Lab 2 Final: Does the percentage of COVID cases by race influence mortality rate more so than a state's hospital preparedness?

Ratan Singh, Kevin Fu, Jamie Smith

12/9/2020

```
if(!require(tidyverse)){install.packages("tidyverse")}
if(!require(stargazer)){install.packages("stargazer")}
library(tidyverse)
library(readxl)
library(data.table)
library(gridExtra)
library(stargazer)
library(car)
library(lmtest)
library(sandwich)
library(tinytex)
```

Introduction

The amorphous nature of Covid-19 has baffled many medical practitioners. The virus has turned the world upside down and has continued to do so throughout 2020. Though the impacts of this virus have been universally felt, our research team postulates that the virus impacts certain demographic groups more acutely than others.

Research Question: Does the percentage of COVID cases by race influence mortality rate more so than a state's hospital preparedness?

To measure these factors, we will be looking at the percentage of COVID cases across the 50 states by White, Hispanic, and Black patients. For hospital preparedness, we will be looking at the available inpatient and ICU beds by state.

Dependent Variable:

Mortality Rate

Data was collected from the Center for Desease Control (CDC) to enable us to perform this research. Throughout the analysis, the Mortality Rate will serve as our Dependent Variable, which we have defined as:

$$\text{Mortality Rate} = \frac{\text{Total Covid-19 Deaths}}{\text{Total Covid-19 Cases}}$$

Total Covid-19 Deaths are recorded by the CDC as well as the Total Covid-19 Cases. Our sample set for these two metrics was captured on October 30, 2020 for each of the 50 sets.

Independent Variables:

Race & Ethnicity Data

The research team will be using the percentage of covid cases by the demographic groups of interest (Black, Hispanic, White). This data is being collected from Kaiser Family Foundation (https://www.kff.org/statedata/ (https://www.kff.org/statedata/)) and has a snapshot date of October 25th, 2020 for each state. The data is provided as a percentage from the KFF, but for illustrative purposes, we've shown what this metric represents below:

$$\%\text{ of Black Covid Cases} = \frac{\text{Total Cases for Individuals Identifying as Black}}{\text{Total Cases}}$$

$$\%\text{ of Hispanic Covid Cases} = \frac{\text{Total Cases for Individuals Identifying as Hispanic}}{\text{Total Cases}}$$

$$\%\text{ of White Covid Cases} = \frac{\text{Total Cases for Individuals Identifying as White}}{\text{Total Cases}}$$

This effectively allows us to see if for certain states with higher percentages of cases for a certain demographic, whether or not there's a difference in the mortality rate. Other considerations in the data set were to try to normalize for the percentage of the population represented by the racial group, however, we found that this additional consideration adds little value. This is because we do not care for the representation percentages, which would influence the percentage breakdown of cases by race, what we care about is if the variation in the percentage of covid cases state by state by demographic group has an effect on the mortality rate.

Available ICU Beds per Covid Patient

There are a number of ways to try to measure the preparedness of the hospitals in a state such as spending, quality and availability of doctors, etc. The research team chose to evaluate one measure of preparedness by the number of available Intensive Care Unit (ICU) beds in a given state over the total number of Covid Patients in the state:

$$\text{ICU Bed Availability Metric} = \frac{\text{Total ICU Beds Available}}{\text{Total Number of Covid Patients}}$$

This metric is being used as a proxy for preparedness in this study. As the cases of covid increases without a rise in the number of ICU beds available, the hospital will be less prepared to handle the potential increase of people needing to go to the ICU. One thing to make note of in this approach is that as ICU Bed availability decreases, that is an indication of more patients needing intensive care and thus influencing the mortality rate. This will be somethig to consider throughout the analysis.

The team collected this data from the CDC. One important factor to note is that the data is captured at the state-wide level, in congruence with other features of the sample. However, it would have been useful for us to have this metric at a more local level, perhaps by zip codes. This local level approach might prove to be a bit more meaningful, however, it might also remove a degree of randomness to our results. Already, there is the possibility that states that are in close proximity share enoguh in common to confound our results a bit.

In-Patient beds per Covid Patient

Similar to the ICU Bed Availability metric, we also will be utilizing a metric to indicate the in-patient beds over the total number of Covid Cases. In-Patient beds are beds occuppied by those in a hospital, but do not need the same level of treatment or care as those in ICU beds which. This metric is defined as:

$$\text{In-Patient Bed Availability Metric} = \frac{\text{Total In-Patient Beds Available}}{\text{Total Number of Covid Patients}}$$

Similar to the ICU Bed availability, the In-Patient bed availability metric should help us understand the preparedness of the hospitals in each state. As the total number of covid cases increases, the total in-patient beds availabel would need to also increase if the hospital were to be more prepared.

EDA & Model Building Process:

Prior to the creation of the model, the team did an extensive exploratory data analysis (EDA) to see if there were any transformations needed for our variables. Where most applicable, we made these transformations. We also did a bit of data clean up along the way, which is detailed in the EDA.

Lastly, the team created three descriptive models to help us understand whether the percent of covid cases by racial demographics had a larger effect on mortality rate than hospital preparedness. This was an iterative process. In model one, we only conisdered the variables pertaining to racial demographics to get a sense for the influence of these factors on mortality rate. In model two we expand on this to also include the variable related to ICU bed availability and finally in model three we introduce the variable related to in-patient bed availability.

Before we start any analysis, we need to load the covid-19 excel file.

```
data <- read_excel("covid-19.xlsx", sheet = "Covid-19",skip=1)
#view(data)
#glimpse(data)
#colnames(data)
```

We then rename the columns to ensure better readability and better access to the variables. We will perform an EDA for these variables later in the report.

```
#variables for dependent variable: mortality rate
names(data)[names(data) == "Total Deaths"] <- "TotDeaths"
names(data)[names(data) == "Total Cases"] <- "TotCases"

#variables for percentage of COVID cases in each race
names(data)[names(data) == "White % of Cases"] <- "CasesWhitePerc"
names(data)[names(data) == "Black % of Cases"] <- "CasesBlackPerc"
names(data)[names(data) == "Hispanic % of Cases"] <- "CasesHispPerc"
```

We also import data from a separate excel file called hospitals_preparedness.xlsx. The data is first downloaded from the CDC website at: https://www.cdc.gov/nhsn/pdfs/covid19/covid19-NatEst.csv (https://www.cdc.gov/nhsn/pdfs/covid19/covid19-NatEst.csv).

The latest date for data available for each of state in this file is July 07, 2020. The data is available for different months from March to July. However, to keep the model simple, we chose to pick the data from the latest date assuming that by July, different states were well aware of the Covid situation and their preparedness should reflect the measures taken in terms of inpatient and ICU bed availability. This filtering of data has already been implemented in the hospitals_preparedness.xlsx file we are using for this analysis. More EDA on these variables later.

We picked these variables because we think these are reliable estimates of the data we need from the excel file. More information can be found on the CDC website (https://www.cdc.gov (https://www.cdc.gov)).

As per the Excel file downloaded from the CDC website:

- InpatBeds_Occ_AnyPat_Est_Avail: Hospital inpatient beds available, estimate

- InpatBeds_Occ_COVID_Est: Number of patients in an inpatient care location who have suspected or confirmed COVID-19, estimate
- ICUBeds_Occ_AnyPat_Est_Avail: ICU beds available, estimate

```
hsp <- read_excel("hospitals_preparedness.xlsx", sheet = "Hospitals",skip=1)
#view(hsp)

#variables for hospital preparedness
names(hsp)[names(hsp) == "InpatBeds_Occ_AnyPat_Est_Avail"] <- "inpatient_beds"
names(hsp)[names(hsp) == "ICUBeds_Occ_AnyPat_Est_Avail"] <- "ICU_beds"
names(hsp)[names(hsp) == "InpatBeds_Occ_COVID_Est"] <- "num_COVID_patients"
```

We join this data from dataframe “hsp” into our initial dataframe “data”.

```
data <- inner_join(data,hsp,by="State")
#glimpse(data)
```

We perform an EDA on our dependent variable, Mortality_Rate, first. We will define mortality rate as the total number of deaths divided the by total number of COVID cases.

```
summary(data$TotDeaths)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      58     641    2268    4454    4942   33247
```

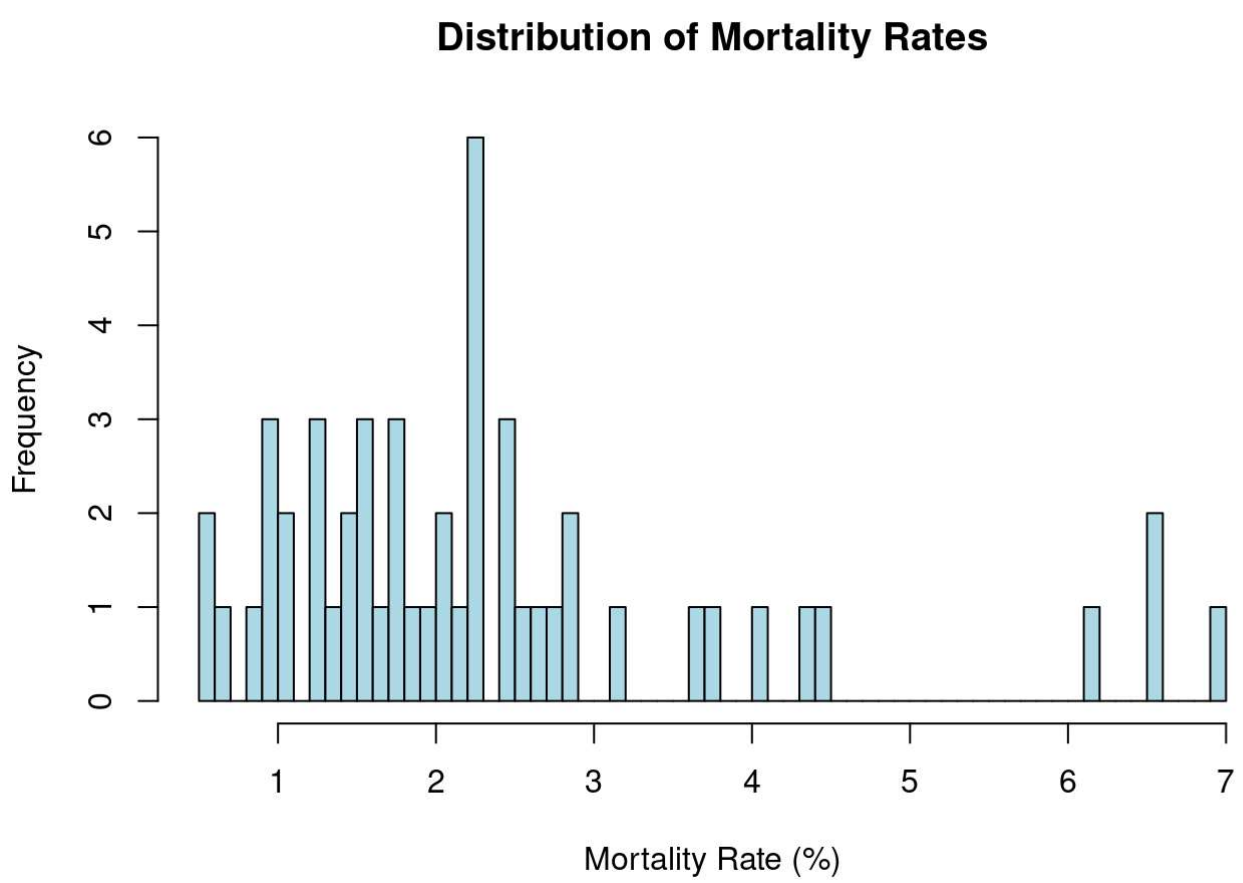
```
summary(data$TotCases)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2155   43396  119336  173583  196460  912904
```

```
data$Mortality_Rate <- (data$TotDeaths/data$TotCases)*100
summary(data$Mortality_Rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5327  1.4002  2.0280  2.3765  2.7401  6.9632
```

```
hist(data$Mortality_Rate, breaks=50,
  main = "Distribution of Mortality Rates",
  xlab = "Mortality Rate (%)", col="lightblue")
```



We have a clean data set with no glaring incorrect values or NAs. We have a non-zero, always positive dependent variable that has a positive skew and outliers - a good candidate to take the log of.

Now, we will perform an EDA on the independent variables that represent a percentage of COVID cases in each race (White, Black, and Hispanic). There is an “Other” category, but since “Other” encompasses so many different races, we did not think that was a good variable to use and make accurate conclusions from. We also do not have datasets broken out for races other than White, Black, and Hispanic.

```
glimpse(data$CasesWhitePerc)
```

```
##  chr [1:51] "0.52" "0.43" "0.37" "0.66" "0.17" "0.46" "0.49" "0.42" "0.23" ...
```

```
glimpse(data$CasesBlackPerc)
```

```
##  chr [1:51] "0.34" "0.05" "0.04" "0.23" "0.04" "0.04" "0.17" ...
```

```
glimpse(data$CasesHispPerc)
```

```
##  chr [1:51] "0.1" "0.12" "0.44" "0.16" "0.61" "0.43" "0.28000000000000003" ...
```

In all 3 data sets, there are a few things that need to be fixed.

- There are a number of data points called “NR” in the data sets. We will set these NRs to NA.

- The data is incorrectly cast as strings; we need to convert them into numbers.
- There are a few data points set to a string of “<.01”. We will need to change this string into a number. Since “<.01” signifies a very small number, we will replace these instances with 0.001.

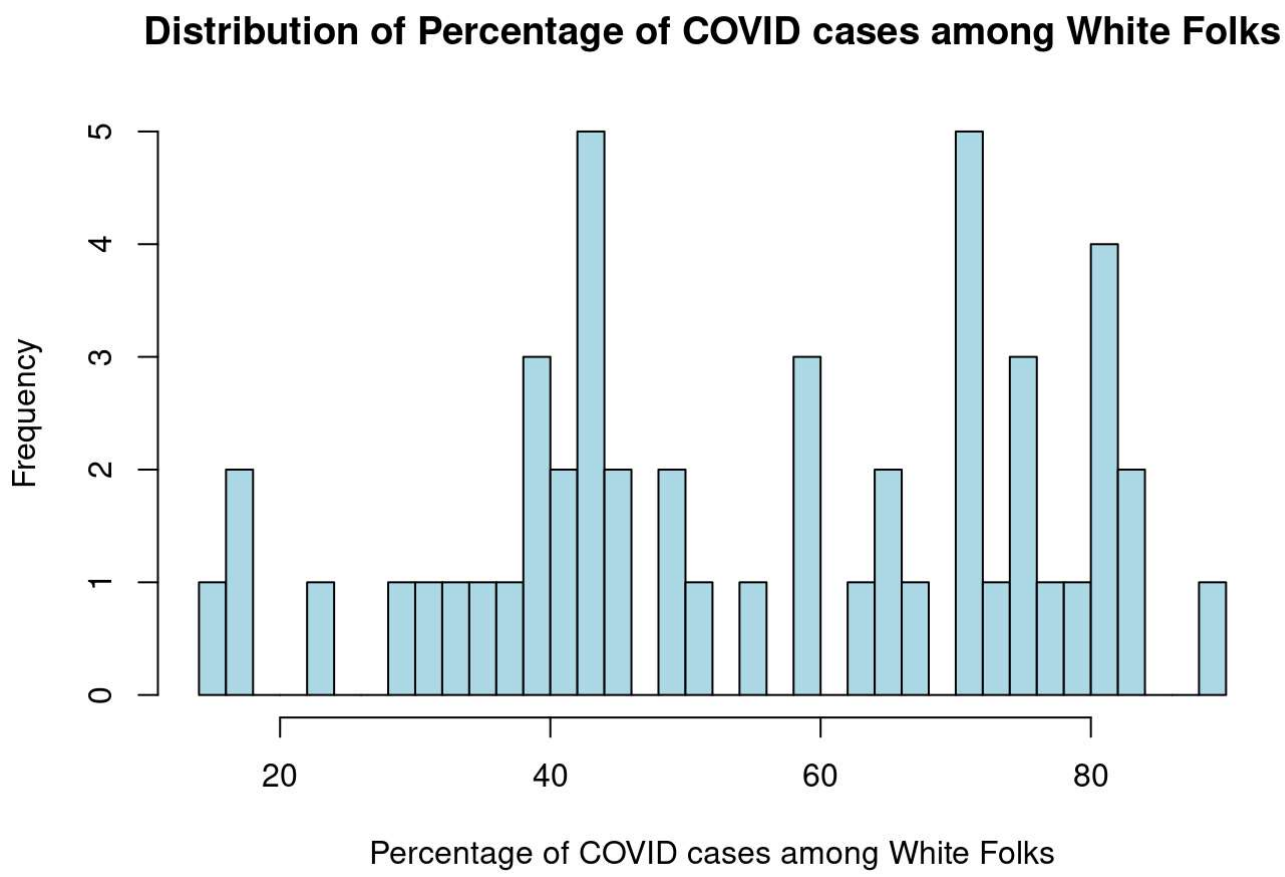
```
data <- data %>%
    mutate(CasesWhitePerc = na_if(CasesWhitePerc, "NR"))
data <- data %>%
    mutate(CasesBlackPerc = na_if(CasesBlackPerc, "NR"))
data <- data %>%
    mutate(CasesHispPerc = na_if(CasesHispPerc, "NR"))

data <- data %>%
  mutate(CasesWhitePerc = str_replace(CasesWhitePerc, "<.01", "0.001"))
data <- data %>%
  mutate(CasesBlackPerc = str_replace(CasesBlackPerc, "<.01", "0.001"))
data <- data %>%
  mutate(CasesHispPerc = str_replace(CasesHispPerc, "<.01", "0.001"))

data$CasesWhitePerc_clean <- as.numeric(data$CasesWhitePerc)*100
data$CasesBlackPerc_clean <- as.numeric(data$CasesBlackPerc)*100
data$CasesHispPerc_clean <- as.numeric(data$CasesHispPerc)*100
(length(data$CasesWhitePerc_clean)+length(data$CasesBlackPerc_clean)+length(data$CasesHispPerc_c
lean))/3
```

```
## [1] 51
```

```
summary(data$CasesWhitePerc_clean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   15.00   42.00   57.00   55.88   72.75   89.00       1
```

```
hist(data$CasesWhitePerc_clean, breaks=50,
  main = "Distribution of Percentage of COVID cases among White Folks",
  xlab = "Percentage of COVID cases among White Folks", col="lightblue")
```

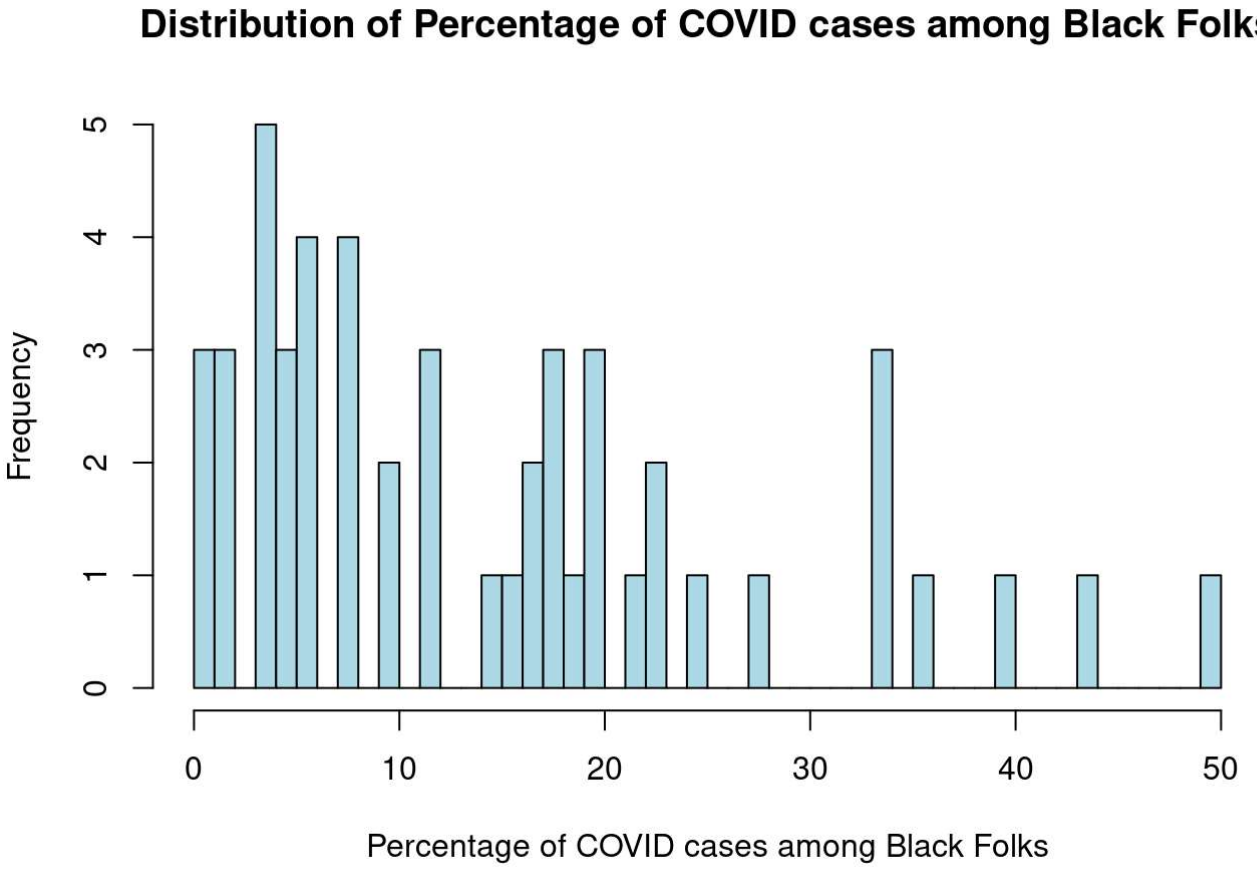


We have a clean data set here. Our histogram of the percentage of COVID Cases represented by the white race looks normally distributed with no significant outliers.

```
summary(data$CasesBlackPerc_clean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.10    5.00   12.00   14.92   20.00   50.00       1
```

```
hist(data$CasesBlackPerc_clean, breaks=50,
  main = "Distribution of Percentage of COVID cases among Black Folks",
  xlab = "Percentage of COVID cases among Black Folks", col="lightblue")
```

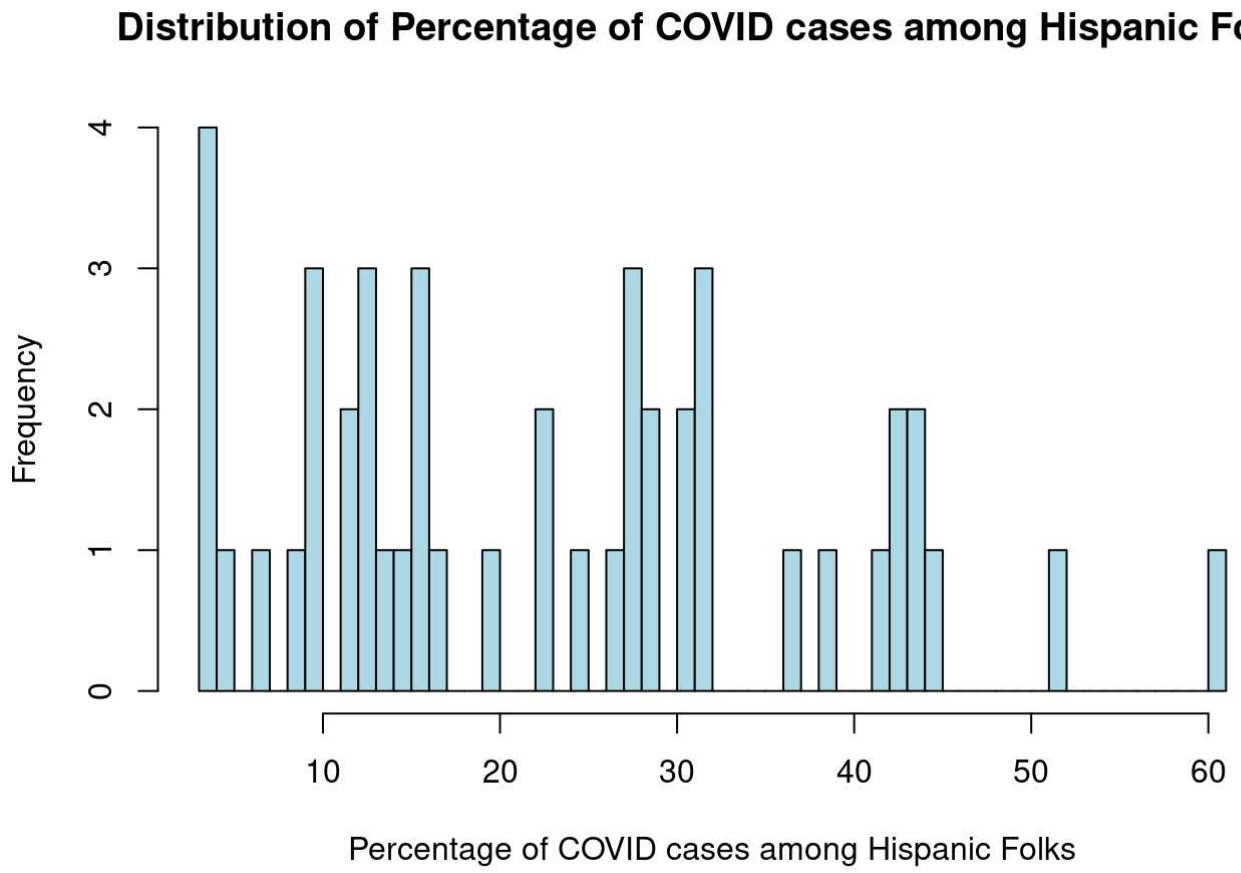


We also have a clean data set here. However, this histogram of the percentage of COVID Cases represented by the black race looks like it is positively skewed with some outliers. It's also a non-zero, always positive data set. Therefore, it is a good candidate to take the log of.

```
summary(data$CasesHispPerc_clean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3.00   12.25   23.00   23.72   32.00   61.00       5
```

```
hist(data$CasesHispPerc_clean, breaks=50,
  main = "Distribution of Percentage of COVID cases among Hispanic Folks",
  xlab = "Percentage of COVID cases among Hispanic Folks", col="lightblue")
```



We have a clean data set here as well. There is also somewhat of a positive skew in this histogram of percentage of COVID Cases represented by the hispanic race, though not as pronounced as the histogram we just discussed for the black race. We will decide whether or not to transform this variable based on what the plot versus our dependent variable (mortality_rate) shows later on in this report.

Next, we perform an EDA on our hospital preparedness independent variables.

```
summary(data$inpatient_beds)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     215    1226    3667    5489    6990   27177
```

```
summary(data$ICU_beds)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.0   246.5   510.0   890.0   853.5  4300.0
```

```
summary(data$num_COVID_patients)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      15     125     487    1184    1009   11043
```

The data looks clean. We will now take the total available inpatient and ICU beds per state and divide them by the total number of COVID patients per state. This way we elliminate the effect of having big states always at the top of our lists (since big states usually have the highest populations, they would always have most number of beds and COVID patients if we did not make this correction).

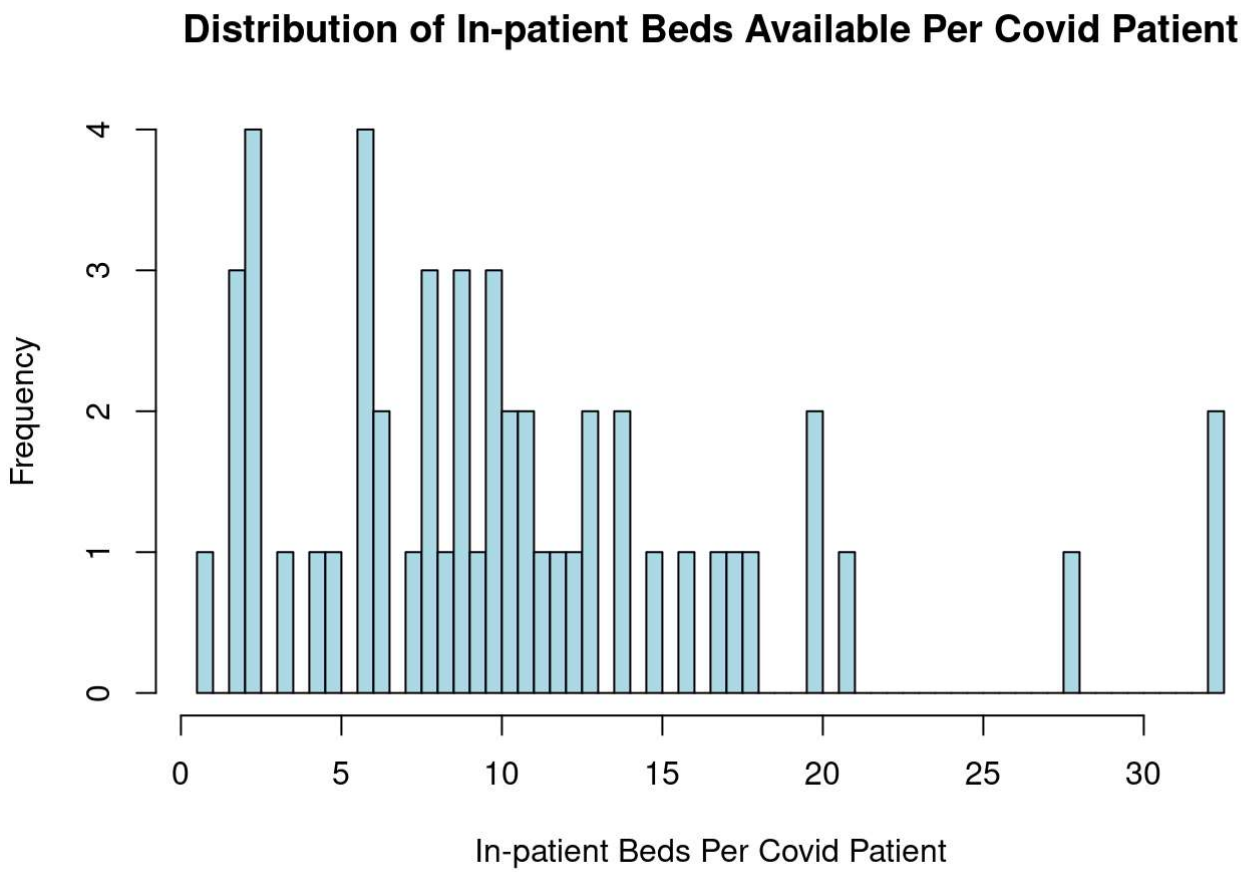
```
data$inpatient_beds_clean = data$inpatient_beds/data$num_COVID_patients
data$ICU_beds_clean = data$ICU_beds/data$num_COVID_patients

summary(data$inpatient_beds_clean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9836  5.6699  9.4200 10.4103 13.2099 32.1875
```

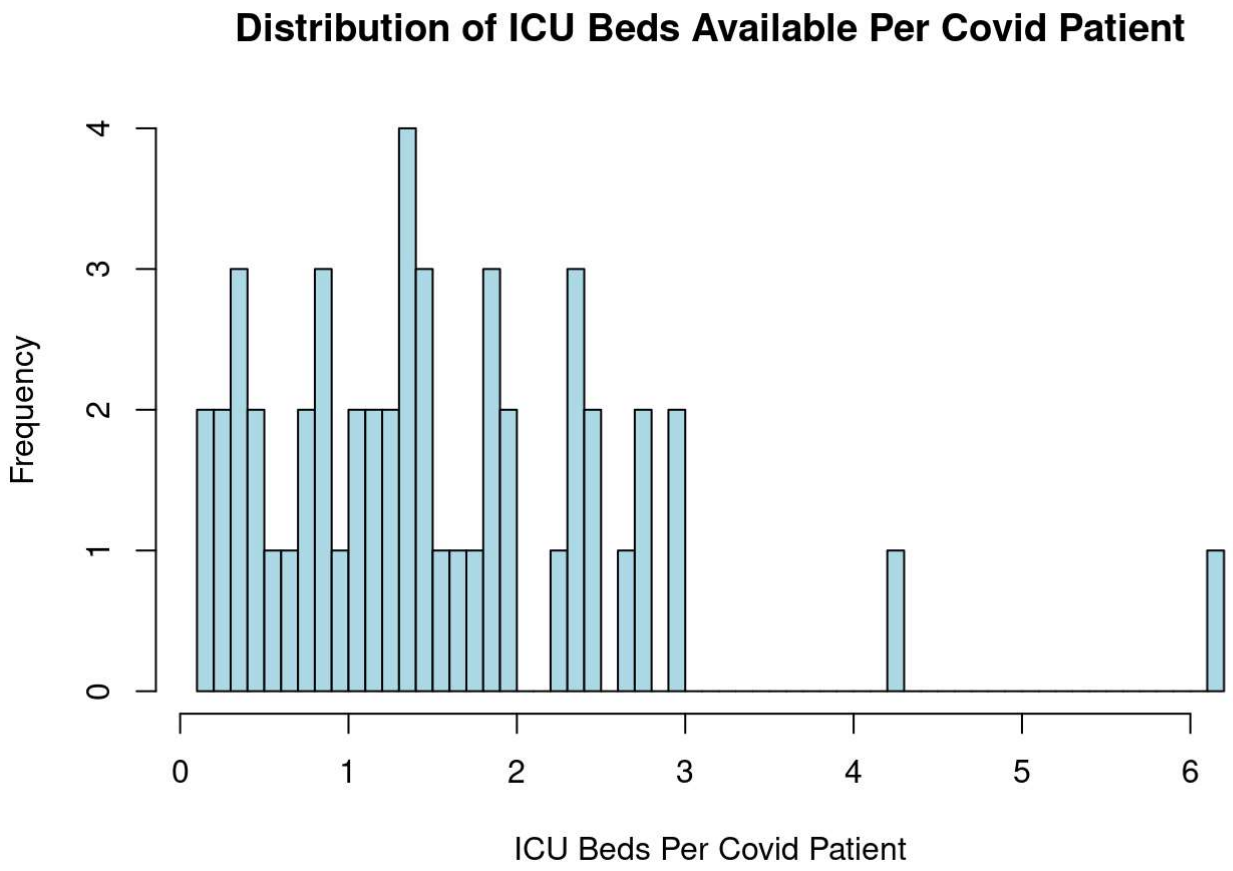
```
summary(data$ICU_beds_clean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1229  0.7964  1.3265  1.5375  2.1045  6.2000
```

```
hist(data$inpatient_beds_clean, breaks=50,
  main = "Distribution of In-patient Beds Available Per Covid Patient",
  xlab = "In-patient Beds Per Covid Patient", col="lightblue")
```



```
hist(data$ICU_beds_clean, breaks=50,
  main = "Distribution of ICU Beds Available Per Covid Patient",
  xlab = "ICU Beds Per Covid Patient", col="lightblue")
```



Our data looks clean and we have a slight positive skew for both variables. In this case, we will also decide whether or not to transform this variable based on what the plot versus our dependent variable (mortality_rate) shows later on in this report.

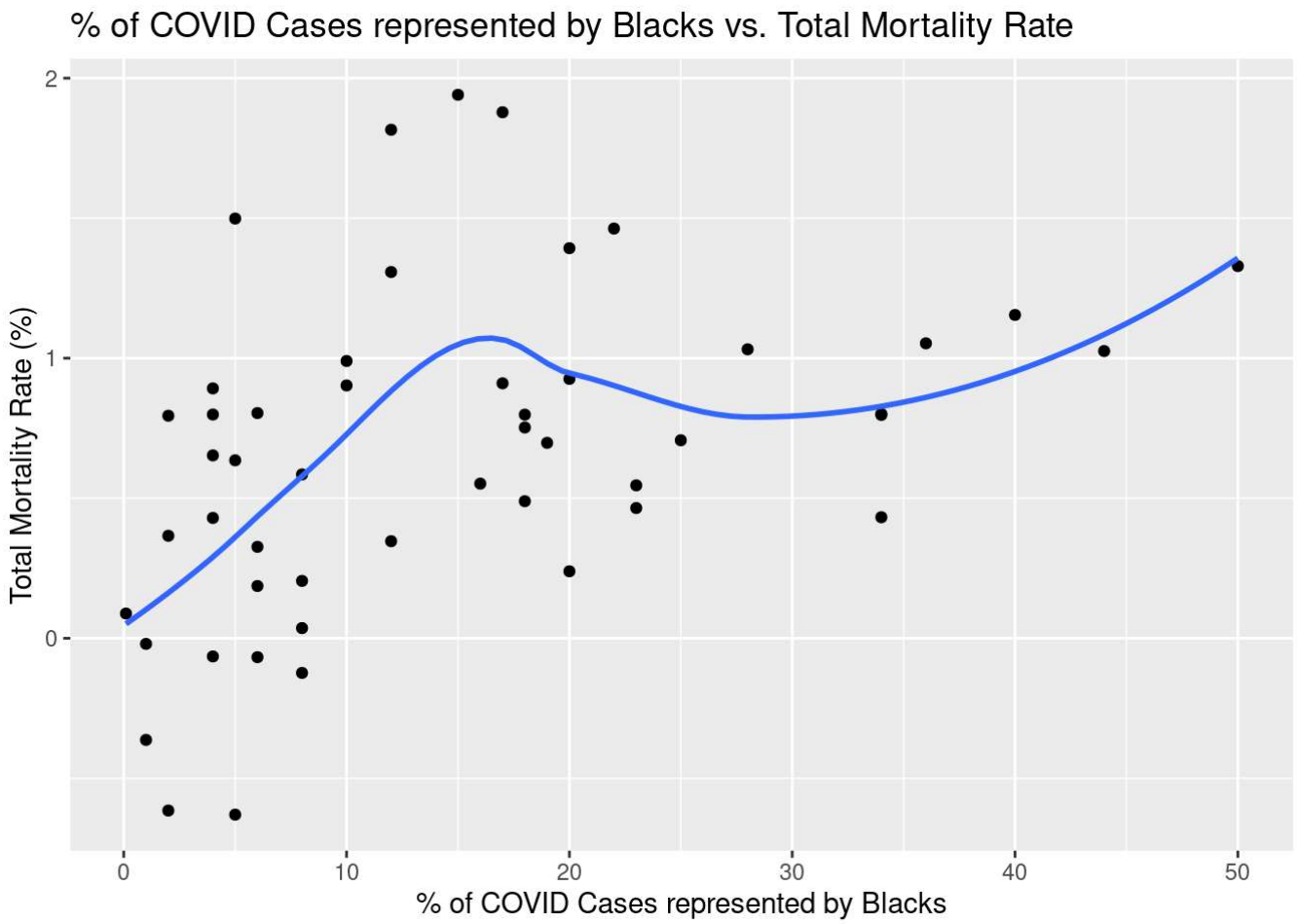
Disclosure: Even though we use the number of COVID patients (“num_COVID_patients”) in both variables’ denominators, our “num_COVID_patients” variable measures the number of COVID patients in an inpatient care location who have suspected or confirmed COVID-19. We do not have the data for number of COVID patients in an ICU care location, so we can only use what we have as a substitute.

Variable Transformation Analysis

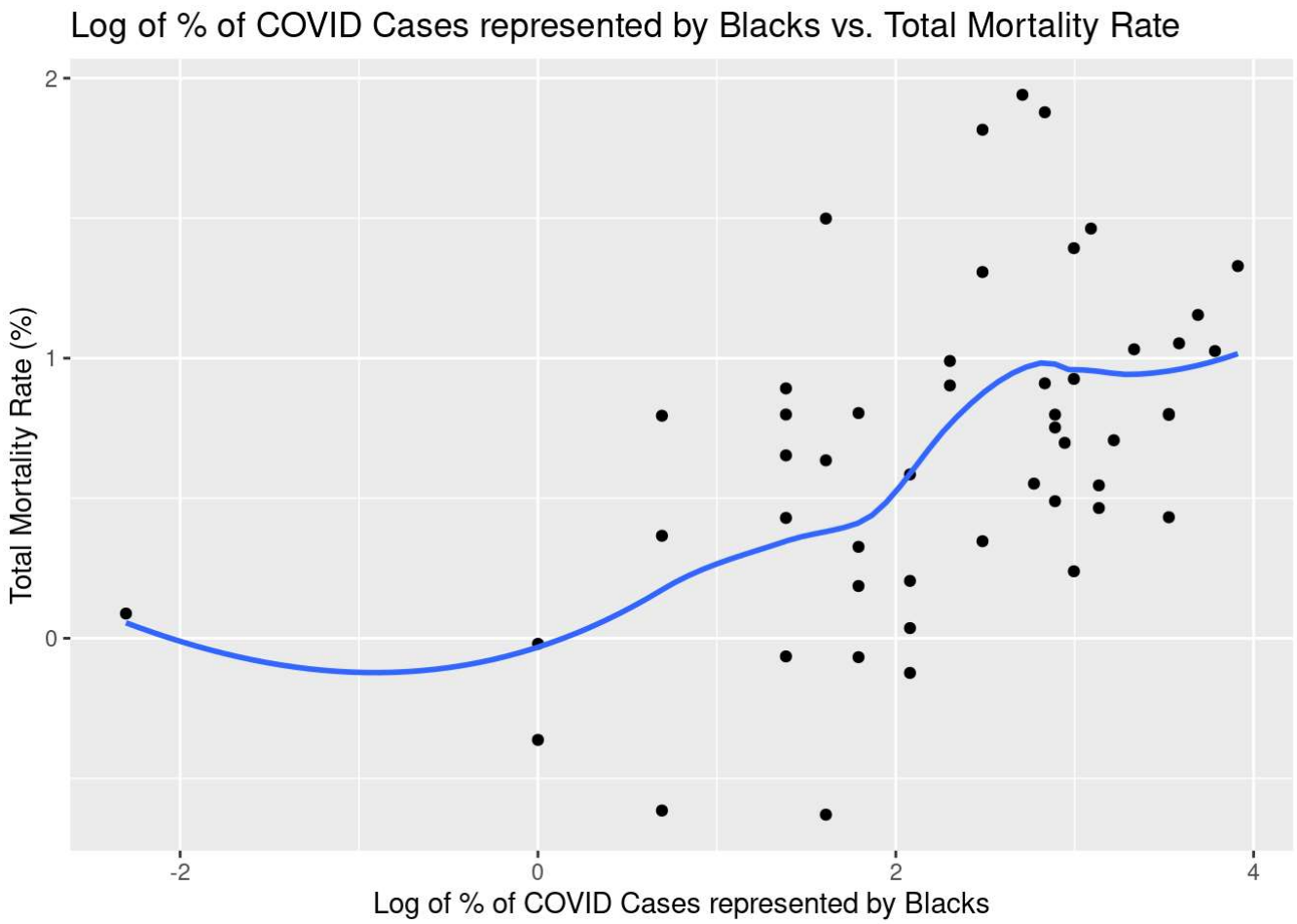
```
data %>%
  ggplot(aes(x = CasesWhitePerc_clean, y = log(Mortality_Rate))) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(
    title = '% of COVID Cases represented by Whites vs. Total Mortality Rate',
    x = '% of COVID Cases represented by Whites',
    y = 'Total Mortality Rate (%)'
  )
```

The % of COVID Cases represented by Whites" variable looks to have a flat/descending linear relationship with total mortality rate, so we are not going to transform this variable any further.

```
data %>%
  ggplot(aes(x = CasesBlackPerc_clean, y = log(Mortality_Rate))) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(
    title = '% of COVID Cases represented by Blacks vs. Total Mortality Rate',
    x = '% of COVID Cases represented by Blacks',
    y = 'Total Mortality Rate (%)'
  )
```

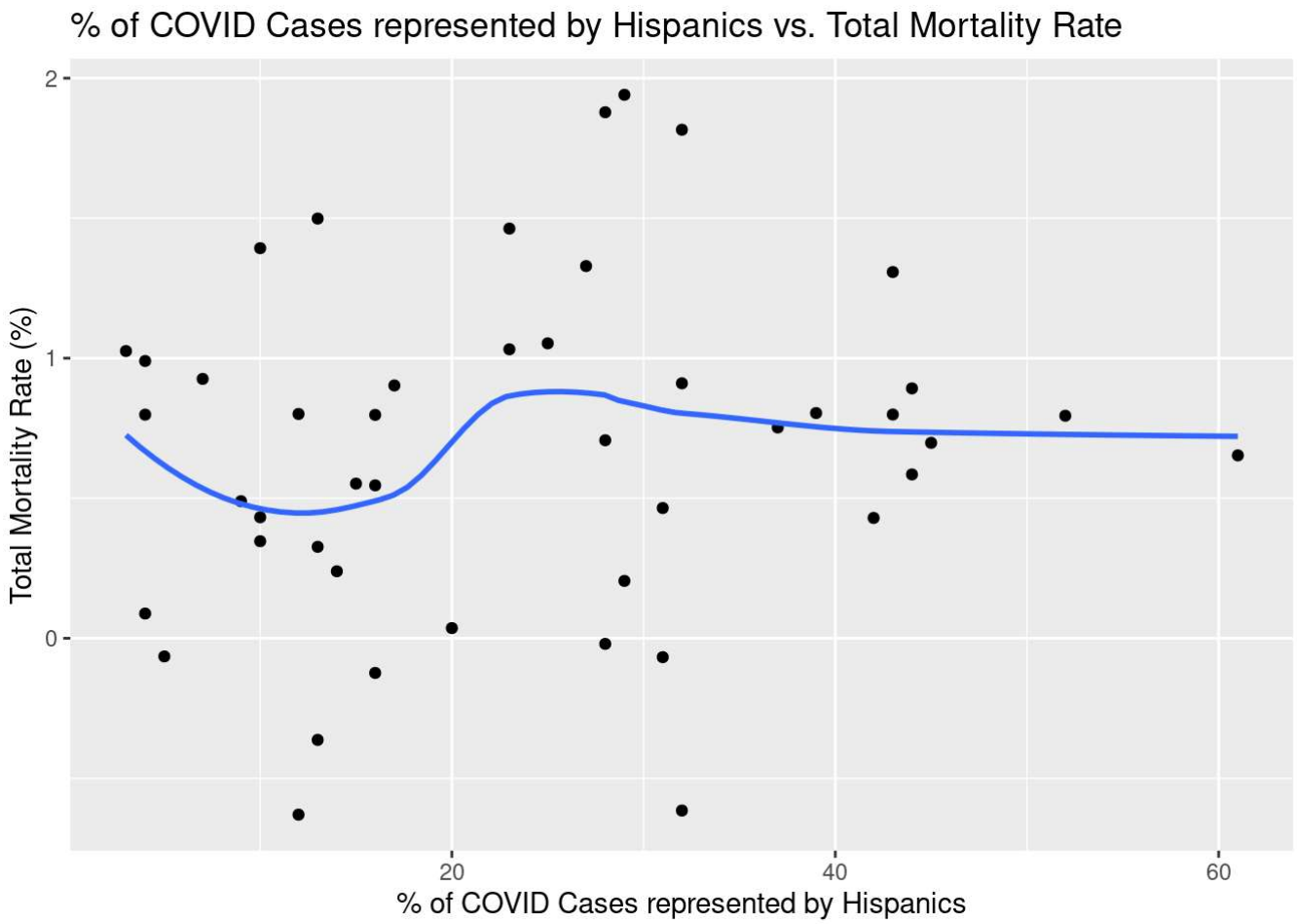



```
data %>%
  ggplot(aes(x = log(CasesBlackPerc_clean), y = log(Mortality_Rate))) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(
    title = 'Log of % of COVID Cases represented by Blacks vs. Total Mortality Rate',
    x = 'Log of % of COVID Cases represented by Blacks',
    y = 'Total Mortality Rate (%)'
  )
```



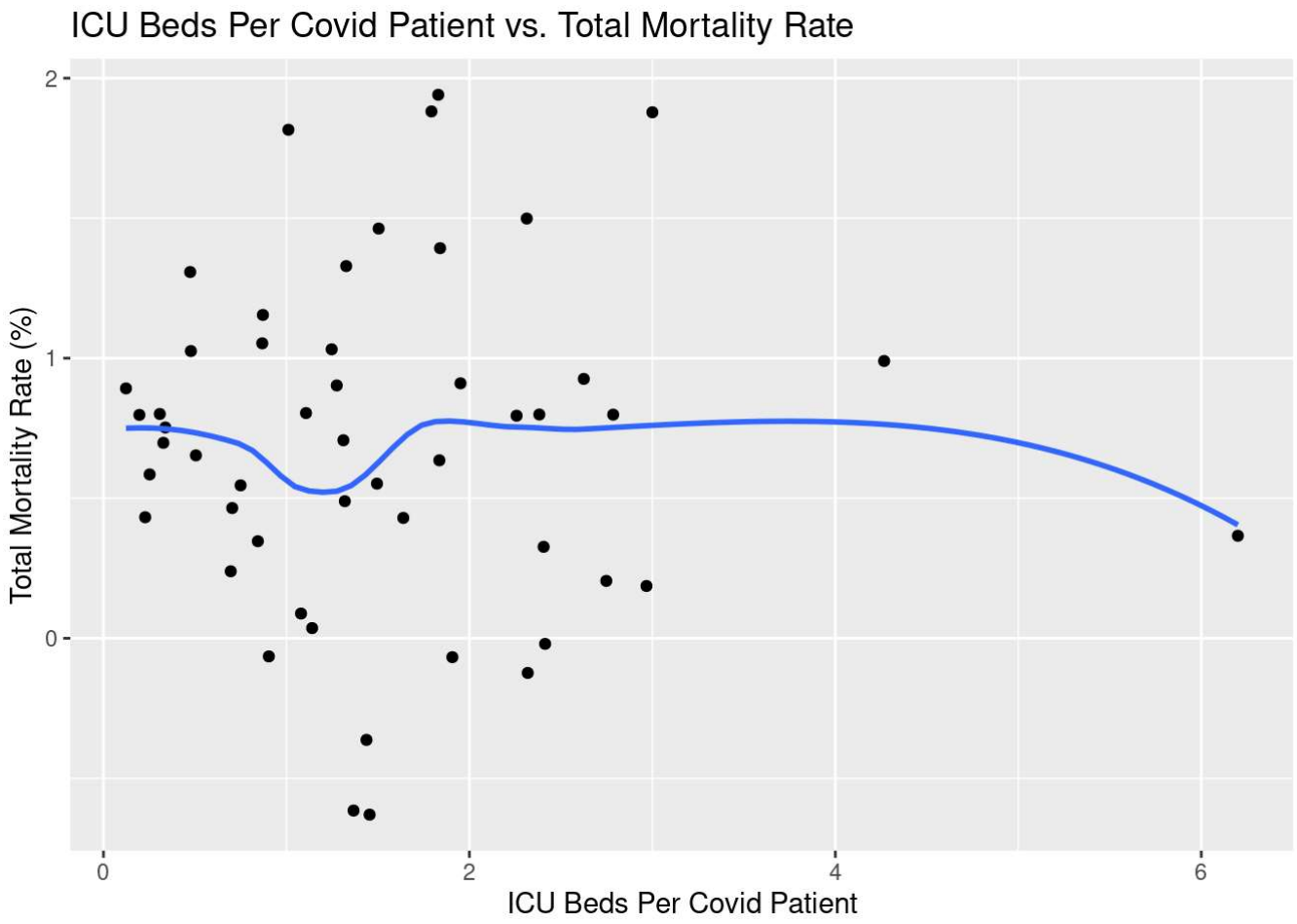
We have a non-linear relationship between the “% of COVID Cases represented by Blacks” variable and our dependent variable. We see more linearity from the log() version of “% of COVID Cases represented by Blacks”, so we will proceed with transforming the variable into its log() version.

```
data %>%
  ggplot(aes(x = CasesHispPerc_clean, y = log(Mortality_Rate))) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(
    title = '% of COVID Cases represented by Hispanics vs. Total Mortality Rate',
    x = '% of COVID Cases represented by Hispanics',
    y = 'Total Mortality Rate (%)'
  )
```



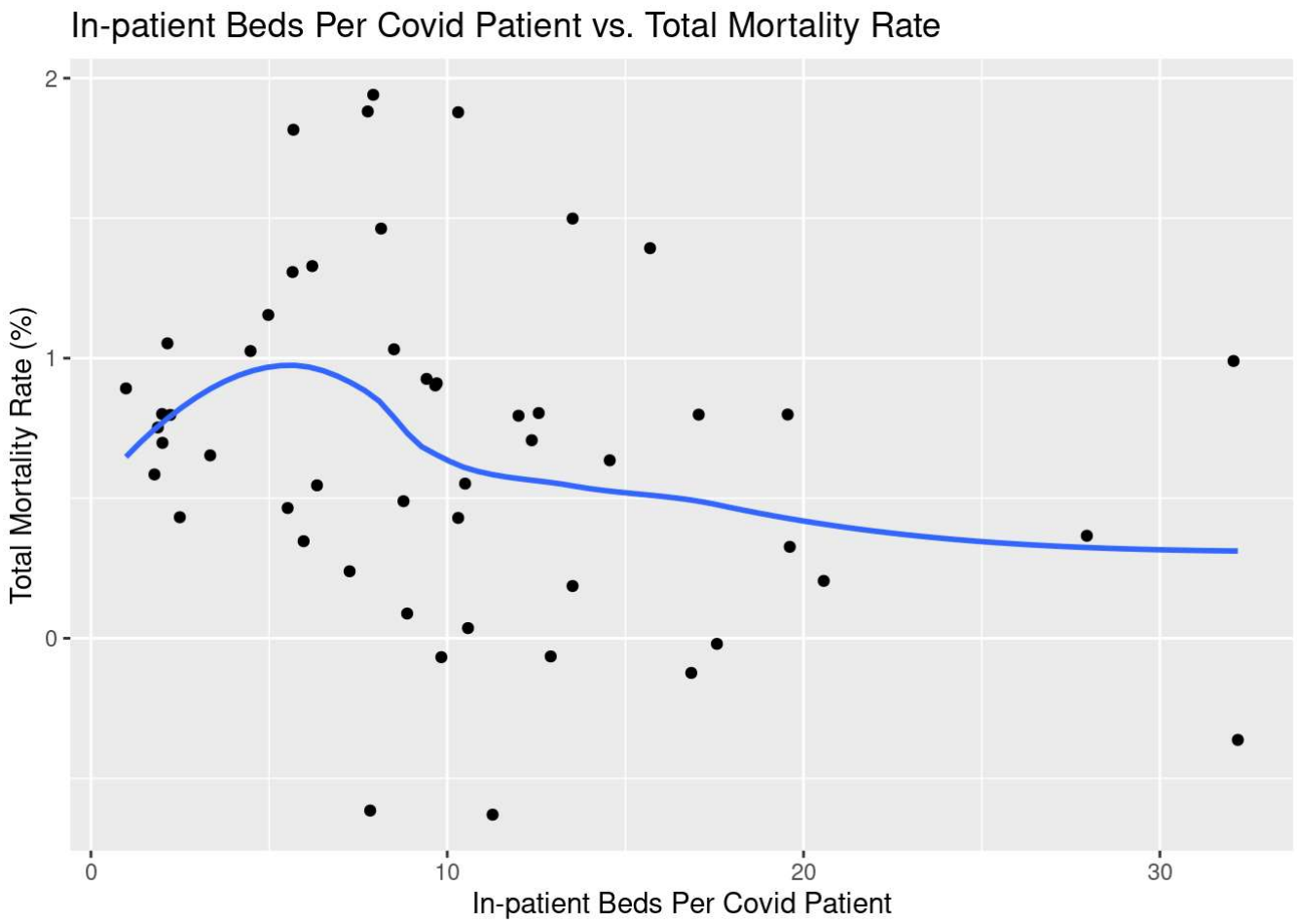
We get a flat, linear line here, so we do not need to further transform the “% of COVID Cases represented by Hispanics” variable.

```
data %>%
  ggplot(aes(x = ICU_beds_clean, y = log(Mortality_Rate))) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(
    title = 'ICU Beds Per Covid Patient vs. Total Mortality Rate',
    x = 'ICU Beds Per Covid Patient',
    y = 'Total Mortality Rate (%)'
  )
```



Excluding the outlier, our graph shows a pretty flat, linear line here, so we do not need to further transform the ICU Beds per Covid Patient variable.

```
data %>%
  ggplot(aes(x = inpatient_beds_clean, y = log(Mortality_Rate))) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(
    title = 'In-patient Beds Per Covid Patient vs. Total Mortality Rate',
    x = 'In-patient Beds Per Covid Patient',
    y = 'Total Mortality Rate (%)'
  )
```



We get a pretty flat/descending, linear relationship here, so we do not need to further transform the In-patient Beds per Covid Patient variable.

Model #1 + Respective CLM Assumptions

```
model1 <- lm(log(Mortality_Rate) ~
  CasesWhitePerc_clean +
  log(CasesBlackPerc_clean) +
  CasesHispPerc_clean
  , data = data,
  na.action=na.omit)
summary(model1)
```

```
##
## Call:
## lm(formula = log(Mortality_Rate) ~ CasesWhitePerc_clean + log(CasesBlackPerc_clean) +
##     CasesHispPerc_clean, data = data, na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2149 -0.2915 -0.1089  0.2111  1.1429
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.497840   0.557039   0.894  0.37656
## CasesWhitePerc_clean      -0.007016   0.005698  -1.231  0.22502
## log(CasesBlackPerc_clean)  0.223718   0.072209   3.098  0.00347 **
## CasesHispPerc_clean        0.002401   0.007574   0.317  0.75276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5175 on 42 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.3013, Adjusted R-squared:  0.2514
## F-statistic: 6.037 on 3 and 42 DF,  p-value: 0.001631
```

Interpretation/analysis of model 1 will be presented on the stargazer section further in our report below.

Discussion of 6 CLM Assumptions for Model #1:

1. Linear in Parameters: We do not need to assess this. This condition is always met if the dependent variable is a linear function (lm) of the explanatory variables.

2. Random iid sampling:

To assess if the data is IID, we need to know more about the sampling process. It is hard to justify this assumption because of the way the virus spreads. There are several reasons we might expect why our COVID data may not be independent of each other.

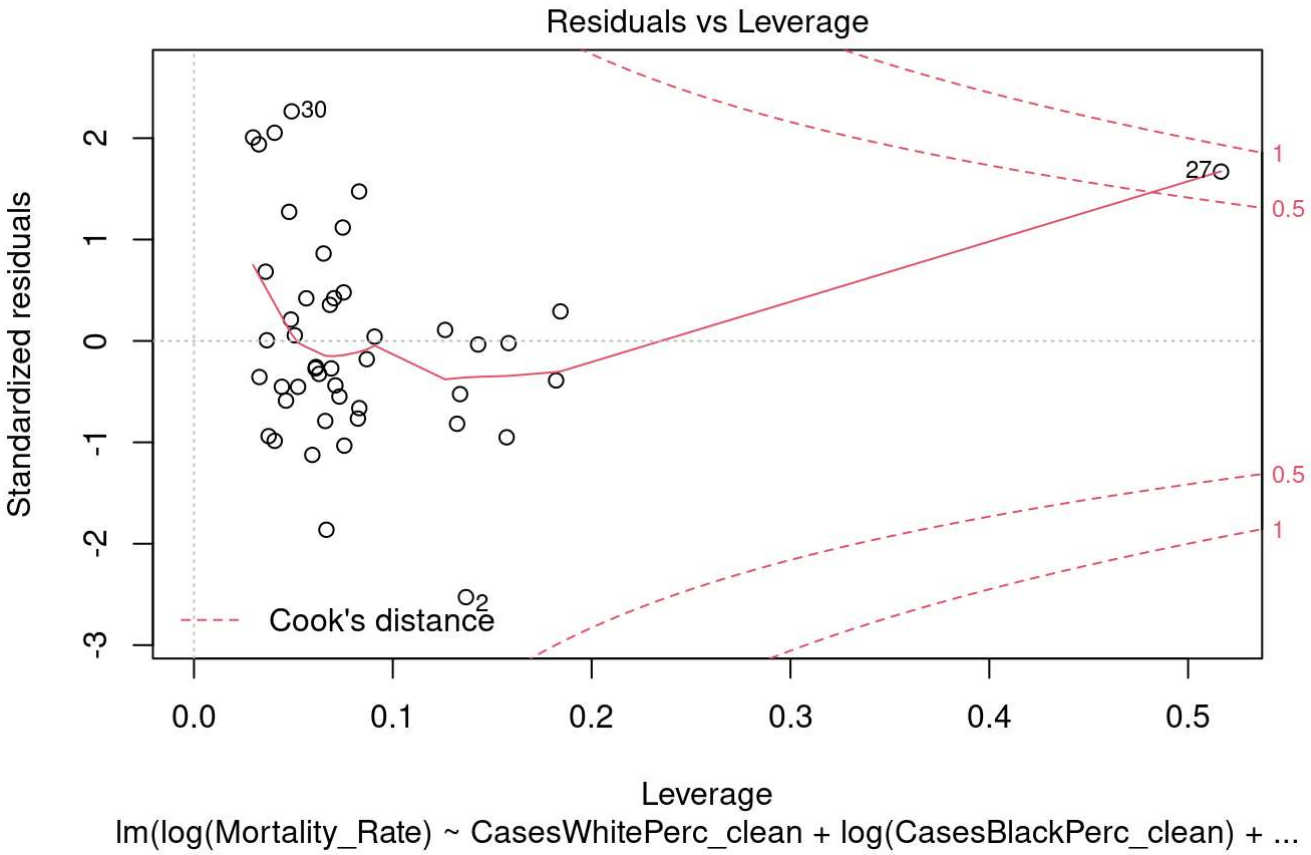
Since COVID is a virus, you can only contract it if you come in contact with someone else with COVID. Therefore, many people in the same vicinity/area will likely contract COVID (clustering).

Since we have data from the 51 states only, we do not have identical distribution, since the states' geographical location are set and cannot be changed. Neighbouring states are most likely to exhibit similar COVID case rate/behvior due to limited access to transportation and movement of people (people preferring cars over planes).

We tried to elliminate the effects of large states vs. small states by dividing by the number of COVID patients, as well as taking percentages rather than nominal numbers whenever possible. However, each state's geographical location cannot be changed.

Our research question, in part, is testing for independent and identically distribution in terms of affecting each human being equally, regardless of race. However, we know that the incubation period for the coronavirus is independent and identically distributed.

```
plot(model1, which=5)
```



We have an outlier (with high leverage) vs. the regression line we predicted, propelling the Cook's distance line higher towards the tail end of the graph. This outlier represents New York, as that state had a considerable spike in COVID cases, hence mortality rate, in the beginning of the COVID pandemic. We opted to keep the outlier since it preserves the integrity of our data set.

However, for the rest of our data points (the other 50 states), the Cooks distance line looks like it is pretty flat (what we wanted to see).

3. No perfect multicollinearity

```
vif(model1)
```

```
##      CasesWhitePerc_clean log(CasesBlackPerc_clean)       CasesHispPerc_clean
##                  2.077124                  1.204228                  2.018241
```

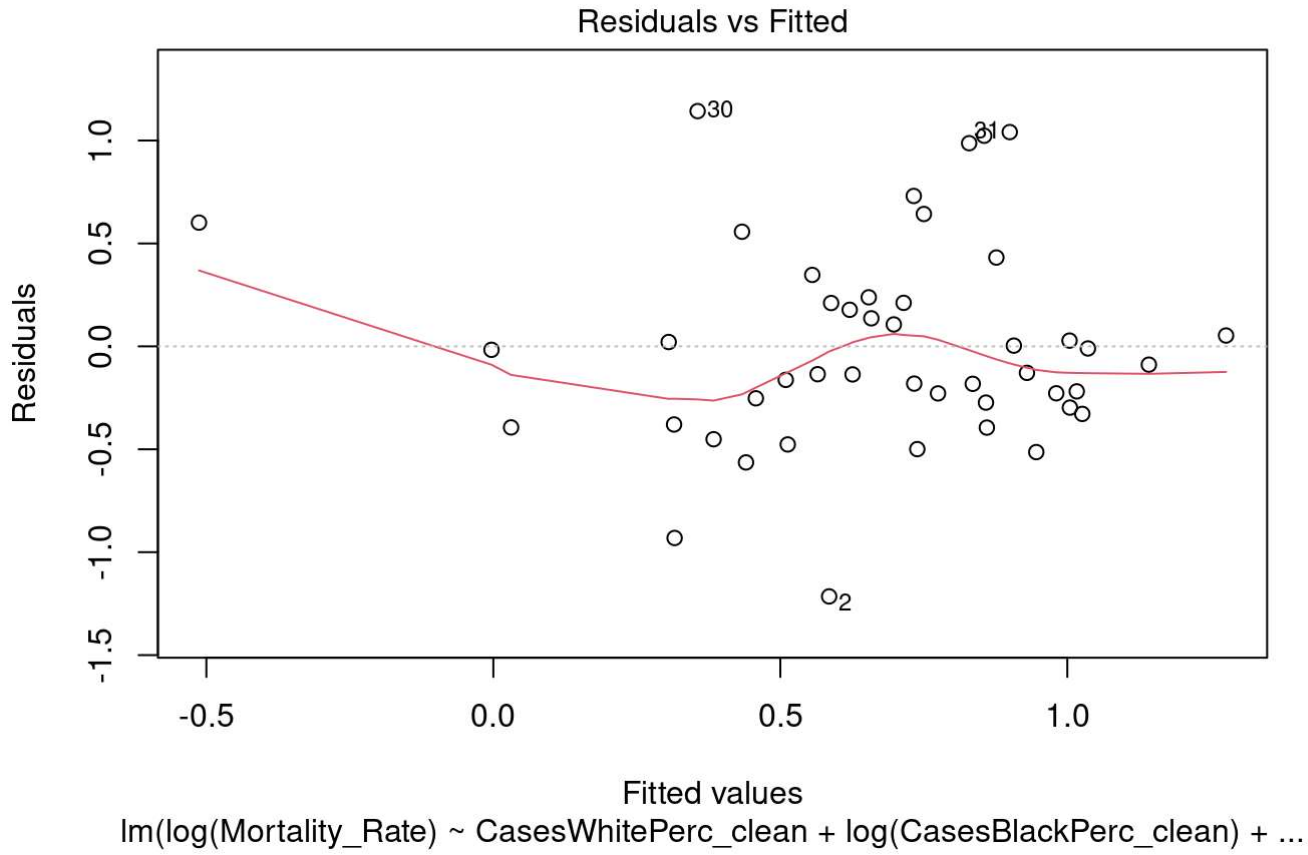
```
vif(model1) > 4
```

```
##      CasesWhitePerc_clean log(CasesBlackPerc_clean)       CasesHispPerc_clean
##                     FALSE                     FALSE                     FALSE
```

Our variables have a VIF less than 4, signalling no issues with multicollinearity.

4. Zero conditional mean

```
plot(model1, which=1)
```



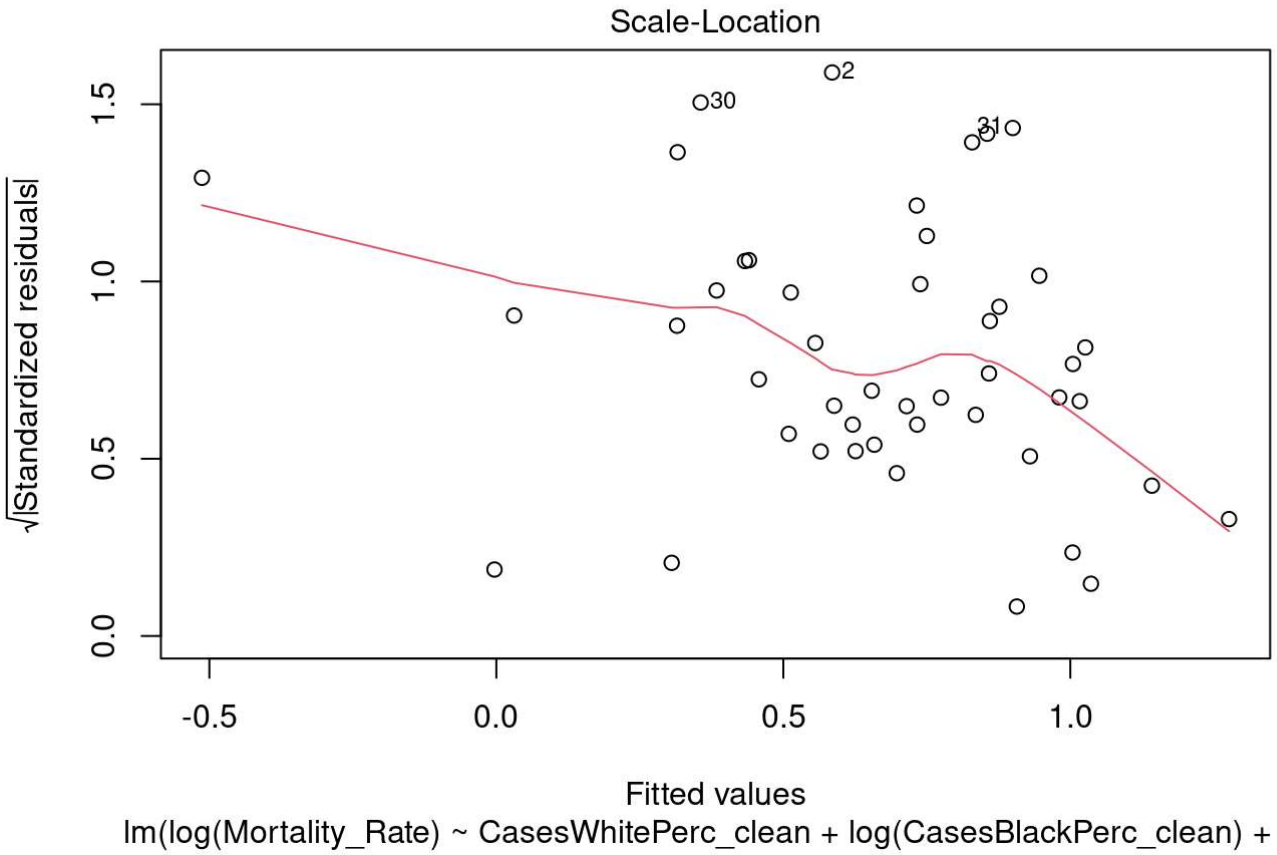
We have a pretty flat red line along 0, signalling we have linear conditional expectation.

5. Homoskedasticity

```
plot(model1, which=1)
```

Other than the front and the back ends of the plot where we don't have that many samples, our Residuals vs. Fitted values has uniform thickness around the errors, signaling homoskedasticity.

```
plot(model1, which=3)
```



Our Scale Location plot does not really have a horizontal band of points all throughout the graph. We also have an-almost linear, descending red line, something that signals we may have heteroskedacity. We will aim to correct this in our 2nd model.

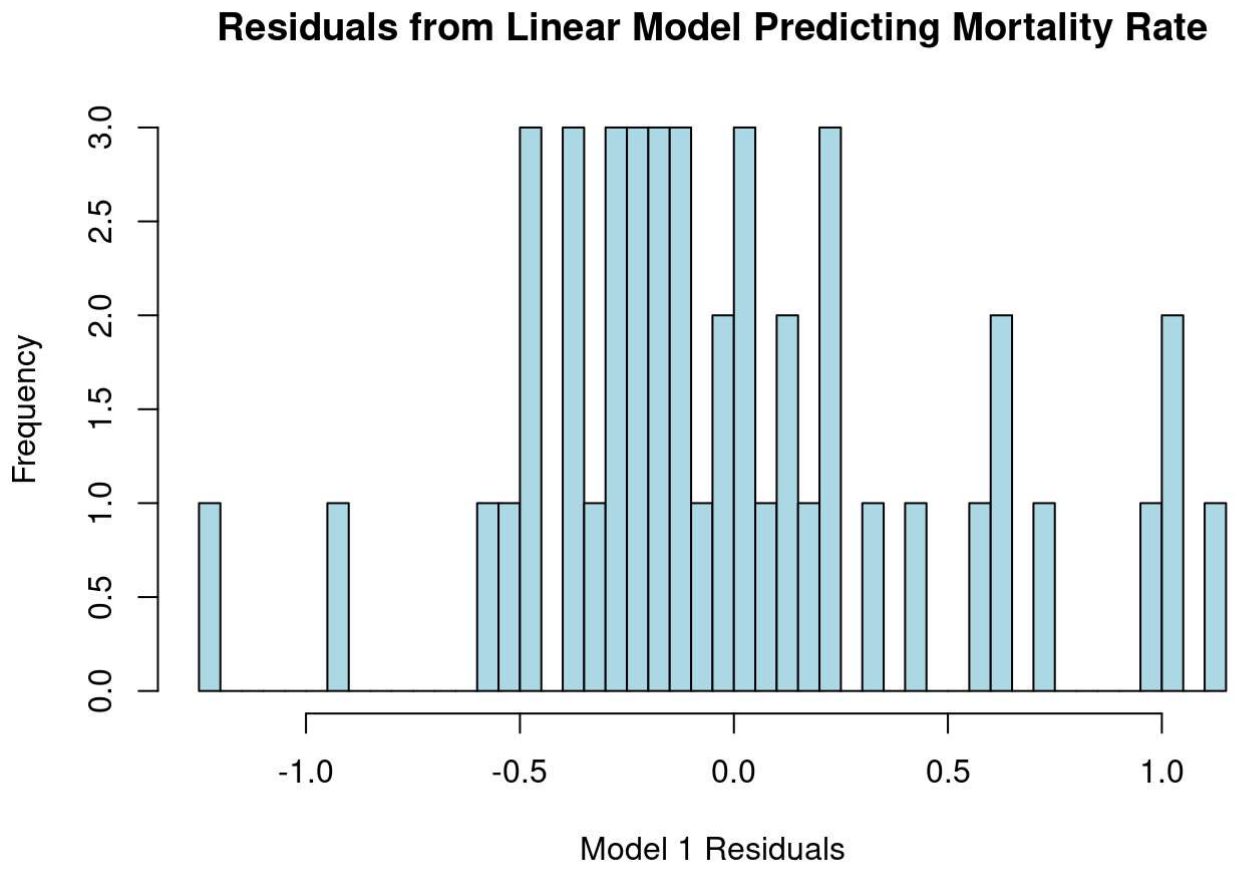
```
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 2.2008, df = 3, p-value = 0.5318
```

Our Breusch-Pagan test generates a high p-value, which means we fail to reject the null. This is a good thing; this signals that our variables are homoskedastic. (Our null hypothesis, (H_0), states that our variables are homoskedastic, and we have failed to reject this null hypothesis.)

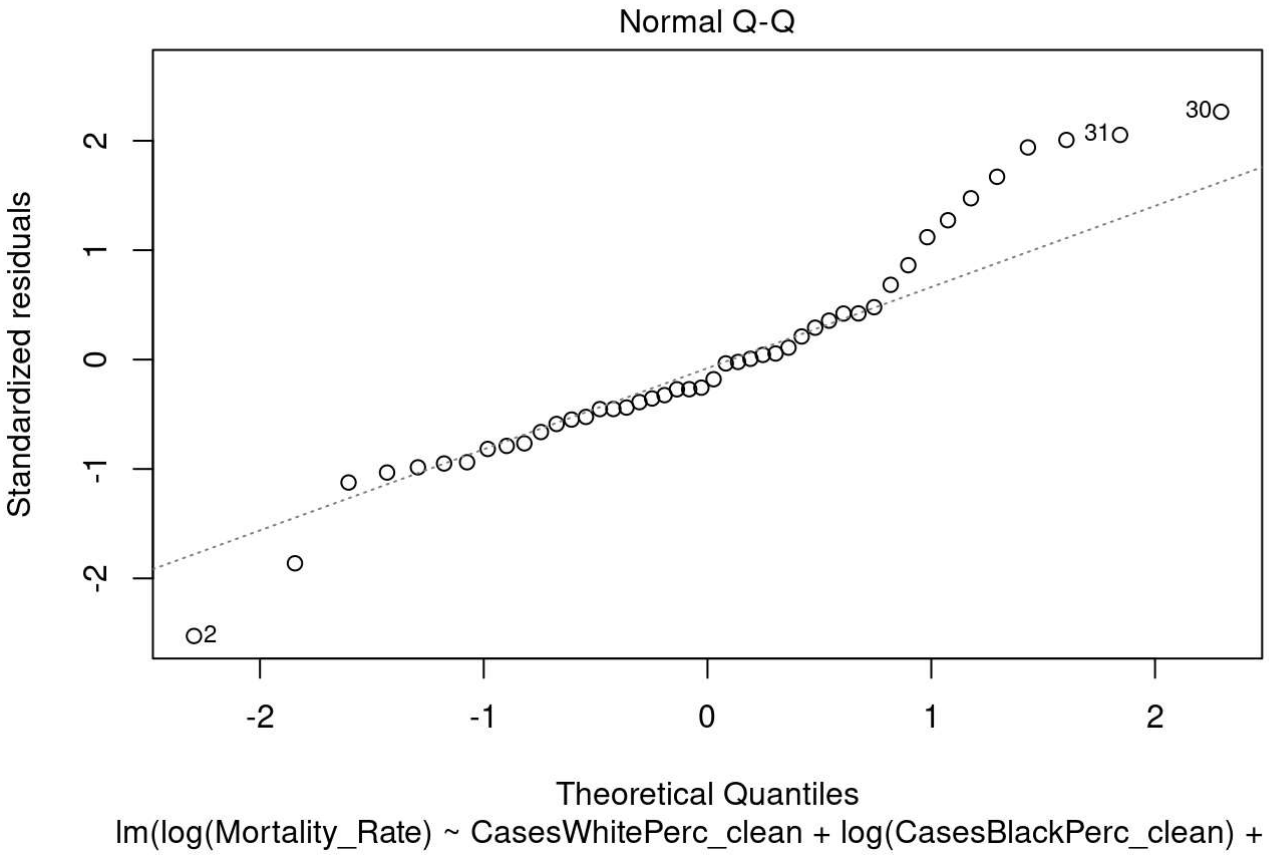
6. Normality of Error Terms

```
hist(model1$residuals, breaks = 50,
  main = "Residuals from Linear Model Predicting Mortality Rate",
  xlab = "Model 1 Residuals",
  col="lightblue")
```



Though we do not have that many data points, we can see the outline of the normal distribution we expect to see in our residuals, which demonstrates that we have normality in our error terms.

```
plot(model1, which=2)
```



In the qqplot graph, we see that most of the residuals fall along the linear, diagonal line. However, there is some deviation at the extremes towards the beginning and the end where there are not that many samples.

```
shapiro.test(model1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.95634, p-value = 0.08255
```

Since we have a p-value greater than 0.05 in the Shapiro-Wilk Test, we again confirm that our residuals are normally distributed. In statistical terms, we have failed to reject the null hypothesis (H_0 = our residuals are normally distributed).

Model #2 + Respective CLM Assumptions

We now proceed with assessing the 6 CLM assumptions with our 2nd model:

```
model2 <- lm(log(Mortality_Rate) ~
  CasesWhitePerc_clean +
  log(CasesBlackPerc_clean) +
  CasesHispPerc_clean +
  ICU_beds_clean
  , data = data,
  na.action=na.omit)
summary(model2)
```

```
##
## Call:
## lm(formula = log(Mortality_Rate) ~ CasesWhitePerc_clean + log(CasesBlackPerc_clean) +
##     CasesHispPerc_clean + ICU_beds_clean, data = data, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31877 -0.20501 -0.07406  0.08626  1.03134
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.5180845  0.5228423   0.991  0.32755
## CasesWhitePerc_clean      -0.0128294  0.0058007  -2.212  0.03262 *
## log(CasesBlackPerc_clean)  0.2341048  0.0678875   3.448  0.00132 **
## CasesHispPerc_clean        0.0003535  0.0071520   0.049  0.96082
## ICU_beds_clean             0.2347399  0.0907930   2.585  0.01338 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4857 on 41 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.3992, Adjusted R-squared:  0.3406
## F-statistic: 6.811 on 4 and 41 DF,  p-value: 0.000267
```

Interpretation/analysis of model 2 will be presented on the stargazer section further in our report below.

Discussion of 6 CLM Assumptions for Model #2:

1. Linear in Parameters: We do not need to assess this. This condition is always met if the dependent variable is a linear function (lm) of the explanatory variables.

2. Random iid sampling:

To assess if the data is IID, we need to know more about the sampling process. It is hard to justify this assumption because of the way the virus spreads. There are several reasons we might expect why our COVID data may not be independent of each other.

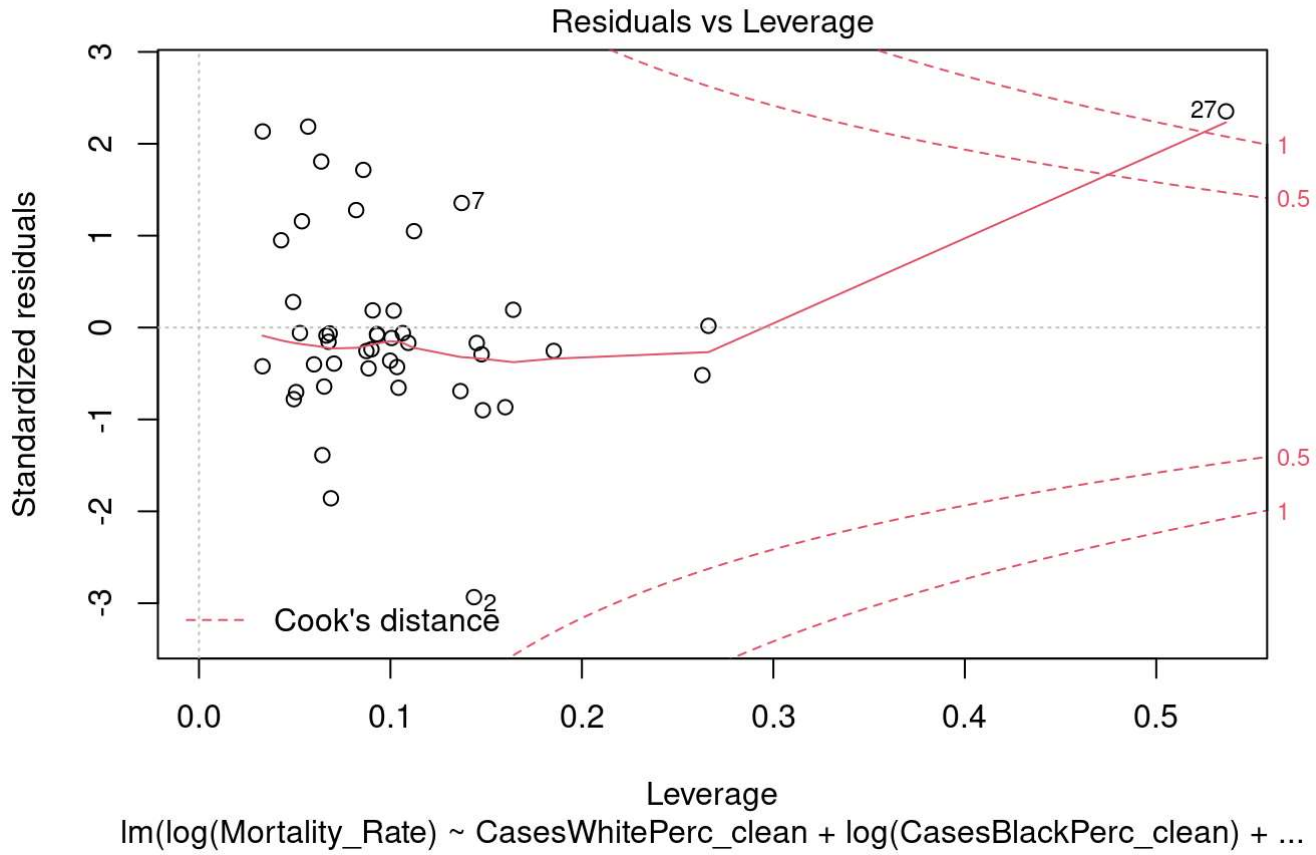
Since COVID is a virus, you can only contract it if you come in contact with someone else with COVID. Therefore, many people in the same vicinity/area will likely contract COVID (clustering).

Since we have data from the 51 states only, we do not have identical distribution, since the states' geographical location are set and cannot be changed. Neighbouring states are most likely to exhibit similar COVID case rate/behvior due to limited access to transportation and movement of people (people preferring cars over planes).

We tried to elliminate the effects of large states vs. small states by dividing by the number of COVID patients, as well as taking percentages rather than nominal numbers whenever possible. However, each state's geographical location cannot be changed.

Our research question, in part, is testing for independent and identically distribution in terms of affecting each human being equally, regardless of race. However, we know that the incubation period for the coronavirus is independent and identically distributed.

```
plot(model2, which=5)
```



We continue to have an outlier (with high leverage) vs. the regression line we predicted, propelling the Cook's distance line higher towards the latter end of the graph.

However, for the rest of our data points (the other 50 states), the Cooks distance line looks like it is pretty flat (what we wanted to see).

3. No perfect multicollinearity

```
vif(model2)
```

```
##      CasesWhitePerc_clean log(CasesBlackPerc_clean)       CasesHispPerc_clean
##                  2.444407                  1.208460                  2.043305
##            ICU_beds_clean
##                  1.290049
```

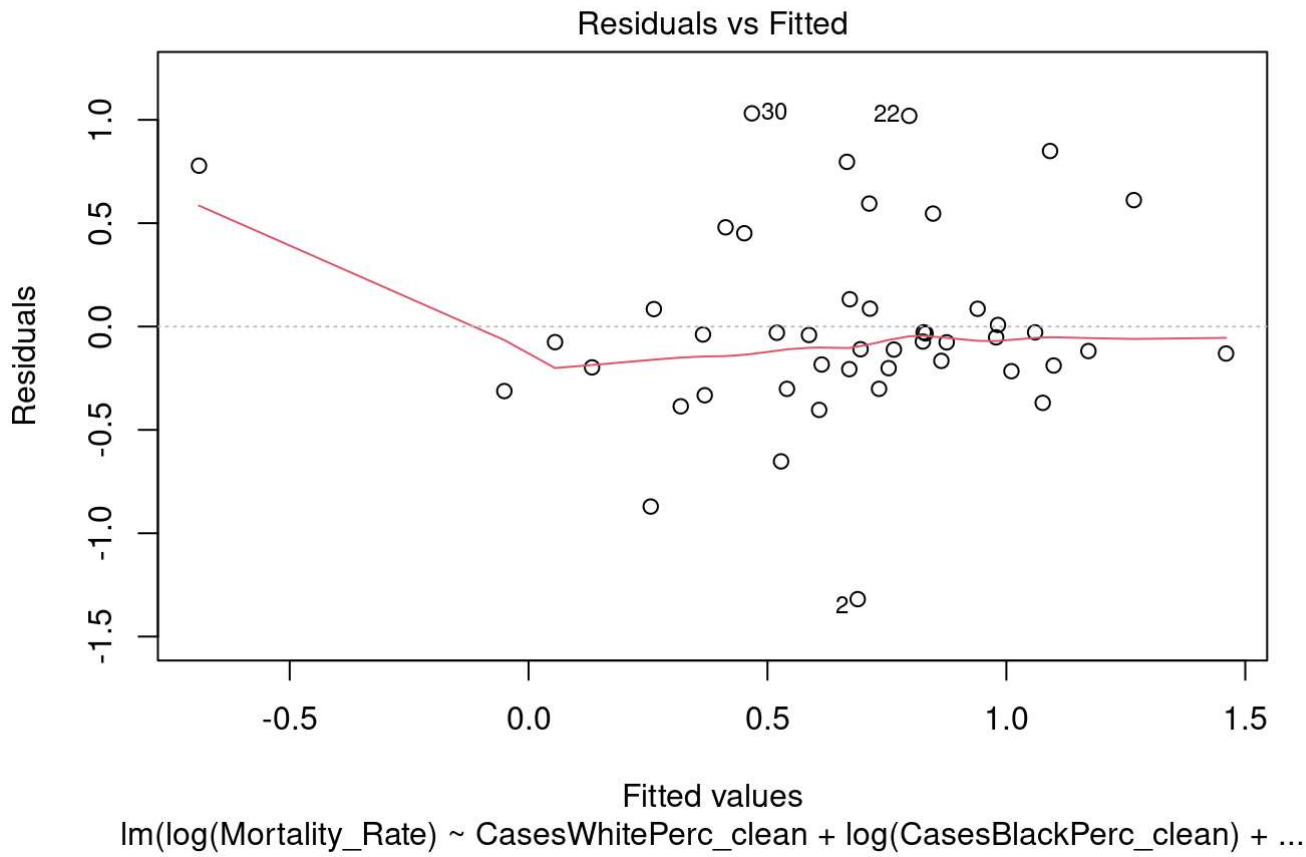
```
vif(model2) > 4
```

```
##      CasesWhitePerc_clean log(CasesBlackPerc_clean)       CasesHispPerc_clean
##                     FALSE                     FALSE                     FALSE
##           ICU_beds_clean
##                     FALSE
```

We do not have multicollinearity issues for any of our variables.

4. Zero conditional mean

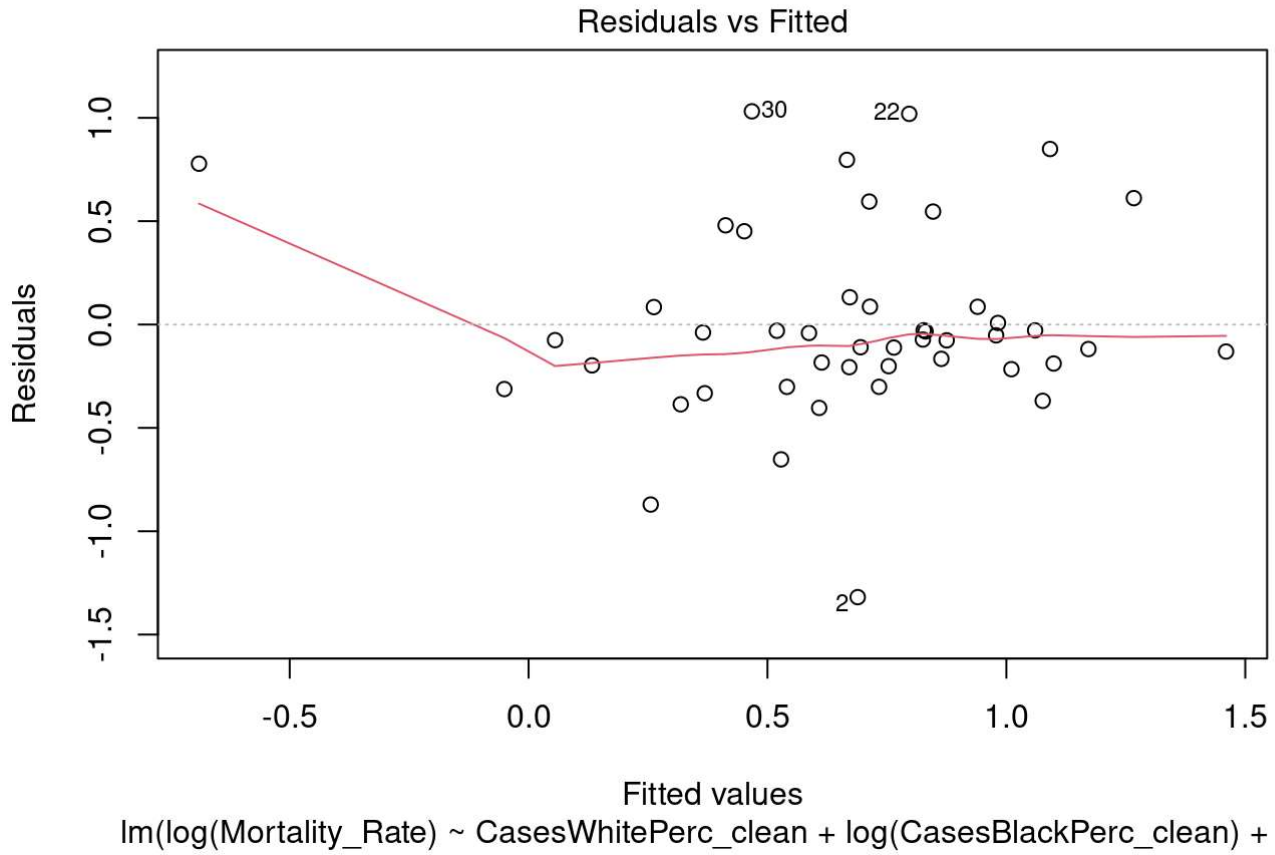
```
plot(model2, which=1)
```



Excluding the outlier in the beginning of the plot, we have a pretty flat red line along 0, signalling we have linear conditional expectation.

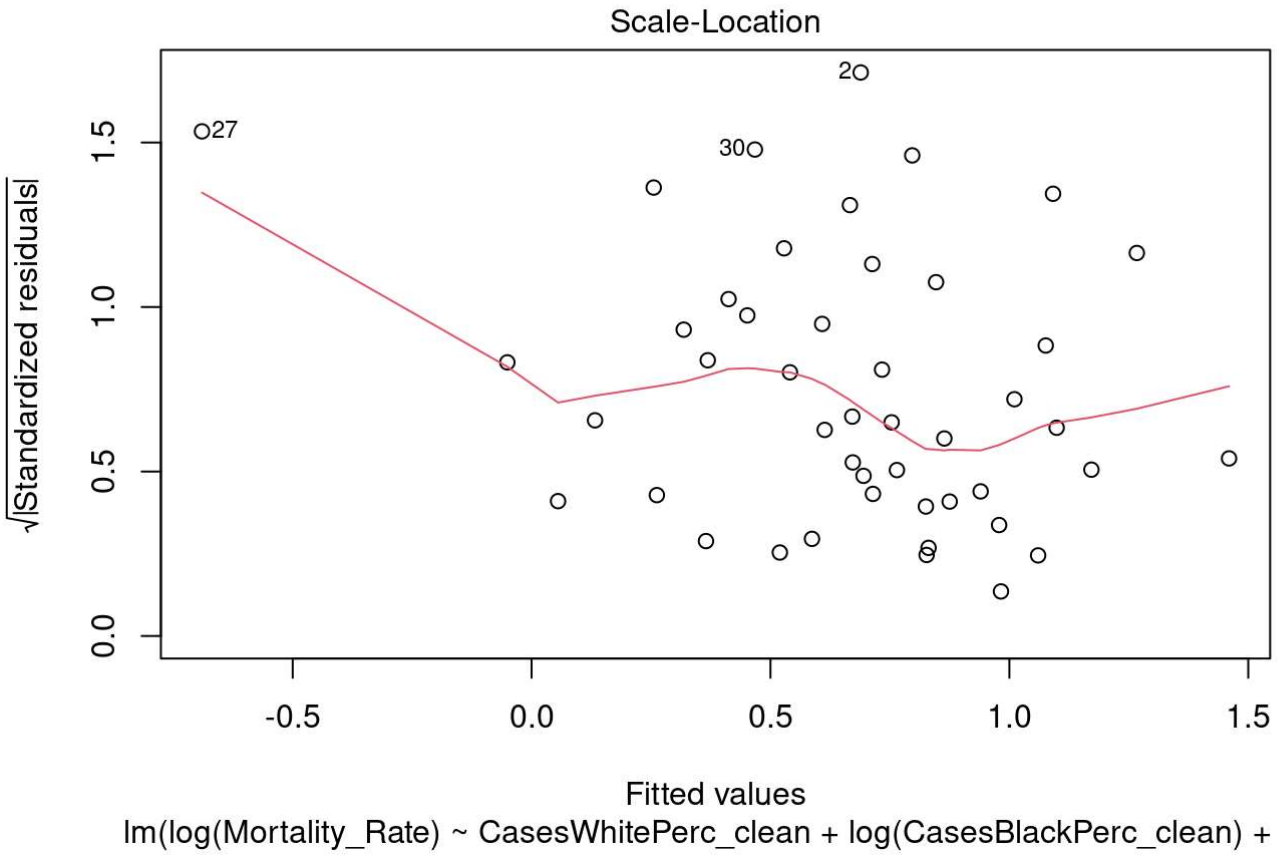
5. Homoskedasticity

```
plot(model2, which=1)
```



Excluding the outlier in the beginning of our plot, our Residuals vs. Fitted values showcase uniform thickness around the errors, signaling homoskedasticity for the most part.

```
plot(model2, which=3)
```



Our Scale Location plot has improved versus model 1. Excluding the outlier in the beginning of the plot, it looks like for the majority of the points in the middle, we are starting to see a horizontal band of points, signaling homoskedasticity.

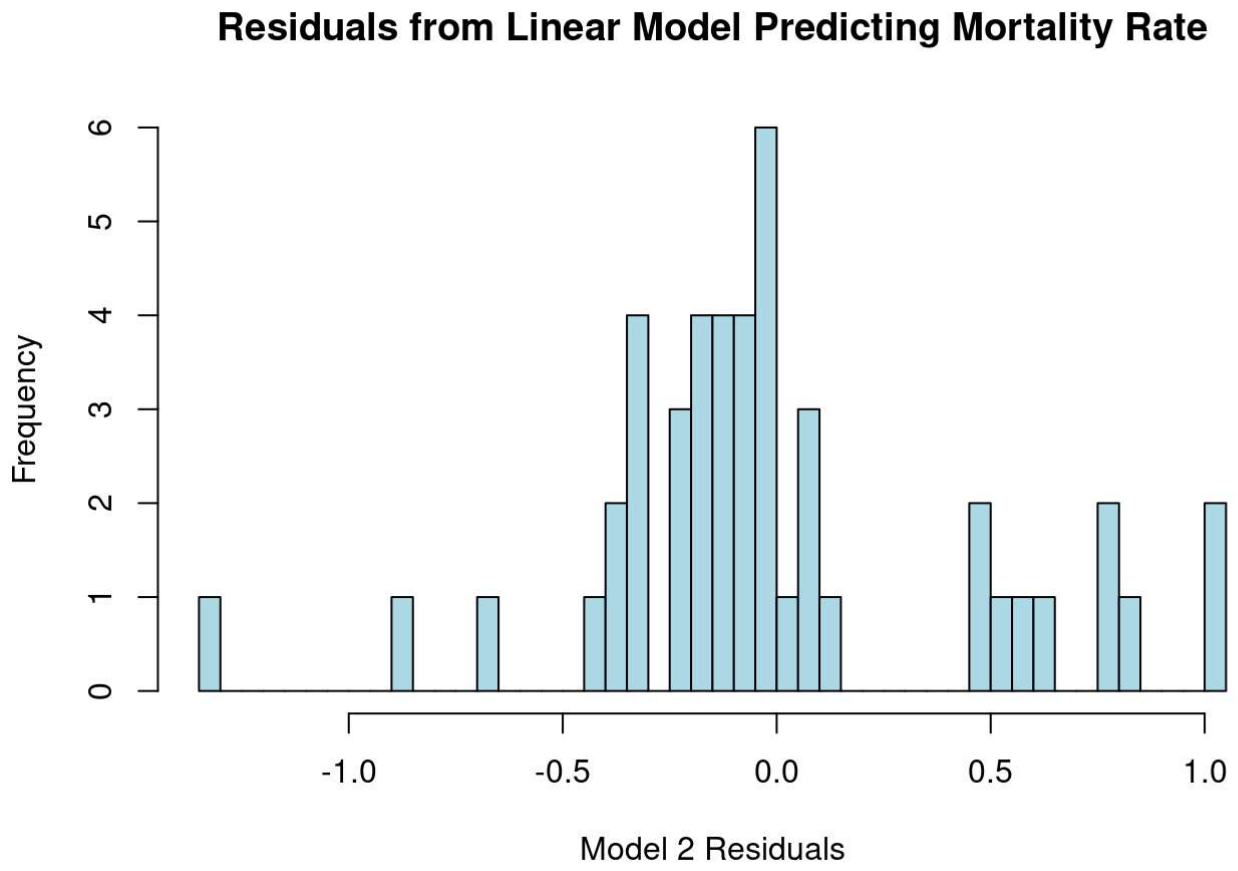
```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 4.9556, df = 4, p-value = 0.2919
```

Our Breusch-Pagan test generates a high p-value, which means we fail to reject the null. This is a good thing; this signals that our variables are homoskedastic. (Our null hypothesis, (H_0), states that our variables are homoskedastic, and we have failed to reject this null hypothesis.)

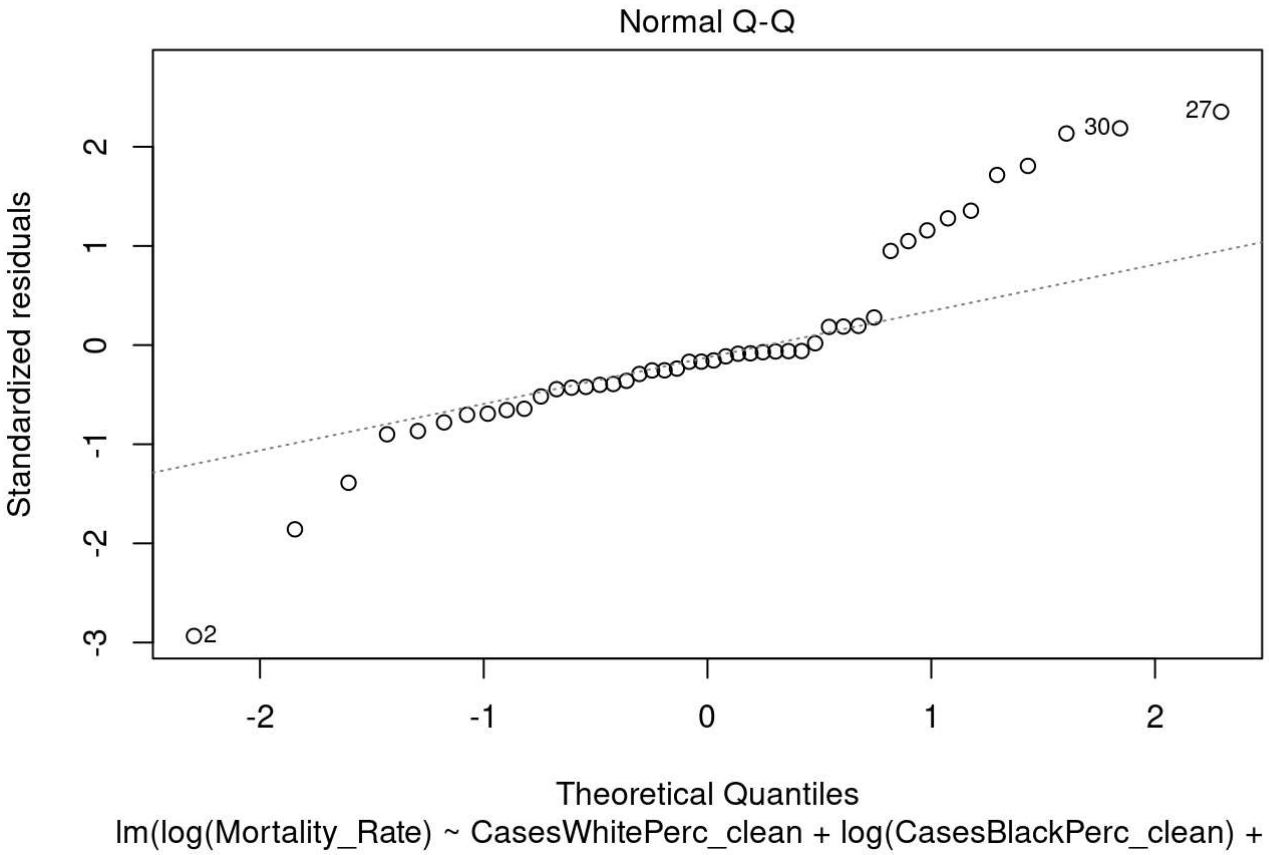
6. Normality of Error Terms

```
hist(model2$residuals, breaks = 50,
  main = "Residuals from Linear Model Predicting Mortality Rate",
  xlab = "Model 2 Residuals",
  col="lightblue")
```



Though we do not have that many data points, we can faintly see the normal distribution we expect to see in our residuals, which demonstrates that we have normality in our error terms.

```
plot(model2, which=2)
```



In the qqplot graph, we see that the residuals, for the most part, still fall along the linear, diagonal line. However, it is flatter when compared to the qqplot in model 1. There is deviation from the linear, diagonal line at the front and the back ends where we do not have a large sample size. Therefore, some concern for the normality of our error terms is warranted.

```
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.91803, p-value = 0.003217
```

We generate a low p-value in the Shapiro-Wilk Test, which means that our residuals may not be normally distributed. In statistical terms, we reject the null hypothesis (H_0 = our residuals are normally distributed).

This might create some pause with using model 2 as opposed to model 1. However, there are many positives to justify preferring model 2 over model 1:

- All of our other diagnostics have improved upon or at least matched model 1's diagnostics with the exception of the normality of error terms.
- All standard errors of our independent variables have improved (decreased) when compared to model 1's standard errors.
- Our adjusted R^2 has increased vs. model 1.

Model #3 + Respective CLM Assumptions

We now proceed with assessing the 6 CLM assumptions with our 3rd model:

```
model3 <- lm(log(Mortality_Rate) ~
  CasesWhitePerc_clean +
  log(CasesBlackPerc_clean) +
  CasesHispPerc_clean +
  ICU_beds_clean +
  inpatient_beds_clean
  , data = data,
  na.action=na.omit)
summary(model3)
```

```
##
## Call:
## lm(formula = log(Mortality_Rate) ~ CasesWhitePerc_clean + log(CasesBlackPerc_clean) +
##     CasesHispPerc_clean + ICU_beds_clean + inpatient_beds_clean,
##     data = data, na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3182 -0.2201 -0.0921  0.1773  0.9897
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.657965   0.521826   1.261   0.2147
## CasesWhitePerc_clean       -0.011533   0.005763  -2.001   0.0522 .
## log(CasesBlackPerc_clean)   0.192760   0.071840   2.683   0.0106 *
## CasesHispPerc_clean        -0.001095   0.007093  -0.154   0.8780
## ICU_beds_clean              0.377400   0.127979   2.949   0.0053 **
## inpatient_beds_clean       -0.028368   0.018237  -1.556   0.1277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4775 on 40 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.4335, Adjusted R-squared:  0.3627
## F-statistic: 6.122 on 5 and 40 DF,  p-value: 0.0002671
```

Interpretation/analysis of model 3 will be presented on the stargazer section further in our report below.

Discussion of 6 CLM Assumptions for Model #3:

1. Linear in Parameters: We do not need to assess this. This condition is always met if the dependent variable is a linear function (lm) of the explanatory variables.
2. Random iid sampling:

To assess if the data is IID, we need to know more about the sampling process. It is hard to justify this assumption because of the way the virus spreads. There are several reasons we might expect why our COVID data may not be independent of each other.

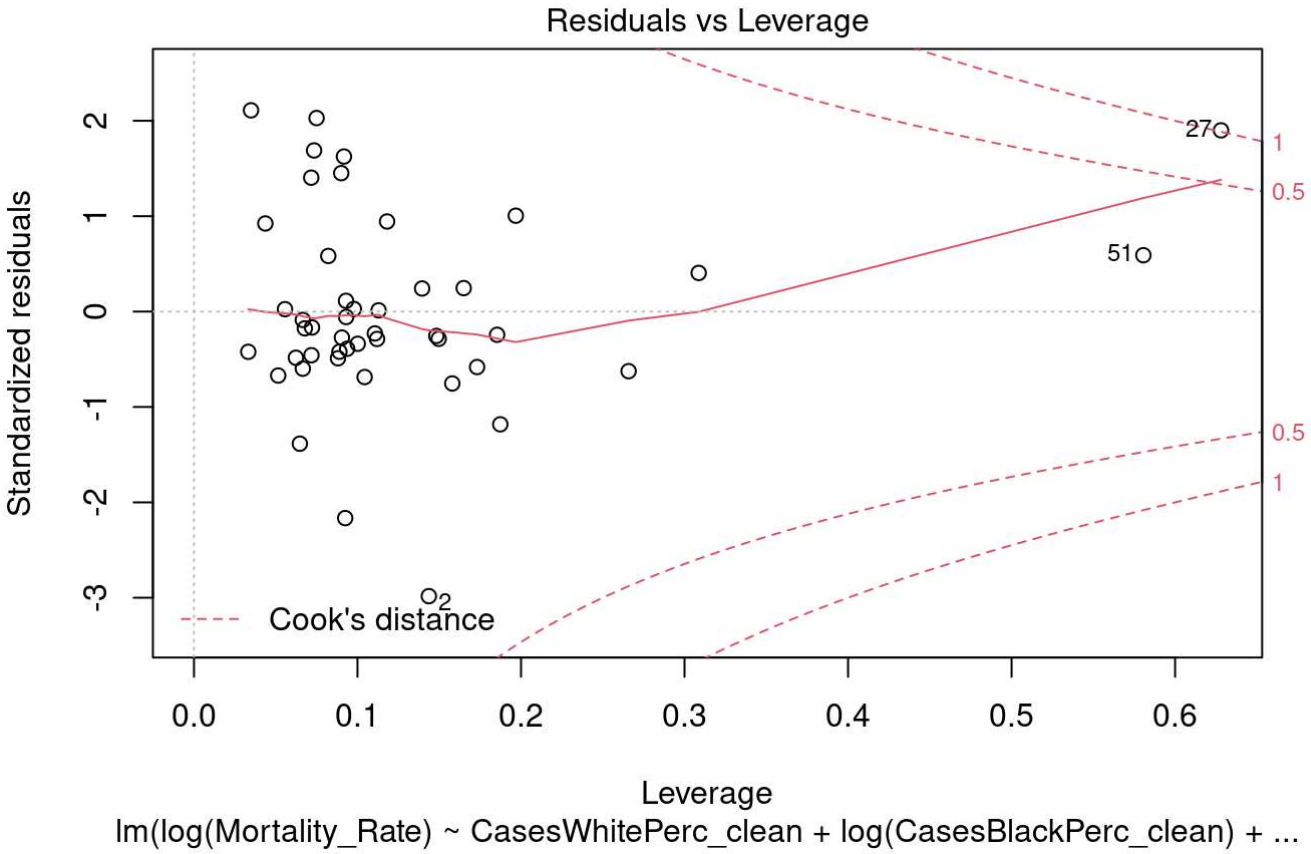
Since COVID is a virus, you can only contract it if you come in contact with someone else with COVID. Therefore, many people in the same vicinity/area will likely contract COVID (clustering).

Since we have data from the 51 states only, we do not have identical distribution, since the states' geographical location are set and cannot be changed. Neighbouring states are most likely to exhibit similar COVID case rate/behvior due to limited access to transportation and movement of people (people preferring cars over planes).

We tried to elliminate the effects of large states vs. small states by dividing by the number of COVID patients, as well as taking percentages rather than nominal numbers whenever possible. However, each state's geographical location cannot be changed.

Our research question, in part, is testing for independent and identically distribution in terms of affecting each human being equally, regardless of race. However, we know that the incubation period for the coronavirus is independent and identically distributed.

```
plot(model3, which=5)
```



We continue to have our outlier (with high leverage) vs. the regression line we predicted, propelling the Cook's distance line higher towards the latter end of the graph.

However, compared to our Cook's distance line from models 1 and 2, it looks like our Cooks distance line is even flatter in model 3, signalling that model 3 is an improvement upon our previous models.

3. No perfect multicollinearity

```
vif(model3)
```

```
##      CasesWhitePerc_clean log(CasesBlackPerc_clean)       CasesHispPerc_clean
##                  2.496617                  1.400118                  2.079169
##            ICU_beds_clean      inpatient_beds_clean
##                  2.651932                  3.243059
```

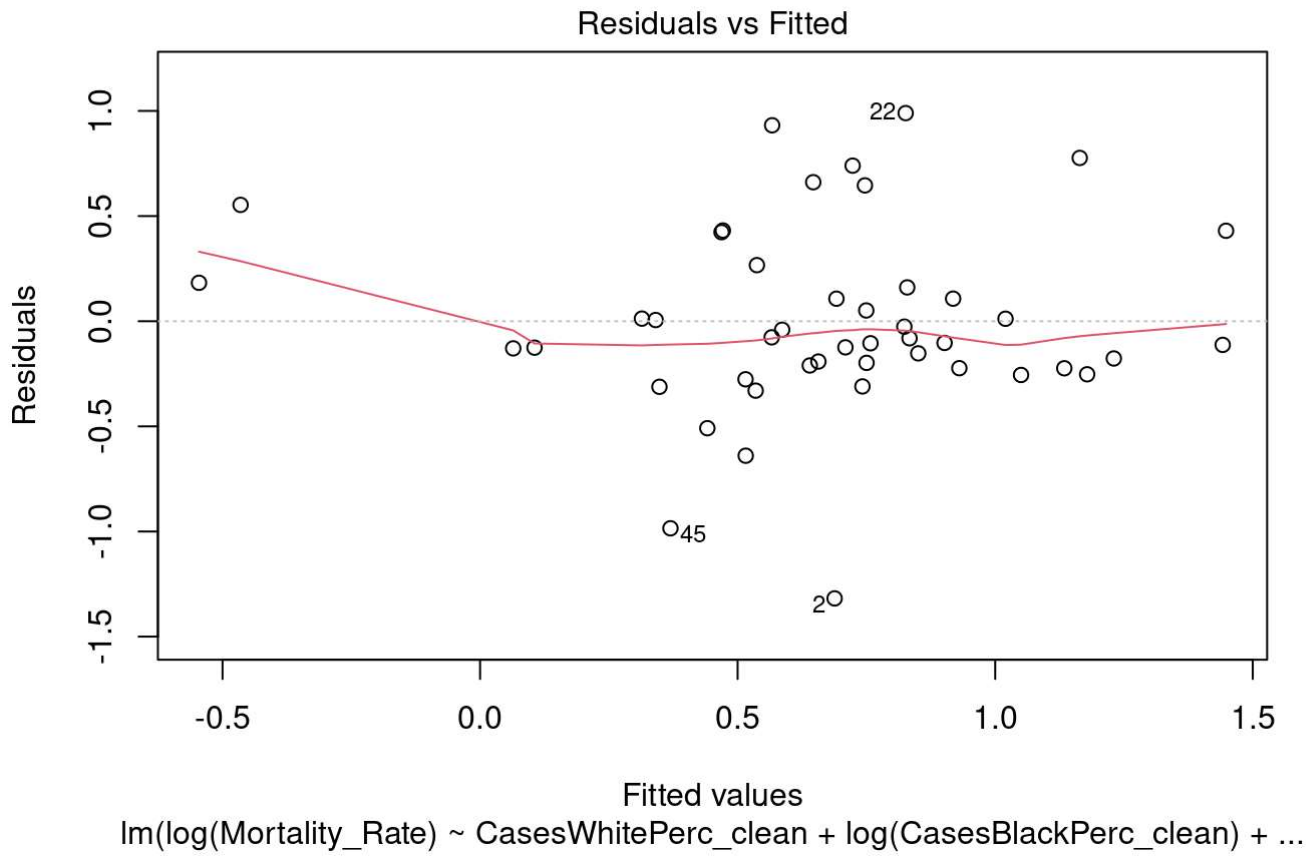
```
vif(model3) > 4
```

```
##      CasesWhitePerc_clean log(CasesBlackPerc_clean)       CasesHispPerc_clean
##                     FALSE                     FALSE                     FALSE
##            ICU_beds_clean      inpatient_beds_clean
##                     FALSE                     FALSE
```

We do not have high multicollinearity issues for any of our variables.

4. Zero conditional mean

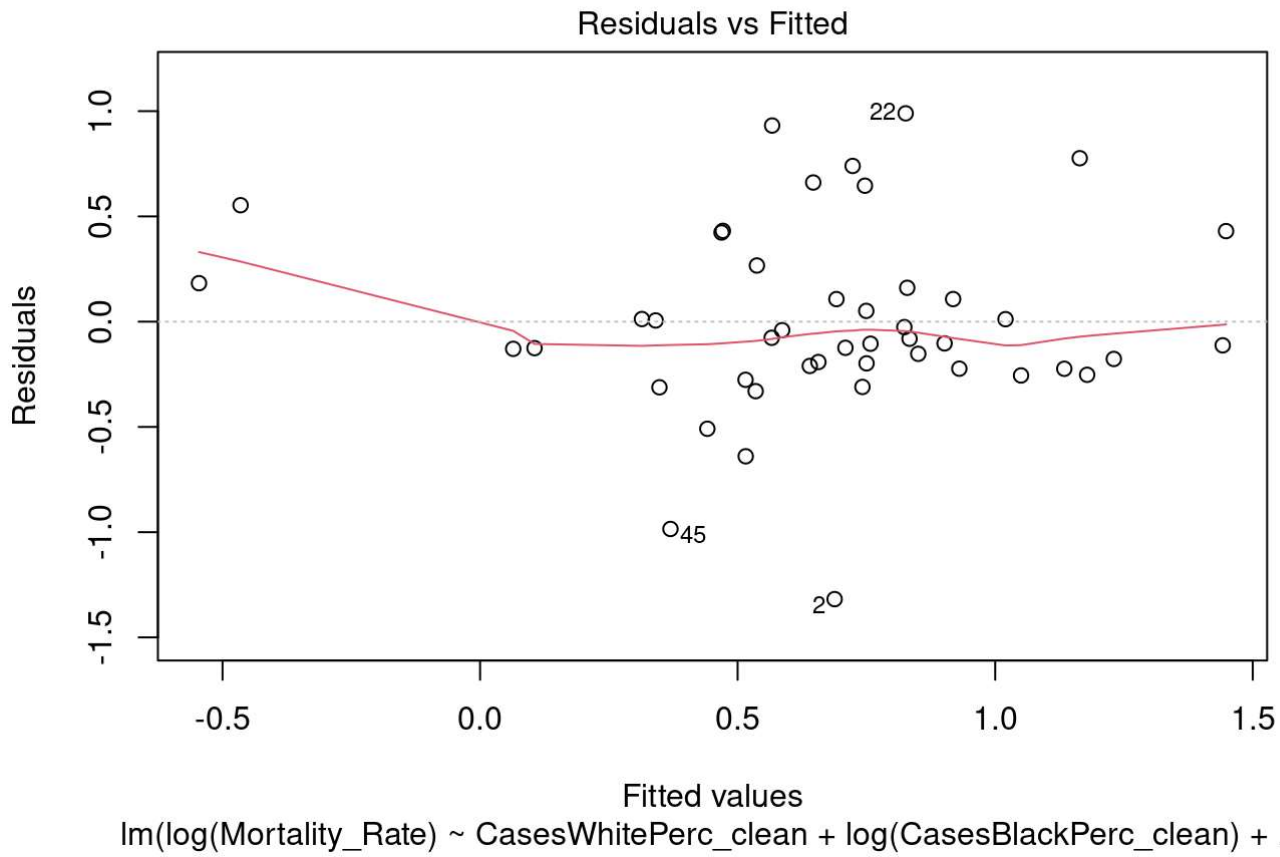
```
plot(model3, which=1)
```



We have a flat red line along 0, signalling we have achieved linear conditional expectation.

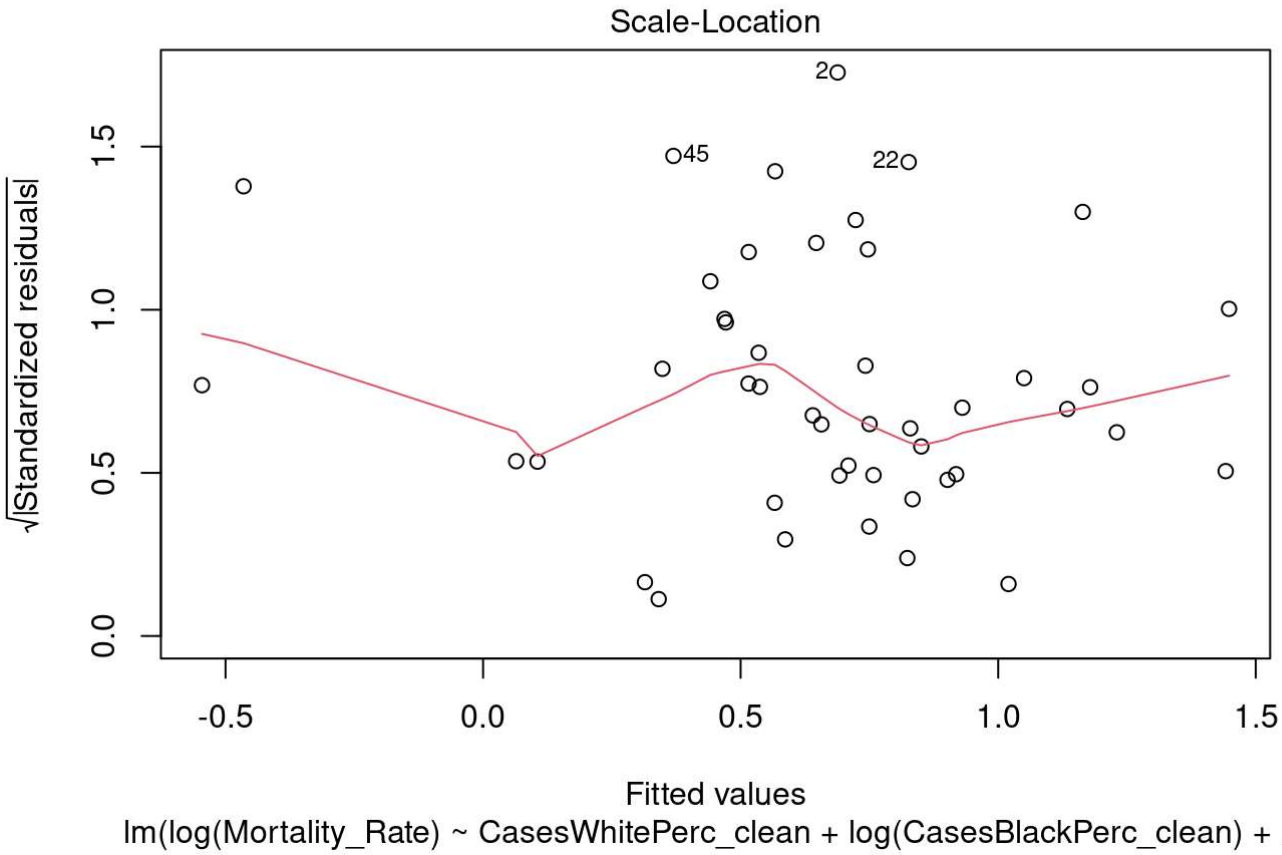
5. Homoskedasticity

```
plot(model3, which=1)
```



Our Residuals vs. Fitted values has uniform thickness around the errors, except for the extreme ends, signaling homoskedasticity.

```
plot(model3, which=3)
```



Our Scale Location plot looks like it has improved versus the plots we saw in model 1 and 2. We see a much flatter red line - a good thing. We see a horizontal band of points for the most part.

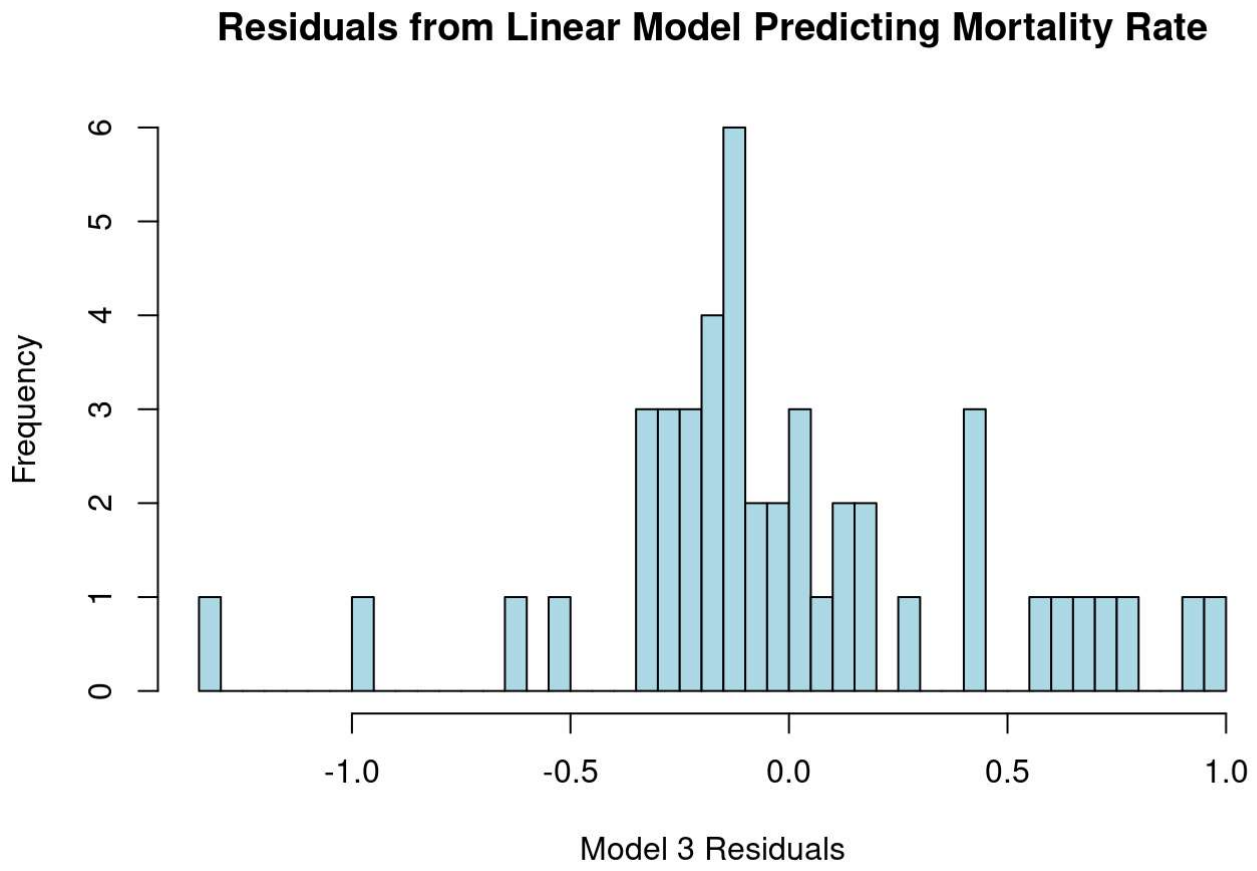
```
bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 4.863, df = 5, p-value = 0.4328
```

Our Breusch-Pagan test generates a high p-value, which means we fail to reject the null. This is a good thing; this signals that our variables are homoskedastic. (Our null hypothesis, (H_0), states that our variables are homoskedastic, and we have failed to reject this null hypothesis.)

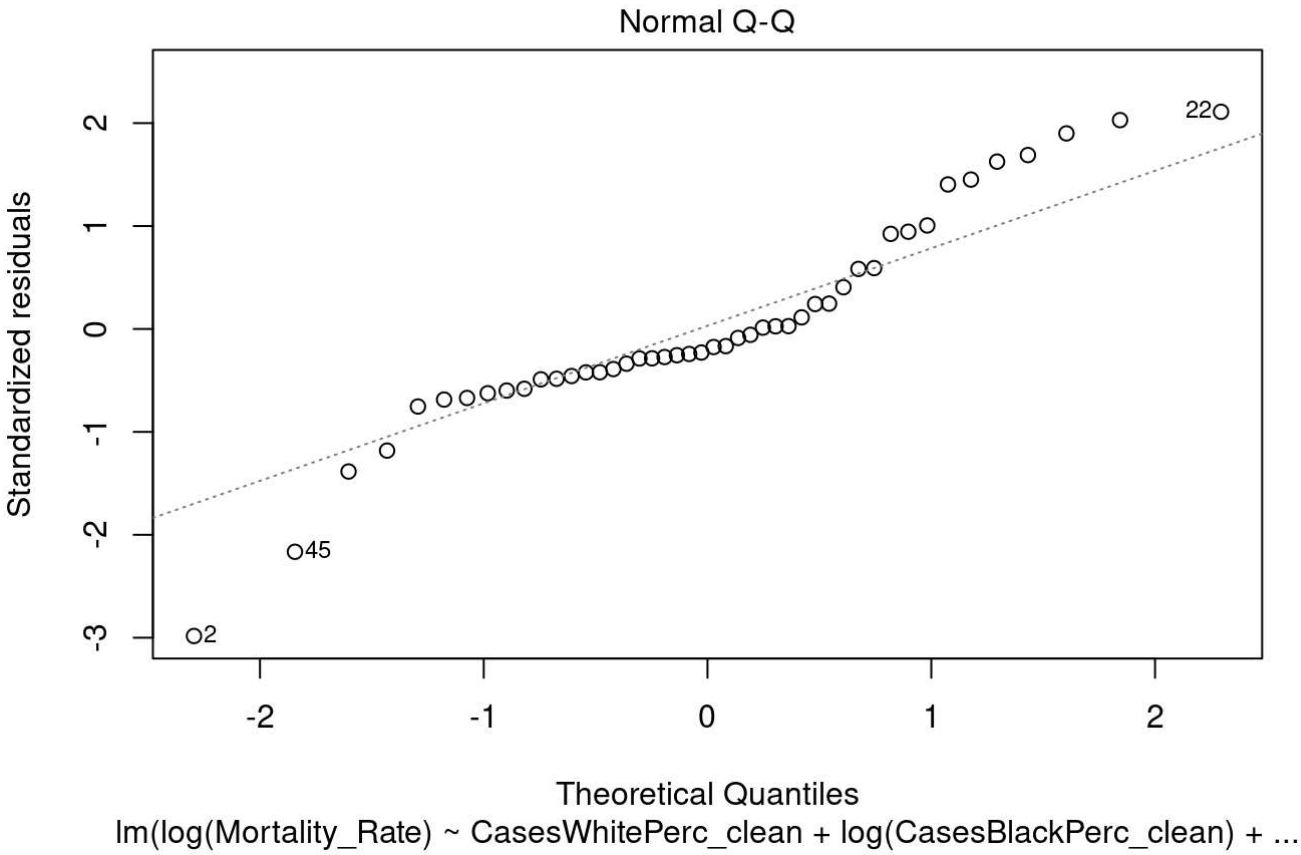
6. Normality of Error Terms

```
hist(model3$residuals, breaks = 50,
  main = "Residuals from Linear Model Predicting Mortality Rate",
  xlab = "Model 3 Residuals",
  col="lightblue")
```



Though we do not have that many data points, we can see the normal distribution we expect to see in our residuals, which demonstrates that we have normality in our error terms.

```
plot(model3, which=2)
```



When compared to model 2, the qqplot of model 3 is even more distorted. The majority of the plots in the middle, which were in a more linear, flatter formation in model 2, have now morphed into a somewhat quadratic function in model 3. There continues to be deviation from the linear, diagonal line at the front and the back ends where we do not have a large sample size. Therefore, more concern for the normality of our error terms is warranted in model 3 when compared to model 2.

```
shapiro.test(model3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.93824, p-value = 0.01684
```

We generate a low p-value in the Shapiro-Wilk Test, which means that our residuals may not be normally distributed. In statistical terms, we reject the null hypothesis (H_0 = our residuals are normally distributed).

These diagnostics create some pause with using model 3 as opposed to model 2, and we believe rightly so. Model 3 generates a mixed bag of diagnostics when compared to model 2 or model 1, and also generates higher standard errors for most of our independent variables versus model 2.

A Regression Table / Stargazer

```
stargazer(model1, model2, model3, type = "text",
  omit.stat = "f",
    star.cutoffs = c(0.05, 0.01, 0.001),
  title =
  "Table 1: Relationship between Mortality Rate vs. Case Rate by Race & Hospital Preparedness")
```

```
##
## Table 1: Relationship between Mortality Rate vs. Case Rate by Race & Hospital Preparedness
## ================================================================================
##                                          Dependent variable:
##                              ---------------------------------------------
##                                          log(Mortality_Rate)
##                                  (1)             (2)             (3)
## --------------------------------------------------------------------------------
## CasesWhitePerc_clean          -0.007          -0.013*         -0.012
##                               (0.006)         (0.006)         (0.006)
##
## log(CasesBlackPerc_clean)     0.224**         0.234**         0.193*
##                               (0.072)         (0.068)         (0.072)
##
## CasesHispPerc_clean            0.002          0.0004          -0.001
##                               (0.008)         (0.007)         (0.007)
##
## ICU_beds_clean                                0.235*          0.377**
##                                               (0.091)         (0.128)
##
## inpatient_beds_clean                                          -0.028
##                                                               (0.018)
##
## Constant                       0.498           0.518           0.658
##                               (0.557)         (0.523)         (0.522)
##
## --------------------------------------------------------------------------------
## Observations                    46              46              46
## R2                             0.301           0.399           0.433
## Adjusted R2                    0.251           0.341           0.363
## Residual Std. Error       0.518 (df = 42) 0.486 (df = 41) 0.477 (df = 40)
## ================================================================================
## Note:                                        *p<0.05; **p<0.01; ***p<0.001
```

Model 1

Model 1 is a basic model that simply aims to capture if there is a significant relationship between the percentage of COVID Cases by race and mortality rate. We quickly see that only the black case rate has a significant relationship with mortality rate. However, we also see that we have a low adjusted R^2, which suggests that we have other missing variables that we have not accounted for.

Approximately 25% of the observed variation can be explained by this model's inputs. For every 1% increase in the percentage of COVID Cases represented by the black population, the overall mortality rate increases by roughly 0.22%, after holding all other variables constant. Going back to our initial research question, it seems like race might have a significant relationship with mortality rate, at least preliminarily.

Model 2

Model 2 builds on Model 1 by adding the “ICU bed availability” variable. We chose the ICU beds variable becuase we hypothesized that there would be a significant relationship between availability of ICU beds vs. mortality rate. ICU beds are supposed to be more critical to mortality rate since only serious patients get admitted into the ICU. Another thing to note is that if the state is already seeing a high COVID mortality rate, the state’s hospitals may already be preparing more ICU beds.

It can be argued that Model 2 is better than Model 1, since it generates significantly lower standard errors vs. model 1, presents better diagnostics across the 6 CLM assumptions, and delivers a higher adjusted R^2.

The “% of COVID Cases represented by the black race” variable continues to show up as statistically significant. Our new variable, ICU bed availability, as well as the percentage of COVID Cases represented by the white race, also shows up as statistically significant.

Approximately 34% of the observed variation can be explained by this model’s inputs.

For every 1% increase in the percentage of COVID Cases represented by the black population, the overall mortality rate increases by roughly 0.23%, after holding all other variables constant. This is a more accurate figure than the 0.22% generated in model 1 since we have a lower standard error.

For every 1% increase in the percentage of COVID Cases represented by the white population, the overall mortality rate decreases by roughly 1%, after holding all other variables constant. This might be because of the widely covered wealth gap between white people and black people (https://cdn.americanprogress.org/content/uploads/2018/02/20130506/RacialWealthGap-fig3-693.png (https://cdn.americanprogress.org/content/uploads/2018/02/20130506/RacialWealthGap-fig3-693.png)). Better access to health care and treatment opportunities, both of which cost money, likely leads to a lower mortality rate, especially when compared to other disadvantaged people of color.

For every 1 unit increase in ICU beds per COVID patient, the overall mortality rate increases by roughly 23%, after holding all other variables constant. This is a much larger coefficient when compared to our race-related coefficients. This is counterinuitive to our initial hypothesis since we thought that more ICU bed availability could mean a lower mortality rate. As discussed above, this mortality rate increase might be driven by the state preparing for an influx of COVID patients in critical condition.

Model 3

Model 3 builds on Model 2 and adds the “inpatient bed availability” variable. Even though model 3 generates a higher adjusted R^2, we see that model 3 does **not** improve upon model 2, with higher standard errors across most of the independent variables and similiar/worse diagnostics across the 6 CLM assumptions. Model 3 represents a maximalist approach, erring on the side of including most variables.

The “% of COVID Cases represented by the black race” variable, as well as our “ICU bed availability” variable, continue to show up as statistically significant, albeit with higher standard errors. The white race case rate is no longer statistically significant.

Approximately 36% of the observed variation can be explained by this model’s inputs. For every 1% increase in the percentage of COVID Cases represented by the black population, the overall mortality rate increases by roughly 0.19%, after holding all other variables constant. For every 1 unit increase in ICU beds per COVID patient, the overall mortality rate increases by roughly 37%, after holding all other variables constant. However, these coefficients presented in model 3 are not as reliable as the coefficients presented in model 2 for the reasons listed above.

Practical Significance

Based on model 2, which aims to be the most parsimonious out of the 3 models, it looks like the percentage of COVID cases represented by white and black populations are indeed factors in influencing overall mortality rate.

However, going back to our original research question, it is unclear the percentage of COVID cases represented by the black population influences mortality rate **more so** than hospital preparedness (availability of ICU beds).

We could argue that the percentage of COVID cases represented by black populations ($p = <0.01$) is technically more significant than ICU bed availability ($p = <0.05$) in model 2. However, in terms of our beta coefficients, ICU bed availability does command a much larger change in our dependent variable when compared to the percentage of COVID cases represented by black populations.

Discussion of Omitted Variables

There are a number of considerations to make when looking at these data and coming to any conclusions about our findings. Below are a few considerations that could impact the results of our analysis which weren't directly included in the analysis itself. To do this exercise, we are applying our second model. In order to assess the omitted variable bias, we leveraged B_2, which has the largest magnitude.

1. People that identify as "Mixed" Race
    - Reasoning: The notion of racial identification by individuals can be subjective in the way that the data was likely collected. Assuming that race is self-reported, there may be those who more closely identify as being of "mixed race" and may have been forced to choose one race over another. Similarly, there may have been individuals who more closely identify as one particular race when in actuality their genetic make up would be considered of "mixed" heritage. This blurs the data set; since there is no option to choose more than one race, we are assuming that respondents chose their "primary" demographic as opposed to checkmarking all of their identifying features. For example, our data may have bias if someone is both white and black, but responded "black" becuase they mostly identified as black instead.
    - Direction of Bias (towards or away from 0): Away from 0
    - Large or Small Effect: Small
    - Any variables that may proxy for the omitted variable: The "Other Race" variable
2. Ages of all COVID-positive patients, particularly ages 50+
    - Reasoning: The data that we've decided to analyze does not include any age data, which is important because we know that it's been scientifically proven that those who are 50+ are more likely to have more serious complications with Covid-19 and higher mortality rates. For states with a higher average age in the population, we could expect a higher mortality rate.
    - Direction of Bias (towards or away from 0): Away from 0
    - Large or Small Effect: **Large**
    - Any variables that may proxy for the omitted variable: Age distribution in state, regardless of COVID infection
3. More Access to COVID treatments
    - Reasoning: In states that are better prepared, tests may be more readily accessible to the populace. Additionally, the acess to these tests is likely skewed towards those who have the means and proper health coverage to afford the tests. This would make it more likely that people with the means to get tested are far more likely to discover that they have the disease. There's also a higher likelihood that

the treatment someone receives if they contract the disease is related to their ability to afford treatment. Therefore, low access/affordability to COVID treatments may have a relationship with a higher mortality rate.

- Direction of Bias (towards or away from 0): Towards 0
- Large or Small Effect: Large
- Any variables that may proxy for the omitted variable: Tests performed, Tests per 100K

4. Urban populations
    - Reasoning: The spread of Covid-19 is driven from proximity to someone who has the disease. For those who live in less proximity to others, their likelihood of contracting Covid-19 is lower. This would suggest that those who live in more urban settings have a higher likelihood than those who live in rural settings to contract and subsequently die from Covid-19.
    - Direction of Bias (towards or away from 0): Away from 0
    - Large or Small Effect: Large
    - Any variables that may proxy for the omitted variable: Low Population density per square miles, Low median household income
5. Having pre-existing health conditions
    - Reasoning: Pre-existing health conditions have been proven to increase the likelihood of dying from Covid-19. This data is not readily available in our data set and would also be very difficult to capture given the spectrum of pre-existing conditions and how they might make the individual more susceptible to Covid-19.
    - Direction of Bias (towards or away from 0): Away from 0
    - Large or Small Effect: **Large**
    - Any variables that may proxy for the omitted variable: % of total cases for individuals with a pre-existing condition “Percent at risk for serious illness due to COVID” This is a column in the data set that we do not use

Though Covid-19 has certainly impacted everyone, there’s been evidence highlighting that the largest contributor to death is an individuals age as well as whether or not the individual has a pre-existing condition. Without being able to appropriately control for these factors, it’s difficult to tell how much bias has been introduced into our models. These are the two factors that we believe would be most beneficial in accounting for if we were to perform a more in-depth analysis, given we were able to collect the data.

Conclusion

In lab2, we aimed at investigating the statistical significance or either percentage of covid cases in each race or hospital preparedness on the overall covid mortality rate. For the data relating percentage of covid cases in each race, we used the “covid19.xls” file provided with lab2 documents and for data relating hospital preparedness, we download “covid19-NatEst.csv” from the CDC website and use the data for the number of in-patient beds available, number of ICU beds available and number of confirmed covid patients in the hospital for the latest month of July provided.

We then operationalize the variables and look out for any unexpected values (“NR”, “<0.01” and etc) and decide on whether to ignore or replace them. All the variables were then analyzed against the covid-19 mortality rate on the graphs to check if we had a linear relationship or whether or not transformations were needed. Detailed analysis related to these transformations was done for each variable as part of the model building process.

We chose 3 different models for the OLS regression analysis. Model1 has only percentage of covid cases in each race as the independent variables. For model2, an extra variable is added on top of model1, which relates to the number of ICU beds available per covid patient in the hospitals. The model3 adds an extra variable on top of model2 which relates to the number of In-Patient beds available per covid patient in the hospital. All 6 CLM assumptions were analyzed in detail for model1, model2 and model3 and the comparison of the these assumptions are talked about briefly with the plots. These 3 models are then fed in to the “stargazer” package and the output is then analyzed and compared for all the 3 models for p-value/statistical significance of coefficients, standard errors and etc. Here are few of the important takeways we find from our analysis:

i. log(CasesBlackPerc_clean) remained statistically significant in terms of the p-value in all three models though its statistical significance decreases significantly in model3. A potential hypothesis that we have is that there are more blacks that have pre-existing conditions, which is a variable that was not investigated as part of this model.
ii. ICU_beds_clean is statistically significant in model2 and model3. Though its statistical significance increases in model3 compared to model2, the standard error of the coefficients also rises in model3. Since our covid mortality rates are captured 4 months after our ICU bed availability a hypothesis that we have is that states that are anticipating a future rise in Covid patients, proactively increase the number of available ICU beds.
iii. A significant increase in adjusted R^2 from model1 (0.251) to model2 (0.341) but lesser increase from model2 to model3 (0.363).
iv. In terms of the CLM assumptions, normality of the residuals seems to be of concern and we see some deviations in the qq-plot. Another major deviation could be relation I.I.D sampling but that is hard to analyze and affects the 3 models equally.

From i) and ii), we can say that for any US state, the percentage of covid cases in the black race and the number of ICU beds available per covid patient in the hospitals seem to be statistically significant in terms of p-value for covid mortality rate in that state. Of course our models are based off a number of assumptions which we have discussed earlier and all of those need to be met for this statement to hold true.