W203 - Lab 2

The Relationship Between Mask-Use and COVID-19 Cases

Devashish Kulkarni, Tiantian Zhao, Yuqiao(Esther) Chen

Investigating the Relationship between Mask Usage and Covid-19 Cases

Intoduction

Throughout the past 18 months, all the nations of the world have been grappling with a once in a century pandemic. The spread of the Covid-19 disease has taken a huge toll on global health, adversely affected the entire global economy and drastically transformed the daily lives of virtually every person on earth, making it a noteworthy event in the course of human history. Experts warn that similar pandemics can and will occur in the future, and hence, must be prepared for. Investigating characteristics of the Covid-19 pandemic provides an opportunity to learn insights that have an incredible potential to mitigate harm from future disease epidemics.

A crucial tool to control the spread of the Coronavirus was the use of face masks by the general population, often mandated by their respective governments. However, due to the political situation in the United States, the federal government failed to put forth a clear message on the critical importance of mask wearing to combat the pandemic. As a result, polarized sections of society made decisions on the use of face masks based on their individual political leanings, rather than the directives of public health agencies, leading to a wide variety of mask use behaviors in different locations around the country. In this paper, we aim to investigate the effect of mask use behaviors on Covid-19 cases. Specifically, we address the following research question:

"How does the usage of masks by the general public affect the number of Covid-19 cases in their respective communities?"

The mechanism behind the usage of face masks on controlling the spread of a respiratory disease seems self evident. Numerous studies[1] have confirmed the efficacy of mask use to reduce transmission of the virus. We present a data driven investigation by counties in the United States into whether or not a more frequent usage of masks by people in a county causes lower Covid-19 case counts in that particular county. In addition, we also explore the effects of various county level attributes on the Covid-19 cases counts, including, education levels, proportion of seniors, percentage of females and classification as urban and rural. We use OLS regression as a strategy to build a series of explanatory models to investigate the effects of our input variables, primarily mask use behavior on our outcome variable, namely Covid-19 cases.

Description of Data

The following section describes the process used to operationalize each of the variables that are used in this analysis. The data for mask use behavior, our primary input variable is provided by the New York Times[2]. The data was collected through a online interviews performed by Dynata, a global data and survey firm, at the request of the New York Times. The survey was conducted between July 2 and July 14, 2020. About

[1] An evidence review of face masks against COVID-19: https://www.pnas.org/content/118/4/e2014564118

[2] Estimates from The New York Times, based on roughly 250,000 interviews conducted by Dynata from July 2 to July 14: https://github.com/nytimes/covid-19-data/tree/master/mask-use

250,000 survey responses were used to estimate frequency of mask use at a county level, through appropriate weighting by age and gender. ZIP codes were used to approximate the locations of the survey respondents. The data for each observation is provided in a categorical fashion representing various levels of frequency of mask use. The mask use frequency levels are specified as NEVER, RARELY, SOMETIMES, FREQUENTLY AND ALWAYS. Numerical values, provided for each of the frequency levels, represent the fraction of survey respondents that answered the question "How often do you wear a mask in public when you expect to be within six feet of another person?" with one of the frequency level. Each observation represents mask use behavior for a county in the United States. In the interest of parsimony, we simplify the variable for use in our regression models. We define a respondent that answer the survey question with FREQUENTLY or ALWAYS as a frequent mask user. We use this definition for frequent mask user as it captures the attitude of the respondent towards mask use. Since the guidance from public health agencies at the time when this data was collected was to always use masks within six feet of another person, we argue that NEVER, RARELY or SOMETIMES wearing a mask does not sufficiently follow this guidance. We use the fraction of frequent mask users in a county to represent that mask use behavior in a particular county, calculated simply by summing the numerical values provided in the FREQUENTLY and ALWAYS categories for each county.

The data source for the outcome variable is the Covid-19 data base maintained by the New York Times[3]. Since the report of the first Coivd-19 cases in Washington state, the Times has tracked cases of Covid-19 in real time as they were identified after testing. Because of the shortage in Covid-19 testing at the time, the data is necessarily limited in the picture it presents of the outbreak. The data is gathered from state and local governments and health departments and is made available for public use for non commercial research purposes. Cumulative Covid-19 case counts are provided for each day and for each county starting from Jan 21, 2020. We use the data from July 14, 2020. We choose this particular date because the mask use data collection was performed from July 2, 2020 to July 14, 2020 and the normal incubation time for Covid-19 is considered to be 2-14 days. July 14 is approximately 7 days after the bisection date of the survey process. Each observation represents the number of Covid-19 cases for each county in the United States. We chose to use the number of cases in a county per 100 thousand residents to normalize for the population differences within different counties. We also choose the total number of Covid-19 cases, rather than a change in the number of cases or multiples of number of cases in some duration. We do this because the spread of a disease is a very complex phenomenon that depends on many factors, such as, movement of the public in and out of the county, lockdown rules, general health of the population and so on. To simplify the analysis, we assume that the mask use behavior data denotes the general attitude of the county towards taking precautions against the disease and the number of cases per capita denotes the outcome of the pandemic in the county. We acknowledge that this treatment of the variable is an oversimplification, but argue that there are still interesting insights to be learnt from this analysis.

Another input variable used is education levels in a county in the United States. Data for this variable is provided by the US Department of Agriculture through it's Economic Research Service(ESR)[4] .ESR compiles the latest data on measures such as poverty rates, population change, unemployment rates and education levels. The data is compiled from U.S. Census Bureau Censuses of Population, and the 2015-19 American Community Survey. Each observation corresponds to a county in the United States and contains the educational attainment for adults age 25 and older. The education attainment levels are classified as less than a high school diploma, high school diploma, some college and four years of college or higher. The number of individuals and percentage of adults are reported for each of the education levels from 1970-2019, with a time point for every 10 years until 2000 and a final time point of 2015-2019. For the purposes of this analysis, we consider the percentage of adults with a bachelor's degree as the education level indicator for every county from 2015-2019. We are argue that completing a bachelor's degree is a sufficient education level for an individual to understand the mechanisms of how face masks prevents the spread of respiratory diseases and provides a sense of the expertise that is required for infectious disease experts and public health officials to make decisions regarding mask use directives. Consequently, these individuals are likely to trust the pandemic response guidance provided by experts and at the same time, have enough skepticism to detect

[3]The New York Times. (2021). Coronavirus (Covid-19) Data in the United States. Retrieved August 1st, 2021, from https://github.com/nytimes/covid-19-data."

[4]Economic Research Service U.S. DEPARTMENT OF AGRICULTURE: https://www.ers.usda.gov/data-products/county-level-data-sets/

any inconsistencies in the guidance.

The third input variable used in this analysis is the Urban/Rural classification of a county. The data source for this variable is also the Economic Research Service from the US Department of Agriculture[5]. The data set provides Urban Influence Codes for every county in the United states. The 2013 Urban Influence Codes are used as a classification scheme that distinguishes each county in the United States as metropolitan county or nonmetropolitan county, both further divided into 12 subgroups. A county is is designated as metropolitan if it contains atleast one city having a population of more than 2.5 million residents as defined by the office of Management and Budget[6]. We use the higher level definition of metropolitan vs nonmetropolitan for every county. We argue that the higher population density in a metropolitan region would be a direct influence on the rate of spread of the Coronavirus as an infected person is likely to come in contact with a larger number of persons in a population dense region. Of the 12 Urban influence codes, a county designated as 1 or 2 is considered metropolitan and counties designated as 3 through 12 are considered nonmetropolitan. The data set specifies 1236 counties and regions as metropolitan and 1984 counties and regions as nonmetropolitan.

The fourth input variable considered is the percentage of people in a county over 60 years of age. This data is provided by Kaiser Health News and is hosted online on kaggle.com[7]. The primary focus of this data set is to evaluate the capacity of ICU beds around the nation, but also includes the percentage of people above the age of 60 in every county. This percentage is collected from figures provided by Census Bureau's American Community Survey. We believe that this variable is likely to have a significant effect on the number of Covid-19 cases in a county. Older individuals are likely to be more seriously infected by Covid-19 and hence are likely to show symptoms. Since a Covid-19 case is only registered after a positive Covid test, we argue that a higher percentage of people over 60 in a county would lead to more individuals being symptomatically infected and thus opt for a Covid test. Persons in this age group are less likely to stay as asymptomatic carriers of the disease.

In addition to the above variables, we consider additional variables for inclusion in the analysis that may not have a direct causal link to Covid-19 case loads in a county, but could potential generate interesting insights by inclusion in the analysis. The first variable we consider is the median age of the county. Though this variable has some information overlap with the percentage of people aged more than 60, it also includes information about how young a county is as a whole. Children and younger adults infected with Covid-19 are likelier to be asymptomatic carriers of the disease, and hence, the distribution of the entire population is expected to affect the Covid-19 case count, in addition to the percentage of seniors. The median age is chosen as the statistic to operationalize this variable. This demographic information is provided on the US county level throught the 2014-2018 release of the American Community Survey and is hosted on kaggle.com[8].There is a risk of high correlation between the median age and percentage of seniors variables, hence, we need to be cautious about this fact during our analysis.

The final input variable that we have considered is the percentage of females in a county. This variable also contained in the data set provided by the American Community Survey and hosted on kaggle.com. We are interested in investigating if there are any systematic differences in Covid-19 case rates with the percentage of female population in a county. Any difference might indicate differences in the lifestyle of females and males on average that cause one gender to be more likely to be infect by the virus, such as possibility of remote work or usage of public transit.

Underlying Causal Model

The causal relationships for the variables are depicted in Figure 1. According to our causal theory, mask use behavior affects Covid-19 cases. Most recommended face masks filter out the virus particles and hence reduce the transmission of air borne virus particles from an infected person or to an uninfected person. We also theorize that a county being urban or rural affects both mask use behaviors and Covid-19 cases. Urban counties have a larger population and population density of people and thus would provide an easier pathway

[5]Economic Research Service U.S. DEPARTMENT OF AGRICULTURE: https://www.ers.usda.gov/data-products/urban-influence-codes/

[6]https://obamawhitehouse.archives.gov/sites/default/files/omb/bulletins/2013/b-13-01.pdf

[7]https://www.kaggle.com/jaimeblasco/icu-beds-by-county-in-the-us

[8]https://www.kaggle.com/headsortails/covid19-us-county-jhu-data-demographics?select=us_county.csv

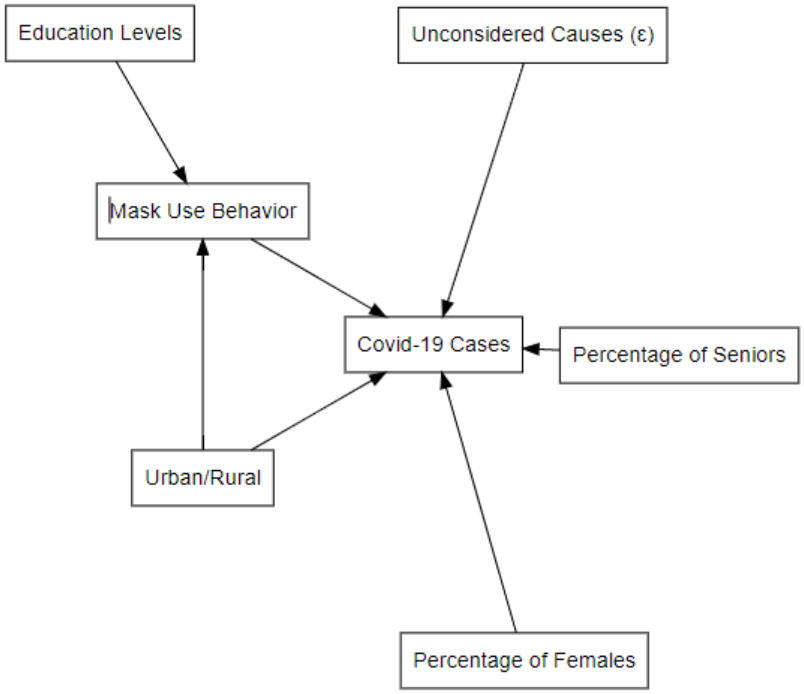

Figure 1: Causal Graph

for virus transmission, compared to their rural counterparts. Urban/rural counties also may affect mask use because of the political atmosphere in the United States during the period under investigation. Rural counties tend to be Republican leaning, and hence are likely to make decisions about mask use based on the political messaging of the party. Many prominent Republican leaders had expressed skepticism about mask usage, which we believe affected the mask use behavior in theie supporters. A similar effect, but in the opposite direction exists in Urban counties. Education is expected to affect mask use behavior. It is reasonable to assume that a higher level of education would contribute to a person's understanding of the mechanism of transmission of the Covid-19 virus. Hence, they are likely to use face masks as a precaution against the disease. Percentage of seniors affect the Covid-19 cases because seniors are likely to have symptomatic infections and hence, opt for a Covid-19 test, which may return positive and add to the Covid-19 case counts. Percentage of females may affect the Covid-19 cases if there is indeed a systematic difference in the lifestyle of females that lead them to be exposed to the virus with a larger probability. All unconsidered causes are grouped together as the epsilon term in the model.

Exploratory Data Analysis

Modeling variables

Our goal of this research is to find "How does the usage of masks by the general public affect the number of COVID-19 cases in their respective communities?". The response variable is the number of COVID-19 cases, which will be standardized to COVID-19 cases per hundred thousand residents for each county. The main explanatory variable is the percentage of masks used by the general population in each county.

Primary Explanatory Variables

Mask use: We consider 'estimated prevalence of mask-wearing' in counties as our primary explanatory variable. There are five categories of the frequency of wearing masks: 'Never', 'Rarely', 'Sometimes', 'Fre-

quently', and 'Always'. We decided to create a new variable called 'frequent_mask_use' by taking the sum of values of 'Frequently' and 'Always' columns, and to use this new variable 'frequent_mask_use' to represent the percentage of people in this county who wear masks most of the time. A histogram of the distribution of frequent.mask.use across counties is shown in Figure 2. We see that the mask use frequency is fairly high for all the counties with the average around 0.75 and minimum around 0.4. This indicates that for most counties, a large fraction of population does frequently use masks.

$$\text{frequent.mask.use} = \text{Frequently} + \text{Always}$$

Response Variable

We considered 'COVID-19 cases per hundred thousand residents' as our primary response variable. We consider two counties to be equally affected by COVID, only if they have the same rate of occurrence of COVID, i.e. the number of COVID cases as a fraction of the total populations is the same. We standardized the total cases per county into total cases per 100,000 residents of the county using the following formula:

$$\text{Cases.Per.100K} = \frac{\text{cases}}{\text{population}} * 100,000$$

The number of cases per county is available as the variable 'cases' from the US COVID-19 dataframe released by New York Times. Each county's population is available as the variable 'population' from the US Demographic dataframe. As the ratio cases/population represents the fraction of population of COVID-19 cases, we multiply it by 100,000 to scale the fraction to one hundred thousand residents. With this transformation Cases.Per.100K now represents the operationalized variable for the primary response variable 'Covid cases per one hundred thousand residents'. We use the cases data for July 14, 2020 as described in the data description section.

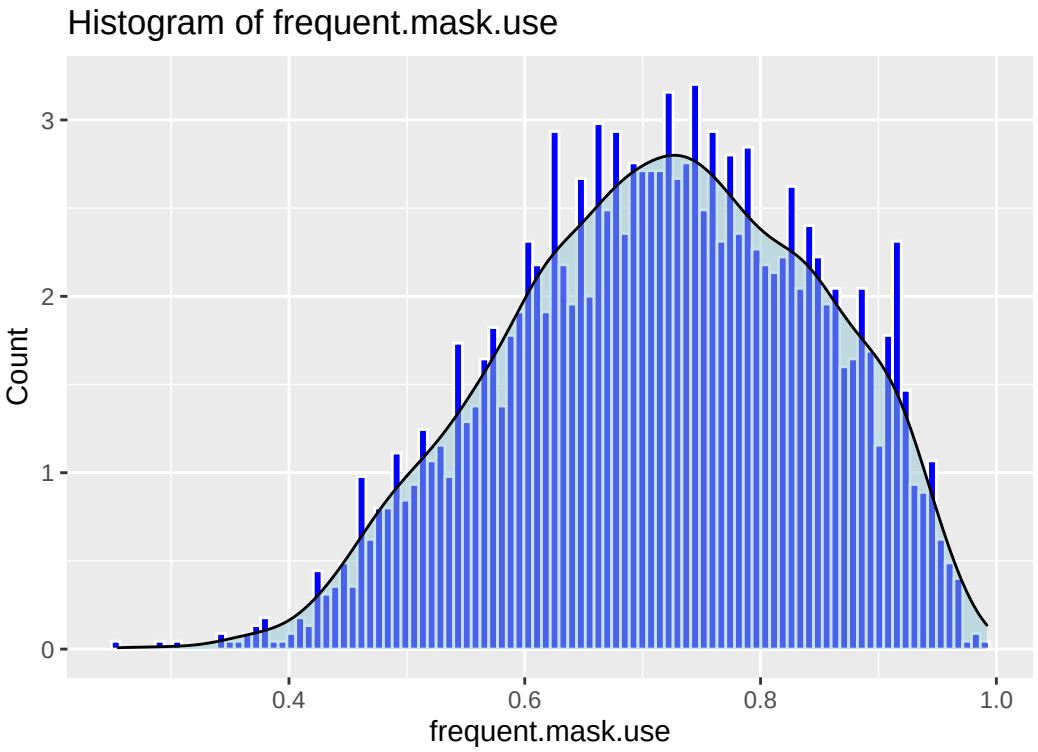

Figure 2: Histogram of mask use behavior

A discussion on the response variable

Figure 3 shows the distribution of Covid-19 cases on the date under consideration across counties in the US. A

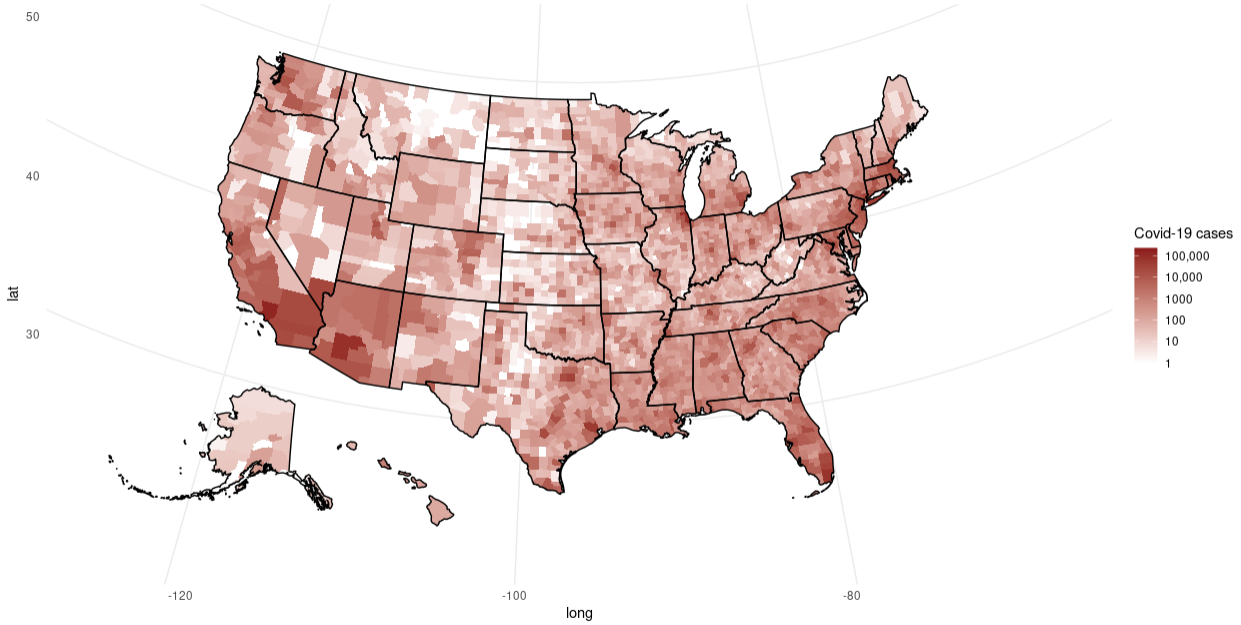

Figure 3: Covid-19 cases distribution by county

log transform is required to visualize the case properly, since there are differences in the orders of magnitude in cases for different counties. It can be seen that county containing major cities in California (LA), Florida (Miami) and New York (NYC) had the highest number of cases at the time. We use a normalization of cases based on the total population of a county to appropriately compare counties with different populations.

Figure 4(a) shows the histogram and density of the response variable 'Cases.Per.100K'. We can observe severe right-biased tails in the histogram of Cases.Per.100K. We applied a log transformation to eliminate these tails, as shown in 4(b). We can see that the curve gets much closer to a normal distribution. We also observe that Case.Per.100K for most of the counties is below 4,000. Using the bar graph on Figure 5, we identify the counties which have more than 4000 cases per hundred thousand residents and we can see that there are 27 counties in total. Interestingly, these are not the counties that had the highest number of total cases as was seen in the map.

Covariate Explanatory Variables

The key categories we had identified and the corresponding variables are listed below.

COVID-19 US County JHU Data & Demographics
The following variables were extracted from the US COVID-19 Demographics dataframe without any transformation: median_age: Overall median age for the county.
population: Total population for the county. This variable is used to calculate the number of COVID-19 cases per 100k people, not as a separate dependent variable. female_percentage: female / population in percent.

Education
The following variable was extracted from the Education dataframe without any transformation.
Percent.of.adults.with.a.bachelor.s.degree.or.higher: The percentage of adults with a bachelor's degree or higher.

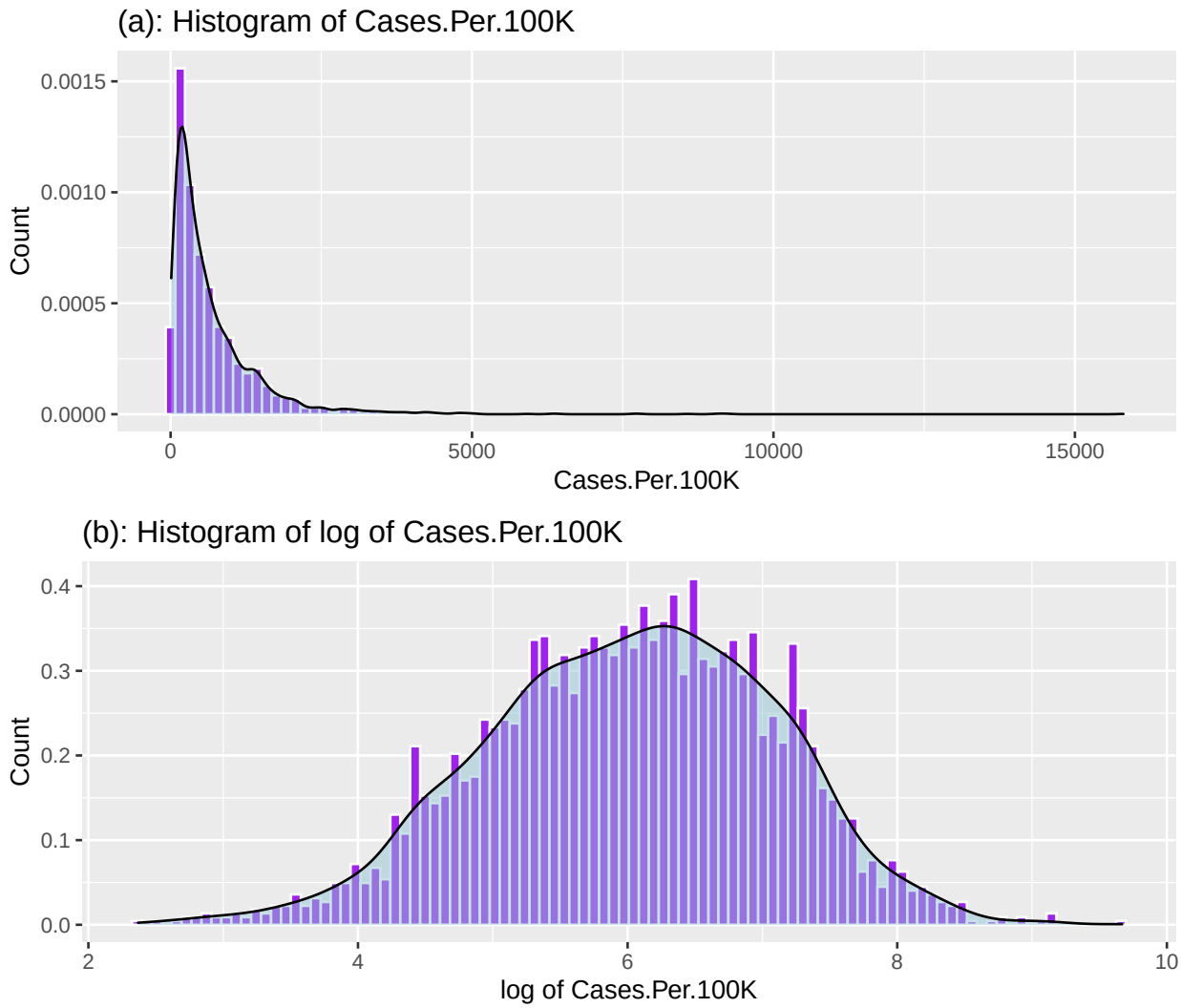

Figure 4: Histogram showing distribution of cases

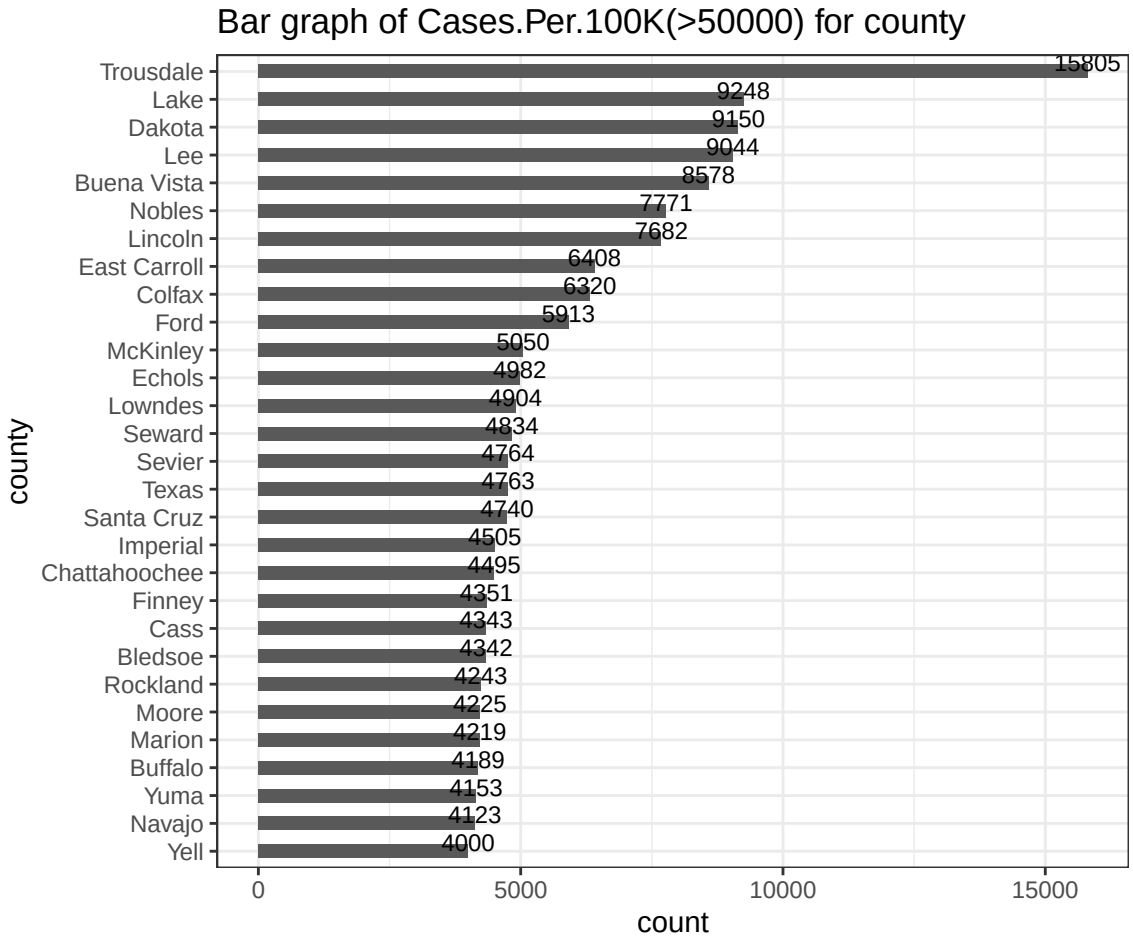

Figure 5: Counties with cases per 100k > 4000

Urban Influence Codes

The following variable was extracted from Urban Influence Codes dataframe without any transformation. We converted the 'code' and binned them into two buckets that are 'Urban' and 'Rural'.

UIC_2013(urban_or_rural): Code (1 & 2) are urban counties and Code(3-12) are rural counties.

Transform Variables

We generated a scatter plot matrices using the 'psych' package in 'R' to visualize the relationship between all numerical variables in the dataset.

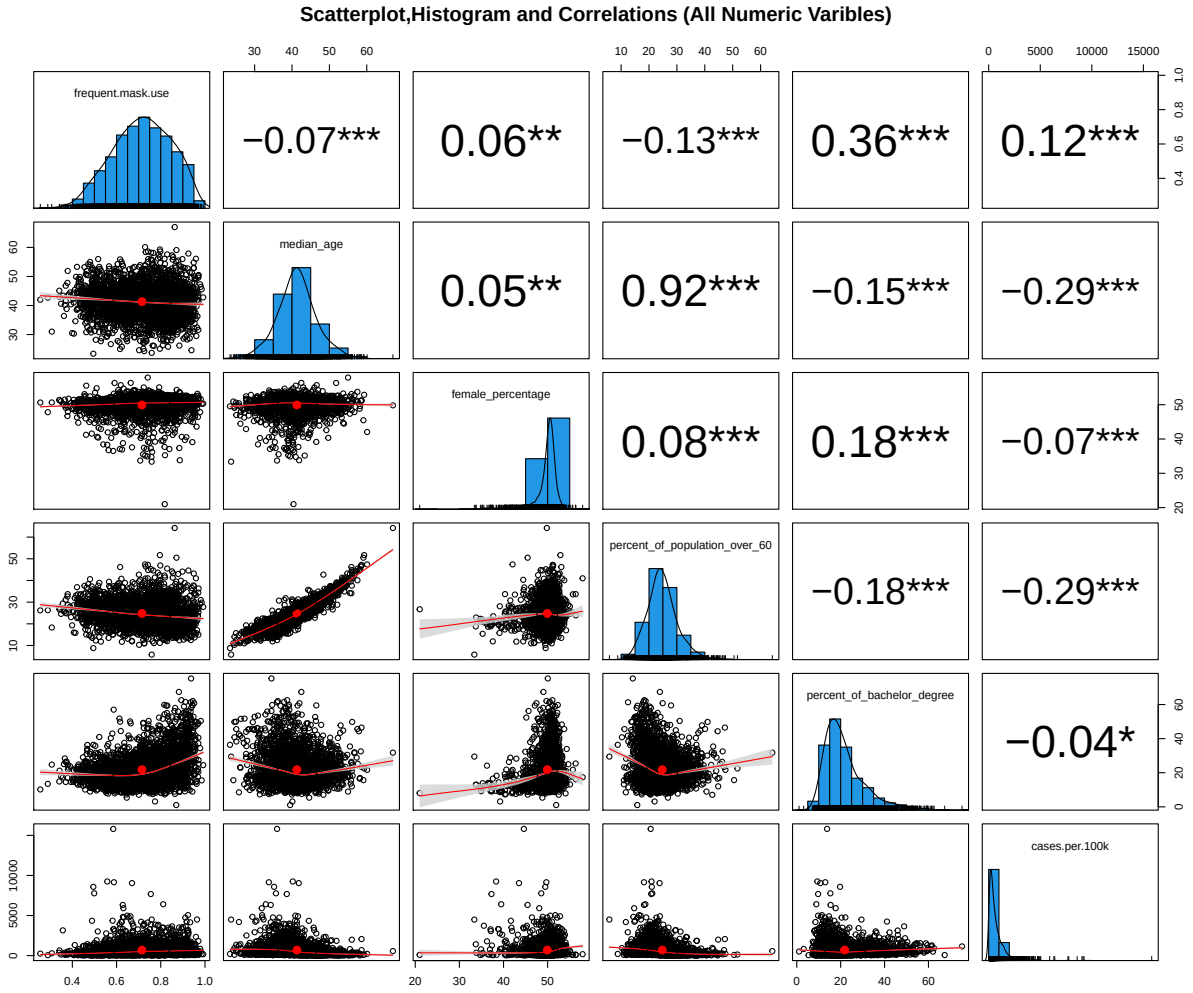

Figure 6: Relationships between Covariates

From the Figure 6, the histograms on the diagonal show the distribution of data to make it easier to detect heavy tails in any distributions. Female_percentage is shewed to the left and we use the square of this variable as a transformation. percent_of_bachelor_degree and Case.per.100K are both right-biased, so sqrt and log downgrade conversions have been performed. These transformations are performed in order to better fir the regression models described subsequently. The scatter matrix chart in Figure 6 shows these variables after transformation.

The upper right corner shows correlation coefficient, we can see the correlation coefficient is 0.92 between

median_age and Percent.of.Population.Aged.60 which indicates that these two variables have a very strong positive correlation. We acknowledge that the high degree of correlation might be an issue for the modeling process, but may still be useful to generate intersting insights.

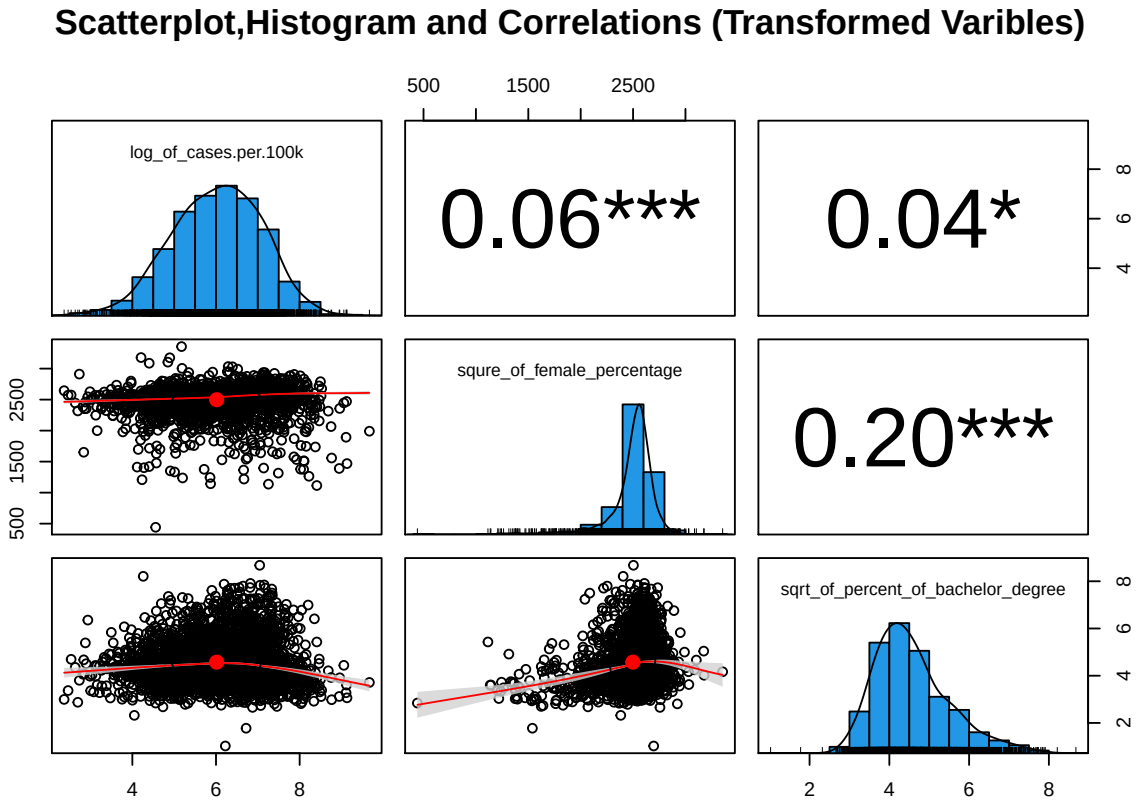

Figure 7: Relationships between transformed Covariates

Model Development

We use OLS regression as a strategy to study the effects of our covariates on the response variable. We will build a total of three models with increasing number of features being included in each subsequent model. For all of our model hypothesis testing we followed standard practice by utilizing a maximum acceptance level with p-value of 0.05 for the statistical significance of the coefficients that are calculated. A regression table is provided following the model section.

Modeling objectives

Our model 1 objective is to observe the association between our primary response variable, namely, COVID cases per hundred thousand residents and the primary explanatory variable, namely, percentage of people who wear masks frequently in county population

Our objective for model 2 is to create a parsimonious model that describes the association between COVID cases per hundred thousand residents and percentage of people who wear masks frequently in county population considering the context of other covariates such as the percentage of county population that is 60 or older, percentage of people with bachelor's degree or higher, and if a county is urban or rural.

Table 1: Summary table of numeric variables

variable_name	variable_decription	min	mean	max
Cases.Per.100K	COVID-19 cases per hundred thousand residents	10.67464	692.73532	15804.86786
frequent.mask.use	The percentage of people who frequently wear masks	0.25500	0.71490	0.99200
median_age	The median age for the county	23.40000	41.31304	67.00000
female_percentage	The female percentage of county population	21.00395	49.92542	57.92269
percent_of_population_over_60	The percentage of county population who is 60 or older	5.80000	24.77243	64.20000
percent_of_bachelor_degree	The percentage of adults with a bachelor's degree or higher	1.04712	21.81167	75.29932

Table 2: Summary table of Categorical variables(Urban/Rural)

Var1	Freq
Rural	1897
Urban	1124

Our objective for model 3 is to create a model with the best predictive power for our primary response variable of COVID cases per hundered thousand residents. Covariates include median_age, female_percentage, percent_of_population_aged_60+, percent_of_bachelor_degree, Urban/Rural.

Model Selection

For the significance level of the hypothesis test, assume $\alpha = 0.05$ in advance. In order to evaluate whether the coefficient of each of our models is significant, decide on the null and alternative hypotheses, where i is the index for each coefficient of the regression equation.

Null hypothesis: $H0 : \beta_i = 0$
Alternative hypothesis: $HA : \beta i \neq 0$

Model 1

Response Variable:
* log of cases per 100K Residents, (log_of_Cases.Per.100K)

Primary Explanatory Variables:
* Percentage of mask use in county, (frequent.mask.use)

Model 1 expression:

$$\text{log of cases per 100,000 Residents} = \beta_0 + \beta_1 * \text{frequent.mask.use} + \epsilon$$

```
##
## Call:
## lm(formula = log_of_cases.per.100k ~ frequent.mask.use, data = covid19)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6643 -0.6903  0.0490  0.7136  3.8877
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.7185     0.1054   44.76   <2e-16 ***
## frequent.mask.use   1.8152     0.1451   12.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.038 on 3019 degrees of freedom
## Multiple R-squared:  0.0493, Adjusted R-squared:  0.04899
## F-statistic: 156.6 on 1 and 3019 DF,  p-value: < 2.2e-16
```

From the result above, we can see that our β_0 and β_1 are far less than the significance level α (0.05), so we reject the null hypothesis and accept the alternative hypothesis, the main explanatory variable frequent.mask.use is statistically significant. Therefore, the expression of model one is :

$$\text{log of cases per 100,000 Residents} = 4.7185 + 1.8152 * \text{frequent.mask.use}$$

For each additional unit of frequent.mask.use, the average value of cases per 100,000 residents will increase by 181.52%. The intercept term is much larger contributor to the outcome even at the maximum possible value of frequent.mask.use of 1. The adjusted R-squared of this model is 0.04899, which means that the model's goodness of fit is very low.

Model 2

Response Variable:
* log of cases per 100K Residents, (log_of_Cases.Per.100K)

Primary Explanatory Variables:
* Percentage of mask use in county, (frequent.mask.use)

Additional Covariate Variables: * Percent of Population Aged 60+. (percent_of_population_over_60)
* Sqrt of Percent of adults with a bachelor's degree or higher, 2015-19.(sqrt_of_percent_of_bachelor_degree)
* Whether the county is a urban or a rural area. (urban_or_rural)

Model two expression:

$$\begin{aligned}\text{log of cases per 100,000 Residents} = \beta_0 + \beta_1 * \text{frequent.mask.use} + \beta_2 * \text{percent.of.population.over.60} \\ + \beta_3 * \sqrt{\text{percent.of.bachelor.degree}} + \beta_4 * \textit{urban.rural} + \epsilon\end{aligned}$$

```
##
## Call:
## lm(formula = log_of_cases.per.100k ~ frequent.mask.use + percent_of_population_over_60 +
##     sqrt_of_percent_of_bachelor_degree + urban_or_rural, data = covid19)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4578 -0.6127  0.0112  0.6526  3.6066
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       7.347965   0.148294  49.550  < 2e-16 ***
## frequent.mask.use                 1.428026   0.146366   9.757  < 2e-16 ***
## percent_of_population_over_60    -0.073155   0.003327 -21.989  < 2e-16 ***
## sqrt_of_percent_of_bachelor_degree -0.136115   0.020487  -6.644 3.61e-11 ***
## urban_or_ruralUrban               0.220917   0.042312   5.221 1.90e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9433 on 3016 degrees of freedom
## Multiple R-squared:  0.2152, Adjusted R-squared:  0.2141
## F-statistic: 206.7 on 4 and 3016 DF,  p-value: < 2.2e-16
```

We can see from the result above, all β_i are far less than the significance level $\alpha(0.05)$, so we reject the null hypothesis and accept the alternative hypothesis, all independent variables are statistically significant. Therefore, the expression of model 2 is :

$$\text{log of cases per 100,000 Residents} = 7.347965 + 1.428026 * \text{frequent.mask.use} - 0.073155 * \text{percent.of.population.over.60} \\ -0.136115 * \sqrt{\text{percent.of.bachelor.degree}} + 0.220917 * Urban$$

```
vif(fit2)
```

```
##                   frequent.mask.use      percent_of_population_over_60
##                            1.231889                            1.127215
## sqrt_of_percent_of_bachelor_degree                       urban_or_rural
##                            1.261362                            1.419925
```

We conducted a Value Inflation Factor (VIF) test, which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model, using the guideline threshold value of 4 indicating strong collinearity between our covariates. The resulting maximum value found was 1.419925 which means that there is no collinearity problem in the variables of our model two. Model 2 was a much better fit for the population as our adjusted R-squared value increased from 0.049 to 0.214. Interestingly, the intercept term has increased significantly between model 1 and model 2. This indicates that inclusion of the additional covariates has resulted in the unconsidered causes absorbing most of the effect from the mask use covariate.

Model 3

Response Variable:
* log of cases per 100K Residents, (log_of_cases.Per.100k)

Primary Explanatory Variables:
* Percentage of mask use in county, (frequent.mask.use)

Additional Covariates Variables: * Overall median age for the county (median_age)
* Square of female / population in percent.(square_of_female_percentage) * Percent of Population Aged 60+. (percent_of_population_over_60)
* Sqrt of Percent of adults with a bachelor's degree or higher, 2015-19.(sqrt_of_percent_of_bachelor_degree)
* Whether the county is a urban or a rural area. (urban_or_rural)

$$\text{log of cases per 100,000 Residents} = \beta_0 + \beta_1 * \text{frequent.mask.use} + \beta_2 * \text{median.age} + \beta_3 * \text{female.percentage}^2 \\ +\beta_4 * \text{percent.of.population.over.60} + \beta_3 * \sqrt{\text{percent.of.bachelor.degree}} + \beta_4 * urban.rural + \epsilon$$

```
##
## Call:
## lm(formula = log_of_cases.per.100k ~ frequent.mask.use + median_age +
##     squre_of_female_percentage + percent_of_population_over_60 +
##     sqrt_of_percent_of_bachelor_degree + urban_or_rural, data = covid19)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2723 -0.6258  0.0106  0.6426  3.5272
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     7.215e+00  2.899e-01  24.887  < 2e-16 ***
## frequent.mask.use               1.526e+00  1.458e-01  10.463  < 2e-16 ***
```

```
## median_age                          -4.326e-02  8.745e-03  -4.947 7.95e-07 ***
## squre_of_female_percentage           4.160e-04  8.226e-05   5.057 4.52e-07 ***
## percent_of_population_over_60       -3.588e-02  8.781e-03  -4.087 4.49e-05 ***
## sqrt_of_percent_of_bachelor_degree -1.615e-01  2.061e-02  -7.835 6.44e-15 ***
## urban_or_ruralUrban                  2.312e-01  4.348e-02   5.317 1.13e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.935 on 3014 degrees of freedom
## Multiple R-squared:  0.2294, Adjusted R-squared:  0.2279
## F-statistic: 149.6 on 6 and 3014 DF,  p-value: < 2.2e-16
```

It can be seen that all β_i are far less than the significance level $\alpha(0.05)$, so we reject the null hypothesis and accept the alternative hypothesis, all independent variables are significant. This left us with a final model of: :

$$\text{log of cases per 100,000 Residents} = 7.215 + 1.52 * \text{frequent.mask.use} - 0.043 * \text{median.age} + 0.0004 * \text{female.percentage}^2 - 0.036 * \text{percent.of.population.over.60} - 0.162 * \sqrt{\text{percent.of.bachelor.degree}} + 0.231 * Urban$$

```
vif(fit3)
```

```
##                  frequent.mask.use                          median_age
##                           1.244516                            7.406259
##         squre_of_female_percentage      percent_of_population_over_60
##                           1.094182                            7.992489
## sqrt_of_percent_of_bachelor_degree                      urban_or_rural
##                           1.299499                            1.525851
```

We also conducted a value inflation factor test with the guideline that a VIF of 4 or larger represents strong collinearity. From the result above, we can see that the values of median_age and percent_of_population_over_60 are both greater than 4, which means that there is a collinearity and the variables(median_age and percent_of_population_over_60) for model 3 are collinear.

Model Regression Table

```
##
## Model comparison
## ================================================================================================
##                                                        Dependent variable:
##                                                ----------------------------------------------
##                                                    Reported COVID-19 Cases per 100K Residents
##                                                      (1)               (2)               (3)
## ------------------------------------------------------------------------------------------------
## Percentage of Residents Wear Masks               1.815***          1.428***          1.526***
##                                                   (0.145)           (0.146)           (0.146)
## Median Age for County                                                                -0.043***
##                                                                                       (0.009)
## Square of Female Percentage                                                          0.0004***
##                                                                                      (0.0001)
## Percentage of Residents 60+                                        -0.073***         -0.036***
##                                                                     (0.003)           (0.009)
## Square Root of Percentage of High Degree                           -0.136***         -0.161***
##                                                                     (0.020)           (0.021)
## County (Urban)                                                     0.221***          0.231***
```

```
##                                                          (0.042)          (0.043)
## Intercept                                4.718***        7.348***         7.215***
##                                          (0.105)         (0.148)          (0.290)
## ------------------------------------------------------------------------------------
## Observations                              3,021           3,021            3,021
## R2                                        0.049           0.215            0.229
## Adjusted R2                               0.049           0.214            0.228
## Residual Std. Error                 1.038 (df = 3019) 0.943 (df = 3016) 0.935 (df = 3014)
## ====================================================================================
## Note:                                                    *p<0.1; **p<0.05; ***p<0.01
```

Discussion on result

1. Model 2 is a much better fit than model one as the adjusted R-squared value increased from 0.049 to 0.214. Model 3 is the best fit as it has the largest adjusted R-squared value which is 0.228.
2. All coefficients show statistical significance because of sample size, but the treatment of practical significance needs to be much more complicated.
3. The coefficient of mask use is positive, meaning that counties with more mask use have more cases. This is counter intuitive, but since most counties have very few cases, it can almost be said that the pandemic does not exist in those counties at all. Hence, mask use behavios may not be making any difference in the spread of Covid-19 and the coefficient that is calculated does not capture the phenomenon we initially planned.
4. Urban/rural has a coefficient of .23, meaning that a urban county is expected to have ~20% higher cases than an identical rural county. This makes sense because the population density in urban counties is higher and covid-19 was most prevelant in big cities during the date under consideration.
5. Comparing the coefficient of 'percent.of.population.over.60+' in model 2 (-0.073) and coefficient of 'percent.of.population.over.60+' & median age in model 3 (-0.036 & - 0.043). It almost looks like -0.073 = -0.036 + (-0.043). It seems as if the effect is split into these two variables which is expected because the two are highly correlated. It is almost like they are the same variable.
6. The coefficient of Education(percent.of.bachelor.degree) is negative(-0.161). This could mean that people with higher education take precautions and stay safe from the virus. It could also mean that they have jobs where they can work from home, and so are less affected by the virus.
7. The intercept term is very large(7.215), so it looks like none of the variables we have considered are strong causes. Other variables we have not considered (epsilon) are more important causes than any we have considered.

Classical Linear Model (CLM) Assumptions

Below is the discussion of CLM assumptions, along with the analysis of how our models fit in each assumption and several limitations.

1. Identically Distributed Population (IID)

Different data draw and report standard and process - In order to meet IID, ideally the process of COVID data collected from each county should be the same. However among 3000 counties in the US, county-level COVID cases data is reported independently by each county, they can be submitted per different county-level processes and standard.

To meet with IID, each county should start counting COVID cases at the same time. However, this is not the case in reality. For example, for counties that actively comply with pandemic mitigation policies at the beginning of the pandemic, more COVID cases can be captured and reported than those counties responded to COVID slowly. Therefore, there is a different level of accuracy across individual county reported COVID data. This also results in the data deviated from IID assumption.

In addition, Covid-19 disease spread depends on the geographical location of each county with cases spreading more easily across neighboring counties. Hence, the data from one county is dependent on the data from other counties.

2. Linear Conditional Expectation

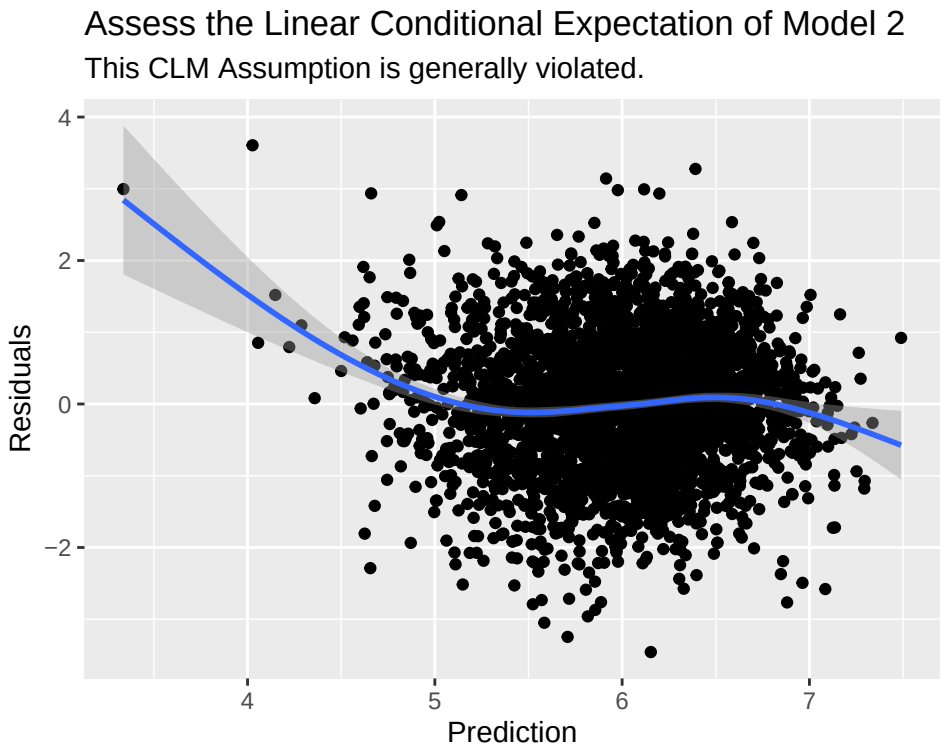

Figure 8: Residuals vs predictions Plot

From the Figure 8, we do not see a flat smoothing line, which indicates that there is no a clear linear relationship in this data. This CLM assumption is generally violated. It can be found in the figure that on the left hand side there is an obvious decreasing residuals before it reaches to a relative flat curve range. An non-constant variance is observed. As a result, the estimated standard errors can potentially be incorrect. Therefore, the confidence intervals and hypothesis tests generated may not be completely reliable. However, within a certain range, the blue curve is relatively flat, where we can still reasonably believe that the model could produce relatively meaningful analysis for exploratory purpose.

3. No Perfect Collinearity

From running the coeftest, R doesn't drop any variables in our model 2. It confirms that this CLM assumption is met. Additionally, we run the test vif(model_2), it simply examines the variance inflation factor (VIF) for each coefficient. This indicates that the standard errors for each variables are relatively similar, for example, one variable standard error (SE) is not few times higher than the SE of other variables. No obvious variance inflation among those variables, which would prevent our mode to produce a precise measurement wanted from the key variable of interest - mask-use.

```
##
## t test of coefficients:
##
##                                    Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)                       7.3479652  0.1482944  49.5498 < 2.2e-16 ***
```

```
## frequent.mask.use                    1.4280262  0.1463663   9.7565 < 2.2e-16 ***
## percent_of_population_over_60      -0.0731549  0.0033269 -21.9888 < 2.2e-16 ***
## sqrt_of_percent_of_bachelor_degree -0.1361148  0.0204873  -6.6439 3.612e-11 ***
## urban_or_ruralUrban                 0.2209175  0.0423118   5.2212 1.899e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                  frequent.mask.use      percent_of_population_over_60
##                           1.231889                           1.127215
## sqrt_of_percent_of_bachelor_degree                     urban_or_rural
##                           1.261362                           1.419925
```

4. Homoskedastic Errors

Looking back to the residual and prediction plot above in section "Linear Conditional Expectation", those observations (dots) do not create a constant shape along the line, which indicates that the variance is not constant. This CLP assumption is violated.

5. Normally Distributed Errors

Our team decides to use histogram and Q-Q Plot to assess this CLM assumption. By observing the histogram of model 2 residuals (every observation in the dataset is associated with a true value of the Y and the fitted value of the Y, the residuals are the difference between them), the histogram in Figure 9 clearly shows a bell shaped curve.

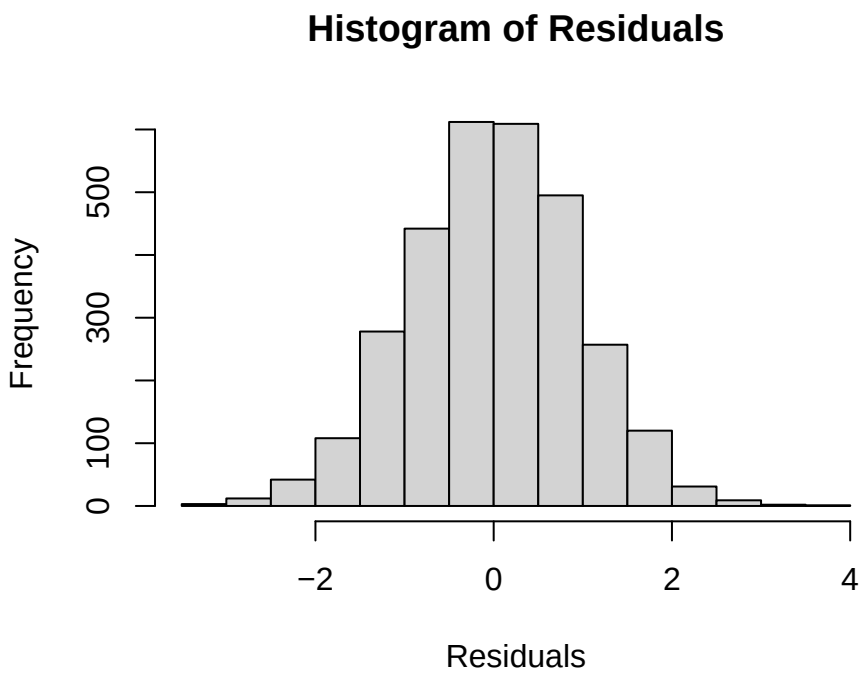

Figure 9: Distribution of Residuals

Additionally, to double confirm, we run a Q-Q Plot as shown in Figure 10. It looks at the theoretical quantile in a normal curve. By observing the qqplot result, the data is mostly aligned on the X=Y straight-line, data

is slightly deviated from that line on bottom left and top right. Based on those observations we believe this assumption is met.

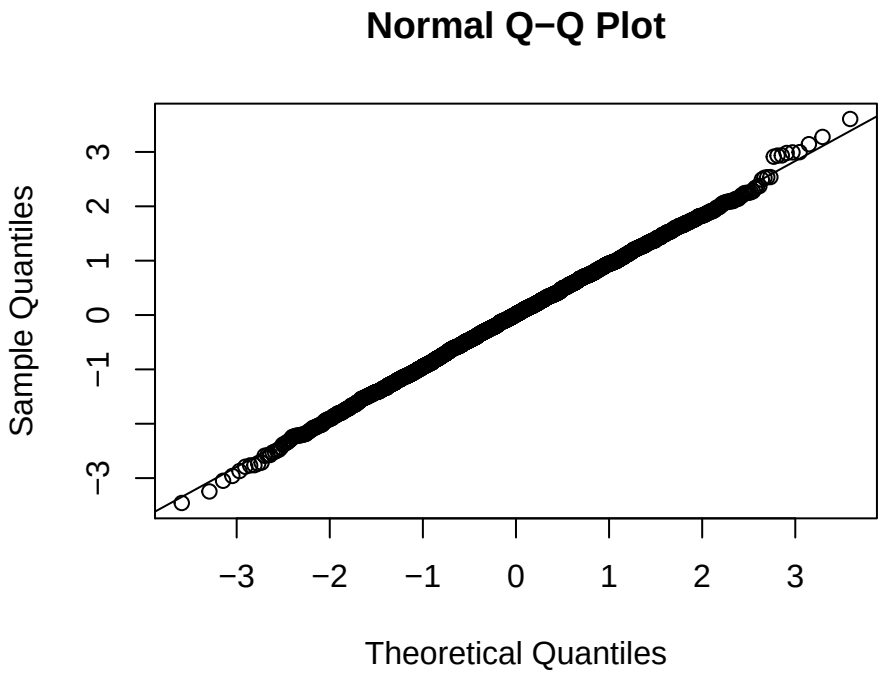

Figure 10: Residuals vs predictions Plot

Limitations

In addition to one limitation specified in IID section, below is a further discussion of limitations in our models.

Potential reverse causality between mask-use and COVID cases - Naturally we assume more mask-use will cause less COVID cases in a county, however in reverse, a county with more COVID cases may cause people to wear mask more frequently. We do not believe our models are sophisticated enough to test this complex causal relationship.

Timing of data collection - Those historical data captured at the beginning of the pandemic, is not captured upon the same standardized processes between counties. By assessing the pandemic roadmap, the CDC Confirmed the First US Coronavirus Case on January 21, however till March 13th Trump Administration Declared COVID-19 a National Emergency. During this gap, there was no a standardized official requirement to each county in terms of how to calculate and submit COVID cases, and when to submit etc. This impacts data accuracy and IID assumption.

It is reasonable to believe there are missing COVID cases from certain counties during this time window. For counties that responded quickly at the beginning of the pandemic, there could be more COVID cases captured. Those records were accumulated as part of the total covid cases count; those counties who actively report covid cases from the early beginning will end up as a high case county. This limitation impacts the population from each county and the standard error result, thus making our models less sensitive to the input variables and introducing potential inaccuracy in our model tests.

Bias that generated from political view influences - Even after the national emergency is issued, counties' responses to pandemic and people's mask-use behaviors can be greatly impacted by political views.

Especially, the data is collected in 2020 - the presidential election year, when different parties frequently expressed their opposit opinions on the pandemic crisis thus creating impact to people's mask-use behaviors. For example, influenced by Trump's public views which underestimated the pandemic severity, a republican county may not respond to covid related policies actively, the cases captured in the county could be less accurate than a Democrats County. And people's mask use behavior also is greatly impacted by President Trump's public speeches and behaviors to Republican counties.

Limited tests done One additional limitation is that, we only looked at COVID cases count reported on 7-14-2020 for simplicity purpose for this particular lab. However, ideally we can test our models upon COVID cases count data on few other days in case there is any outlier or data quality from certain day's COVID data.

Further discussion of limitations

Even with the limitations above, our team still believe there are sufficient justifications to use the County-Level Covid Cases as our observation variable as follows.

County-level observation is relatively the most precise method to assess COVID data County is a political subdivision of a state with governmental authority. How an individual or community response to the pandemic is primarily based on county issued policy. For example, even with nearby counties, per their transmission level differences, the mask-use policy and people's mask-use behavior can be completely different. Therefore, county-level data instead of state-level data, can create more precision and effectiveness in our models, especially to hour model 2 – that includes mask-use behavior which is dramatically influenced by county-level policy. County-level observation is the smallest breakdown observation unit team believes can best fit in the IID assumption. County consists of "a geographic region with specific boundaries" 2, therefore the observations in data are exclusive to each other, and at the same time large enough to be regarded as a large random sample, generally fit in the "independent draw" requirement of the IID assumption while providing large enough sample in order to conduct meaningful analysis.

County-level observation empowers our models In our model 2, we added two variables – a county is rural/suburb and county-level college degree rate. Geographically, whether a location is considered a rural or suburb, can only be precisely defined at county-level instead of other broader scale such as state-level. Same state could have only very few counties that are extremely urbanized while most other counties are rural places. Same as the college degree rate by county, since in one state there could be certain very well-educated counties versus other counties. Therefore, county-level observations successfully justified our model 2 in order to answer the research question.

Based on the limitations and justifications above, team still decide to choose the models and variables proposed with limitations taken consideration.

Omitted Variables Discussion

The primary research question of this analysis is to understand the relationship between mask-use with covid cases count at county-level. Naturally people may believe mask-use will reduce COVID cases, however, instead of mask-use being a cause, it could be a result from a county COVID cases count high as more COVID cases may make people want to use masks more. This implies a reverse causality relationship may exist. And there are a variety of variables that have correlations with both the mask-use variable and covid cases count. Our team does not believe those three simple linear models are sophisticated enough to present such complex causal relationships between mask-use and covid cases count. Based on this consideration above, the following variables are the omitted variables we wish to include but did not have in our models.

2020 annual average COVID transmission rate of each county – This could be a key variable that blocks out the bi-direction between mask-use and covid cases. If we can add transmission rate as a control, the model result could be more sensitive to the input variable – mask-use. The Beta1 in our models can be more powerful for achieving a meaningful result.

Political Views – We argue that whether a county is republican or demoncrat will impact both mask use

and COVID cases. In year 2020 when president election occurs, both parties have sent clear message to the public how they would like people to response to the pandemic. Demoncrats have been urging people to stay home and wear masks; whereas the Republican Administration has underestimated the pandemic and delayed pandemic mitigation approaches. Hence, we argue that a Republican leaning county would have lower mask use compared to a Democrat leaning county. Also, since Republican leaning counties impose fewer restrictions on businesses such as restaurants, Covid-19 would spread more readily in these counties and hence, increase the Covid-19 cases. If we were to use an indicator variable that took a value of 0 for a Democratic leaning county and a value of 1 for a Republican leaning county, an increse in the indicator variable would correspond to a decrease in mask use and an increase in Covid-19 cases. In this case, the omitted variable bias would be negative. Since, the estimate coefficient for mask use is positive, our estimate are biased towards zero, relative to the true coefficient. This suggests that our estimate may have lost statistical significance if this omitted variable was included.

Monthly average temperature of each county – This is a variable that the team believes could impact both the input variable mask-use and the outcome variable – covid cases. For example, if a county average temperature in July 2020 (the mask-use data collection time) is greater than 85 F, because of the hot and/or humid feeling, people may choose to not wear masks regardless of other analysis factors. On the contrary, if a county is cold, people may feel more comfortable wearing a mask. And "Influenza virus is more likely to be transmitted during winter on the way to the subway than in a warm room", [9]. Therefore, if the temperature is low, it will result in higher transmission rate thus leading to more covid cases. In this case, the Omitted Variable bias is positive as increase in temperature reduces mask use and also reduces covid-19 cases. Since, the estimated coefficient of mask use in our model is positive, our estimate is biased away from zero, relative to the true coefficient. This suggests that our results will not lose statistical significance due to the omission of the temperature variable.

Conclusion

Based on the discussion presented in this paper, we have attempted to answer the research question regarding the relationship between mask use and Covid-19 cases. The results from our regression analysis have failed to prove the expected relationship between these variables There are numerous limitations in the data used for this analysis that contribute to the uncertainty. This analysis underscores the importance of using appropriate data for answering complex research questions. In addition, limited tools of OLS regression are not sufficient for the treatment of variables that we had planned. Despite these limitations, we believe that this paper was a productive exercise in understanding the pitfalls that may occur during the process of gleaning insights from data and provides numerous considerations for future work. Specifically, a better treatment of the outcome variable was warranted. A change in number of Covid cases between two time points may have been more appropriate. Additionally, we may need to add variable to handle the large differences in case counts in any county. There could also be more appropriate covariates that could have been used that have a more direct effect on Covid cases and mask use behavior. For example, state policy on Covid restrictions could be one such variable.

[9] "Study shows why the flu likes winter"_ - The New York Times, https://www.nytimes.com/2007/12/05/health/05iht-05flu.8591550.html