Dhruv Shah
10/13/2020
MA346 - Carter

COVID/Election Correlation Project

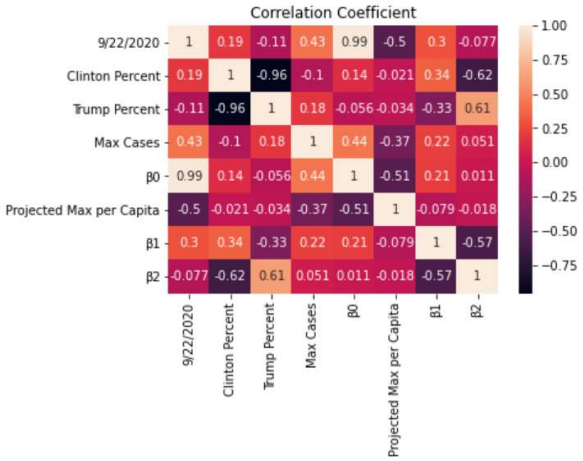
The data is showing a strong positive correlation between time to max spread for Republican states and a negative correlation for Democratic states. This would make sense because democrat states have been quicker to shutting down high risk operations while republican states have been more lenient on closing and have opened earlier. This would lead republican states to reach there peak, or time to max spread, faster.

Began by downloading the appropriate csv files from NPR and USA facts. I then deleted the extra Maine and Nebraska rows and added column names to the election data. After that I grouped the Covid confirmed cases and county populations, and then summed both. I then made the state the index, and merged the two datasets. This gave us the total population and the total number of covid confirmed cases per state in a singular dataset. I then merged the election dataset to give me df_final, which included populations, confirmed cases, and 2016 voting results for each state.

I then defined the logistic curve function to find my 3 betas. I performed a for loop in order to find the 3 betas of all states which returned a list within a list which contained the betas for each state. I then used another for loop to append the betas to beta_frame, which is the final dataset that includes all the info needed prior to creating a heatmap.

When assessing the correlation coefficient heatmap there seems to be a positive correlation between states that voted for Clinton and the rate of spread, alternatively there is a negative correlation (.34) for states that voted for Trump(-.33). This does surprise me, I would expect this to be the opposite considering the precautions most democrat states took.

There also seems to be a negative correlation between the most recent number of cases (9/22/20) and the projected max cases per capita. This is interesting because people would assume that most cases recently would mean more cases to come, although the spread of the virus is a logistic curve which mean as the cases are increasing the population is moving further down the curve. The strength of the correlation was (-.5). This does surprise me, I would have expected them to be positively correlated.



Decided to run a hypothesis test on whether the states political leaning had any bearing on the rate of spread. The null hypothesis is The states political affiliation First, I split the states into two subpopulations. The first subpopulation represented the states that voted for trump (republican states). The other subpopulation represented the states that voted for Clinton (democratic states). The null hypothesis was that the states affiliation was correlated to the time of maximum spread. The p-value proved to be higher than our alpha, which means we reject the null hypothesis, which mean the populations seem to be different.

I was surprised to see democrat states have a high rate of spread positive correlation. I would suspect this might be because many blue states are tightly packed like California, New York, and Massachusetts. Populous dense areas make it easy for the virus to spread quicker. Republican states are usually more spread out, like Texas, which might make the rate of spread much lower.