Lab 2 Second Draft

Lucas Brossi, Amar Chatterjee, Daniel Chow, Sandip Panesar

11/30/2020

Introduction

As of October 2020, more than 10 million Americans have been infected with the novel coronavirus, of which more than 240,000 have perished. Governments from the local to state to federal level have scrambled to enact policies to regain some semblance of control. Despite all of their efforts, the United States currently leads the globe in both the number of cases and the number of deaths by a long shot.

One of the earliest recommendations by health officials to help protect against contracting the virus was to don Personal Protective Equipment (PPE), more specifically face masks. However, not all face masks are created equally. The highly effective N-95 face masks (which filter 95% of airborne particles) were scarcely available and rightfully reserved primarily for frontline healthcare workers, resulting in a boom in production of the next best public alternative: cloth face mask coverings. While not as effective as the medical-grade N-95, when combined with social distancing cloth face masks were said to drastically reduce the risk of the virus spreading. In the absence of sophisticated testing, containment, and contact tracing techniques, the adoption of face masks in the United States became an essential strategic component in the COVID-19 containment efforts.

As shown in the below diagram, the biggest beneficiaries of wearing a mask are actually other people. While it is hard to quantify the exact efficacy, wearing a mask aids considerably in reducing the spread of airborne particles of the mask-wearer. Given that many COVID-19 carriers remain asymptomatic for at least some period of time, the messaging from health officials centered around a moral and social obligation to help contain the virus spread.

On April 3, 2020, the Center for Disease Control (CDC) issued an official recommendation advising all persons to wear a cloth face mask or covering in public to help slow the spread of the coronavirus. Following this guidance, almost every state went on to enact a policy requiring people to wear face masks at all times in public settings. In fact, only 7 states to date have proceeded with no such policy (although many have since ended their order).

Despite all of these recommendations, the use of face masks has become politicized and undermined by large swaths of the country's population. Conflicting messaging from government officials, including the President himself, has resulted in a loss of credibility and trust in the CDC. Whether as a result of denial, distrust, or a desire to feel in control, the fact remains that tens of millions of Americans would rather take the risk over wearing a face mask in public. And for all we know, they could be justified in doing so!

Accordingly, as a team we decided to leverage the provided dataset to validate the guidance from the CDC and answer the following question:

Does the implementation of a mandatory face mask policy for all individuals aid in reducing the case rate of COVID-19 in the United States?

We hypothesize that face masks do indeed have a measurable & causal impact on containing the spread of COVID-19, even when taking into account socioeconomic, demographic, alternate government policies, and other potential competing factors. Our measurement goal is to assess the statistical significance and practical significance of mandatory face mask usage policies on reducing the COVID-19 case rate in the United States. Over the course of this report, we will include other covariates in our regression modeling which we deem to be important in reducing the COVID-19 case rate in an effort to isolate the portion of variability actually explained by the implementation of a mandatory face mask policy for all individuals in the United States.

These other covariates, while important, will help absorb some of the "noise" not associated specifically with the implementation of a face mask policy.

Data

The data used in the model is taken from the provided covid_19 dataset. This dataset comes from a very large collection of sources and collected using various methods. The dataset is up-to-date as of October 30th, 2020. Additionally, the covid_19 dataset uses the Google Human mobility metrics. Google Human mobility data is compiled daily by Google and includes information on the amount of time spent at various public locations compared to Google's baseline data. Some of this data is included in the dataset and assumed that it was taken the same day the rest of the dataset was compiled on October 30th, 2020. Below are the adjustments made to variables that were either created or supplemented.

There are a total of 6 data types in the dataset: character, numeric, integer, factor, dates, and logical. Any variable with "date" in the name is read in as a date. Logical variables include mask_use, mask_legal, and maskbus_use. Variables read in as factors include gov_party, and tests_positive. The only character variable is the state name. Finally, all other variables are read in as either numeric or an integer. All numeric and integer values are real and nonnegative.

talk about data sources, talk about inclusion of data points - citation from Amar https://www.nbcnews.com/health/health-news/here-are-stay-home-orders-across-country-n1168736

Variable Operationalization

Mask Use

This binary/logical variable was created by assigning a 1 if the state had a mask mandate and 0 if it did not (based on the mask mandate date column).

Percent Age Below 25

This column was created by combining the 0-18 and 18-25 age groups. No other adjustments were made.

Percent Age Above 55

This column was created by combining the 55-64 and 65+ age groups. No other adjustments were made.

Days in Shelter-in-Place

The number of days each state was under the Shelter-in-Place mandate. This data was missing some data and supplemented by researching and populating the missing data. The column was creating by subtracting the end and start dates.

Days Businesses Closed

The number of days each state closed non-essential businesses. Similar to the days in SIP, missing data was populating through research of the state's specific mandates, then calculated by subtracting the end and start dates.

Percentage of Population: Black

Observations that were marked as "< 0.01" were rounded down to 0 so that they could be treated as numeric values.

Model 1

Objective

Model 1 is our simplest model. It aims to measure the strength of the relationship between the presence of mandatory mask use policy for all individuals and the COVID-19 case rate per 100k in US states. It has no

other covariate, with the exception of test rate, which we included as a way to control for the impact that different test availabilities might have on the reported case rate by state.

```
df_mod1 <- df %>%
  select(case_rate, mask_use, test_rate)
```

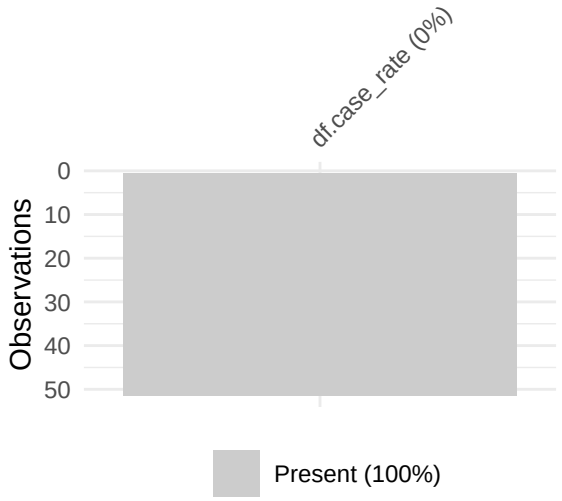
Exploratory Data Analysis

First, we assessed the data provided for the variables of interest. This specifically pertained to any spurious values, and whether there was any missing data. We discuss the distribution of the data in dedicated sections below.

Case Rate

For the main outcome variable, we can see that there are no missing values in the columns of interest:

```
vis_miss(data.frame(df$case_rate))
```



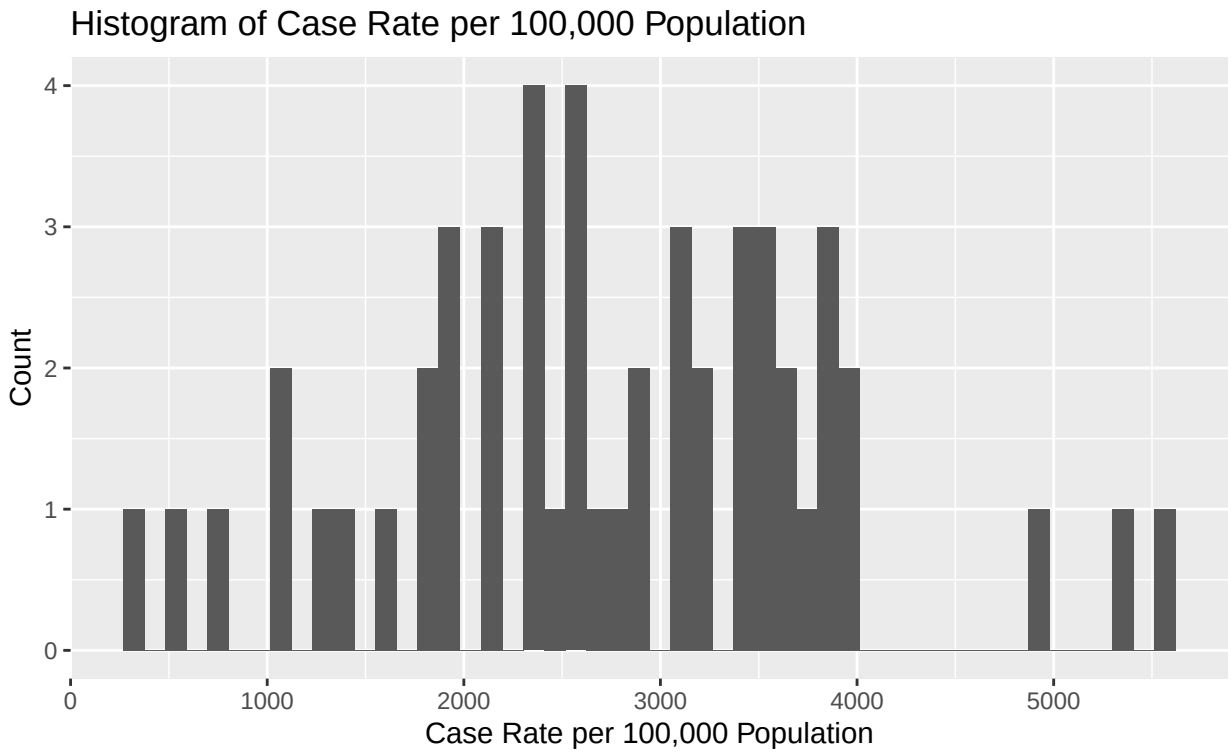
We can then analyze the summary statistics for the case rate from which we can see there are no negative or observably spurious values:

```
summary(df_mod1$case_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     344    2040    2633    2749    3516    5589
```

Together with the summary statistics presented in the previous section, we can further isolate the case rate variable and visualize its distribution:

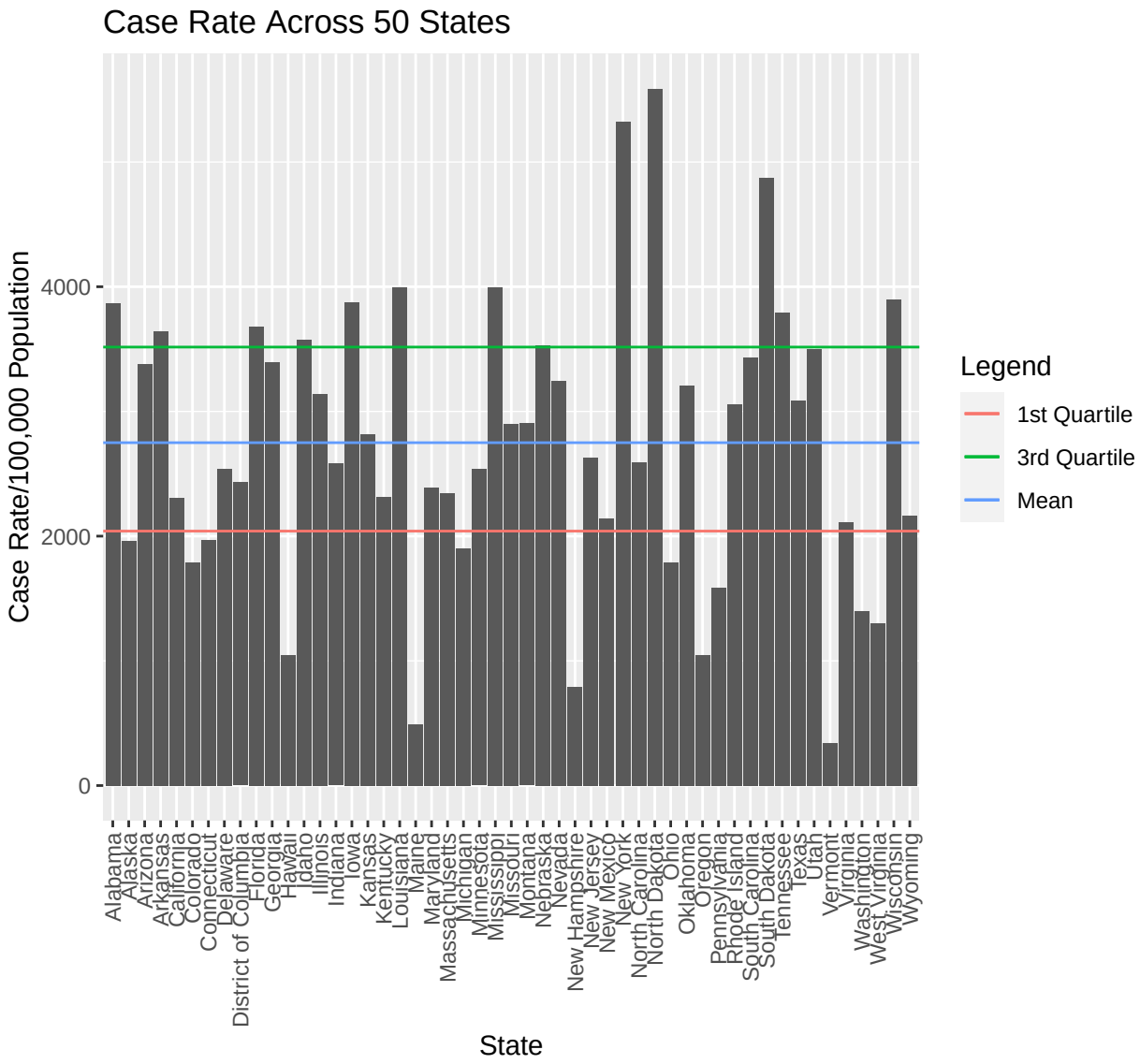
```
hist_case_rate <- df %>%
  ggplot(aes(x = case_rate)) +
  geom_histogram(bins = 50) +
  labs(title = "Histogram of Case Rate per 100,000 Population",
       y = "Count",
       x = "Case Rate per 100,000 Population")

hist_case_rate
```



From the above histogram, we can see that the case rate is approximately normally distributed, though there are some States that are obvious outliers, with very high and very low case rates per 100,000.

Next, we can visually assess the case rate across the 50 States by plotting a bar graph:

```
hist_state_rate <- df %>%
  ggplot(aes(x = state, y = case_rate)) +
  geom_bar(stat="identity") +
  labs(title = "Case Rate Across 50 States",
       y = "Case Rate/100,000 Population",
       x = "State",
       color = "Legend") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  geom_hline(aes(yintercept=mean(df$case_rate), color='Mean')) +
  geom_hline(aes(yintercept=quantile(df$case_rate, 0.25), color='1st Quartile')) +
  geom_hline(aes(yintercept=quantile(df$case_rate, 0.75), color='3rd Quartile'))

hist_state_rate
```



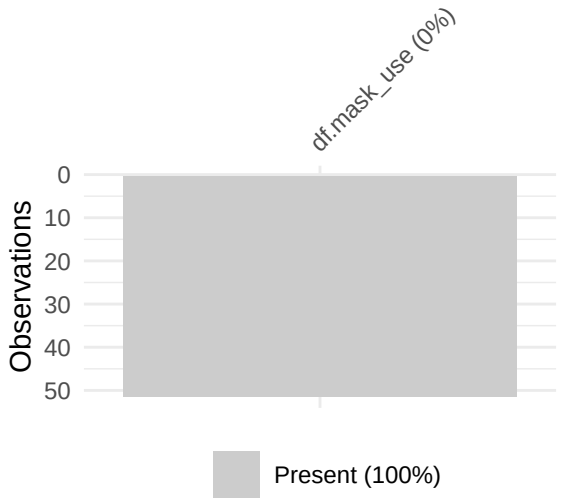
We can see that there are a substantial number of States with case rates below the mean (2749), and a few below the 1st (2040) and 3rd (3516) quartiles.

Next, we look at the correlation between case rate and our causal variables of interest.

Case Rate vs. Mandatory Mask Policy

Before we conduct a comparison between the dependent and main independent variable, we must ensure that there are no spurious or missing data in the mandatory mask policy category. As we can see from below, there is no missing data for this variable:

```
vis_miss(data.frame(df$mask_use))
```



As the mask use policy variable is binary, we only need to ensure that there are 51 observations (51 as the District of Columbia is included):

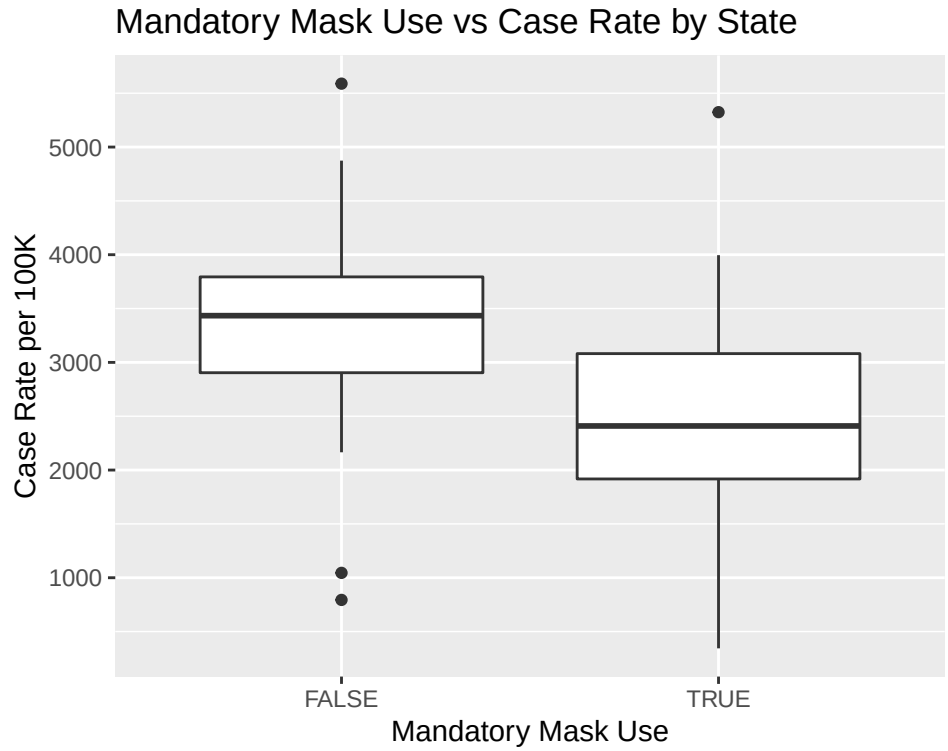
```
summary(df_mod1$mask_use)
```

```
##    Mode   FALSE    TRUE
## logical      17      34
```

We can also see that only 1/3 of all the States surveyed did not implement a mask use policy.

Based on our hypothesis, implementation of a Statewide mask-use policy should lead to a decrease in the number of COVID cases by preventing their spread. From the box plot comparison below we can visually discern an apparent positive correlation between mask use policy categories and case rates. This broadly falls in line with our hypothesis, as we can see that the mean case rate in the 'True' (i.e. enforced mask policy) category (2473/100,000) is substantially lower than that of the 'False' (i.e. no enforced mask policy) category (3303/100,000):

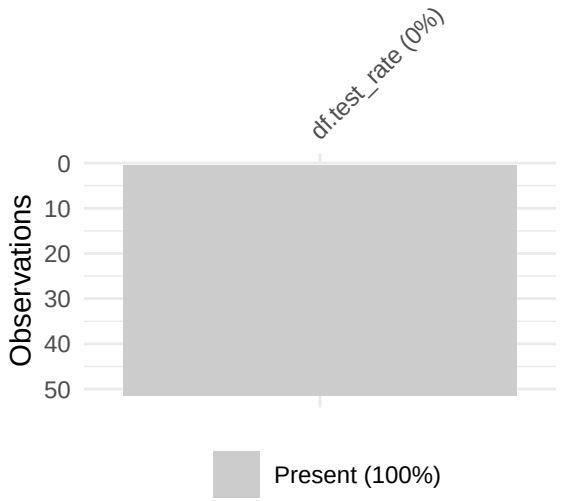
```
df_mod1 %>%
  ggplot(aes(y = case_rate, x = mask_use)) +
  geom_boxplot() +
  labs(
    title = "Mandatory Mask Use vs Case Rate by State",
    x = "Mandatory Mask Use",
    y = "Case Rate per 100K"
  )
```



Case Rate vs. Test Rate

First, we assess for any potential missing data for the test rate variable. As we can see from below, there is none:

```
vis_miss(data.frame(df$test_rate))
```



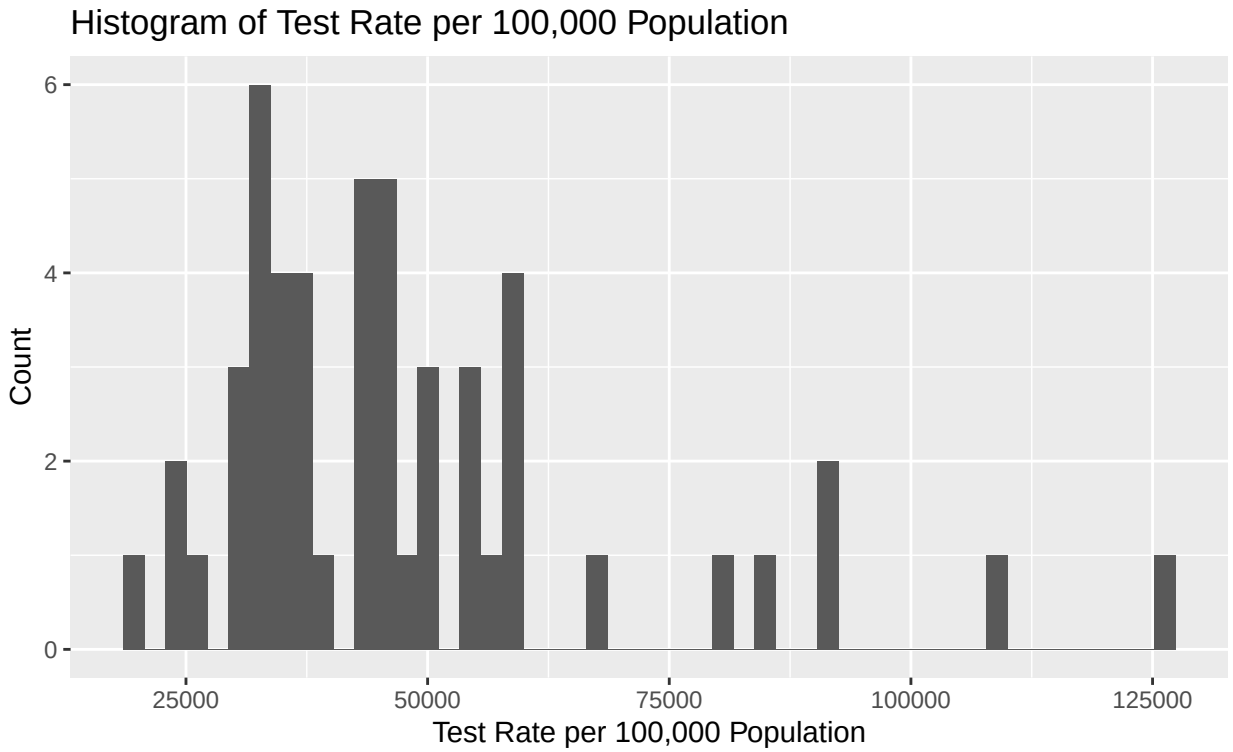
Next, we ensure there are no spurious values by calculating summary statistics:

```
summary(df_mod1$test_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19206   33940   43413   48074   54302  125894
```

Together with the summary statistics presented previously, we can visualize the distribution of test rates using a histogram:

```
hist_test_rate <- df %>%
  ggplot(aes(x = test_rate)) +
  geom_histogram(bins = 50) +
  labs(title = "Histogram of Test Rate per 100,000 Population",
       y = "Count",
       x = "Test Rate per 100,000 Population")

hist_test_rate
```



From the above histogram, we can see that there is a definite positive skew of the data, meaning that the "worst performing" States had test rates were approximately 1/3 of the "top performing" States in terms of tests administered per 100,000 of population. This is an interesting observation, and warrants a comparison between the test rates and case rates to see if there is any potential relationship, which we will next perform.

An important control variable in our causal models is the test rate per 100,000 of the population: We hypothesize that the more tests are performed, the higher the overall case rate. This is a contentious issue due to its potential for becoming politicized and utilized as a means to explain a lack of governmental control of COVID-19. Nevertheless, some suggest that the assumption that more tests results in a higher number of cases is not as simple as a simple correlation. As such we believe that including the test rate per 100,000 is an important control variable in our model.

https://www.politico.eu/article/does-more-coronavirus-testing-mean-more-cases/

```
df_mod1 %>%
  ggplot(aes(y = case_rate, x = test_rate)) +
  geom_point() +
  geom_smooth(method = "lm", level = 0) +
  labs(
```

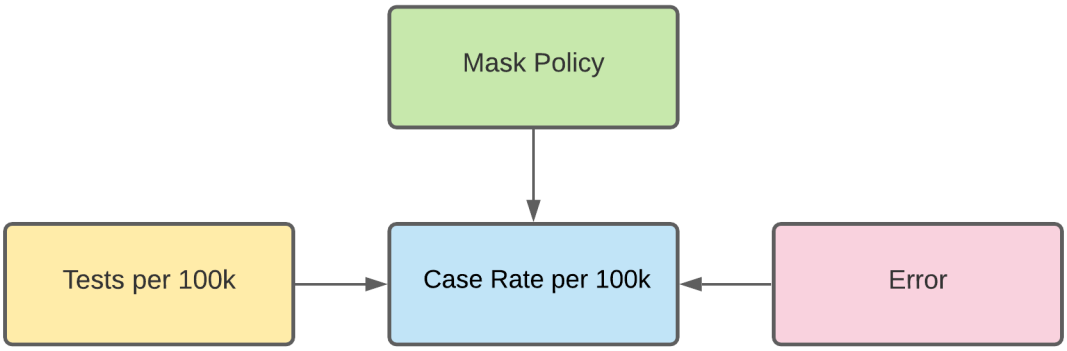
```
    title = "Case Rate vs Test Rate by State",
    x = "Test Rate per 100k",
    y = "Case Rate per 100K"
  )
```

```
## `geom_smooth()` using formula 'y ~ x'
```

The above diagram visualizes the relationship between test and case rates per 100,000 population (below). In line with the more simple hypothesis, there is a slight positive relationship (as evidenced by the slope of the line) between the test rate and case rates. Due to the prior reasoning and demonstrable positive relationship between the variables, we choose to include it in our model.

Casual Diagram for Model 1

We found that our initial assumptions have been validated by a cursory analysis and are confident that the causal diagram below holds true. Colored in blue at the center is the main dependent variable, case rate per 100k. In green, pointing to the case rate (signifying a causal effect on the dependent variable) is the main independent variable, a mask use policy. In yellow is the test rate per 100k as the first control variable. Finally, in red, is the error term that contains all other variables.



Model specification

Our first regression model has **COVID-19 Case Rate per 100,000 Population** as the outcome variable and two covariates: our variable of interest (**Mandatory Mask Use**) and **Test Rate per 100,000 Population**.

Test Rate has been included because there is an evident positive relationship between test rates and case rates. Subsequently, it should be included as a control variable in our subsequent models in order to prevent potential misattribution of effects to other included variables.

Model 1 is subsequently defined as:

$$\text{case rate per } 100,000 \text{ pop.} = \beta_0 + \beta_1(\text{mandatory mask use policy}) + \beta_2(\text{test rate per } 100,000 \text{ pop.})$$

Model summary

```
model_1 <- lm(case_rate ~ mask_use + test_rate, data = df)
std_errors = sqrt(diag(vcovHC(model_1)))
stargazer(model_1, se = std_errors, type = "latex", title = "Model 1 Summary")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Dec 04, 2020 - 07:32:24 AM

Overall model significance (F-test)

This compares the null hypothesis, where $H_0 : \beta_1 = \beta_2 = 0$, against an alternative hypothesis where $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$ at a significance level of 0.05:

```
model_0 <- lm(case_rate ~ 1, data = df)
anova(model_0, model_1, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: case_rate ~ 1
## Model 2: case_rate ~ mask_use + test_rate
```

Table 1: Model 1 Summary

	Dependent variable:
	case_rate
mask_use	−990.470
test_rate	0.018
Constant	2,530.239***
	(501.044)
Observations	51
R^2	0.236
Adjusted R^2	0.204
Residual Std. Error	1,013.835 (df = 48)
F Statistic	7.416*** (df = 2; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
##   Res.Df      RSS Df Sum of Sq     F   Pr(>F)
## 1     50 64582577
## 2     48 49337332  2  15245245 7.416 0.001561 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-Statistic = 7.416, and the p-value < 0.01 with an adjusted R-squared of 0.204. From the F-test, we can reject the null hypothesis (H_0) in favor of a more complete model (H_1) which now includes the covariates **mask_use** and **test_rate**. This is the model we will henceforth build upon.

Coefficient significance (t-test)

In order to assess the performance of the model, we can look at the other model coefficients:

```
coeftest(model_1, vcovHC)
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  2530.239162  501.044176  5.0499 6.799e-06 ***
## mask_useTRUE -990.469625  324.753111 -3.0499   0.00372 **
## test_rate       0.018295    0.010272  1.7811   0.08123 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Under a significance level of 0.05, we can accept the alternative hypotheses $H_a : \beta_1 \neq 0$, which means **mandatory_mask_use** explains at least a part of the variability observed in the **case_rate**.

On the other hand, for **test_rate** we failed to reject the null hypothesis that $H_0 : \beta_2 = 0$. It seems that **test_rate** is not absorbing a significant part of the variability observed in the outcome, **case_rate**.

Our estimate for β_1 (the coefficient of our variable of interest) is $\tilde{\beta}_1 \approx -990.5$, with a standard error of ~324.8 and a p-value of 0.004.

Practical significance

According to Model 1, enforcement of mask use policies would be expected to reduce the case rate by ~990.5 cases/100,000 or by ~1%, *ceteris paribus*. Given that the median COVID-19 case rate among US states is 2,633 per 100,000 population, the coefficient estimate has practical significance, with an effect size corresponding to a reduction of 37.6% of the median case rate among states.

Model 2

Objective

Model 2 is designed to be our optimal model, which strikes a balance between accuracy and parsimony. It is intended to reflect our best understanding of the relationships among key variables. It includes the same covariates used in Model 1, in addition to new covariates related to structural demographics and behavioral differences among US States that might partly correlate with the variability observed in the case rate.

It is difficult to make *a priori* assumptions regarding the variables we will utilize for Model 2. We hope to select one causal variable from each of the three broader categories that we believe align with our hypotheses regarding the factors influencing COVID-19 case rates. The broad categories include:

- Age demographics
- Socioeconomic demographics
- Actual social distancing

Model selection will be based upon our EDA. For each one of these categories, we will look for variables that correlate more strongly with case rate, but without a high degree of collinearity with the other variables already included in the model.

```
df_race <- df %>%
  select(state, case_rate, white_pop, black_pop, hispanic_pop, other_pop)

df_socio <- df %>%
  select(state, case_rate, homeless_total, poverty_rate, household_income, life_expectancy, unemployment

df_dist <- df %>%
  select(state, case_rate, `mob_R&R`, `mob_G&P`, mob_P, mob_TS, mob_WP, mob_RES)

df_age <- df %>%
  select(state, case_rate, age_0_18, age_19_25, age_26_34, age_35_54, age_55_64, age_65)

df_socio$life_expectancy <- df_socio$life_expectancy/10
```

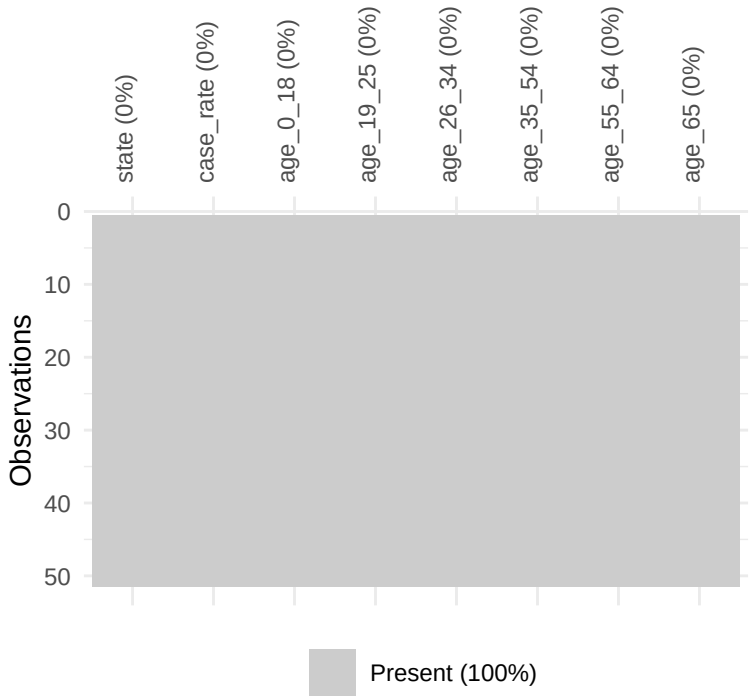
Exploratory Data Anaysis

Age Demographics

In order to select the most appropriate age groups (i.e. % of population of a particular age group in a State) to include in our model, we must conduct a more detailed exploratory analysis of the individual independent variables, their relationship with the dependent variable and their fellow categories.

First we ensure that there is no missing data in any of the age groups. The below figure confirms that there is no missing data:

```
df_age %>%
  vis_miss() +
  theme(axis.text.x = element_text(angle = 90))
```



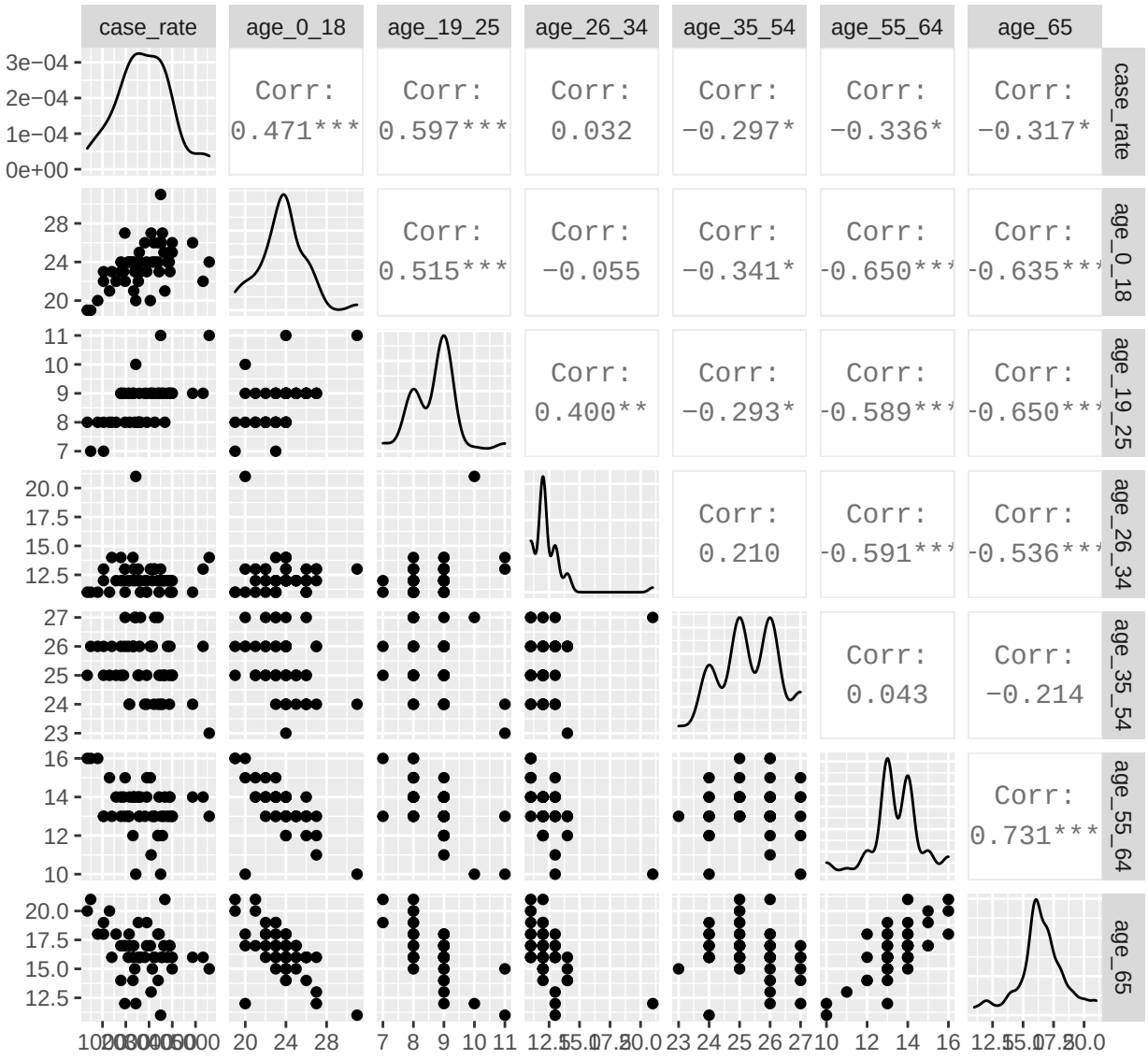
Next we ensure there is no spurious data contained in any of the age categories. From below, we can see that all the categories look to be well behaved in terms of the data they contain:

```
df_age %>%
  select(where(is.numeric)) %>%
  summary()
```

```
##    case_rate       age_0_18       age_19_25        age_26_34
##  Min.   : 344   Min.   :19.00   Min.   : 7.000   Min.   :11.00
##  1st Qu.:2040   1st Qu.:22.50   1st Qu.: 8.000   1st Qu.:12.00
##  Median :2633   Median :24.00   Median : 9.000   Median :12.00
##  Mean   :2749   Mean   :23.65   Mean   : 8.706   Mean   :12.31
##  3rd Qu.:3516   3rd Qu.:25.00   3rd Qu.: 9.000   3rd Qu.:13.00
##  Max.   :5589   Max.   :31.00   Max.   :11.000   Max.   :21.00
##    age_35_54       age_55_64        age_65
##  Min.   :23.00   Min.   :10.00   Min.   :11.00
##  1st Qu.:25.00   1st Qu.:13.00   1st Qu.:16.00
##  Median :25.00   Median :13.00   Median :16.00
##  Mean   :25.33   Mean   :13.43   Mean   :16.47
##  3rd Qu.:26.00   3rd Qu.:14.00   3rd Qu.:17.50
##  Max.   :27.00   Max.   :16.00   Max.   :21.00
```

Now that we have ensured the integrity of our data, we examine how each particular age group is related to the dependent variable and to each of the other age variables:

```
ggpairs(df_age[,-1])
```

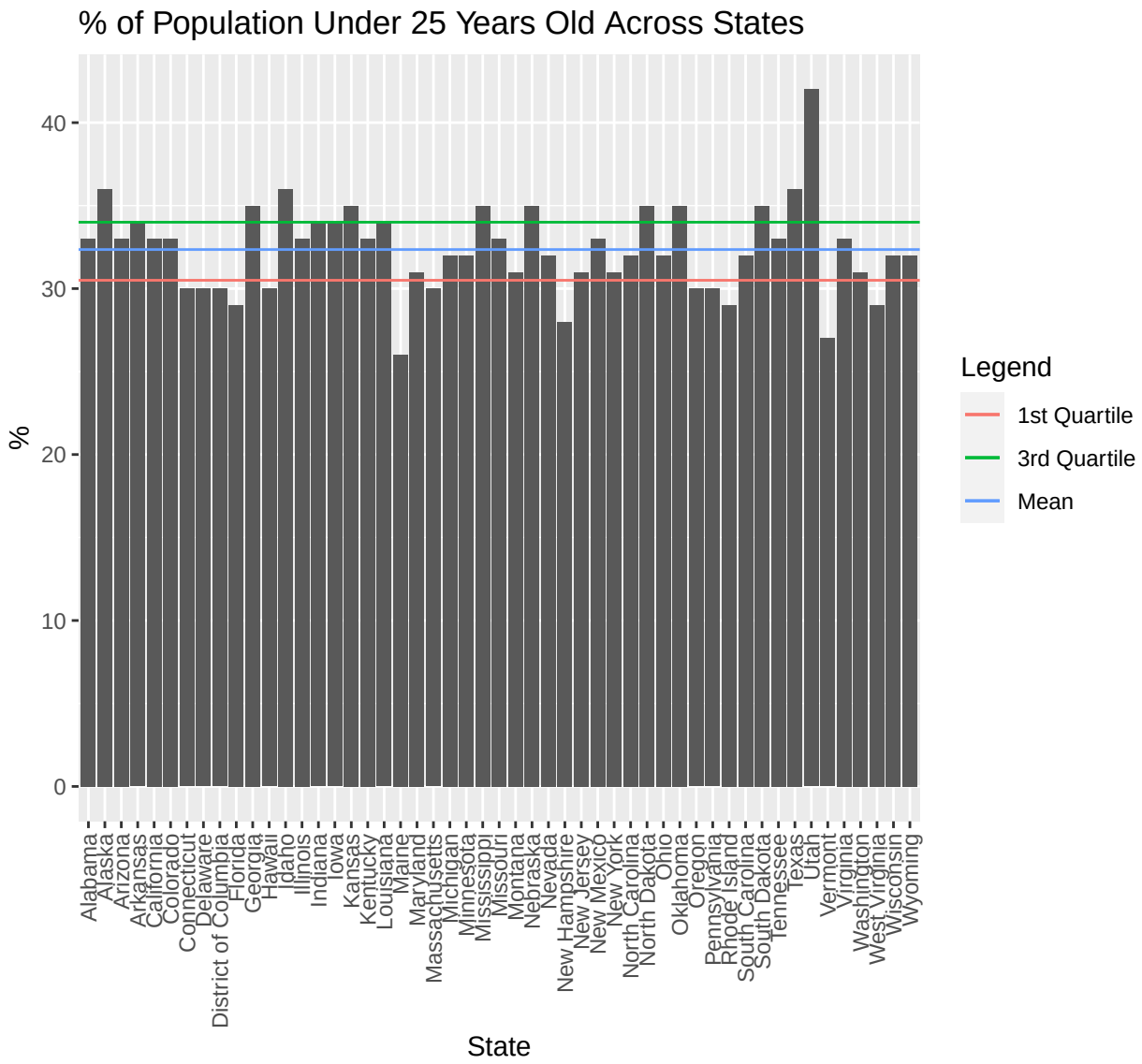


From the above scatterplot matrix we can see that the age groups that have highest correlation with our dependent variable (case rate) are below 25 (r = ~0.6) and 0-18 (r = ~0.5). These might both be valuable causal factors in terms of case rate and could be combined to enhance the model, while preventing the colinearity occuring if the two variables were added separately (as they also strongly correlate with each other, r = ~0.5). We can subsequently look at our newly transformed '<25' category in more detail:

First we can plot a bar graph to see what percentge of the population <25 exists in each State:

```
df_age$age_below_25 = df$age_0_18 + df$age_19_25

df$age_below_25 = df$age_0_18 + df$age_19_25

hist_state_age <- df %>%
  ggplot(aes(x = state, y = age_below_25)) +
  geom_bar(stat="identity") +
  labs(title = "% of Population Under 25 Years Old Across States",
```

```
       y = "%",
       x = "State",
       color = "Legend") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  geom_hline(aes(yintercept=mean(df$age_below_25), color='Mean')) +
  geom_hline(aes(yintercept=quantile(df$age_below_25, 0.25), color='1st Quartile')) +
  geom_hline(aes(yintercept=quantile(df$age_below_25, 0.75), color='3rd Quartile'))

hist_state_age
```

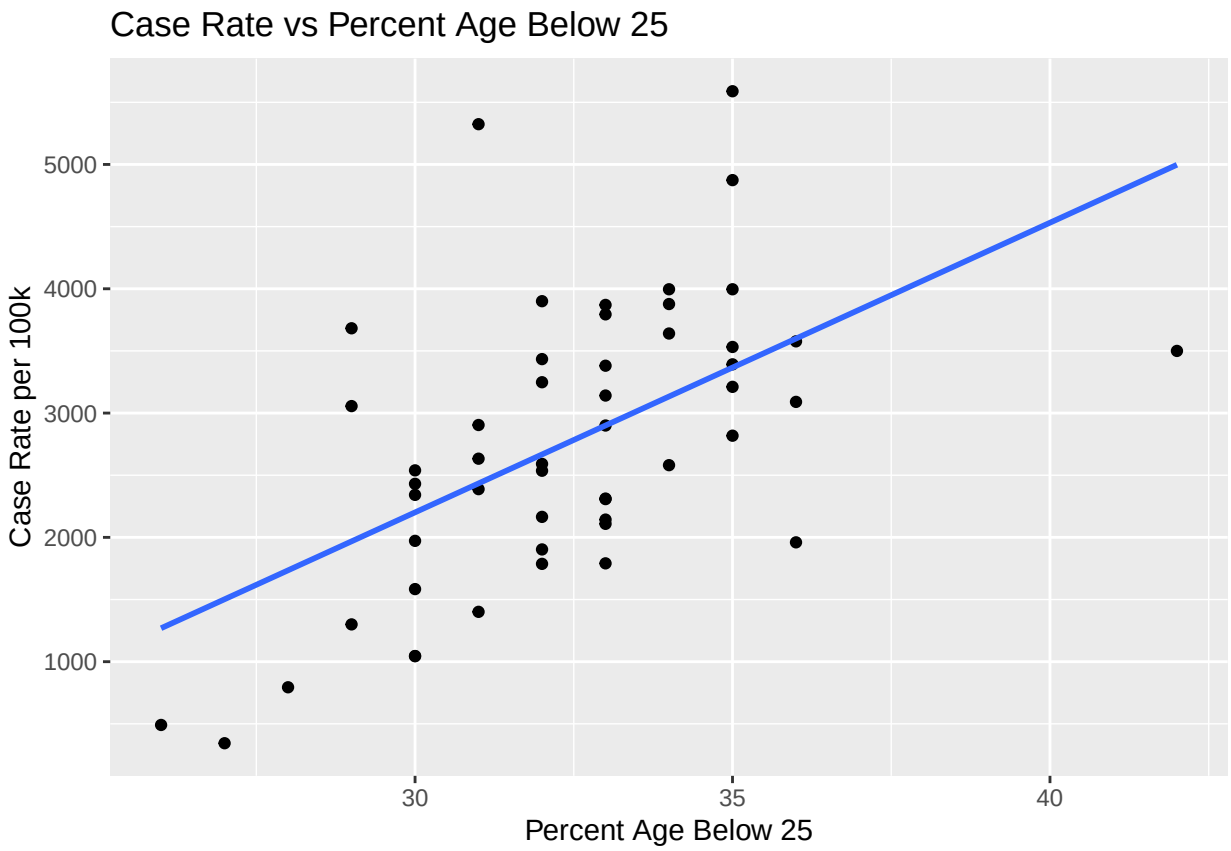


As we can see from the above, all of the States have similar proportions of their population < 25, with the interquartile distance being relatively small.

We can subsequently move on to the comparison between our new 'age <25' category and our dependent variable:

```
df_age %>%
  select(case_rate, age_below_25) %>%
```

```
  ggplot(aes(y = case_rate, x = age_below_25)) +
  geom_point() +
  geom_smooth(method = "lm", level = 0) +
  labs(
    title = "Case Rate vs Percent Age Below 25",
    x = "Percent Age Below 25",
    y = "Case Rate per 100k"
  )
```



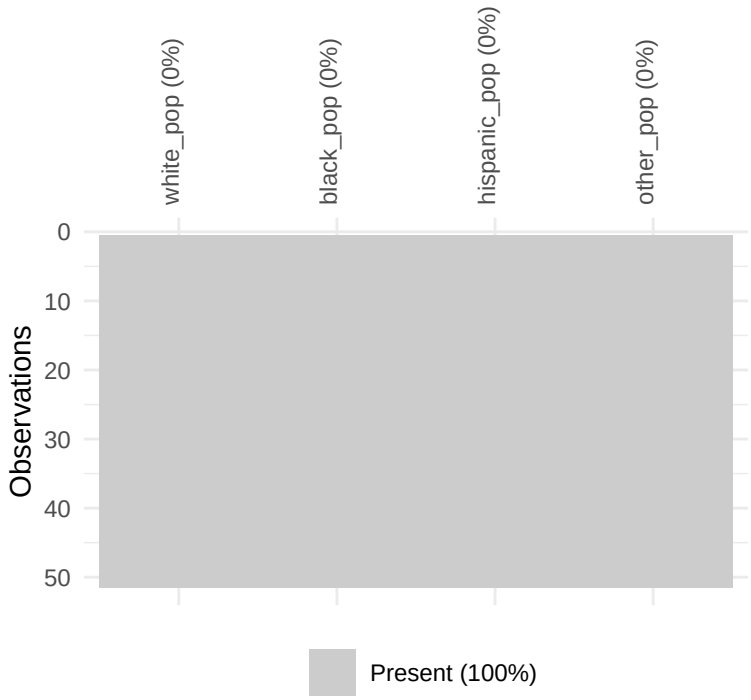
As we can see from the above, there is a definite positive relationship between the percent of people in the <25 age category and the case rate per 100. As such, it is unlikely that further transformation to this variable is required, and we can utilize it as a representation of an age demographic factor in our causal model moving forward.

Case Rate vs. Socio-Economic Demographics

It is well known and documented that COVID-19 has affected different ethnic groups differently, with figures demonstrating that case rates and death rates are higher among certain minority groups. This may be partially attributable to genetic factors. It may also be related to socioeconomic factors, such as poverty, inability to work from home, education, among many other reasons, which are also related to race.

Again, we begin by confirming the data has been recorded properly and that there are no missing values:

```
df_race[,c(-1,-2)] %>%
  vis_miss() +
  theme(axis.text.x = element_text(angle = 90))
```



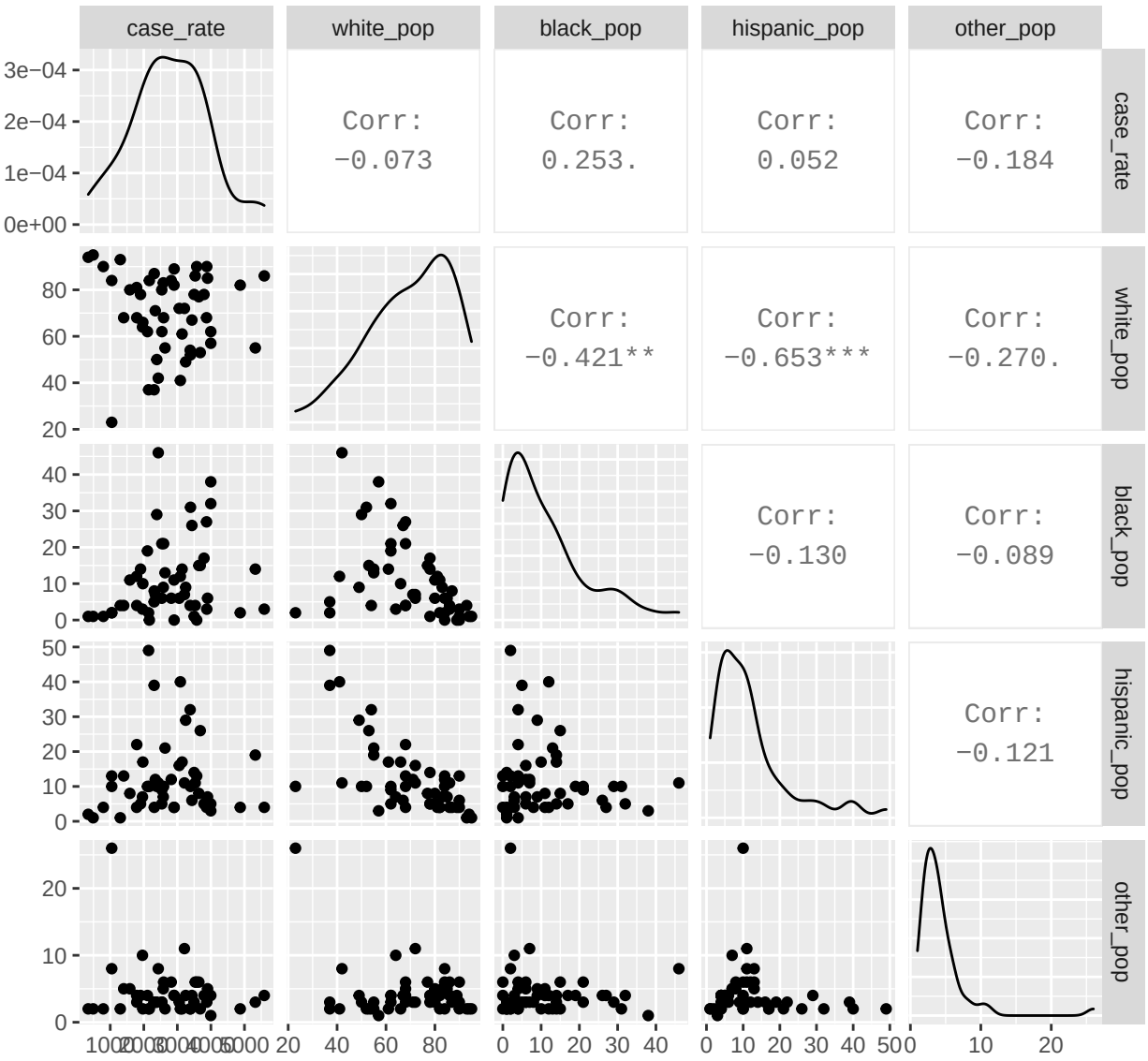
Next we ensure there is no spurious data contained in any of the race categories. From below, we can see that all the categories look to be well behaved in terms of the data they contain:

```
df_race[,c(-1,-2)] %>%
  select(where(is.numeric)) %>%
  summary()
```

```
##    white_pop       black_pop      hispanic_pop     other_pop
##  Min.   :23.00   Min.   : 0.00   Min.   : 1.00   Min.   : 1.000
##  1st Qu.:59.00   1st Qu.: 3.00   1st Qu.: 5.00   1st Qu.: 2.000
##  Median :72.00   Median : 7.00   Median :10.00   Median : 3.000
##  Mean   :70.04   Mean   :10.94   Mean   :12.04   Mean   : 4.294
##  3rd Qu.:84.00   3rd Qu.:14.50   3rd Qu.:13.50   3rd Qu.: 5.000
##  Max.   :95.00   Max.   :46.00   Max.   :49.00   Max.   :26.000
```

Now that we have ensured the integrity of our data, we examine how the State percentage of particular race category might be related to the dependent variable and to each of the other race groups:

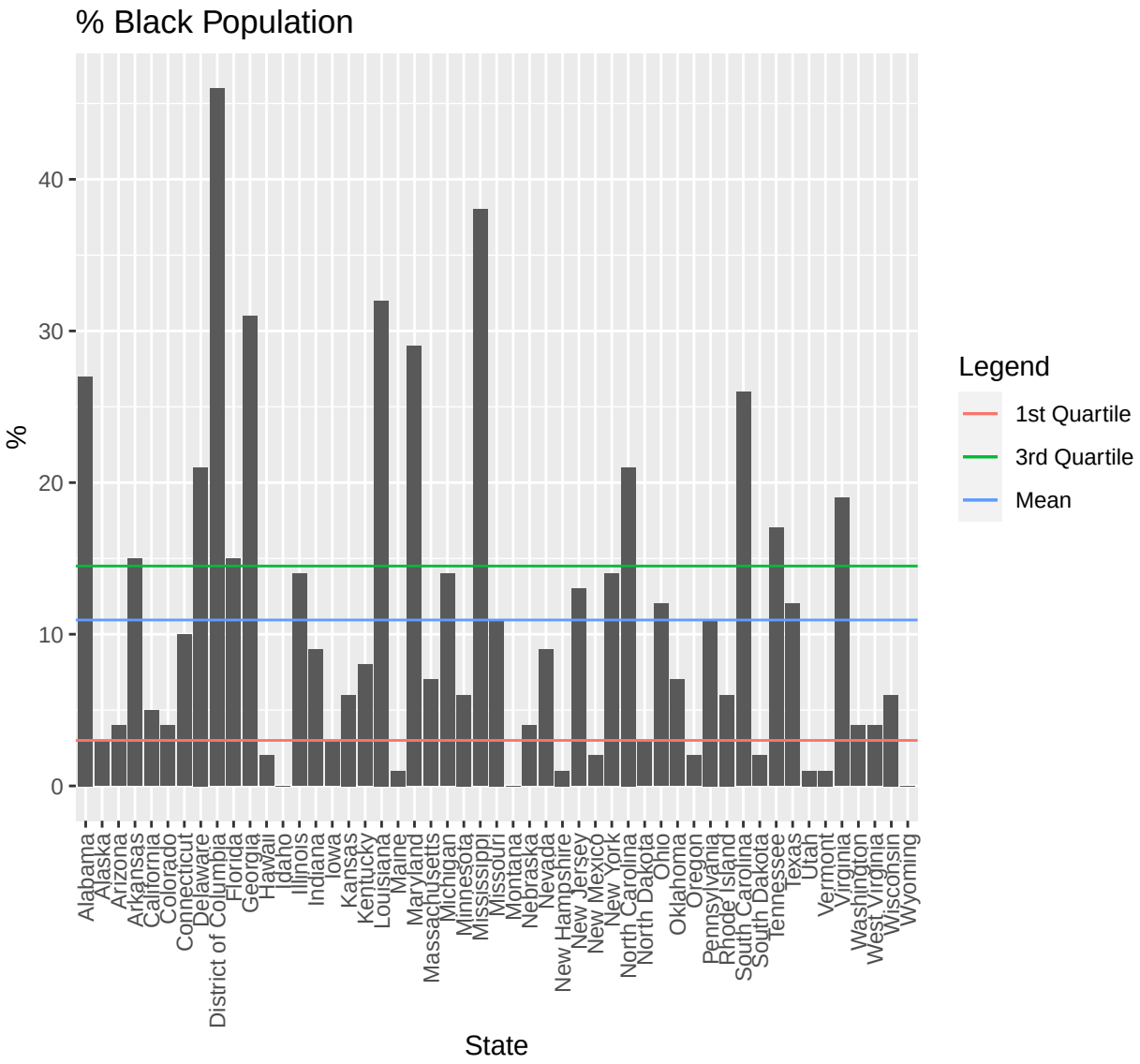
```
ggpairs(df_race[,-1])
```



As we can see from the above scatterplot matrix, in general there are generally weak relationships between case rate per 100,000 population and percentage of the particular racial group per State, in fact the majority show a negative relationship. Nevertheless, we can observe that the strongest absolute value for relationship is positive, and demonstrated by the percentage of black population per State, correlating with case rate with an r = ~0.3. We can visualize the distribution of black people by percentage of State population using a bar graph:

```
hist_state_black <- df %>%
  ggplot(aes(x = state, y = black_pop)) +
  geom_bar(stat="identity") +
  labs(title = "% Black Population",
       y = "%",
       x = "State",
       color = "Legend") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  geom_hline(aes(yintercept=mean(df$black_pop), color='Mean')) +
  geom_hline(aes(yintercept=quantile(df$black_pop, 0.25), color='1st Quartile')) +
  geom_hline(aes(yintercept=quantile(df$black_pop, 0.75), color='3rd Quartile'))
```

```
hist_state_black
```

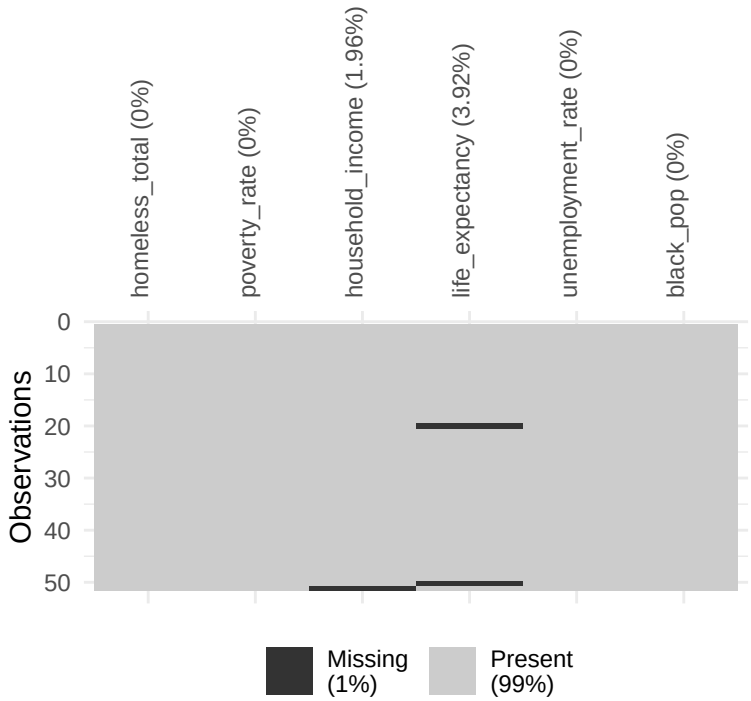


As we can observe from the above graph, there is a wide variance in the percentage of black population by State.

As we mentioned in the introduction, race may potentially act as a proxy for socioeconomic factors such as poverty, access to healthcare and education, among others. As such, if there is potential colinearity observed between dedicated socioeconomic categories in the dataset and a particular racial category, including them both as dependent variables might cause problems with the model. As such we have included the percentage of black population per State in our scatterplot analysis of the relationship between various socioeconomic factors and case rate. Our reasoning for this is because we already know from previously that the black ethnicity variable is strongly correlated with case rate.

First we assess for missing data in the socioeconomic categories we will be assessing:

```
df_socio[,c(-1,-2)] %>%
  vis_miss() +
  theme(axis.text.x = element_text(angle = 90))
```



Next we do an analysis to ensure there are no spurious values in the data for any of the variables:

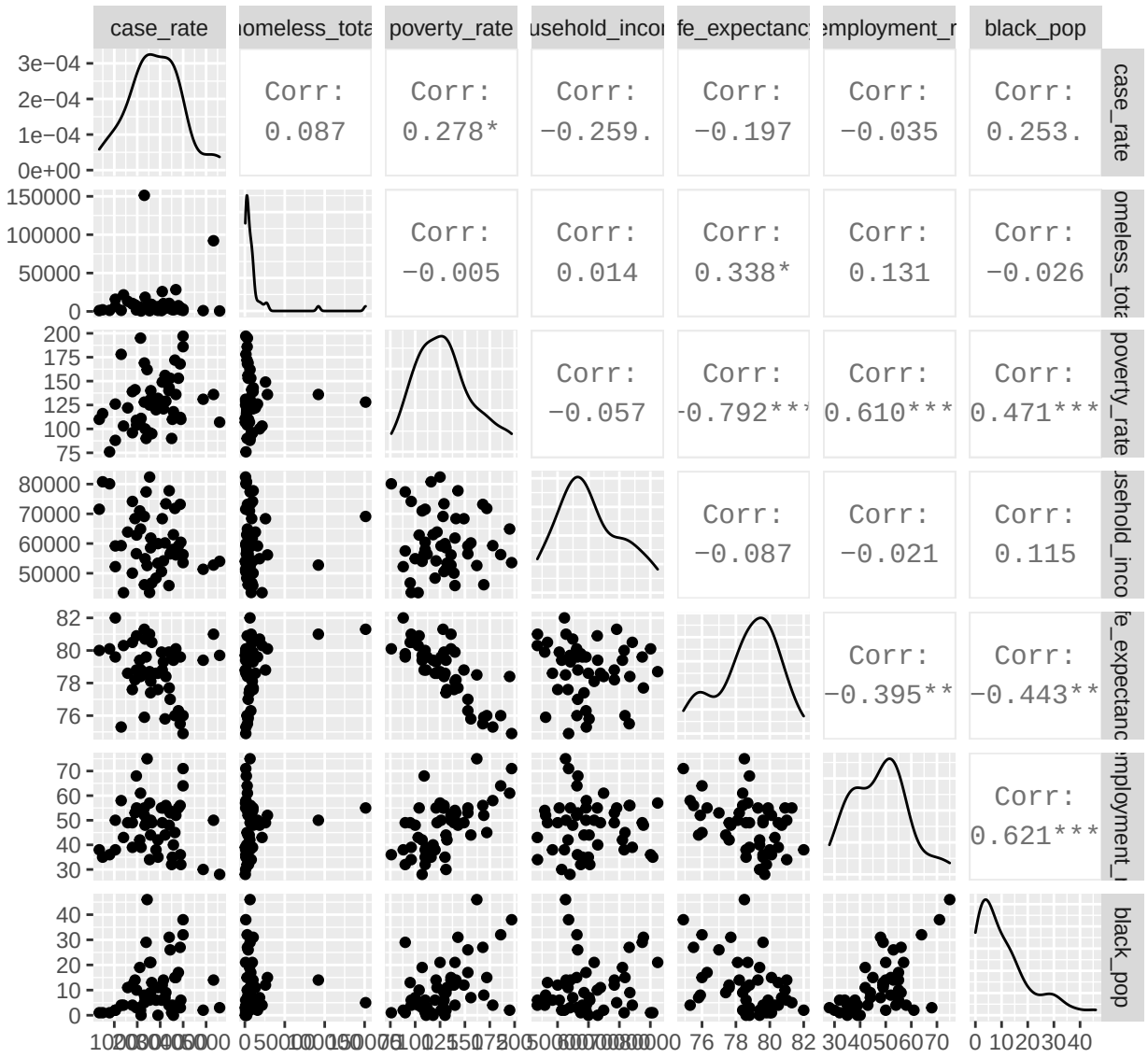
```
df_socio[,c(-1,-2)] %>%
  select(where(is.numeric)) %>%
  summary()
```

```
##  homeless_total    poverty_rate   household_income life_expectancy
##  Min.   :   548   Min.   : 76.0   Min.   :43469    Min.   :74.90
##  1st Qu.:  2315   1st Qu.:109.5   1st Qu.:53434    1st Qu.:77.70
##  Median :  4538   Median :128.0   Median :58882    Median :78.80
##  Mean   : 11023   Mean   :129.1   Mean   :60478    Mean   :78.76
##  3rd Qu.:  9466   3rd Qu.:142.0   3rd Qu.:68380    3rd Qu.:79.90
##  Max.   :151278   Max.   :197.0   Max.   :82372    Max.   :82.00
##                                   NA's   :1        NA's   :2
##  unemployment_rate   black_pop
##  Min.   :28.00     Min.   : 0.00
##  1st Qu.:38.50     1st Qu.: 3.00
##  Median :49.00     Median : 7.00
##  Mean   :47.47     Mean   :10.94
##  3rd Qu.:55.00     3rd Qu.:14.50
##  Max.   :75.00     Max.   :46.00
##
```

We have already identified a few missing values in the household income and life expectancy column. Nevertheless, the rest of the variables look to be well behaved and contain summary values concordant with what one would expect for these data types.

Next we can create the scatterplot matrix for the socioeconomic variables of interest (+ percentage of blacks):

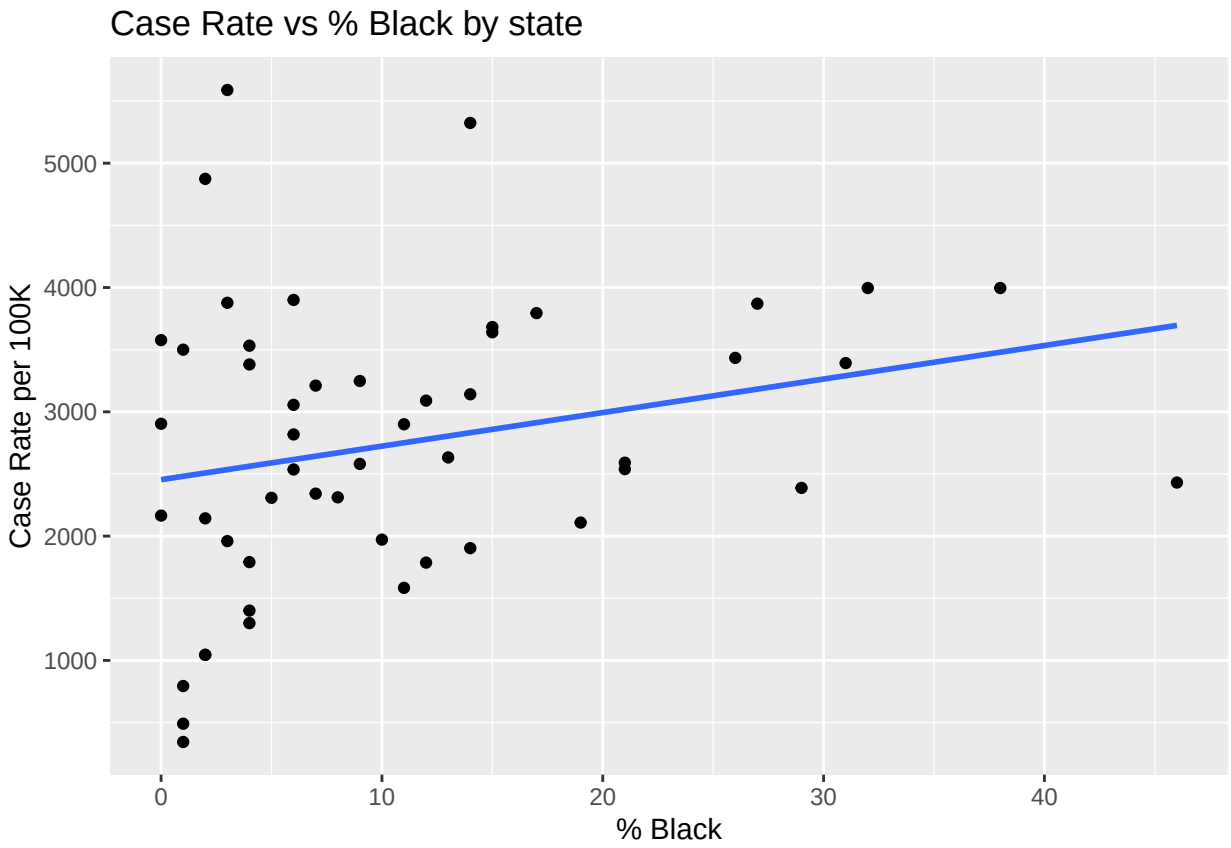
```
ggpairs(df_socio[,-1])
```



From the pairs plot above, we see that poverty rate and household income have the highest absolute correlation with case rate per 100k. They are comparable to the correlation of percentage of black population variable seen above. Additionally, the black population variable has high collinearity with household income, poverty rate, life expectancy, and unemployment rate. Because of this, the percent black population variable may act as a variable that can control for many of these factors, while also being representative of racial factors. We have previously demonstrated a bar plot for black race by State.

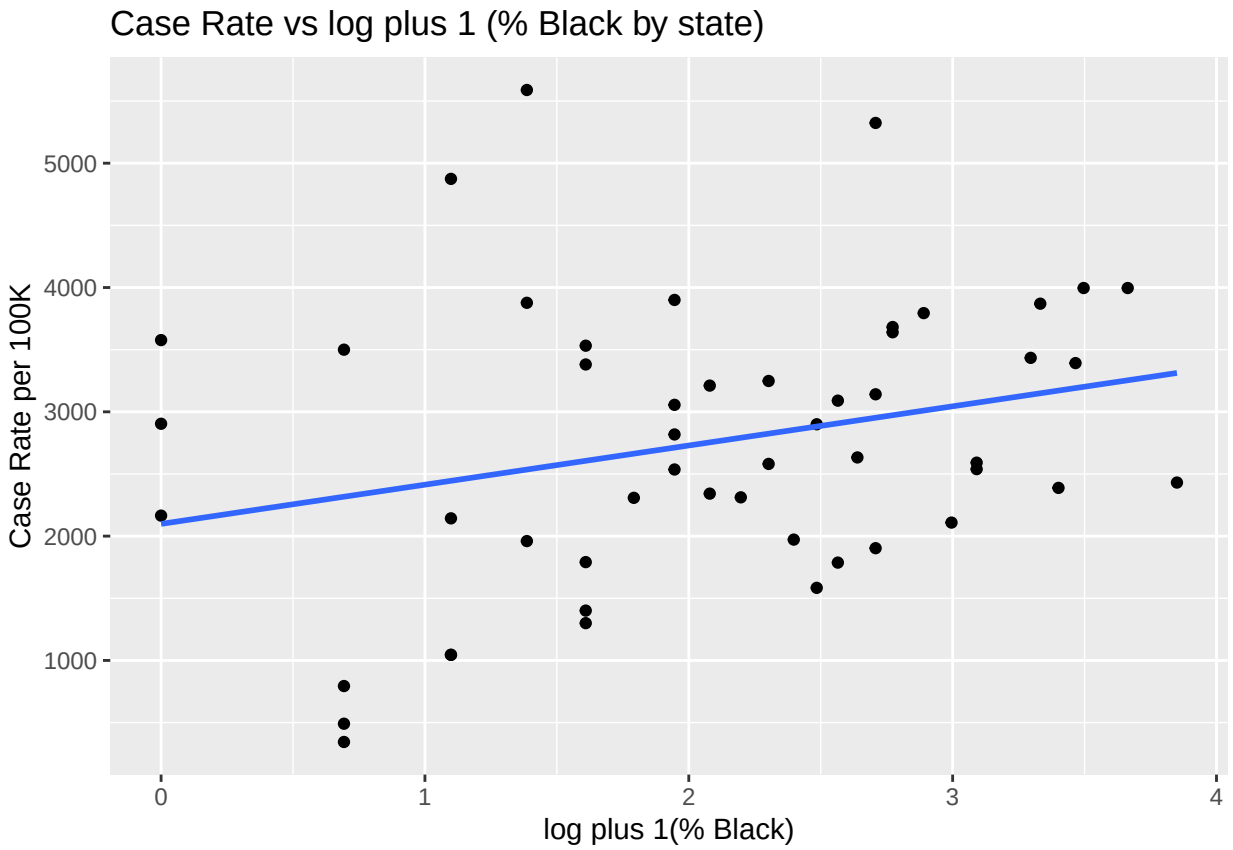
We explore the bivariate relationship between case rate and percentage of black population below:

```
df_race %>%
  ggplot(aes(y = case_rate, x = black_pop)) +
  geom_point() +
  geom_smooth(method = "lm", level = 0) +
  labs(
    title = "Case Rate vs % Black by state",
    x = "% Black",
    y = "Case Rate per 100K"
  )
```



The untransformed relationship shows a rather weak positive correlation with case rate. Moreover, the points seem to become further spaced as the value of X increases. Disregarding the regression line, we can observe a somewhat non-linear relationship between the data points. As such, this variable could benefit from transformation. Because some of our black population values are equal to 0, a log transformation would not apply. To avoid this, each value is increased by 1 prior to the log transform:

```
df_race %>%
  ggplot(aes(y = case_rate, x = log1p(black_pop))) +
  geom_point() +
  geom_smooth(method = "lm", level = 0) +
  labs(
    title = "Case Rate vs log plus 1 (% Black by state)",
    x = "log plus 1(% Black)",
    y = "Case Rate per 100K"
  )
```



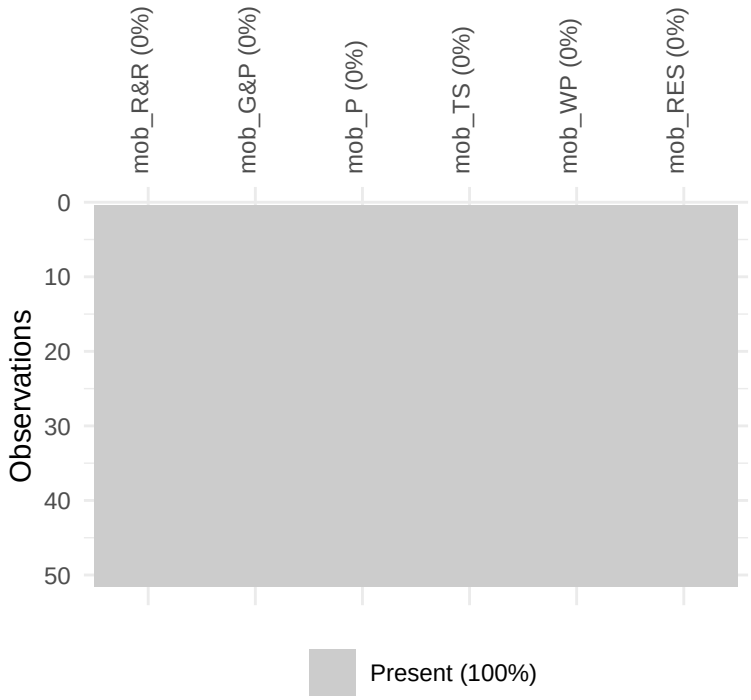
The graph above demonstrates that a transformation distributes the points more evenly along the X axis, making the potential linear relationship more discernable.

Case Rate vs. Social Distancing Adherence

The Google Human mobility data includes information on the amount of time spent at various public locations compared to Google's baseline data. The values are recorded as percentage changes with possible values ranging from -100 to 100. In general this data is a proxy for adherence to social distancing regulations. We would expect to see reductions in mobility upon enforcement of shelter-in-place, work-from-home or quarantine orders. Moreover, we would expect to see drastic reductions at transit stations (e.g. busy commuter hubs). As social distancing measures are designed to reduce the transmission of COVID-19, we would expect there to be a positive relationship between mobility and case rate, i.e. reductions in mobility are linked to lower case rates and vice versa. Nevertheless, there are several different social distancing metrics to choose from, which we deal with in this section.

We can see that there are no missing data points in this category:

```
df_dist[,c(-1,-2)] %>%
  vis_miss() +
  theme(axis.text.x = element_text(angle = 90))
```



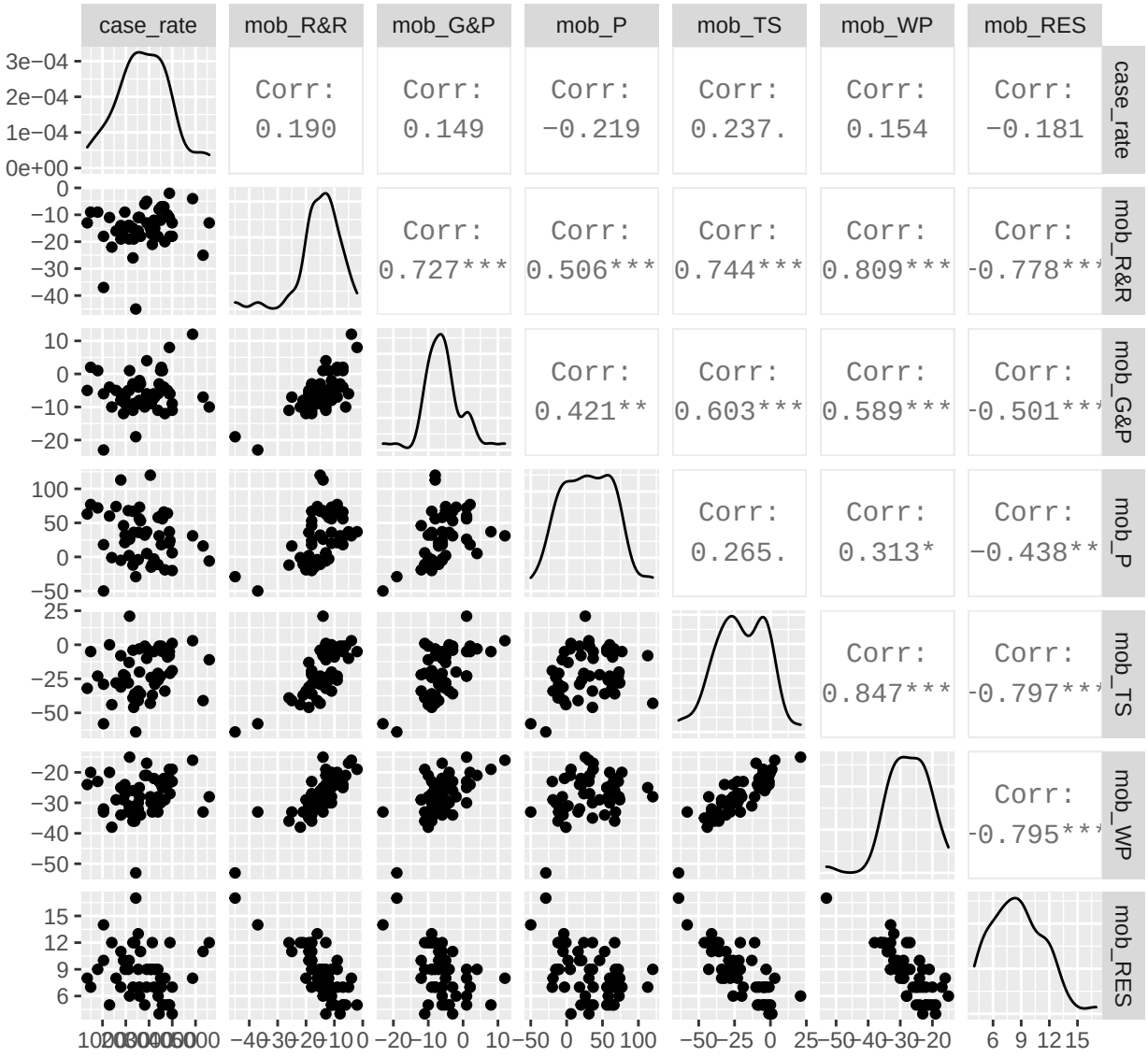
A look at the summary statistics of the values fail to show any significant errors in values:

```
df_dist %>%
  select(where(is.numeric)) %>%
  summary()
```

```
##    case_rate       mob_R&R          mob_G&P           mob_P
##  Min.   : 344   Min.   :-45.00   Min.   :-23.000   Min.   :-50.00
##  1st Qu.:2040   1st Qu.:-18.00   1st Qu.: -9.000   1st Qu.: -2.00
##  Median :2633   Median :-14.00   Median : -6.000   Median : 31.00
##  Mean   :2749   Mean   :-14.82   Mean   : -5.765   Mean   : 30.27
##  3rd Qu.:3516   3rd Qu.:-11.00   3rd Qu.: -3.500   3rd Qu.: 60.00
##  Max.   :5589   Max.   : -2.00   Max.   : 12.000   Max.   :120.00
##      mob_TS           mob_WP          mob_RES
##  Min.   :-64.00   Min.   :-53.00   Min.   : 4.000
##  1st Qu.:-33.00   1st Qu.:-32.00   1st Qu.: 7.000
##  Median :-22.00   Median :-28.00   Median : 9.000
##  Mean   :-20.55   Mean   :-27.45   Mean   : 8.608
##  3rd Qu.: -5.00   3rd Qu.:-23.00   3rd Qu.:10.500
##  Max.   : 21.00   Max.   :-15.00   Max.   :17.000
```

Below is the pairs plot for the mobility data and case_rate:

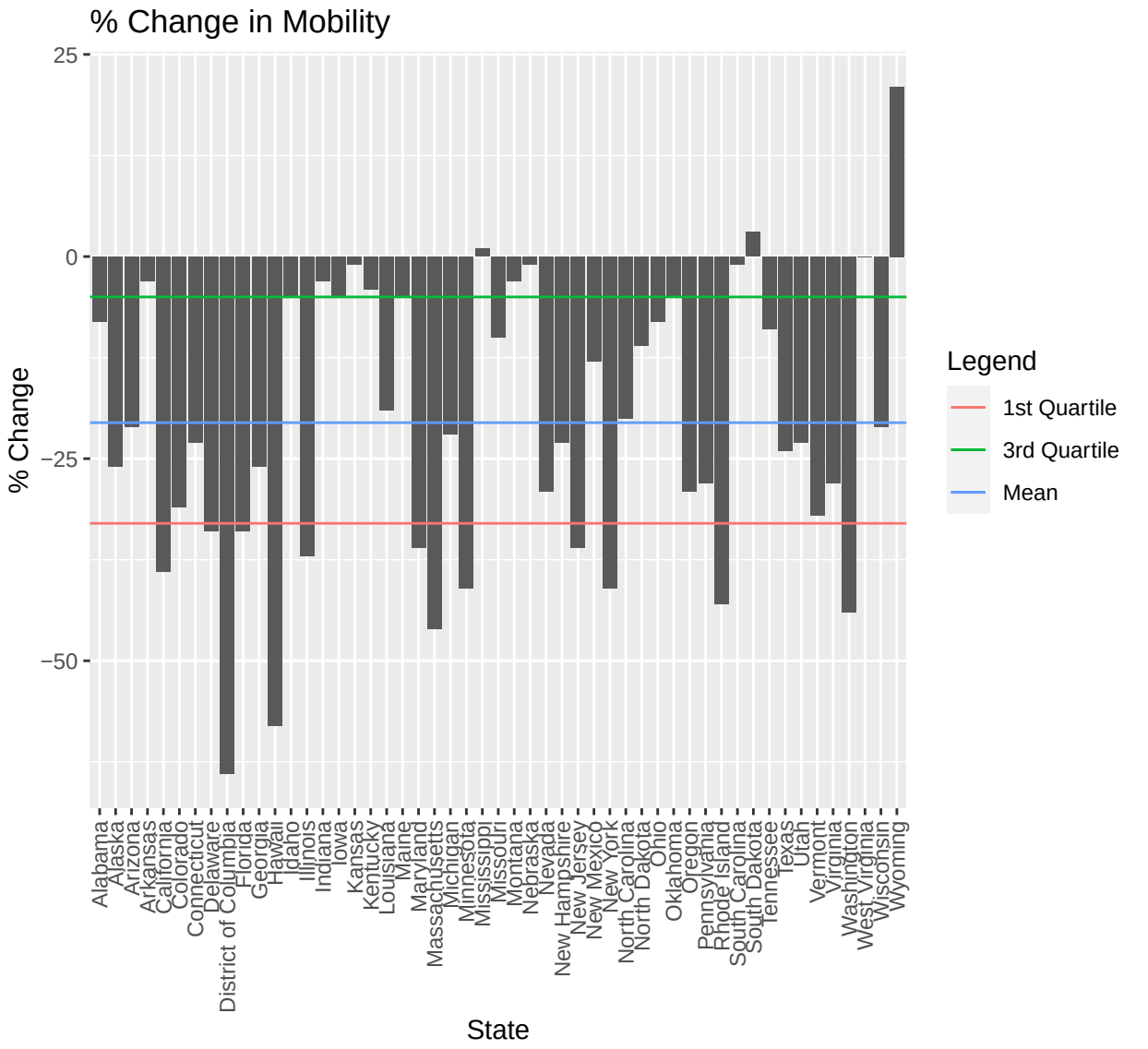
```
ggpairs(df_dist[,-1])
```



All the mobility data are highly collinear with one another, so for the model, we will select mob_TS (i.e. change in mobility at transit stations) as it has the highest correlation with case rate, with an r of 0.237. We can visualize the changes in mobility by State to see if there are any interesting trends worth exploring:

```
hist_state_mob <- df %>%
  ggplot(aes(x = state, y = mob_TS)) +
  geom_bar(stat="identity") +
  labs(title = "% Change in Mobility",
       y = "% Change",
       x = "State",
       color = "Legend") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  geom_hline(aes(yintercept=mean(df$mob_TS), color='Mean')) +
  geom_hline(aes(yintercept=quantile(df$mob_TS, 0.25), color='1st Quartile')) +
  geom_hline(aes(yintercept=quantile(df$mob_TS, 0.75), color='3rd Quartile'))
```

```
hist_state_mob
```



As expected, the majority of States have witnessed a decline in mobility at their respective transit centers. In general, the data is congruent with expectations, for example the District of Columbia has seen the largest decrease in transit station mobility at >70%. Interestingly, there are 3 States that have seen increases, with Wyoming seeing a substantial increase in transit station mobility. Nevertheless, these trends are unlikely to affect our further analysis using the variable in our model.

The figure below shows the bivariate relationship between case rate variable and the change in mobility at Transit Stations:

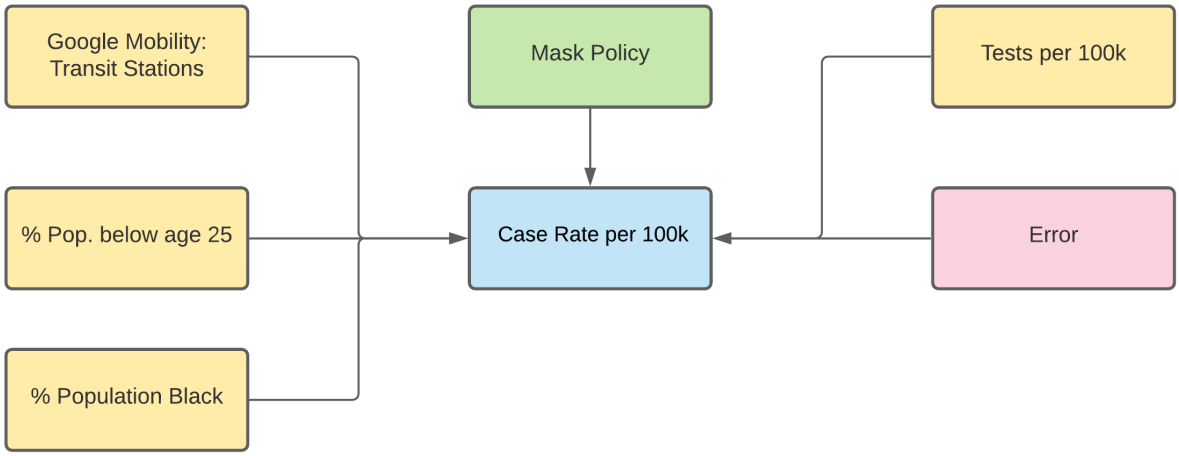
```
df_dist %>%
  ggplot(aes(y = case_rate, x = mob_TS)) +
  geom_point() +
  geom_smooth(method = "lm", level = 0) +
  labs(
    title = "Relationship Between Human Mobility: Transit Stations and Case Rate by state",
    x = "Percentage Change in Human Mobility: Transit Stations",
```

```
    y = "Case Rate per 100K"
)
```

From the above figure, we can discern a clear positive relationship between the X and Y variables. As the points are approximately evenly distributed within the axes, there is no clear and explainable transformation that can be applied. The variable will be left as is for the second model.

Casual Diagram for Model 2

The causal diagram for Model 2 further resembles the diagram from the introduction section. We have now explored the effect of age, race, socioeconomic factors, and mobility as it pertains to policies for mask use and the COVID-19 case rate per 100k. For the age category, percent population <25 had strongest correlation with case rate. The percentage of black people per State served as the strongest proxy to control for socioeconomic and racial characteristics. Using Google's mobility data as a proxy for adherence to social distancing measures, we found that the percentage change in traffic through transit stations had strongest correlation with case rate per 100k.



Model specification

The second regression model has **COVID-19 Case Rate per 100,000 population** as the outcome variable and 5 covariates: The variable of interest (**Mandatory Mask Use**), and 4 control variables: **Test Rate per 100,000 Population**, **Percentage of Population Below 25 Years Old**, **Log of Percentage of Black Ethnicity in Total Population + 1**, and **Human Mobility Change in Transit Stations**.

The variable of interest continues to be **Mandatory Mask Use** and the primary measurement goal remains to assess the significance and practical impact of **Mandatory Mask Use** on the **Case Rate**. We predict this model (with more control variables) to be better at capturing the actual significance and practical relevance of the **Mandatory Mask Use** on **Case Rate**.

Model 2 has the format:

$$\begin{aligned}\text{case rate per } 100,000 \text{ pop.} = \beta_0 + \beta_1(\text{mandatory mask use policy}) + \beta_2(\text{test rate per } 100,000 \text{ pop.}) + \\ \beta_3(\%\text{ pop. } < 25) + \beta_4(\ln(\%\text{ black pop.} + 1)) + \beta_5(\text{mobility } \Delta \text{ transit stations})\end{aligned}$$

Model Summary

```
model_2 <- lm(case_rate ~ mask_use + test_rate + age_below_25 + log(black_pop + 1) + mob_TS, data = df)
std_errors = list(
  sqrt(diag(vcovHC(model_1))),
  sqrt(diag(vcovHC(model_2)))
)
stargazer(model_1, model_2, se = std_errors, type = "latex", title = "Model 2 Summary")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Dec 04, 2020 - 07:32:43 AM

Overall model significance (F-test)

This compares the null hypothesis, where H_0 : **Model 1**, against an alternative hypothesis where H_a : **Model 2** at a significance level of 0.05:

Table 2: Model 2 Summary

	Dependent variable:	
	case_rate	
	(1)	(2)
mask_use	−990.470***	−919.251***
	(324.753)	(227.028)
test_rate	0.018*	0.024**
	(0.010)	(0.011)
age_below_25		190.555***
		(46.660)
log(black_pop + 1)		461.650***
		(110.822)
mob_TS		14.832**
		(6.575)
Constant	2,530.239***	−4,603.593***
	(501.044)	(1,745.175)
Observations	51	51
R^2	0.236	0.670
Adjusted R^2	0.204	0.633
Residual Std. Error	1,013.835 (df = 48)	688.111 (df = 45)
F Statistic	7.416*** (df = 2; 48)	18.279*** (df = 5; 45)
Note:		*p<0.1; **p<0.05; ***p<0.01

```r
anova(model_1, model_2, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: case_rate ~ mask_use + test_rate
## Model 2: case_rate ~ mask_use + test_rate + age_below_25 + log(black_pop +
##     1) + mob_TS
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     48 49337332
## 2     45 21307370  3  28029962 19.733 2.602e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this F-test, we can reject the null hypothesis (**Model 1**) in favor of our optimized **Model 2**, which now includes the covariates **Mask Use**, **Test Rate per 100,000 Population**, **% of Age <25**, **ln(% of Black Population + 1)**, and **% Mobility Change at Mobile Transit Stations**. The F-Statistic is 19.733, and the p-value < 0.01. Model 2 has an adjusted R-squared of 0.633.

Coefficient significance (t-test)

Under a significance level of 0.05, we can accept all the alternative hypotheses: $H_{a1} : \beta_1 \neq 0$, $H_{a2} : \beta_2 \neq 0$, $H_{a3} : \beta_3 \neq 0$, $H_{a4} : \beta_4 \neq 0$, and $\beta_5 \neq 0$. This means all 5 covariates assist in explaining a part of the variability observed in the **case_rate**.

Our estimate for β_1 (the coefficient of our variable of interest) is $\tilde{\beta}_1 = -919.3$, with a standard error of 227.0 and a p-value of 0.0002. It continues to be statistically significant, and with an estimated value that did not change a lot from **Model 1** (-990.5) to **Model 2** (-919.2).

```r
coeftest(model_2, vcovHC)
```

```
##
## t test of coefficients:
##
##                       Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)        -4.6036e+03  1.7452e+03 -2.6379 0.0114146 *
## mask_useTRUE       -9.1925e+02  2.2703e+02 -4.0491 0.0002002 ***
## test_rate           2.3974e-02  1.0969e-02  2.1856 0.0340918 *
## age_below_25        1.9055e+02  4.6660e+01  4.0839 0.0001795 ***
## log(black_pop + 1)  4.6165e+02  1.1082e+02  4.1657 0.0001388 ***
## mob_TS              1.4832e+01  6.5746e+00  2.2559 0.0289818 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Practical significance

According to Model 2, suggests that enforcing a mandatory mask use policy would expect to reduce the COVID-19 positive case rate by ~919.2 cases per 100,000 population, with all other variables being held constant. Given that the median case rate across US states is 2,633 per 100,000 population, the coefficient estimate has practical significance, with an effect size corresponding to a reduction of 34.9% of the median State case rate value.

Model 3

Objective

Model 3 includes all the previous covariates, and several other covariates, erring on the side of inclusion. A key purpose of this model is to demonstrate the robustness of our coefficient for mandatory mask use. We do not expect our model performance to increase significantly compared to Model 2.

Subsequently, we will include other COVID-related measures adopted by States, which might also contribute towards variability of the dependent variable i.e. case rate per 100,000. Moreover, some of these variables may be colinear with our primary independent variable, mandated mask use. This may potentially reduce the overall explanatory performance of our model. We want to verify that even under harsh conditions our coefficient for mandatory mask use remains statistically significant, and with a practical significance close to that of Model 2.

Model 3 therefore acts as an *acid test* to further validate Model 2 as the most optimized model.

```
df_mod3 <- df %>%
  select(state, case_rate, bus_close_days, shelter_days, mask_legal, maskbus_use)
```

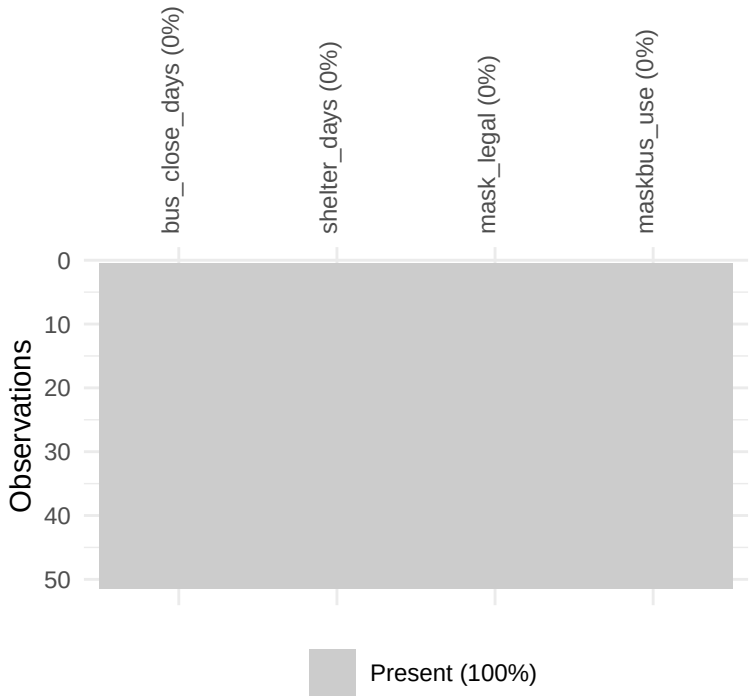
Case Rate vs. Other COVID Related Policies

The variables selected in the exploratory phase of Model 3 regard other COVID-19 specific policies. We are using the number of days a business had to close, the duration of shelter-in-place, and whether there was legal enforcement to wear a mask.

From the summary statistics, we can see that the numbers mostly make sense. There are no negative values, and the max number of days in shelter-in-place (since october 30th) would have the shelter-in-place start from March 19th, 2020, a full 8 days after the WHO declared COVID-19 a pandemic.

First we can assess the integrity of our data, and confirm that there are no missing datapoints:

```
df_mod3[,c(-1,-2)] %>%
  vis_miss() +
  theme(axis.text.x = element_text(angle = 90))
```



Next we can ensure that the data are well behaved and do not contain any spurious values:

```
df_mod3[,c(-1,-2)] %>%
  select(where(is.numeric)) %>%
  summary()
```

```
##  bus_close_days   shelter_days
##  Min.   : 0.00   Min.   :  0.00
##  1st Qu.:33.00   1st Qu.: 27.50
##  Median :43.00   Median : 46.00
##  Mean   :43.43   Mean   : 49.29
##  3rd Qu.:53.00   3rd Qu.: 59.50
##  Max.   :78.00   Max.   :225.00
```

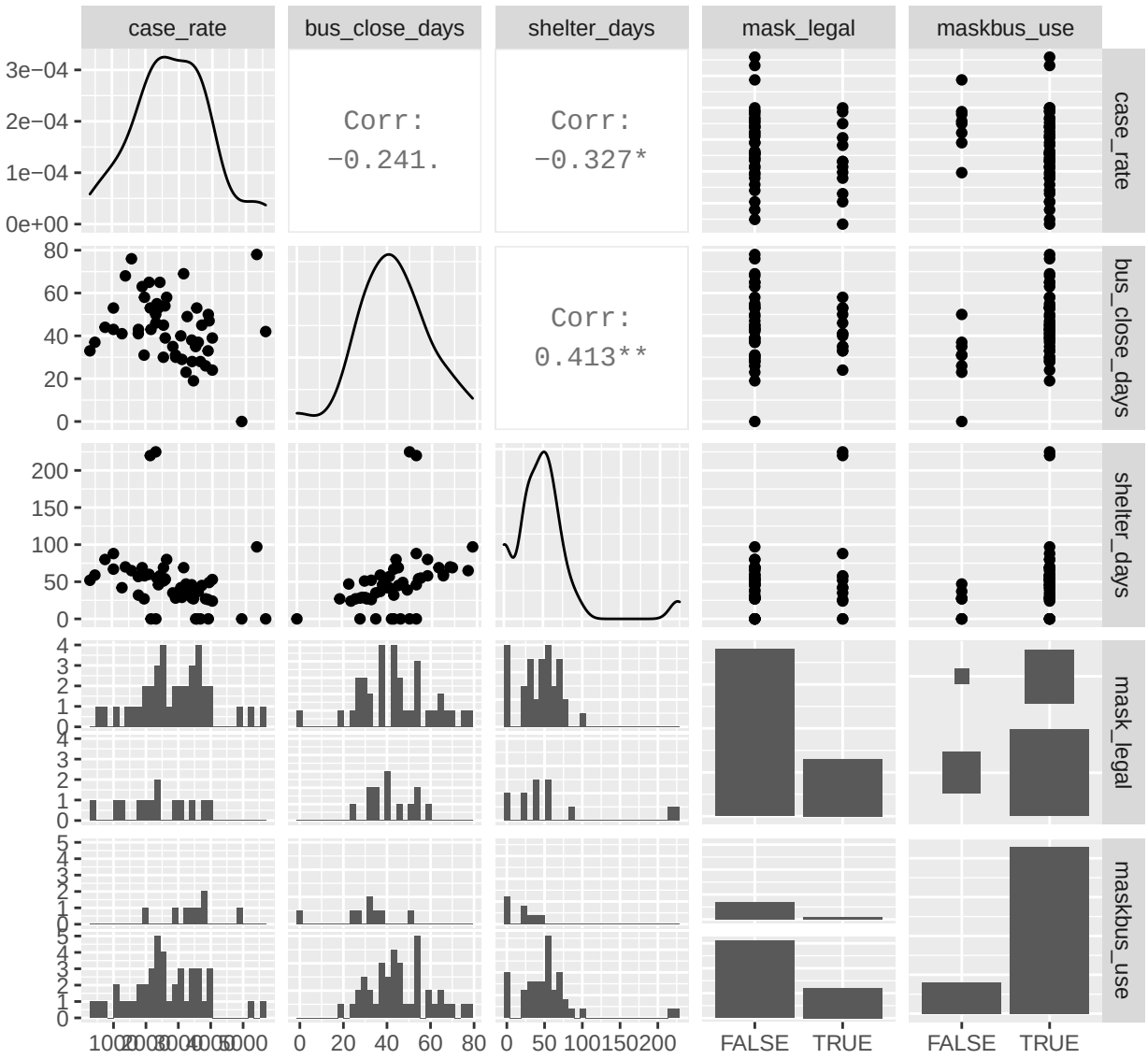
The continues variables corresponding to business close days and days of shelter in place seem to contain data concordant with what we would expect them to contain. Next we can assess the column containing data pertaining to enforcement of mask wear, which we can see contains 51 values, with no missing data:

```
summary(df_mod3$mask_legal)
```

```
##    Mode   FALSE    TRUE
## logical      38      13
```

Next we can utilize a scatterplot matrix to visualize the relationships between the case rate and other COVID-19 related policies:

```
ggpairs(df_mod3[,-1], upper = list(combo = "points"))
```



As we cannot calculate a Pearson r correlation score to quantify the relationship between a categorical variable and a continuous variable, instead we can conduct a point biserial test to assess for the relationship between the legal enforcement of mask use and case rate variables:

```
df_mod3 <- df_mod3 %>%
  mutate(binary_legal = as.integer(as.logical(mask_legal)))

cor.test(df_mod3$case_rate, df_mod3$binary_legal)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_mod3$case_rate and df_mod3$binary_legal
## t = -1.5153, df = 49, p-value = 0.1361
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.46031389  0.06797836
## sample estimates:
```

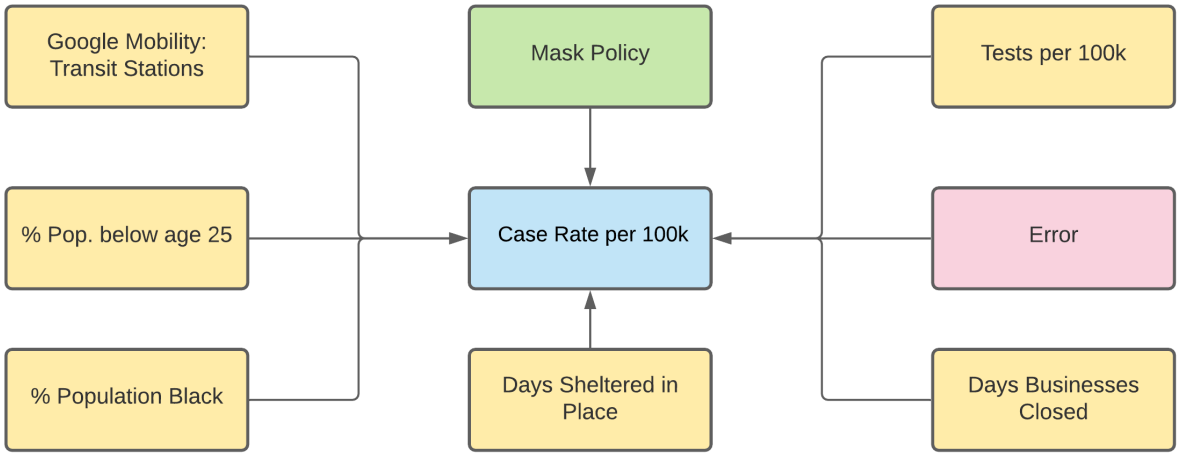
```
##        cor
## -0.2115688
```

As we can see from the above scatterplot matrix, variables pertaining to business closure days and shelter days are both negatively correlated with case rate, with Pearson r of ~-0.2 and ~-0.3, respectively. They are also both colinear, with a correlation of r = ~0.4. Due to the fact that these variables are negatively related to case rate (whereas the previous models have exclusively utilized variables that were positively related to case rate), and colinear, adding them separately will be useful to test our model by introducing variables that might interact strongly with the dependent and independent variables. Moreover, as the point biserial estimate of correlation between case rate and legal enforcement of mask wear is ~-0.2, it does not seem to be any more useful than the business closure and shelter in place day variables, and will henceforth be discarded.

As these two variables are specifically chosen to conflict with our existing variables, we elect not to do any further exploratory data analysis upon them for the sake of brevity. We have already assured that they do not contain any values that might be considered spurious nor do they contain a substantial amount of missing data.

Casual Diagram for Model 3

The causal diagram for Model 3 is the most "complete" model, and is thus identical to the diagram included in the introduction. We believe that Model 2 is the most optimal in regards to containing variables proxying for the complex demographic, socioeconomic, social, policy and other factors that contribute to case rate per 100,000 of population. The additional variables pertaining to shelter in place days and enforced business closures are both negatively correlated with case rate (as opposed to positively correlated for all other variables) and colinear with each other. This is essentially an *acid test* of Model 2 to ensure that it is robust.



Model specification

Model 3 has **COVID-19 Case Rate per 100,000 Population** as the primary outcome variable and seven covariates: the primary variable of interest (**Mandatory Mask Use**), **Test Rate per 100,000 Population**, **Percentage of Population Below 25 Years Old**, **Log of Percentage of Black Ethnicity in Total Population + 1**, **Human Mobility Change in Transit Stations**, **Number of Days of Shelter in Place**, and **Number of Days of Non-Essential Businesses Closure**.

Model 3 will demonstrate robustness of **Model 2** ($\tilde{\beta}_1$). New variables on **Model 3** represent other common policies US states have adopted to combat the virus spread. They have some collinearity with mask use as would be expected, since typically States enact a set of policies against COVID-19 concurrently.

Despite losing some explanatory power due to inclusion of the new variables, the result we would like to highlight is that our coeficient of interest ($\tilde{\beta}_1$) continues to be both statistically significant, and with an estimated value having practical significance in terms of informing public policies in the combat to the virus.

Model 3 has the format:

$$\begin{aligned}\text{case rate per } 100,000 \text{ pop.} = \beta_0 + \beta_1(\text{mandatory mask use policy}) + \beta_2(\text{test rate per } 100,000 \text{ pop.}) + \\ \beta_3(\%\text{ pop. } < 25) + \beta_4(\ln(\%\text{ black pop.} + 1)) + \beta_5(\text{mobility } \Delta \text{ transit stations}) + \\ \beta_6(\text{number days shelter in place}) + \beta_7(\text{number days businesses closed})\end{aligned}$$

Model summary

```
model_3 <- lm(case_rate ~ mask_use + test_rate + age_below_25 + log(black_pop + 1) + mob_TS + shelter_da
std_errors = list(
  sqrt(diag(vcovHC(model_1))),
  sqrt(diag(vcovHC(model_2))),
  sqrt(diag(vcovHC(model_3)))
)
stargazer(model_1, model_2, model_3, type = "latex", se = std_errors, title = "Model 3 Summary")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Dec 04, 2020 - 07:32:46 AM

Overall model significance (F-test)

```
anova(model_2, model_3, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: case_rate ~ mask_use + test_rate + age_below_25 + log(black_pop +
##     1) + mob_TS
## Model 2: case_rate ~ mask_use + test_rate + age_below_25 + log(black_pop +
##     1) + mob_TS + shelter_days + bus_close_days
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     45 21307370
## 2     43 21301841  2    5529.1 0.0056 0.9944
```

At the significance level of 0.05, we cannot reject the null hypothesis (H_0 :**Model 2**) in favor of our fuller H_a :**Model 3**, which now includes the covariates **Number of Days of Shelter in Place** and **Number of Days of Non-Essential Businesses Closure**. Our residual standard error remains almost unchanged, even with inclusion of these new variables. This demonstrates collinearity between the variables and the existing ones from **Model 2**. The inclusion of the new variables did not enhance explained variability of the outcome variable. The adjusted R-squared of **Model 3** decreased to 0.616, which would be expected due to the interactions.

As such we have demonstrated that **Model 3** serves as a reliable acid test for the robustness of **Model 2** and the coefficient of its primary explanatory variable pertaining to mask use ($\tilde{\beta}_1$).

Table 3: Model 3 Summary

	Dependent variable:		
	case_rate		
	(1)	(2)	(3)
mask_use	−990.470***	−919.251***	−913.750***
	(324.753)	(227.028)	(271.775)
test_rate	0.018*	0.024**	0.024**
	(0.010)	(0.011)	(0.012)
age_below_25		190.555***	189.940***
		(46.660)	(44.506)
log(black_pop + 1)		461.650***	462.123***
		(110.822)	(119.141)
mob_TS		14.832**	14.663*
		(6.575)	(7.820)
shelter_days			0.079
			(1.628)
bus_close_days			−0.871
			(11.988)
Constant	2,530.239***	−4,603.593***	−4,561.508***
	(501.044)	(1,745.175)	(1,705.048)
Observations	51	51	51
R^2	0.236	0.670	0.670
Adjusted R^2	0.204	0.633	0.616
Residual Std. Error	1,013.835 (df = 48)	688.111 (df = 45)	703.841 (df = 43)
F Statistic	7.416*** (df = 2; 48)	18.279*** (df = 5; 45)	12.481*** (df = 7; 43)
Note:			*p<0.1; **p<0.05; ***p<0.01

Coefficient significance (t-test)

```
coeftest(model_3, vcovHC)
```

```
##
## t test of coefficients:
##
##                       Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)        -4.5615e+03  1.7050e+03 -2.6753 0.0105189 *
## mask_useTRUE       -9.1375e+02  2.7177e+02 -3.3622 0.0016327 **
## test_rate           2.4050e-02  1.2092e-02  1.9889 0.0530957 .
## age_below_25        1.8994e+02  4.4506e+01  4.2678 0.0001066 ***
## log(black_pop + 1)  4.6212e+02  1.1914e+02  3.8788 0.0003550 ***
## mob_TS              1.4663e+01  7.8196e+00  1.8752 0.0675712 .
## shelter_days        7.8936e-02  1.6283e+00  0.0485 0.9615605
## bus_close_days     -8.7101e-01  1.1988e+01 -0.0727 0.9424163
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At 0.05 significance, we can accept an alternative hypotheses: $H_{a1} : \beta_1 \neq 0$, $H_{a3} : \beta_3 \neq 0$, and $H_{a4} : \beta_4 \neq 0$, meaning that only 3 out of 7 of Model 3 covariates have ability to partially explain the variability observed in the case rate per 100,000 population.

The estimate for β_1 (the coefficient for our variable of interest) is $\tilde{\beta}_1 = -913.8$, with a standard error of 271.8 and a p-value of 0.0016. It continues to be statistically significant, with an estimated value that changes little from **Model 2** (-919.3) to **Model 3** (-913.8), indicating that mask use policies are a reliable explanator for differences in case rates between the States.

Practical significance

According to Model 3, states that have adopted mandatory mask use would expect to have 913.8 fewer COVID cases per 100,000 population, all other variables held constant. Given that the median case rate among US states is 2,633 per 100,000 populations, the coefficient estimate has practical significance, with an effect size corresponding to a reduction of 34.7% of the median rate.

CLM Assumptions & Limitations

In theory, our EDA process should have helped us choose the most optimal variables (and subsequent transformations) to use in our models. Nevertheless, we must assess how good our models are at explaining the causal relationship between the dependent variable (case rate per 100,000), and our primary independent variable of interest (State implementation of a mask mandate) or whether they require further modification and optimization. As such we will consider whether they meet the 5 key assumptions required for the classic linear model (CLM). Moreover, we can also utilize CLM assessment techniques to demonstrate how our iterative model building process has optimized our causal model.

1) Independent & Identically Distributed Random Variables

As it is aggregated by State, our data may not be independent and identically distributed:

a. Clustering Effect

States in close proximity to each other may have similar population characteristics. There may also be frequent movements of populations between neighboring States. Moreover, these States may have similar population demographics (ethnicities, ages) or geographical characteristics (e.g. climate, see Omitted Variable Bias section) which lead to a clustering effect in terms of case rates.

b. Strategic Effect

Similar to clustering, socioeconomic and behavioral characteristics of populations may effect public health policies and case rates. Moreover, adjacent States or States with similar population characteristics (and behaviors) may be encouraged to adopt similar public health policies such as implementing shelter-in-place orders, quarantines, mask use mandates and other regulations.

2) Linear Conditional Expectation

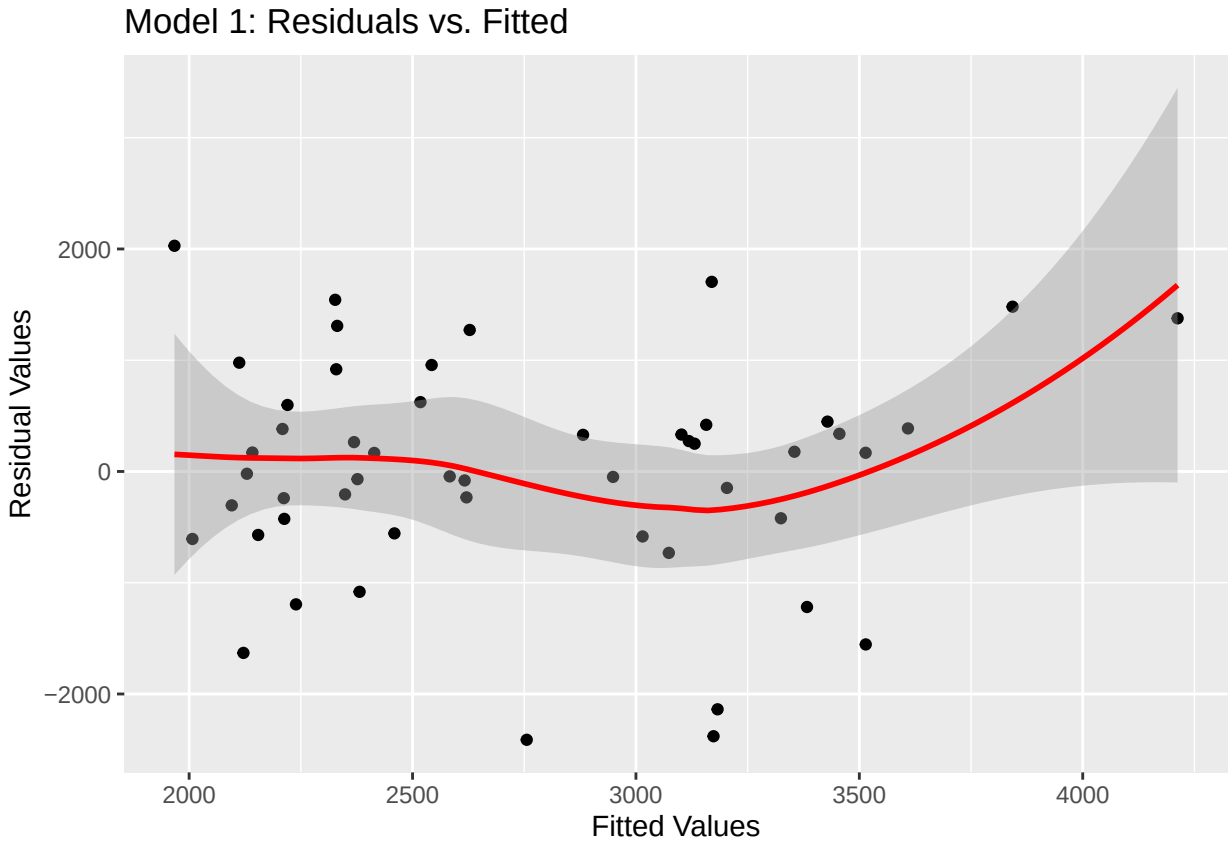
The linear regression model assumes a straight-line relationship between the predictors and the response.

Residuals vs. Fitted

```
model_1_residuals = resid(model_1)
model_2_residuals = resid(model_2)
model_3_residuals = resid(model_3)

model_1_predicteds = predict(model_1)
model_2_predicteds = predict(model_2)
model_3_predicteds = predict(model_3)

plot_1_predicts <- model_1 %>%
  ggplot(aes(model_1_predicteds, model_1_residuals)) +
  geom_point() +
  stat_smooth(color="red") +
  labs(
    title = "Model 1: Residuals vs. Fitted",
    x = "Fitted Values",
    y = "Residual Values"
  )

plot_2_predicts <- model_2 %>%
  ggplot(aes(model_2_predicteds, model_2_residuals)) +
  geom_point() +
  stat_smooth(color="blue") +
  labs(
    title = "Model 2: Residuals vs. Fitted",
    x = "Fitted Values",
    y = "Residual Values"
  )

plot_3_predicts <- model_3 %>%
  ggplot(aes(model_3_predicteds, model_3_residuals)) +
  geom_point() +
  stat_smooth(color="green") +
  labs(
    title = "Model 3: Residuals vs. Fitted",
    x = "Fitted Values",
    y = "Residual Values"
  )
```

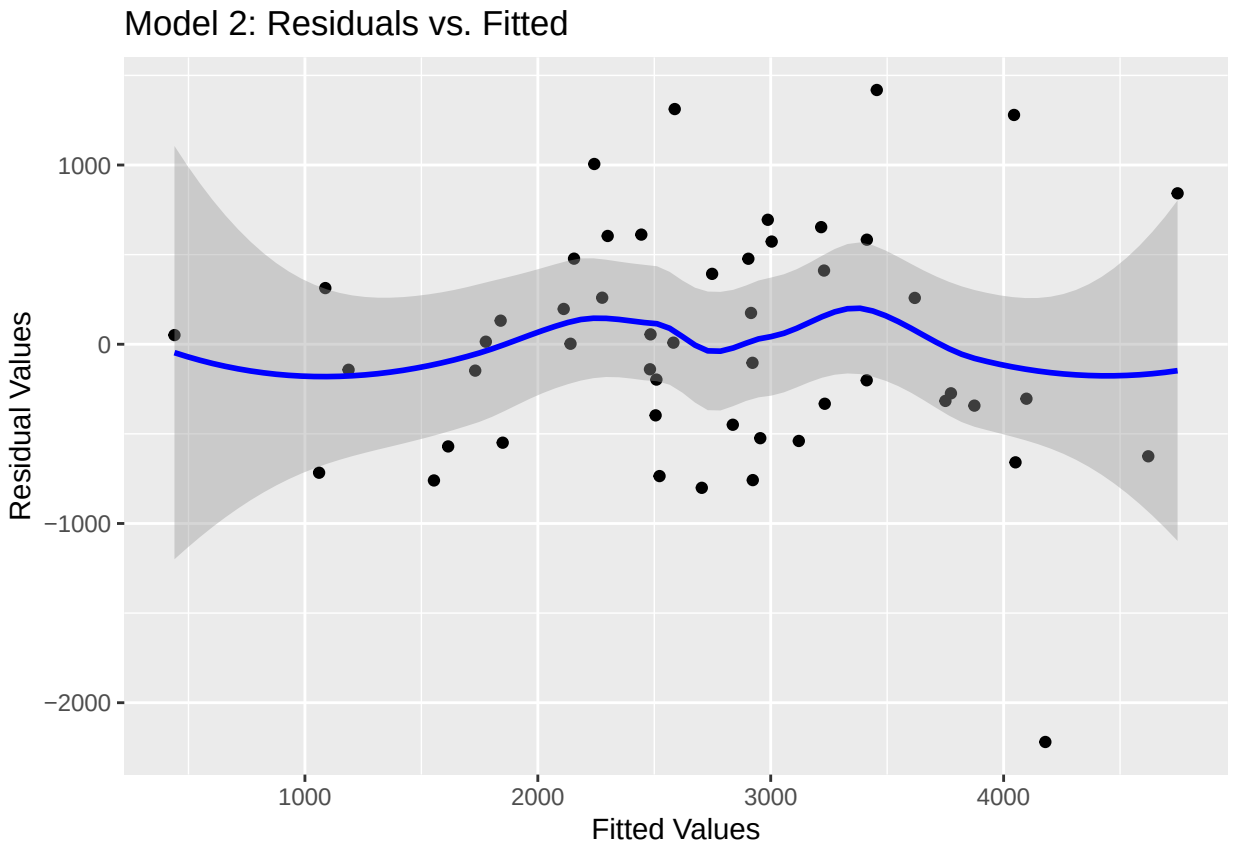
```
plot_1_predicts
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



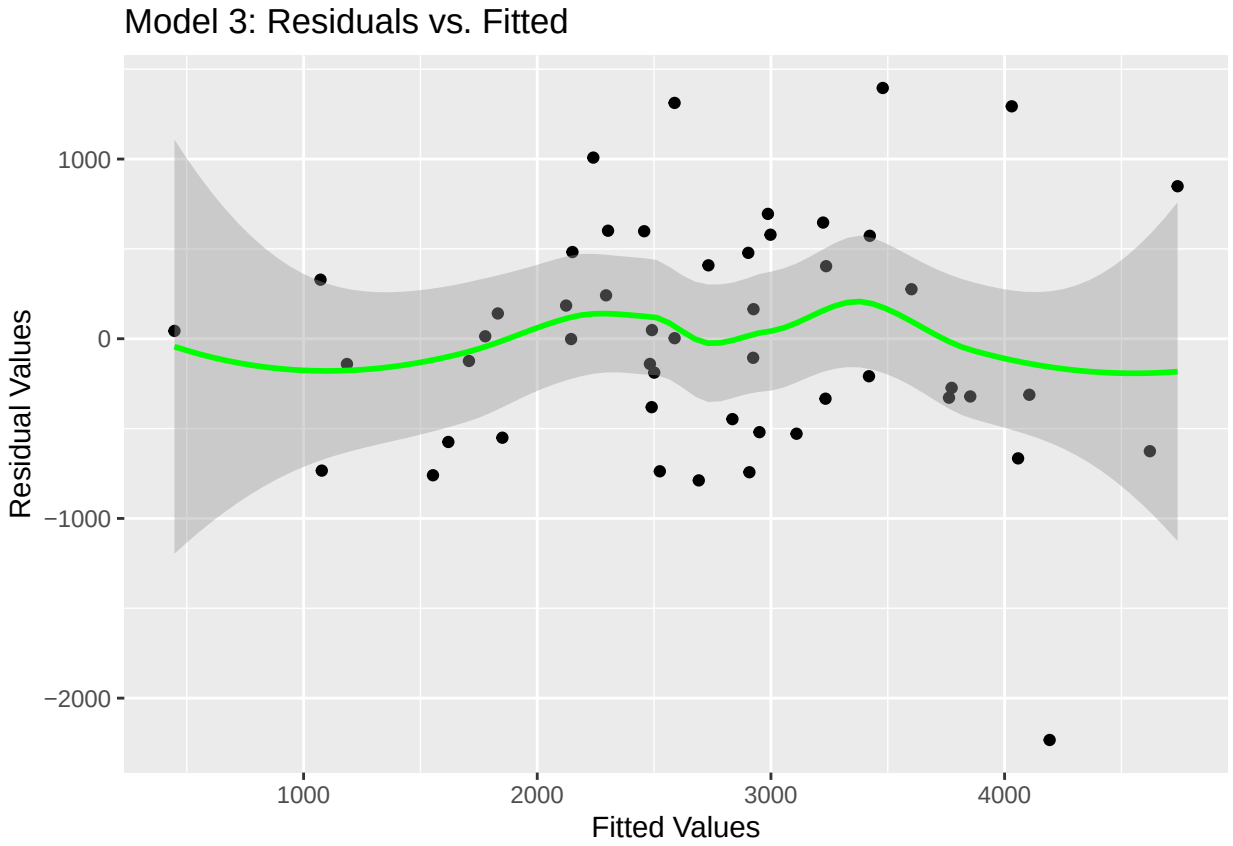

```
plot_2_predicts
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
plot_3_predicts
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Looking at the fitted vs. residuals plot for Model 1, we can see that the residuals demonstrate an element of non-linearity (evidenced especially at fitted values > 3500), indicating problems with the model. In model 2, with the addition of several more control variables, we can see that the plot assumes a more linear pattern, with the line much closer to 0. There appears to be very little difference (almost indiscernable to the naked eye) between the fitted vs. residual line between Model 2 and 3, indicating that Model 3, despite the addition of several more variables, does little to improve the overall model.

Moreover, comparing Model 1 to Model 2, the residuals seem to be more evenly distributed (i.e. randomly about the) about the line in the latter model, with fewer outliers. There is almost no change in residual distribution about the line between Model 2 and 3. We will discuss homoskedasticity in a subsequent section.

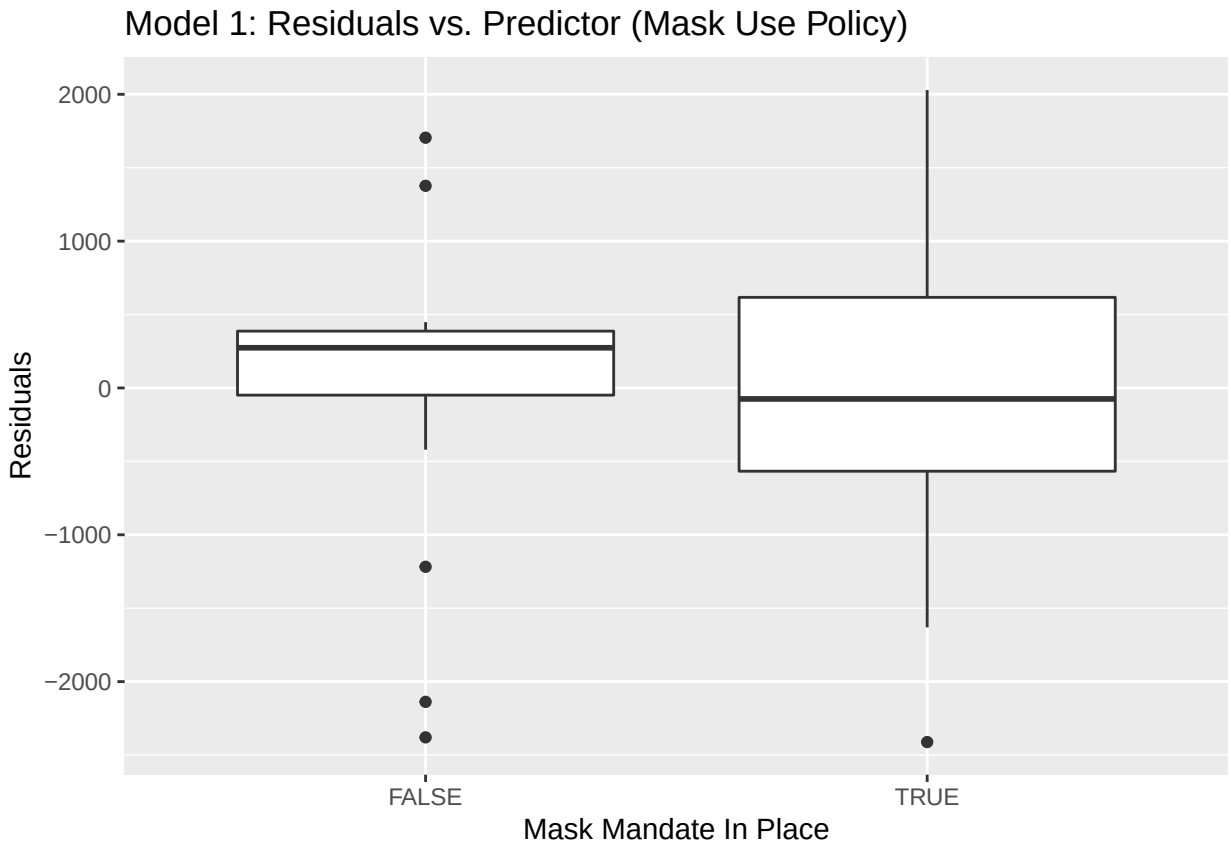
Altogether we can see that compared to Model 1, Model 2 does a better job at meeting the fundamental assumption that the error term has a conditional mean of 0. Moreover, Model 3 does not seem to contribute further to meeting this assumption, despite the addition of further variables.

Residuals vs. Explanatory Variable

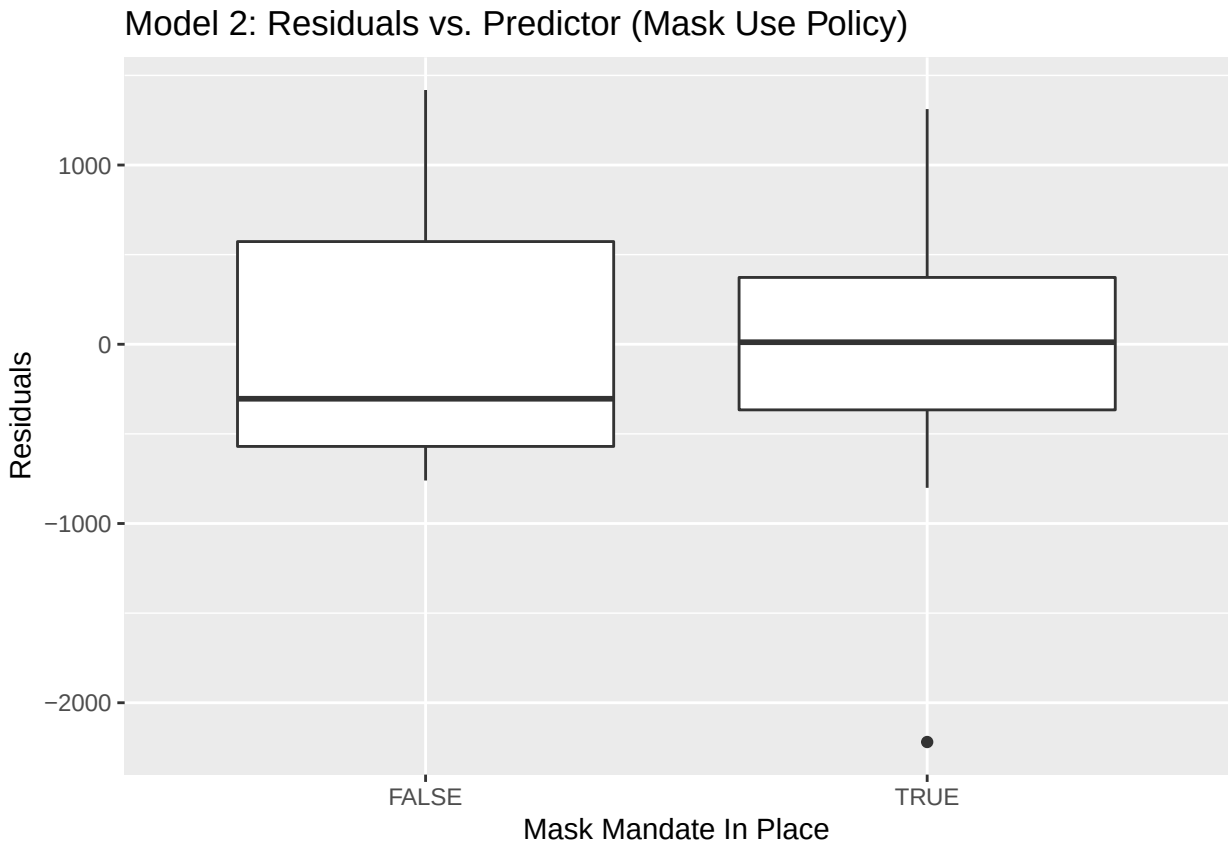
As our main explanatory variable (mask use) is a binary variable rather than a continous one, we cannot easily demonstrate the effect of the changing model upon the predictor values, however we can see whether the distribution of residuals is constant across the binary categories:

```
plot_1_residuals <- model_1 %>%
  ggplot(aes(x = mask_use, y = model_1_residuals)) +
  geom_boxplot() +
  labs(
    title = "Model 1: Residuals vs. Predictor (Mask Use Policy)",
    x = "Mask Mandate In Place",
    y = "Residuals"
  )
```

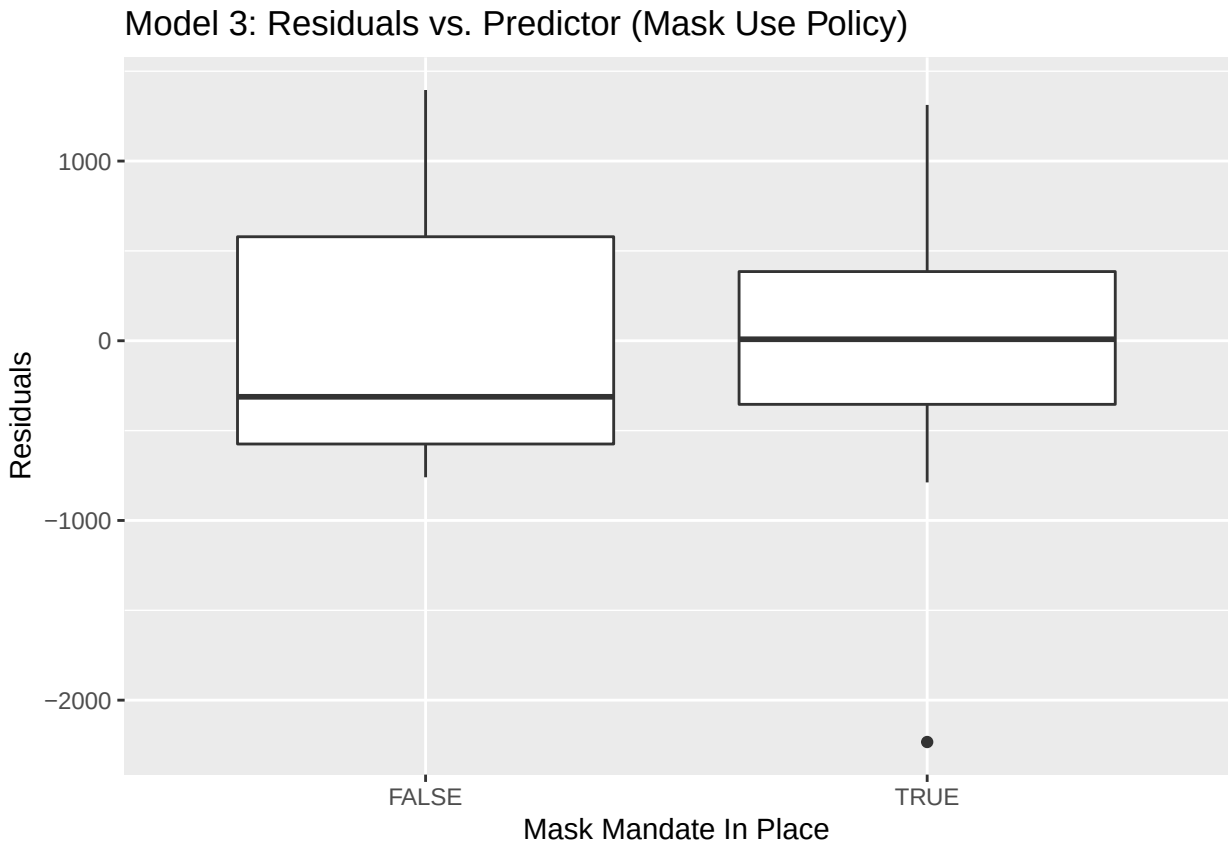
```r
plot_2_residuals <- model_2 %>%
  ggplot(aes(x = mask_use, y = model_2_residuals)) +
  geom_boxplot() +
  labs(
    title = "Model 2: Residuals vs. Predictor (Mask Use Policy)",
    x = "Mask Mandate In Place",
    y = "Residuals"
  )

plot_3_residuals <- model_3 %>%
  ggplot(aes(x = mask_use, y = model_3_residuals)) +
  geom_boxplot() +
  labs(
    title = "Model 3: Residuals vs. Predictor (Mask Use Policy)",
    x = "Mask Mandate In Place",
    y = "Residuals"
  )

plot_1_residuals
```




```r
plot_2_residuals
```




```
plot_3_residuals
```

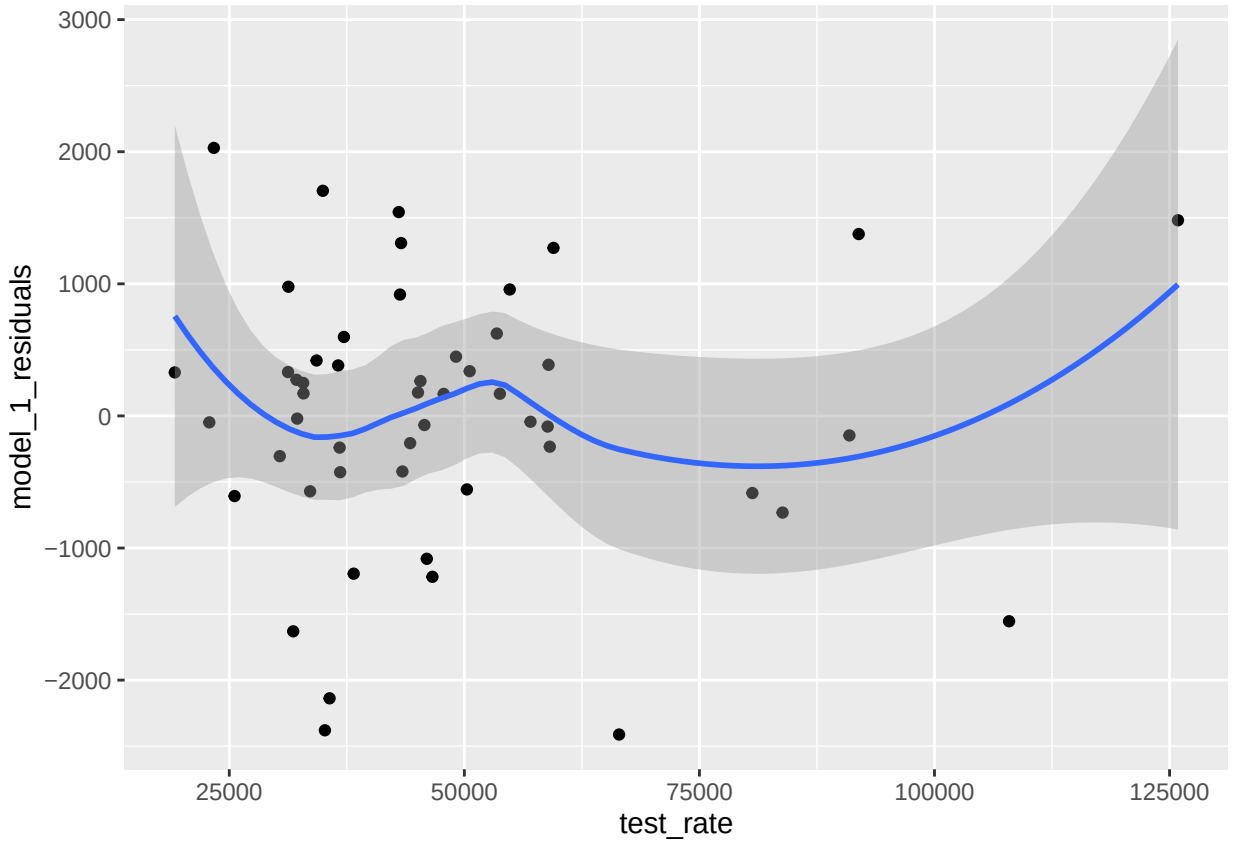


As we can see, in Model 1 there is a large difference in residual spread between the groups. Moreover, the expected value for residuals within the False category deviate substantially from 0, compared to the True category. Moreover, the spread of residual errors seems to be much narrower (and with more outliers) compared to residuals from the True category. By adding more control variables to the model, we can see that the expected value of the residuals for the True category is approximately 0, however the expected value for the False residuals has now become negative. Nevertheless, the spread between the two groups is now more similar. There is almost no discernable change in the expected value of the residuals between Model 2 and Model 3.

?delete

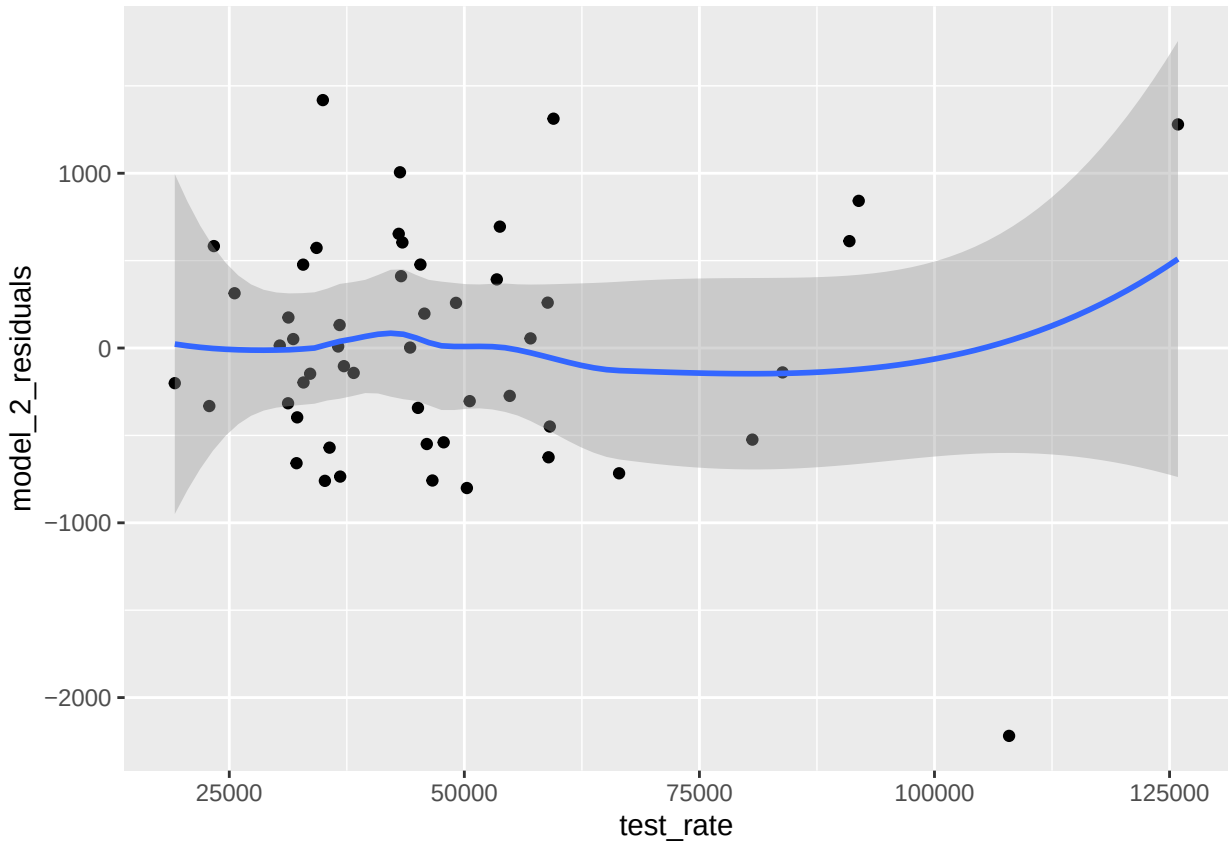
```
plot_1_residuals <- model_1 %>%
  ggplot(aes(x = test_rate, y = model_1_residuals)) +
  geom_point() +
  stat_smooth(se = TRUE)

plot_2_residuals <- model_2 %>%
  ggplot(aes(x = test_rate, y = model_2_residuals)) +
  geom_point() +
  stat_smooth(se = TRUE)

plot_3_residuals <- model_3 %>%
  ggplot(aes(x = test_rate, y = model_3_residuals)) +
  geom_point() +
  stat_smooth(se = TRUE)

plot_1_residuals
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



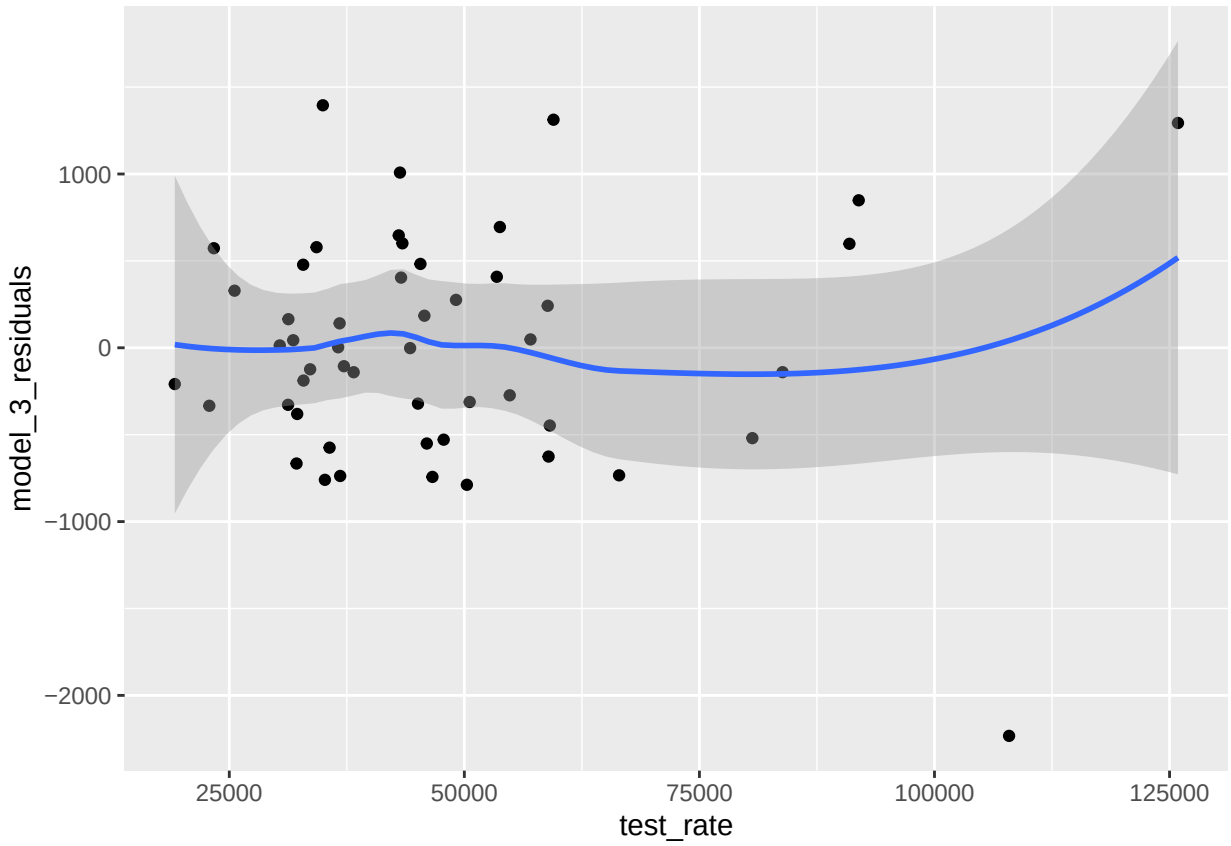

```
plot_2_residuals
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
plot_3_residuals
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



3) No Perfect Collinearity

Dropped Coefficients

We can assess for perfect collinearity by looking to see if there are any dropped coefficients in any of the 3 models:

```
std_errors = list(
  sqrt(diag(vcovHC(model_1))),
  sqrt(diag(vcovHC(model_2))),
  sqrt(diag(vcovHC(model_3)))
)
stargazer(model_1, model_2, model_3, type = "latex", se = std_errors, title = "Model 3 Summary")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Dec 04, 2020 - 07:32:49 AM

As we can see, none of the variables have been dropped from any of the models, indicating that there is no perfect colinearity between any of the independent variables.

Variance Inflation Factor

We can quantify the degree of colinearity between the independent variables by conducting a variance inflation factor (VIF) test. This is the quotient of the variance of a model with multiple terms by by the variance of a model with only one term. It is given by the formula:

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

Table 4: Model 3 Summary

	Dependent variable:		
	case_rate		
	(1)	(2)	(3)
mask_use	−990.470***	−919.251***	−913.750***
	(324.753)	(227.028)	(271.775)
test_rate	0.018*	0.024**	0.024**
	(0.010)	(0.011)	(0.012)
age_below_25		190.555***	189.940***
		(46.660)	(44.506)
log(black_pop + 1)		461.650***	462.123***
		(110.822)	(119.141)
mob_TS		14.832**	14.663*
		(6.575)	(7.820)
shelter_days			0.079
			(1.628)
bus_close_days			−0.871
			(11.988)
Constant	2,530.239***	−4,603.593***	−4,561.508***
	(501.044)	(1,745.175)	(1,705.048)
Observations	51	51	51
R^2	0.236	0.670	0.670
Adjusted R^2	0.204	0.633	0.616
Residual Std. Error	1,013.835 (df = 48)	688.111 (df = 45)	703.841 (df = 43)
F Statistic	7.416*** (df = 2; 48)	18.279*** (df = 5; 45)	12.481*** (df = 7; 43)
Note:			*p<0.1; **p<0.05; ***p<0.01

Where R_i^2 is the coefficient of determnation of of a regression equation with X_i on the left hand side, and all other predictor variables on the right hand side. Subsequently, it will produce a VIF index value for the coefficient estimator of the particular variable we are analyzing (VIF($\hat{\beta}_i$)). According to several sources, VIF($\hat{\beta}_i$) > 10 (some also say values >5 indicate high colinearity) is considered indicative of multicolinearity.

```
car::vif(model_1)
```

```
##  mask_use test_rate
##  1.039238  1.039238
```

The VIF values for both variables in Model 1 are <2, indicating no high degree of colinearity between the two independent variables.

```
car::vif(model_2)
```

```
##           mask_use          test_rate       age_below_25 log(black_pop + 1)
##           1.185361           1.193469           1.143912           1.120042
##             mob_TS
##           1.485480
```

The VIF values for all variables in Model 2 are <2, indicating no high degree of multicolinearity between the 5 independent variables.

```
car::vif(model_3)
```

```
##           mask_use          test_rate       age_below_25 log(black_pop + 1)
##           1.326697           1.244233           1.203190           1.147285
##             mob_TS       shelter_days     bus_close_days
##           1.751725           1.409882           1.579966
```

The VIF values for all variables in Model 3 are <2, indicating no high degree of multicolinearity between the 7 independent variables.

From the two tests above, we can safely say our models meet the assumption of having no substantial amount of colinearity or multicolinearity between the independent variables.

4) Homoscedastic Errors

There are two methods we can employ to test for homocedasticity of the error terms:

Scale-Location Plots

This is a method to visually assess for homoscedasticity of the error terms.

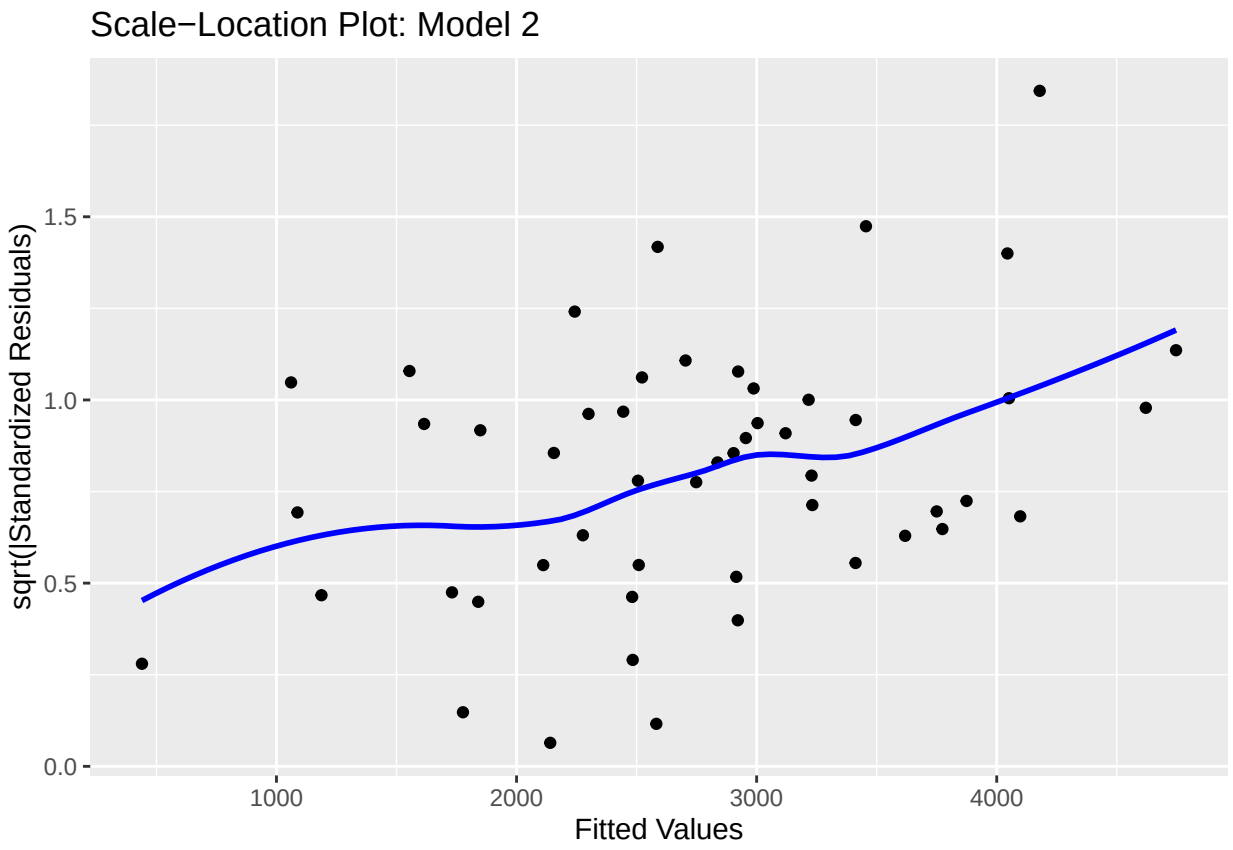
```
plot_1_sl <- model_1 %>%
  ggplot(aes(x = model_1_predicteds, y = sqrt(abs(model_1_residuals/sd(model_1_residuals))))) +
  geom_point() +
  stat_smooth(color="red", se=FALSE) +
  labs(
    title = "Scale-Location Plot: Model 1",
    x = "Fitted Values",
    y = "sqrt(|Standardized Residuals)"
```

```
  )

plot_2_sl <- model_2 %>%
  ggplot(aes(x = model_2_predicteds, y = sqrt(abs(model_2_residuals/sd(model_2_residuals))))) +
  geom_point() +
  stat_smooth(color="blue", se=FALSE) +
  labs(
    title = "Scale-Location Plot: Model 2",
    x = "Fitted Values",
    y = "sqrt(|Standardized Residuals)"
  )

plot_3_sl <- model_3 %>%
  ggplot(aes(x = model_3_predicteds, y = sqrt(abs(model_3_residuals/sd(model_3_residuals))))) +
  geom_point() +
  stat_smooth(color="green", se=FALSE) +
  labs(
    title = "Scale-Location Plot: Model 3",
    x = "Fitted Values",
    y = "sqrt(|Standardized Residuals)"
  )

plot_1_sl
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

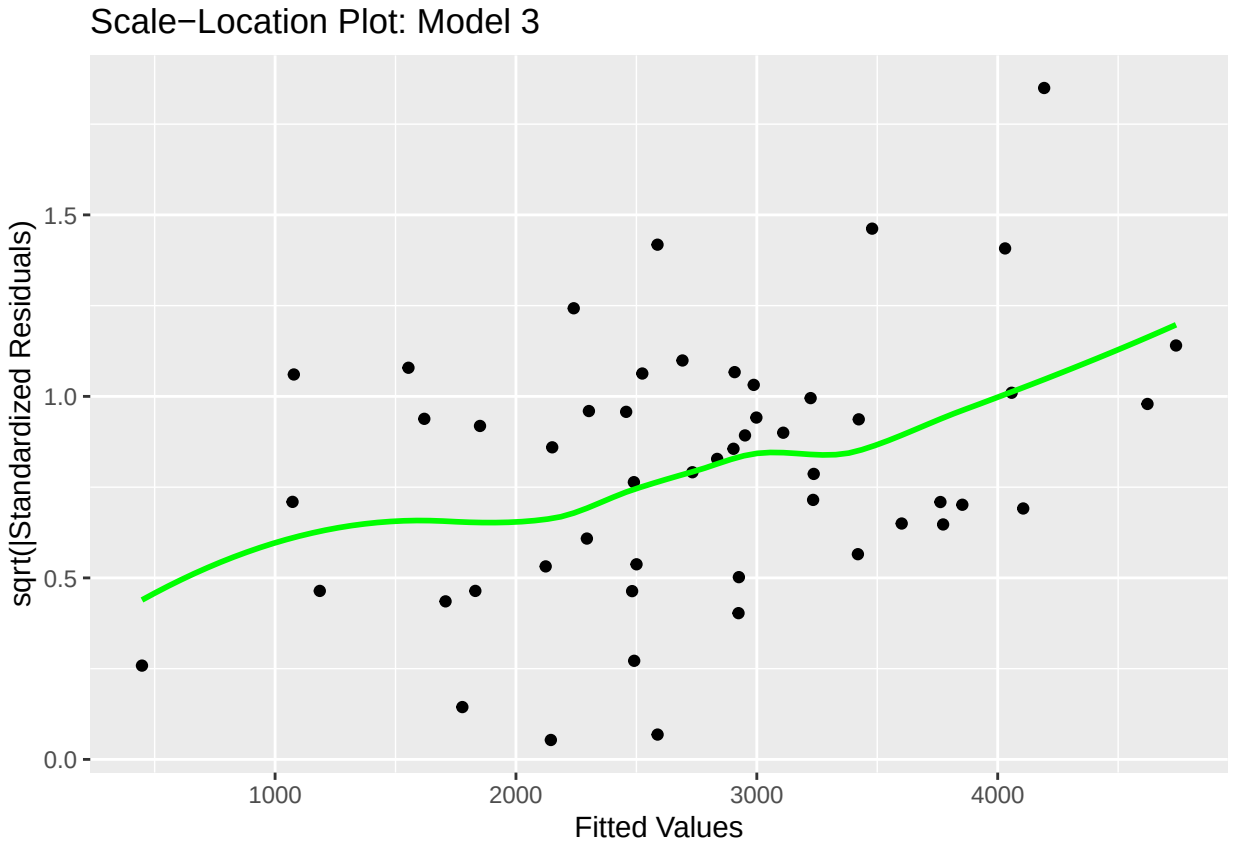
```
plot_2_sl
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
plot_3_sl
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



In Model 1, it seems like there are a greater concentration of points below the line, which demonstrates evident curvature at the lower and upper X-scale, with a greater upward slope at Fitted values > 3500. It is difficult to test for heteroscedasticity in this plot. In Model 2, the errors seem much more evenly distributed above and below the line, which is also straighter than that in Model 1. However it seems that the error magnitude is increasing as X increases. There is almost no discernable change in any of the aforementioned parameters between model 2 and 3, indicating that the extra added variables do nothing to increase the efficacy of our model.

As it is difficult to discern homoscedasticity in the scale-location plot for model 1, and it appears errors are increasing in magnitude in both model 2 and 3, we can also perform a quantitative assessment for non-constant variance in the form of the Breusch-Pagan test. Our null hypothesis is that there is no evidence for heteroscedastic variance.

Breusch-Pagan Test

Running the test for Model 1, we observe a high p-value of 0.7. While we cannot assertively state that there is no heteroscedastic variance, we can safely assume that we fail to reject the null hypothesis:

```
lmtest::bptest(model_1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_1
## BP = 0.71226, df = 2, p-value = 0.7004
```

For Model 2, we now observe a very low p-value of ~0.008, and reject the null hypothesis. Although this indicates that our data is heteroscedastic, it is important to note that we are utilizing Robust Standard Errors versus Classical Standard Errors in not just Model 2, but all of our models and their associated tests:

```
lmtest::bptest(model_2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_2
## BP = 17.365, df = 5, p-value = 0.003857
```

Finally in Model 3, running the BP test yields a low p-value of ~0.02, and we reject the null hypothesis. This again indicates that our data is heteroscedastic, but as mentioned previously we solve for this by utilizing Robust Standard Errors in our model and associated tests:

```
lmtest::bptest(model_3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_3
## BP = 19.723, df = 7, p-value = 0.0062
```

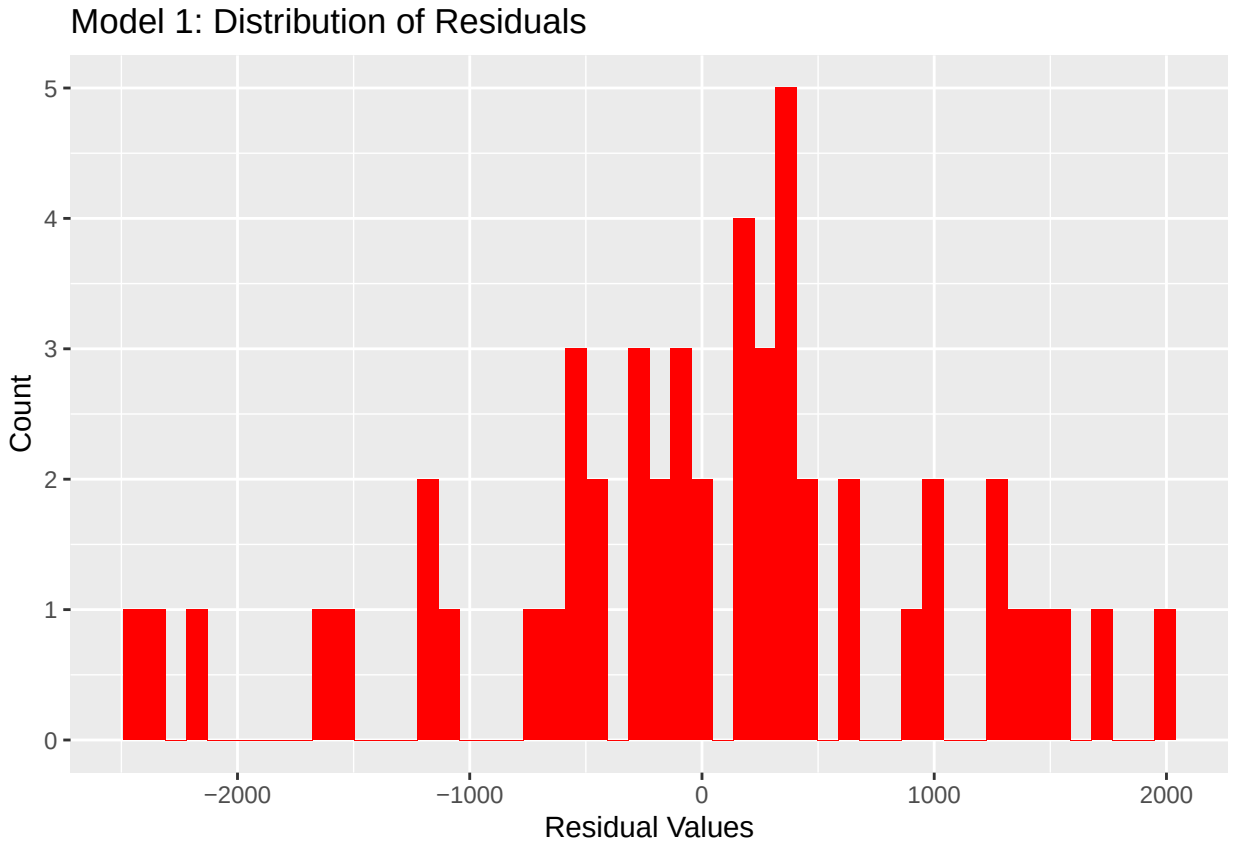
5) Normally Distributed Errors

The final CLM assumption we must test for is normally distributed errors. We can utilize both quantile-quantile (Q-Q) plots and histograms of residuals for each of our models to examine whether the residuals are normally distributed:

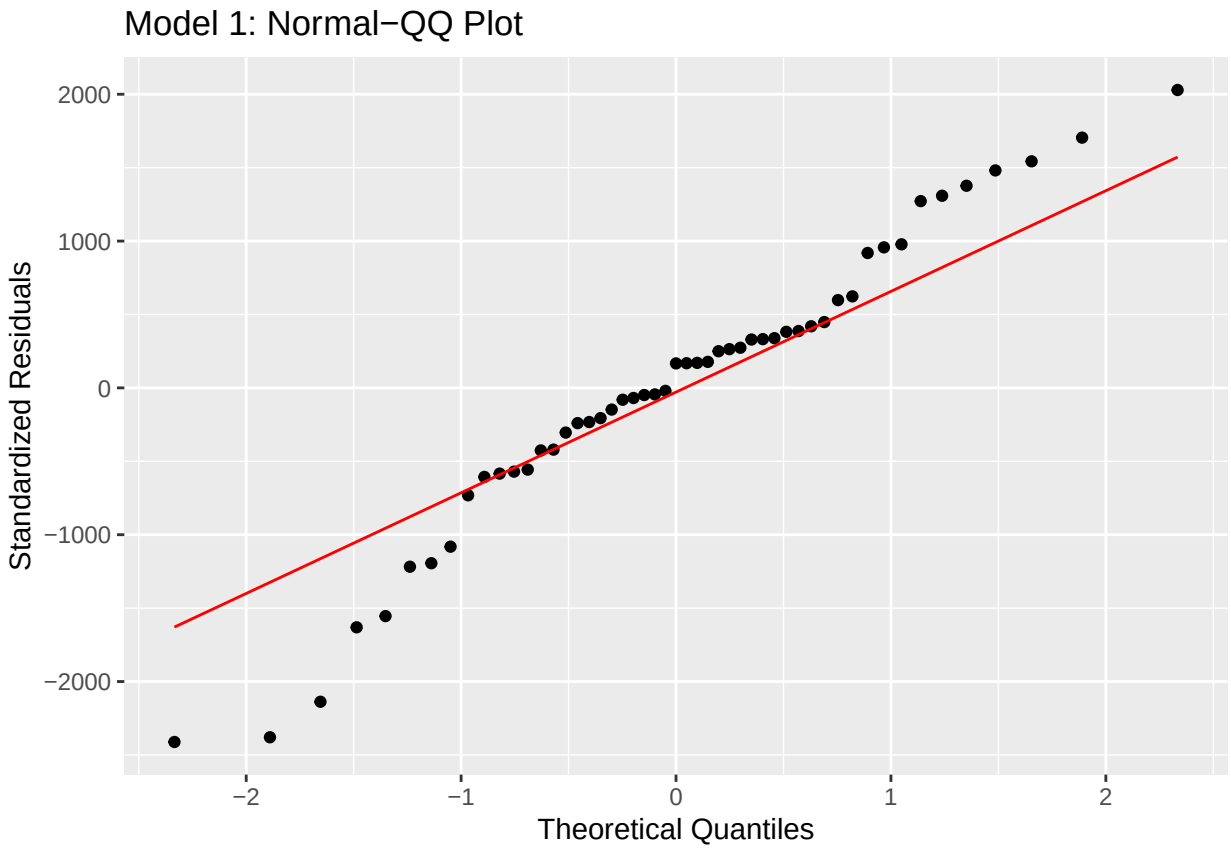
```
mod_1_hist <- model_1 %>%
  ggplot(aes(x = model_1_residuals)) +
  geom_histogram(fill="red", bins=50) +
  labs(
    title = "Model 1: Distribution of Residuals",
    x = "Residual Values",
    y = "Count"
  )

mod_1_qq <- model_1 %>%
  ggplot(aes(sample = model_1_residuals)) +
  stat_qq() + stat_qq_line(color="red") +
  labs(
    title = "Model 1: Normal-QQ Plot",
    x = "Theoretical Quantiles",
    y = "Standardized Residuals"
  )

mod_2_hist <- model_2 %>%
  ggplot(aes(x = model_2_residuals)) +
  geom_histogram(fill="blue", bins=50) +
  labs(
    title = "Model 2: Distribution of Residuals",
    x = "Residual Values",
    y = "Count"
```

```
  )

mod_2_qq <- model_2 %>%
  ggplot(aes(sample = model_2_residuals)) +
  stat_qq() + stat_qq_line(color="blue") +
  labs(
    title = "Model 2: Normal-QQ Plot",
    x = "Theoretical Quantiles",
    y = "Standardized Residuals"
  )

mod_3_hist <- model_3 %>%
  ggplot(aes(x = model_3_residuals)) +
  geom_histogram(fill="green", bins=50) +
  labs(
    title = "Model 3: Distribution of Residuals",
    x = "Residual Values",
    y = "Count"
  )

mod_3_qq <- model_3 %>%
  ggplot(aes(sample = model_3_residuals)) +
  stat_qq() + stat_qq_line(color="green") +
  labs(
    title = "Model 3: Normal-QQ Plot",
    x = "Theoretical Quantiles",
    y = "Standardized Residuals"
  )
```

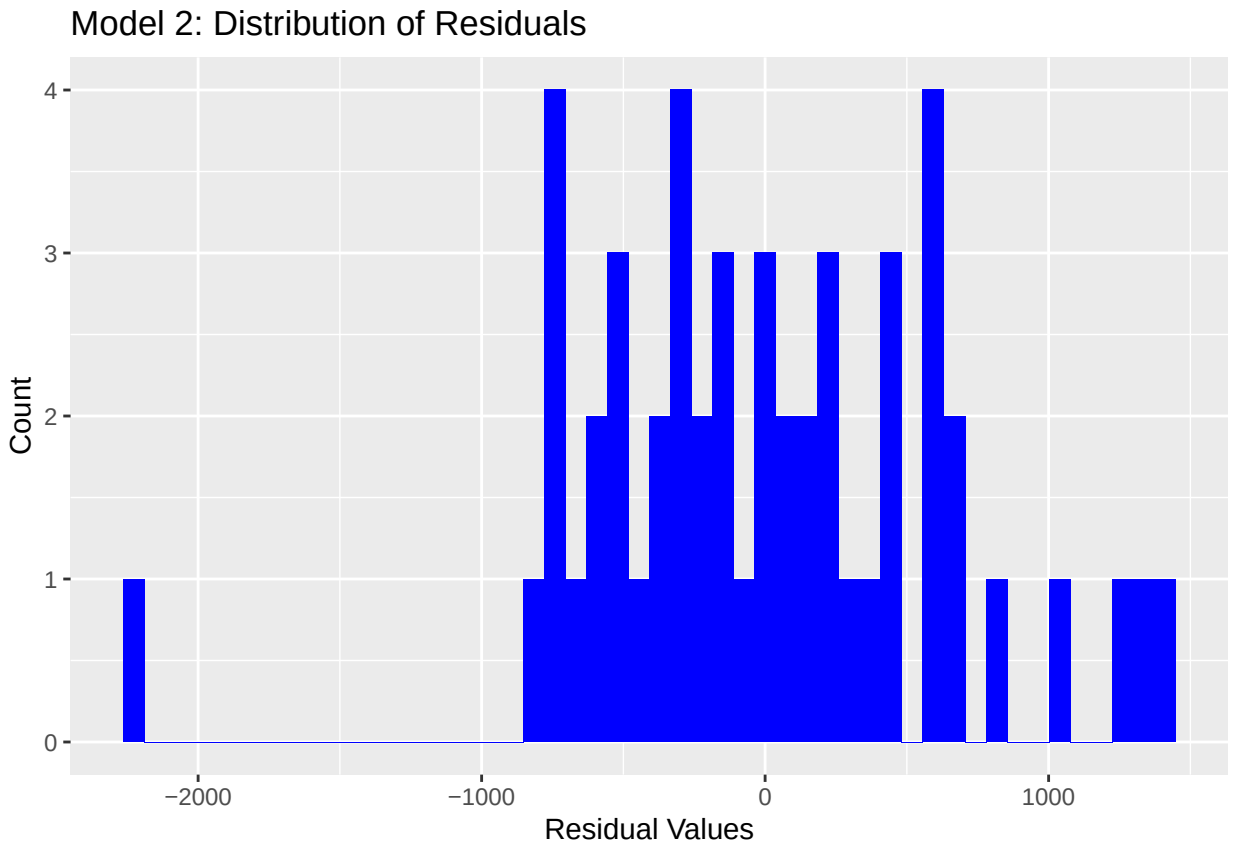
```
mod_1_hist
```


Model 1: Distribution of Residuals
Count
Residual Values
0
1
2
3
4
5
−2000
−1000
0
1000
2000

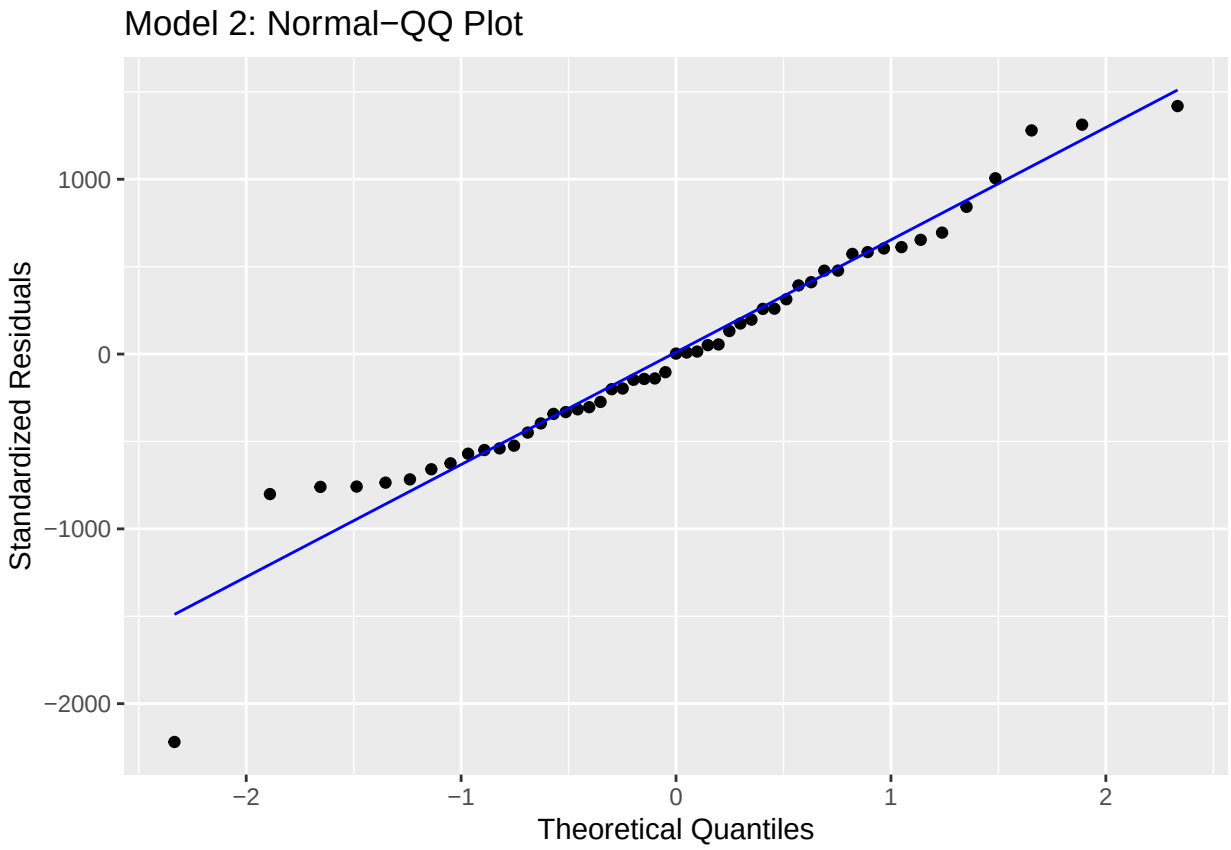
```
mod_1_qq
```


Model 1: Normal−QQ Plot
Standardized Residuals
2000
1000
0
−1000
−2000
−2
−1
0
1
2
Theoretical Quantiles

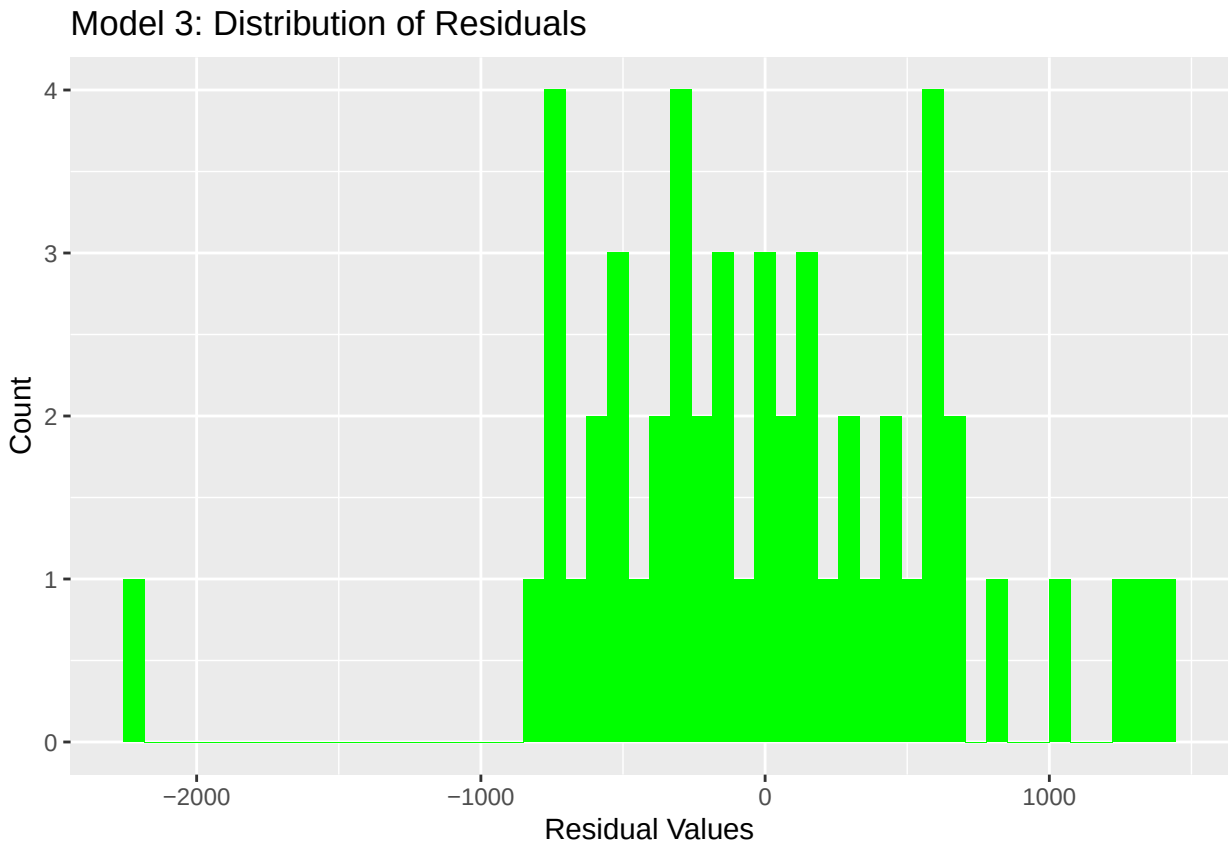
```
mod_2_hist
```


Model 2: Distribution of Residuals
Count
Residual Values
0
1
2
3
4
−2000
−1000
0
1000

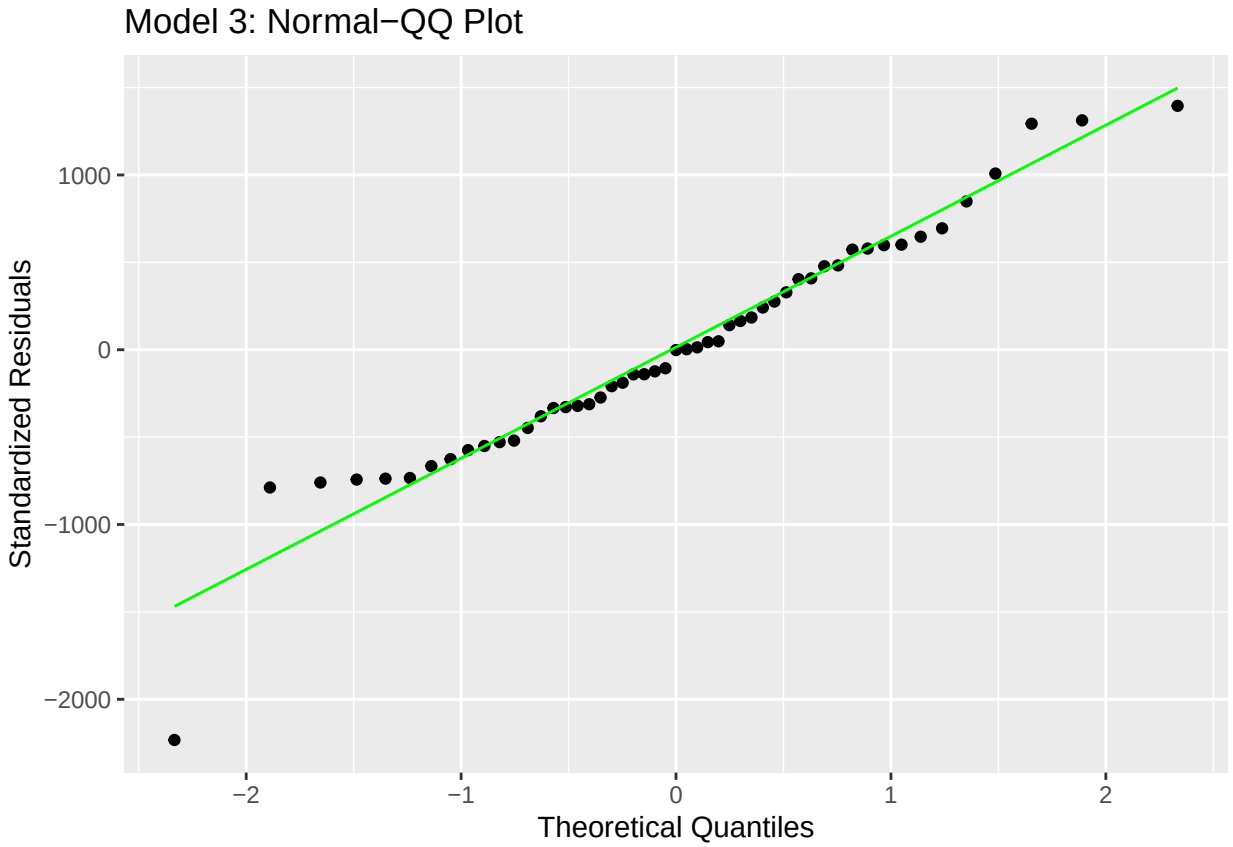
```
mod_2_qq
```


Model 2: Normal−QQ Plot
Standardized Residuals
1000
0
−1000
−2000
−2
−1
0
1
2
Theoretical Quantiles

```
mod_3_hist
```


Model 3: Distribution of Residuals
Count
Residual Values
0
1
2
3
4
−2000
−1000
0
1000

```
mod_3_qq
```



As we can see, in model 1 even though the distribution of residuals represented by the histogram seems somewhat evenly distributed about 0, with a slight negative skew, the associated Q-Q plot shows that the points fall very far from the line at the theoretical quantile values -1 and 1, indicating a heavy-tailed distribution of residuals. In comparison, it is difficult to make the claim that the residuals are normally distributed in Models 2 and 3 (which are almost identical), with an extreme negative outlier obvious in both. This is reflected in the Q-Q plot for these models, however with the caveat that the points are generally more well behaved and congruent with the Q-Q line versus that of Model 1. Nevertheless, Models 2 and 3 also suffer from a heavy-tailed residual distribution, albeit on that is uni-directional rather than the bi-directional heavy tails of Model 1's residual distribution.

Omitted Variables Discussion

```
quar <- read.csv("ovb.csv")

quar <- data.frame(quar)

df2 <- merge(df, quar, by="state")
```

As we are attempting to study an exceedingly-complex real life phenomenon, naturally there will be an element of bias in our models. We selected what we believed to be the most appropriate explanatory variables influencing the COVID-19 case rate per 100,000, which included a range of logistical, social, ethnic and population-based metrics (as provided in the dataset). Nevertheless, even though the dataset was extensive, it is impossible to completely predict every factor that might influence both our dependent variable and independent variables - this is reflected in the model performance metrics (for example R^2, which would be 1 if our model perfectly represented the real world phenomenon we were trying to model) and error parameter.

As such, we have considered 5 potential omitted variables that we predict might exert a hidden effect upon the regression model, specifically upon the dependent variable (i.e. cases per 100,000) and the primary independent variable of interest (i.e. whether the state implemented a mask use policy). These are all envisaged real-world phenomena. For 2 of these, we can use proxy variables as provided in the dataset or data from external sources for bias estimation. We can predict the potential relationship between the omitted variables and the dependent and independent variables, and the direction of bias for all of our suggested omitted variables. We cannot estimate the size of the bias for the omitted variables that we do not have a proxy or direct information for, however.

For our 'best' model (model 2), the coefficient for Mandatory Mask Use is ~-919, with a S.E. of of ~227 and a p-value of < 0.01. This can be interpreted to mean that *ceteris paribus*, a state that enforces mandatory mask use laws reduces the overall case rate by ~919/100,000 or ~1%. We will subsequently discuss how we predict the omitted variables could influence this coefficient and what it might mean.

For the 3 omitted variables that we either have proxy (Travel restrictions, % Republicans) or direct data (Average annual temperature) for, we will calculate the omitted variable bias using the base model usng the following system:

$$\tilde{Y} = \tilde{\beta_0} + \tilde{\beta_1} X_1 + \epsilon_1$$

This specifies the biased estimation. Where $\tilde{Y}$ is the dependent variable (i.e. positive cases per 100,000 population), $\tilde{\beta_0}$ is the y-intercept term, $\tilde{\beta_1}$ the coefficient of the independent variable of interest (i.e. whether the state implemented a mask use policy, X_1) and ϵ_1 the associated error term.

By including the omitted variable, the theoretically 'true' or unbiased estimator becomes:

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X_1 + \hat{\beta_2} X_2 + \epsilon_2$$

Where $\hat{Y}$ is the dependent variable (i.e. positive cases per 100,000 population), $\hat{\beta_0}$ is the y-intercept term, $\hat{\beta_1}$ the unbiased coefficient of the independent variable of interest (i.e. whether the state implemented a mask use policy, X_1), $\hat{\beta_2}$ the unbiased coefficient of the omitted variable X_2, and ϵ_2 the modified associated error term.

Subsequently, $\tilde{\beta_1} - \hat{\beta_1}$ is used to calculate the magnitude and direction of the actual omitted variable bias.

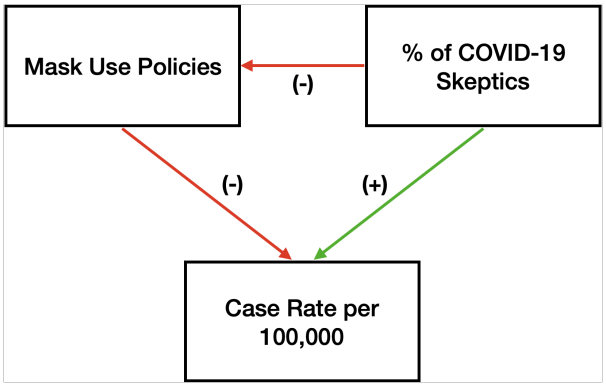
Skepticism Towards COVID-19 Control Policies

There is a documented segment of the population that are skeptical towards the existence of COVID-19[1]. They may view subsequent governmental efforts to control its spread as a conspiracy. In the provided dataset, we do not have any data pertaining to the proportion of people in any State that are COVID-19 skeptics, however potential surrogate measures could relate to whether the State had elected Republican politicians (see $1, 4$). Subsequently, we predict that the more skeptics there are in a State, the less likely they will be to observe preventative public-health measures and will thus contract COVID-19 at a higher rate. Therefore, we predict a positive relationship between the State population of COVID-19 skeptics and the case rate per 100,000.

Similarly, the more COVID-19 skeptics there are among a State's poulation, the more likely it is that their elected local government will align themselves with their views and resist enforcement of mandatory mask usage policies. We therefore predict a negative relationship between the population of COVID-19 skeptics in a State and State enforcement of mandatory mask laws.

As the relationship between the omitted variable and the dependent variable is positive, while the relationship between the omitted variable and the explanatory variable in question is negative, the overall effect of the omitted variable bias will be negative. If we could gather information about a State's population of COVID-19 skeptics and add this variable to our model, doing so would cause the coefficient for mandatory mask use

to increase, or move towards 0. In other words, the effect-reinforcing influence of a hypothetical 'skeptics' omitted variable on the dependent variable would offset the potential effect-mitigating impact of a mask use policy on positive case rates, ultimately leading to a smaller reduction (i.e. >-919) in positive cases attributed to a mandatory mask use policy.

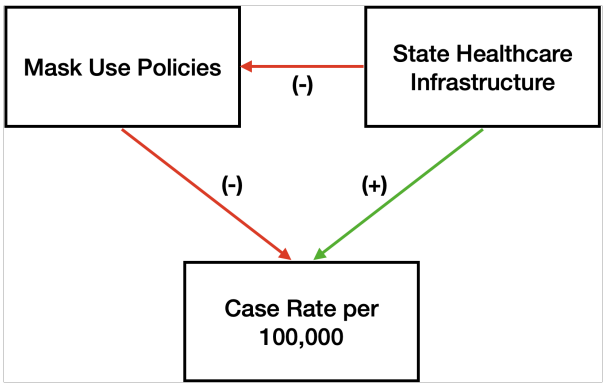


State Healthcare Infrastructure

The standard of healthcare infrastructure may vary across States. For example, there may be fewer healthcare facilities in largely rural States, compared to those that are more urban. As a result, populations in States with poorer healthcare infrastructure may not have easy access to COVID testing or diagnosis. A possible proxy variable that we do have access to could be looking at a State's population density, however specific transformations would likely have to be performed to somehow translate it to a State's healthcare infrastrucutre (e.g. doctors per X unit of population). If we could measure and quantify this State healthcare infrastructure metric, we predict that it would be postively related to COVID-19 case rate per 100,000 population. That is, the higher the theoretical "State healthcare infrastructure" score, the higher number of COVID-19 tests performed, translating to a higher positive case rate.

In theory, governments of States with poor healthcare infrastructure might be worried about their population being unable to access treatment for COVID-19, and might be more likely to enact efforts to prevent its spread among the population, which could overwhelm a under-funded/under-resourced healthcare system (even if this hasn't been the observed trend in the real world). For the purposes of analysis, however, we therefore predict a negative relationship between a theoretical "State healthcare infrastructure" score and enforcement of laws requiring use of face masks.

As the relationship between the omitted variable and the independent variable is positive, and the relationship between the omitted variable and and dependent variable is negative, overall omitted variable bias effect is predicted to be negative. If we could compute a "State healthcare infrastructure" metric and include it in our model its addition would cause the mandatory mask use coefficient to increase, or move towards 0. In other words, by taking into account the effect of healthcare infrastructure on the regression system, we would expect to see a reduction in purported benefit of implementing mask use policy as demonstrated by our 'best' model: This would mean a mask use coefficient >-919.

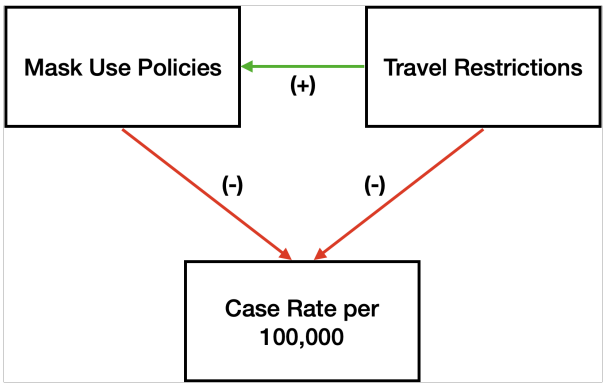


Travel Restrictions

Certain states have imposed travel restrictions to prevent the spread of COVID-19[2]. These may vary from recommending visitors to quarantine, to requiring them to provide proof of a negative COVID-19 test before being granted entry. The State-mandated quarantine variables are the closest potential proxies, and we can potentially use these to calculate the effect of the omitted variable bias for this variable. We predict that these measures to prevent the spread of COVID-19 are negatively related to the case rate per 100,000 population. That is, States that have enforced some kind of travel restriction have fewer overall cases.

Similarly, if a State government is ready to impose travel restrictions, they are also likely to enforce other public health measures such as mandatory wearing of face masks in public places. Therefore we predict a positive relationship between State enforcement of travel restrictions and implementation of mandatory face mask policies.

As the relationship between the omitted variable and the dependent variable is negative, and the relationship between the omitted variable and the independent variable is positive, we predict an overall negative omitted variable bias on the model. By adding a variable pertaining to State enforcement of travel restrictions to the model, we would expect to see an increase in the mandatory mask use coefficient. As it is negative in the existing model, we would expect to see it move towards 0. In other words, the amount of effect (i.e. reduction in number of positive cases) that can be attributed to the mask use policy in our 'best' model will be reduced (i.e. smaller reduction in number of positive cases) due to the effect of quarantine measures and the coefficient is expected to be >-919.



We prove our omitted variable bias predictions in this instance, by using a proxy from the 'COVID-19 US State Policy Dataset' specifically pertaining to whether a state required all visitors entering from another state to quarantine and adding it to a regression equation between the main variable of interest (mask use policies) and the dependent variable (positive cases per 100,000). Due to the potentially complicated effects exerted upon the dependent and independent variables in a model with >1 variable, we will only use the dependent and independent variables of interest (plus the omitted variable in the unbiased estimation) to demonstrate omitted variable bias:

```r
base_model <- lm(case_rate ~ mask_use, data = df2)
model_quar <- lm(case_rate ~ mask_use + state_quarantine, data = df2)

stargazer(base_model, model_quar, type = "latex", title = "Omitted Variable Bias Comparison: Travel Res
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Dec 04, 2020 - 07:32:53 AM

Table 5: Omitted Variable Bias Comparison: Travel Restrictions

	Dependent variable:	
	case_rate	
	(1)	(2)
mask_use	−830.000**	−808.456**
	(319.744)	(308.088)
state_quarantine		−732.492**
		(333.254)
Constant	3,302.765***	3,475.116***
	(261.070)	(263.369)
Observations	51	51
R^2	0.121	0.201
Adjusted R^2	0.103	0.168
Residual Std. Error	1,076.417 (df = 49)	1,036.653 (df = 48)
F Statistic	6.738** (df = 1; 49)	6.048*** (df = 2; 48)
Note:		*p<0.1; **p<0.05; ***p<0.01

As we can see here, the overall effect of including quarantine information upon our variables of interest is concordant with predictions, and we observe an increase (i.e. moving towards 0) in the overall coefficient value for our independent variable of interest (mask use policies) meaning that lesser effect is attributed to it in a more 'true' system. We can further calculate and prove that our predictions of negative omitted variable bias were correct by subtracting the value of the coefficient of interest in the unbiased estmation (i.e. $\hat{\beta}_1$) from that of the 'false' model (i.e. $\tilde{\beta}_1$):

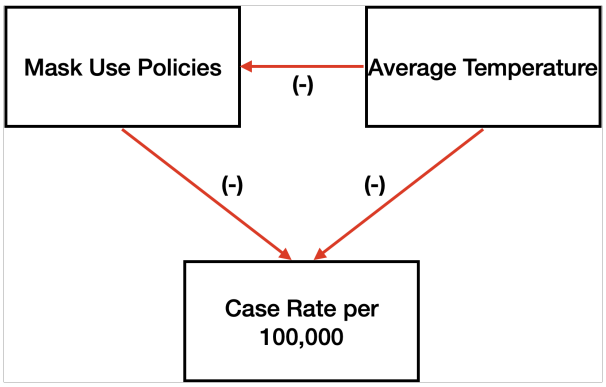
$$\tilde{\beta}_1 - \hat{\beta}_1 = -830.0 - (-808.5) \approx -21.5$$

Average Temperature

It is thought that the dry air occuring in cold weather enhances the spread of the flu virus[3]. Certain States in the U.S. have cooler average climates than others, and may also experience colder winters. We therefore predict that there is a negative relationship between a State's average temperature and the positive case rate per 100,000 population. As acquiring annual State temperature data is relatively easy, we have compiled this information from external sources, to be utilized to prove our predictions.

Similarly, a local government is likely aware of the link between cooler temperatures and viral spread, and are therefore more likely to enact mandatory mask use policies to mitigate this phenomenon. We therefore predict that the lower a State's average temperature, the more likely the government will be to enforce mandatory mask use policies, a negative relationship.

As the relationship between the omitted variable and both the dependent and independent variables is negative, the overall effect of the average temperature omitted variable will be positive. That is, if we add an average temperature variable to our model, we would expect to see a decrease in the coefficient for mandatory mask use. As the mandatory mask use coefficient is already negative, we would expect it to become more negative or move further away from 0. In other words, by including State temperature data we would expect to see an increase in the purported benefit of mask use policies (i.e. a coefficient <-961, or moving further away from 0 in the negative direction) in reducing the number of positive COVID-19 cases.



We prove our omitted variable bias prediction in this instance, by using externally-sourced data pertaining to a State's average annual temperature and adding it to a regression equation between the main variable of interest (mask use policies) and the dependent variable (positive cases per 100,000). Due to the potentially complicating effects exerted upon the dependent and independent variables in a model with >1 variable, we will only use the dependent and independent variables of interest (plus the omitted variable in the 'true' model) to demonstrate omitted variable bias:

```
model_climate <- lm(case_rate ~ mask_use + temp, data = df2)

stargazer(base_model, model_climate, type = "latex", title = "Omitted Variable Bias Comparison: Average
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Dec 04, 2020 - 07:32:53 AM

As we can see here, the overall effect of including average annual State temperature information upon our variables is concordant with predictions, and we observe an increase (i.e. moving towards 0) in the overall coefficient value for our independent variable of interest (mask use policies) meaning that lesser effect is attributed to it in a more 'true' system. We can further calculate and prove that our predictions of negative omitted variable bias were correct by subtracting the value of the coefficient of interest in the unbiased estmation (i.e. $\hat{\beta}_1$) from that of the 'false' model (i.e. $\tilde{\beta}_1$):

$$\tilde{\beta}_1 - \hat{\beta}_1 \approx -830.0 - (-775.0) \approx -55.0$$

Percentage of Republican Voters

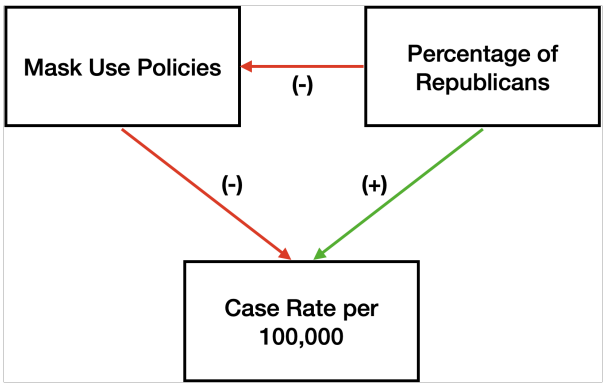
There is evidence to show that States that are majority Republican are experiencing the majority of new COVID-19 infections. Though the reasons for this are likely complex, it has been demonstrated, for example, that Republican voters are less likely to observe social distancing regulations, relative to Democratic voters[4]. We have information provided on the State's ruling officials, which we can take as a proxy for the majority political inclination. We predict that there is a positive relationship between proportion of Republican-voting population in a State and the positive case rate per 100,000 people.

Table 6: Omitted Variable Bias Comparison: Average Annual Temperature

	Dependent variable:	
	case_rate	
	(1)	(2)
mask_use	−830.000**	−775.464**
	(319.744)	(325.062)
temp		17.153
		(17.933)
Constant	3,302.765***	2,374.398**
	(261.070)	(1,005.178)
Observations	51	51
R^2	0.121	0.137
Adjusted R^2	0.103	0.101
Residual Std. Error	1,076.417 (df = 49)	1,077.354 (df = 48)
F Statistic	6.738** (df = 1; 49)	3.821** (df = 2; 48)
Note:		*p<0.1; **p<0.05; ***p<0.01

Similarly, States with a majority of Republican-leaning voters are likely to elect Republican officials who will pander to their voter base. We therefore predict that States with a higher percentage of Republican voters, will be less likely to enforce mandatory mask use laws i.e. a negative relationship.

As there is a positive relationship between the omitted variable (% of Republican voters) and the dependent variable, and a negative relationship between the omitted variable and the independent variable, the overall omitted variable bias will be negative. If we include the proportion of Republican voters as a variable in our model, we would expect to see the coefficient for mandatory mask use to increase or move towards 0. In other words, by including information regarding the proportion of Republican affiliates in a State, we would expect to see an decrease in the purported benefit of implementing mask use policites (i.e. the coefficient moves towards 0, or becomes >-961) in reducing the number of positive cases.



We can actually calculate the omitted variable bias in this instance, by using the political affiliation of the elected officials as proxies for majority political inclination in the State. States with Republican officials are represented by the binary variable 1, and Democrats 0. Due to the potentially complicating effects exerted upon the dependent and independent variables in a model with >1 variable, we will only use the dependent and independent variables of interest (plus the omitted variable in the 'true' model) to demonstrate omitted variable bias:

```
model_party <- lm(case_rate ~ mask_use + political_party, data = df2)

stargazer(base_model, model_party, type = "latex", title = "Omitted Variable Bias Comparison: Republica
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Dec 04, 2020 - 07:32:53 AM

Table 7: Omitted Variable Bias Comparison: Republican Majority

	Dependent variable:	
	case_rate	
	(1)	(2)
mask_use	−830.000**	−688.018*
	(319.744)	(357.526)
political_party		301.713
		(337.144)
Constant	3,302.765***	3,054.295***
	(261.070)	(381.476)
Observations	51	51
R^2	0.121	0.135
Adjusted R^2	0.103	0.099
Residual Std. Error	1,076.417 (df = 49)	1,078.611 (df = 48)
F Statistic	6.738** (df = 1; 49)	3.756** (df = 2; 48)
Note:		*p<0.1; **p<0.05; ***p<0.01

As we can see here, the overall effect upon the variables of interest caused by including information about a State's political inclination is concordant with predictions. We observe an increase (i.e. moving towards 0) in the overall coefficient value for our independent variable of interest (mask use policies) meaning that lesser effect is attributed to it in a more 'true' system. We can further calculate and prove that our predictions of negative omitted variable bias were correct by subtracting the value of the coefficient of interest in the unbiased estmation (i.e. $\hat{\beta}_1$) from that of the 'false' model (i.e. $\tilde{\beta}_1$):

$$\tilde{\beta}_1 - \hat{\beta}_1 \approx -830.0 - (-688.0) \approx -142.0$$

Conclusion

COVID-19 is not a disease to take lightly. As the death toll and case rate continues to mount, we must collectively work together to slow the spread. Our modeling above concludes that the implementation of a mandatory face mask policy is effective in reducing the COVID-19 case rate by upwards of 900 cases per 100,000 residents within the United States.

Our exploratory data analysis helps inform how we chose the variables we did, ranging from age to race to mobility data elements. Leveraging those variables, our second and most optimal model achieved an adjusted R^2 value of nearly 60%. Further controlling for other policy variables such as shelter-in-place and closure of businesses did not erode the adjusted R^2 value very much (57%), and the residual standard error remained nearly constant. As demonstrated in the progression of our modeling, the statistical significance

and practical significance remain robust even with the addition of several other variables. They also meet the assumptions of the CLM. Finally, we chose a healthy variety of omitted variables to assess potential bias on our primary model coefficients.

The next few months will be crucial for the country to course correct until the widespread availability and administration of a safe vaccine. We have already lost far too many of our fellow citizens to this disease.

Wear a mask. Slow the spread.

References

1. Whatley, Z., Shodiya, T., **Why So Many Americans Are Skeptical of a Coronavirus Vaccine**, https://www.scientificamerican.com/article/why-so-many-americans-are-skeptical-of-a-coronavirus-vaccine/

2. **Thinking of Traveling in the U.S.? These States Have Travel Restrictions.**, https://www.nytimes.com/2020/07/10/travel/state-travel-restrictions.html

3. Lowen, AC., Steel, J., **Roles of Humidity and Temperature in Shaping Influenza Seasonality**, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4097773/

4. Gollwitzer A. et al., **Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic**, https://www.nature.com/articles/s41562-020-00977-7