The Effects of Mobility Shifts on COVID-19 Active Cases

Mychelle Hale and Nicholas Roy

STAT 510: Regression Analysis

Final Project

California State University, Long Beach

Abstract

The COVID-19 pandemic is actively spreading throughout the US. Different counties in the US have different risks due to health concerns, different responses with stay at home order timing and compliance with statewide orders. In this paper we seek to determine what factors reduced the number of cases in each county within the four week period following statewide stay-at-home orders. We also want to see how different categories of mobility shifts affect the case change. We limit our analysis to states that have announced stay-at-home orders as of May 12^{th}, 2020. Our findings indicate that higher income areas have reduced effects of health issues increasing case counts. We also conclude that mobility is the most significant determinant during our observation period. Our analysis is limited due to data collection issues as well as some excluded potential variables.

I. Introduction

Our primary research question is how do characteristics of a county's population and residents' health, income, and mobility affect COVID-19 case numbers four weeks after stay-at-home orders. We merge three datasets into a single cross-sectional dataset where our observation unit is a U.S. county. To make the time-series case data compatible with the cross-sectional analysis, we generate a variable for the change in active cases four weeks after the statewide stay-at-home orders were announced in each county (our observation period). For county health and demographic information we use the 2020 US County Health Rankings dataset. Finally, we include Google's public mobility report data by creating variables for change in mobility during our observation period. This report shows the daily change in mobility from baseline for six different categories of destinations for a subset of U.S. counties.

II. Questions of Interest

A. How well do previous public health and demographic values explain county case increases for COVID-19 for 4 weeks after stay at home orders are announced?
B. What are the interaction effects of median income and health variables?
C. Did changes in mobility during our observation period affect case changes? If so by how much?

III. Regression Method

For the first question, we regress the change in county active cases four weeks after stay-at-home orders were announced on several health statistics of counties. We use the Akaike Information Criteria (AIC) variable selection method to determine which variables to be included. After confirming the multiple OLS linear regression assumptions are satisfied, we use the AIC criteria again to find which interactions with median income improve the model. Using this regression as our reduced model, we apply a general F-test to see if marginal mobility shifts affect the number of active cases. After testing for overall significance, we interpret the parameter estimates of the added variables as the marginal change in log cases due to a percentage point increase in mobility for a given category during our observation period. There is further discussion about this interpretation in the next section. We exclude counties in states that did not have any stay-at-home orders (Wyoming, Utah, South Dakota, Oklahoma, North Dakota, Nebraska, Iowa, and Arkansas) as well as where some data is missing.

Hypotheses we want to test by analyzing the data:

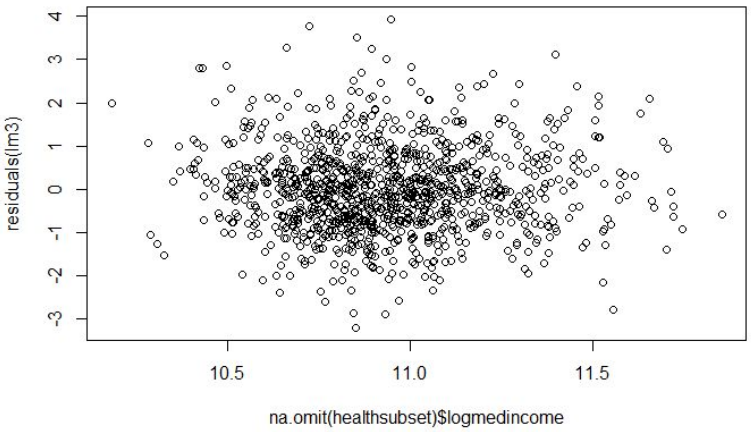
- Higher infant mortality rates, lower life expectancy, lower infant birth weights, lower vaccination rates, and lower insurance coverage will be found in counties with higher increases in active cases because they serve as proxies for general health of a population.
- Lower income populations will be associated with a greater increase in the number of active cases than average.
- Locations with decreases in mobility will have lesser increase in active cases across all categories of mobility.
- As median income increases, the marginal effects of poorer health on cases will decrease.

IV. Regression Analysis, Results and Interpretation

Before building our models we perform a Box-Cox test on a preliminary model and determine that the case change variable needs to be log transformed (see Figure 1). We can also see the need to transform this in the scatterplot matrix (Figure 2). For our variable selection, we choose several candidate variables that we think might best represent the counties' overall health: infant mortality rate, life expectancy, percent uninsured, percent age 65 or older, smoking percent, particulate matter count, and other similarly health related variables. We use AIC to determine which variables would work best for the model. The variables we select for our base model using this method can be seen in model (1) of Table 1 in the appendix (variables that are log transformed were not transformed at this stage in the process).

After the AIC variable selection, we check for equal variance and linearity using residuals versus fit and residuals versus variable plots, applying log transformations where necessary. Below in Figure 3 we can see the common change in residual plots after the log transformation. We find no non-linearity issues that would motivate polynomial terms. For the independence condition, it is reasonable to assume that the counties' information are independent of each other. The only cases where a county might have case or variable dependence on other counties is for magnet counties, such as Los Angeles, that attract a lot of cross county travel/commute. However, these counties tend to have lower urban populations and also shut down sooner than the whole state.

We apply the AIC criteria again to identify interactions with median household income. This leads to an additional five estimated parameters for the interaction median household income with log of child mortality rate, percent of rural population, life expectancy, and percent food insecure (See model (2) of Table 1). We still fail to satisfy the normality condition, so we now move to remove influential points. Using Cook's Distance, we identify any points that have a magnitude greater than $\frac{4}{n}$ and checked the QQ-plot to



see if there were more high leverage points included in the data. A plot of the Cook's Distance magnitudes for each point can be found in Figure 4 of the appendix.

Final Results

	Dependent variable:	
	logcasechange	
	(1)	(2)
loghouseholds	1.131***	1.200***
	(0.033)	(0.067)
log_percent_low_birthweight	2.676***	3.214***
	(0.212)	(0.507)
log_percent_food_insecure	8.876	1.849
	(5.488)	(10.603)
log_percent_uninsured	-0.330***	-0.166
	(0.075)	(0.157)
life_expectancy	1.749**	1.176
	(0.687)	(1.373)
logmedincome	22.871***	13.349
	(6.623)	(12.858)
percent_vaccinated	0.010**	0.025**
	(0.005)	(0.011)
log_child_mortality_rate	23.365***	7.926
	(5.911)	(13.275)
workplaces_pctpoint_change		-0.008
		(0.019)
retail_and_rec_pctpoint_change		-0.021*
		(0.011)
grocery_and_pharmacy_pctpoint_change		0.022**
		(0.010)
parks_pctpoint_change		-0.002
		(0.001)
transit_stations_pctpoint_change		-0.006
		(0.006)
residential_pctpoint_change		-0.010
		(0.028)
log_percent_food_insecure:logmedincome	-0.865*	-0.252
	(0.498)	(0.945)
logmedincome:log_child_mortality_rate	-2.119***	-0.822
	(0.541)	(1.191)
life_expectancy:logmedincome	-0.155**	-0.117
	(0.063)	(0.124)
Constant	-265.622***	-146.826
	(72.273)	(142.908)
Observations	1,031	293
R^2	0.712	0.747
Adjusted R^2	0.709	0.732
Residual Std. Error	0.853 (df = 1019)	0.835 (df = 275)
F Statistic	228.617*** (df = 11; 1019)	47.827*** (df = 17; 275)
Note:		*p<0.1; **p<0.05; ***p<0.01

Following the removal of 64 influential points and running the variable selection procedure using AIC, it seems that most of the influential points were feeding explanatory power to the percent rural variable. Removing them from the analysis and removing the percent rural variable better explains the variation in cases overall. We also ran the Shapiro Test to check for normality (see Figure 5). With a p-value of 0.2418, the model now satisfies the normality assumption. We retest the other conditions and confirm that all four assumptions of multiple linear OLS regression are satisfied with our final health model. This model can be seen as model (1) of Table 3 on the left and in the appendix (also in model (4) in Table 1).

After developing a model of health variables and interactions that satisfies the assumptions, we then add our mobility variables. These variables are the county level change in mobility for each category over the four week period. To see if any of the mobility variables have an effect on case change we use a general F-test (See Figure 6). With 90% confidence we can say that at least one mobility variable has a nonzero effect on the case change.

For the results of the regressions we run, please see Table 1 and Table 2 in the appendix. According to the R^2 and adjusted R^2, we explain 64-74% of the variation across all our models. In our final health model (model(1) in the above table and model (4) in Table 1 of the appendix), we see several variables that have intuitive explanations. Increases in log transformed number of households, log percent low birthweight, log percent food insecure, log percent uninsured, and log child mortality rate are all associated with increased number of cases. Household numbers are a good proxy for both population and number of transmissions locations, so we are mostly using it as a control. A 1 log percentage increase in food insecurity of a county leading to 8.876 more log cases is also sensible, as individuals likely stock up frequently during this time if they were in a county with more food insecurity. Low birthweight, uninsured rates,

and child mortality rates are good indications of overall public health conditions in a county that result from the environment, sex education, and standard of living, so lower health values should indicate an increase in the number of cases.

The two interaction variables we have in our model also make sense. As median household income increases by one log median income (in $1000), the marginal effect of food insecurities on the change in cases decreases by -.865 change in log cases. Wealthier areas are more likely to have the capacity to address food insecurities during a pandemic compared to lower income areas. For child mortality, increases in log median income (in $1000) decrease the effect of child mortality on case change by -2.119 log cases. For similar reasons as food insecurity, this can be explained by the overall public health capacity of wealthier areas to reduce the pandemic growth that results from issues that cause higher child mortality rates.

Some variables have more questionable coefficients. Increase in median income was shown to positively affect the change in cases by 22.871 log cases for every log thousand dollars. In fact, this was the largest marginal effect compared to the other variables other than child mortality rate (23.365 log cases for a one logged percentage point increase in child mortality rates). This shows that income alone has a positive effect on COVID-19 cases, but reduces the effect of health risks on the spread of the virus. Possible explanations for this is that wealthier areas have less issues with under reporting of tests or propensity to travel. Vaccination rates also increase the number of cases, but by only .01 log cases for one log percentage point increase in vaccination rates. This doesn't have an obvious reason, but it could also be due to reporting bias.

The most complex effect we have is with life expectancy and its interactions. Counties with larger life expectancies have a greater increase in cases over the observation period. For every percentage point increase in life expectancy, there is a 1.749 increase in log cases. This could be because higher life expectancy is correlated with counties with greater elderly populations, but this variable was excluded by our AIC selection. However, the interaction term is negative implying that as areas get wealthier, this effect begins to reverse.

After establishing the base health model, we can test the overall significance of the mobility shifts. It is important to note that due to data limitations, the mobility analysis was with a smaller subset of data with only 293 of our original 1031 observations (with influential points removed). To test mobility variables joint significance, we take the null hypothesis that all of the parameter estimates are zero and the alternative hypothesis that at least one mobility variable has a nonzero effect on case changes. Using a general F-test, we get a p-value of 0.05787 when moving from the model (3) to model (4) as seen in Table 2 of the appendix. This means that we can say with almost 95% confidence that at least one mobility variable has a non-zero effect on the number of cases. With a 90% confidence interval we can reject the null in favor of the alternative. It is important to note that once including the mobility variables, almost all pre-existing public health factors other than percent low birthweight are no longer significant.

When we look at the mobility variables, we see that only two categories are significant - retail/recreation and grocery/pharmacy. A percentage point decrease in the trips to groceries and pharmacies led to a .022 decrease in the log case change during the four week period after the stay-at-home order. With retail and recreation, we see an almost equal and opposite effect with a one percentage point increase during the observation period leading to a .021 decrease in the log case change. The mechanism we think we are seeing here is that counties that are stocking up on essentials are making less trips to the store; thus are having less cases due to less exposure. This

can further be argued by noticing that before removing influential points, park and residential areas travel had a negative effect on the number of cases. It is unclear why we see a similar effect with recreation and retail. One potential explanation is that areas that haven't closed retail or recreation facilities had less strict orders due to the lower public health risk of coronavirus. This is potentially where the rural variable might have been valuable if it wasn't mainly explaining influential data points.

To synthesize these results we return to our original hypotheses. Firstly, we mostly saw the effects of key health variables behaving how we expected. Counties that were already susceptible to public health issues did see increases in cases. While we weren't able to measure our rural population, as it seemed to be the only variable explaining influential points, we saw that higher income areas were actually likely to have more cases, proving our hypothesis wrong. However, increased median income further reduced the number of cases caused by health factors which was predicted by our hypothesis. Finally, decreased mobility to grocery stores and pharmacies showed a decrease in cases, while the opposite was true for recreation and retail areas. Perhaps the most striking result is that the pre-existing health factors lose almost all their explanatory power when applying the mobility variables to the analysis. This means that even areas with very high risk are not substantially different from other areas in that mobility reductions are the most statistically significant way to reduce cases.

Here we see that several of the effects we observed were influenced by some unexplained factors such as data limitations, variables determining influential points, and potential interactions between these variables and the mobility variables. These questions are outside of the scope of our model and research questions. With more refined questions and more complete data, we can get a better and more realistic understanding of these effects.

V. Conclusion

Our analysis shows a few important behaviors of case growth during the pandemic. A key finding is that poorer counties are more severely impacted by coronavirus due to health conditions than higher income counties. We also can see that mobility factors explain more of the change in cases over the past month than the pre-existing public health variables. These two main findings reveal to us that we should be concerned about reducing mobility to reduce the amount of cases; then, once the effect of mobility is no longer significant, we should target lower income and unhealthier counties.

This analysis took an originally time-series data set and restructured it to be cross-sectional. Future analysis of this particular topic should perform a more robust time series analysis of the mobility changes and case counts. We also ignored the impacts of deaths and hospital capacity because it was outside of the scope of our dataset. Including these variables and exploring interactions with mobility could better explain the behavior of the pandemic in individual counties.

References

CDC SVI Documentation. (2020, March 27). Retrieved from https://svi.cdc.gov/Documents/Data/2018_SVI_Data/SVI2018Documentation.pdf.

COVID-19 Community Mobility Report. (n.d.). Retrieved from https://www.google.com/covid19/mobility/

Kutner, M. H., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models*. McGraw-Hill/Irwin.

The New York Times. (2020, March 27). We're Sharing Coronavirus Case Data for Every U.S. County. Retrieved from https://www.nytimes.com/article/coronavirus-county-data-us.html

If you'd like to download the code feel free to access it at the github we transition to for collaboration around the time of the presentation.

NicholasHRoy. (n.d.). NicholasHRoy/covidanalysis. Retrieved from https://github.com/NicholasHRoy/covidanalysis

VI. Appendix

Appendix

Table 1: Health Regressions

Health Models

	Dependent variable:			
	logcasechange			
	(1)	(2)	(3)	(4)
percent_rural	-0.0004	0.199***	0.121*	
	(0.002)	(0.073)	(0.071)	
loghouseholds	1.086***	1.071***	1.121***	1.131***
	(0.048)	(0.047)	(0.043)	(0.033)
logmedincome	0.661**	20.498***	27.247***	22.871***
	(0.259)	(6.894)	(7.098)	(6.623)
log_percent_low_birthweight	2.927***	2.487***	2.639***	2.676***
	(0.229)	(0.230)	(0.215)	(0.212)
percent_vaccinated	0.009*	0.013**	0.010**	0.010**
	(0.005)	(0.005)	(0.005)	(0.005)
log_infant_mortality_rate	-0.209			
	(0.202)			
log_percent_uninsured	-0.278***	-0.234***	-0.332***	-0.330***
	(0.084)	(0.082)	(0.075)	(0.075)
life_expectancy	0.065***	1.718**	2.302***	1.749**
	(0.024)	(0.737)	(0.759)	(0.687)
log_average_daily_pm2_5	-0.078			
	(0.187)			
log_child_mortality_rate	0.391*	15.613***	22.730***	23.365***
	(0.205)	(5.824)	(5.930)	(5.911)
log_percent_food_insecure:logmedincome				-0.865*
				(0.498)
log_percent_food_insecure	-0.509***	9.309*	10.021*	8.876
	(0.185)	(5.288)	(5.547)	(5.488)
percent_smokers	0.024			
	(0.016)			
logmedincome:log_child_mortality_rate		-1.401***	-2.062***	-2.119***
		(0.533)	(0.543)	(0.541)
percent_rural:logmedincome		-0.019***	-0.011*	
		(0.007)	(0.007)	
logmedincome:life_expectancy		-0.152**	-0.206***	
		(0.067)	(0.070)	
logmedincome:log_percent_food_insecure		-0.897*	-0.968*	
		(0.479)	(0.503)	
life_expectancy:logmedincome				-0.155**
				(0.063)
Constant	-24.794***	-239.581***	-312.951***	-265.622***
	(3.705)	(75.292)	(77.347)	(72.273)
Observations	1,101	1,101	1,031	1,031
R^2	0.642	0.654	0.712	0.712
Adjusted R^2	0.638	0.649	0.709	0.709
Residual Std. Error	1.004 (df = 1088)	0.989 (df = 1087)	0.852 (df = 1017)	0.853 (df = 1019)
F Statistic	162.693*** (df = 12; 1088)	157.709*** (df = 13; 1087)	193.853*** (df = 13; 1017)	228.617*** (df = 11; 1019)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2: Mobility Variable Choice and Models

Mobility Results

	Dependent variable:			
	logcasechange			
	(1)	(2)	(3)	(4)
loghouseholds	1.218***	1.152***	1.227***	1.200***
	(0.070)	(0.072)	(0.064)	(0.067)
log_percent_low_birthweight	2.628***	2.897***	3.191***	3.214***
	(0.507)	(0.537)	(0.475)	(0.507)
log_percent_food_insecure	5.108	1.040	3.446	1.849
	(10.344)	(10.240)	(10.511)	(10.603)
log_percent_uninsured	-0.298*	-0.178	-0.242*	-0.166
	(0.157)	(0.167)	(0.145)	(0.157)
life_expectancy	1.116	0.146	1.519	1.176
	(1.440)	(1.444)	(1.348)	(1.373)
logmedincome	15.385	5.400	16.526	13.349
	(13.738)	(13.859)	(12.569)	(12.858)
percent_vaccinated	0.022**	0.028**	0.026**	0.025**
	(0.011)	(0.012)	(0.011)	(0.011)
log_child_mortality_rate	14.768	7.539	9.967	7.926
	(13.363)	(13.394)	(13.136)	(13.275)
workplaces_pctpoint_change		-0.032		-0.008
		(0.021)		(0.019)
retail_and_rec_pctpoint_change		-0.031***		-0.021*
		(0.012)		(0.011)
grocery_and_pharmacy_pctpoint_change		0.025**		0.022**
		(0.011)		(0.010)
parks_pctpoint_change		-0.002*		-0.002
		(0.001)		(0.001)
transit_stations_pctpoint_change		-0.007		-0.006
		(0.006)		(0.006)
residential_pctpoint_change		-0.061**		-0.010
		(0.030)		(0.028)
log_percent_food_insecure:logmedincome	-0.515	-0.149	-0.401	-0.252
	(0.918)	(0.909)	(0.936)	(0.945)
logmedincome:log_child_mortality_rate	-1.397	-0.744	-1.009	-0.822
	(1.196)	(1.199)	(1.178)	(1.191)
life_expectancy:logmedincome	-0.106	-0.021	-0.146	-0.117
	(0.130)	(0.130)	(0.121)	(0.124)
Constant	-175.279	-62.998	-183.804	-146.826
	(152.552)	(153.937)	(139.655)	(142.908)
Observations	312	312	293	293
R^2	0.669	0.692	0.736	0.747
Adjusted R^2	0.656	0.674	0.726	0.732
Residual Std. Error	0.977 (df = 300)	0.952 (df = 294)	0.844 (df = 281)	0.835 (df = 275)
F Statistic	55.027*** (df = 11; 300)	38.861*** (df = 17; 294)	71.176*** (df = 11; 281)	47.827*** (df = 17; 275)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3: Final Models

Final Results

	Dependent variable:	
	logcasechange	
	(1)	(2)
loghouseholds	1.131***	1.200***
	(0.033)	(0.067)
log_percent_low_birthweight	2.676***	3.214***
	(0.212)	(0.507)
log_percent_food_insecure	8.876	1.849
	(5.488)	(10.603)
log_percent_uninsured	-0.330***	-0.166
	(0.075)	(0.157)
life_expectancy	1.749**	1.176
	(0.687)	(1.373)
logmedincome	22.871***	13.349
	(6.623)	(12.858)
percent_vaccinated	0.010**	0.025**
	(0.005)	(0.011)
log_child_mortality_rate	23.365***	7.926
	(5.911)	(13.275)
workplaces_pctpoint_change		-0.008
		(0.019)
retail_and_rec_pctpoint_change		-0.021*
		(0.011)
grocery_and_pharmacy_pctpoint_change		0.022**
		(0.010)
parks_pctpoint_change		-0.002
		(0.001)
transit_stations_pctpoint_change		-0.006
		(0.006)
residential_pctpoint_change		-0.010
		(0.028)
log_percent_food_insecure:logmedincome	-0.865*	-0.252
	(0.498)	(0.945)
logmedincome:log_child_mortality_rate	-2.119***	-0.822
	(0.541)	(1.191)
life_expectancy:logmedincome	-0.155**	-0.117
	(0.063)	(0.124)
Constant	-265.622***	-146.826
	(72.273)	(142.908)
Observations	1,031	293
R^2	0.712	0.747
Adjusted R^2	0.709	0.732
Residual Std. Error	0.853 (df = 1019)	0.835 (df = 275)
F Statistic	228.617*** (df = 11; 1019)	47.827*** (df = 17; 275)
Note:		*p<0.1; **p<0.05; ***p<0.01

Figure 1: Box-Cox Test

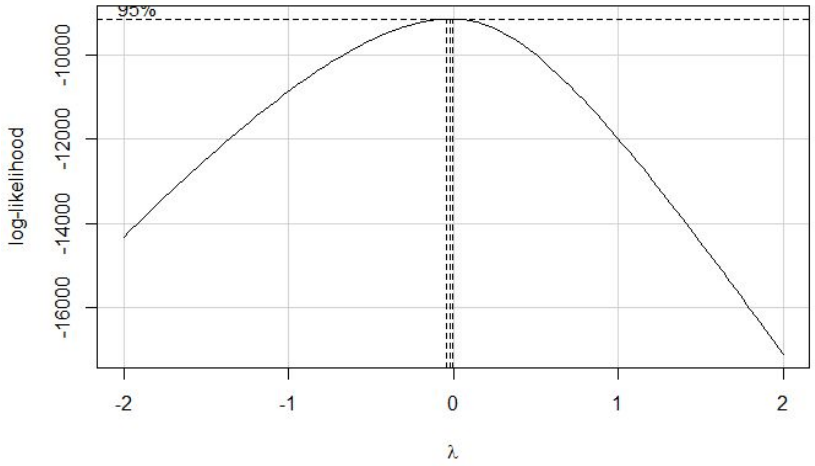
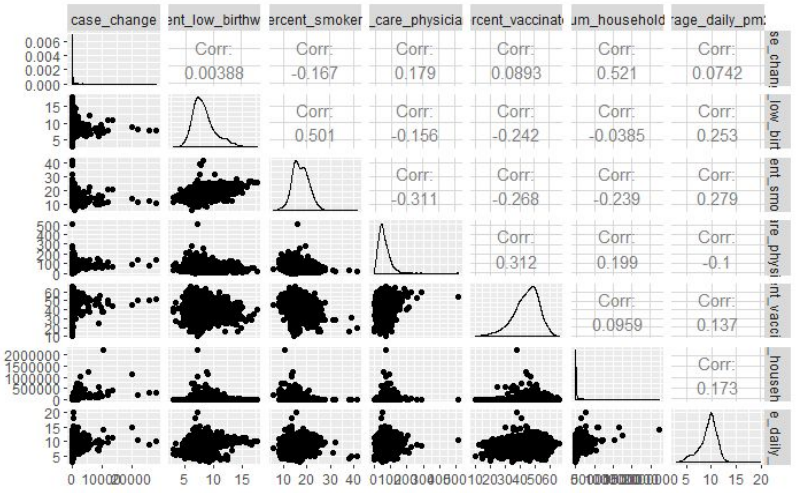
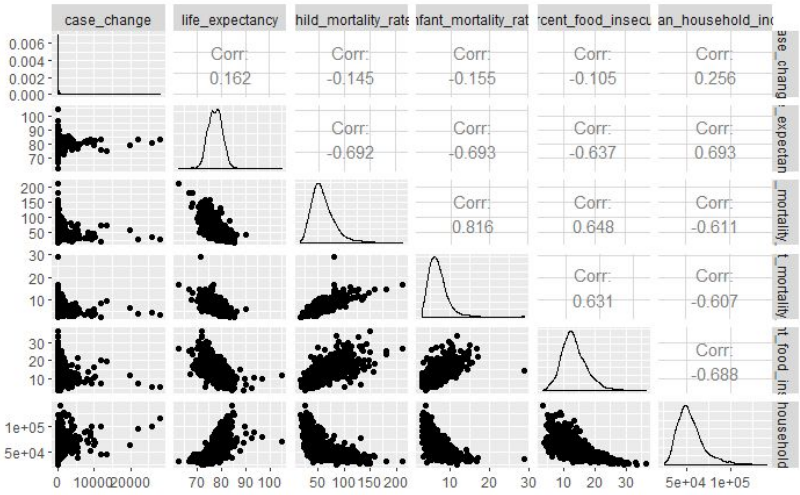


Figure 2: Cases Month vs. Independent Variables




case_change
life_expectancy
hild_mortality_rate
nfant_mortality_rat
rcent_food_insecu
an_household_inc
Corr:
0.162
Corr:
-0.145
Corr:
-0.155
Corr:
-0.105
Corr:
0.256
Corr:
-0.692
Corr:
-0.693
Corr:
-0.637
Corr:
0.693
Corr:
0.816
Corr:
0.648
Corr:
-0.611
Corr:
0.631
Corr:
-0.607
Corr:
-0.688

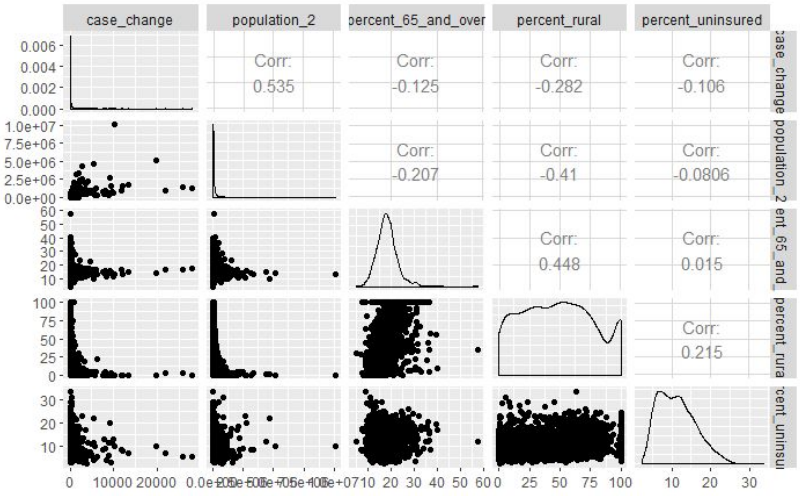

case_change
population_2
percent_65_and_over
percent_rural
percent_uninsured
Corr:
0.535
Corr:
-0.125
Corr:
-0.282
Corr:
-0.106
Corr:
-0.207
Corr:
-0.41
Corr:
-0.0806
Corr:
0.448
Corr:
0.015
Corr:
0.215

Figure 3a: Residual vs. Median Household Income

Figure 3b: Residual vs. Log(Median Household Income)

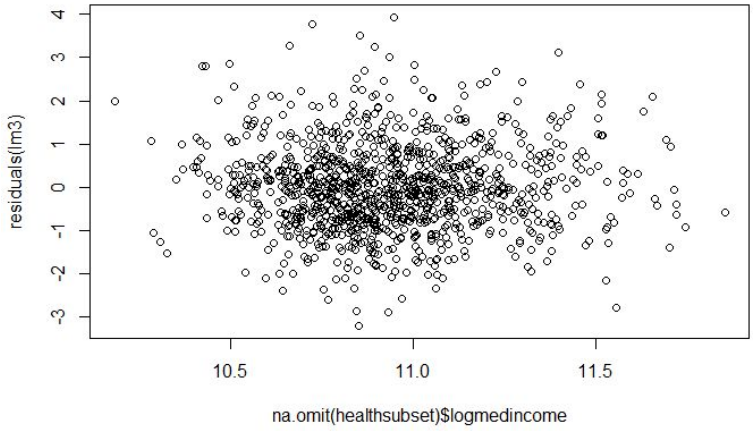


Figure 4: Cook's Distance

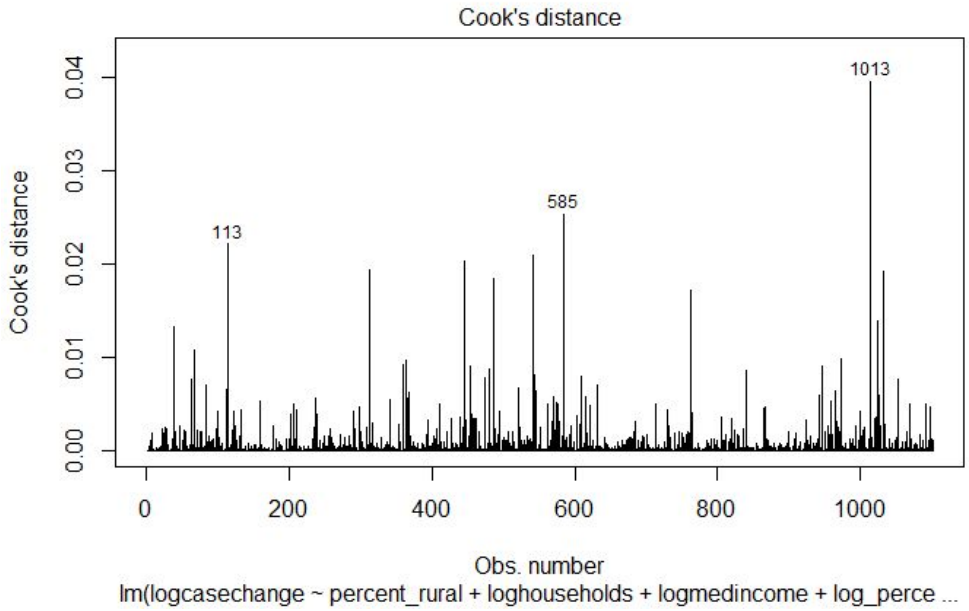


Figure 5a: Shapiro-Wilk Test and Other Plots for Normality (No High Leverage Points)

```
        Shapiro-Wilk normality test

data:  residuals(lm10)
W = 0.99095, p-value = 2.698e-06
```

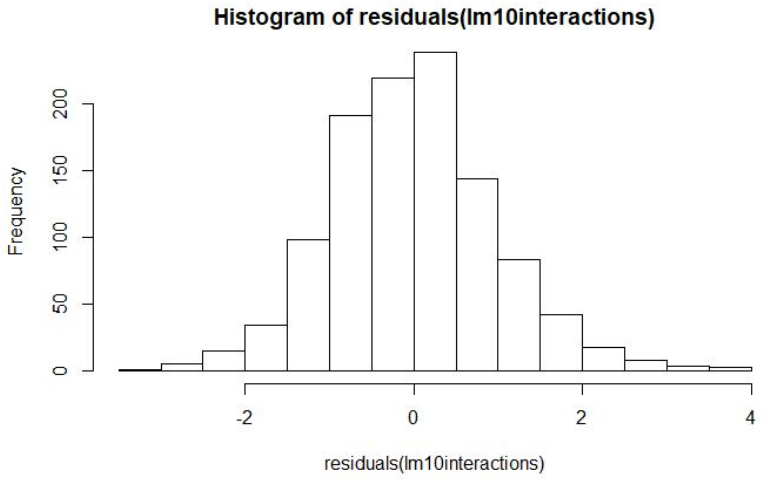


Figure 5b: QQ-plot with Interactions with High Leverage Points

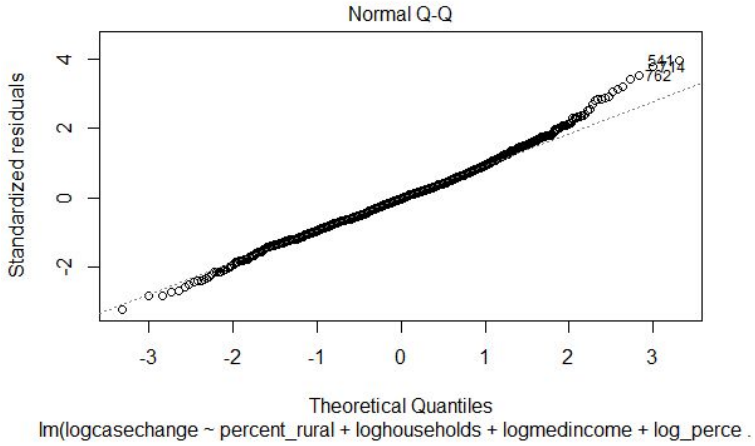


Figure 5c: QQ-plot with Interactions with No High Leverage Points

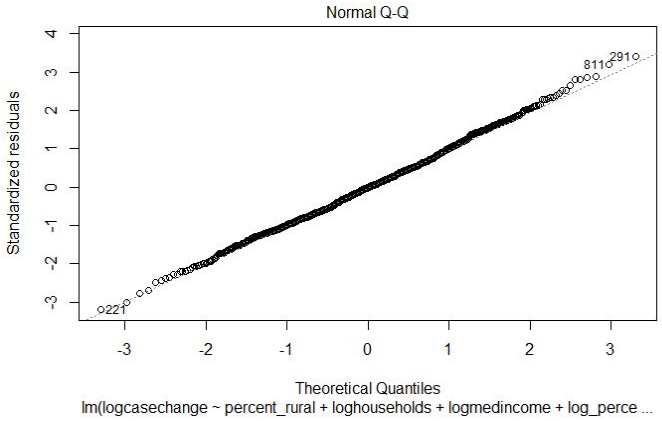


Figure 5d: Shapiro-Wilk Test and Other Plots for Normality (No High Leverage Points)

```
        Shapiro-Wilk normality test

data:  residuals(lm10interactions.new)
W = 0.99795, p-value = 0.2418
```

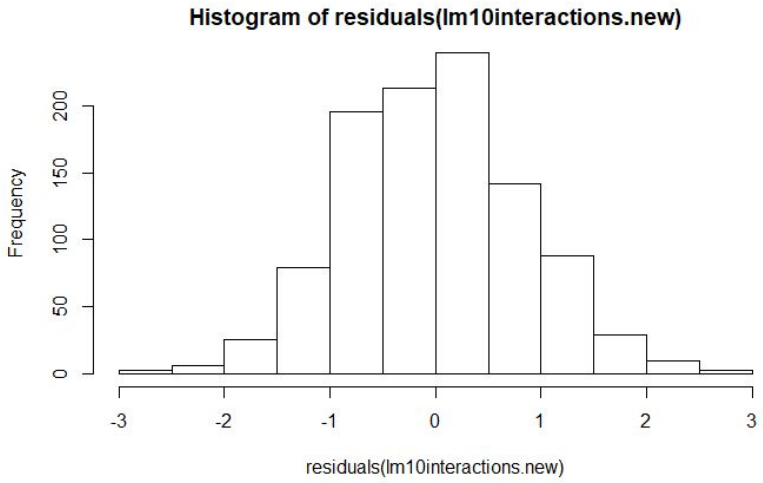


Figure 6: General F-Test for Overall Significance of Mobility Variables

```
Analysis of Variance Table

Model 1: logcasechange ~ loghouseholds + log_percent_low_birthweight +
    log_percent_food_insecure + log_percent_uninsured + life_expectancy +
    logmedincome + percent_vaccinated + log_child_mortality_rate +
    log_percent_food_insecure:logmedincome + logmedincome:log_child_mortality_rate +
    life_expectancy:logmedincome
Model 2: logcasechange ~ loghouseholds + log_percent_low_birthweight +
    log_percent_food_insecure + log_percent_uninsured + life_expectancy +
    logmedincome + percent_vaccinated + log_child_mortality_rate +
    log_percent_food_insecure:logmedincome + logmedincome:log_child_mortality_rate +
    life_expectancy:logmedincome + workplaces_pctpoint_change +
    retail_and_rec_pctpoint_change + grocery_and_pharmacy_pctpoint_change +
    parks_pctpoint_change + transit_stations_pctpoint_change +
    residential_pctpoint_change
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    281 200.22
2    275 191.60  6    8.6203 2.0621 0.05787 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: Life Expectancy vs. Other Independent Variables

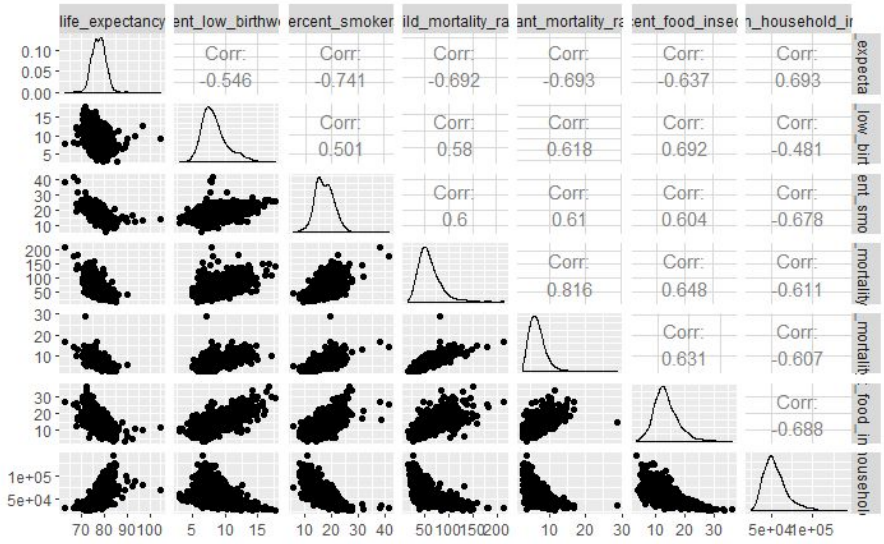


Figure 8a: All Points in Initial Dataset (Value shown is Case Change)

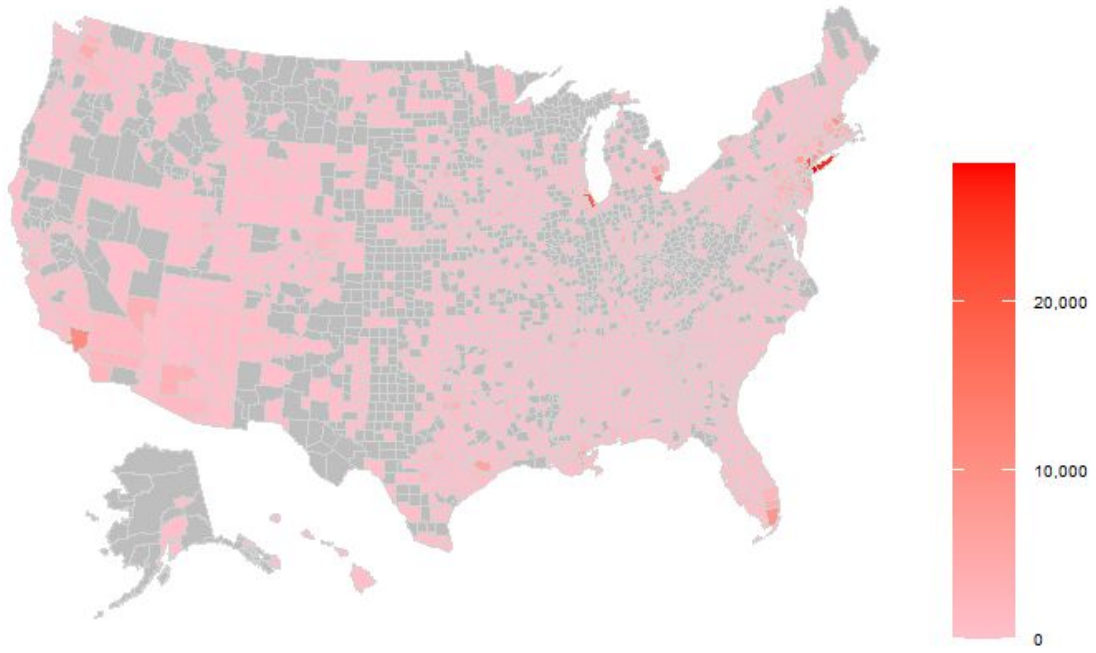


Figure 8b: All Points in Health Model (Influential Points Removed) (Value shown is Case Change)

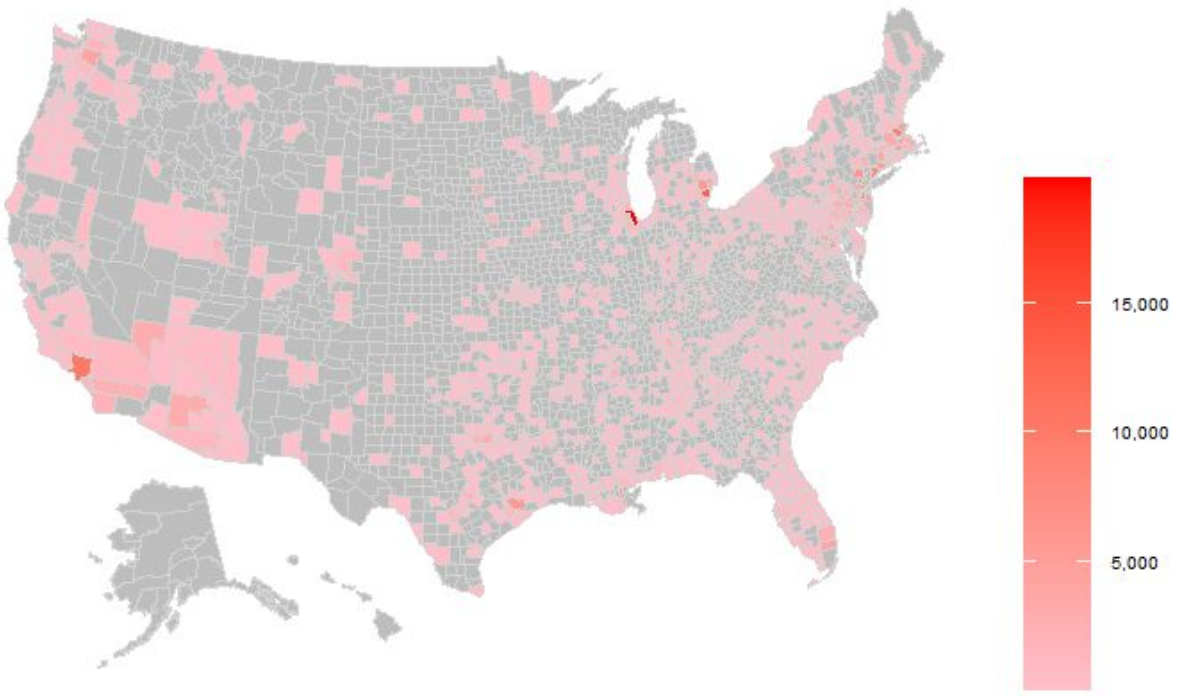
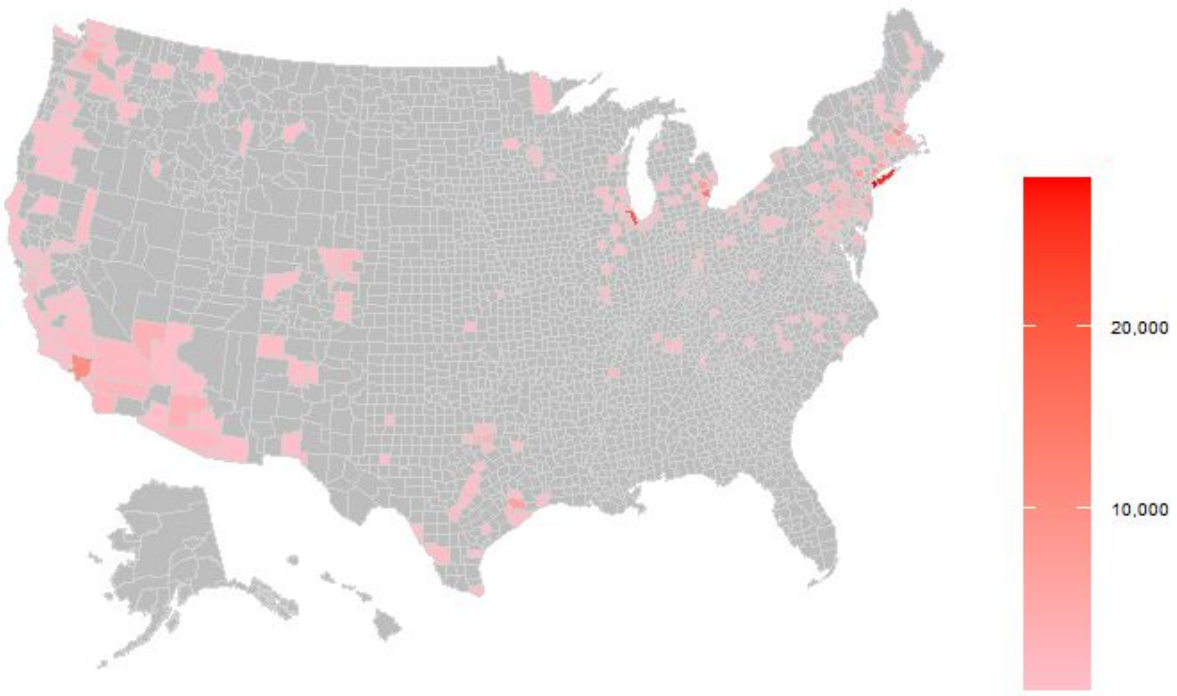


Figure 8c: All Points in Mobility analysis (Influential Points Removed) (Value shown is Case Change)



ALL R- CODE (FOR MARKDOWN)

````
---
title: "STAT510_Project"
author: "Nicholas Roy and Mychelle Hale"
date: "4/22/2020"
output: html_document
---

```{r setup, include=FALSE}

knitr::opts_chunk$set(echo = TRUE)

```

```{r, echo=FALSE}
pkgs = c("alr4","rmarkdown", "faraway", "datasets", "sos", "car", "plotly", "ggplot2",
"tidyverse", "broom", "SemiPar", "RegularExpressions","GGally", "leaps", "usmap", "maps",
"plotly", "devtools", "stargazer", "MASS")
new.packages = pkgs[!(pkgs %in% installed.packages()[,"Package"])]
if(length(new.packages)) install.packages(new.packages)

library(GGally)
library(ggplot2)
```

```{r}
##Working Directory
#setwd("C:\Users\mnhale\OneDrive - Urban Science\STATS\Spring 2020\STAT
510\covidanalysis\project")

#setwd("C:/Users/Nicho/OneDrive/2_Mathematics/School/courses/STAT510_intro_regression_
analysis/project/github/covidanalysis/project")
```

#Data Preparation

##Step 1: Import Raw Datasets

```{r}

#casesanddeaths = read.csv(file = "csvs/us-counties.csv")
casesanddeaths =
read.csv(url("https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv"))
````
````

```
data_health = read.csv(file = "csvs/us-county-health-rankings-2020.csv")
SAHorder = read.csv(file = "csvs/state-stay-home.csv")
global_mobility = read.csv(file = "csvs/Global_Mobility_Report.csv")
#global_mobility =
read.csv(url("https://www.gstatic.com/covid19/mobility/Global_Mobility_Report.csv?cachebust
=a88b56a24e1a1e25"))

#***REMOVE ONCE NEW DATA IS WORKING***
#data_mobility_old = read.csv(file = "csvs/us-mobility.csv")
#data_mobility_old$county = stringr::str_remove(data_mobility$region, " County")
#***REMOVE ONCE NEW DATA IS WORKING***
```

```

##Step 2: Create US County Level Data
```{r}

#Clean
us_mobility = global_mobility

#Remove Non-US Countries
othercountries<-which(us_mobility$country_region_code!="US")
us_mobility<-us_mobility[-othercountries,]

#Remove Observations with Missing Country Data
nacountries<-which(is.na(us_mobility$country_region_code))
us_mobility<-us_mobility[-nacountries,]

#Remove non-county subregion_2 observations (i.e. cities, state level observations)

us_mobility = us_mobility[grepl("County", us_mobility$sub_region_2),]

```

##Step 3: Remove non SAH states

For this class we are limited to cross-sectional analysis. To address the inherent time-series nature of this data, we are going to perform a transformation that simplifies the data to a 4 week period and only looks at states that stayed at home. Because each state implemented the stay at home orders at different times, our observations are simplified to the change over the 4 week period immediately following the stay at home orders. According to the Kaiser Family Foundation (KFF) , Missouri and South Carolina were the last
```

states to implement these orders on April 7th, 2020. We used the timeline from KFF to determine our observation period. Note according to USA Today , we see that Wyoming, Utah, South Dakota, Oklahoma, North Dakota, Nebraska, Iowa, and Arkansas have not declared stay at home orders. Since these states have very few counties and cases, we will be excluding them from the mobility data prior to the analysis as they could be potential outliers. This only applies for the mobility variables as they are they are the only independent variables that are most likely to be affected by stay at home orders. The health variables are based on pre-pandemic data and thus will not suffer from this issue. For the dependent variable we look at the 4 week change in cases/population after the most day following the most recent statewide order in the nation (4/8/20). A separate analysis excluding these states will also be included.

```{r}
#Remove non SAH states

nonSAH_states = which(us_mobility$sub_region_1 == "Wyoming" | us_mobility$sub_region_1 == "Utah" | us_mobility$sub_region_1 == "South Dakota" | us_mobility$sub_region_1 == "Oklahoma" |us_mobility$sub_region_1 == "North Dakota" | us_mobility$sub_region_1 == "Nebraska" |us_mobility$sub_region_1 == "Iowa" | us_mobility$sub_region_1 == "Arkansas")
#States w/ no stay at home order

us_mobility<-us_mobility[-nonSAH_states,]

```

##Step 4: Merge Mobility & Case Data
```{r}

#Clean Mobility Data to Merge w/ COVID Data
us_mobility$state = us_mobility$sub_region_1
us_mobility$county = stringr::str_remove(us_mobility$sub_region_2, " County")

CD_mobility = merge(x = casesanddeaths, y = us_mobility, all.x = TRUE, by = c("county", "state", "date"))

```

##Step 5: Adjust data based on "Stay at Home" orders

```{r}
```

```
SAHdates_withdata = merge(x = CD_mobility, y = SAHorder, all.x=TRUE, by.x = "state", by.y
= "STATE")

#Get rid of pre stay at home order dates
earlydates = which(as.integer(as.Date(SAHdates_withdata$date)) <
as.integer(as.Date("2020-03-18")))
SAHdates_withdata2 = SAHdates_withdata[-earlydates,]

#Focus on key dates
unimportant_dates = which(as.Date(SAHdates_withdata2$date) !=
as.Date(SAHdates_withdata2$EFF_DATE) & (as.Date(SAHdates_withdata2$date) -
as.Date(SAHdates_withdata2$EFF_DATE) != 28)) #Only have dates for a state based on day of
stay at home order and 3 weeks after.
impdates = SAHdates_withdata2[-unimportant_dates,]

pseudoSAH_dates =  which((impdates$date != "2020-04-08" & impdates$date != "2020-05-06")
& (impdates$state == "Wyoming" | impdates$state == "Utah" | impdates$state == "South
Dakota" | impdates$state == "Oklahoma" | impdates$state == "North Dakota" | impdates$state
== "Nebraska" |impdates$state == "Iowa" | impdates$state == "Arkansas" |impdates$state ==
"Guam" | impdates$state == "Northern Mariana Islands" | impdates$state == "Puerto Rico" |
impdates$state == "Virgin Islands")) #For locations where we don't have SAHOrder data

twodatesastate = impdates[-pseudoSAH_dates,]
```

##Step 6: Make variables differences over the past month to make cross sectional dataset from time-series.

```{r}
#Generating differences
sorteddata = twodatesastate[order(as.integer(twodatesastate$fips),
as.Date(twodatesastate$date)),] #sort by fips and date
head(sorteddata)

require("data.table")
sorteddata <- data.table(sorteddata)
sorteddata[, case_change := c(NA,diff(cases)), by="fips"]
sorteddata[, retail_and_rec_pctpoint_change :=
c(NA,diff(retail_and_recreation_percent_change_from_baseline)), by="fips"]
sorteddata[, grocery_and_pharmacy_pctpoint_change :=
c(NA,diff(grocery_and_pharmacy_percent_change_from_baseline)), by="fips"]
sorteddata[, parks_pctpoint_change := c(NA,diff(parks_percent_change_from_baseline)),
by="fips"]
sorteddata[, transit_stations_pctpoint_change :=
```

```
c(NA,diff(transit_stations_percent_change_from_baseline)), by="fips"]
sorteddata[, workplaces_pctpoint_change :=
c(NA,diff(workplaces_percent_change_from_baseline)), by="fips"]
sorteddata[, residential_pctpoint_change :=
c(NA,diff(residential_percent_change_from_baseline)), by="fips"]

missingcasedata = which(is.na(sorteddata$case_change))
almostfinalobs = sorteddata[-missingcasedata,]
missingfipsdata = which(is.na(almostfinalobs$fips))

finalobs = almostfinalobs[-missingfipsdata,]
```

```
##Step7: Add Health Variables

```{r}

allmerged = merge(x=finalobs, y = data_health, all.x = TRUE, by = "fips")
allmerged$state = allmerged$state.x

```


##Step8: Make datasets with variables of interest.
```{r}

## Full Linear Models for Cases

alldata = subset(allmerged,
select=c("fips","state","case_change","percent_low_birthweight","percent_smokers","primary_care_physicians_rate","percent_vaccinated","num_households","average_daily_pm2_5","life_expectancy","child_mortality_rate","infant_mortality_rate","percent_food_insecure","median_household_income","average_traffic_volume_per_meter_of_major_roadways","population_2","percent_65_and_over","percent_rural","percent_uninsured","workplaces_pctpoint_change",
"retail_and_rec_pctpoint_change","grocery_and_pharmacy_pctpoint_change","parks_pctpoint_change","transit_stations_pctpoint_change","residential_pctpoint_change",
"average_traffic_volume_per_meter_of_major_roadways"))

healthsubset =
subset(allmerged,select=c("fips","state","case_change","percent_low_birthweight","percent_smokers","primary_care_physicians_rate","percent_vaccinated","num_households","average_daily_pm2_5","life_expectancy","child_mortality_rate","infant_mortality_rate","percent_food_insecure","median_household_income","population_2","percent_65_and_over","percent_rural","percent
```
```

```
_uninsured"))

#For checking which mobility variable will be og greatest interest
# summary(alldata$retail_and_rec_pctpoint_change)
# summary(alldata$grocery_and_pharmacy_pctpoint_change)
# summary(alldata$parks_pctpoint_change)
# summary(alldata$transit_stations_pctpoint_change)
# summary(alldata$workplaces_pctpoint_change)
# summary(alldata$residential_pctpoint_change)
```

```

## Step 9: Check Correlation between variables and also linearity between independent variables and the dependent variable

```{r}
#based on the output here, we will probably need to log transform casesmonth
ggpairs(healthsubset, columns = c(3, 4:9))
ggpairs(healthsubset, columns = c(3, 10:14))
ggpairs(healthsubset, columns = c(3, 15:18))
#head(cov19_project_nm)
```

We can see that there is non-normality of the case change per capita variable. We will likely log transform this variable. Immdiately we see some variables have unequal variance, but we will perform variable selection before making any transformations.

As far as variables that are collinear, we make the following observations:
We see that life expectancy is collinear with many of the variables. This is to be expected since it is a more general term that is determined by the other variables.This wouldn't be useful as an interaction term, so we will look at the scatterplot matrix of variables correlated with life expectancy.

Step 10: Look at correlation with life expectancy.

```{r}
ggpairs(healthsubset, columns = c(10, 4,5,11,12,13,14))
#head(cov19_project_nm)
```

Based off this information we can see that median household income might be the interaction
```

with these variables. Life expectancy could be collinear with percent smokers and percent food insecure, but applying our variable selection criteria should remove the variable with the least amount of explanatory power. We should look for interactions with these variables after the variable selection.If interactions from here are not significant later in the selection process than we will remove collinear terms based on their signifigance.

Step 11: Use Box-cox method to determine best transformation for the dependent variable. The variable tempcases is simply to be positive to prevent errors for the box cox transformation

```{r}
library(car)

summary(healthsubset$case_change)
hist(healthsubset$case_change)

healthsubset$tempcases = healthsubset$case_change + 1.5 #Only to remove negative issue from boxcox and for future log transform

boxcoxtestreg = lm(data = na.omit(healthsubset), tempcases ~  percent_low_birthweight +percent_smokers +primary_care_physicians_rate + num_households +average_daily_pm2_5 +life_expectancy +infant_mortality_rate +median_household_income +percent_65_and_over + percent_rural )

bc = boxCox(boxcoxtestreg)

lambda.opt = bc$x[which.max(bc$y)]
lambda.opt

#Confirms that log transform is the best for cases


healthsubset$logcasechange = log(healthsubset$tempcases) #usingtemp cases to remove places that reduced cases

#rechecking the scatterplot matrix after log transforming y
ggpairs(healthsubset, columns = c(20, 4:9))
ggpairs(healthsubset, columns = c(20, 10:14))
ggpairs(healthsubset, columns = c(20, 15:18))
```

From the box-cox method, we can see that a log transformation is most appropriate.

Step 12: We can now do variable selection using the AIC variable selection criteria.

```{r}

mod0.lower_cases = lm(data = na.omit(healthsubset), logcasechange ~ 1)
mod0.upper_cases = lm(data = na.omit(healthsubset), logcasechange ~  percent_low_birthweight + percent_smokers +primary_care_physicians_rate +percent_vaccinated + num_households +average_daily_pm2_5 +life_expectancy +child_mortality_rate +infant_mortality_rate +percent_food_insecure +median_household_income +percent_65_and_over +percent_rural +percent_uninsured)
step(mod0.lower_cases, scope = list(lower = mod0.lower_cases, upper = mod0.upper_cases))

```

Step 13: We now have the base model for what variables we will be using. Now we must check the 4 assumptions of Multiple Linear Regression: Linearity, Independence, Normality, and Equal Variance

```{r}

lm1 = lm(formula = logcasechange ~ percent_rural + num_households +
    median_household_income + percent_low_birthweight + percent_vaccinated +
    infant_mortality_rate + percent_uninsured + life_expectancy +
    average_daily_pm2_5 + child_mortality_rate + percent_food_insecure,
    data = na.omit(healthsubset))

plot(residuals(lm1), fitted(lm1))

shapiro.test(residuals(lm1)) #Doesn't pass normality
hist(residuals(lm1)) #Looks relatively normal
plot(lm1, which = 2) #Looks relatively normal

```

We appear to have decent equal variance, linearity, and assume independence. However our model still does not satisfy the normality condition. To resolve this issue we will look to see if each variable satisfies equal variance conditions and then recheck normality.

```{r}

plot(y = residuals(lm1), x = na.omit(healthsubset)$percent_rural) # relatively equal variance. appears there is one particular outlier

plot(y = residuals(lm1), x= na.omit(healthsubset)$num_households) # clear funnelling patter that could use a log transform.
```

```
healthsubset$loghouseholds = log(healthsubset$num_households)

lm2 = lm(formula = logcasechange ~ percent_rural + loghouseholds +
   median_household_income + percent_low_birthweight + percent_vaccinated +
   infant_mortality_rate + percent_uninsured + life_expectancy +
   average_daily_pm2_5 + child_mortality_rate + percent_food_insecure,
   data = na.omit(healthsubset))

plot(y = residuals(lm2), x=na.omit(healthsubset)$loghouseholds) #Still some minor funnelling,
but less severe than in the previous graph

plot(y = residuals(lm2), x= na.omit(healthsubset)$median_household_income) #some minor
funnelling, log transform might help


healthsubset$logmedincome = log(healthsubset$median_household_income)

lm3 = lm(formula = logcasechange ~ percent_rural + loghouseholds +
   logmedincome + percent_low_birthweight + percent_vaccinated +
   infant_mortality_rate + percent_uninsured + life_expectancy +
   average_daily_pm2_5 + child_mortality_rate + percent_food_insecure,
   data = na.omit(healthsubset))

plot(y = residuals(lm3), x=na.omit(healthsubset)$logmedincome) #Equal Variance Acheived

plot(y = residuals(lm3), x=na.omit(healthsubset)$percent_low_birthweight) #some minor
funnelling, log transform might help


healthsubset$log_percent_low_birthweight = log(healthsubset$percent_low_birthweight)

lm4 = lm(formula = logcasechange ~ percent_rural + loghouseholds +
   logmedincome + log_percent_low_birthweight + percent_vaccinated +
   infant_mortality_rate + percent_uninsured + life_expectancy +
   average_daily_pm2_5 + child_mortality_rate + percent_food_insecure,
   data = na.omit(healthsubset))

plot(y = residuals(lm4), x=na.omit(healthsubset)$log_percent_low_birthweight) #Equal
Variance Acheived

plot(y = residuals(lm4), x=na.omit(healthsubset)$percent_vaccinated) #Seems like there is equal
variance
plot(y = residuals(lm4), x=na.omit(healthsubset)$infant_mortality_rate) #Needs log transform
```

```
healthsubset$log_infant_mortality_rate = log(healthsubset$infant_mortality_rate)

lm5 = lm(formula = logcasechange ~ percent_rural + loghouseholds +
   logmedincome + log_percent_low_birthweight + percent_vaccinated +
   log_infant_mortality_rate + percent_uninsured + life_expectancy +
   average_daily_pm2_5 + child_mortality_rate + percent_food_insecure,
   data = na.omit(healthsubset))

plot(y = residuals(lm5), x=na.omit(healthsubset)$log_infant_mortality_rate) #equal variance
acheived

plot(y = residuals(lm5), x=na.omit(healthsubset)$percent_uninsured) #Minor Funnelling, could
use log transform

healthsubset$log_percent_uninsured = log(healthsubset$percent_uninsured)

lm6 = lm(formula = logcasechange ~ percent_rural + loghouseholds +
   logmedincome + log_percent_low_birthweight + percent_vaccinated +
   log_infant_mortality_rate + log_percent_uninsured + life_expectancy +
   average_daily_pm2_5 + child_mortality_rate + percent_food_insecure,
   data = na.omit(healthsubset))

plot(y=residuals(lm6), x=na.omit(healthsubset)$log_percent_uninsured) #Improved funnelling
pattern to reach equal variance

plot(y=residuals(lm6), x=na.omit(healthsubset)$life_expectancy) #Looks like equal variance is
acheived

plot(y=residuals(lm6), x=na.omit(healthsubset)$average_daily_pm2_5)#Funnelling pattern

healthsubset$log_average_daily_pm2_5 = log(healthsubset$average_daily_pm2_5)

lm7 = lm(formula = logcasechange ~ percent_rural + loghouseholds +
   logmedincome + log_percent_low_birthweight + percent_vaccinated +
   log_infant_mortality_rate + log_percent_uninsured + life_expectancy +
   log_average_daily_pm2_5 + child_mortality_rate + percent_food_insecure,
   data = na.omit(healthsubset))

plot(y=residuals(lm7), x=na.omit(healthsubset)$log_average_daily_pm2_5) #Improved
funnelling pattern to reach near equal variance

plot(y=residuals(lm7), x=na.omit(healthsubset)$child_mortality_rate) #Needs log transformation
```

```
healthsubset$log_child_mortality_rate = log(healthsubset$child_mortality_rate)

lm8 = lm(formula = logcasechange ~ percent_rural + loghouseholds +
    logmedincome + log_percent_low_birthweight + percent_vaccinated +
    log_infant_mortality_rate + log_percent_uninsured + life_expectancy +
    log_average_daily_pm2_5 + log_child_mortality_rate + percent_food_insecure,
    data = na.omit(healthsubset))

plot(y=residuals(lm8), x=na.omit(healthsubset)$log_child_mortality_rate) #Fixed funnelling and
achieved equal variance

plot(y=residuals(lm8), x=na.omit(healthsubset)$percent_food_insecure) #Funnelling pattern

summary(lm8)

healthsubset$log_percent_food_insecure = log(healthsubset$percent_food_insecure)

lm9 = lm(formula = logcasechange ~ percent_rural + loghouseholds +
    logmedincome + log_percent_low_birthweight + percent_vaccinated +
    log_infant_mortality_rate + log_percent_uninsured + life_expectancy +
    log_average_daily_pm2_5 + log_child_mortality_rate + log_percent_food_insecure,
    data = na.omit(healthsubset))

plot(y=residuals(lm9), x=na.omit(healthsubset)$log_percent_food_insecure) #Fixed funnelling
and achieved equal variance

plot(lm9,which = 2)
hist(residuals(lm9))
plot(residuals(lm9), fitted(lm9))
shapiro.test(residuals(lm9))

```
We see that we satisfy equal variance and linearity conditions, but we have not acheived normality. We will check for if any variables previously excluded should now be included.

Step 14: Perform variable selection test again but for all newly transformed variables

```{r}

lm9.upper_cases = lm(data = na.omit(healthsubset),logcasechange ~ percent_rural +
loghouseholds + logmedincome + log_percent_low_birthweight + percent_vaccinated +
log_infant_mortality_rate + log_percent_uninsured + life_expectancy +
```

```
log_average_daily_pm2_5 + log_child_mortality_rate + log_percent_food_insecure +
log_percent_uninsured + percent_smokers + primary_care_physicians_rate +
percent_65_and_over)
step(lm9, scope = list(lower = lm9, upper = lm9.upper_cases))
```

```

Now we should include percent_smokers.

```{r}

lm10 = lm(formula = logcasechange ~ percent_rural + loghouseholds +
    logmedincome + log_percent_low_birthweight + percent_vaccinated +
    log_infant_mortality_rate + log_percent_uninsured + life_expectancy +
    log_average_daily_pm2_5 + log_child_mortality_rate + log_percent_food_insecure +
    percent_smokers, data = na.omit(healthsubset))

plot(y = residuals(lm10), x = na.omit(healthsubset)$percent_smokers) #It appears there is equal
variance with a few outliers.

shapiro.test(residuals(lm10)) #Still have not solved the normality problems

```

Now we will test interaction terms with variable selection.

#Step 15: Use step function with interaction terms with log median income.

```{r, echo=FALSE}

library(MASS)

help(stepAIC)
lm10interactions = stepAIC(lm10, scope=list(upper= logcasechange ~ percent_rural +
loghouseholds + logmedincome + log_percent_low_birthweight + percent_vaccinated +
log_infant_mortality_rate + log_percent_uninsured + life_expectancy +
log_average_daily_pm2_5 + log_child_mortality_rate + log_percent_food_insecure +
percent_smokers + logmedincome:percent_rural + logmedincome:logmedincome +
logmedincome:log_percent_low_birthweight + logmedincome:percent_vaccinated +
logmedincome:log_infant_mortality_rate + logmedincome:log_percent_uninsured +
logmedincome:life_expectancy + logmedincome:log_average_daily_pm2_5 +
logmedincome:log_child_mortality_rate + logmedincome:log_percent_food_insecure +
```
```

```
logmedincome:percent_smokers, lower= ~1))

```

Step 16: Check normality with new terms considered.

```{r}

summary(lm10interactions)
shapiro.test(residuals(lm10interactions))
plot(lm10interactions, which = 2)
hist(residuals(lm10interactions))

```

We still did not achieve normality, so we must remove outliers.

Step 17: Remove influential points

```{r}
cooks = cooks.distance(lm10interactions)
n = nrow(na.omit(healthsubset))

plot(lm10interactions, which = 4)
abline(h = 1, lty = 2)

highLeverage = cooks.distance(lm10interactions)> (4/n)
highLeverage

healthsubset.new = na.omit(healthsubset)[!highLeverage,]

lm10interactions.new = lm(logcasechange ~ percent_rural + loghouseholds + logmedincome +
   log_percent_low_birthweight + percent_vaccinated + log_percent_uninsured +
   life_expectancy + log_child_mortality_rate + log_percent_food_insecure +
logmedincome:log_child_mortality_rate + percent_rural:logmedincome +
logmedincome:life_expectancy + logmedincome:log_percent_food_insecure, data =
healthsubset.new)
lm10interactions.lower = lm(logcasechange ~ 1, data = healthsubset.new)

shapiro.test(residuals(lm10interactions.new))
summary(lm10interactions.new)
summary(lm10interactions)

which(highLeverage == TRUE)
```

```
hist(residuals(lm10interactions.new)) #Looks relatively normal
plot(lm10interactions.new, which = 2)
```

```

## Step 18: AIC with Influential Points Removed

```{r}

retestlower = lm(logcasechange ~ 1, data = na.omit(healthsubset.new))

step(retestlower, scope = list(lower = retestlower, upper = lm10interactions.new))

```
We see that percent rural was what previously controlling for some of the effect of the influential points. Now percent rural is not selected because the influential points have been removed.

Step 18: Add New Variables and Remove Influential Points from all dataset.

```{r}

finalhealthmodel = lm(formula = logcasechange ~ loghouseholds +
log_percent_low_birthweight +
    log_percent_food_insecure + log_percent_uninsured + life_expectancy +
    logmedincome + percent_vaccinated + log_child_mortality_rate +
    log_percent_food_insecure:logmedincome + logmedincome:log_child_mortality_rate +
    life_expectancy:logmedincome, data = na.omit(healthsubset.new))

plot(residuals(finalhealthmodel), fitted(finalhealthmodel)) #satisfies equal variance and linearity
assumption

alldata$logcasechange = log(alldata$case_change + 1.5)
alldata$loghouseholds = log(alldata$num_households)
alldata$logmedincome = log(alldata$median_household_income)
alldata$log_percent_low_birthweight = log(alldata$percent_low_birthweight)
alldata$log_percent_uninsured = log(alldata$percent_uninsured)
alldata$log_average_daily_pm2_5 = log(alldata$average_daily_pm2_5)
alldata$log_child_mortality_rate = log(alldata$child_mortality_rate)
alldata$log_percent_food_insecure = log(alldata$percent_food_insecure)

head(alldata)

basemobilitymodel = lm(formula = logcasechange ~ loghouseholds +
log_percent_low_birthweight +
```
```

```
    log_percent_food_insecure + log_percent_uninsured + life_expectancy +
    logmedincome + percent_vaccinated + log_child_mortality_rate +
    log_percent_food_insecure:logmedincome + logmedincome:log_child_mortality_rate +
    life_expectancy:logmedincome, na.omit(alldata))

mobilitymodel = lm(formula = logcasechange ~ loghouseholds + log_percent_low_birthweight +
    log_percent_food_insecure + log_percent_uninsured + life_expectancy +
    logmedincome + percent_vaccinated + log_child_mortality_rate +
    log_percent_food_insecure:logmedincome + logmedincome:log_child_mortality_rate +
    life_expectancy:logmedincome + workplaces_pctpoint_change +
retail_and_rec_pctpoint_change + grocery_and_pharmacy_pctpoint_change +
parks_pctpoint_change + transit_stations_pctpoint_change + residential_pctpoint_change,
na.omit(alldata))


cooks = cooks.distance(basemobilitymodel)
n = nrow(na.omit(alldata))

plot(basemobilitymodel, which = 4)
abline(h = 1, lty = 2)

highLeverage = cooks.distance(basemobilitymodel)> (4/n)
highLeverage

alldata.new = na.omit(alldata)[!highLeverage,]

basemobilitymodel.new = lm(formula = logcasechange ~ loghouseholds +
log_percent_low_birthweight +
    log_percent_food_insecure + log_percent_uninsured + life_expectancy +
    logmedincome + percent_vaccinated + log_child_mortality_rate +
    log_percent_food_insecure:logmedincome + logmedincome:log_child_mortality_rate +
    life_expectancy:logmedincome, data = alldata.new)

mobilitymodel.new = lm(formula = logcasechange ~ loghouseholds +
log_percent_low_birthweight +
    log_percent_food_insecure + log_percent_uninsured + life_expectancy +
    logmedincome + percent_vaccinated + log_child_mortality_rate +
    log_percent_food_insecure:logmedincome + logmedincome:log_child_mortality_rate +
    life_expectancy:logmedincome + workplaces_pctpoint_change +
retail_and_rec_pctpoint_change + grocery_and_pharmacy_pctpoint_change +
parks_pctpoint_change + transit_stations_pctpoint_change + residential_pctpoint_change, data =
alldata.new)

shapiro.test(residuals(basemobilitymodel.new))
```

```
summary(basemobilitymodel)
summary(basemobilitymodel.new) #R-Squared still goes up


which(highLeverage == TRUE)
```

```

## Step 19: General F-test for mobility variables

```{r}

summary(basemobilitymodel.new)
summary(mobilitymodel.new)
anova(basemobilitymodel.new, mobilitymodel.new)
```

```

We see that w/ 95% confidence, we know that at least one mobility variable is significant.

## Step 20: Output tables

```{r}

library(stargazer)
help(stargazer)

#Health Model Building
table = stargazer(lm10, lm10interactions, lm10interactions.new, finalhealthmodel, title = "Health Models", align = TRUE, type ="html")

#Mobility Analysis

table = stargazer(basemobilitymodel, mobilitymodel, basemobilitymodel.new, mobilitymodel.new, title = "Mobility Results", align = TRUE, type ="html")

#Final Models


table = stargazer(finalhealthmodel, mobilitymodel.new, title = "Final Results", align = TRUE,
```

type ="html")

```

## Step 21: Maps

```{r}
library(usmap)

states = alldata$state

plot_usmap(regions = c("states"), data = alldata,
          values = "case_change", color = "light gray", labels = FALSE,
          label_color = "black") + scale_fill_continuous(low = "pink", high = "red", na.value = "
gray", name ="", label = scales::comma) +
  theme(legend.key.width = unit(2, "line"), legend.key.height = unit(3, "line"), legend.position =
"right", legend.title = element_text(size=16))


states = healthsubset.new$state

plot_usmap(regions = c("states"), data = healthsubset.new,
          values = "case_change", color = "light gray", labels = FALSE,
          label_color = "black") + scale_fill_continuous(low = "pink", high = "red", na.value = "
gray", name ="", label = scales::comma) +
  theme(legend.key.width = unit(2, "line"), legend.key.height = unit(3, "line"), legend.position =
"right", legend.title = element_text(size=16))

states = alldata$state

plot_usmap(regions = c("states"), data = alldata.new,
          values = "case_change", color = "light gray", labels = FALSE,
          label_color = "black") + scale_fill_continuous(low = "pink", high = "red", na.value = "
gray", name ="", label = scales::comma) +
  theme(legend.key.width = unit(2, "line"), legend.key.height = unit(3, "line"), legend.position =
"right", legend.title = element_text(size=16))


```
```