Descriptive Analysis of COVID-19 Spread and Mortality

Code

Introduction

High density urban areas such as Wuhan, Madrid and New York City have seen devastating impacts from COVID-19. However, Singapore, a country with very high population density has had low COVID mortality, and Peru, a country with a significantly lower population density has currently one of the highest COVID mortality rates. To further investigate the relationship between population distribution and COVID-19 mortality rates, we will analyze data from the United States, a country with a mixture of urban, suburban and rural populations. We wish to answer a single question.

Is there a relationship between population density and deaths from COVID-19?

The covid_19 dataset was compiled by Majid Maki-Nayeri in October 2020. It draws many variables from the COVID-19 US state policy database (Raifman J, Nocka K, Jones D, Bor J, Lipson S, Jay J, and Chan P.). There are 51 observations in the dataset covering each state and the District of Columbia. We will operationalize the several fields in the data to represent mortality, population density, poverty, and community mobility. For mortality, we will use a single variable representing total statewide deaths from COVID per 100 thousand inhabitants. For population density, we will use one variable representing the statewide average of number of inhabitants per square mile. For poverty, we will use one variable representing the statewide percentage of residents who live below the federal poverty line. For community mobility, we will use four variables representing retail and recreational mobility, grocery and pharmacy mobility, transit mobility, and workplace mobility. Exploration and operationalization of these variables is laid out in greater detail in their corresponding sections of this report.

The aim of the report is to determine the relationship between population density and mortality rate. We attempt this by creating three models. The first is a minimalist regression comparing population density and COVID-19 mortality. The second expands upon the first by including statewide poverty rate into consideration. The third takes a maximalist approach, expanding upon the second by including the wide range of community mobility variables discussed earlier.

As a result of our limited data and limited knowledge of causal order, we don’t aim to make claims of causation between population density, COVID-19 mortality, or any of the other variables used to construct this model. This model was nonetheless constructed with the understanding that, despite only being a descriptive investigation, the relationships described within it likely correspond to meaningful relationships in the real world. For this reason, we punctuate our analysis with discussion of ommitted variables and model limitations as a way to guide areas of future study.

Team

John Andrus
Jenny Pyon
Prathyusha Charagondla
Sharon Wu

Model Generation

We developed our model by starting with a narrow scope of study and broadening that scope upon each iteration to include a greater number of potentially significant variables. This section lays out the R packages and libraries used to generate our model, a exploratory analysis of each of the variables used, and the methodology by which our three models were created.

Packages and Libraries

Our model and report were generated on the UC Berkeley R datahub. All code is in written in the R programming language and executes within an R notebook. A full listing of the libraries used is included below.

- tidyverse
- plotly
- ggplot2
- dplyr
- sandwich
- lmtest
- MASS
- stargazer

Hide

```
install.packages("grid")
```

```
There were 21 warnings (use warnings() to see them)
```

Hide

```
library(tidyverse)
library(plotly)
library(ggplot2)
library(dplyr)
library(tidyverse)
library(ggplot2)
library(sandwich)
library(lmtest)
library(MASS)
library(stargazer)
```

Data

The dataset used to generate our model is the Covid-19 State and County Policy Orders dataset provided by the US Department of Health and Human Services and compiled by Virtual Student Federal Service Interns Raifman, Nocka, Jones, Bor, Lipson, Jay, and Chan. This resource compiles information from the fifty states and Washington D.C. about COVID-related states of emergency, stay at home orders, face mask mandates, quarantine mandates, vote by mail policies, and many other state policies enacted in responce to the COVID-19 Pandemic.

Hide

```
load("covid_19.RData")
covid_all <- covid_19_edit
```

Model 1

The first, and simplest version of our model includes only the relationship between the key variables, population density and COVID mortality rates. This model is essential for understanding the relationship between population density and COVID deaths.

Mortality Rate

Operationalization

We will operationalize the concept of mortality rate by using the variable `deaths_per_100k`. This variable represents the statewide total of COVID-19 deaths normalized to a population of 100 thousand individuals. We believe that the `deaths_per_100k` realistically represents the concept we are trying to measure because it is calculated on the state level, represents a population averaged value, and is expressed in common units that are easily interpretable by most readers.

Summary

The distribution of `deaths_per_100k` can be observed in the histogram below. Mortality rates are being used for this analysis in order to compare states with different population values. This field ranges from 9 deaths per 100k in the state of Vermont and 368 deaths per 100k in New York. The distribution appears fairly normal with most states grouping near the median rate of 47 deaths per 100k population. This grouping includes states such as Missouri, New Mexico, Tennessee, and Iowa, to name a few. The mean is slightly higher at approximately 63 deaths per 100k population with a standard deviation of 56.54, indicating a rightward skew. This skew is driven in part by New York's relatively high mortality rate which represents the single major outlier in this category. Although an outlier, this data reflects New York as an early (first half of 2020) epicenter for COVID-19, so the analysis will not remove New York as a data point but will eventually perform a log transform to help meeting CLM assumption of normally distributed errors.

Summary of Mortality Rate

Field	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
Mortality Rate	9.00	33.50	47.00	63.08	75.50	368.00

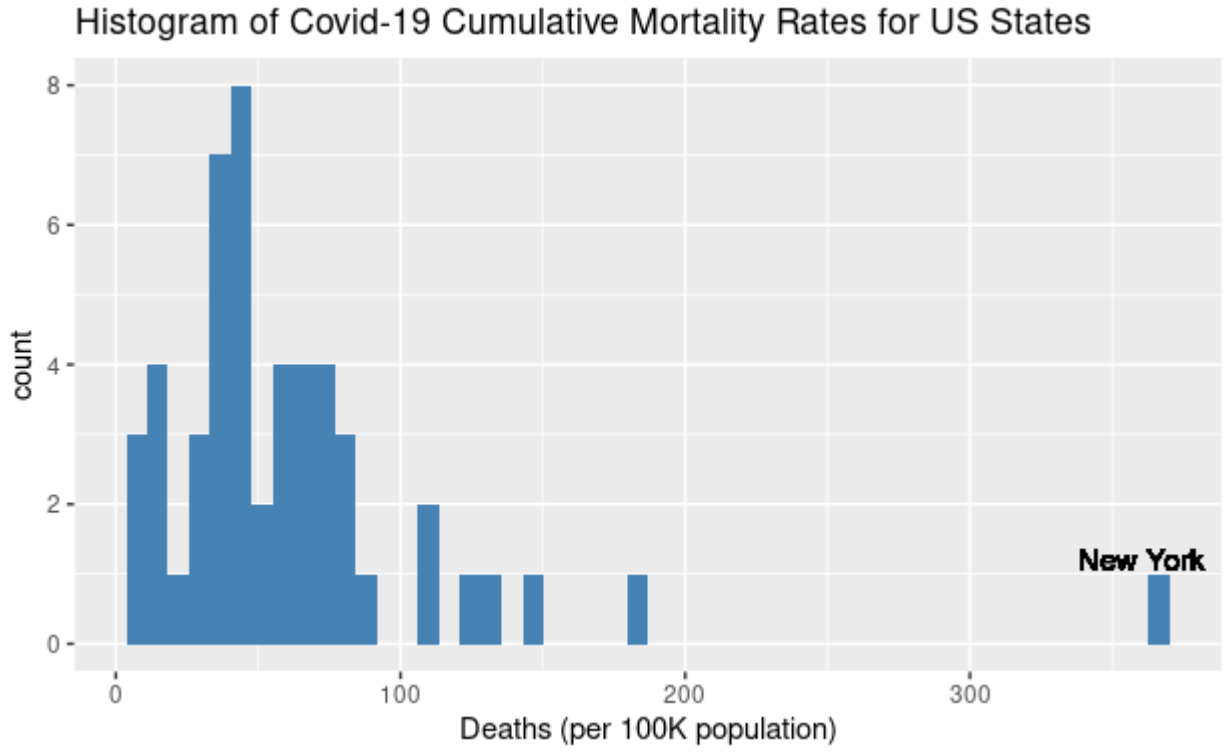
Hide

```
ggplot(data = covid_all,
```

```
There were 50 or more warnings (use warnings() to see the first 50)
```

Hide

```
       mapping = aes(x= deaths_per_100K)) +
       geom_histogram(bins = 50, fill = 'steelblue') + ggtitle("Histogram of Covid-19 Cumulativ
e Mortality Rates for US States") +
       labs(x = "Deaths (per 100K population)") +  geom_text(label="New York", x = 360, y = 1.2
)
```



Population Density

Operationalization

We will operationalize the concept of population density by using the variable `pop_density`. This variable represents the statewide average of number of inhabitants per square mile. We believe that the `pop_density` realistically represents the concept we are trying to measure because it is calculated on the state level, represents the area density of number of inhabitants, and is expressed in common units that are easily interpretable by most readers.

Summary

This field ranges from 1.11 per mi^2 in Alaska to 11500 per mi^2 in The District of Columbia (Washington D.C). This field has considerable right skew, with a median value of 93.24 and a mean value of 392.64, likely as a result of Washington D.C. being the single significant outlier in this field. For this reason, two histograms are provided, the first including the value for Washington D.C., the second ommitting. We are using observations of US States for this analysis. Our analysis will omit Wanshington DC as it is the nation's capital, a major city but not a state.

Hide

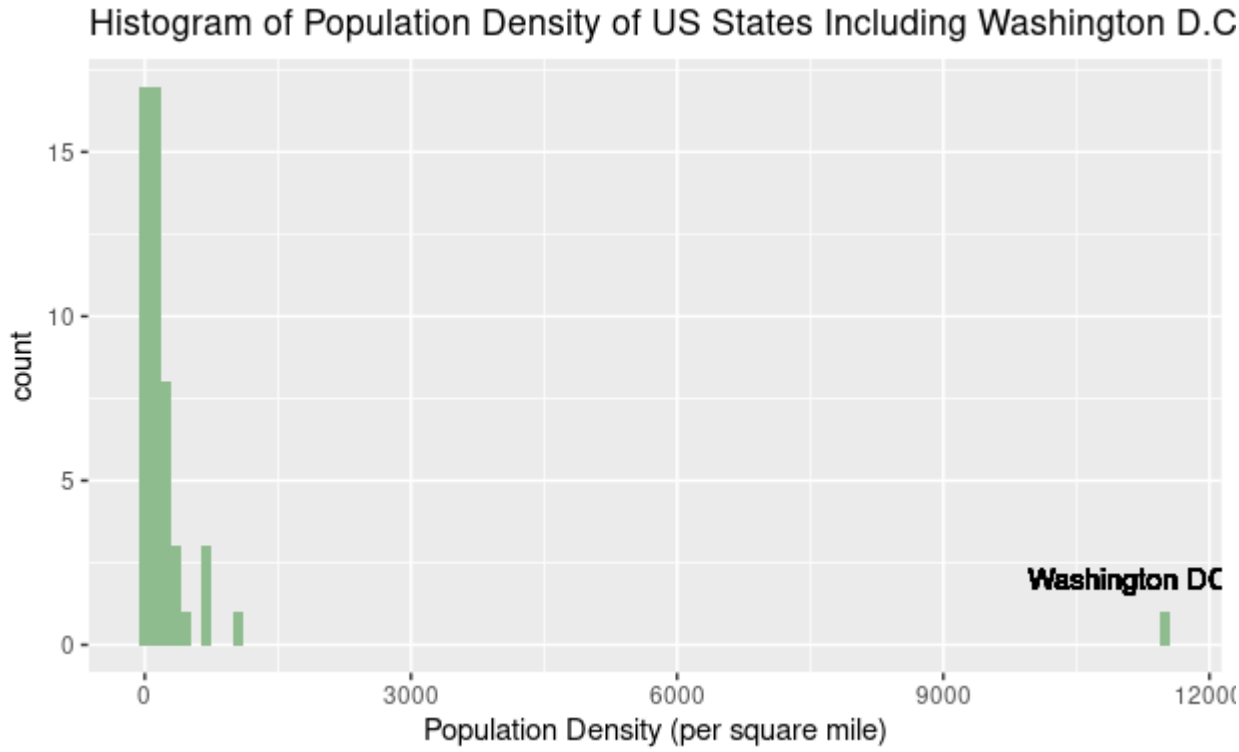
```
summary(covid_all$pop_density)
```

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   1.11    48.66    93.24   392.64   209.56 11496.81
```

Field	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
Population Density	1.11	48.66	93.24	392.64	209.56	11496.81

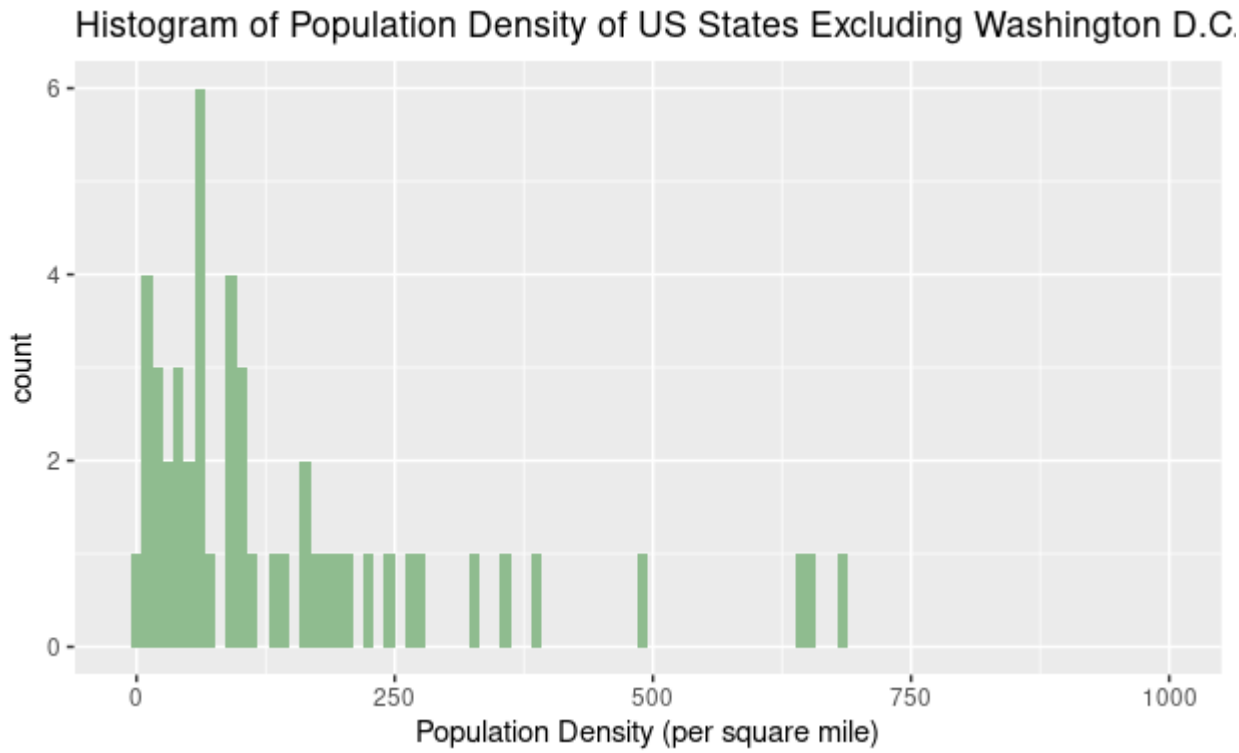
Hide

```
ggplot(data = covid_all,
      mapping = aes(x= pop_density)) +
      geom_histogram(bins = 100, fill = "darkseagreen") +
      ggtitle("Histogram of Population Density of US States Including Washington D.C.") +
      labs(x = "Population Density (per square mile)") +  geom_text(label="Washington DC", x = 1
1100, y = 2)
```



Hide

```
ggplot(data = covid_all,
      mapping = aes(x= pop_density)) +
      geom_histogram(bins = 100, fill = "darkseagreen") +xlim(-10, 1000) +
      ggtitle("Histogram of Population Density of US States Excluding Washington D.C.") +
      labs(x = "Population Density (per square mile)")
```



Hide

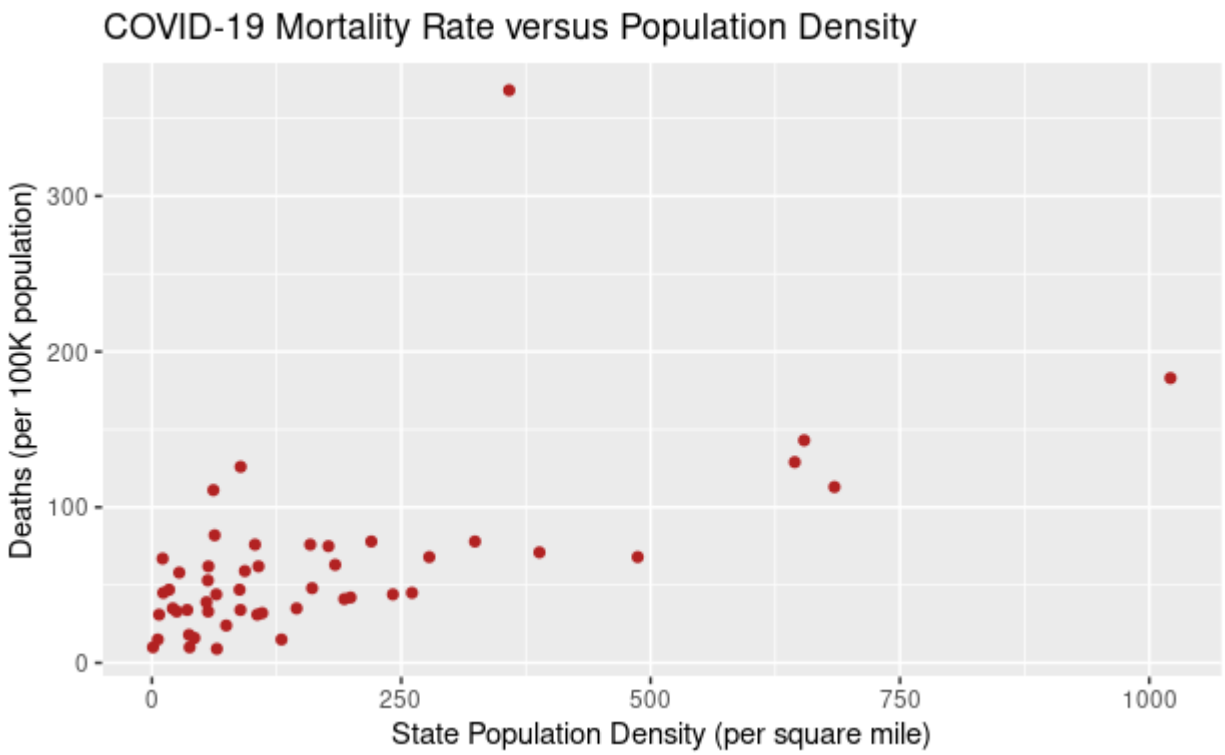
```
#New dataset covid that does not contain Washington DC
covid <- covid_all[-c(9), ]
```

Death Rates vs. Population Density

Finally, we examined the relationship between state mortality rates and population density. As discussed in the previous section we remove, Washington DC, as it is the US capital, a major city and not a US State. As shown in the scatterplot of mortality rate versus state population density below, there is a linear pattern in the data.

Hide

```
covid %>%
  ggplot(aes(x = (pop_density), y = (deaths_per_100K)))+
  geom_point(color = "firebrick") +
  ggtitle("Scatterplot") +
  labs(
    title = 'COVID-19 Mortality Rate versus Population Density',
    x ="State Population Density (per square mile)",
    y = "Deaths (per 100K population)")
```



Model Design

The regression model below, lm1, suggests that there is a strong correlation between the population density and the cumulative COVID-19 death rate in the United States. An increase of density (per square mile) by a hundred is associated with almost 16 more deaths (per 100K population). The R^2 value of 0.3145 indicates that population density alone does not completely explain the variation in the mortality rate; however, it is a good start.

Hide

```
lm1 <- lm(deaths_per_100K ~ pop_density, data = covid)
summary(lm1)
```

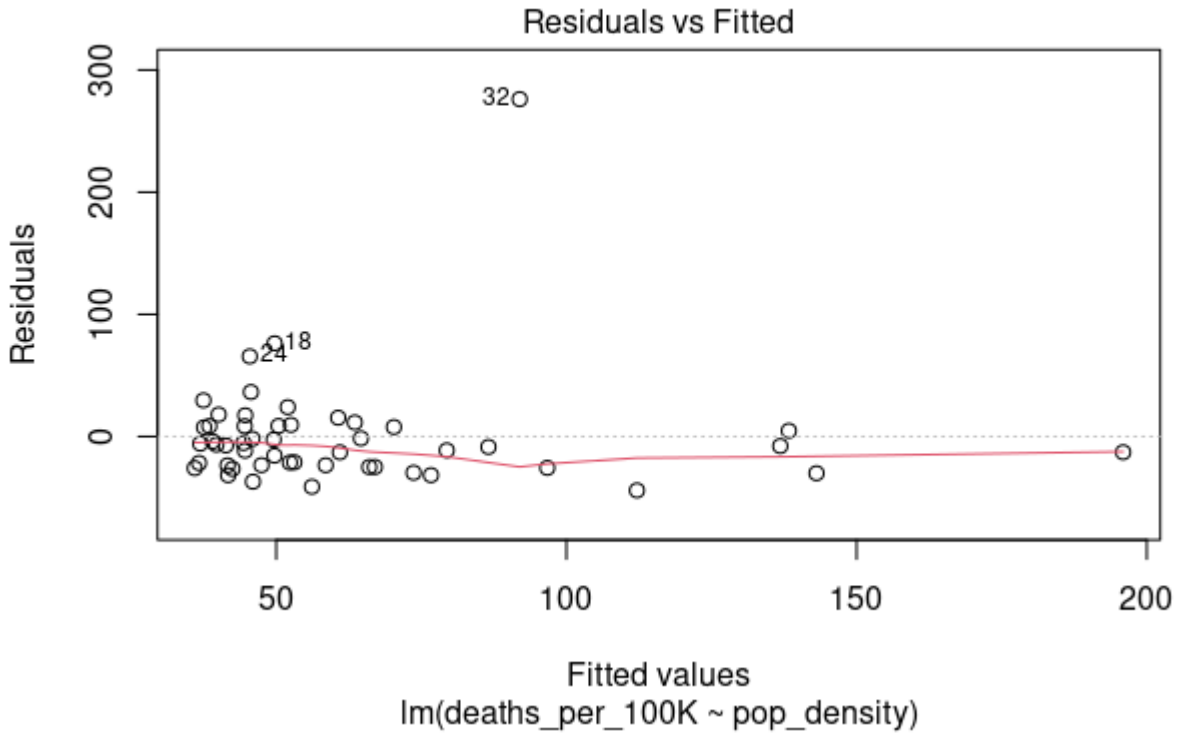
```

Call:
lm(formula = deaths_per_100K ~ pop_density, data = covid)

Residuals:
    Min      1Q  Median      3Q     Max 
-44.156 -23.591  -7.586   8.500 276.053 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 35.77426    8.65886   4.132 0.000143 ***
pop_density  0.15682    0.03236   4.846 1.36e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.17 on 48 degrees of freedom
Multiple R-squared:  0.3285,    Adjusted R-squared:  0.3145 
F-statistic: 23.48 on 1 and 48 DF,  p-value: 1.362e-05
```

CLM Assumptions

- IID
  We are interested in undsertanding the relationship between population density and deaths from COVID-19 using a descriptive model. The reference population is the world population with COVID deaths. Our dataset is aggregated data for each US State. The samples could reasonably be considered identically distributed with the removal of the District of Columbia. However, states are not independent of each other (and not random from the population). There are regional similarities and influences from neighboring states. A different dataset with randomly selected populations from around the world would be a better approximation of IID for the population of interest. However, we will proceed with our analysis with the data we have available.
- Linear Conditional Expectation
  The fitted model, lm1, is level-level. The EDA indicated that there was a linear pattern in the data. The residuals vs. fitted plot looks fairly centered at zero - although there are some extreme outliers.

Hide

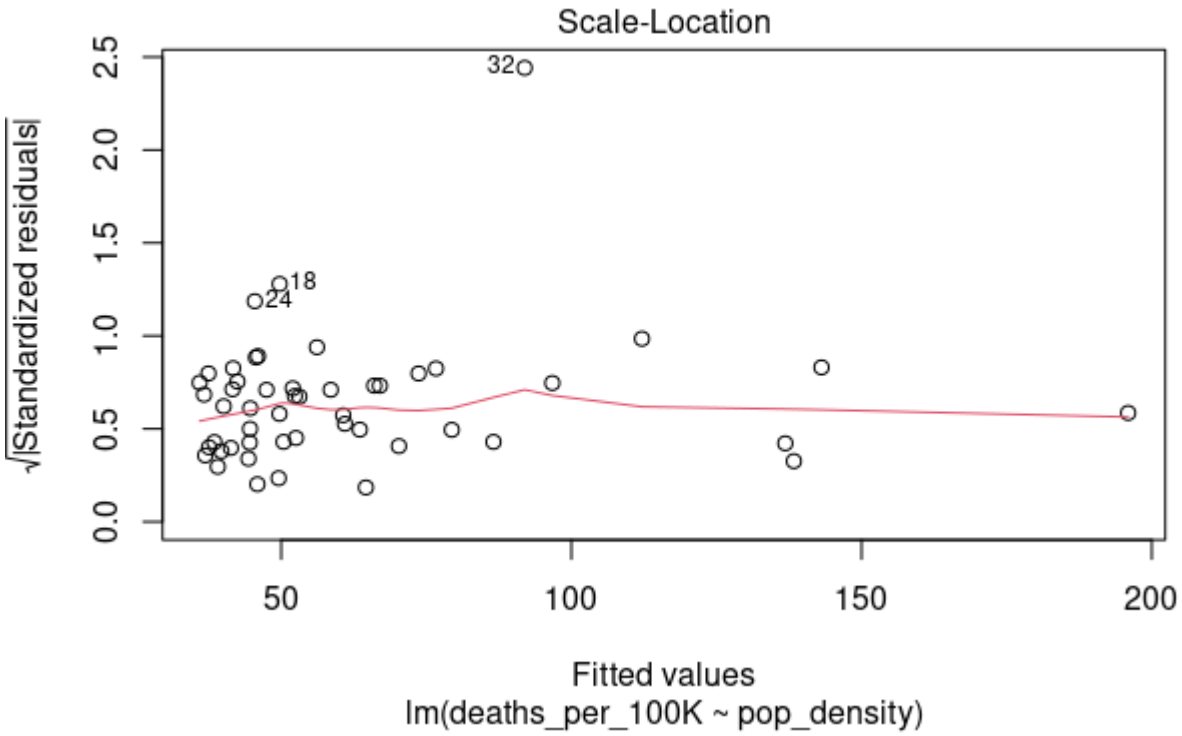
```
plot(lm1, which = 1)
```



- No Perfect Collinearity
  Perfect collinearity between the COVID-19 mortality rates and population denisty was not found while running the model. The f statistic and p values are aligned, so there does not appear to be an issue with collinearity (vif is not valid for models with less than two terms).
- Constant Error Variance (Homoskedastic errors)
  Checking for heteroskedastic variance using the standardized residuals vs. fitted value plot. There does not appear to be fanning or heteroskedastic behavior.

Hide

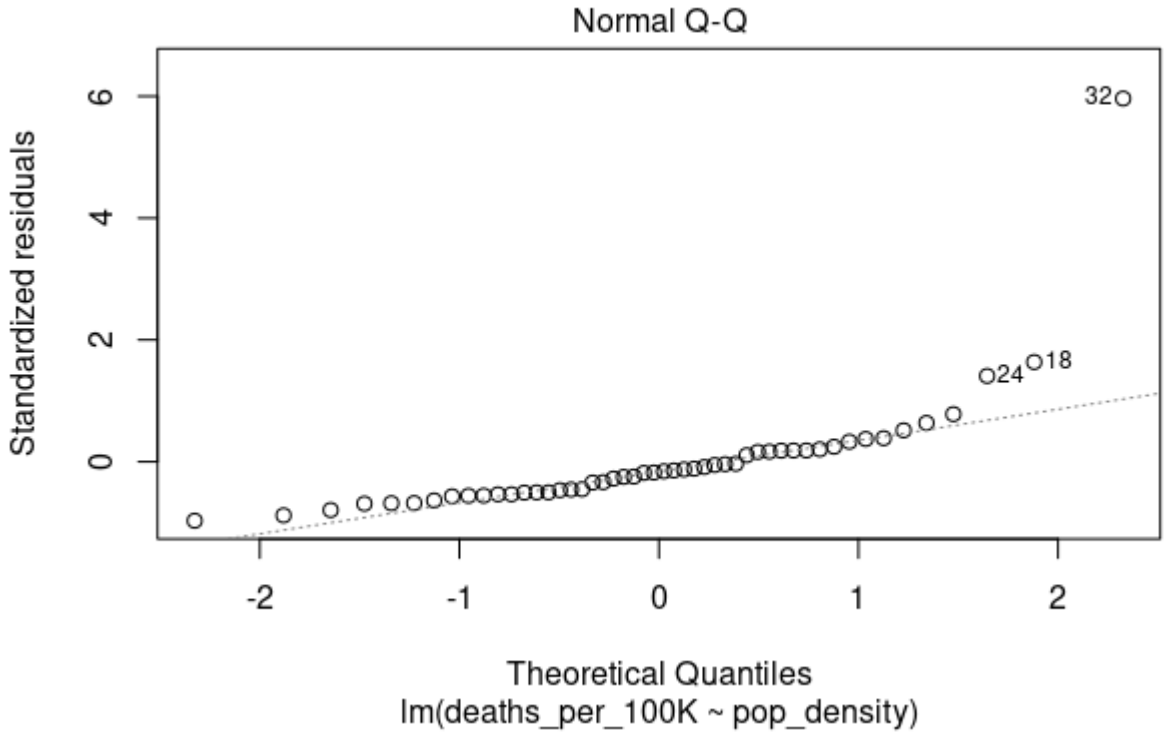
```
plot(lm1, which = 3)
```



- Normally Distributed Errors
  The Q-Q plot shows the errors appear generally follow the normal line, there is a strong deviation for the right-most values. The histogram of residuals has errors that are generally normal but their is a tail due to an outlier (New York). Applying a log transformation may help our model with meeting this CLM assumption.

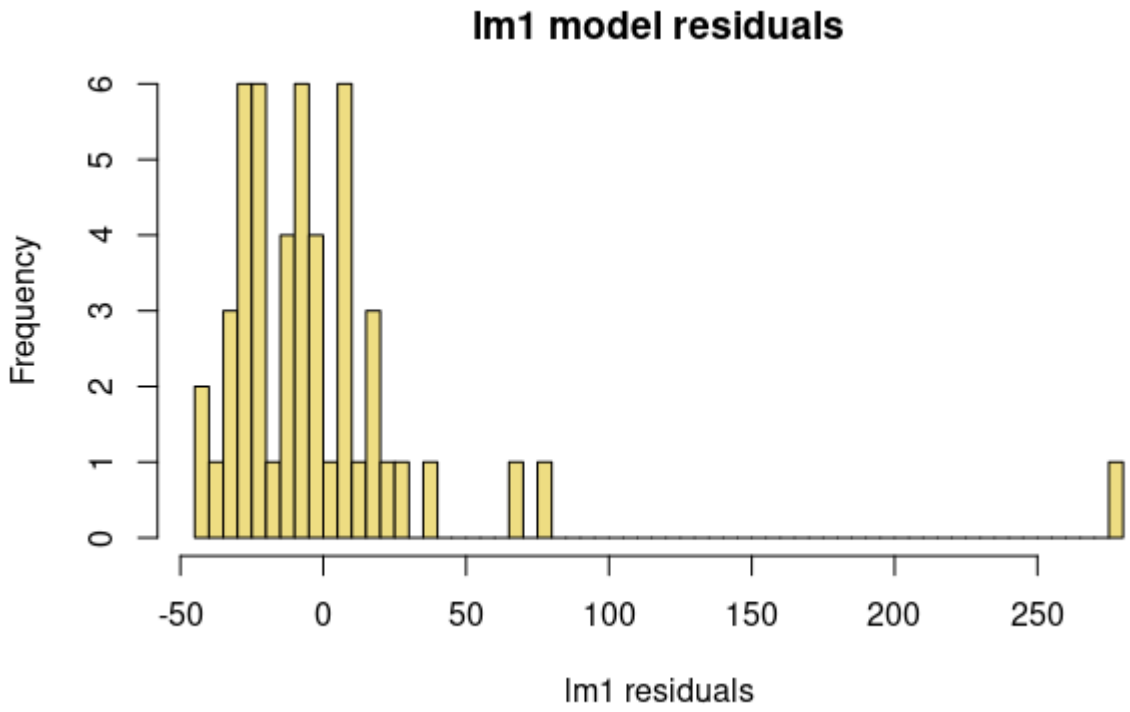
Hide

```
plot(lm1, which = 2)
```



Hide

```
lm1_resid = resid(lm1)
hist(lm1_resid, col = "lightgoldenrod", main = "lm1 model residuals", xlab = "lm1 residuals", br
eaks = 100)
```

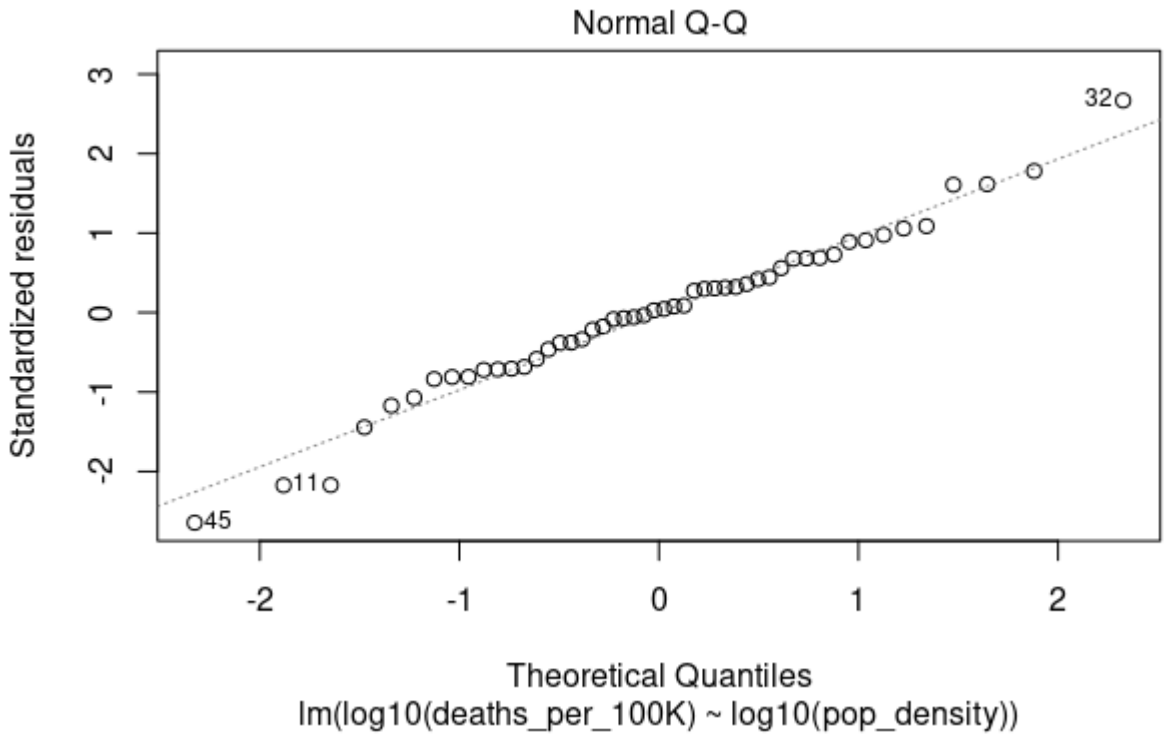


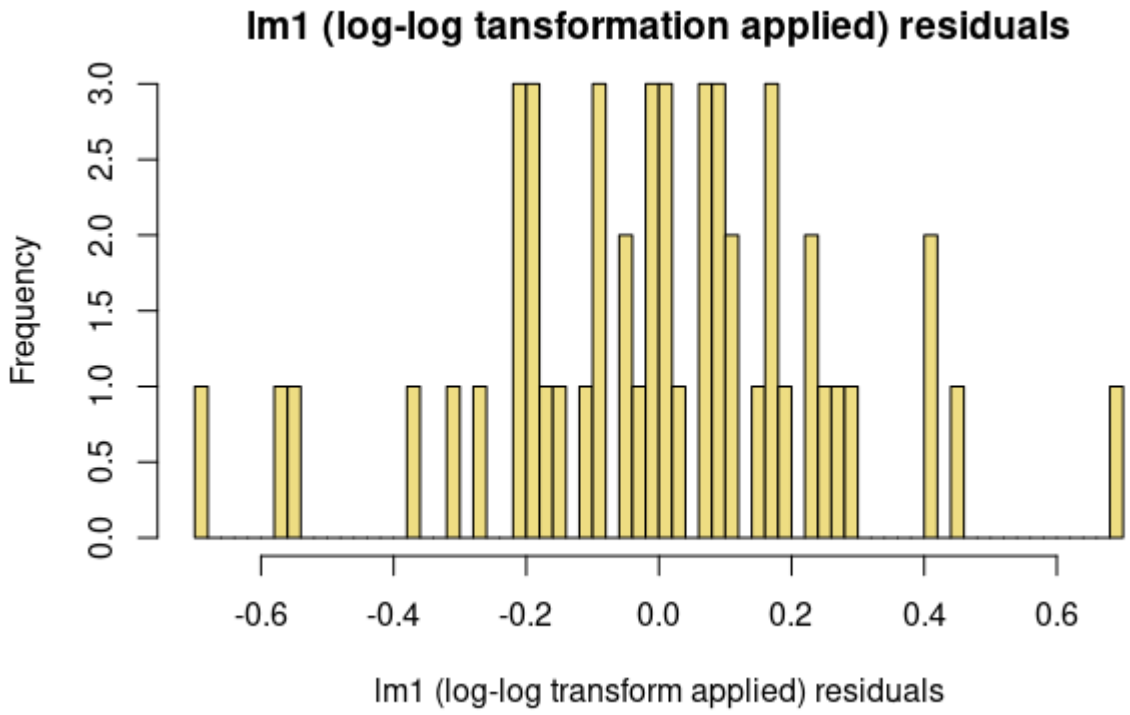
What *transformations*, if any, should you apply to each variable? These transformations might reveal linearities in scatterplots, make your results relevant, or help you meet model assumptions.

A log-log transformation for lm1 will be applied to help with meeting CLM assumption for normally distributed errors in order to draw a valid conclusion from the hypothesis testing.

```
Call:
lm(formula = log10(deaths_per_100K) ~ log10(pop_density), data = covid)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68361 -0.16993  0.00922  0.16399  0.68150

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.03396    0.12938   7.991 2.27e-10 ***
log10(pop_density)  0.33295    0.06414   5.191 4.20e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2612 on 48 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.3462
F-statistic: 26.94 on 1 and 48 DF,  p-value: 4.198e-06
```

After applying the log-log transformation, the errors now appear to be normally distributed as shown in the Q-Q plot and histogram of the residuals





Model 2

The second, and slightly more detailed version of our model includes the relationship between the key variables that we wish to measure as well as select covariates that we believe advance our modeling goals. Specifically, in addition to modeling the relationship between COVID-19 mortality rate and population density, we introduce the a feature to represent poverty.

Poverty Rate

Operationalization

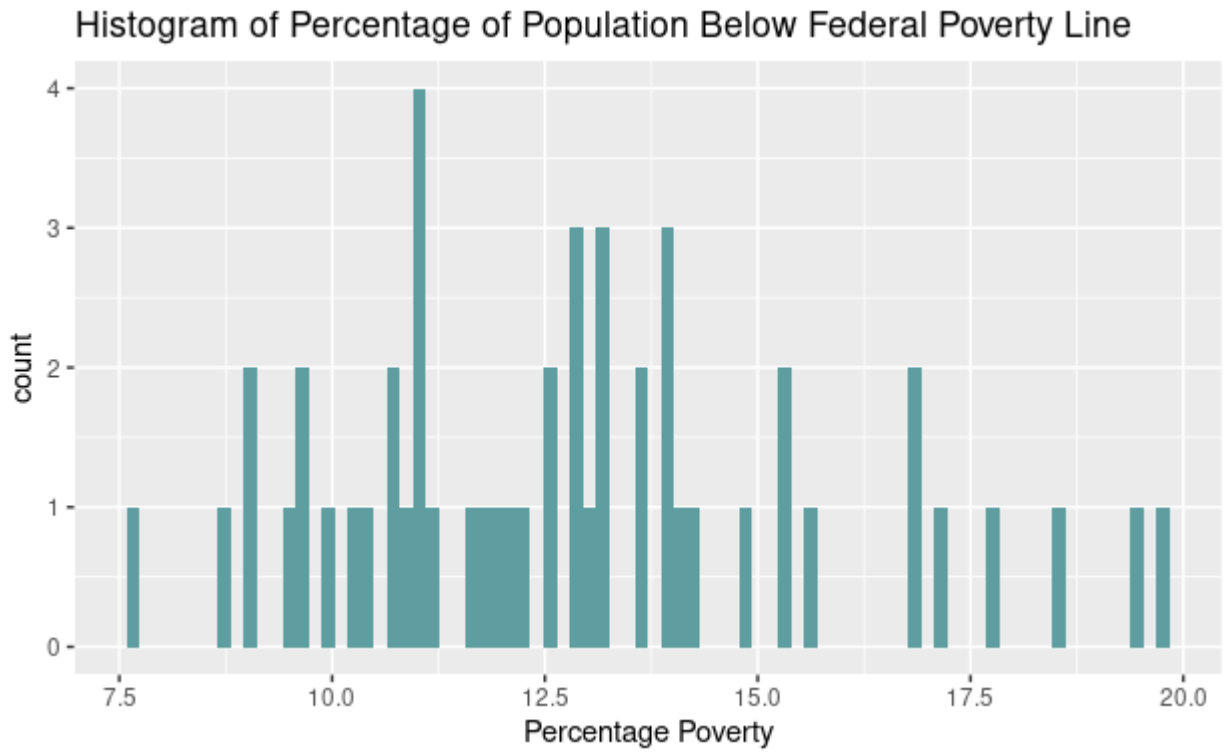
We will operationalize the concept of poverty rate by using the variable `poverty_percent`. This variable represents the statewide average of number of inhabitants who live in income below the federal poverty line. We believe that the `poverty_percent` realistically represents the concept we are trying to measure because it is calculated on the state level, is normalized for population, and is expressed in common units that are easily interpretable by most readers. Using the federal poverty line potentially creates issues as a result of cost of living differences between states, but also provides a consistent way of isolating each state's poverty level.

Summary

The histogram below shows the distribution of homelessness population across the fifty states. It is evident from the figure that there is considerable variation in the percentage of the population living in poverty from state to state, which leads us to believe that this could be an indicator that describes the difference between total mortality rate between states.

Hide

```
ggplot(data = covid, mapping = aes(x= poverty_percent)) +
  geom_histogram(bins = 80, fill="cadetblue") +
  ggtitle("Histogram of Percentage of Population Below Federal Poverty Line") +
  labs(x = "Percentage Poverty")
```



Model Design

We generate our second model by summarizing the linear model function and interpreting its results.The last column of the table above contains the p-values for each of the independent variables. The p-value for percentage of poverty is small than 0.05, providing evidence that percentage of poverty is a significant predictor of the number of deaths. The Estimate column in the coefficients table, gives us the coefficients for each independent variable in the regression model.

The R2 value increases with the number of independent variables so it is better to use the adjusted R squared value when comparing models. The adjusted R2 indicates that 40.88% of the variation in the number of deaths can be explained by the model.

Hide

```
lm2 <- lm(log10(deaths_per_100K) ~  log10(pop_density)+ (poverty_percent), data = covid)
summary(lm2)
```

```
Call:
lm(formula = log10(deaths_per_100K) ~ log10(pop_density) + (poverty_percent),
    data = covid)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62528 -0.13211  0.01713  0.12959  0.65232

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.61776    0.20878   2.959  0.00482 **
log10(pop_density)  0.34230    0.06111   5.601 1.07e-06 ***
poverty_percent     0.03099    0.01256   2.467  0.01731 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2484 on 47 degrees of freedom
Multiple R-squared:  0.433, Adjusted R-squared:  0.4088
F-statistic: 17.94 on 2 and 47 DF,  p-value: 1.621e-06
```

After performing the regression analysis, for the following analysis, we will closely diagnostic the regression model in order to detect potential problems and to check whether the following four assumptions made by the linear regression model are met or not.

- IID

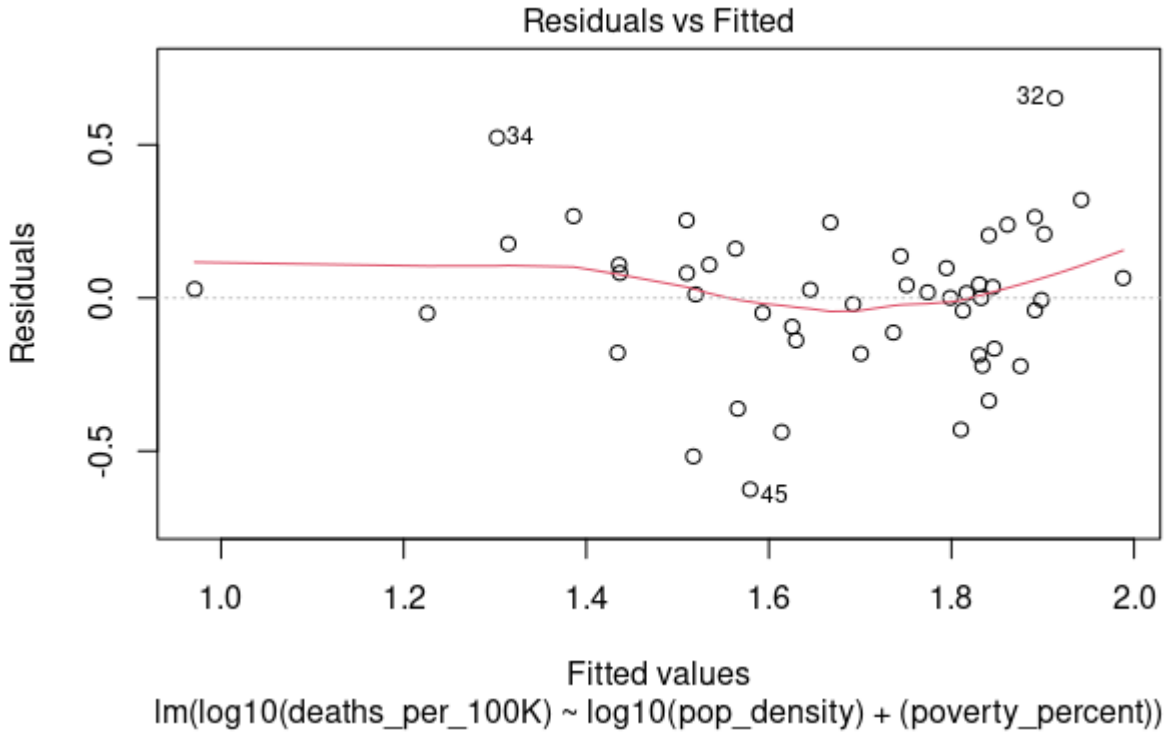
We are interested in undsertanding the relationship between population density and deaths from COVID-19 using a descriptive model.

The reference population is the world population with COVID deaths. Our dataset is aggregated data for each US State. The samples could reasonably be considered identically distributed with the removal of the District of Columbia. However, states are not independent of each other (and not random from the population). There are regional similarities and influences from neighboring states. A different dataset with randomly selected populations from around the world would be a better approximation of IID for the population of interest. However, we will proceed with our analysis with the data we have available.

- Linear relationship: The relationship between the independent variable and the dependent variable is linear.

Hide

```
plot(lm2,1)
```

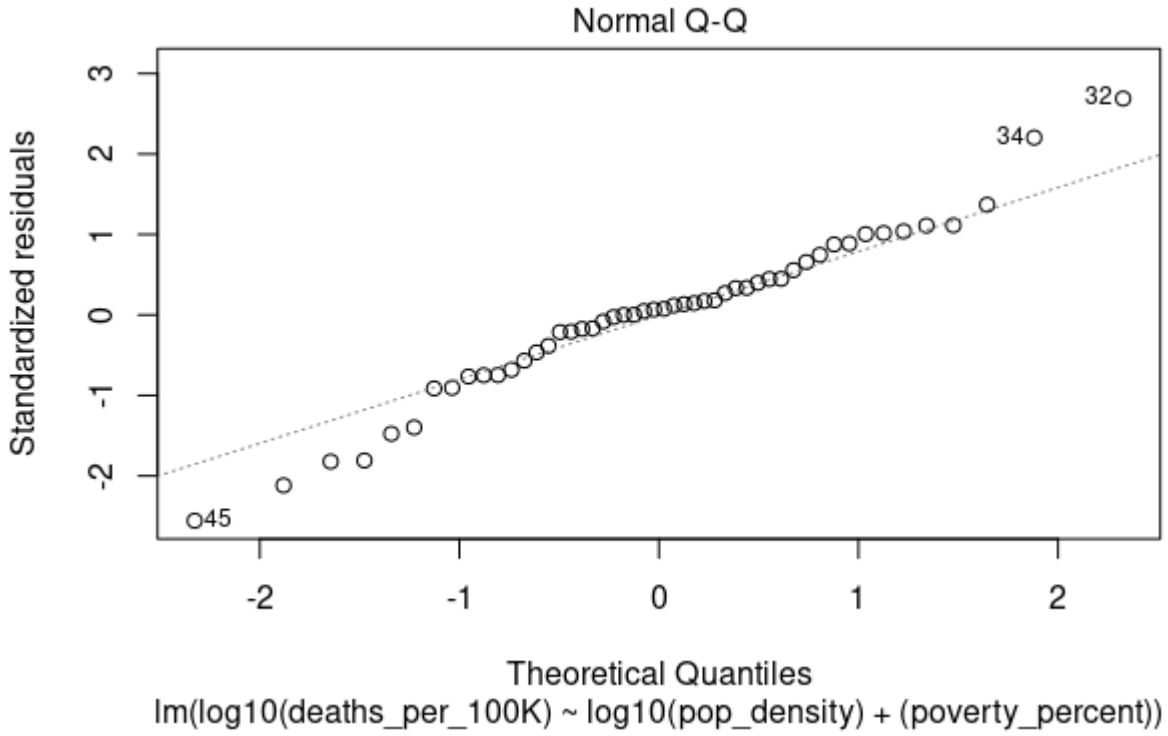


We used the plot of residuals versus predicted values to check the assumption of linearity. Ideally, the residual plot will show no fitted pattern. That is, the red line should be approximately horizontal at zero. The presence of a pattern may indicate a problem with some aspect of the linear model. In graph above, there is no pattern in the residual plot. This suggests that we can assume linear relationship between the independent and the dependent variables.

- Normality of Residual: Residuals are normally distributed

Hide

```
plot(lm2,2)
```



The QQ plot above shows that not all the residuals are normally distributed. We perform a Shapiro-Wilk Normality Test for further investigation.

Hide

```
sresid <- studres(lm2)
shapiro.test(sresid)
```

```
    Shapiro-Wilk normality test

data:  sresid
W = 0.96768, p-value = 0.186
```

From the p-value, which is larger than 0.05, it is reasonable to assume that residuals from the model is normally distributed.

- Multicollinearity: The independent variables are not highly correlated

Hide

```
car::vif(lm2, warning=FALSE,message=FALSE)
```

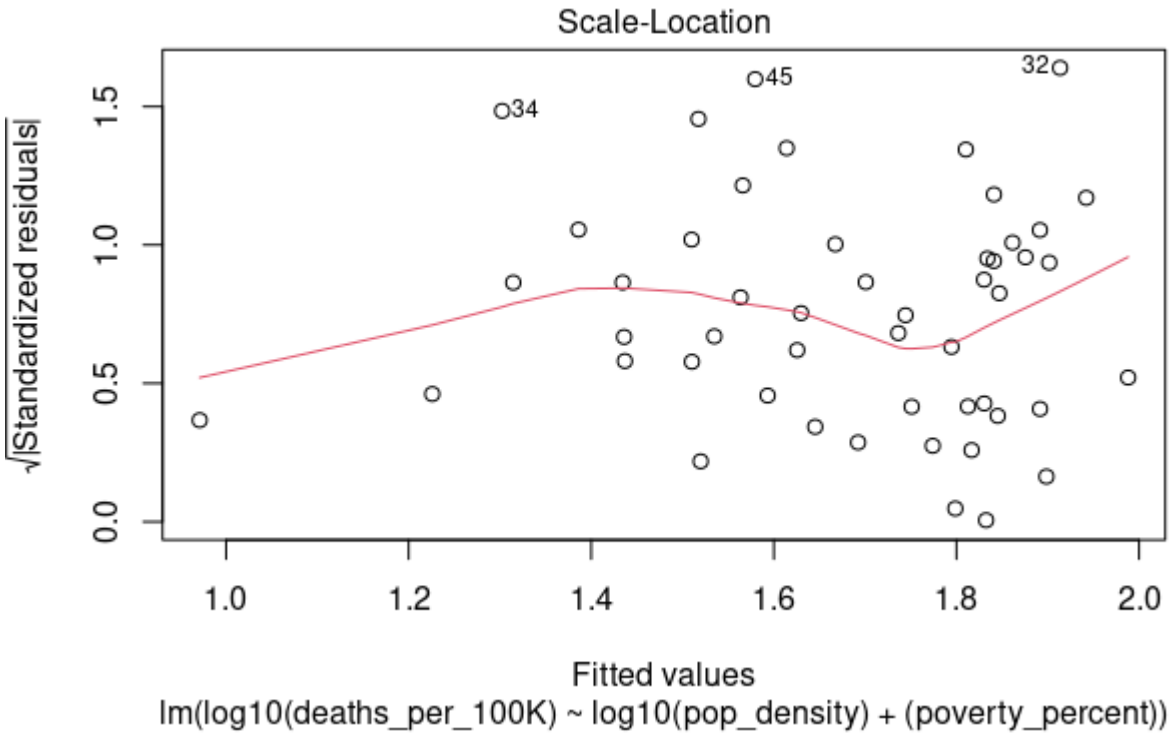
```
log10(pop_density)    poverty_percent
          1.003863           1.003863
```

We used the most widely-used diagnostic for multicollinearity, the variance inflation factor (VIF), which estimates how much the variance of a coefficient is inflated because of linear dependence with other predictors. For the data we are using, the VIF of percentage of poverty and population density is 1, which indicates that there is essentially no correlation between these two indicators.

- Homoskedastic: The residuals are assumed to have a constant variance

Hide

```
plot(lm2,3)
```



The final test is to use Scale Location Plot to determine if the error terms is the same across all values of the independent variable. The red line above has a positive slope and the data points are not randomly spread out, which mean this assumption is violated. Thus further adjustments are needed by either including or excluding predictors or weighting the measurements.

Model 3

Mobility Variables

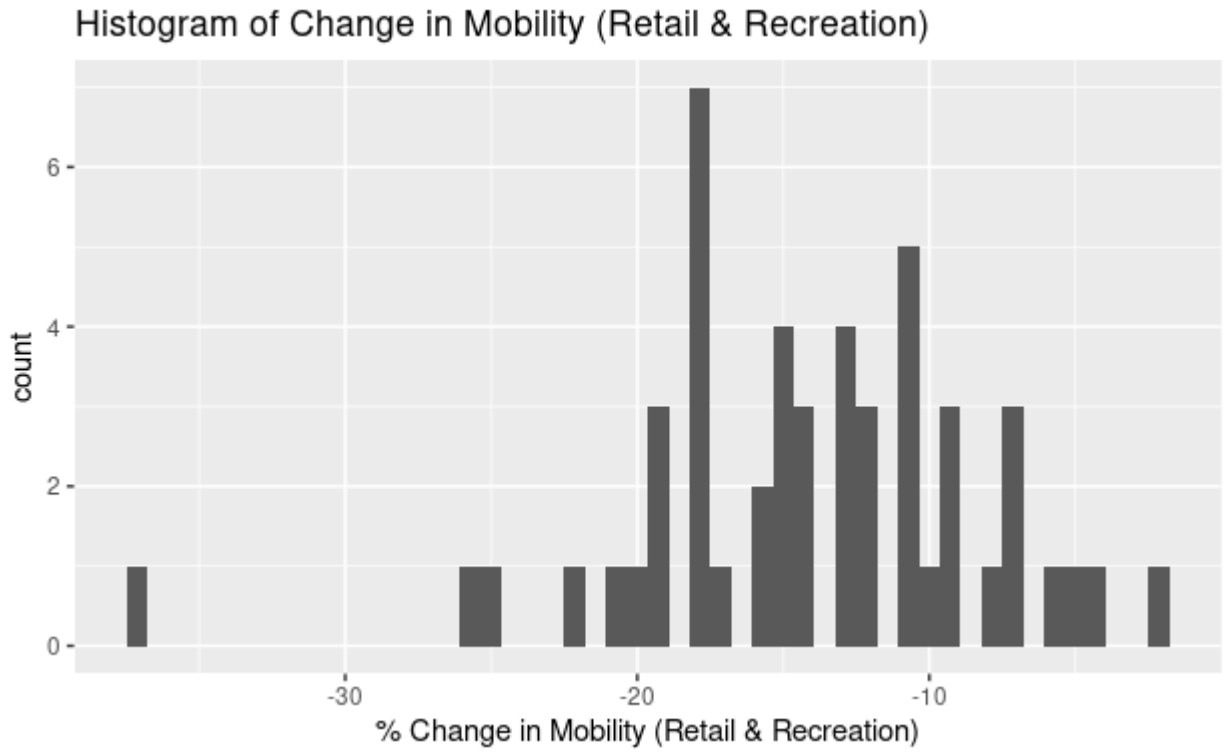
Operationalization

We will operationalize the concept of the mobility variables, specifically the Retail & Recreation, Workplaces and Grocery & Pharmacy by using the variables `retail_rec` , `grocery_pharm` , and `workplaces` from Google's COVID-19 Community Mobility Reports. These variables represents the statewide average of percent change in mobility to these types of places, in other words we are looking at the movement trends during the 5-week period Jan 3–Feb 6, 2020. We believe that these mobility variables represent the concept we are trying to measure because they are calculated at the state level and are expressed in common units that are easily interpretable by most readers. As the categories of specific places that we've chosen are common ones that a majority people visit regularly - Retail & Recreation, Groceries & Pharmacies and Workplaces, the changes or lack of changes in trends could expose some patterns with regards to population density and deaths from COVID-19. #### Summary For the model 3, we will take a maximalist approach and added more covariates to those in model 2, specifically the mobility variables - `retail_rec` , `grocery_pharm` , and `workplaces` . We are looking at these variables, as we

believe the people's movement patterns may affect their exposure to the novel coronavirus and vulnerability to COVID-19, thus we are looking at home they may or may not have have changed. Thus we believe this variables may also be responsible for mortality variation.

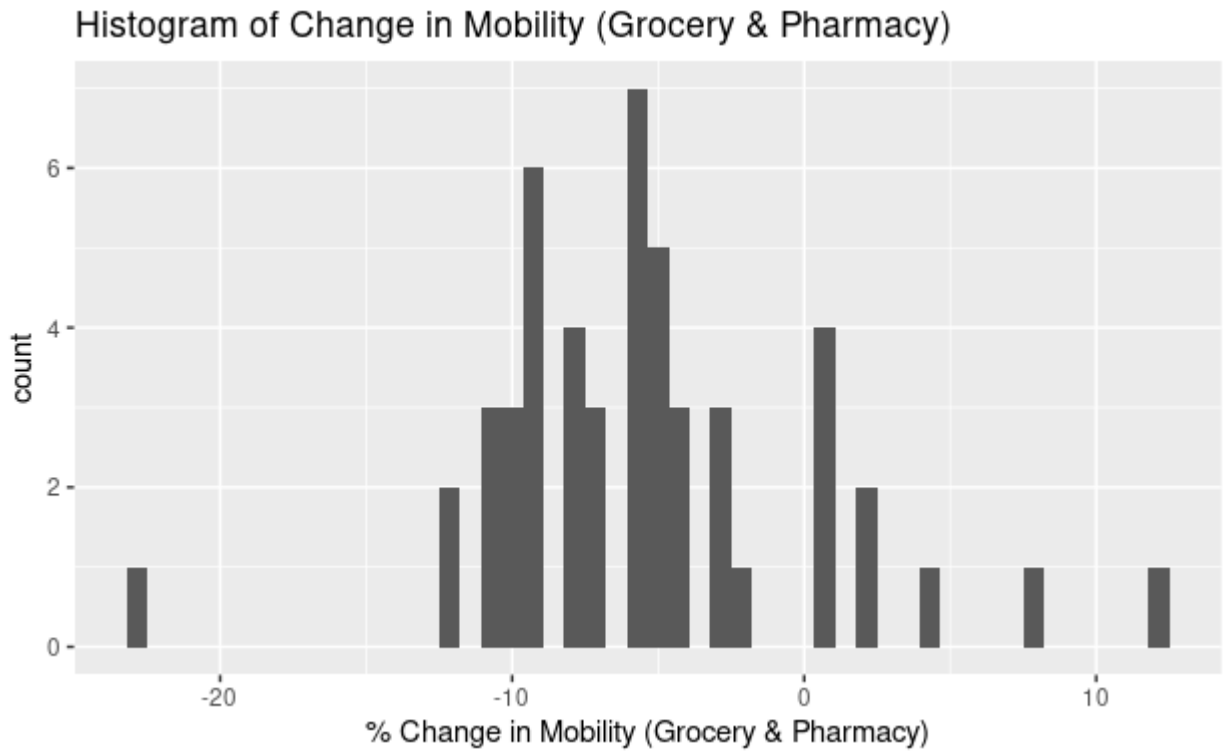
Hide

```
ggplot(data = covid, mapping = aes(x= retail_rec)) +
  geom_histogram(bins = 50) +
  ggtitle("Histogram of Change in Mobility (Retail & Recreation)") +
  labs(x = "% Change in Mobility (Retail & Recreation)")
```



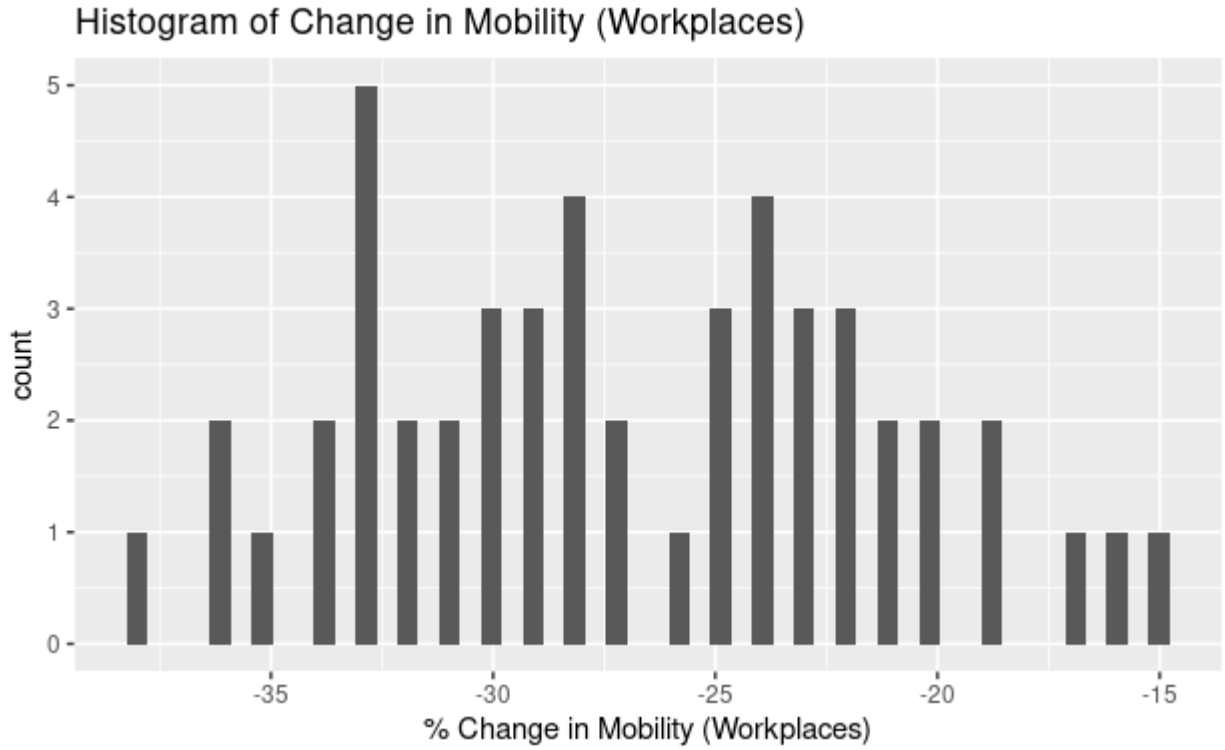
Hide

```
ggplot(data = covid, mapping = aes(x= grocery_pharm)) +
  geom_histogram(bins = 50) +
  ggtitle("Histogram of Change in Mobility (Grocery & Pharmacy)") +
  labs(x = "% Change in Mobility (Grocery & Pharmacy)")
```



Hide

```
ggplot(data = covid, mapping = aes(x= workplaces)) +
  geom_histogram(bins = 50) +
  ggtitle("Histogram of Change in Mobility (Workplaces)") +
  labs(x = "% Change in Mobility (Workplaces)")
```



From the histograms above, while the percent change does vary, the distributions for each of the histrograms - `grocery_pharm` and `workplaces` are relatively normal, while `retail_rec` is slightly skewed right due to one datapoint on the left. It seems from the histograms that we may not need to apply any transformations.

Hide

```
lm3 <- lm(log10(deaths_per_100K) ~ log10(pop_density) + poverty_percent +  retail_rec + grocery_
pharm + workplaces, data = covid)
```

Hide

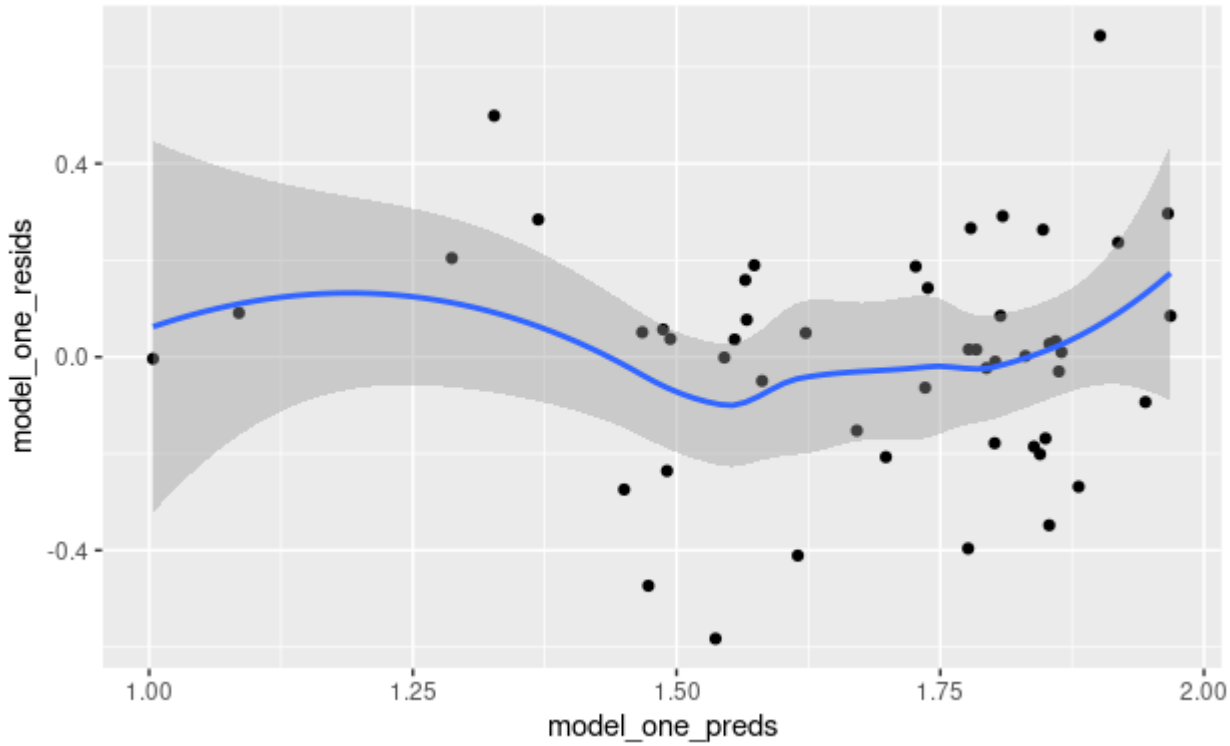
```
summary(lm3)
```

```

Call:
lm(formula = log10(deaths_per_100K) ~ log10(pop_density) + poverty_percent +
    retail_rec + grocery_pharm + workplaces, data = covid)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5827 -0.1645  0.0153  0.1297  0.6644

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.306559   0.323774   0.947  0.34889
log10(pop_density)  0.315184   0.070508   4.470 5.42e-05 ***
poverty_percent     0.038433   0.014059   2.734  0.00899 **
retail_rec          0.007994   0.010061   0.795  0.43111
grocery_pharm       0.001997   0.009249   0.216  0.83006
workplaces         -0.014577   0.010383  -1.404  0.16736
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.251 on 44 degrees of freedom
Multiple R-squared:  0.4578,    Adjusted R-squared:  0.3962
F-statistic: 7.431 on 5 and 44 DF,  p-value: 3.991e-05
```

We can see from the above results that that only poverty_percent and log10(pop_density) is statistically significant, similar to the results seen in model 2. Using the estimates column the model is log10(deaths_per_100K) = 0.306 + 0.038*(poverty_percent) – 0.315*log10(pop_density) + 0.008*(retail_rec) + 0.002* (grocery_pharm) - 0.015* (workplaces). As with Model 2, we will be using the adjusted R2. The adjusted R2 for this model shows that the model can explain 39.62% of the variation in the number of deaths per 100K population. This set of independent variables are not explaining the variation in the number of deaths per 100K population. Next, we will now look at the 5 Assumptions of CLM. 1. IID Sampling We are interested in understanding the relationship between population density and deaths from COVID-19 using a descriptive model. The reference population is the world population with COVID deaths. Our dataset is aggregated data for each US State. The samples could reasonably be considered identically distributed with the removal of the District of Columbia. However, states are not independent of each other (and not random from the population). There are regional similarities and influences from neighboring states. A different dataset with randomly selected populations from around the world would be a better approximation of IID for the population of interest. However, we will proceed with our analysis with the data we have available. 2. Linear Conditional Expectation To assess whether there is a linear conditional expectation, we will look at the predicted vs. residuals of the model.

Hide

```
model_one <- lm(log10(deaths_per_100K) ~ log10(pop_density) + poverty_percent +  retail_rec + gr
ocery_pharm + workplaces, data = covid)
model_one_preds = predict(model_one)
model_one_resids = resid(model_one)
covid <- covid %>% mutate(
    model_one_preds,
    model_one_resids
  )
covid %>%
  ggplot(aes(model_one_preds, model_one_resids)) +
  geom_point() + stat_smooth()
```



The residual plot above shows no fitted pattern. Thus, we can assume that there is a linear relationship between the independent and the dependent variables. If this assumption had been violated, then to correct this, we would need to add variable transformations. 3. No Perfect Collinearity To observe if there is no collinearity, we will be looking at the variance inflation factor (VIF).

Hide

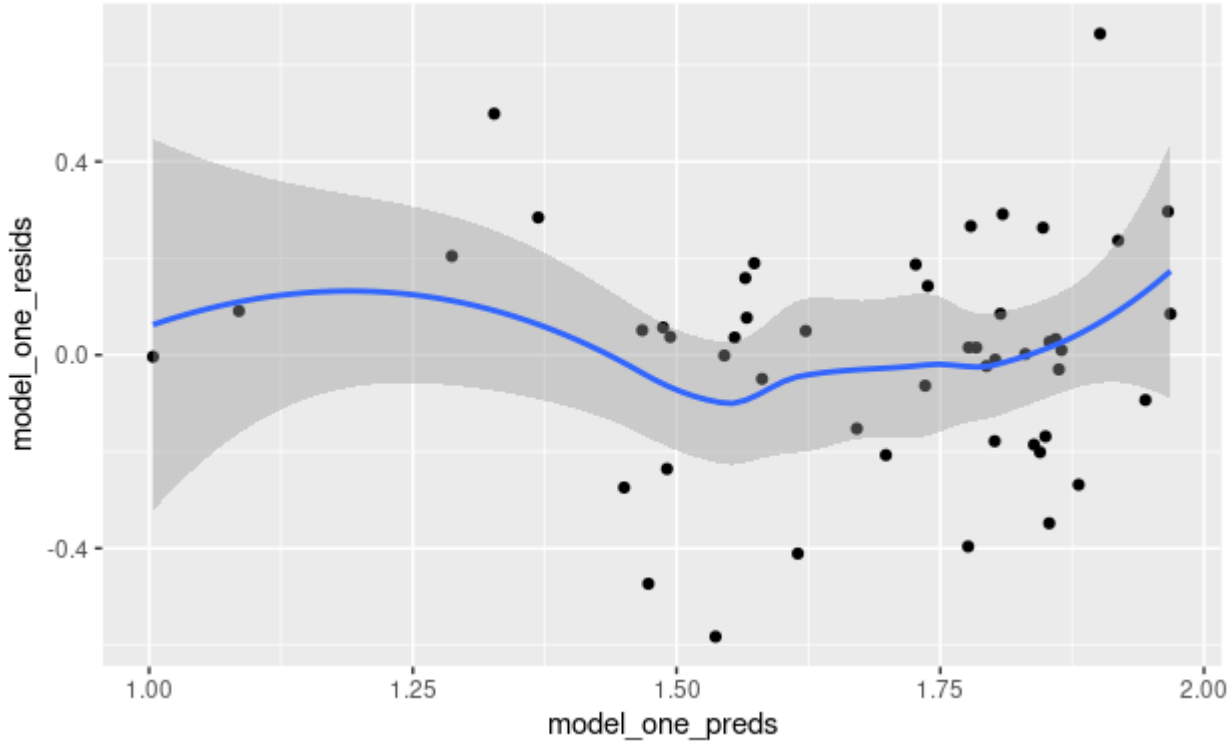
```
car::vif(lm3)
```

```
log10(pop_density)    poverty_percent         retail_rec      grocery_pharm
          1.308490           1.231417           3.008661           2.102905
        workplaces
          2.711908
```

From the output, we can see that poverty_percent (1.2), pop_density (1.3), retail_rec (3.0), grocery_pharm (2.1) and workplaces (2.7) have a VIF of less than 4, so there is little to no correlation between the variables. 4. Homoskedastic Errors

Hide

```
covid %>%
  ggplot(aes(model_one_preds, model_one_resids)) +
  geom_point() + stat_smooth()
```



To observe whether the distribution of the errors is homoskedastic, we will examine the residuals versus fitted plot from earlier. There is not a band of even thickness from left to right. There is an area between x=1.5 and x = 2.0, where there are many points bunched up and not even. Thus, the distribution of the errors is not randomly spread out and is not homoskedastic. As it seems that this model is not homoskedastic, we will run a Breusch-Pagan test to verify. The null hypothesis is that there is homoskedastic error variance.

Hide

```
lmtest::bptest(lm3)
```

```
    studentized Breusch-Pagan test

data:  lm3
BP = 4.6111, df = 5, p-value = 0.4652
```

From the results of the Breusch-Pagan test, as the p-value is greater than 0.05, we fail to reject the null hypothesis that there is homoskedasticity and there might not be heteroskedasticity. To fix this, in future iterations, we will adjust the standard errors for the coefficients by using robust standard errors. 5. Normally Distributed Errors

Hide

```
plot_one <- covid %>%
  ggplot(aes(x = model_one_resids)) +
  geom_histogram()

plot_two <- covid %>%
  ggplot(aes(sample = model_one_resids)) +
  stat_qq() + stat_qq_line()
plot_one / plot_two
```

```
Error in plot_one/plot_two : non-numeric argument to binary operator
```

The histogram of residuals is relatively normal. The qqplot is also relatively normal and linear, with some deviation on the right tail, but it is still acceptable. The errors are normally distributed.

Limitations

Our model has two main limitations that we would like to identify. First, while the model seeks to use variables that likely have real-world causal relationships with COVID-19 mortality rate, it is still a descriptive model. Second, this model was generated after investigating many variable pairs and selecting those which had a strong relationship to one another. While we did not engage in any cherry picking to remove inconvenient data points, these variable pairs were intentionally selected with correlation coefficient in mind.

Regression Table

Hide

```
library(stargazer)
```

Hide

```
lm1 <- lm(log10(deaths_per_100K) ~ log10(pop_density), data = covid)

lm2 <- lm(log10(deaths_per_100K) ~ log10(pop_density) + poverty_percent, data = covid)

lm3 <- lm(log10(deaths_per_100K) ~ log10(pop_density) + poverty_percent + retail_rec + grocery_p
harm + transit + workplaces, data = covid)
```

Hide

```
stargazer(lm1, lm2, lm3,
          type="text",
          #se = list(sqrt(diag(vcovHC(lm1))), sqrt(diag(vcovHC(lm2))), sqrt(diag(vcovHC(lm3)))),
                    column.labels = c("Model 1", "Model 2", "Model 3"))
```

```
length of NULL cannot be changedlength of NULL cannot be changedlength of NULL cannot be changed
length of NULL cannot be changedlength of NULL cannot be changednumber of rows of result is not
a multiple of vector length (arg 2)number of rows of result is not a multiple of vector length
(arg 2)
```

	Dependent variable:		
	log10(deaths_per_100K)		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
log10(pop_density)	0.333***	0.342***	0.320***
	(0.064)	(0.061)	(0.075)
poverty_percent		0.031**	0.037**
		(0.013)	(0.016)
retail_rec			0.007
			(0.010)
grocery_pharm			0.002
			(0.010)
transit			0.001
			(0.005)
workplaces			-0.016
			(0.012)
Constant	1.034***	0.618***	0.289
	(0.129)	(0.209)	(0.336)
Observations	50	50	50
R2	0.360	0.433	0.458
Adjusted R2	0.346	0.409	0.383
Residual Std. Error	0.261 (df = 48)	0.248 (df = 47)	0.254 (df = 43)
F Statistic	26.944*** (df = 1; 48)	17.944*** (df = 2; 47)	6.068*** (df = 6; 43)
Note:			*p<0.1; **p<0.05; ***p<0.01

Omitted Variables

Although this is not a causal analysis, we selected the variables in our model with the assumption that they had a real relationship to COVID-19 mortality. For this reason, discussing the possibility of omitted variable bias can be a useful exercise to explore relationships within the model and provide potential areas of focus for future research. Identified below are five areas of potential omitted variable bias.

- Our model found a positive correlation between population density and COVID-19 mortality. However, it could be the case that areas of high population density have a lower hospital capacity when adjusted for population. Given that there is a positive relationship between population density and COVID-19 mortality, and likely a positive correlation between lower hospital capacity and COVID-19 mortality, the OVB would be away from zero.

- Regarding the relationship in the previous bullet, individuals in areas of high population density areas may be more likely to consume left-leaning media and thus take lockdown and social distancing policies more seriously. Since population density is positively coordinated with mortality and quarantine adherence is probably negatively correlated with mortality, the OVB would be toward zero.
- Our model found a positive correlation between poverty rate and COVID-19 mortality. However, it could be the case that the low income population is greater in states with a higher occurrence of health risk factors, which is associated with higher COVID-19 mortality. Since there is a positive correlation between these two variables and COVID-19 mortality, the OVB would be away from zero.
- Our model found a negative correlation between grocery/pharmacy mobility and COVID-19 mortality. However, it could be the case that areas of higher grocery/pharmacy mobility tend to be higher income and thus have better resources to prevent and treat illness. In this case, the omitted variable bias would be away from zero.
- Our model found a positive correlation between transit mobility and COVID-19 mortality. This could be similar to the previous case, where higher mobility is correlated with higher income, and thus individuals in this area have better resources to prevent and treat illness. In this case, the omitted variable bias would be toward zero.

Conclusion

The COVID-19 pandemic is an issue on the minds scientists, policymakers, and billions of individuals worldwide who have a stake in the health and safety of their communities. Creating actionable plans and policies to navigate this current global crisis is complicated by the panoply of geographic, demographic, and social characteristics that make each of these communities unique. In this project we endeavored to investigate several of these charastics that we believed had the potential to impact COVID-19 mortality in the United States.

At the beginning of this project, we set out to study the relationship between population density and COVID-19 mortality in the United States and we constructed and initial model to do just that. We identified a strong, positive relationship between the average population density of a state and that state's COVID-19 mortality rate. In our second model, we expanded upon the first to include data representing each states poverty rate, which increased the strength of the relationship and reinforced our understanding that, as in many aspects of life, negative consequences are most likely to fall on those in our society who are most vulnerable. Finally, in our third model, we tried to more comprehensively capture the diversity of COVID-19 impacts on communities. By further expanding the scope to include several types of community mobility, we were able to account for the fact that general changes in policies, compliance, and public attitudes have an impact on which members of society are contracting and dying from COVID-19.

It is our hope that this report can provide foundational information for any individual who has concern about the health and safety of their community. By understanding the relationship between population density and COVID-19 mortality, policymakers on the national level can make more effective plans to distribute medical supplies, PPE, and vaccines. By understanding the relationship between poverty and COVID-19 mortality, state officials can direct attention to the struggling communities who need help the most. And finally, by understanding the effects of community mobility factors on COVID-19 mortality, individuals can look at their own community with the knowledge of how their visits to their local grocery stores, hospitals, and workplaces affect the ability of their neighbors to fight, and survive, this deadly pandemic.