Regression analysis to study the Spread of Covid-19

Jyoti Kumari

4/12/2021

1. An Introduction

SARS-CoV-2 was first identified in December 2019 in China. This new coronavirus has caused a worldwide pandemic of respiratory illnesses known as COVID-19. It accounts for more than 31 million cases in the United States, with more than 560,576 deaths as of April 14th, 2021(1). SARS-CoV-2 can be spread very easily from person to person. The three most common ways that this can happen are direct contact and spread via droplets or aerosols. Physical distancing is considered as one of the most effective means to combat the spread of the virus, however, this measure is hard to meet in areas with high poverty where several families must live under one roof (compound houses) and/or in high density areas.

In this study, I examine the following research question: *Is there a causal relationship between different socio-economic conditions such as poverty, unemployment rate and the number of cases related to Covid 19?* This research question is motivated by news, reports, and updates over the past 12 months about different socio-economic sections of the society affected differently and sometimes more adversely by COVID -19.

The objective here is to better understand this relationship so health and socioeconomical policies can be implemented at the state level and help reduce the spread of COVID-19. The expectation here is that the individuals living under the poverty line and/or people living in high-density areas are going to have more COVID -19 cases. Three different models will be developed to explore our research question, additional variables such as number of homeless and face mask mandate and population density will also be included to understand their effect on the number of COVID-19 cases.

Two different datasets were used to analyze our research question: Covid-19 US State Policy Database and New York Times Covid Case Tracker dataset which will be described below. All data was compiled into state-by-state metrics for regression analysis.

Out of 50-sample dataset, following variables have been operationalized:

1) *covid_cases:* The number of covid cases has been operationalized as our outcome variable. Last retrieved on April 15th, 2021 from the New York Times Covid Case Tracker dataset.

2) *poverty_rate:* Poverty rate [2018] has been operationalized as the key variable. In line with the above research question, I believe that this variable will help me understand how number of people living below federal poverty line affects covid cases. I think that among the population, this group is likely to be contributing to the spread of COVID cases, since they will have less facilities to control COVID cases by following the state policies. Extracted from the Covid-19 US State Policy Database.

3) *unemployment_rate:* Unemployed rate [2018] has also been operationalized as a covariate. This will help us understand the association between unemployment and poverty as well as the relationship between unemployment and covid cases. Extracted from the Covid-19 US State Policy Database.

4) *pop_density:* Population density per square miles. I expect that states with denser populations will have higher spread of COVID-19 cases. Extracted from the Covid-19 US State Policy Database.

5) *face_mask:* A binarized variable based on whether the state has legal enforcement of mask mandate. Extracted from the Covid-19 US State Policy Database. The thinking here is, that the enforcement of

masks would yield lower cases and spread.

6) *homeless*: Number Homeless [2019] has also been operationalized as a covariate. Extracted from the Covid-19 US State Policy Database. I believe that areas with higher number of homeless will yield higher number of cases and spread.

It has been assumed that the data such as population, population density, number of homeless people, and the number of people living under federal poverty line does not change much over the years. These variables contain data from the year 2018 and 2019.

In addition to this, District of columbia has been filtered out from the Covid-19 US State Policy dataset. Since the data has been compiled into state-by-state metrics for regression analysis, and "District of columbia" is not a US state, it has been decided to filter out "District of columbia" from the entire analysis.

2. Data loading & Data Wrangling

The following data sources have been used:

- COVID-19 US State Policy Database A database of state policy responses to the pandemic, compiled by researchers at the Boston University School of Public Health. From this website, excel sheet "COVID-19 US state policy database 4_13_2021" has been downloaded and then has been uploaded to R for further usage.(https://github.com/USCOVIDpolicy/COVID-19-US-State-Policy-Database/blob/master/COVID-19%20US%20state%20policy%20database%204_13_2021.xlsx)
- New York Times Covid case tracker dataset This historical database includes the daily number of cases and deaths nationwide, indlucing states, U.S. territories and the District of Columbia. The data begins with the first reported case in Washington State on January 21st, 2020. This data is updated on daily bases, for this study the data was retrieved on April 13th, 2021.

2.1. Pull data from US state policy changes database

Creating list of features to be operationalized from US policy changes dataset:

```
used_features <- c("State",
                   "No legal enforcement of face mask mandate",
                   "Extend the amount of time an individual can be on unemployment insurance",
                   "Weekly unemployment insurance maximum amount (dollars)",
                   "Weekly unemployment insurance maximum amount with extra stimulus (through July 31, 
                   "Population density per square miles",
                   "Population 2018", "Number Homeless (2019)",
                   "Percent Unemployed (2018)",
                   "Percent living under the federal poverty line (2018)")
```

Selecting key variables of interest from the Covid-19 state policy database and applying necessary filtering: (here, District of Columbia has been filtered out to only stick to states):

```
covid_data_usp <- read_excel("COVID-19 US state policy database 4_13_2021.xlsx",
                             sheet = 1,
                             range = "State policy changes !A2:HO56") %>%
  select(used_features) %>%
  filter(State != ('category'),
         State != ('type'), State != ('unit')) %>%
  rename(
        face_mask_char = 'No legal enforcement of face mask mandate',
        UI_extended_char = 'Extend the amount of time an individual can be on unemployment insurance',
        UI_max_char = 'Weekly unemployment insurance maximum amount (dollars)',
        stimulus_char = 'Weekly unemployment insurance maximum amount with extra stimulus (through July
```

```
        pop_density_char = 'Population density per square miles',
        population_char = 'Population 2018',
        homeless_char = 'Number Homeless (2019)',
        unemployment_rate_char = 'Percent Unemployed (2018)',
        poverty_rate_char = 'Percent living under the federal poverty line (2018)') %>%
  filter(State != 'District of Columbia')  %>%
  mutate(unemployment_rate = as.numeric(unemployment_rate_char),
         population = as.numeric(population_char),
         poverty_rate = as.numeric(poverty_rate_char),
         stimulus = as.numeric(stimulus_char),
         homeless = as.numeric(homeless_char),
         pop_density = as.numeric(pop_density_char),
         face_mask = as.numeric(face_mask_char),
         UI_extended = as.numeric(UI_extended_char),
         UI_max = as.numeric(UI_max_char)) %>%
select(-c('unemployment_rate_char',
          'population_char',
          'poverty_rate_char',
          'stimulus_char',
          'homeless_char',
          'pop_density_char',
          'face_mask_char',
          'UI_extended_char',
          'UI_max_char'))
```

2.2. Pull data from NYT COVID Database

- Data here is available from Jan-21-2020 onwards till Apr-13-2021.

```
NYT_Data <- fread("https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv")
```

```
#Create subset of NYT Data
subset_NYT =  data.frame(NYT_Data$state, NYT_Data$cases)
names(subset_NYT) = c("State", "Cases")

# creating a subset of NYT dataset by doing group by state and aggregating total number of cases.
subset_NYT <- subset_NYT %>% group_by(State) %>% summarise(covid_cases = sum(Cases))

# filtering out islands and territories from US state list
NYT_Data <- subset_NYT %>%
filter(
    State != 'District of Columbia',
    State != 'Guam',
    State != 'Northern Mariana Islands',
    State != 'Puerto Rico',
    State != 'Virgin Islands'
  )
```

2.3. Merging datasets

Merging the two datasets above, here's what the preview looks like:

```
# joining NYT dataframe(NYT_Data) and the main dataframe - US policy(covid_data_usp)
covid_data <- covid_data_usp %>%
  left_join(NYT_Data, by = 'State')

head(covid_data)
```

```
## # A tibble: 6 x 11
##   State   unemployment_ra~ population poverty_rate stimulus homeless pop_density
##   <chr>              <dbl>      <dbl>        <dbl>    <dbl>    <dbl>       <dbl>
## 1 Alabama              5.6    4887871         16.8      875     3261        93.2
## 2 Alaska               6.8     737438         10.9      970     1907         1.11
## 3 Arizona              5.4    7171646         14        840    10007        62.9
## 4 Arkans~              4.5    3013825         17.2     1051     2717        56.7
## 5 Califo~              5.5   39557045         12.8     1050   151278       242.
## 6 Colora~              3.9    5695564          9.6     1218     9619        54.7
## # ... with 4 more variables: face_mask <dbl>, UI_extended <dbl>, UI_max <dbl>,
## #   covid_cases <int>
```

```
#Discard na values
covid_data = na.omit(covid_data)
```

Now, we have got main dataframe *covid_data* to be used for our research.

Table 1 summarizes the original and final number of rows and columns in our datasets, and reasoning behind the data reduction. As mentioned before, both dataframes were merged to create a final dataframe *covid_data* to be used for our research project. This dataframe has a total of 50 rows that represent the number of states and eleven variables relevant to our project.

Dataset	Original number of rows	Original number of columns	Final number of rows	Final number of columns	Goal
Covid-19 US State policy (State policy changes)	51	232	50	10	Remove District of Columbia, and select relevant columns
New York Times Covid case tracker dataset	22509	5	50	2	Added total number of cases by State, and islands and territories were removed.
covid_usp_cases	--	--	50	11	Merged State policy dataset and NY Times dataset, and removed null values
Final dataframe: covid_data	--	--	50	11	Final database that we are going to use to our model

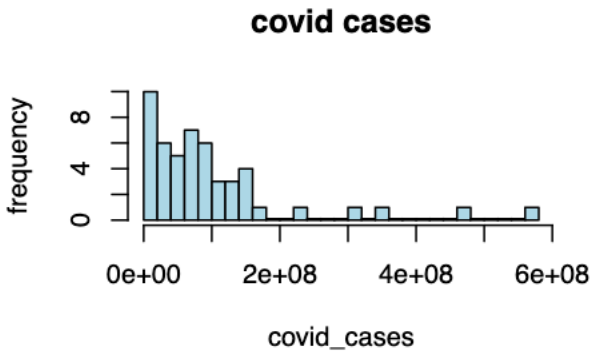
Table 1: Data reduction summary

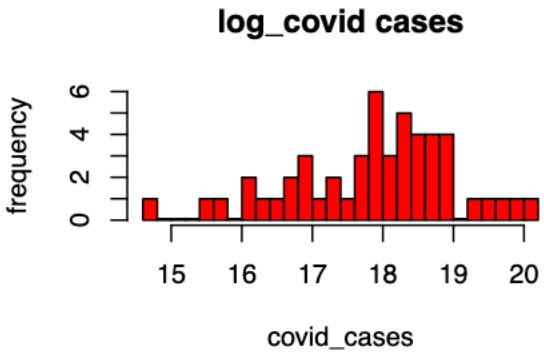
The *summary()* function was used to obtain a summary statistics of our variables and find out if some of them require any data transformation. In addition histograms were generated to help us identified the variables that would benefit from those transformations.

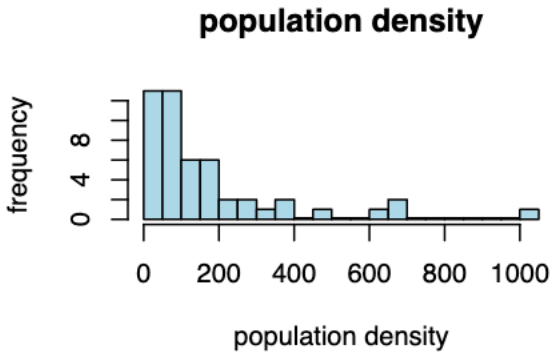
```
summary(covid_data)
```

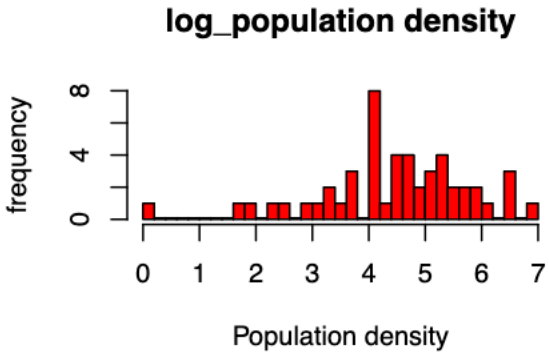
```
##     State           unemployment_rate   population        poverty_rate
##  Length:50          Min.   :2.800     Min.   :  577737   Min.   : 7.60
##  Class :character   1st Qu.:3.825     1st Qu.: 1836691   1st Qu.:10.93
##  Mode  :character   Median :4.900     Median : 4564190   Median :12.70
##                     Mean   :4.692     Mean   : 6529300   Mean   :12.85
```

```
##                    3rd Qu.:5.475     3rd Qu.: 7444605   3rd Qu.:14.07
##                    Max.   :7.100     Max.   :39557045   Max.   :19.70
##     stimulus       homeless       pop_density        face_mask
##  Min.   : 835   Min.   :   548   Min.   :   1.11   Min.   :0.00
##  1st Qu.: 972   1st Qu.:  2315   1st Qu.:  45.63   1st Qu.:0.00
##  Median :1060   Median :  4355   Median :  91.11   Median :0.00
##  Mean   :1080   Mean   : 11113   Mean   : 170.56   Mean   :0.34
##  3rd Qu.:1167   3rd Qu.:  9543   3rd Qu.: 197.58   3rd Qu.:1.00
##  Max.   :1423   Max.   :151278   Max.   :1021.27   Max.   :1.00
##   UI_extended       UI_max       covid_cases
##  Min.   :0.00   Min.   :235.0   Min.   :  2380165
##  1st Qu.:0.00   1st Qu.:372.0   1st Qu.: 27483300
##  Median :0.00   Median :460.0   Median : 66858840
##  Mean   :0.06   Mean   :480.1   Mean   : 99850384
##  3rd Qu.:0.00   3rd Qu.:567.0   3rd Qu.:120155665
##  Max.   :1.00   Max.   :823.0   Max.   :578892907
```









As mention before, histograms were generated to analyze the distribution of the data in our variables with and without data transformations. The histograms on the left (light blue color) don't have any transformations and the histograms on the right (red color) were log transformed.

After analyzing the statistics summary and the histograms, it was determined that the variables *covid cases* and *population density* were left skewed and after applying the log transformation, both variables displayed a relative normal distribution. For this reason, log transformations will be used in our outcome variable (covid_cases) and input variable population density for model building. No transformations will be done in the remaining variables.

3. Exploratory Data Analysis

With all the necessary variables in place, we should look at how the outcome variable (number of covid cases) is distributed along with our key predictor variable poverty_rate.

The *summary()* function was again used to closely analyze the covid cases variable which is our outcome variable and the poverty rate variable that was identified as the key predictor variable to answer the research question.

```
summary(covid_data$covid_cases)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   2380165  27483300  66858840  99850384 120155665 578892907
```

```
summary(covid_data)
```

```
##     State           unemployment_rate   population        poverty_rate
##  Length:50          Min.   :2.800     Min.   :  577737   Min.   : 7.60
```

```
##  Class :character   1st Qu.:3.825     1st Qu.: 1836691   1st Qu.:10.93
##  Mode  :character   Median :4.900     Median : 4564190   Median :12.70
##                     Mean   :4.692     Mean   : 6529300   Mean   :12.85
##                     3rd Qu.:5.475     3rd Qu.: 7444605   3rd Qu.:14.07
##                     Max.   :7.100     Max.   :39557045   Max.   :19.70
##     stimulus       homeless       pop_density        face_mask
##  Min.   : 835   Min.   :   548   Min.   :   1.11   Min.   :0.00
##  1st Qu.: 972   1st Qu.:  2315   1st Qu.:  45.63   1st Qu.:0.00
##  Median :1060   Median :  4355   Median :  91.11   Median :0.00
##  Mean   :1080   Mean   : 11113   Mean   : 170.56   Mean   :0.34
##  3rd Qu.:1167   3rd Qu.:  9543   3rd Qu.: 197.58   3rd Qu.:1.00
##  Max.   :1423   Max.   :151278   Max.   :1021.27   Max.   :1.00
##   UI_extended       UI_max       covid_cases
##  Min.   :0.00   Min.   :235.0   Min.   :  2380165
##  1st Qu.:0.00   1st Qu.:372.0   1st Qu.: 27483300
##  Median :0.00   Median :460.0   Median : 66858840
##  Mean   :0.06   Mean   :480.1   Mean   : 99850384
##  3rd Qu.:0.00   3rd Qu.:567.0   3rd Qu.:120155665
##  Max.   :1.00   Max.   :823.0   Max.   :578892907
```

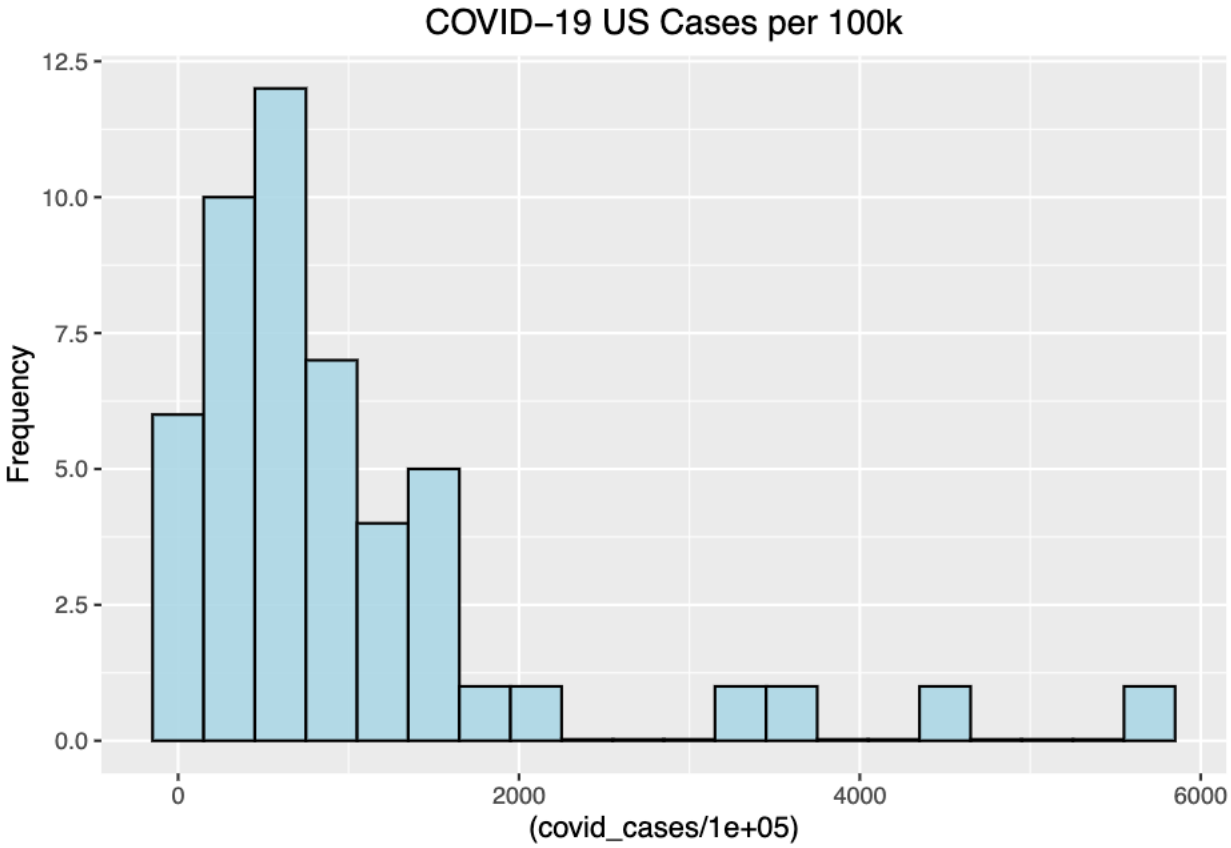
```
summary(covid_data$poverty_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.60   10.93   12.70   12.85   14.07   19.70
```

The summary statistics of the outcome variable - covid_cases tells us that our outcome variable (covid cases) has a mean around 90,342,206, median around 60,197,368 and 52,3024,989 maximum cases, given a state.

Also, the summary stats of the key predictor variable "poverty_rate" with the mean and median around 12% and max of 20% per state.

The following graph shows the distribution of Covid cases across US states between Jan-2020 to Apr 2021. This also shows that the number of Covid case across US states has been up to 1200,000 (1.2 million) for around 35 states.



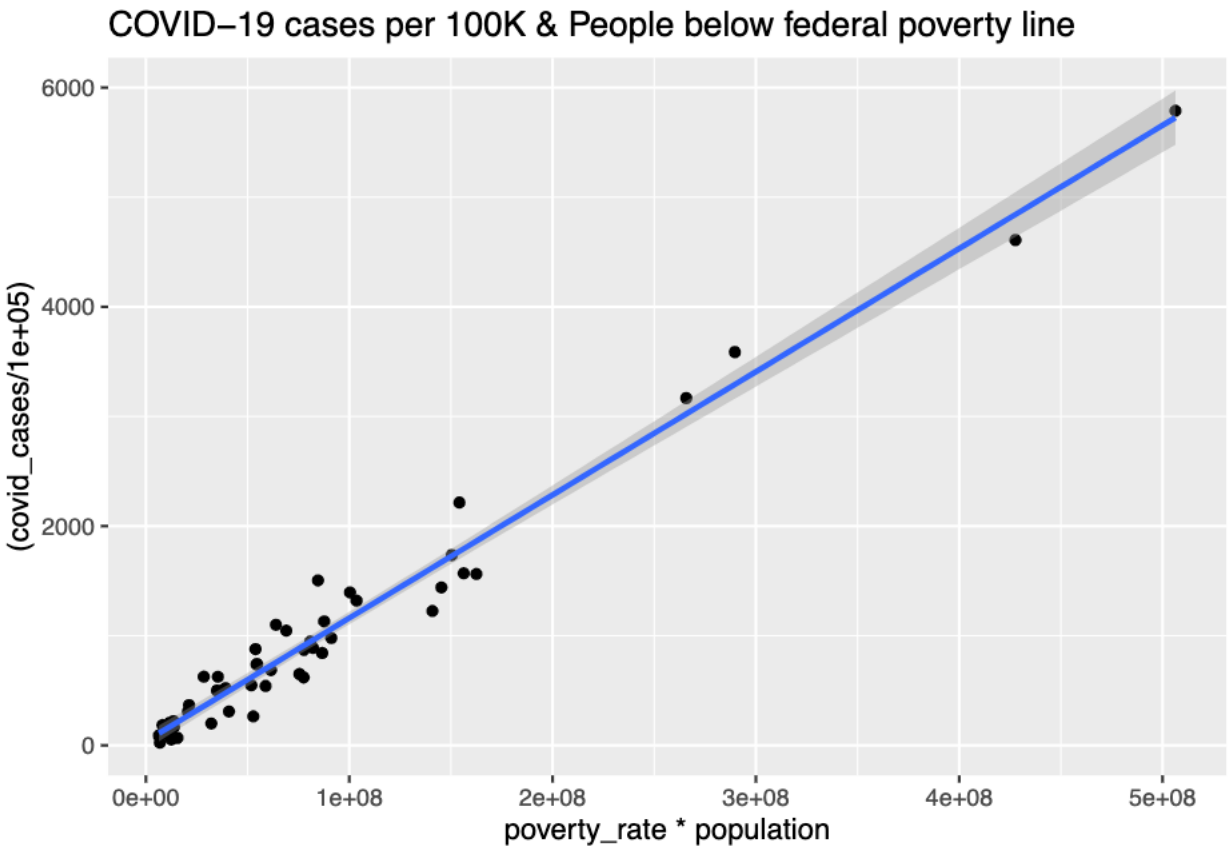
The graph also shows a sparse representation of a normal distribution. There are around 5 to 6 states that have significantly larger number of Covid cases that can be seen from the above graph on the right side. This aligns with the fact that some states in the US are highly populated and most of them are relatively less populated.

There are 2 states with highest cases in the far right with over 400,000,000 cases.

```
## # A tibble: 2 x 3
##   State      covid_cases poverty_rate
##   <chr>            <int>        <dbl>
## 1 California   578892907         12.8
## 2 Texas        460947259         14.9
```

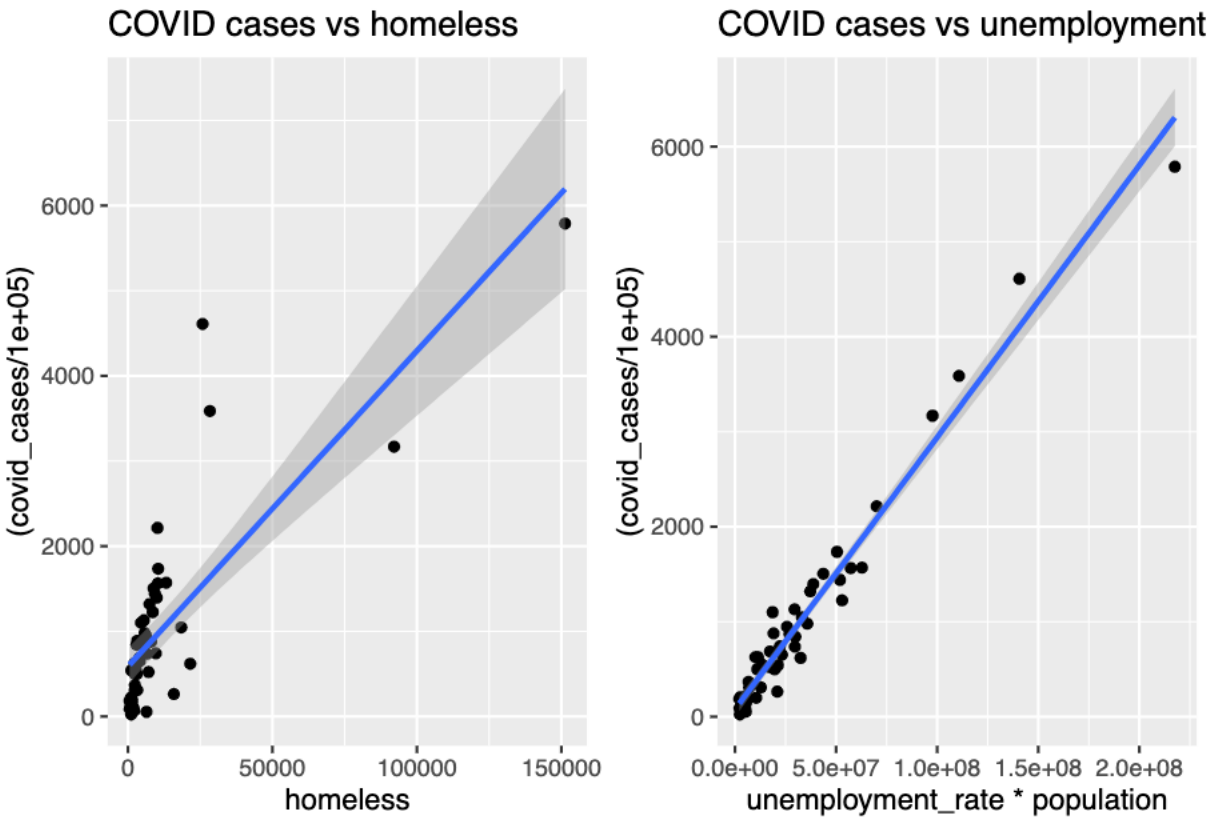
After querying the number of states with the higher number of cases (more than 4000 cases per 100K people), it can be clearly identified that California and Texas takes the lead. These are the states in the far right of the above histogram.

Let's draw a scatter plot to see the trend between cases and poverty rate



The scatterplot shows a positive relation between Covid cases and people below federal poverty line. We can infer from this graph that the states with more people below federal poverty line are more prone to get Covid. This may be because of several reasons such as lack of basic amenities, hygiene, cleanliness, less awareness of their surroundings. They would be more busy in fulfilling their basic necessities than taking precautionary steps for maintaining their health.

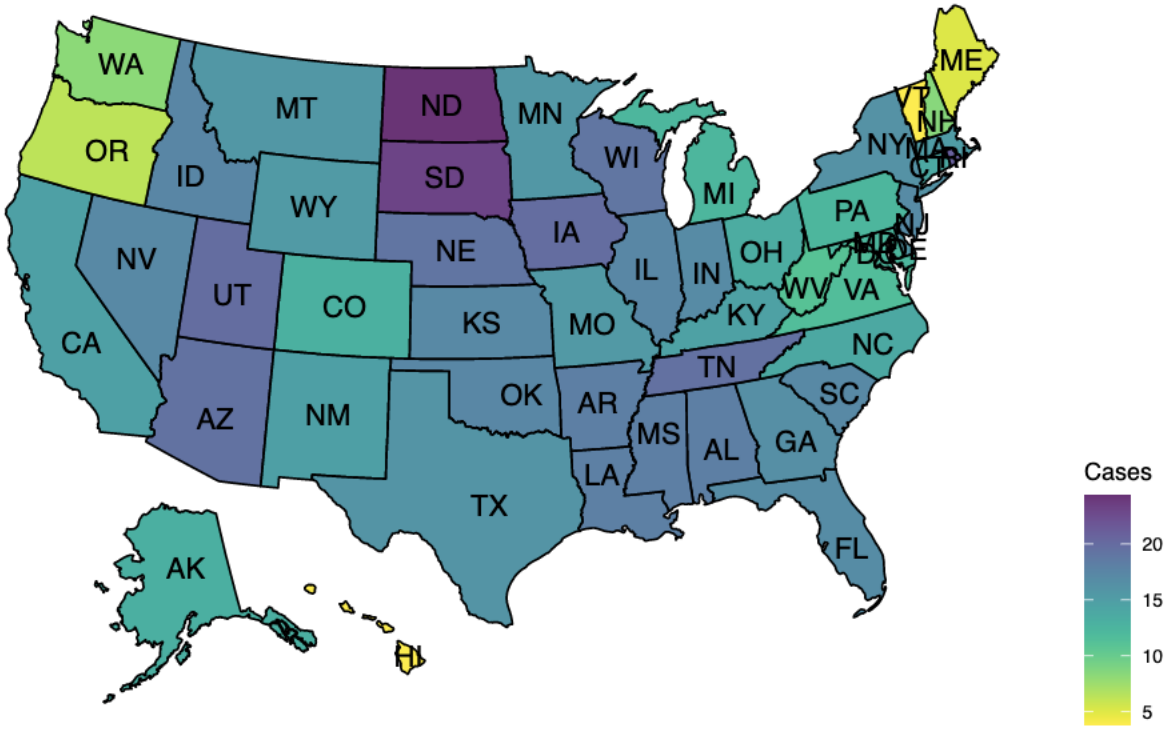
Let's explore further to see if there is any relationship between poverty and unemployment as well as between poverty and being homeless.



The above scatter plot again shows a strong and positive relationship between unemployment and Covid cases per state. This may be due to the reason that the unemployed people would be more worried about their livelihood and paying bills than taking precautionary steps for maintaining their health, which would also require some additional dollars.

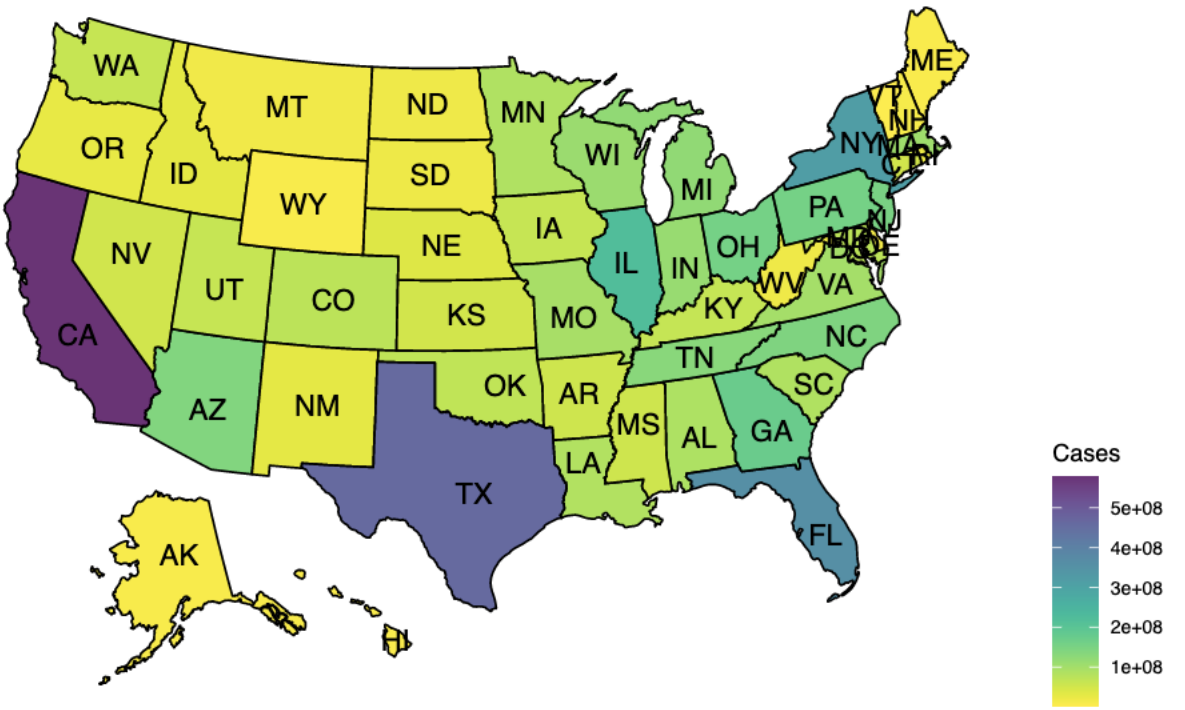
However, the same inference cannot be drawn about the relationship between Covid cases and homeless people per state. Some states can have implemented better and timely state better policies for protection and benefits for homeless people compared to another state.

Map Figure 1: COVID−19 Cases by State population



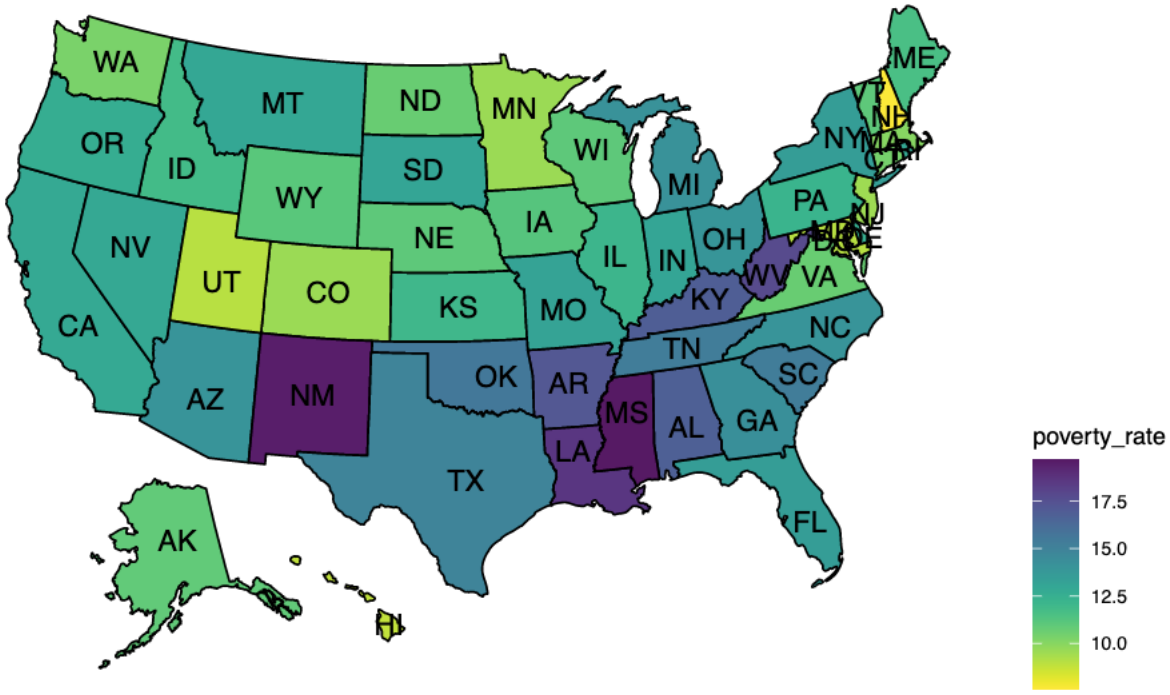
Map Figure 1 shows the distribution of case ratio per state i.e. the total number of covid cases/total population. It clearly shows that North Dakota and South Dakota has highest number of ratio for Covid cases for their population.

Map Figure 2: COVID−19 Cases by State



Map Figure 2 shows the total number of covid cases by state that clearly shows California, Texas, New York and Florida in the high range.

Map Figure 3: Percent below federal poverty line by State



Map Figure 3 shows the distribution of percent below federal poverty line by State. There are 4 states - New Mexico, Mississippi, Louisiana and West Virginia Louisiana that can be seen in the US state policy dataset as well.

Let's look at a correlation matrix of the variables of interest as it may help explain the spread of COVID-19. These variables are:

- population density
- UI max (unemployment insurance max)
- Stimulus
- UI extended (unemployment insurance extended)
- Face mask
- Poverty rate
- unemployment rate
- Covid cases
- Homeless

The correlation matrix can help us understand how correlated our variables are to covid_cases variable and to each other.

From this correlation matrix, we can infer that the number of COVID-19 cases has a positive correlation with poverty and unemployment rate. In contrast, the covid_cases variable is negatively correlated with stimulus provided by the state government as well as with the Weekly unemployment insurance maximum amount (UI_max).

4. A Model Building Process

For all the models, robust standard errors has been used, due to the heteroskedastic nature of errors

4.1. **Limited Model**: This model includes *only the key predictor variable* that is *poverty_rate*. This variable was not transformed, as determined by the EDA above.

Model 1 has been specified as:

$$\log(covid\_cases) = \beta_0 + \beta_1 poverty\_rate$$

```
model1 <- lm(log(covid_cases) ~ poverty_rate, data = covid_data)
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.684726   0.743523 22.4401  < 2e-16 ***
## poverty_rate  0.091640   0.054198  1.6908  0.09735 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(model1)
```

```
##
## Call:
## lm(formula = log(covid_cases) ~ poverty_rate, data = covid_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0101 -0.8867  0.1417  0.7733  2.3189
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.68473    0.76075  21.932   <2e-16 ***
## poverty_rate  0.09164    0.05786   1.584     0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.146 on 48 degrees of freedom
## Multiple R-squared:  0.04966,    Adjusted R-squared:  0.02986
## F-statistic: 2.508 on 1 and 48 DF,  p-value: 0.1198
```

Model one is the model that includes the one key variable `poverty rate` that has been measured to understand the impact on covid cases. It can be derived from the results of model 1 that if the poverty rate increases by 1 unit, the number of covid cases will increase by 9%.

Even though the coefficient t test returns the coefficient of poverty rate to not be significant, an increase of 9% in the number of covid cases holds practical significance. The R Squared values indicate that poverty rate doesn't do a significantly good job of adding explanatory power to number of covid cases.

4.2. **Model Two**: for model 2, let's add a few more variables that is expected to be significant in explaining the variation in number of COVID-19 cases among states. The additional covariates include unemployment_rate, log(pop_density), and face_mask.

$$log(covid\_cases) = \beta_0 + \beta_1 * poverty\_rate + \beta_2 * unemployment\_rate + \beta_3 * log(pop\_density) + \beta_4 * face\_mask + \text{u}$$

```r
model2 <- lm(log(covid_cases) ~ unemployment_rate + poverty_rate +
               log(pop_density) + face_mask, data = covid_data)

coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       14.433330   0.872913 16.5347 < 2.2e-16 ***
## unemployment_rate  0.016888   0.206182  0.0819 0.9350829
## poverty_rate       0.098416   0.068721  1.4321 0.1590200
## log(pop_density)   0.454605   0.114741  3.9620 0.0002625 ***
## face_mask          0.180886   0.346451  0.5221 0.6041545
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(model2)
```

```
##
```

```
## Call:
## lm(formula = log(covid_cases) ~ unemployment_rate + poverty_rate +
##     log(pop_density) + face_mask, data = covid_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9769 -0.3057  0.1858  0.5227  1.8426
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       14.43333    0.87621  16.473  < 2e-16 ***
## unemployment_rate  0.01689    0.19362   0.087 0.930882
## poverty_rate       0.09842    0.06655   1.479 0.146124
## log(pop_density)   0.45460    0.11327   4.013 0.000224 ***
## face_mask          0.18089    0.30941   0.585 0.561729
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 45 degrees of freedom
## Multiple R-squared:  0.3216, Adjusted R-squared:  0.2613
## F-statistic: 5.334 on 4 and 45 DF,  p-value: 0.001329
```

The model results show that only log(pop_density) has a significant explanatory power in the model when controlling for the other variables. This is expected as in states where population density is low, the risk of spreading covid is also low - the opposite of high density states.

From the coefficients, it can be seen that if the population density increases by 1%, the number of covid cases would increase by 0.45%.

In this model, both unemployment_rate and poverty_rate has been used as the two main proxy variables for socio-economic conditions in the state. In addition, a binary variable, `face_mask` has also been used, expecting that in states where face mask rules were enforced, the covid cases would be significantly lower, controlling for all other variables in the model. The face_mask is not a significant variable in this model. The reason for that might be due to the fact that face mask rules were imposed after the covid cases were already high in some of the big states.

There is a significant improvement from model 1 to model 2 in terms of explanatory power of the model. The F-statistic has a p-value of 0.00142 compared to 0.118 in the first model.

4.3. **Model Three**:

This model includes the additional covariate of the number of homeless people in a state.

$$log(covid\_cases) = \beta_0 + \beta_1 * poverty\_rate + \beta_2 * unemployment\_rate + \beta_3 * log(pop\_density)$$
$$+\beta_4 * face\_mask + \beta_5 * homeless + \text{u}$$

```r
model3 <- lm(log(covid_cases) ~ unemployment_rate + poverty_rate + log(pop_density) +
               face_mask + homeless, data = covid_data)

coeftest(model3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                      Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)        1.4797e+01  7.8909e-01 18.7525 < 2.2e-16 ***
## unemployment_rate -6.7287e-02  1.9067e-01 -0.3529  0.725850
## poverty_rate       1.1632e-01  6.3303e-02  1.8375  0.072901 .
## log(pop_density)   3.7511e-01  1.1414e-01  3.2864  0.001998 **
## face_mask          3.4187e-02  3.1587e-01  0.1082  0.914305
## homeless           1.8425e-05  1.1225e-05  1.6415  0.107833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(model3)
```

```
##
## Call:
## lm(formula = log(covid_cases) ~ unemployment_rate + poverty_rate +
##     log(pop_density) + face_mask + homeless, data = covid_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7593 -0.3883  0.1265  0.5467  1.5195
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.480e+01  8.018e-01  18.455  < 2e-16 ***
## unemployment_rate -6.729e-02  1.773e-01  -0.379 0.706202
## poverty_rate       1.163e-01  6.056e-02   1.921 0.061254 .
## log(pop_density)   3.751e-01  1.055e-01   3.556 0.000914 ***
## face_mask          3.419e-02  2.840e-01   0.120 0.904719
## homeless           1.843e-05  5.611e-06   3.284 0.002011 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9065 on 44 degrees of freedom
## Multiple R-squared:  0.4552, Adjusted R-squared:  0.3933
## F-statistic: 7.352 on 5 and 44 DF,  p-value: 4.406e-05
```

In this model, another additional variable **homeless** has been added, which is the number of homeless people in each state. This variable is correlated to population and would have an impact on covid cases. Considering the homeless population is one of the high risk groups and is highly correlated with the population size - since the states with bigger population have a bigger homeless population as well - it can be seen that **homeless** is a significant variable in the model. In addition, in this model the poverty_rate has a 0.05 significance level and the model Adjusted R-squared has increased to 0.39, while F-statistic to 7.27.

Comparing models 2 and 3 to analyze the variance through an f-test:

```r
anova(model3, model2, test = 'F')
```

```
## Analysis of Variance Table
```

```
##
## Model 1: log(covid_cases) ~ unemployment_rate + poverty_rate + log(pop_density) +
##     face_mask + homeless
## Model 2: log(covid_cases) ~ unemployment_rate + poverty_rate + log(pop_density) +
##     face_mask
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     44 36.156
## 2     45 45.019 -1   -8.8624 10.785 0.002011 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above test has a 0.001 significance level, telling us that the new coefficient would improve the model performance overall.

Despite the fact that homeless was a significant variable, and the model performance improved, there is a chance of some potential collinearity problems in the model. As previously stated, homeless is a variable that is highly correlated with population. Along with that, poverty rate and unemployment would be somewhat related to the number of homeless as well.

5. A Regression Table

```
se.classic1 <- data.frame(model1$coefficients)$std..Error
se.classic2 <- data.frame(model2$coefficients)$std..Error
se.classic3 <- data.frame(model3$coefficients)$std..Error

stargazer(model1, model2, model3, type = "text",
          se = list(se.classic1, se.classic2, se.classic3),
          title = "Table 1: OLS models for COVID-19 Spread")
```

```
##
## Table 1: OLS models for COVID-19 Spread
## ================================================================================
##                                        Dependent variable:
##                     ---------------------------------------------------------
##                                          log(covid_cases)
##                            (1)                 (2)                  (3)
## --------------------------------------------------------------------------------
## unemployment_rate                           0.017                -0.067
##                                            (0.194)              (0.177)
##
## poverty_rate              0.092             0.098                0.116*
##                          (0.058)           (0.067)              (0.061)
##
## log(pop_density)                          0.455***             0.375***
##                                            (0.113)              (0.105)
##
## face_mask                                   0.181                0.034
##                                            (0.309)              (0.284)
##
## homeless                                                       0.00002***
##                                                                (0.00001)
##
## Constant                 16.685***         14.433***            14.797***
##                           (0.761)           (0.876)              (0.802)
##
## --------------------------------------------------------------------------------
## Observations                50                50                   50
## R2                         0.050             0.322                0.455
## Adjusted R2                0.030             0.261                0.393
## Residual Std. Error  1.146 (df = 48)  1.000 (df = 45)      0.906 (df = 44)
## F Statistic         2.508 (df = 1; 48) 5.334*** (df = 4; 45) 7.352*** (df = 5; 44)
## ================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

With model 1, the primary goal was to explore the causal relation between poverty rate and number of covid cases. With model 2, additional variables were used as proxies to socioeconomic conditions along with the population density. With model 3, correlated covariate `homeless` was added to study the relationship.

It can be seen that from model 1 to model 3, there are significant improvements. Model 2 explains about 25% of the variance in the number of covid cases while model 3 explains about 39% of the variance in the number of covid cases. Even so, out of all the three models, model 2 has been considered to be the primary model.

Model 3 includes the number of homeless as a variable which would practically be related to poverty rate and unemployment rate. With model 2, taking population density, unemployment rate, poverty rate and face

mask as the predictor variables adds sufficient interpretability to the impact on the number of covid cases per state. Thus with this, model 2 is our most parsimonious and more interpretable out of all the models.

For models 2 and 3, we can see that log(population_density) is statistically significant. This goes along with intuition since the number of covid cases would be affected greatly by crowded places and an increase in population in an area. Thus, population density emerges as an important predictor variable.

However, on the other hand, the variable of number of homeless in model 3 shows as significant but with a coefficient as small as 0.00002, there is no practical significance.

Even with these estimates and interpretations for the primary causal question, there are biases included in the model which have been discussed in the next section, the primary one being geographical clustering and dependency.

6. Limitations of Model (CLM)

There are 5 assumptions for a Classic Linear Model, as elaborated below:

1. IID Sampling
2. Linear Conditional Expectation
3. No Perfect Collinearity
4. Homoskedastic Errors
5. Normally Distributed Errors

These assumptions are being evaluated in reference to the following model from above:

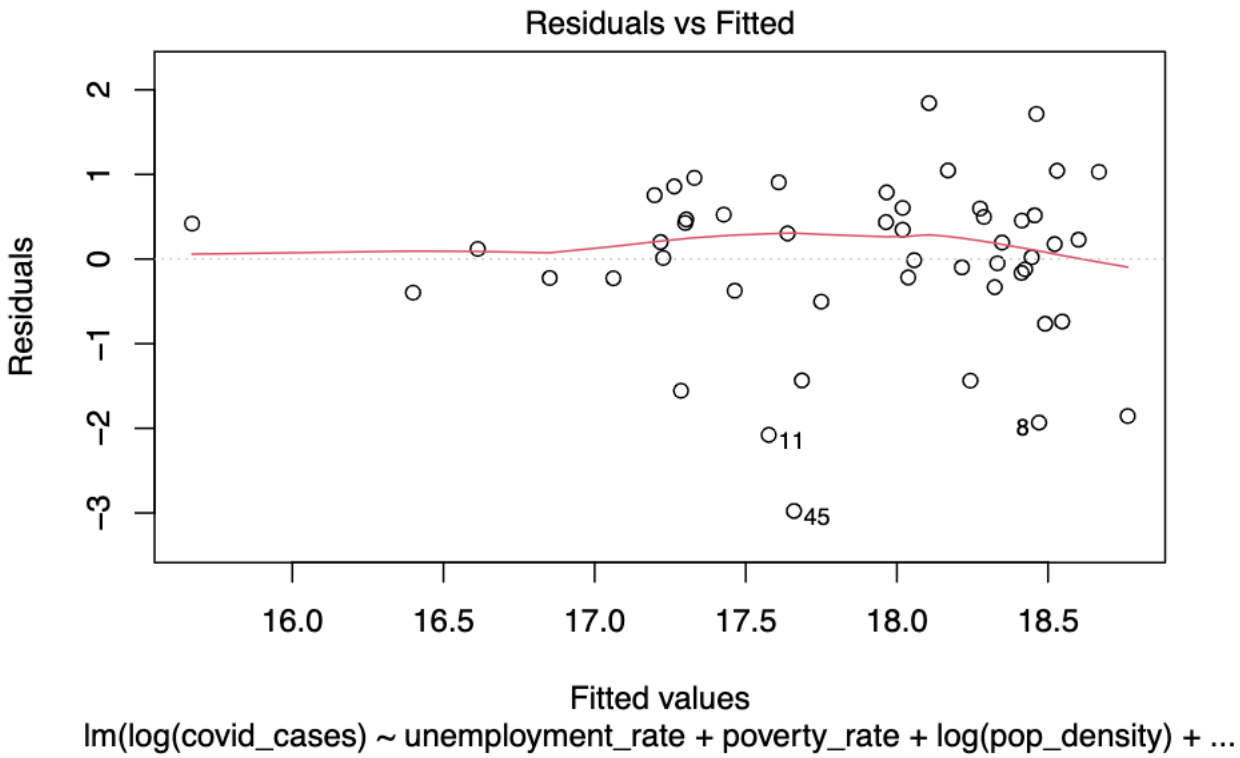
$$log(covid\_cases) = \beta_0 + \beta_1 * poverty\_rate + \beta_2 * unemployment\_rate + \beta_3 * log(pop\_density) + \beta_4 * face\_mask + \text{u}$$

Evaluating the assumptions individually:

1. IID Sampling: IID sampling assumption requires the data to be independently and identically distributed. The data we are using is being collected at the state level. This would lead to geographical clustering of some degree. Moreover, different states depending upon their population density, economic conditions and political distributions will not have identical distributions. Some states might be similar but all the states will not have the same distribution. Neighboring states will also have dependency due to movement of people between the states. With this, IID assumption does not hold here. Even though it doesn't stand, Let's still continue with the regression.

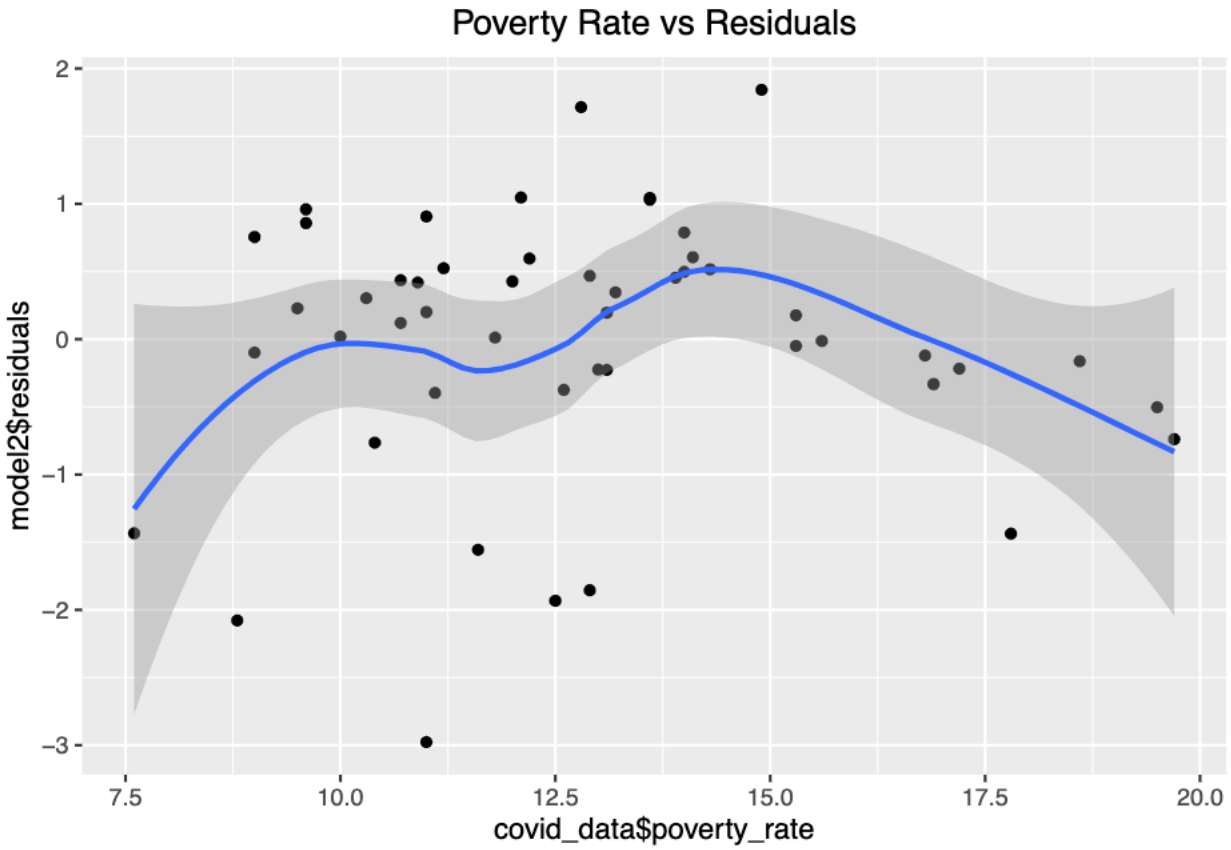
2. Linear Conditional Expectation and Zero Conditional Mean: To validate these assumptions, let's first look at the plot of the residuals vs fitted values for model 2(which has been specified above as well).

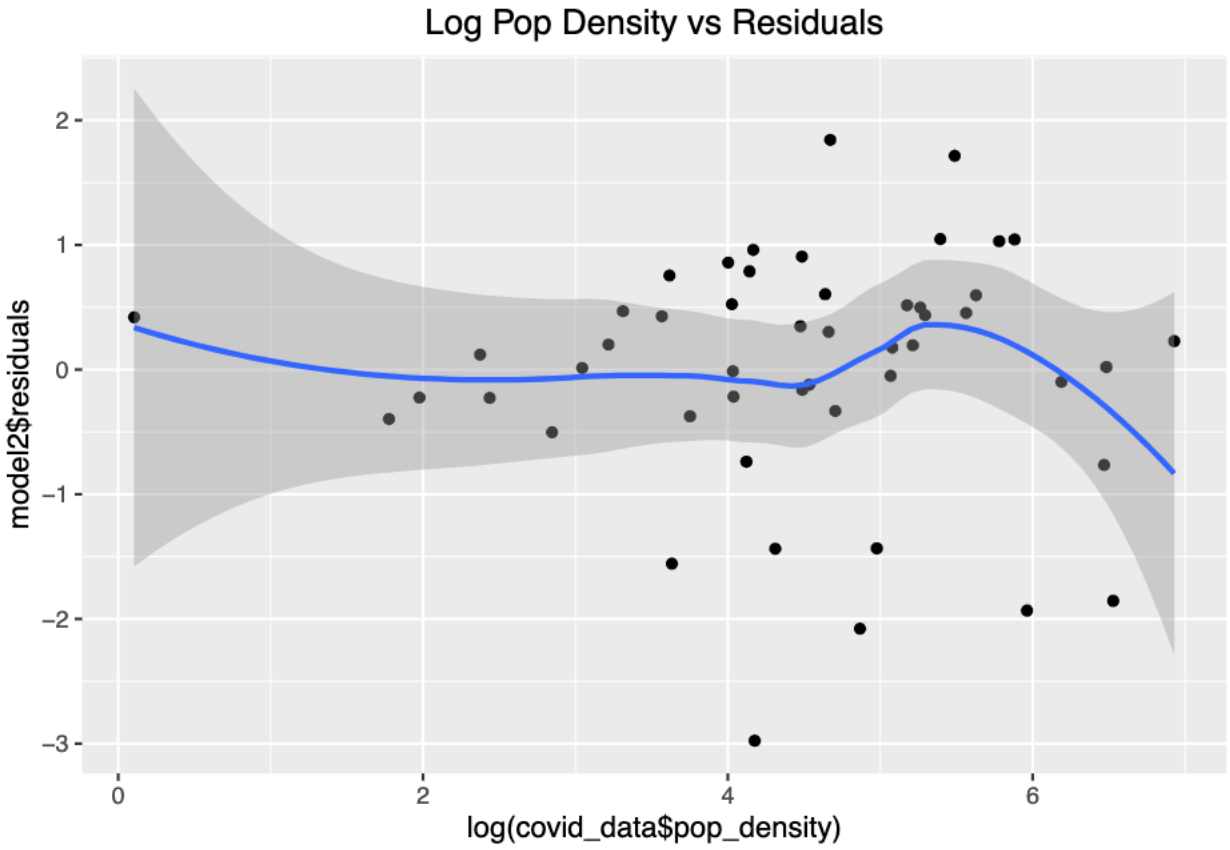
The residuals vs fitted shows an almost flat line(red line), except with a slight curvature. This is probably because of the outliers we have seen earlier in the covid cases and poverty rate across some of the states like California, North Dakota, Texas, and Florida.



Residuals vs Fitted Values for Model

Plotting our key predictor variables of poverty rate and population density with the residuals from the model:





Both of these graphs are indicating that there are outliers causing curvatures in the flat line. For poverty rate, the curvatures are more than for population density. This is probably because poverty rate has extremes in few states with extremely large values in some while extremely small values in others.

3. No Perfect Collinearity

To check for perfect collinearity, Let's check if R dropped any coefficients from the ones we used while creating the model

```
##       (Intercept) unemployment_rate      poverty_rate  log(pop_density)
##       14.43333020        0.01688797        0.09841630        0.45460496
##         face_mask
##        0.18088551
```

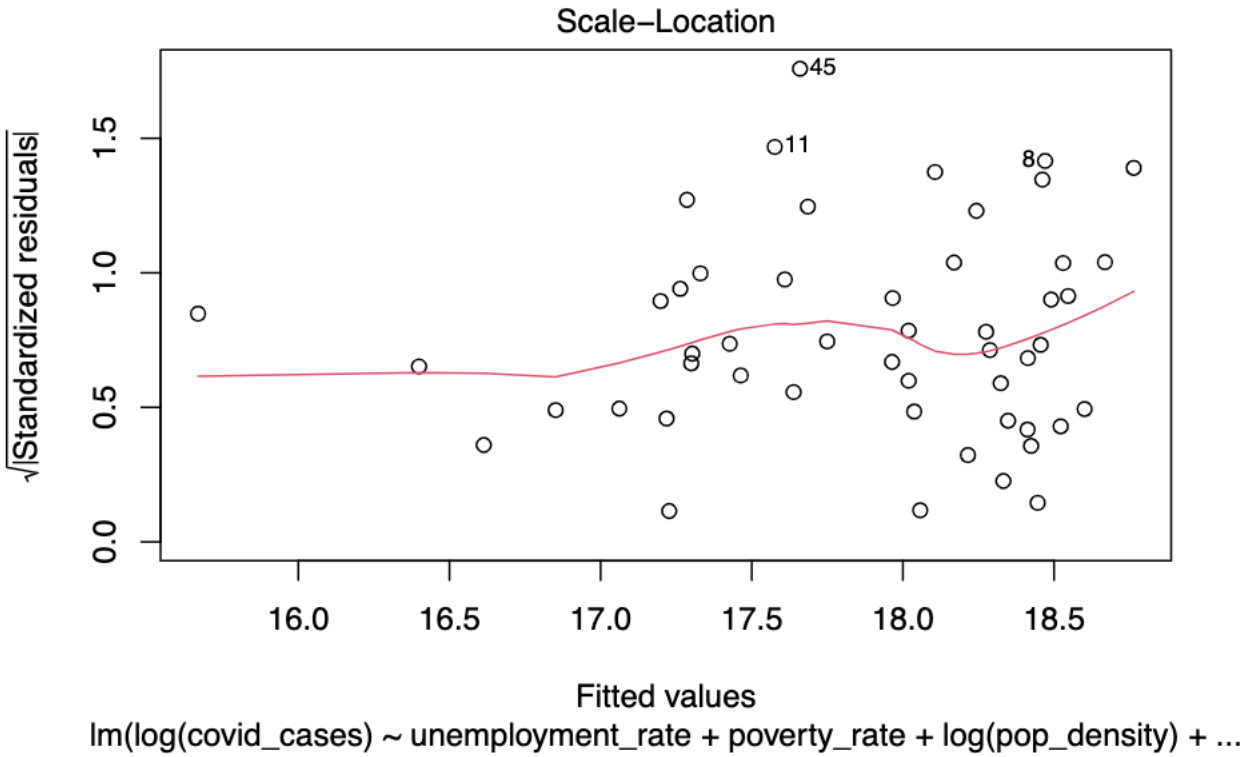
As we can see, R didn't drop any any input predictors and thus, there is no perfect collinearity between the variables.

We can also check this by looking at the values returned by the variance inflation factor. As most of these values are around 1, we can safely assume that there is no concern of perfect collienarity affecting our model.

```
## unemployment_rate      poverty_rate  log(pop_density)         face_mask
##          1.827106          1.737226          1.127483          1.073712
```

4. Homoskedastic Errors

We can investigate the homoskedasticity of the errors by looking at the graph of fitted values vs the square root of standardized residuals.

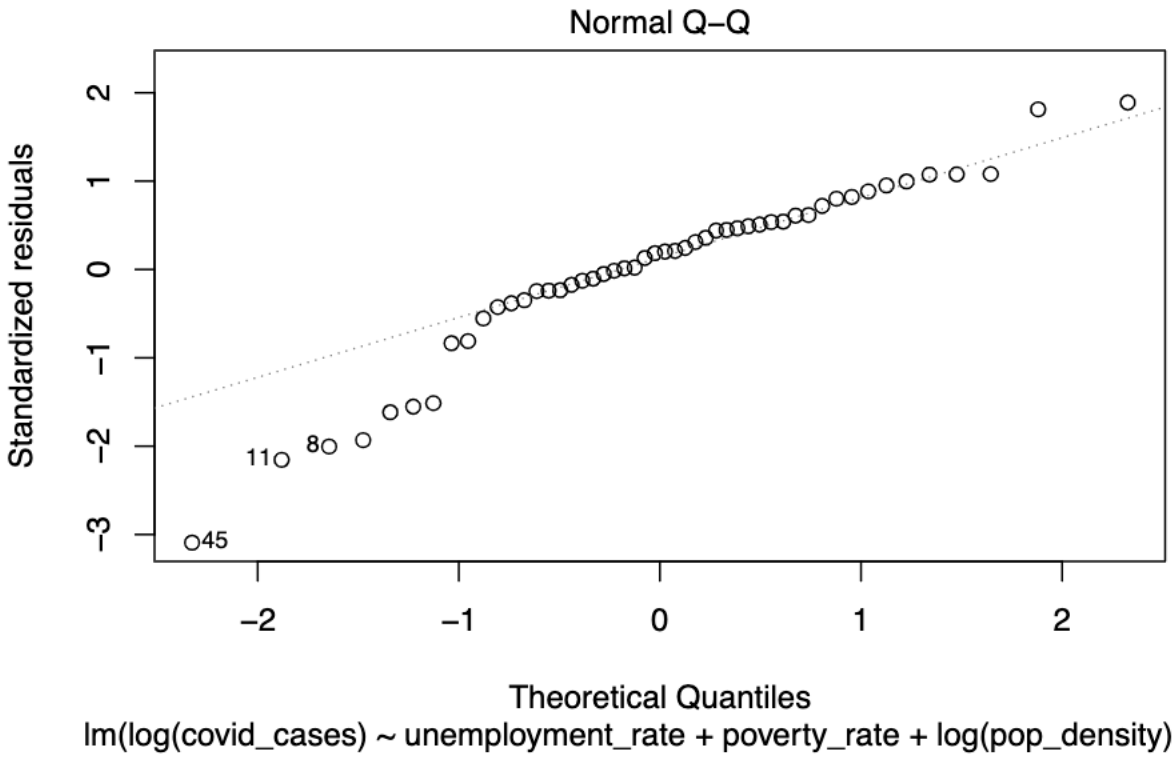


The scale-location plot starts with being flat, generally linear, but there are few curves post that. This is probably because of the outliers been identified above such as, California and Florida.

In addition to this, Robust standard errors has also been used in all our models.

5. Normally Distributed Errors

To investigate the ditribution of errors, let's look at the Normal Q-Q plot.



As we can see that the values are mostly lying on the linear line, just like we would want it to be. However, at the two ends, we can see some outliers and this is probably cause of the outliers we noticed across few states previously mentioned.

7. Discussion of Omitted Variables

Some of the omitted variables that could affect our primary specification (model 2) are the following:

As a reminder, this is our primary model

$$log(covid\_cases) = \beta_0 + \beta_1 * poverty\_rate + \beta_2 * unemployment\_rate + \beta_3 * log(pop\_density) + \beta_4 * face\_mask + u$$

Omitted Variable	Correlation to Outcome Variable (number of covid cases)	Correlation to Primary Explanatory Variable (poverty rate)	Direction of Bias
Public Transportation	Positive	Positive	Away from zero
Pre Existing Medical Condition	Positive	Positive	Away from zero
Homeless Shelter	Positive	Positive	Away from zero
Covid-Safe Information Awareness	Negative	Negative	Away from zero
Medical Facilities	Negative	Negative	Away from zero

1. Public Transportation: A feature indicating the average usage of public transportation would definitely impact our outcome variable of number of covid cases. It would be positively related to the number of covid

cases since the spread of this diseases closely relates to proximity between individuals. Public transportation would also be positively related to the poverty rate. States with a higher poverty rate would have a higher average usage of public transportation. With this, the omitted variable bias is moving away from zero.

Writing down both the equations:

$$log(covid\_cases) = \beta_0 + \beta_1 * poverty\_rate + \beta_2 * unemployment\_rate +$$

$$\beta_3 * log(pop\_density) + \beta_4 * face\_mask + \beta_5 * public\_transporation + \text{ u}$$

$$public\_transporation = \alpha_0 + \alpha_1 * poverty\_rate$$

$$\text{if } \beta_5 > 0 \text{ and } \alpha_1 > 0, \text{then OMVB} = \beta_5 * \alpha_1 > 0,$$

and since β_1 is greater than zero will be scaled away from zero(more positive) gaining statistical significance

2. Pre Existing Medical Condition: A feature indicating the prevelance of pre existing medical conditions will be positively correlated to the number of covid cases, since Covid-19 affects individuals with certain pre-existing conditions more adversely. Moreover, it will also be positively correlated to the poverty rate since individuals belogingin to the lower strata of economic status will probably have more pre existing health conditions due to the limitations of medicial facilities available to them. With this, we can see that the bias is moving away from zero.

Writing down both the equations:

$$log(covid\_cases) = \beta_0 + \beta_1 * poverty\_rate + \beta_2 * unemployment\_rate + \beta_3 * log(pop\_density) + \beta_4 * face\_mask +$$

$$\beta_5 * pre\_medical + \text{ u}$$

$$pre\_medical = \alpha_0 + \alpha_1 * poverty\_rate$$

$$\text{if } \beta_5 > 0 \text{ and } \alpha_1 > 0, \text{then OMVB} = \beta_5 * \alpha_1 > 0,$$

and since β_1 is greater than zero will be scaled away from zero(more positive) gaining statistical significance

3. Homeless Shelters: The count of homeless shelters in the state is positively corelated to the poverty rate of the state. Moreover, homeless shelters, due to their density and inability to provide distance between individuals, are positively correlated to number of covid cases as well. Given this, the omitted variable bias would be away from zero.

Writing down both the equations:

$$log(covid\_cases) = \beta_0 + \beta_1 * poverty\_rate + \beta_2 * unemployment\_rate + \beta_3 * log(pop\_density) + \beta_4 * face\_mask +$$

$$\beta_5 * homeless\_shelter + \text{ u}$$

$$homeless\_shelter = \alpha_0 + \alpha_1 * poverty\_rate$$

$$\text{if } \beta_5 > 0 \text{ and } \alpha_1 > 0, \text{then OMVB} = \beta_5 * \alpha_1 > 0,$$

and since β_1 is greater than zero will be scaled away from zero(more positive) gaining statistical significance

4. Covid-Safe Information Awareness: Awareness of covid-safe information would be negatively correlated to the number of covid cases in the state. With more awareness and acceptance, there will be lesser covid cases. This omitted variable would be negatively correlated to the poverty rate as well since information awareness among the lower income strata is less. With this, the omitted variable bias will be positive and it will be away from zero.

Writing down both the equations:

$$log(covid\_cases) = \beta_0 + \beta_1 * poverty\_rate + \beta_2 * unemployment\_rate + \beta_3 * log(pop\_density) + \beta_4 * face\_mask +$$

$$\beta_5 * covid\_awareness + \text{ u}$$

$$covid\_awareness = \alpha_0 + \alpha_1 * poverty\_rate$$

$$\text{if } \beta_5 < 0 \text{ and } \alpha_1 < 0, \text{then OMVB} = \beta_5 * \alpha_1 > 0,$$

and since β_1 is greater than zero will be scaled away from zero(more positive) gaining statistical significance

5. Availability of Medical Facilities: The availability of medical facilities in a state is negatively correlated to the number of covid cases. On the other hand, medical facilities are not readily available for individuals belonging to the lower income strata of the society. With that, it would be negatively correlated to the poverty rate. And thus, the omitted variable bias will be away from zero.

Writing down both the equations:

$$log(covid\_cases) = \beta_0 + \beta_1 * poverty\_rate + \beta_2 * unemployment\_rate + \beta_3 * log(pop\_density) + \beta_4 * face\_mask +$$

$$\beta_5 * medical\_facilities + \text{ u}$$

$$medical\_facilities = \alpha_0 + \alpha_1 * poverty\_rate$$

$$\text{if } \beta_5 < 0 \text{ and } \alpha_1 < 0, \text{then OMVB} = \beta_5 * \alpha_1 > 0,$$

and since β_1 is greater than zero will be scaled away from zero(more positive) gaining statistical significance

8. Conclusion

The primary research question was "Is there a causal relationship between different socio-economic conditions such as poverty, unemployment rate and the number of cases related to Covid 19?". The question stemmed from news, reports, and updates over the past 12 months about different sections of the society being affected differently and sometimes more adversely by COVID -19.

$$log(covid\_cases) = \beta_0 + \beta_1 * poverty\_rate + \beta_2 * unemployment\_rate + \beta_3 * log(pop\_density) + \beta_4 * face\_mask + \text{ u}$$

Through this primary model(model 2), I studied the relationship between the number of covid cases and socio economic conditions along with population density and the mandate of face mask. The model explained 25% of the variance in the number of covid cases. Population density emerged as a significant predictor variable. This aligns with the information around the pandemic that Covid-19 spreads more with crowds and thus, population density. Even though poverty rate didn't emerge statistically significant, a 1% increase in poverty rate leads to a 0.45% increase in the number of poverty case. Thus, a higher poverty rate definitely affects the number of covid cases practically and in the real world.

I also considered the number of homeless per state as a predictor variable as well but due to its close relation with poverty rate and unemployment rate, I decided to go with the model without this. and also avoided using this highly correlated variable to avoid inflating the outcome.

For the model 2, that emerged as the primary model, CLM assumptions were measured and it has been found that the slight deviations in the assumptions were owing to outlier states like Florida, California and North Dakota. The primary CLM assumption that was violated was independent and identical sampling

of data. With this, I am aware that the coefficients could potentially be biased. I am also aware that the models suffer from omitted variable bias.

The problems associated with this model and data powering the models can be addressed by probably collecting samples at a further granular scale, allowing for a bigger sample set, and along with that, resampling for independence.

I believe that there is still room for more research to understand the true significance of socio-economic conditions. This further exploration and study includes doing a time series regression analysis, considering that most of the essential factors in the number of covid cases have changed in time. With those covariates in the model as time series,I would further test how socio-economic conditions, which are more static, contribute to the number of covid cases.

With this, I would like to conclude with stating that their is a causal relation between the number of covid cases in a state and the poverty rate of that state. And this understanding can be further enhanced with more research.