W203 Lab 2 - COVID-19 Modeling

Jun Qian, Lucas Schroyer, Ryan Mitchell, Oliver Chang

4/14/2021

Introduction

As of the writing of the document, the COVID-19 coronavirus (COVID-19) has been spreading throughout the United States for nearly 14-15 months, with the initial cases identified as having entered the country in January 2020. This report uses data from the United States Census Bureau (including state level demographics and county level population and population density data), the New York Times for COVID-19 case counts, a Google dataset on state-level mobility data, and related COVID-19 policy data from the US State Policy Database. All data used in the project was pulled in on April 10th, 2021.

Our team's primary research question was "How does mobility impact the spread of COVID-19?" To begin to answer that question, our research team decided to conduct an exploratory observational analysis using OLS regression to measure the complex relationships that exist between changes in a states' population mobility and state-level COVID-19 case counts per 100K people in the 365 days following each state's declaration of a state of emergency (SOE). As part of this endeavor, we also explore whether other variables, such as state age demographics, population density, and state level policies on mask mandates, stay at home orders, quarantine restrictions, enhanced unemployment benefits, and business closures might also have an impact on COVID case counts.

This research question was initially motivated as an attempt to understand if the preventive measures that states have enacted in response to the pandemic were associated with statistically significant changes in case counts. Initially, we had hoped to find a causal relationship between mobility and COVID case counts. However, we ultimately decided against this because of the high likelihood of reverse causality between these variables (i.e. a change in mobility causes a change in case counts which in turn will cause a change in mobility). Additionally, given how complex the nature of pandemics are, there was a strong possibility for omitted variable bias (such as state level differences in temperature and humidity, behavioral differences in terms of mask compliance, COVID testing availability, the percentage and absolute numbers of people using mass transit, and the percentage and absolute numbers of people who are able to work from home, to name a few). These omitted variables will be discussed in the model limitations section of this report. We decided to focus on COVID cases as our dependent variable, as opposed to COVID deaths, because we inferred that COVID deaths counts are directly dependent on the number of COVID cases and other causal inputs, such as genetic predisposition, underlying health conditions/comorbidities, hospital utilization rates, the availability of various treatment options (ventilators, experimental drugs, etc.). Given COVID cases are a direct input to COVID deaths and much of the data needed to analyze these supplemental variables is not easily available or likely does not exist, a study on COVID death counts may suffer from substantial omitted variable bias. This would raise questions about the interpretation of our model in terms of both statistical and practical significance.

Our motivation for normalizing the dependent variable per 100K people 365 days after each state's declaration of a state of emergency was as follows:

- Absolute population counts vary substantially from state to state, and population being the most important factor in absolute COVID case counts is not an interesting finding.

- One year was a natural ending point given that we recently passed the one year mark for each state's declaration of a state of emergency. Additionally, we believe that the mere act of the state declaring an emergency may have contributed to shifts in the behaviors of the residents of each state, so we wanted to index our analysis against that moment in time.

We decided to regress on population mobility as our key variable of interest because:

- Although we did not conduct a causal study, mobility fits well into a causal regression framework for COVID cases, because physical proximity is a requirement for disease transmission, and proximity is a function of mobility (and population density, which we will also explore in our modeling efforts). As such, we expected that it could have a strong relationshio with case counts and thus included it in our exploratory study.
- The mobility data captures the actual effects of multiple correlated policies intended to reduce COVID transmission (such as stay-at-home orders, quarantine requirements after traveling or possible exposure to a person who tested positive, business and school closures, etc.). This is effectively a method of dimensionality reduction that contributes to model parsimony.

Feature Selection for OLS Regression

For this study, we aggregated the different data sources across their individual study horizons (in most cases, more than one year) to generate a small sample of 50 state-level observations. The table below summarizes the nine features we ultimately included for our descriptive OLS regression models to understand their association with the dependent variable (i.e. case count per 100,000 people):

Feature	Present in Which Model	Source
Mobility (% change)	1, 2, 3	Google Dataset
Population Density (People/sq. mi.)	1, 2, 3	US Census Data
Percent of Population Under the Age of 25 (%)	2, 3	US Policy Database
Mask Mandate Days	2, 3	US Policy Database
Unemployment Benefit Days	3	US Policy Database
Increased Weekly Unemployment Insurance Amount Through July 31	3	US Policy Database
Business Close To Open Days	3	US Policy Database
Travel Quarantine Mandate Days	3	US Policy Database
Stay at Home Days	3	US Policy Database

`Case Count Per 100,000 people`: This is our outcome variable, operationalized as discussed in the introduction. We decided to normalize the case counts around the population as we were concerned that a OLS regression absolute case count would not result in any meaningful association discoveries.

`Mobility (% Change)`: We decided to regress on population mobility because physical proximity is a requirement for disease transmission and because mobility data captures the actual effects of multiple correlated policies intended to reduce COVID transmission (such as stay-at-home orders, quarantine requirements after traveling or possible exposure to a person who tested positive, business closures, etc.). This is effectively a method of dimensionality reduction that contributes to model parsimony. This Google dataset we used tracks the changes in mobility for the following sectors: (1) Grocery and Pharmacy, (2) Parks, (3) Residential, (4) Retail and Recreation, (5) Transit Stations, and (6) Workplaces. Ultimately, we decided to use the percent change in transit mobility in our models for two primary reasons (described in detail later). Firstly, most of the other mobility features are highly correlated with transit. Secondly, we couldn't determine a

method for aggregating these features into a statewide change in mobility data as the raw data was captured in percent changes and not absolute changes. We theorized that changes in transit station mobility would be the most associated with changes in case counts from this list as transit stations are often identified as major spreaders of disease.

`Population Density (People/sq. mi.)`: Normalized the state's population by its area (given in square miles).

`Percent of Population Living in a High Population Density County (%)`: While normalizing a state's case count by the population is useful, we decided to add an additional feature illustrating what percentage of the population lives in a high density county, which we defined as the top 50 counties in the state when ranked by county population per square mile. An illustrative example of why we added this feature can be seen in the state of New York, which was one of the earliest states to be plagued with the virus. According to the New York Times COVID-19 tracking database, New York City alone has had 882K of New York's 1.94M (45%) total cases since the beginning of the pandemic (as of April 10th, 2021).

`Percent of Population Under the Age of 25 (%)`: The media and research studies alike have suggested that young adults and children are less likely to be impacted by the virus or may be asymptomatic compared with older adults. As a result, this young age group is more likely to continue spreading COVID-19. We hope to uncover a relationship between a state's percentage of the population with young people and a corresponding change in case counts.

`Mask Mandate`: the number of days that a state had a mask mandate.

Other model 3 policy features: We selected the following policy-related features from the COVID-19 US State Policy Database (www.tinyurl.com/statepolicies) that we thought related to COVID cases including, the date the state of emergency declared, length of mask mandate, length of stay-at-home orders, length of business closure, length of travel quarantine mandate, length of increased unemployment benefits, and increased unemployment insurance amount. All the policy-related features were recorded as dates, except for the increased unemployment insurance amount that was recorded as integers. Ultimately, what we were interested in was the total days a given policy was active for during the one year period after a state of emergency was announced for each state.

Initial Data Loading and Cleaning

The primary data loading and cleaning steps for this report included the following:

1. First, we read in the NYT COVID database from an excel workbook (see the data folder in our repository for all of the raw data files used in this analysis).

```
## Pull data from NYT COVID Database and plot summary
NYT_Data <- fread("https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv")

## Convert date strings to dates.
NYT_Data[,date:=as.Date(date)]

## Calculate cases on a time interval
# start.date <- as.Date("2021-01-01")
# end.date <- as.Date("2021-02-28")

interval.cases <- NYT_Data %>%
  arrange(state, date) %>%
#          date == start.date) %>%
  group_by(state) %>%
```

```
  mutate(cases_inc = cases - lag(cases),
         deaths_inc = deaths - lag(deaths)) %>%
  ungroup() %>%
  filter(cases_inc >= 0)
```

2. Then, we read in the policy data from an excel workbook downloaded from the US State Policy Database.

```
tab_names <- excel_sheets(path = "data/US_Covid_Policy_data.xlsx")

list_all <- lapply(tab_names,
                   function(x) read_excel(path = "data/US_Covid_Policy_data.xlsx",
                                                    sheet = x))
#Rename list elements
names(list_all) <- tab_names %>%
    tolower() %>%
    gsub(pattern = "_", replacement = " ") %>%
    str_to_title() %>%
    gsub(pattern = " ", replacement = ".")

#Rename colums in the DF
for (df in 1:length(list_all)){
  name_qc <- names(list_all)[df]
  #print(name_qc)
  if(name_qc %in% c("Stay.At.Home", "Unemployment.Benefits")){
    names(list_all[[df]]) <- list_all[[df]][1,]
    list_all[[df]] <- list_all[[df]][-1,]
  }

  names(list_all[[df]]) <- gsub(" ", "_", names(list_all[[df]]) %>% tolower())
}
```

Note: the rest of the R code used to read in the remaining observations/datasets has been suppressed for report readability. From this point onward, we only print code that is necessary for the report narrative. Please refer do this project's GitHub repository for the full R-markdown file

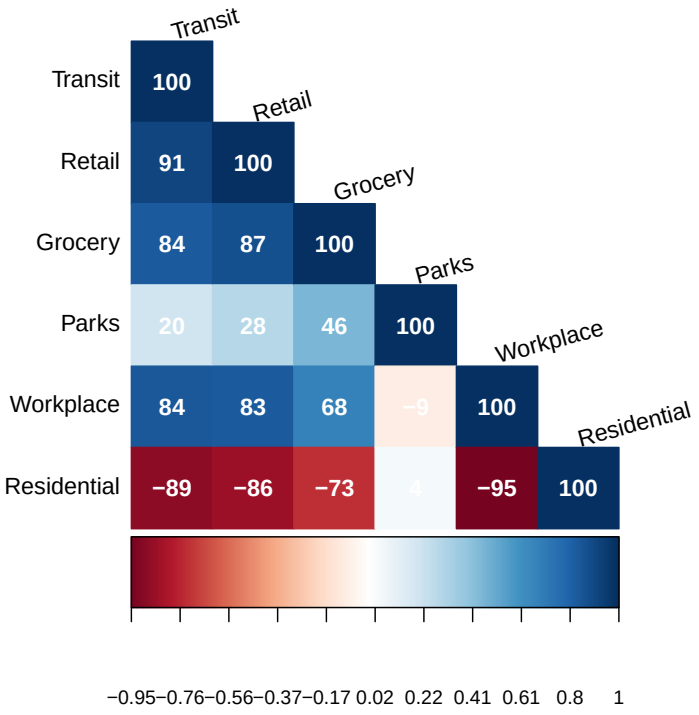
3. Then, we read in the Google dataset (from an excel workbook) on state-level changes in mobility data.
4. Next, we read in three spreadsheets available for download from the U.S. Census Bureau to obtain county-level demographics, including population and area.
5. Now that we have read in all of the raw datasets we will use in this study, we joined the mobility data with the county-level population and area data.
6. With the county level population observations and mobility observations, we then calculated the weighted average change in mobility at the state level.
7. Finally, we join the state-level mobility observations with the added population and percentage of population in a high density area features to the NYT Covid database by state and date.

Exploratory Data Analysis

The first step of our exploratory data analysis (EDA) was to determine which mobility metric to use from the Google mobility dataset in order to answer our intial secondary research question, which was: "How does

mobility impact the spread of COVID-19?" We began by exploring the correlations that existed in between the Google observations between Transit, Retail, Grocery, Workplace, Residential, and Parks.

Correlation Matrix: Median Mobility Changes by Category



After inspection of the correlation matrix (above) between the different measures of mobility changes, we chose to use the state-level median change in transit mobility in the 365 days after each state declared an emergency as our main variable of interest for Model 1. Our motivation for choosing transit over the other mobility features was as follows:

1) Transit is most closely aligned with our understanding of how viruses spread, particularly from one locality or population center to another.

2) The mobility changes in Transit, Retail, Grocery, and Workplace are highly positively correlated with each other (>0.68 in each pair), and highly negatively correlated with Residential mobility changes (absolute value of >0.7 or above). Highly correlated features should be avoided in descriptive and explanatory linear regression modeling as they tend to increase the standard error estimates on the model parameter estimates for the correlated features.

3) We did not want to take an aggregation of the set of highly correlated features (Transit, Retail, Grocery, and Workplace) because the raw data provided are relative numbers and we do not have access to the underlying absolute mobility data, so we would be unable to derive a correct weighted average of these features.

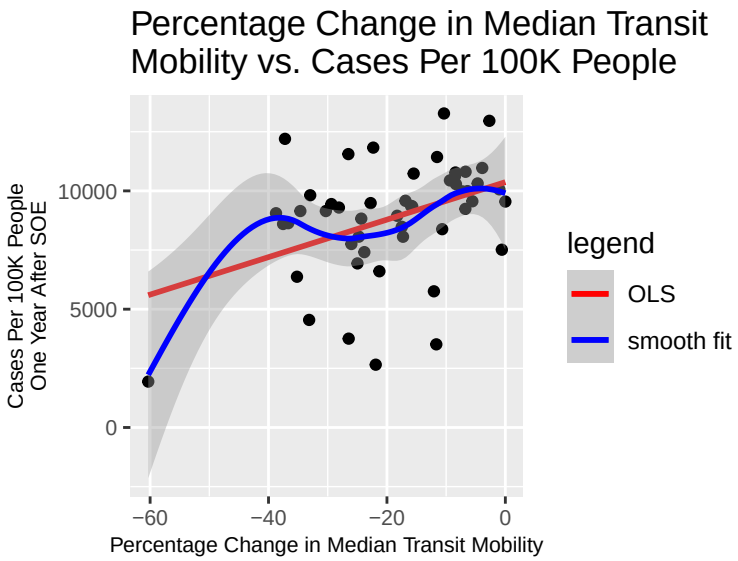
4) Given that the distribution of relative transit mobility change had a left skew, our team decided to use the median value for each state within the 365 day window as a better measure of central tenancy.

5) The only mobility metric that was not strongly correlated with the others was Parks. Since the median change in Park mobility was not strongly correlated with any other mobility changes, we examined its

relationship with our target variable but did not observe a positive or negative linear relationship. A t-Test on our calculated simple regression coefficient failed to reject the null hypothesis that there was no evidence that the coefficient for change in median Parks mobility was measurably different than zero.

```
## Park Mobility and COVID CASES per 100K People Regression for EDA

##
## Call:
## lm(formula = cases_per_100k_at_365d ~ parks, data = final_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6177.2  -861.8   528.6  1468.5  4564.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8711.05     429.77  20.269   <2e-16 ***
## parks          13.17      22.20   0.593    0.556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2468 on 48 degrees of freedom
## Multiple R-squared:  0.007277,   Adjusted R-squared:  -0.0134
## F-statistic: 0.3519 on 1 and 48 DF,  p-value: 0.5559
```

Having decided on using the median change in transit mobility, the next step in our EDA was to examine what type (if any) of relationship existed between transit and our outcome of interest, which is COVID cases per 100K people. The scatter plot below shows that there is a positive linear relationship between change in median transit and COVID cases. Consequently, we decided to move forward with change in median transit as the key variable for our Model 1.



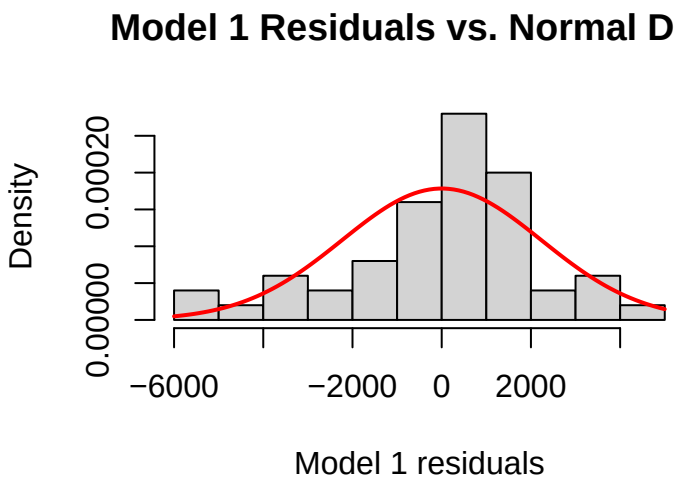
Model 1 - OLS Regression Development and Continued EDA

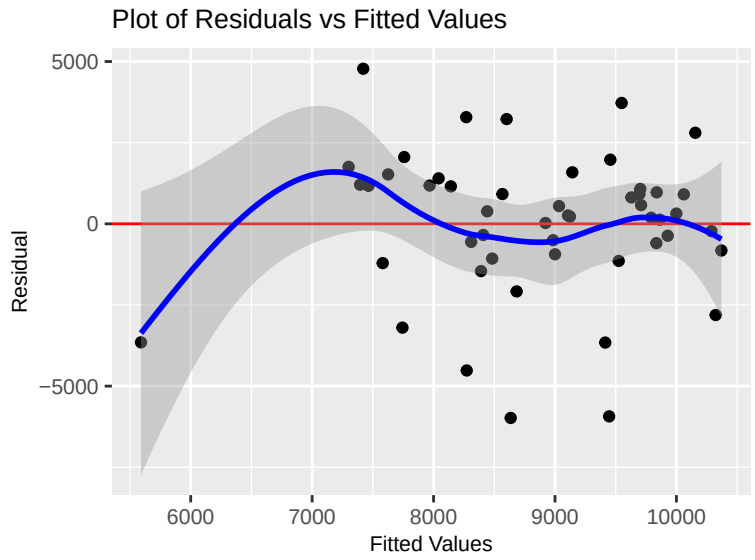
For our first model that explores the association between transit mobility and changes in COVID case counts, we used R's lm function to create an OLS regression between case counts conditional on percentage change in median transit mobility. We then generated some summary plots and statistics to analyze the result of our OLS regression, including:

- summary statistics of the OLS regression
- the distribution of the error terms
- the scatter plot of the fitted values with the residuals
- the Breusch-Pagan test to assess Homoskedasticity
- the calcualted robust standard errors for the OLS
- Shapiro-Wilk normality of residuals test

Throughout the report, we continued to run these tests (in additon to some other graphics and statistics). The result summary for these include:

```
##
## Call:
## lm(formula = cases_per_100k_at_365d ~ median_transit_change,
##     data = final_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5981.0  -909.3   286.3  1177.2  4778.3
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           10370.16     580.47  17.865  < 2e-16 ***
## median_transit_change    79.25      25.43   3.117  0.00309 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2259 on 48 degrees of freedom
## Multiple R-squared:  0.1683, Adjusted R-squared:  0.151
## F-statistic: 9.714 on 1 and 48 DF,  p-value: 0.003086
```






```
##
## -------Homoskedasticity Test-------

##
##  studentized Breusch-Pagan test
##
## data:  model_1_final
## BP = 1.5589, df = 1, p-value = 0.2118

##
## -------Robust Standard Errors-------

##           (Intercept) median_transit_change
##             610.18386              32.94262
```

```
##
## ------Normality of Residuals Test-------

##
##  Shapiro-Wilk normality test
##
## data:  model_1_final$residuals
## W = 0.94719, p-value = 0.02617
```

Model 1 Discussion

The first parameter estimate returned by Model 1 is an intercept estimate of 10,370, which represents the expected number of COVID-19 cases per 100K people (in the first year after declaration of a state of emergency) if there were no change in transit mobility levels. Although the Breusch-Pagan test does not indicate that our Model 1 suffers from heteroskedasticity, we will err on the side of caution and use robust standard errors for claims of statistical significance throughout this report. The robust standard error on the intercept estimate is 610, which suggests that the intercept parameter estimate is statistically significant using a standard Type I error rate of 0.05.

The results from Model 1 also tell us that a one percent increase in median transit mobility is associated with 79 additional COVID-19 cases per 100K people. Equivalently, we can state that a one percent decrease in median transit mobility is associated with 79 fewer COVID-19 cases per 100K people. The robust standard error is 33, which indicates that the transit mobility parameter estimate is statistically significant using a standard Type I error rate of 0.05.

Taken together, these two parameter estimates tell a story about what could have happened if people did not change their transit behavior in response to COVID-19. The observed average number of COVID-19 cases per 100K people was 8,860, which is lower than the intercept estimate of 10,370 because 1) on average, states recorded a 19% decrease in median transit mobility and 2) each percent decrease is associated with 79 fewer cases per 100K people. This paints a promising picture about how behavior and policy changes may reduce the spread of the virus.

Although the Shapiro-Wilk normality test returned a p-value that rejects the null hypothesis of normally distributed residuals, we would like to note that we will be adding additional features to this model specification that should address the non-normality concern (discussed in detail in later sections).

Model 2 - OLS Regression Development and Continued EDA

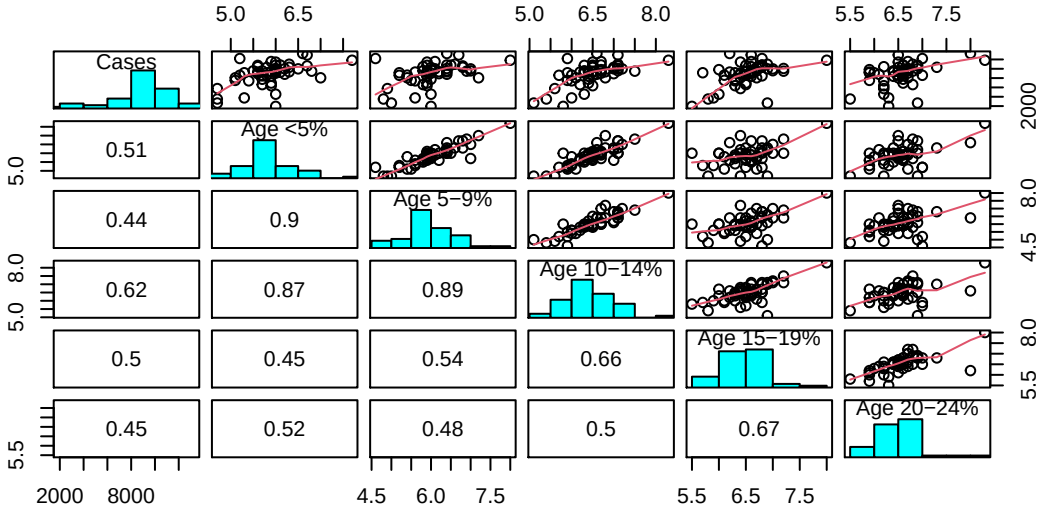
As we thought about an approach to building our Model 2 specifications, the general approach our group decided to take was to incrementally test the addition of new features to our Model 1 specification in descending order of expected importance, according to our general understanding of how viruses spread. At each iteration, we began by examining the relationship (using scatterplots) between a new demographic or policy-related feature and our outcome of interest: COVID cases per 100K people 365 days after each state declared a state of emergency. If we observed a relationship, we ran a combination of t-Tests and/or ANOVA F-Tests to determine whether or not to add the feature to our Model 1 specification. We proceeded in this 'greedy' algorithmic fashion by adding feature to our Model 1 specification until we could no longer justify further additions based on results from ANOVA F-Tests. The first incremental feature we tested was derived from observational data from the census bureau on population age distributions by state.

One of the contributing factors to the spread of COVID-19 is asymptomatic spread. Younger people have been shown to have milder symptoms and therefore it stands to reason that they may be less likely to get tested, and ultimately end up spreading the disease at a greater rate than older age groups. Hence, for our second model, we were interested in testing a general hypothesis that age demographics may play a role in the

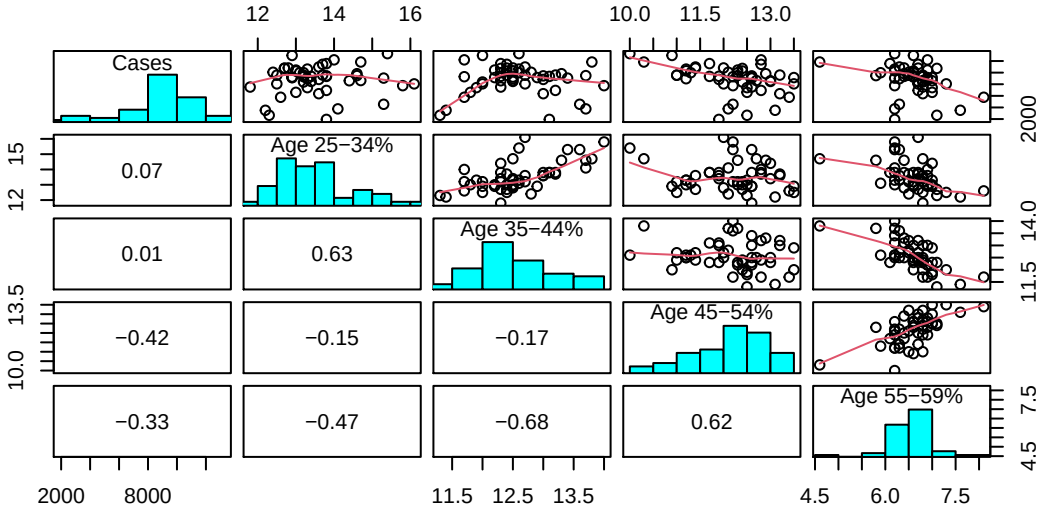
spread of COVID-19. We began by doing some basic EDA with respect to age distributions and our target variable, cumulative COVID-19 case counts per 100K. We include the EDA code for our first added feature for the reader's reference. We omit the rest for readability but the steps follow the same basic pattern.

```
#Pull the features of interest from our master data frame (i.e. percent age < 25)
census_df<-final_df[c('cases_per_100k_at_365d',
                      'Percent.Sex.And.Age.Total.Population.Under.5.Years',
                      'Percent.Sex.And.Age.Total.Population.5.To.9.Years',
                      'Percent.Sex.And.Age.Total.Population.10.To.14.Years',
                      'Percent.Sex.And.Age.Total.Population.15.To.19.Years',
                      'Percent.Sex.And.Age.Total.Population.20.To.24.Years')]

# rename columns
names(census_df) <- names(census_df) %>%
  gsub(pattern = "Percent.Sex.And.Age.Total.Population.",
       replacement = "Pct. ") %>%
  gsub(pattern = ".To.", replacement = "- \n") %>%
  gsub(pattern = ".Years", replacement = " Yrs") %>%
  gsub(pattern = 'cases_per_100k_at_365d',
       replacement = " Cases per \n 100k")

# Create a series of plots illustrating the feature's observation distribution
# and a scatter plot with other features to assess colinearity
panel.hist <- function(x, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}

panel.cor<-function(x,y)
  {
    usr<-par("usr"); on.exit(par(usr))
    par(usr=c(0,1,0,1))
    r=round(cor(x,y),digits=2)
    text(0.5,0.5,r)
  }

# census_df %>% names()
pairs(census_df,lower.panel = panel.cor,
      upper.panel = panel.smooth,
      diag.panel = panel.hist,
      cex.labels=1,
      gap = 0.5,
      labels = c("Cases", "Age <5%", "Age 5-9%", "Age 10-14%",
                 "Age 15-19%", "Age 20-24%"),
      main = "Case vs Age Bins Distribution and Scatter Plots")
```

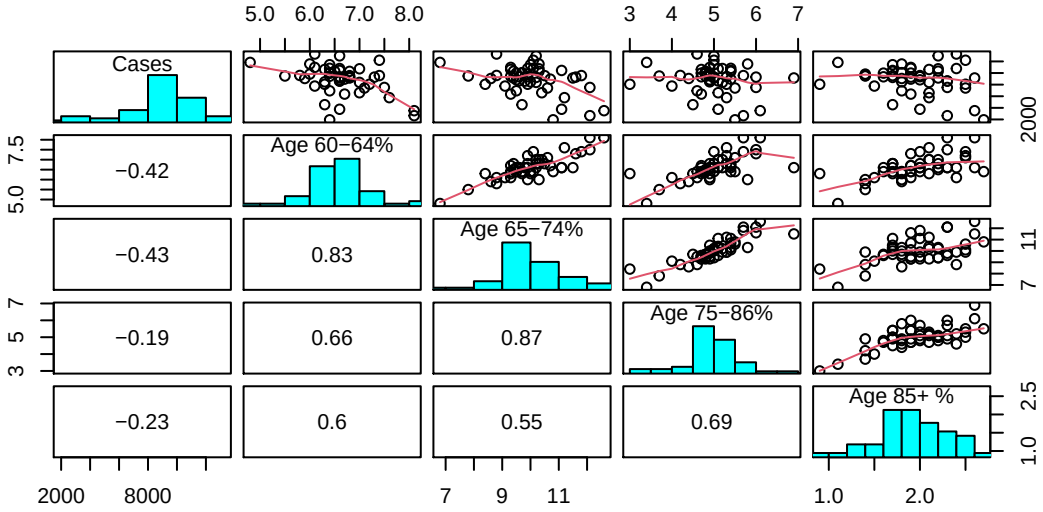
Case vs Age Bins Distribution and Scatter Plots



Case vs Age (Contin.) Bins Distribution and Scatter Plots



Case vs Age (Contin. 2) Bins Distribution and Scatter Plots



When performing EDA on the categorical age group distributions and their relationship to our target variable, we noticed a strong positive correlation to the target among age groups 0-24. Age groups in the 24+ range did not appear to follow a consistent pattern with respect to correlation with our target variable and were therefore not a key focus area for our analysis.

Next, we performed a series of ANOVA F-tests on a nested set of linear regression models with parameter estimates for this set of age groups (Under 5, 5-9, 10-14, 15-19, and 20-24) as well as our main variable of interest (median transit mobility change). Our motivation here was to understand whether the regression residuals were measurably different from one another between the different (nested) model specifications. For context, the null hypothesis for an F-test is that fitting additional coefficients for a longer model does not measurably reduce the residuals relative to a nested model with fewer parameters.

Our first ANOVA F-test compared our Model 1 with a new model that had an additional parameter estimate for the percentage of the population under 5 years old. With a p-value of 0.0003, we rejected the null hypothesis and used this new model as the baseline model for subsequent comparisons with additional parameter estimates for different age categories.

As shown in the Anova tables below, we failed to reject the null hypothesis when adding an estimator for ages 5-9, and succeeded in rejecting the null hypothesis when adding an estimator for ages 10-14, with a p-value of 0.001. Adding additional estimators for 15-19 and 20-24 failed to reject the null hypothesis that these models were measurably better at reducing residuals. These results suggested creating two new features for the percentage of the population ages 0-9 and 10-24 and adding them to our base Model 1 to create the first specification for our Model 2.

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years
## Model 2: cases_per_100k_at_365d ~ median_transit_change
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 47 184378849
## 2 48 245004785 -1 -60625936 15.454 0.0002765 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

```
' ' 1

## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years
## Model 2: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 46 170183297
## 2 47 184378849 -1 -14195553 3.837 0.05621 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years
## Model 2: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 45 132785225
## 2 47 184378849 -2 -51593624 8.7424 0.00062 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years +
## Percent.Sex.And.Age.Total.Population.15.To.19.Years
## Model 2: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 44 132369352
## 2 45 132785225 -1 -415873 0.1382 0.7118

## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years +
## Percent.Sex.And.Age.Total.Population.15.To.19.Years +
```

```
Percent.Sex.And.Age.Total.Population.20.To.24.Years
## Model 2: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years +
## Percent.Sex.And.Age.Total.Population.15.To.19.Years
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 43 124070515
## 2 44 132369352 -1 -8298836 2.8762 0.09713 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1


## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years +
## Percent.Sex.And.Age.Total.Population.15.To.19.Years +
Percent.Sex.And.Age.Total.Population.20.To.24.Years
## Model 2: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 43 124070515
## 2 45 132785225 -2 -8714710 1.5102 0.2324
```

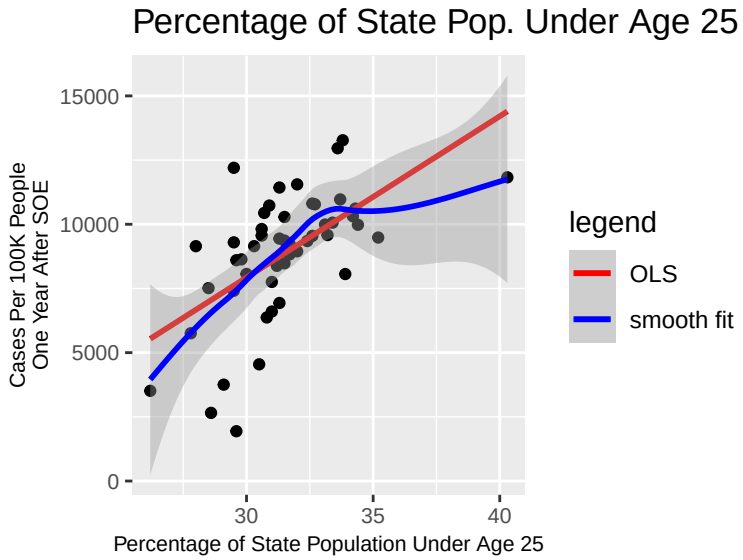
Here we calculate the correlation between our two binned age groups 0-9 and 10-24 years of age.

```
#create two new features that bins together ages 0-9 and 10-24
final_df <- final_df %>%
  mutate(pop_pct_age_0_9 = Percent.Sex.And.Age.Total.Population.Under.5.Years +
           Percent.Sex.And.Age.Total.Population.5.To.9.Years,
         pop_pct_age_10_24 =
           Percent.Sex.And.Age.Total.Population.10.To.14.Years +
           Percent.Sex.And.Age.Total.Population.15.To.19.Years +
           Percent.Sex.And.Age.Total.Population.20.To.24.Years)

#check the correlation between these two age groups
cor(final_df$pop_pct_age_0_9, final_df$pop_pct_age_10_24)
```

```
## [1] 0.7701834
```

```
### strongly correlated - bin these two groups into a new feature
final_df <- final_df %>%
  mutate(pop_pct_age_0_24 = Percent.Sex.And.Age.Total.Population.Under.5.Years +
           Percent.Sex.And.Age.Total.Population.5.To.9.Years +
           Percent.Sex.And.Age.Total.Population.10.To.14.Years +
           Percent.Sex.And.Age.Total.Population.15.To.19.Years +
           Percent.Sex.And.Age.Total.Population.20.To.24.Years)
```

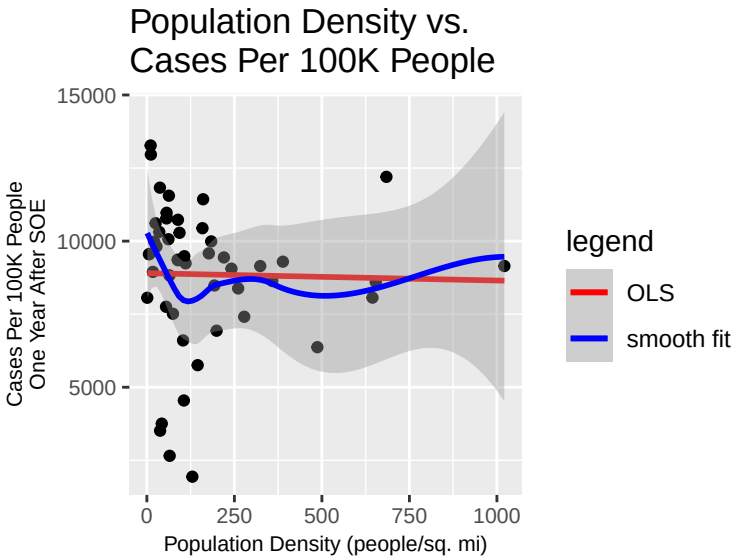
However, after measuring a 0.77 correlation between these two features - and with the goal of increased model parsimony - we decided to group them together to prevent the standard errors for their respective coefficients from increasing substantially. Let's take a look and assess whether this new variable for percentage of the population < 25 years old visually satisfies the conditional linearity expectation with respect to our target.



The conditional linear expectation between our population percentage aged 0-24 and our target appears to be met. Indeed, including a feature for the percentage of the population aged < 25 years old appears to have improved our model's performance. An ANOVA F-test returned a p-value of 0.0001, suggesting that we reject the null hypothesis that the (interim) Model 2's residuals were not measurably different from the residuals of Model 1.

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24
## Model 2: cases_per_100k_at_365d ~ median_transit_change
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1     47 176654365
## 2     48 245004785 -1 -68350419 18.185 9.595e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After accounting for changes in mobility and demographic age differences between states, the next variable we wanted to explore as part of our descriptive model for COVID case counts was population density. According to the World Health Organization, "COVID-19 virus is primarily transmitted between people through respiratory droplets and contact routes." In other words, COVID spreads primarily through physical interactions between infected and uninfected hosts, regardless of whether the actual mechanism of transmission is airborne or surface based contact. Hence, it stands to reason that more densely populated areas would see greater rates of infections, because the frequency of these physical interactions will increase with population density. This was the motivation for our group exploring whether a relationship existed between state population density and our outcome variable of interest.



To our surprise, however, there was no clear relationship between population density and our outcome variable. A t-Test on the parameter estimate for population density's relationship with our target variable failed to reject the null hypothesis that the coefficient value was not measurably different from zero.

```
##
## Call:
## lm(formula = cases_per_100k_at_365d ~ population_density, data = final_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6930.5  -799.2   487.9  1416.9  4372.2
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8902.5248   454.6106  19.583   <2e-16 ***
## population_density   -0.2508     1.6991  -0.148    0.883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2477 on 48 degrees of freedom
## Multiple R-squared:  0.0004535,  Adjusted R-squared:  -0.02037
## F-statistic: 0.02178 on 1 and 48 DF,  p-value: 0.8833
```

In spite of this test, our group decided to move forward and try including the population density feature in Model 2, as we believed it to be a conceptually meaningful variable in describing the population prevalence of COVID in each state one year after each state declared a state of emergency. Therefore, we proceeded with an ANOVA F-Test to test whether this added population density feature measurably improved model performance (via reduction of residuals) relative to our interim Model 2 with features for median transit mobility change and percentage of the population < 25 years old. This ANOVA F-test returned a p-value of 0.002, enough to reject the null hypothesis that the model residuals were not measurably different from one another.

```
##
## -------ANOVA F-Test Significance Test Relative To Interim Model 2-------
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24
##   Res.Df       RSS Df Sum of Sq      F  Pr(>F)
## 1     46 144077125
## 2     47 176654365 -1 -32577240 10.401 0.00232 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the population density feature only becomes statistically significant at the $p < 0.05$ level when we include features in our model for median change in transit mobility and the percentage of the population under age 25, we say that the population density has a conditional relationship with our outcome of interest. Both of these two co-variates (median transit mobility change and percentage of the population under age 25) are negatively correlated with population density and positively correlated with our outcome variable.

However, we are only interested in the unique variation of population density with respect to our outcome variable. When the OLS regression algorithm calculates the parameter estimate for population density, it starts by regressing population density on the other model co-variate (input) features. The residuals from that regression represent the portion of population density that is *not* colinear with median transit mobility change and percentage of the population under age 25. Then OLS regresses our target values on those residuals to derive an estimate for the population density parameter. The model summary tells us that if we hold the percentage of the population under 25 and the median transit mobility change for a state constant (we do not allow them to co-vary), that there exists a positive correlation between population density and our outcome variable at a statistically significant level ($p = 0.002$) using classical standard errors.

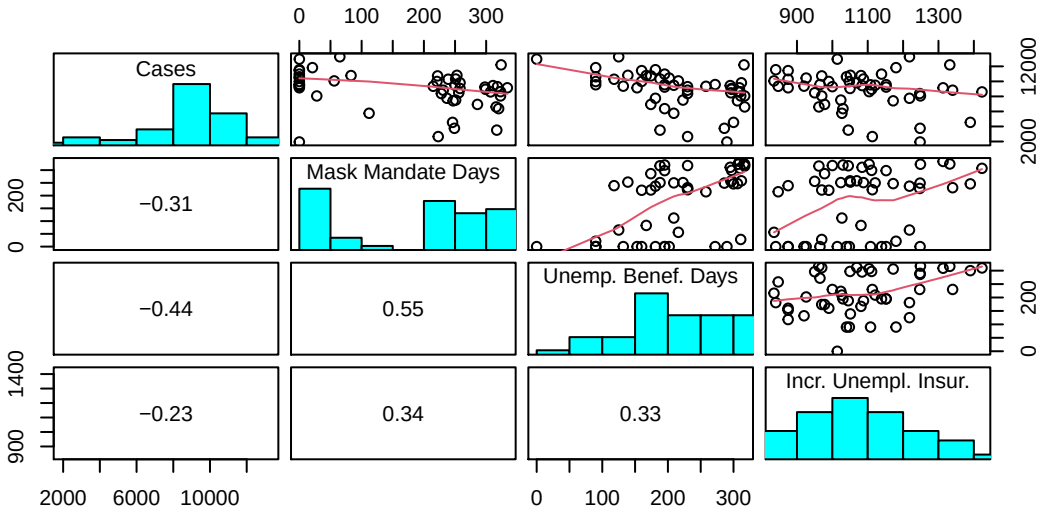
At this point, we wanted to explore whether any state policy changes aimed at reducing the spread of COVID-19 added incremental descriptive power beyond our current Model 2 specification. In particular, we wanted to examine whether the duration of mask mandates, duration and amount of increased government assistance via enhanced unemployment benefits, duration of the first wave of business closures, duration of stay at home mandates, and duration of travel quarantine restrictions had measurable effects on our outcome of interest after accounting for the features already in our Model 2 specification (which included state-level features for median change in transit mobility, percentage of the population under 25 years of age, and population density). All the policy-related features were recorded as dates, except for the increased unemployment insurance amount, which was recorded as an integer.

To align with our target variable of COVID-19 Cases per 100K one year after state of emergency declaration, we encoded the date related policy variables as the total number of days each policy was in place for after the state of emergency was announced for each state (up to 365 days). Steps to transform this total days feature included:

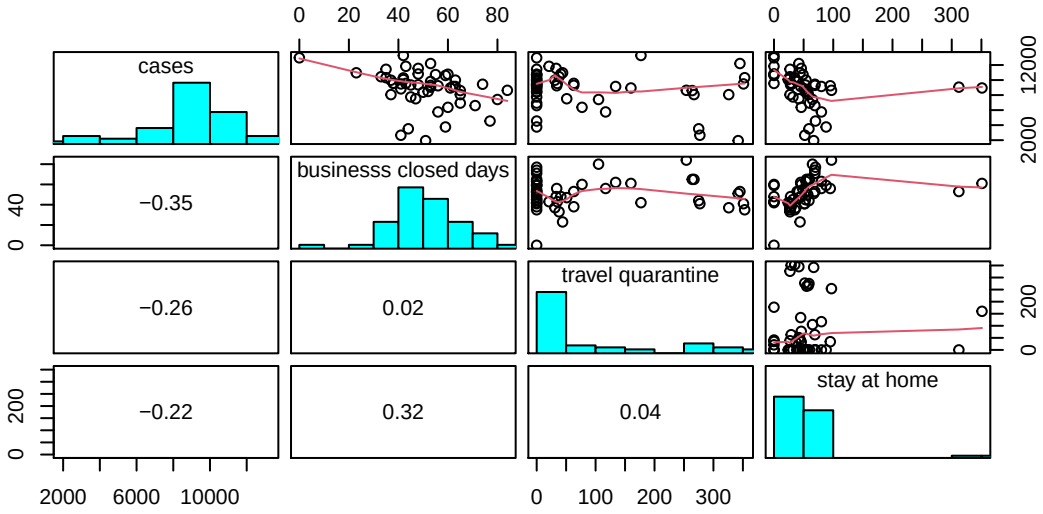
1. If a policy had no beginning and end dates, the total days were assigned as zero.
2. If a policy had beginning but not end dates, the total days were calculated by the date of state emergency declared + 364 days - the date of the beginning of the policy.
3. If a policy had both beginning and end dates, the total days were calculated by the difference in days of the two dates.

Once the policy-related features of our interest were transformed into the days a policy was enforced, we conducted EDA on the policy-features and COVID cases per 100K population using scatter plots and correlation matrices. Here we show the distribution of our policy features and some scatter plots indicitating their colinearity levels.

Policy Days vs Cases Per Capita Distribution and Scatter Plots

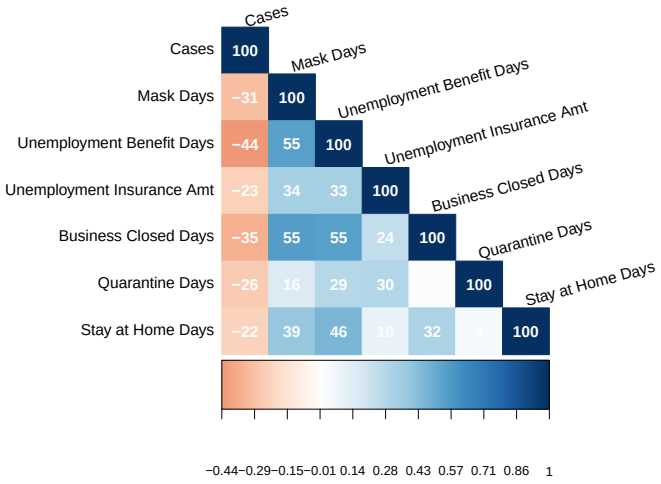


Policy Days (Contin.) vs Cases Per Cap. Distrib. and Scatter Plots



From these scatterplots (above), we see varying degrees of linearity between the policy features and our outcome of interest. Next, we provide a correlation heatmap plot to identify how correlated *all* of the policy features are with each other (the scatter plots above could not fit all the features into one figure).

Correlation Matrix: State Policies vs. Cases



We tested each of these features iteratively in the same manner as before, using significance from ANOVA F-tests as the benchmark to decide whether or not to include incremental policy features as part of our final Model 2.

Failed ANOVA F-Tests on Policy Variables

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + mask_mandate_days
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 138865334
## 2     46 144077125 -1  -5211792 1.6889 0.2004

## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + unemployment_benefits_days
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 137886393
## 2     46 144077125 -1  -6190732 2.0204 0.1621

## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + increased_weekly_unemployment_insurance_amt_thru_jul31
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
```

```
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 135803753
## 2     46 144077125 -1  -8273373 2.7415 0.1047
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + business_closed_days_round1
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 140581017
## 2     46 144077125 -1  -3496108 1.1191 0.2958
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + travel_quarantine_mandate_days
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 140733564
## 2     46 144077125 -1  -3343562 1.0691 0.3067
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + stay_at_home_days
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 144017384
## 2     46 144077125 -1    -59741 0.0187 0.8919
```

Ultimately, none of these policy variables (length of mask mandates, length and amount of increased government assistance via enhanced unemployment benefits, length business closures, length of stay at home mandates, and length of travel quarantine restrictions) returned a p-value that would allow us to reject the null hypothesis that the model residuals had not measurably improved.

Still, it is worth noting that all of the aforementioned policy features, which were declared to mitigate COVID spread, demonstrated a negative correlation (from approximately -0.5 to -0.2) with our outcome variable of COVID cases per 100k people. The fact that these policy features failed to reject the null hypothesis in the ANOVA F-Test against the features already in our Model 2 was not entirely unexpected. Conceptually, several of the features share a significant amount information with median transit mobility change. One could make the argument that business closures, quarantine mandates, and stay at home mandates are all captured, to some extent, in the transit mobility change.

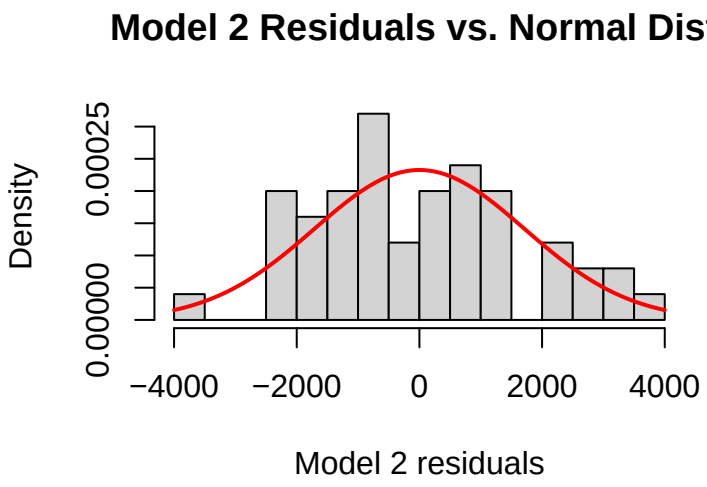
Unemployment insurance extensions and benefit increases can also be conceptually linked with reduced transit mobility. Near the outset of the pandemic in April 2020, unemployment rates peaked at 14.7% - up from a baseline of 3.5% in February of 2020 (Source). By increasing the length and amount of unemployment benefits, states kept a substantial number of these laid off employees, totaling 11.2% of the work-force, from needing to find a new means of employment in the short term. This, combined with the fact that these employees no longer needed to commute to work, likely contributed to reduced transit mobility.

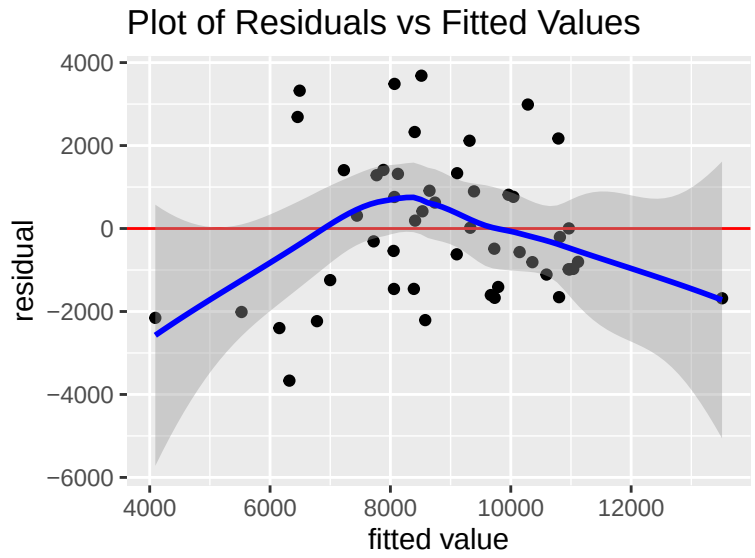
Our team was surprised, however, that information on the presence and length of state mask mandate policies did not incrementally benefit our baseline Model 2 performance as mask policies did not, in our estimation, share an obvious link to transit mobility or the percentage of the population under age 25. We did expect to see a material positive correlation between the duration of the mask mandate policy and population density, but even if we were to exclude population density from our model specification and substitute in the duration of the mask mandate feature, a t-Test indicates that the parameter estimate is not significant at the 0.05 level.

```
##
## Call:
## lm(formula = cases_per_100k_at_365d ~ median_transit_change +
##     pop_pct_age_0_24 + mask_mandate_days, data = final_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4506.6 -1037.6  -143.2  1396.1  5331.9
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -6943.7874  4547.3362  -1.527 0.133608
## median_transit_change    46.7185    24.0034   1.946 0.057738 .
## pop_pct_age_0_24        533.3941   135.9031   3.925 0.000288 ***
## mask_mandate_days        -0.5726     2.4368  -0.235 0.815262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1958 on 46 degrees of freedom
## Multiple R-squared:  0.401,  Adjusted R-squared:  0.362
## F-statistic: 10.27 on 3 and 46 DF,  p-value: 2.722e-05
```

Hence, by process of elimination, we have arrived at our final Model 2 specification. As with Model 1, we provide some summary plots and statistics to analyze the result of our OLS regression for Model 2.

```
##
## Call:
## lm(formula = cases_per_100k_at_365d ~ median_transit_change +
##     pop_pct_age_0_24 + population_density, data = final_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3666.7 -1367.9  -257.5  1191.5  3684.0
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -10160.586   3917.035  -2.594  0.01268 *
## median_transit_change     82.223     23.514   3.497  0.00106 **
## pop_pct_age_0_24         628.592    119.313   5.268 3.55e-06 ***
## population_density         4.665      1.446   3.225  0.00232 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1770 on 46 degrees of freedom
## Multiple R-squared:  0.5109, Adjusted R-squared:  0.479
## F-statistic: 16.02 on 3 and 46 DF,  p-value: 2.875e-07
```






```
##
## -------Homoskedasticity Test-------

##
##  studentized Breusch-Pagan test
##
## data:  model_2_final
## BP = 8.3386, df = 3, p-value = 0.03951

##
## -------Robust Standard Errors-------

##           (Intercept) median_transit_change      pop_pct_age_0_24
##           5070.544143             31.480733             154.756274
##    population_density
##              2.212617
```

```
##
## ------Normality of Residuals Test-------

##
##  Shapiro-Wilk normality test
##
## data:  model_2_final$residuals
## W = 0.97275, p-value = 0.2984
```

Model 2 Discussion

The first parameter estimate returned by Model 2 is an intercept estimate of -10,160, which represents the expected number of COVID-19 cases per 100K people (in the first year after declaration of a state of emergency) if there were no change in transit mobility levels, 0% of the population under 25 years of age, and 0 population density. Clearly the latter two scenarios are not possible given our context and sample observations. As a result of adding additional features to the model specification, the intercept term has lost its intuitive interpretation from Model 1 and therefore we will not comment further on the intercept estimates for this and subsequent models.

After introducing two additional co-variates to the model, the parameter estimate for change in median transmit mobility has changed such that a one percent increase is now associated with 82 additional COVID-19 cases per 100K people (up slightly from our estimate of 79 in Model 1), holding all other co-variates constant. The Breusch-Pagan test returned a p-value < 0.05, meaning that we reject the null hypothesis of constant residual variance. As discussed previously, we will consequently rely on robust standard errors in our discussions of statistical significance. The robust standard errors for this parameter estimate have decreased slightly to 31 in Model 2 (down from 33 in Model 1). This reduction in the standard errors occurred because the two additional co-variates we added uniquely explain a portion of the variance of our target variable without being highly colinear with the change in transit mobility feature. In Model 1, all of that unique variation with respect to the percentage of the population under age 25 and population density was being captured in the error term, giving us less certainty around our parameter estimate for change in median transit mobility. It is also worth noting that, thus far, the parameter estimate for change in median transit mobility has been robust to changes in the model specification as we moved from Model 1 to Model 2, which gives us increased confidence that the relationship we observe between it and the outcome is not spurious. The robust standard error estimates suggest that the change in median transit mobility parameter estimate is statistically significant using a standard Type I error rate of 0.05.

Model 2 also tells us that a one percent increase in the percentage of the population under age 25 is associated with 629 additional COVID-19 cases per 100K people, holding median transit mobility and population density constant. The robust standard error is 155, indicating that the percentage of the population under age 25 parameter estimate is statistically significant using a standard Type I error rate of 0.05. What this tells us is that the COVID-19 case count per 100K people is *far* more sensitive to changes in the percentage of the population under age 25 than it is to changes in median transit mobility. The practical significance of this outcome could be substantial, as it suggests that governing bodies should consider policies that specifically target the youth and/or potentially fund new studies that seek to better understand the causal pathways between young people and the spread of COVID-19.

Finally, Model 2 estimates an increase of 4.7 COVID-19 cases per 100K people for each additional person per square mile, after holding all other co-variates constant. The robust standard error is 2.2, indicating that the population density parameter estimate is statistically significant using a standard Type I error rate of 0.05. As previously discussed, this estimate and its significance is conditional on the inclusion of change in median transit mobility and the percentage of the population under age 25 in the model specification. And while a coefficient of 4.7 may seem small in comparison to the other features in this model specification, it is important to consider that population density has a much wider range of potential values than the other co-variate features in this model, with a range of 1.1 people per square mile in Alaska to 1021.3 in New

Jersey. The practical implications here are that some COVID-19 policies and behavior changes may be far more impactful in densely populated states than in sparsely populated ones, which suggests that a careful consideration of costs and benefits should be conducted and decisions and outcomes may justifiably vary by state.

Note: The Shapiro-Wilk normality test failed to reject the null hypothesis that the residuals for Model 3 were normally distributed ($p > 0.05$).

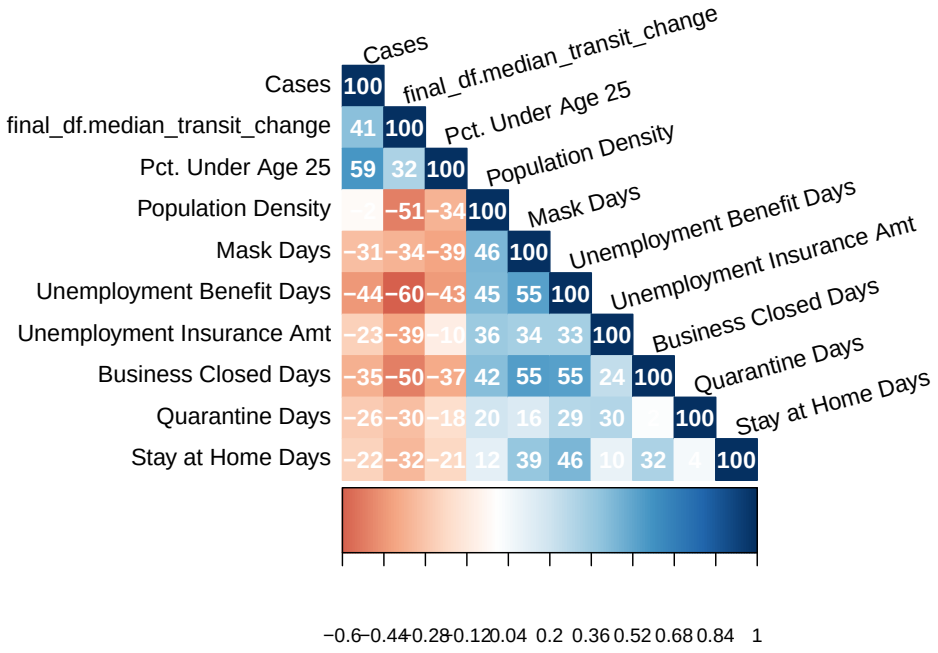
Model 3 - OLS Regression Development and Continued EDA

For Model 3, we added the six additional policy co-variates as described in the previous EDA section to further explore whether including information on government-enacted COVID-19 policies could measurably reduce residuals relative to our Model 2 specification. These six features were as follows: duration of mask mandates, duration and amount of increased government assistance via enhanced unemployment benefits, duration of the first wave of business closures, duration of stay at home mandates, and duration of travel quarantine restrictions.

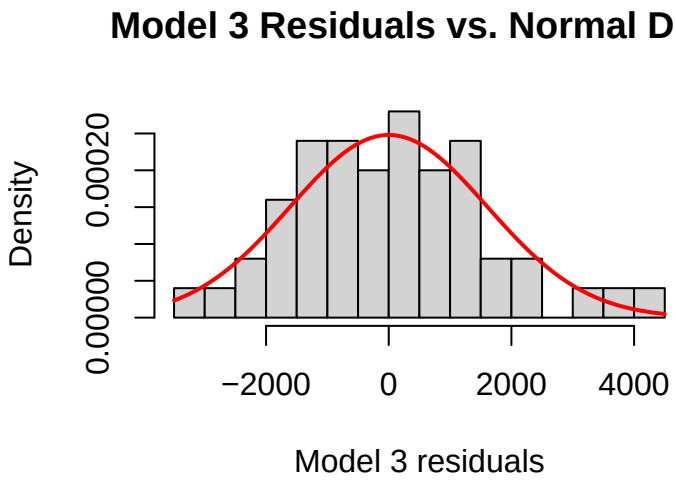
We previously considered the addition of each of these six variables incrementally (in isolation) to our Model 2 specification, but did not have statistical justification to include them. We also discussed the rationale for why these additional policy features may be correlated with the features we have already created and included in Model 2, and recognize that the colinearities with existing features in the model will limit the amount of unique variance of our target variable that each of the features can explain. Still, we have decided to include these six policy features in our specification for Model 3 because they do not demonstrate perfect colinearity with our existing features and all of the policies can be conceptually linked to the number COVID cases per 100K people. As with Model 1 and 2, we again provide summary statistics and plots for our Model 3 OLS regression result.

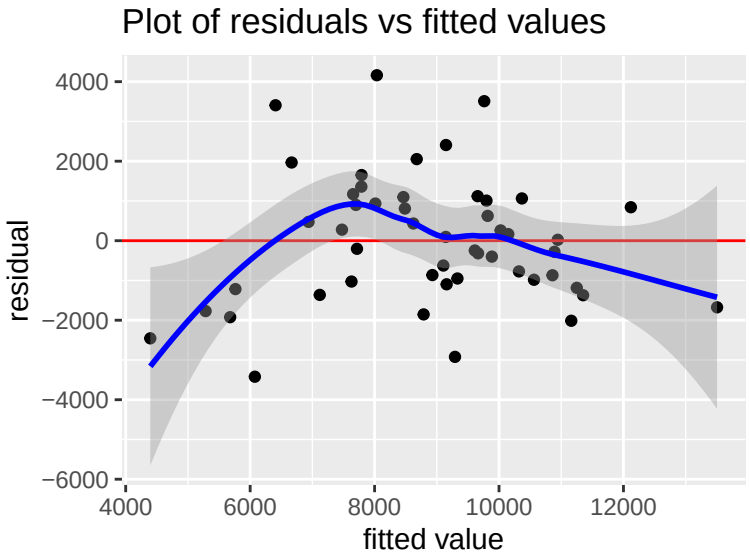
```
##
## Call:
## lm(formula = cases_per_100k_at_365d ~ median_transit_change +
##     pop_pct_age_0_24 + population_density + mask_mandate_days +
##     unemployment_benefits_days + increased_weekly_unemployment_insurance_amt_thru_jul31 +
##     business_closed_days_round1 + travel_quarantine_mandate_days +
##     stay_at_home_days, data = final_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3421.8 -1079.3   -90.2   990.0  4162.1
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       -4556.676   4720.588
## median_transit_change                                54.348     28.765
## pop_pct_age_0_24                                    569.302    129.214
## population_density                                    5.790      1.575
## mask_mandate_days                                    -1.188      2.766
## unemployment_benefits_days                           -3.873      4.991
## increased_weekly_unemployment_insurance_amt_thru_jul31   -2.339      2.031
## business_closed_days_round1                         -16.598     23.816
## travel_quarantine_mandate_days                       -1.779      2.379
## stay_at_home_days                                     1.841      4.816
##                                                   t value Pr(>|t|)
## (Intercept)                                        -0.965 0.340206
```

Correlation Matrix: All Features vs. Cases



```
## median_transit_change                                        1.889 0.066103 .
## pop_pct_age_0_24                                             4.406 7.68e-05 ***
## population_density                                           3.677 0.000694 ***
## mask_mandate_days                                           -0.429 0.669918
## unemployment_benefits_days                                  -0.776 0.442269
## increased_weekly_unemployment_insurance_amt_thru_jul31      -1.151 0.256369
## business_closed_days_round1                                 -0.697 0.489895
## travel_quarantine_mandate_days                              -0.748 0.458904
## stay_at_home_days                                            0.382 0.704251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1781 on 40 degrees of freedom
## Multiple R-squared:  0.5695, Adjusted R-squared:  0.4727
## F-statistic:  5.88 on 9 and 40 DF,  p-value: 3.369e-05


##
## -------ANOVA F-Test Significance Test Relative To Model 2-------


## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + mask_mandate_days + unemployment_benefits_days +
##     increased_weekly_unemployment_insurance_amt_thru_jul31 +
##     business_closed_days_round1 + travel_quarantine_mandate_days +
##     stay_at_home_days
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     40 126808331
## 2     46 144077125 -6 -17268794 0.9079 0.4991
```






```
##
## -------Homoskedasticity Test-------

##
##  studentized Breusch-Pagan test
##
## data:  model_3_final
## BP = 14.249, df = 9, p-value = 0.1137

##
## -------Robust Standard Errors-------

##                                             (Intercept)
##                                            5891.584751
##                                  median_transit_change
##                                              45.057309
##                                       pop_pct_age_0_24
##                                             160.444566
##                                     population_density
##                                               2.789427
##                                      mask_mandate_days
##                                               3.116445
##                                unemployment_benefits_days
##                                               5.251836
## increased_weekly_unemployment_insurance_amt_thru_jul31
##                                               2.109691
##                                business_closed_days_round1
##                                              23.840701
##                           travel_quarantine_mandate_days
##                                               3.121299
##                                        stay_at_home_days
##                                               5.387842

##
## ------Normality of Residuals Test-------
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_3_final$residuals
## W = 0.98344, p-value = 0.703
```

Model 3 Discussion

In Model 3, the coefficients of all six added policy co-variates returned as being statistically insignificant using robust standard errors. Model 3 also failed to reject the null hypothesis that Model 3's residuals were not measurably reduced from Model 2's in an ANOVA F-Test. Although the Breusch-Pagan test does not indicate that our Model 3 suffers from heteroskedasticity ($p > 0.05$), we will err on the side of caution and use robust standard errors for claims of statistical significance for Model 3. Additionally, the Shapiro-Wilk normality test also failed to reject the null hypothesis that the residuals for Model 3 were normally distributed ($p > 0.05$).

With Model 3, we observed some interesting changes to the parameter estimates for our nested features from Model 2. In particular, the coefficient for change in median transit decreased to 54 while the robust standard errors increased to 45, which meant that we could no longer reject the null hypothesis that this parameter estimate was measurably different than zero. Our group anticipated this effect, recognizing that many of the policy-related features were conceptually linked to change in transit mobility, as discussed previously. Indeed, this intuition was confirmed by looking at the correlation between change in transit mobility and these six policy variables, with the correlations ranging from -0.30 to -0.60. Although no single correlation was extremely high, we conclude that the amount of unique variation of our outcome variable explained by change in transit mobility is considerably reduced with the inclusion of the policy variables in the specification for Model 3. Conceptually, with fixed variance in the target variable, increasing the number of colinear parameters to be estimated means that there is less unique information available for each parameter estimate, which results in increased standard errors. Hence we conclude that our parameter estimate for change in median transit mobility was not robust to additions of multiple moderately colinear features in the Model 3 specification.

The parameter estimate for population density increased to 5.8 from 4.7 in Model 2. Just as we saw with the parameter estimate for change in transit mobility, the standard error grew from 2.2 to 2.8. However, the parameter estimate for population density remained significant in Model 3, meaning that population density continued to explain enough unique variance of our outcome of interest (COVID-19 cases per 100K people one year post declaration a state of emergency) to remain significant despite moderate correlations with some policy co-variates (which ranged from 0.12 to 0.46).

Finally, the parameter estimate for the percent of population below age 25 feature decreased to 569 in Model 3 from 629 in Model 2, while the standard error increased slightly to 160 from 155. In spite of this, the parameter estimate for the percentage of the population below age 25 remained significant using a standard Type I error rate of 0.05. Just as we saw with population density, the percent of population below age 25 feature continued to explain enough unique variance of our outcome variable to remain significant despite moderate correlations with some policy co-variates, which ranged from -0.10 to -0.43.

Because the change in median transit mobility was the only feature from Model 2 to become insignificant in the Model 3 specification, we decided to dig a bit deeper - beyond the bivariate correlations between this feature and all others in Model 3 - and calculated the variance inflation factor (VIF) for change in median transit mobility. The VIF is a singular measure of each feature's multicolinearity with all other co-variates in the model specification. The VIF for change in median transit mobility was 2.1, which was higher than any of the other features included in the Model 2 specification, but not high enough to justify dropping it from the model specification (according to our research, the rule of thumb is that this is only necessary for features with VIF values > 5). Ultimately, while we can partially attribute this feature's loss of statistical significance from Model 2 to Model 3 to multi-colinearity with the new policy features, the fact that the VIF

was only 2.1 should give us pause when evaluating the statistical significance of the relationship between change in median transit mobility and COVID-19 cases in Model 2.

Still, based on the totality of the Model 3 results, we conclude that adding the policy variables did not measurably reduce the model residuals in comparison with Model 2. Indeed, Model 3 recorded an adjusted R^2 of 0.473, less than the 0.479 recorded for Model 2. In light of this, and the fact that Model 3 also failed to reject the null hypothesis of an ANOVA F-Test against Model 2, we will move forward with Model 2 for the purposes of our CLM assumption assessment.

```
##                                    median_transit_change
##                                                 2.060399
##                                       pop_pct_age_0_24
##                                                 1.354473
##                                     population_density
##                                                 1.662260
##                                      mask_mandate_days
##                                                 1.967246
##                                unemployment_benefits_days
##                                                 2.315319
## increased_weekly_unemployment_insurance_amt_thru_jul31
##                                                 1.362193
##                               business_closed_days_round1
##                                                 1.930239
##                           travel_quarantine_mandate_days
##                                                 1.263924
##                                       stay_at_home_days
##                                                 1.396814
```

Regression Table

stargazer output

Table 1: OLS models for COVID-19 Spread

	Dependent variable:		
	cases_per_100k_at_365d		
	(1)	(2)	(3)
median_transit_change	79.250** (32.943)	82.223*** (31.481)	54.348 (45.057)
pop_pct_age_0_24		628.592*** (154.756)	569.302*** (160.445)
population_density		4.665** (2.213)	5.790** (2.789)
mask_mandate_days			-1.188 (3.116)
unemployment_benefits_days			-3.873 (5.252)
increased_weekly_unemployment_insurance_amt_thru_jul31			-2.339 (2.110)
business_closed_days_round1			-16.598 (23.841)
travel_quarantine_mandate_days			-1.779 (3.121)
stay_at_home_days			1.841 (5.388)
Constant	10,370.170*** (610.184)	-10,160.590** (5,070.544)	-4,556.676 (5,891.585)
Observations	50	50	50
R2	0.168	0.511	0.570
Adjusted R2	0.151	0.479	0.473
Residual Std. Error	2,259.262 (df = 48)	1,769.777 (df = 46)	1,780.508 (df = 40)
F Statistic	9.714*** (df = 1; 48)	16.018*** (df = 3; 46)	5.880*** (df = 9; 40)
Note:			*p<0.1; **p<0.05; ***p<0.01

Assessment of the CLM

Per instructions from the W203 staff, we have limited our CLM assumption discussion to what we consider to be our "best" model, which for us is Model 2. As previously mentioned, we aggregated the year-plus worth of county level observations across many datasets to a final data frame with 50 observations (i.e. one observation per state). We then performed an OLS regression to generate three separate models to try to capture the relationship between our features of interest and the change in cases. In order to generate unbiased estimators in our regression, we assess the applicability of the first three classical linear model (CLM) assumptions.

In order to fail to reject whether our model estimators would not be biased, we evaluated the applicability of the first three CLM assumptions:

Assumption 1: Independent and Identically Distributed Data (IID):

In assessing IID, let's first assume that the observations are not IID. When thinking through how COVID-19 spreads, as cases in one state surge, there will be COVID-positve people who travel to other states and spread the disease. This is more likely to impact states that are in close proximity to one another. In fact, this New York Times article describes the travel restrictions (as of April 11th) that states have put in place to try to mitigate the spread of the virus from one state to another. The impact of state's case counts influencing other states will likely result in some amount of clustering, suggesting that the observations are not independent. Thus, we fail to reject that our observations are not IID. Additionally, many of the state's policy decisions were influenced by what states were doing, such as NY and NJ coordinating on quarantine policies near the beginning of the pandemic.

In assessing identically distributed data, we note that mobility data is based on Google-Maps cell phone users. It is important to note that not everyone has access to a smart phone or uses Google Maps and allows their location to be traced, resulting in a different underlying distribution for change in mobility for each state from which our samples are drawn.

These IID infringements will likely result in biased model estimates as we build out our model. Despite this model limitation, we continue with our analysis as there are still useful insights and trends to be drawn despite biases in the model for even when the CLM assumptions are not met, the OLS estimator is consistent

for the parameter values that minimizes Kullback-Leibler (KL) divergence between the true distribution and the model distribution, but what we learn from the sample has no guarantees about the population.

Assumption 2: Linear Conditional Expectation:

In assessing whether our observations have a Linear Conditional Expectation (LCE), let's first assume that the LCE assumption has been met. We then examine the following plots of model predictions vs. model residuals. For Model 2, we find that the plot of model predictions vs model residuals results in a fitted line that is centered around zero. The smooth fitted line similarly residuals are significantly different than zero for both models. Thus we conclude that nothing in this data suggests there is not a nonlinear relationship and we fail to reject the hypothesis that LCE has been met.

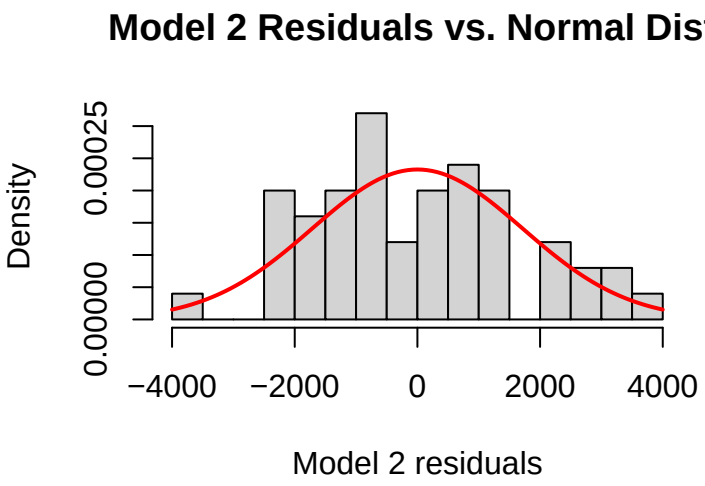
Assumption 3: No Perfect Colinearity:

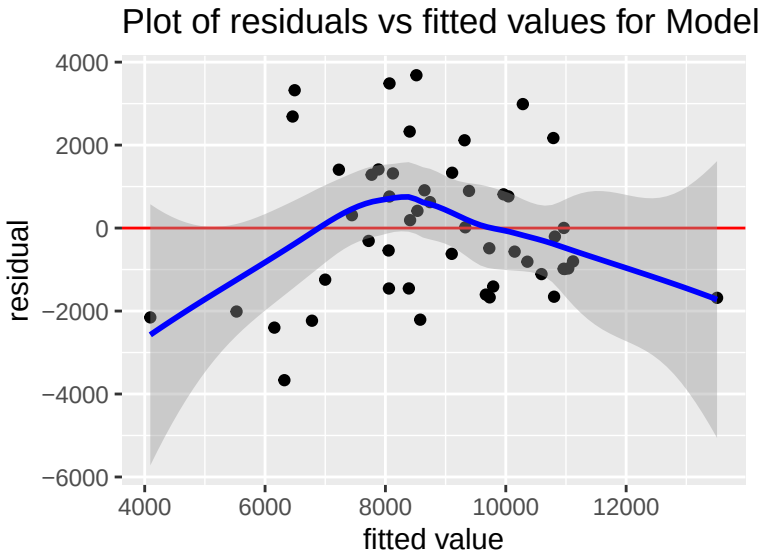
Perfect Colinearity means no independent variable can be expressed as a linear function of any other independent variables. It is clear there is no perfect colinearity since all feature paramaters had an estimation (i.e. the OLS regression function in R automatically drops `n-1` features in the instance where there are `n` perfectly colinear features).

The values of variance inflation factors show that the colinearity between the independent variables in Model 2 is not high (all values are < 2). Hence, we observed no evidence of perfect or near perfect colinearity in Model 2.

Assumption 4: Homoskedastic Errors:

We assessed the possibility of homoskedastic errors by visually assessing the plot between fitted values and residuals of the model, together with the Breusch-Pagan test.






```
##
##  studentized Breusch-Pagan test
##
## data:  model_2_final
## BP = 8.3386, df = 3, p-value = 0.03951
```

The plot shows the variance of error term of Model 2 is likely to be non constant since the residuals widens toward both ends of the plot. The Breusch-Pagan test's p-value is 0.040, meaning that we reject the null hypothesis that the residuals have constant variance. Although heteroskedasticity is likely to exist, we have remedied the problem by making all of our statistical significance claims for model parameter estimates using robust standard errors instead of classical standard errors.

Assumption 5: Normally Distributed Errors:

To assess whether our model errors were normally distributed, we conducted a Shapiro-Wilk normality test. The p-value of 0.30 suggests that we cannot reject the null hypothesis that the residuals are normally distributed.

```
##
##  Shapiro-Wilk normality test
##
## data:  model_2_final$residuals
## W = 0.97275, p-value = 0.2984
```

Model Limitations & Omitted Variables Discussion

In addition to the features we have selected to conduct the regression analysis, there could be other features that are correlated to the target feature (COVID cases per 100K population) or the selected features. These omitted features could cause directional biases to the regression coefficients. This could change the value and significance of the regression coefficients and the interpretation of the models. Two common reasons that the omitted features are not included in the analysis:

- The features are difficult to measure.

- The features are not available during the analysis.

The effect of omitted features can be further organized into the following four scenarios:

- If the omitted feature is positively correlated with the selected and target feature, it generates a positive bias to the regression coefficient. A positive bias means the regression coefficient of the selected feature is overestimated.
- If the omitted feature is negatively correlated with the selected and target feature, it also generates a positive bias to the regression coefficient.
- If the omitted feature is negatively correlated with the selected but positively correlated with the target feature, it generates a negative bias to the regression coefficient. A negative bias means the regression coefficient of the selected feature is underestimated.
- If the omitted feature is positively correlated with the selected but negatively correlated with the target feature, it also generates a negative bias to the regression coefficient.

During the regression analysis, we have tried our best to include all the features available to us that we thought were related to our outcome of interest. However, in the real-world, it is almost impossible to include all the related features. In the following table, we have listed several features that were not available to us but could have potential impact to the target feature.

Omitted Feature	Correlation to Target Feature	Correlation to Selected Feature	Bias Direction
Temperature	Positive	Positive (Mobility)	Positive
Mask Compliance	Negative	Positive (Population Density)	Negative
COVID Testing Availability	Positive	Positive (Population Density)	Positive
Percentage People with Mass Transit	Positive	Positive (Population Density)	Positive
Percentage People WFH	Negative	Negative (Mobility)	Positive

Temperature: According to the existing scientific evidence reported by the MedicalNewsToday, COVID-19 was more dormant under higher temperature, which reduced its ability to infect human cells. In Addition, based on the mobility data, mobility seems to have a moderate negative correlation to temperature in 2020. Therefore, we think the temperature has a negative correlation to the COVID cases per 100k population and also a negative correlation to mobility and could potentially create a positive bias to the mobility regression coefficient.

Mask compliance: Masks are proven to reduce the range to aerosol particles from a sneeze or cough and reduce the spread of COVID. In the regression analysis, the mask mandate dates are available to us but the mask compliance data is not. That is, even though the government requires people to wear masks from a particular date, we don't know how many people actually follow the mandate and wear masks in the public. We also suspect that the mask compliance is positively correlated to population density. If people live in a densely-populated area, such as New York and other big cities, and have less space between individuals,

they may prefer to wear masks to reduce the risk to infection. Hence, the mask compliance could impose a negative bias on the population density regression coefficient.

COVID testing availability: There are two ways of testing COVID. One is a viral test that tells you if you have a current infection. The second test is an antibody test (also known as a serology test) that tells you if you had a past infection. However, both tests have to be executed by health professionals. Therefore, the availability of the tests depends on the medical resources in the area. If an area has less medical resources, such as the country areas in the less populated states, it can be more difficult to obtain such tests. On the contrary, big cities with more medical resources have more test kits for the residents. We suspect there is a positive correlation between the COVID testing availability with the population density. Also with more testing available to the public, the more potential COVID cases can be identified. Hence, there is also a positive correlation between the test availability and COVID cases per 100K population. The COVID testing availability could generate a positive bias to the regression coefficient for population density.

Percentage of the state population that commutes via mass transit: Many people take public transportation (buses, subways, etc.) to work daily. Typically the spacing between passengers on public transportation is limited to maximize capacity. Such a setting could facilitate the spread of COVID due to close contact of passengers. Therefore, we think the percentage of the state population who commute via public transportation is another important factor to understand the spread of COVID. However, this data is not available to us. In densely populated areas, a greater portion of people prefer taking public transport as their primary commute choice due to lack of parking spaces, higher cost, bad traffic, etc. In the less populated areas, people prefer to drive to work due to less traffic and little to no cost for parking. In some areas, mass transportation is not even available to the public. Hence, we believe that the percentage of the state population that commutes via mass transit has a positive correlation with the population density and COVID cases per 100K people. Therefore, the percentage of the state population that commutes via mass transit could generate a positive bias to the regression coefficient for population density.

Percentage of the state population that could work from home: During the pandemic, more and more companies have modified their working policies to allow employees to work from home if feasible. Working from home prevents close contact at workplaces and could potentially reduce the risk of getting COVID. Therefore, we think the percentage of the state population that is able to work from home is an important factor to model the spread of COVID. With more people working from home, transit mobility will decrease. We therefore suspect that the percentage of the state population that is able to work from home has a negative correlation with the COVID cases per 100K population, and also a negative correlation with the change in transit mobility. The double negative correlations potentially give a positive bias to the regression coefficient for the transit mobility feature.

Other Model Limitations: Potential Impact from the Family-Wise Error Rate (FWER)

In this regression analysis Lab, we have conducted a series of tests for different purposes, including coefficient t-tests, BP tests, and F-tests. When multiple tests are conducted simultaneously, the type I error can inflate with an increasing total number of tests. The family-wise error rate (FWER) is the overall error rate under a series of hypothesis tests. In other words, the FWER means the probability of making at least one type I error in a series of hypothesis tests. There are a couple of methods to control inflation of family-wise error rate, including (1) a single step method, and (2) a sequential method.

One single step method is the Bonferroni correction. The method divides the alpha level (p-value) by the total number of tests you run and apply the new alpha level to each individual test. Extended from the single step method, sequential FWER correction method adaptively adjusts the p-value at each individual stage. One common sequential method is the Holm-Bonferroni correction method. The method first orders the p-values for the hypothesis tests from the smallest to greatest. The method then calculates a corrected

p- value using the Holm-Bonferroni formula (HB = Target / (n – rank + 1)). After the correct p-value is obtained, the method compares it with the first-rank (the smallest) p-value. If the corrected p-value is smaller, the method rejects the null hypothesis for this individual test. Continue this process for the next p-value until the first non-rejected hypothesis test is met. From here, all subsequent hypothesis tests are non-significant.

These tests were not in the scope of the course material and not conducted in this regression analysis lab. However, we kept the idea in mind that with increasing numbers of hypothesis tests comes a higher likelihood of committing a Type I error. In addition to the steps described above, another way to assess the robustness of your parameter estimates is to observe how they change in response to multiple model specifications. Our regression table suggests that our regression coefficient estimations are relatively robust, aside from the change in median transit mobility feature which was more sensitive to colinear variables than we had hoped. In the future though, we should consider applying these additional methods to control type I error rate inflation and properly track the total number of tests conducted in the project.

Conclusion.

In this regression analysis lab, we collected different types of data/observations, including mobility, demographics, and policies, that we thought were related to COVID cases in the U.S. from various databases. Once we extracted the raw data/observations from the databases, we performed a data cleaning process to identify and remove errors and missing values, and a data wrangling process to calculate/transform the raw data/observations into the features that we were interested in. Once the features were obtained, we further conducted a rigorous Exploratory Data Analysis (EDA) to investigate the relationships/correlations among different features and the target. After the EDA, we decided to select the target as the COVID cases per 100k people in each state after 364 days after the state of emergency was declared by the state government, and the descriptive features including transit mobility change, population density, percent of population under the age of 25, mask mandates total days, stay-at-home total dyas, travel quarantine mandate total days, closure of business total days, increased unemployment benefit total days, and the increased amount of unemployment benefit. We continued to sequentially build 3 regression models with different specifications for this regression analysis lab.

Model Summaries

Across all three models, our key variable coefficient and robust standard errors perform the best in model 2 and then in model 1, while in model 3 none of the additional policy covariates were statistically insignificant and the robust standard error for our main predictor, median transit, actually increased to a point where this parameter was no longer statistically different than zero. This was expected, however, given we hypothesized that that the median transit parameter captured many of the policy-related features inherently (e.g. quarantine policies we hypothesized would be highly associated with lower mobility). The slight reduction in the standard errors between from model 1 (cases ~ mobility) to model 2 (case ~ mobility, age below 24, and population density) occurred because the additional covariates explain some of the variance of our target variable without being highly collinear with the change in transit mobility feature, thus capturing some of the unique variation with respect to the Model 1 error term. With this in mind, we drew the following associative conclusions from our OLS regressions:

Model 1:

the number of COVID cases per 100,000 people in the first year after the declaration of a state of emergency

$$\approx 10370 + 79 * \text{percent change in transit mobility levels}$$

In our first (base) model, we selected the mobility change as the primary feature for our regression analysis. The analysis results showed that the number of COVID cases per 100,000 people in the first year after

the declaration of a state of emergency 10370 + 79 * percent change in transit mobility. The regression coefficient of mobility change suggested that a 1% change in mobility was associated with approximately 80 case counts increase in the first year. The intercept term suggested that an average of 10,370 case increase per 100k people within the year after the state of emergency was declared was expected if there was no mobility change. Our base model had the poorest model fit (adjusted R-squared) value (0.151) and the largest residual standard error (2,260) among the three models. The base model with mobility change explained approximately less than 20% of the total variance of the target.

Model 2:

$$\begin{aligned}\text{change in cases per 100,000 people...} \approx -10,160 + \\ 82 * \text{percent change in transit mobility levels} + \\ 628 * \text{percent of the population under 25 years of age} + \\ 5 * \text{population density}\end{aligned}$$

In our second model, we added two more features, population density and percentage of population under age 25, in an attempt to improve the performance/fit of the regression model. We hypothesized that younger people's higher chance for reduced or no COVID symptoms would result in the spreading the disease at greater rates than older age groups. Hence, for our second model, we were interested in testing our general hypothesis that age demographics, as well as population density, may play a role in the spread of COVID-19.

The regression analysis showed that the cases per 100,000 people -10,160 + 82 * percent change in transit mobility levels + 628 * percent of the population under 25 years of age + 5 * population density. The regression coefficients suggested that a 1% change in mobility was associated with approximately 80 case counts increase in the first year, a 1% change in the population under age 25 was associated with 628 cases increase per 100,000 people, and a 1% change in the population density was associated with 5 cases increase per 100,000 people when the other two features were kept constant. It would seem that a change in the percentage of the population ages 0-24 is associated with the the greatest change in COVID cases.

Unlike our first model result, our second model results in a negative intercep, the intercept term was negative in this case, which meant the added features successfully explained the variance included in the intercept from the first model. Since there was no state with 0 population density and percent population under 25, it was reasonable that we got a negative intercept in model 2. With that in mind, this intercept isn't useful as we wouldn't expect such a negative change in cases per capita to be associated with no change in the independent features. Thus, from this point on, we focused our analysis towards the parameter coefficients. The model fit of the second model improved significantly (adjusted R-squared increased from 0.15 to 0.48) and the residual standard error also reduced pronouncedly (from 2,260 to 1,769). By adding the two non-highly correlated features appeared to significantly improve the efficiency of our descriptive model. The second model was able to explain almost half of the variance in the target, which is a significant improvement over Model 1.

Model 3:

$$\begin{aligned}\text{change in cases per 100,000 people...} \approx -4556 + \\ 54 * \text{percent change in transit mobility levels} + \\ 569 * \text{percent of the population under 25 years of age} + \\ 6 * \text{population density} + f(\text{policy features})\end{aligned}$$

In the third model, we added a series of additional policy related features with multicollinearity on top of our second model, which included duration of mask mandates, duration and amount of increased government assistance via enhanced unemployment benefits, duration of the first wave of business closures, duration of stay at home mandates, and duration of travel quarantine restrictions. The addition of six policy features, ,

did not particularly improve the efficiency of our model. In fact, the adjusted R-squared actually decreased from 0.48 to 0.47.

This excercise of adding more and more features was mostly an exercise to help us understand the effect of adding excessive features to a regression model. In this model, the regression coefficients of the primary features used in model two remained at a similar level (mobility: 82 to 54, population density: 628 to 570, and percent population under age 25: 4.6 to 5.7), which suggested that the robustness of the descriptive power of the features. The policy related feature all came out insignificant with very low regression coefficients (mostly between -2 to 2). The adjusted R square value also reduced from 0.48 to 0.47 and the residual standard error remained almost the same (1,780) compared to our second model. Model 3 also failed to reject the null hypothesis that Model 3's residuals were not measurably reduced from Model 2's in an ANOVA F-Test. In addition, the mobility feature became statistically insignificant in the third model potentially due to collinearity donated by the excessive additional features. The results suggested that adding excessive features with multicollinearity didn't improve the efficiency of the model by the slightest but could potentially mask the actual descriptive features.

Given the jump in performance in going from Model 1 to Model 2, and the subsequent decrease in model efficiency and increase in interpretobility in going from Model 2 to Model 3 (along with the fact that we failed to reject the null hypothesis that Model 3's residuals were not measurably reduced from Model 2's in an ANOVA F-Test.), we selected Model 2 as our preferred model for exploring the associations between mobility, demographic, and policy data with a dependent change in COVID cases counts as it balanced improved statistical power with explanatory efficiency and interpretability.

Study Limitations

Because we had only 50 sets of observations in this study, the assumptions of the classical linear model were extremely important. Therefore, we checked if our models satisfied the five assumptions of the classical linear model. In this study, we failed to reject that our observations are not IID due to data collection processes or clustering effect of neighboring states. We further investigated if the linear conditional expectation was met in our models. The residual vs. model fit plot confirmed that we had a linear conditional expectation so the assumption was met. Among all the features we selected, we conducted VIF tests to check if near-perfect collinearity existed. The values of VIF were all below 2, so we concluded that there was no evidence of existence of perfect or near perfect collinearity. We continued to assess the possibility of homoskedastic errors by plotting the fitted values and residuals of the models, together with the Breusch-Pagan test(BP test) and obtained p-values of 0.05. Therefore, we rejected the null hypothesis that the errors were homoskedastic. Although heteroskedasticity was likely to exist, we proceeded with the model estimation using robust standard errors. Lastly, we assessed the normality of the residual distribution using Shapiro-Wilk normality test. P-values under 0.05 failed to reject the null hypothesis that the model errors were not normally distributed. When we analyze the findings of our models, we need to keep in mind that some of the unsatisfied assumptions could potentially affect the integrity of our results and proceed our interpretation with care.

Discussion

In conclusion, based on our regression analyses, we think that mobility, population density, and percentage of younger people are important descriptive features for understanding COVID cases spread in different states in the U.S.. On the contrary, we didn't find the policy-related features are statistically significant to describe COVID cases increase. However, it doesn't mean that the policies designed to reduce virus spread are not important. We believe that the primary features we used in the second model actually reflect certain policies. For example, the mobility change could be related with the travel quarantine mandate and stay-at-home order, and the population density could be related with the mask mandate. In addition, there could be features that are not included in this analysis as mentioned in the omitted feature sections that can impact the interpretation of the model. Therefore, it is not justifiable to denounce the effect of COVID spread mitigation policies due to statistical insignificance. We hope this study can help people understand

more about the relationships among COVID cases and related features, provide insights to support COVID mitigation strategies, and serve as a reference for future public health researches.