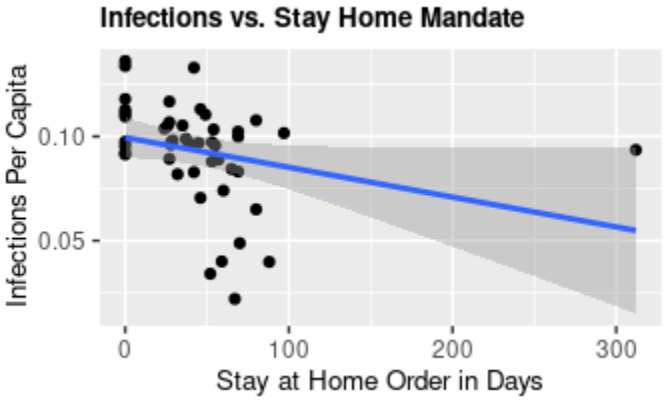
Movie Theater Closures: Movie closures have a negative correlation with the infections per capita. The degree of correlation as observed in the scatter plot below is however not very high.

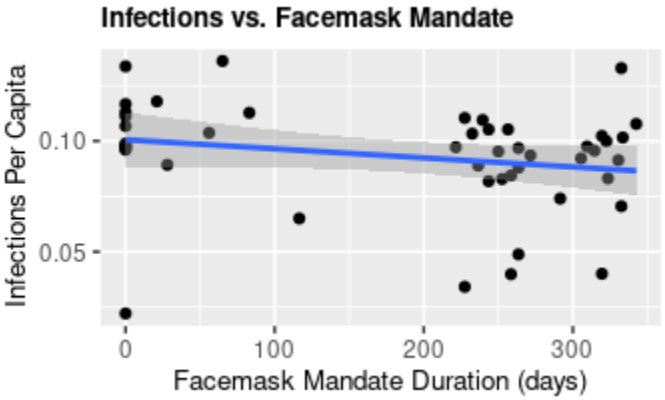




Stay Home Mandate: The state with the longest stay at home mandate is California which creates the outlier in the data as shown below. The linear model shows that the error in the correlation is significant. The data also has clustering which shows that a good linear correlation between staying at home and infections per capita does not exist.

Facemask Mandate: The data for facemask mandate implemented by different states has a relatively even distribution. Wearing facemasks is associated with lowering the transmission of Covid. A negative correlation does exist between the facemask mandate and infections as shown.





Model Building

During our research, we utilized an iterative methodology to come up with our final linear regression model. We took a bottoms up approach, and started with a base model that only included (*is.rep*.) Each model built upon the previous iteration, and we ultimately selected **model #4** to utilize in production. For each model, we will explain the rationale behind the variable selection and briefly touch upon some of the key model statistics. Lastly, we will go through the coefficients of the final model to explain the practical significance of our model.

	Dependent variable:			
	infections.per.capita			
	(1)	(2)	(3)	(4)
is.rep	0.025***	0.027**	0.021	0.029***
	(0.007)	(0.011)	(0.015)	(0.011)
mental.health.professionals.per.100.000.population.in.2019			-0.00003	
			(0.00004)	
rest.duration.days			0.00001	
			(0.0001)	
population.density.per.square.mile		0.0001***	0.0001***	-0.0001
		(0.00002)	(0.00002)	(0.0001)
movies.duration.days			-0.00001	
			(0.0001)	
stayhm.duration.days			-0.0001	
			(0.0003)	
I(population.density.per.square.mile2)				0.00000
				(0.00000)
I(population.density.per.square.mile3)				-0.000
				(0.000)
percent.living.under.the.federal.poverty.line..2018.		0.001	0.001	0.002
		(0.001)	(0.002)	(0.001)
facemsk.duration.days		-0.00002	-0.00002	-0.00001
		(0.00004)	(0.00005)	(0.00003)
Constant	0.078***	0.052**	0.066*	0.059***
	(0.007)	(0.020)	(0.040)	(0.018)
Observations	47	47	47	47
R2	0.246	0.394	0.416	0.466
Adjusted R2	0.230	0.336	0.293	0.385
Residual Std. Error	0.021 (df = 45)	0.020 (df = 42)	0.021 (df = 38)	0.019 (df = 40)
F Statistic	14.719*** (df = 1; 45)	6.818*** (df = 4; 42)	3.383*** (df = 8; 38)	5.807*** (df = 6; 40)

Limited Model (#1)

We started off by fitting a model that only included the variable of a State's party affiliation (Democrat v.s. Republican) against infections per capita. This is the main variable of interest, and from our EDA correlation analysis, we hypothesized that the indicator variable alone would be highly significant. With a p value of 0.0038, we found this input variable to be highly significant in predicting the output variable. Additionally, with an adjusted R^2 of 0.23, our baseline model gave us a good benchmark of improving the prediction accuracy.

Briefly scanning across each of the four models, it's worth noting that our party affiliation variable is consistently (¾ models) significant at the 99% level. The F statistic of model one (14.72) is also the best value across the four models, indicating that the party affiliation alone has a strong relationship with infections per capita.

Model Two (#2)

With the understanding from model one that we had a high significance utilizing just the party affiliation variable, we wanted to add additional inputs that didn't have a high colleatnity to it; while improving the model accuracy at the same time. We added population density (per square mile), percent living under the poverty line (2018 data), and facemask public mandate (days). More specifically, these variables didn't have a high correlation to the is.rep variable (< 0.5 pearson correlation coefficient), but also showed a relationship against our dependent variable.

Of the three additional variables, population density was the only incremental significant one in predicting infections per capita. Adjusted R^2 increased to 0.336, while the F statistic actually lowered to 6.818. The F statistic for model one actually performed the best across all four models, which does indicate to the fact that a state's party affiliation has a highly significant relationship with our response variable. Additionally, while we increased more variables after model one, the decrease in the F statistic indicates that the added variables actually have a low impact on the response variable. This shows that we lost more degrees of freedom than squared errors, ultimately lowering the F statistic.

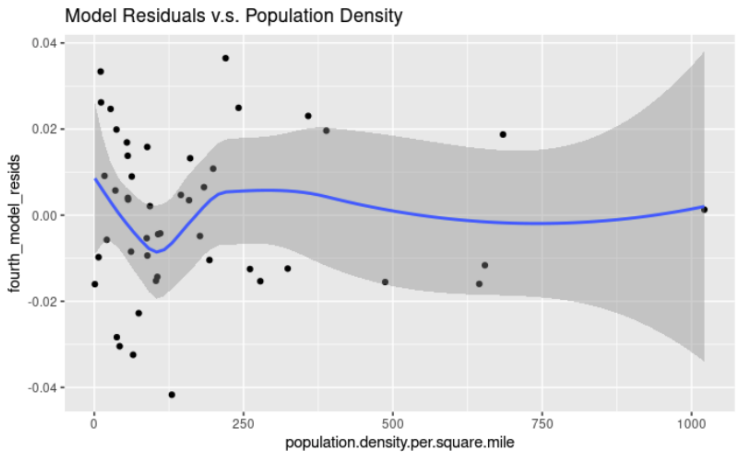
Model Three (#3)

For our model three, we essentially included all the variables that we narrowed down from EDA. This included all the variables from model two, and rest duration (days), movie theatre closure (days), stay at home orders (days), and mental health professionals per 100,000 population. Similar to model two, population density returned to be significant, however, party affiliation actually became non significant. This is an interesting observation as this shows that there are additional confounding variables that we are not accounting for which are associated with a state's party affiliation. We see this as by adding more variables from model two to three, we lose significance in our main variable of interest.

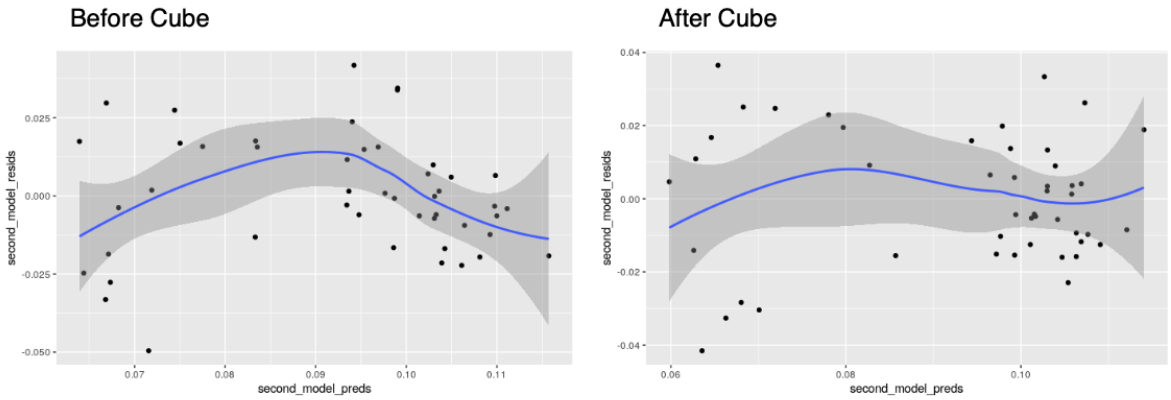
Looking at the adjusted R^2 of this model, we actually see this value drop to 0.293. The F statistic also drops to 3.383, which doesn indicate that we're crowding our model with variables that aren't strongly correlated with our response variable, while adding additional noise to the overall model construction.

Final Model (#4)

Our model four, and our final production model was built after we evaluated the CLM assumption from model two. More specifically, we noticed that we failed the linear conditional expectation assumption, with the root cause of our population density right skewed. This shows that the majority of states had a similar value (< 250 population density per square mile) while some populated small states such as NJ, RI, and MA resulted in significant outliers.



As we can see in the previous graph, there seems to be a polynomial relationship between the residuals and population density. Ultimately, we decided to cube the variable to help take on the shape seen previously. This meant we included population density, population density^2, and population density^3. As seen in the new linear conditional expectation check, we see the relationship flatten out more to 0, than it did previously.



This adjustment to cube this variable also showed very good results in our model summary statistics. Adjusted R^2 was the best performing here with 0.385 and our variable of interest (state's party affiliation) also came back to be significant.

Interpreting the party affiliation coefficient

As you can see in model four, the party variable is the only real significant variable that has a coefficient big enough to interpret. We can say, holding all other inputs constant, **changing from a democratic to republican state will increase it's infections per capita by roughly 0.029.**

More on the residual standard error

For model four, we observe a residual standard error of 0.019 with the intercept term coming out to be 0.059 infections per capita. Through these two values we can say that any prediction we have from this model can still be off by ~32.2% (SE/intercept). This prediction error is actually higher than our model one's error of 26.92%. It's interesting to see that our simplest model (#1) had quite good results versus the other three.

Model Analysis

AIC/BIC

Analyzing the two information criteria, we get two distinct results. AIC points to model four as the best one, while BIC points to model one. These two metrics are optimizing for different results so we need to interpret them differently. With BIC, the metric looks to find the true model in the set of choices, while AIC aims to find the model that best fits an unknown. It's a slight distinction, but AIC essentially assumes that the true model isn't necessarily in the selections.

With BIC selecting model one, we essentially are experiencing occam's razor and the single variable is explaining enough of the infections to warrant the selection. On the other hand, AIC selects model four which has more explanatory variables and probably predicts the output more accurately than the first. However, given that we're aiming to research a casual relationship, BIC does indicate that a party's affiliation has a strong enough relationship with infections to trump the other three models.

Model <chr>	AIC <dbl>	BIC <dbl>
#1	-223.6512	-218.1007
#2	-227.9208	-216.8200
#3	-221.6949	-203.1934
#4	-229.7775	-214.9763

CLM Assumptions

To be able to trust the results from our model, we need to ensure that our data meets the five Classical Linear Model assumptions, since our dataset contains only 47 observations. Satisfying the first three assumptions (I,I.D data, Linear Conditional Expectation, No Perfect Collinearity) will give us confidence that our model is unbiased. The fourth assumption (Homoskedastic Errors) is required to trust our standard errors. The fifth assumption (Normally Distributed Errors) allows us to trust the results of the hypothesis test used to determine whether our coefficients are statistically significant. In the following section, we will analyze each of these assumptions for our final model (Model #4), to determine if we can extract meaning from our model. We will also describe the limitations of our model, if any of these assumptions are not satisfied.

1) I,I.D

   Each observation in our dataset represents a state and contains information for the state's population density, number of infections per capita, closures of restaurants and movie theaters, and the amount of people living under the poverty line. Since we have data for all fifty states and our goal is to draw conclusions on a state level, we are dealing with the entire population, as opposed to a sample. Thus, there isn't a sampling process to analyze. However, we still have to determine whether the states are I,I.D.

   There are several reasons to believe that states are not I,I.D:

   - Clustering due to the location of a particular state could provide insights on states that are nearby. Residents from one state can also easily travel to neighboring states to visit friends or

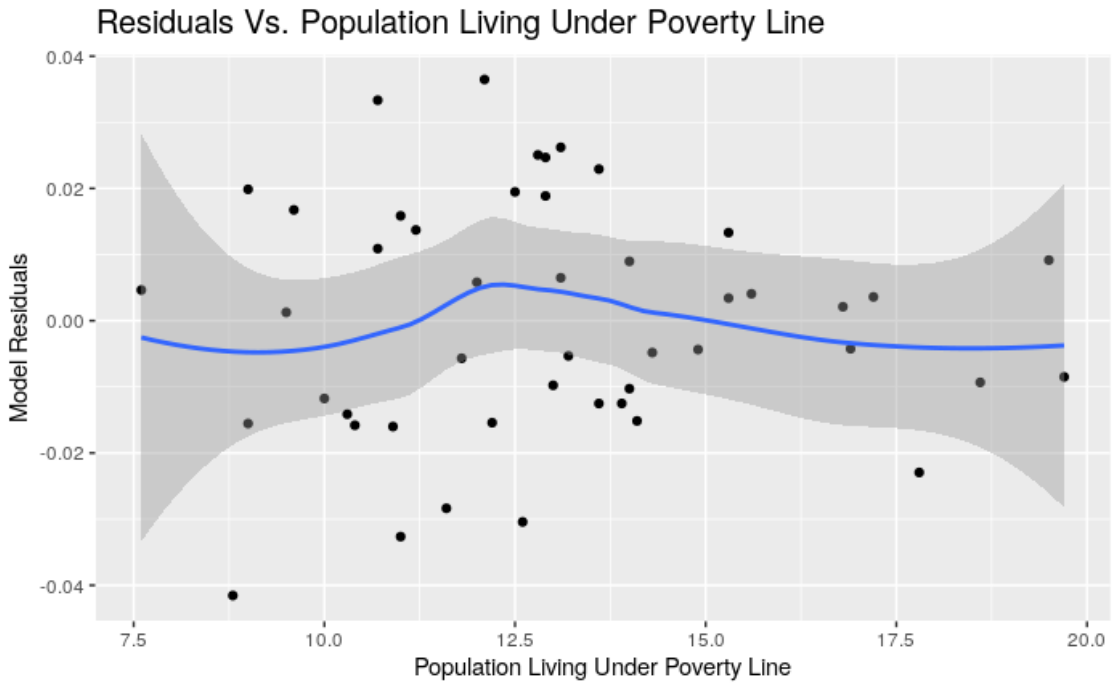
family, creating similarities between them. For example, neighboring states may share very similar political ideologies, which dictate the state legislature, because people are sharing ideas across state borders.

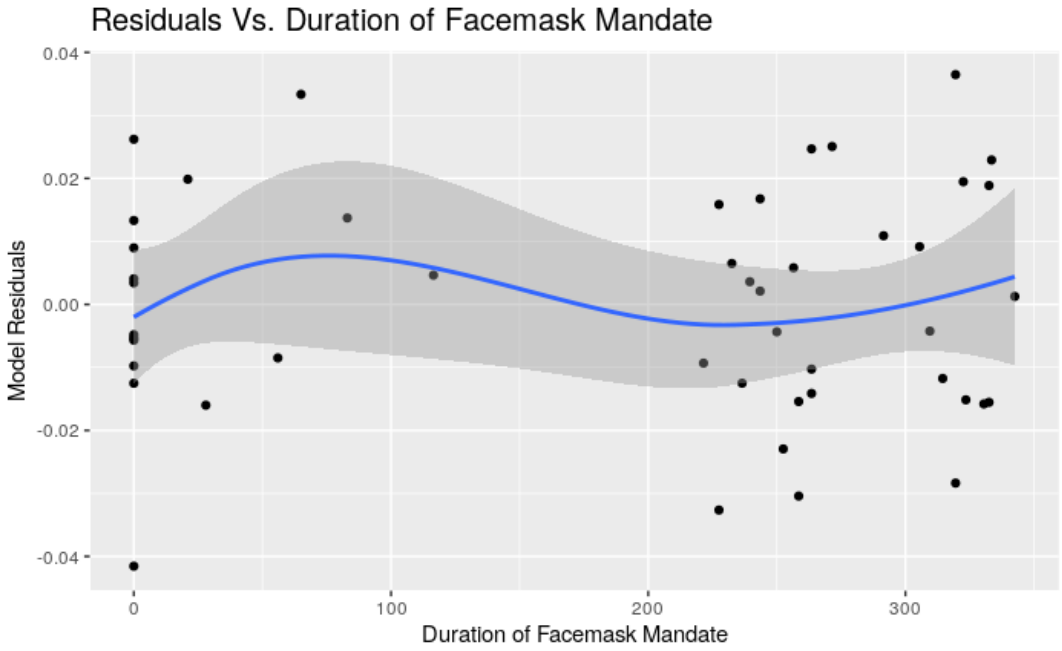
- Although each state may have different opinions regarding the severity of Covid-19, especially in the early stages of the pandemic, most states aimed to control the virus and reduce the number of cases/deaths. Thus, if one state implemented policies that were successful, another state may try to implement similar policies.
- The Covid-19 pandemic has impacted many businesses. Since businesses operate between states, either with other businesses or within the same company, they will have some influence on state policies. They will also have an effect on the wealth of residents, since they can directly impact jobs.

2) Linear Conditional Expectation

Since we're using a linear model, we need to ensure that the conditional expectation function (CEF) that we're estimating is also linear. By plotting the residuals vs. our predicted values, we can uncover any nonlinearity in the underlying joint density function. From the analysis on Model #4 above, we explained why we used a cubic transformation on the population density variable, after looking at the residuals plotted against the fitted values. Although most of the nonlinearity seems to be coming from the population density variable, the line in that plot does seem to flatten out around zero when looking at the fitted values.

Below are the same plots for the variables corresponding to the percentage of the population living under the poverty line and the duration of the facemask mandates.





3) No Perfect Collinearity

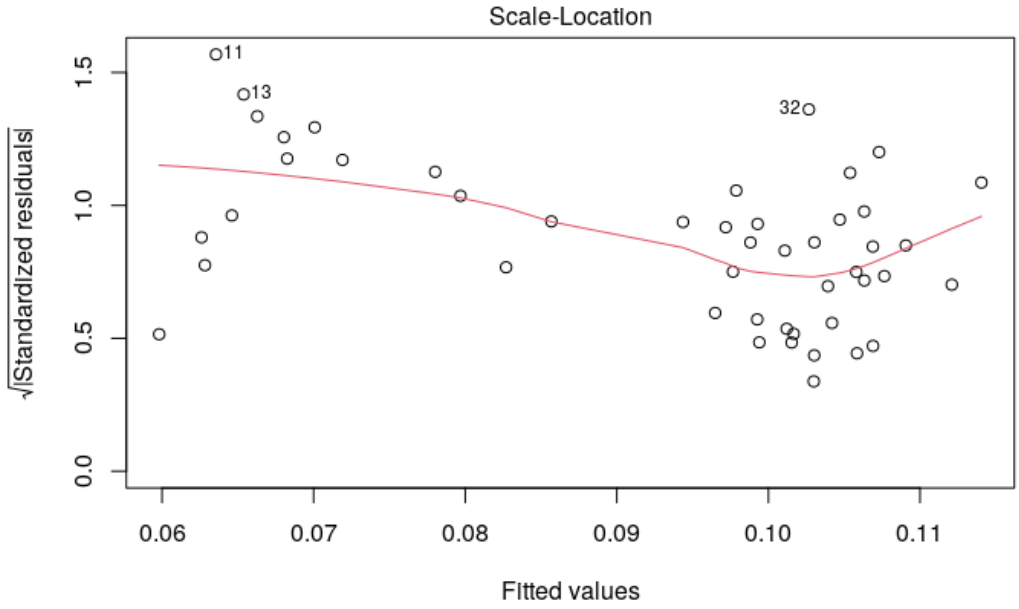
When two or more independent variables are correlated, redundant information can cause the results of the model to be skewed. Although perfect collinearity will prevent a model from being generated by R, it is important to assess the degree of correlation between independent variables, which was done by looking at each variable's variance inflation factor. As shown in the table below, the main variable of interest, corresponding to the state's governing party, has very low correlation between the other variables with a VIF of 2.037284.

The VIF for the remaining independent variables are also low, except for the population density. The VIF for population density is high because we did a polynomial transformation on it. However, since this is not a variable of interest, we decided that we do not need to resolve this.

Republican Legislature	Population Density (per sq mile)	Duration of Facemask Mandate	Percentage of Residents Living Under the Poverty Line
2.037284	^1: 35.282225 ^2: 199.036664 ^3: 84.419265	1.778564	1.442570

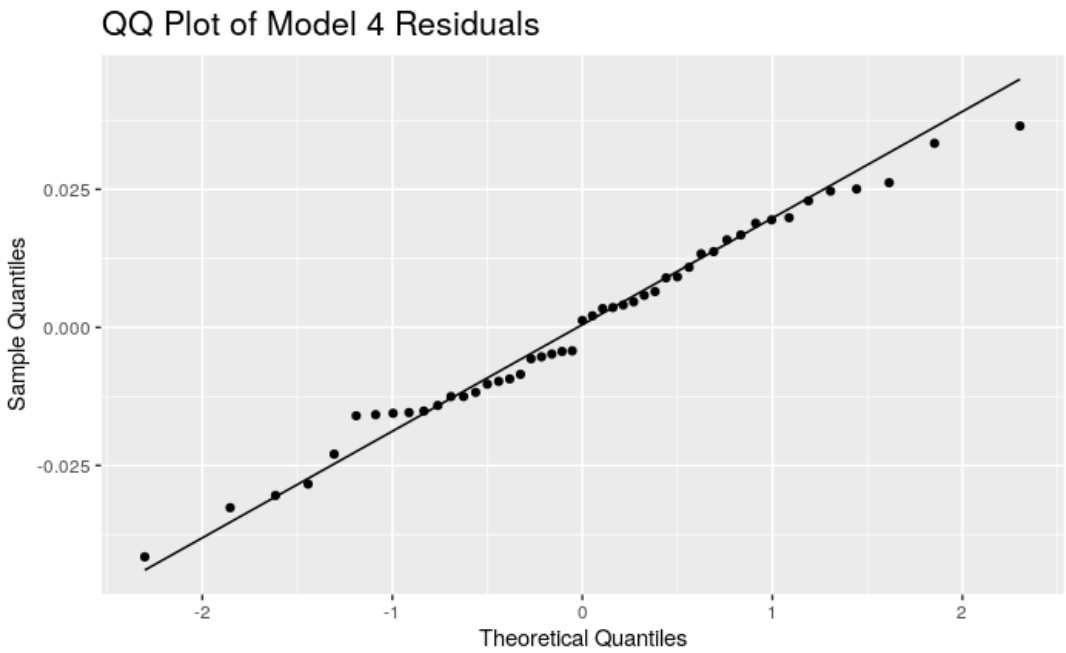
4) Homoskedastic Errors

To determine whether we have homoskedastic errors, we used the Breusch-Pagan Test, which yielded a p-value of 0.01666. Since the null hypothesis of this test is that we do have homoskedastic errors, we will reject the null hypothesis. Thus, we have heteroskedastic errors. This can also be seen in the plot below, where the red line is not flat.



To address this problem, we used robust standard errors when doing our hypothesis testing on the model coefficients.

5) Normally Distributed Errors
Since we are using a t-test to determine whether our model coefficients are statistically significant, we require that our model's standard errors are normally distributed. This was done by visually checking the QQ plot below. From this plot, we see that the data points fall fairly closely to the straight line, which represents a theoretical normal distribution. Thus, we can trust the results above, which states that the variable representing the state's party affiliation is statistically significant.



Omitted Variables

The model that we created using the features available does not have good explainability as evidenced by the low R^2 score of 0.466. This implies that there are significant omitted variables that are not included in the model. In fact adding more variables to the model does not increase explainability as seen by the R^2 value in model 3 that has a higher number of variables and yet a lower R^2 score of 0.416.

Notable significant omitted variables are:

Behavior Data: Infections in the wild are spread through human interaction. Even though we have state guidelines (sometimes enforceable by fines), they are a poor substitute for variables that provide data on actual population behavior for the year 2020. For example, a state could have a population that continues to visit malls and restaurants despite government mandates and have higher rates of Covid infections.

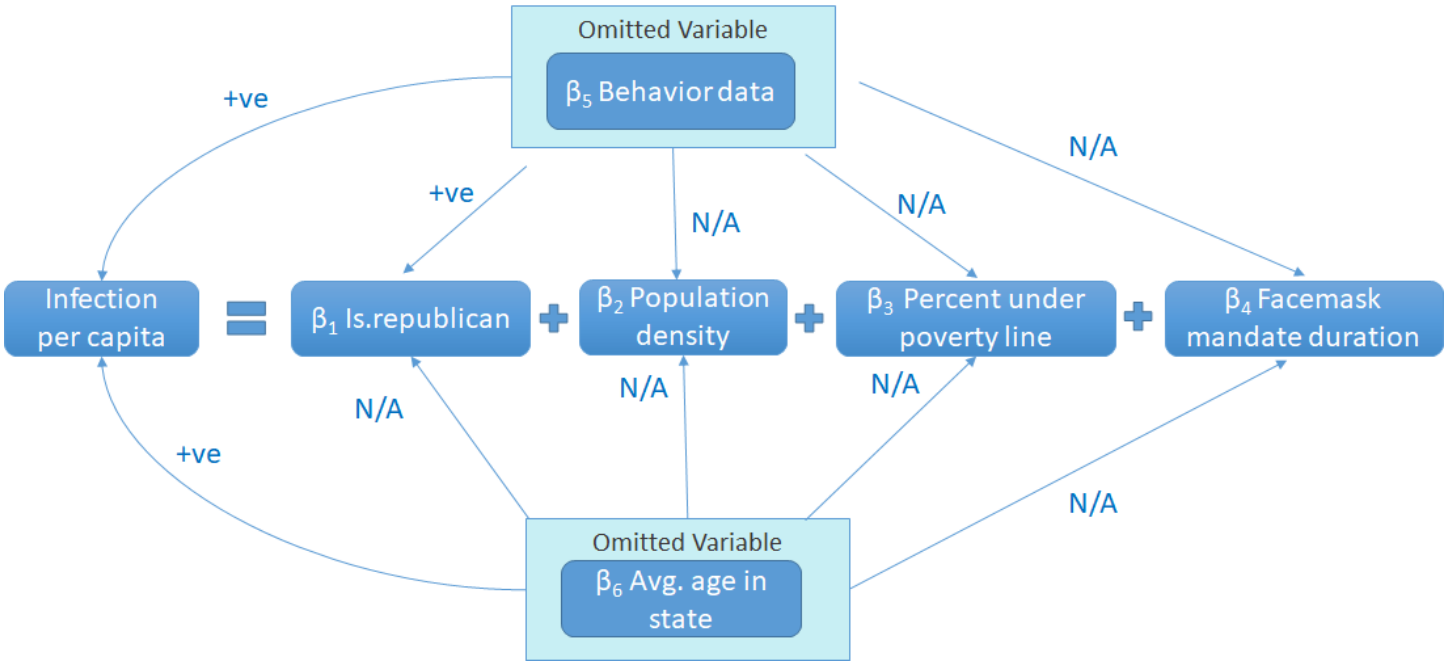
Average Age of each States' Population: Younger population is less likely to experience severe effects from Covid. Therefore this population is less likely to visit a medical facility and be screened for Covid. Average age of a state could therefore have a negative correlation to the number of infections in a state.

Unreported Infections: the number of medical facilities, availability of health insurance, and transportation are all factors that influence the number of infections reported within a state.

Using background knowledge the relationship between Infections per capita and the omitted variable:

- Behavior data is positive (higher interaction among the population, higher the infections)
- Average age in the state is positive (higher the age, higher the infections)

Map of the relationships between the variables:



When we assess the omitted variables, it emerges that their effect on the outcome variable may be profound however they are not measurable.

Using background knowledge, and given that there is a strong correlation between a republican legislature (“Is.Republican”) and the higher covid infections per capita, we can assume that behavior data of the population in majority republican states will also indicate higher degrees of socialization leading to higher infection rates. Therefore we can assume that there is a positive effect (i.e. infections will be higher) of the omitted variable which is being captured by other variables in the regression. The estimator for “Is.Republican” is therefore pushed away from zero.

The effect of the “Average age in a state” is more difficult to discern. A higher average age will lead to higher infections being reported since the older population is more susceptible to Covid. However, we cannot know what kind of relationship should exist between “”Average in a state” and other variables until we acquire the data and perform an analysis..

The confounding variables reduce the explainability of our regression and their bias is also not known (except for between “Behavior data" and “Is.Republican”).

Conclusion

In model one alone (simplest model) we rejected the null hypothesis, and concluded that a state’s party affiliation had a significant impact to it’s infections per capita. This wasn’t a surprise as we saw in our EDA that the two variables have a high pearson's correlation of 0.52. Additionally, through our model building process, we were able to see consistently that the state’s party affiliation continued to show statistically significance in predicting our outcome variable; even when we added more inputs. Furthermore, through the CLM assumption checks and specifically ensuring for normally distributed errors, we were confident in the models returned p value that ultimately convinced us to reject the null hypothesis.

Through this analysis we saw statistically significant results to conclude that a state’s party affiliation does impact their number of infections per capita. That being said, we recognize that there are confounding variables that may have contributed to omitted variable bias. A state political affiliation in itself includes many different policies and actions that are both difficult to quantify as well as finding an exhaustive list to run a regression on. For example, in the chart outlined above, a state’s political affiliation is ultimately composed of individual people who vote, however, that doesn’t mean the people living there actually act according to the rules and regulations set by the state. With this in mind, we’re still confident in our results that we observed, with an understanding that there are many other variables that could either remove significance or increase it.

As we are always thinking about iterating and improving, we’ve outlined some additional next steps that we would carry out if we had more time for research.

- F-test on each incremental variable
- Additional EDA to determine whether or not we can further transform our current dataset to return a better conditional expectation graph
- Further research on potential omitted variables