Lab-3-Report-Final

Eric Lundy, Gabriel Ohaike, Javed Roshan

8/2/2020

Contents

Chapter 1: Introduction

1.1: Overview

According to The Center for Disease Control and Prevention (CDC) "COVID-19 is mostly spread by respiratory droplets released when people talk, cough, or sneeze." Due to the rapid spread and high mortality rate of COVID-19, the CDC recommends the following guidelines:

- Wash your hands often
- Avoid close contact
- Cover your mouth and nose with a cloth face cover when around others
- Cover coughs and sneezes
- Clean and disinfect
- Monitor Your Health Daily

In the United States, one of the guidelines on the list above "Cover your mouth and nose with a cloth face cover when around others" has been a controversial guideline. Some citizens argue that requiring a mask in public places is an attack on their personal freedom while others believe that simply wearing a mask could save lives. As more businesses are gradually reopening, most state government have mandated that face mask be worn in public places. Yet, there are many who have failed to comply with this directive. This leads us to our research question.

1.2: Research Question

Does wearing a mask help prevent the spread of COVID-19?

1.3: Operationalization

We are operationalizing this definition by assuming that the states who have mandated face mask adoption will impact the spread of the COVID-19. We also assume that a higher number of cases per square mile is a good indicator of the spread of the disease. Therefore, we normalize the total cases by the population density. Following the mandate for employees in public-facing businesses to wear face masks, we would like to consider the wider adoption of this mandate to all residents of a state. Thereby, measuring how masks are actually affecting the spread of the disease.

Chapter 2: Exploratory Data Analysis

2.1: Data Sources

The data was compiled by Majid Maki-Nayeri, a professor at UC Berkeley. He extracted 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).

The dataset includes variables representing:

- Spread of the disease
- State-level policy responses
- General state-level characteristics

In order for us to answer the research question, we have added two new variables to the dataset: `Fixed_Mandate_Face_Mask` and `First_Case_Date`. We sourced this data from the CDC website.

2.2: Data Exploration

Let's examine all data from the dataset provided

```
library(car)
library(lmtest)
library(sandwich)
library(stringr)
library(stargazer)
library(dplyr)

library(ggplot2)
library(gridExtra)
library(grid)
library(lattice)
library(ggimage)
```

```
dataSet = read.csv("/Users/javed/Documents/UCB/covid-19_dataset.csv")
ds <- dataSet
sapply(ds, function(x) paste0(head(x),  collapse = ", "))
```

```
##                                                          State
##      "Alabama, Alaska, Arizona, arizona, Arkansas, California"
##                                                    Total.Cases
##                      "44909, 1138, 83376, 14713, 23814, 260155"
##                                                    Total.Death
##                                 "1009, 16, 1538, 271, 287, 6331"
##                                                     Death_100k
##                                 "20.6, 2.2, 25.2, 25.2, 9.5, 16"
##                                               CasesInLast7Days
##                            "9804, 284, 28038, 28038, 4504, 53722"
##                                                  RatePer100000
##                     "918.8, 154.3, 1367.7, 1367.7, 790.2, 657.7"
##                                                totalTestResults
##                "449886, 122732, 604362, 604362, 338893, 4680138"
##                                             State.of.emergency
##           "3/13/20, 3/11/20, 3/11/20, 3/11/20, 3/11/20, 3/4/20"
##                                  Stay.at.home..shelter.in.place
##                  "4/4/20, 3/28/20, 3/31/20, 3/31/20, 0, 3/19/20"
##                           End.relax.stay.at.home.shelter.in.place
##                      "4/30/20, 4/24/20, 5/16/20, 5/16/20, 0, 0"
```

```
##                                Closed.non.essential.businesses
##               "3/28/20, 3/28/20, 3/30/20, 3/30/20, 0, 3/19/20"
##                             Began.to.reopen.businesses.statewide
##             "4/30/20, 4/24/20, 5/8/20, 5/8/20, 5/4/20, 5/8/20"
## Mandate.face.mask.use.by.employees.in.public.facing.businesses
##            "5/11/20, 4/24/20, 5/8/20, 5/8/20, 5/11/20, 5/5/20"
##                                          Fixed_Mandate_Face_Mask
##            "5/11/20, 4/24/20, 5/8/20, 5/8/20, 5/11/20, 5/5/20"
##         Weekly.unemployment.insurance.maximum.amount..dollars.
##                                  "275, 370, 240, 240, 451, 450"
##                               Population.density.per.square.miles
##                      "93.24, 1.11, 62.91, 62.91, 56.67, 241.65"
##                                                 Population.2018
##         "4887871, 737438, 7171646, 7171646, 3013825, 39557045"
##           Percent.living.under.the.federal.poverty.line..2018.
##                                "16.8, 10.9, 14, 14, 17.2, 12.8"
##                Percent.at.risk.for.serious.illness.due.to.COVID
##                             "43.1, 32.8, 39.1, 39.1, 43.5, 33.3"
##                                               All.cause.deaths.2018
##                     "54352, 4453, 59282, 59282, 32336, 268818"
##                                                   Children.0.18
##                             "0.24, 0.27, 0.24, 0.24, 0.25, 0.24"
##                                                    Adults.19.25
##                             "0.09, 0.09, 0.09, 0.09, 0.09, 0.09"
##                                                    Adults.26.34
##                             "0.12, 0.13, 0.12, 0.12, 0.12, 0.14"
##                                                    Adults.35.54
##                             "0.25, 0.26, 0.24, 0.24, 0.25, 0.26"
##                                                    Adults.55.64
##                             "0.14, 0.13, 0.12, 0.12, 0.13, 0.12"
##                                                             X65.
##                             "0.17, 0.12, 0.18, 0.18, 0.17, 0.14"
##                                                  First_Case_Date
##         "3/13/20, 3/12/20, 1/26/20, 1/26/20, 3/11/20, 1/25/20"
```

There are 27 variables and 52 observations. As seen in the dataset, all `date` variables are treated as character. Therefore, we will have to convert them into a date field so we can perform date arithmetic.

```
summary(ds$Total.Cases)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     939   11882   31305   55332   67052  398828
```

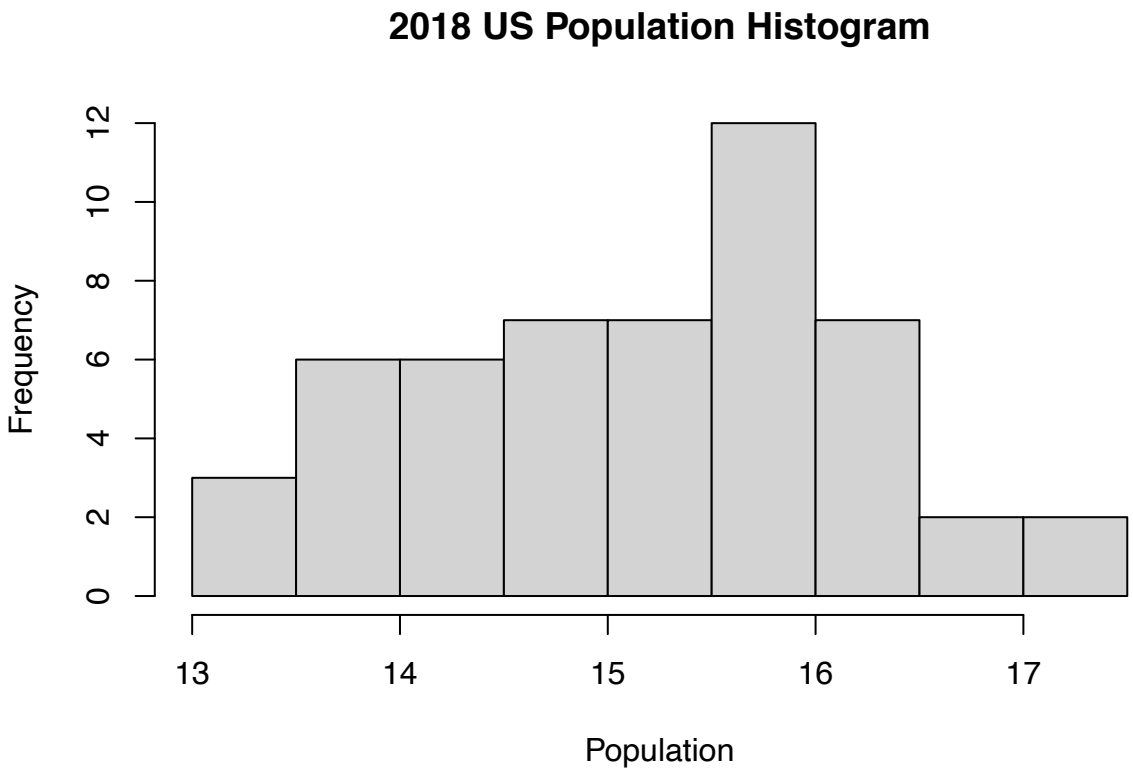
```
table(ds$Mandate.face.mask.use.by.employees.in.public.facing.businesses)
```

```
##
##       0 4/10/20 4/15/20 4/16/20 4/17/20 4/18/20 4/19/20 4/23/20 4/24/20 4/26/20
##      10       1       1       1       2       2       1       1       1       1
## 4/27/20 4/28/20 4/29/20  4/3/20  4/8/20  5/1/20 5/11/20 5/29/20  5/4/20  5/5/20
##       1       1       1       1       1       7       4       1       3       1
##  5/6/20  5/7/20  5/8/20  5/9/20  6/1/20 6/26/20
##       2       1       3       2       1       1
```

Total Cases is the dependent variable of our analysis. We will apply logarithm function in our model as its mean value is >55k. Logarithmic transformation is a convenient means of transforming a highly skewed variable into a more normalized dataset.

A look at `Mandate.face.mask.use.by.employees.in.public.facing.businesses` shows that it has 10 states where face masks were not mandated.

```
pop <- ds$Population.2018
options(scipen=999)
hist(log(pop), main = "2018 US Population Histogram", xlab = "Population")
```




```
summary(ds$Percent.living.under.the.federal.poverty.line..2018.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.60   10.97   12.85   12.93   14.15   19.70
```

```
summary(ds$Percent.at.risk.for.serious.illness.due.to.COVID)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   35.98   38.65   38.16   40.58   49.30
```

We see relatively on average, the percent of the population at risk for serious illness is 38.15 compared to 12.91% of population living under federal poverty line.

2.3: Data Cleansing & Transformation

There are two rows with data for state "Arizona". Upon assessment the data for row #4 looks incorrect, so we decided to remove that row from the dataset.

```
ds <- ds[-c(4),]
```

let's examine the two new variables introduced.

- `Fixed_Mandate_Face_Mask` provides a date when wearing face mask was mandated by a state
- `First_Case_Date` provides COVID-19 first case date in a state

Computing `Fixed_Mandate_Face_Mask - First_Case_Date` gives us number of days since the first case identified and until face mask was mandated by a state.

We know that the policies and state characteristics related variables' data was compiled as of 07/02/2020 and Covid-19 related variables' data was compiled as of 07/06/2020. We have assumed that for any `Fixed_Mandate_Face_Mask` date with a value 0, that the mandate was implemented as of 07/03/2020 i.e., a day after the last data update to the current dataset

```
maskStart <- as.Date(ds$Fixed_Mandate_Face_Mask, format="%m/%d/%y")
firstCase <- as.Date(ds$First_Case_Date, format="%m/%d/%y")
# the load of date data with 0 gets loaded as NA; find all NA - there are 10 such values
ndx <- which(is.na(maskStart))
ndx
```

```
##  [1] 13 16 17 26 27 37 41 42 43 50
```

```
# set all of them to 07/07/20 date
maskStart[ndx] <- as.Date("07/03/20", format="%m/%d/%y")
mm <- as.data.frame(maskStart-firstCase)
mm <- mm$`maskStart - firstCase`
mm <- strtoi(mm)
# days until mask was mandated
summary(mm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.00   48.00   60.00   73.08  102.00  149.00
```

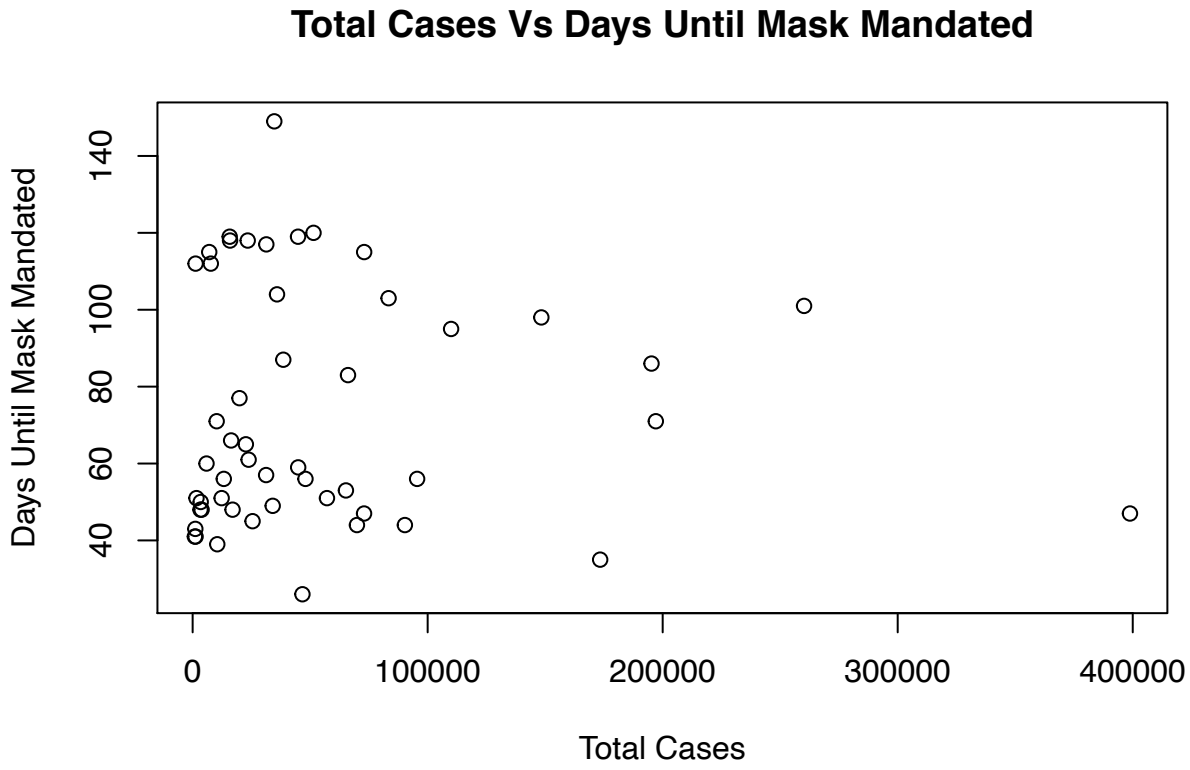
The mask mandated date was enforced in a state after an average of 73 days after the detection of the first instance of COVID-19.

We are also interested in the variable, `Total.Cases`, which signifies total cases of COVID-19 in a state.

```
# total cases
tc <- ds$Total.Cases
title <- "Total Cases Vs Days Until Mask Mandated"
xtitle <- "Total Cases"
ytitle <- "Days Until Mask Mandated"
summary(tc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     939   11415   31353   56128   68003  398828
```

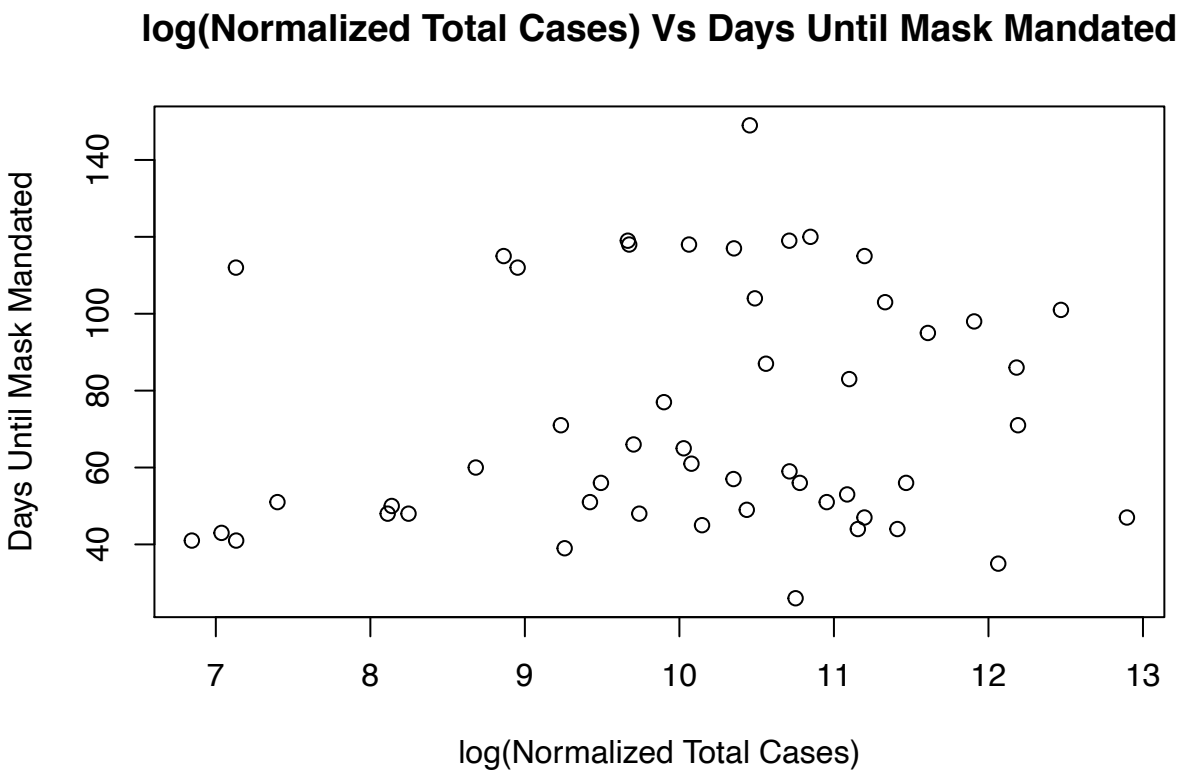
```
plot(tc, mm, main=title, xlab=xtitle, ylab=ytitle)
```




```
# total cases density
tcd <- tc/ds$Population.density.per.square.miles
title <- "log(Normalized Total Cases) Vs Days Until Mask Mandated"
xtitle <- "log(Normalized Total Cases)"
ytitle <- "Days Until Mask Mandated"
summary(tcd)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.9117  171.3911  354.9767  445.7422  619.9640 1827.0541
```

```
plot(log(tc), mm, main=title, xlab=xtitle, ylab=ytitle)
```



We have decided to divide the total number of cases by the `Population.density.per.square.miles`. We believe this should help account for people who live in highly populated states being difficult to observe social distancing. We also are applying the natural log to `Total.Cases` / `Population.density.per.square.miles`. This will help scale the variable since the number of cases is high compared to the days before the mandate.

Let's examine following pairs of attributes:

- Dates when `Closed.non.essential.businesses` and `Began.to.reopen.businesses.statewide`
- Dates when `Stay.at.home..shelter.in.place` and `End.relax.stay.at.home.shelter.in.place`

```
# get the reopen and close dates
reopen <- ds$Began.to.reopen.businesses.statewide
close <- ds$Closed.non.essential.businesses
# find the dates that have both dates = 0
# Knowing very clearly we did not see any dates that have same date for both variables
ndx1 <- which(reopen == close)
# let's see how many dates have both dates = 0
ndx1
```

```
## [1] 42
```

```
ds$State[ndx1]
```

```
## [1] "South Dakota"
```

```
reopen[ndx1]
```

```
## [1] "0"
```

```
close[ndx1]
```

```
## [1] "0"
```

There is only one row at index 42 (for state South Dakota) that has a zero date in both the

`Closed.non.essential.businesses` and `Began.to.reopen.businesses.statewide` variables. It implies that this state did not close non-essential business and therefore, it did not had to reopen business.

Performing a date difference in R with incorrect date will result in a NA value. Let's preserve this index to fix it back to zero later.

```
closeZero <- ds[ds$Closed.non.essential.businesses == 0,]
closeZero$State
```

```
##  [1] "Arkansas"     "Florida"      "Georgia"      "Missouri"     "Nebraska"
##  [6] "North Dakota" "Oregon"       "South Dakota" "Utah"         "Virginia"
## [11] "Wyoming"
```

```
closeZero$Closed.non.essential.businesses
```

```
##  [1] "0" "0" "0" "0" "0" "0" "0" "0" "0" "0" "0"
```

```
closeZero$Began.to.reopen.businesses.statewide
```

```
##  [1] "5/4/20"  "5/18/20" "5/1/20"  "5/4/20"  "5/18/20" "5/1/20"  "5/15/20"
##  [8] "0"       "5/1/20"  "5/29/20" "5/1/20"
```

```
paste("Total number of states with 0 dates: ", length(closeZero$State))
```

```
## [1] "Total number of states with 0 dates:  11"
```

```
which(ds$Closed.non.essential.businesses == 0)
```

```
##  [1]  4 10 11 26 28 35 38 42 45 47 51
```

We have already examined the situation with South Dakota above. However, there are 10 additional states that have bad data as they have valid reopen date without a corresponding close date. For these states, we will consider using a `mode` of the number of days.

```
# convert string into a date format for reopen date
finish <- as.Date(reopen, format="%m/%d/%y")
# as well when busineeses were closed
start <- as.Date(close, format="%m/%d/%y")
# get a date diff to get number of days
drs <- as.data.frame(finish-start)
ind1 <- which(is.na(drs))
print("Indices of this error: ")
```

```
## [1] "Indices of this error: "
```

```
paste(ind1)
```

```
##  [1] "4"  "10" "11" "26" "28" "35" "38" "42" "45" "47" "51"
```

It is confirmed that the date difference for the bad data resulted in NA values.

```
# function that returns mode given a vector input
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

drs_noNA <- drs[!is.na(drs)]
drs_noNA = str_remove_all(drs_noNA, "[days ]")
drs_noNA = strtoi(drs_noNA)
drsMode <- getmode(drs_noNA)
```

```
drs <- drs$`finish - start`
drs[is.na(drs)] <- drsMode

drs[ndx1] = 0
paste("Mode of # of days since reopen for states with valid data: ", getmode(drs_noNA))
```

```
## [1] "Mode of # of days since reopen for states with valid data:  39"
```

We used 39 as a replacement for all 10 states which had bad data.

Let's look at second set of dates: `Stay.at.home..shelter.in.place` and `End.relax.stay.at.home.shelter.in.place`

```
shelterEnd <- ds$End.relax.stay.at.home.shelter.in.place
shelterStart <- ds$Stay.at.home..shelter.in.place
ndx2 <- which(shelterEnd == shelterStart)
ds$State[ndx2]
```

```
##  [1] "Arkansas"     "Connecticut"  "Iowa"         "Kentucky"     "Nebraska"
##  [6] "New Jersey"   "North Dakota" "Oklahoma"     "South Dakota" "Texas"
## [11] "Utah"         "Wyoming"
```

```
ds$End.relax.stay.at.home.shelter.in.place[ndx2]
```

```
##  [1] "0" "0" "0" "0" "0" "0" "0" "0" "0" "0" "0" "0"
```

```
ds$Stay.at.home..shelter.in.place[ndx2]
```

```
##  [1] "0" "0" "0" "0" "0" "0" "0" "0" "0" "0" "0" "0"
```

```
paste("Number of states that did not had shelter in place: ", length(ndx2))
```

```
## [1] "Number of states that did not had shelter in place:  12"
```

This initial assessment shows that above listed 12 states did not implement Shelter in Place therefore, they did not also relax that government mandate. We will ensure that we will appropriately replace data with 0s for these 12 states.

```
finishS <- as.Date(shelterEnd, format="%m/%d/%y")
startS <- as.Date(shelterStart, format="%m/%d/%y")
dsp <- as.data.frame(finishS - startS)
dsp <- dsp$`finishS - startS`
# fix the 12 states with 0 days for the shelter in place
dsp[ndx2] = 0

t1 <- shelterEnd[which(is.na(dsp))]
t2 <- shelterStart[which(is.na(dsp))]
t3 <- ds$State[which(is.na(dsp))]

cat("States that did not end Shelter in Place: ", t3, "\n")
```

```
## States that did not end Shelter in Place:  California Hawaii New Mexico New York
```

```
cat("Shelter start date: ", t2, "\n")
```

```
## Shelter start date:  3/19/20 3/25/20 3/24/20 3/22/20
```

```
cat("Shelter end date: ", t1, "\n")
```

```
## Shelter end date:  0 0 0 0
```

```r
cat("Indices of these state in dataset: ", which(is.na(dsp)), "\n")
```

```
## Indices of these state in dataset:  5 12 32 33
```

These 4 states did not end Shelter in Place order. 07/02/20 is the last day when the Covid-19 state policy data was pulled. We go with the assumption that Shelter in Place in these 4 states ended on 07/03/20 i.e., a day later the data was updated.

```r
# get the 4 indexes updated
endDate <- rep(as.Date("07/03/20", format="%m/%d/%y"), each = 4)
stDate <- c(startS[5], startS[12], startS[32], startS[33])
replDate <- as.data.frame(endDate - stDate)
replDate <- replDate$`endDate - stDate`
replDate <- strtoi(replDate)
dsp[5] <- replDate[1]
dsp[12] <- replDate[2]
dsp[32] <- replDate[3]
dsp[33] <- replDate[4]
# convert string into integers
dsp <- strtoi(dsp)
# check if there are any NA left
paste("Are there any NAs?: ", shelterEnd[which(is.na(dsp))])
```

```
## [1] "Are there any NAs?:  "
```

```r
paste("Number of days since Shelter in Place lifted")
```

```
## [1] "Number of days since Shelter in Place lifted"
```

```r
dsp
```

```
##  [1]  26  27  46   0 106  32   0  69  58  45  28 100  37  69  54   0  35   0  53
## [20]  60  46  55  69  51  24  28  29   0  39  80   0 101 103  53   0  57   0  88
## [39]  65  42  27   0  29   0   0  51  60  70  41  49   0
```

Let us also examine the following variables:

- Percent.living.under.the.federal.poverty.line..2018
- Percent.at.risk.for.serious.illness.due.to.COVID
- Population.2018

We can get the people living under poverty line by multiplying the population with the percent living under poverty line. Similar is the case for the people at risk for serious illness.

```r
pdp <- ds$Population.2018 * ds$Percent.living.under.the.federal.poverty.line..2018.
prs <- ds$Population.2018 * ds$Percent.at.risk.for.serious.illness.due.to.COVID
summary(ds$Percent.living.under.the.federal.poverty.line..2018.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.60   10.95   12.80   12.91   14.20   19.70
```

```r
summary(ds$Percent.at.risk.for.serious.illness.due.to.COVID)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   35.95   38.30   38.15   40.65   49.30
```

```r
summary(ds$Population.2018)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   577737  1780020  4468402  6415048  7353618 39557045
```

```
ds$Percent.at.risk.for.serious.illness.due.to.COVID
```

```
##  [1] 43.1 32.8 39.1 43.5 33.3 31.3 36.0 41.3 31.8 42.1 36.2 39.1 36.2 36.2 39.9
## [16] 36.9 38.0 43.6 42.1 42.5 37.1 34.6 41.2 33.9 42.5 40.5 39.0 36.6 36.1 40.5
## [31] 34.6 39.4 36.9 39.0 34.6 39.8 40.8 39.8 39.8 38.3 41.4 35.3 41.6 34.8 30.0
## [46] 39.1 35.9 35.1 49.3 36.5 36.4
```

```
ds$Percent.living.under.the.federal.poverty.line..2018.
```

```
##  [1] 16.8 10.9 14.0 17.2 12.8  9.6 10.4 12.5 16.2 13.6 14.3  8.8 11.8 12.1 13.1
## [16] 11.2 12.0 16.9 18.6 11.6  9.0 10.0 14.1  9.6 19.7 13.2 13.0 11.0 12.9  7.6
## [31]  9.5 19.5 13.6 14.0 10.7 13.9 15.6 12.6 12.2 12.9 15.3 13.1 15.3 14.9  9.0
## [46] 11.0 10.7 10.3 17.8 11.0 11.1
```

```
ds$Population.2018
```

```
##  [1]  4887871   737438  7171646  3013825 39557045  5695564  3572665   967171
##  [9]   702455 21299325 10519475  1420491  1754208 12741080  6691878  3156145
## [17]  2911505  4468402  4659978  1338404  6042718  6902149  9995915  5611179
## [25]  2986530  6126452  1062305  1929268  3034392  1356458  8908520  2095428
## [33] 19542209 10383620   760077 11689442  3943079  4190713 12807060  1057315
## [41]  5084127   882235  6770010 28701845  3161105   626299  8517685  7535591
## [49]  1805832  5813568   577737
```

There does not seem to be any data anomalies for the 3 variables considered for this model improvement #2.

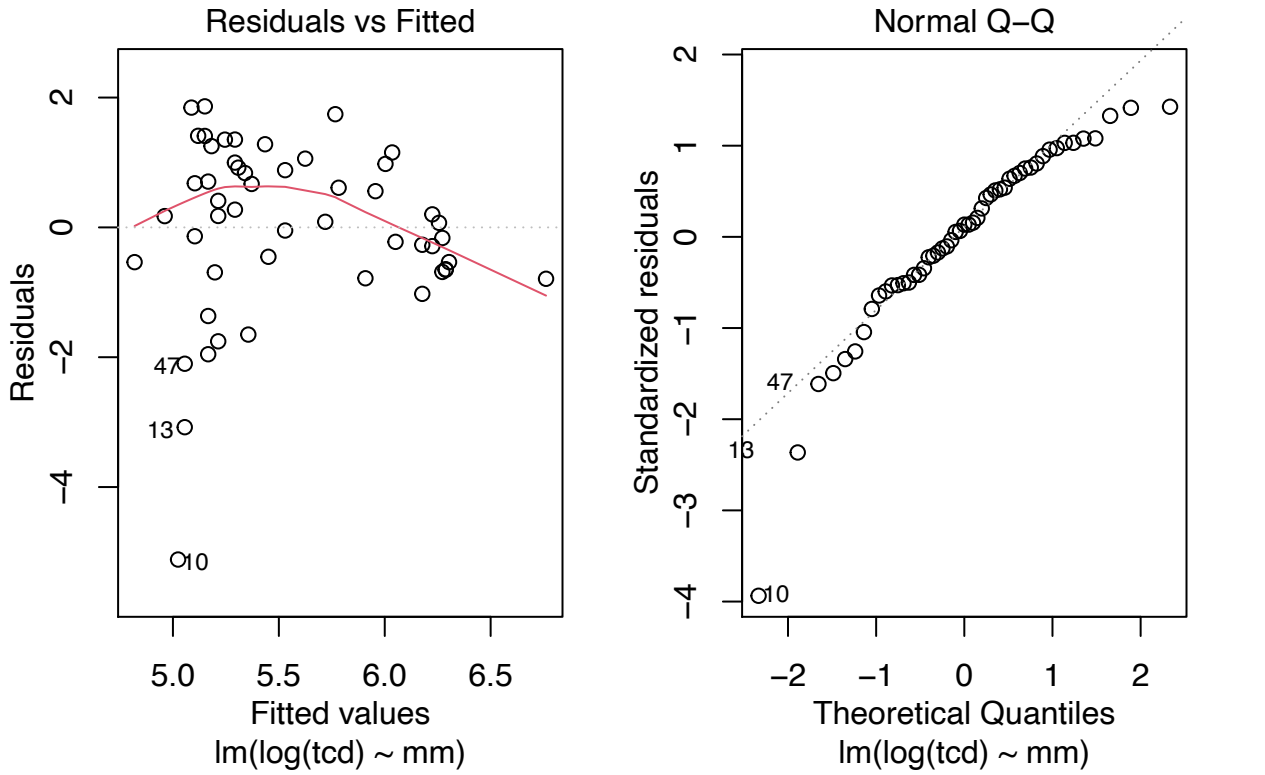
Chapter 3: Model Building Process

3.1: Model 1

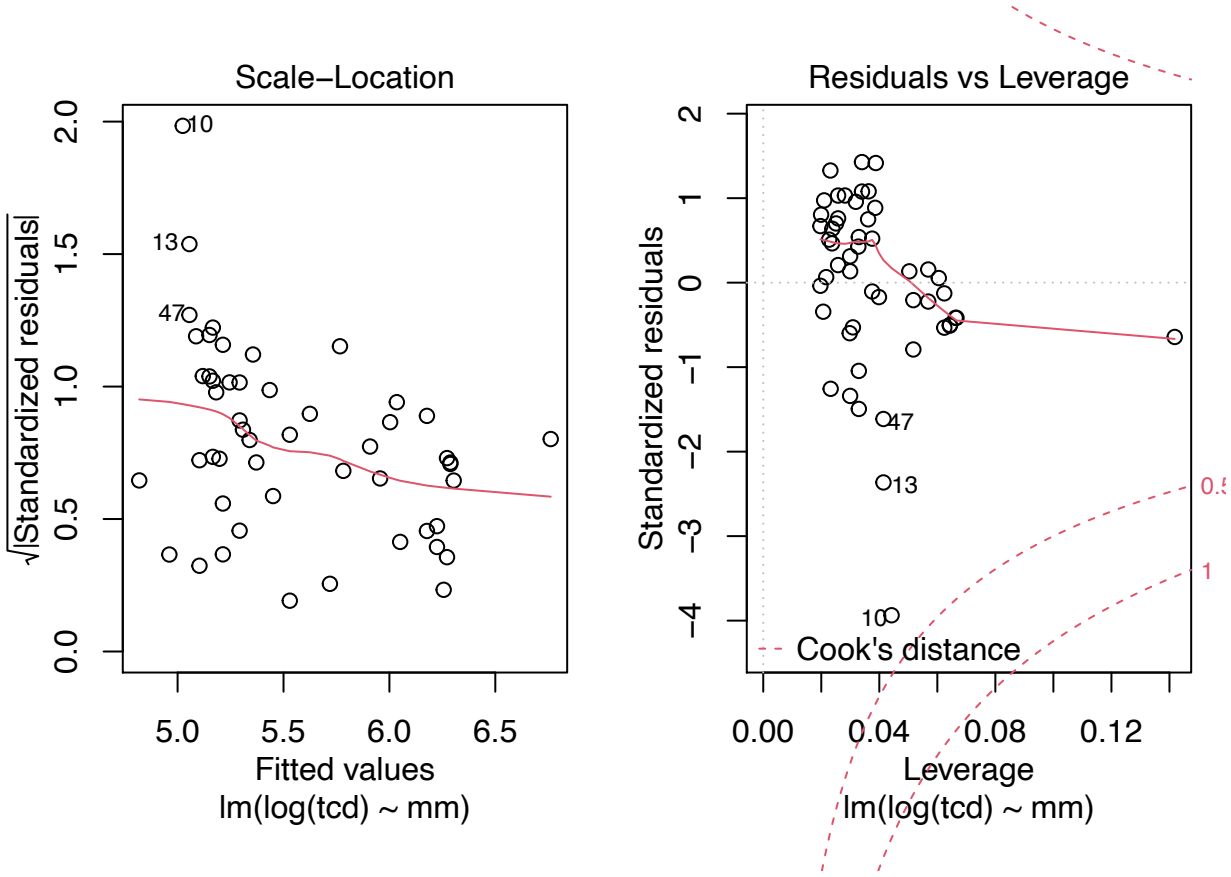
The Basic Model is:

$log(NormalizedTotalCases) = \beta_0 + \beta_1.DaysUntilMaskMandated + u$

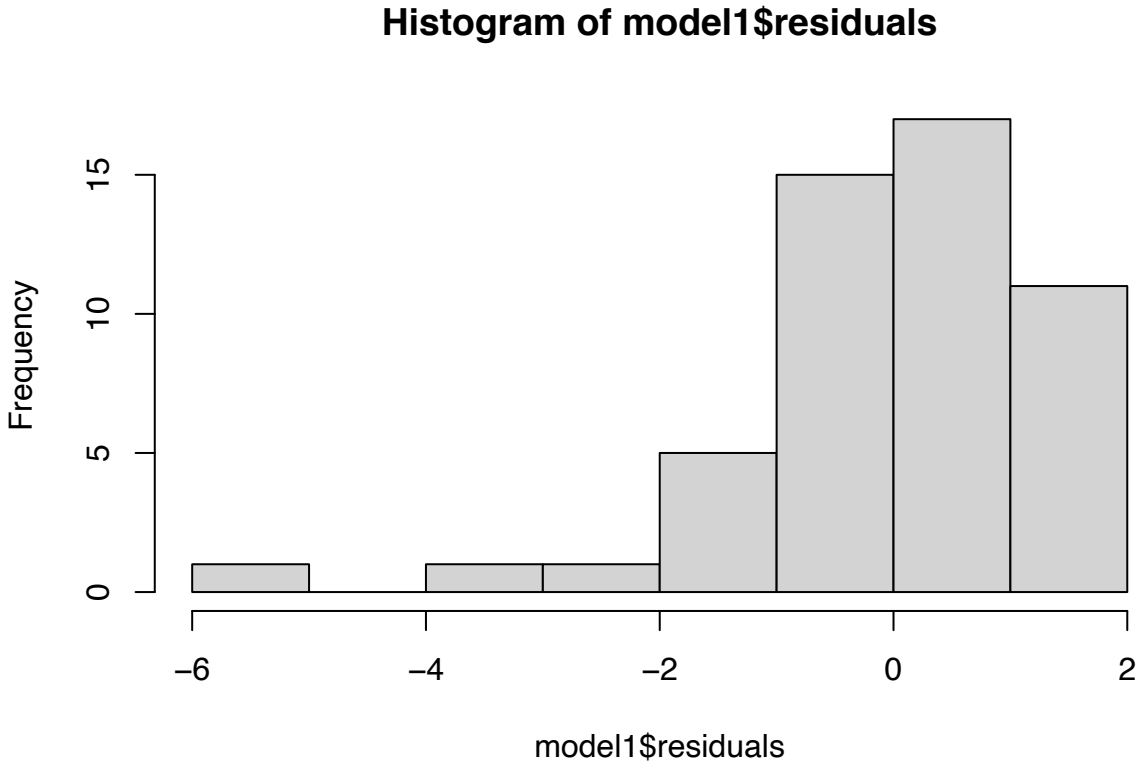
```
#model original
model1 <- lm(log(tcd) ~ mm, data = ds)

plot1 <- as.ggplot(~plot(model1, which = 1))
plot2 <- as.ggplot(~plot(model1, which = 2))
plot3 <- as.ggplot(~plot(model1, which = 3))
plot4 <- as.ggplot(~plot(model1, which = 5))

grid.arrange(plot1, plot2, layout_matrix = rbind(c(1,1,2,2),c(1,1,2,2)))
```




```
grid.arrange(plot3, plot4, layout_matrix = rbind(c(1,1,2,2),c(1,1,2,2)))
```




```
hist(model1$residuals)
```



`mm` is the masked mandated date - first case of COVID-19 date

3.1.1: CLM Assumptions

Let us examine the following 6 six CLM Assumptions for the basic model based on the above plots:

1. Linear in Parameters

We will assume that all models are linear in their parameters as we do not have a way to verify them.

2. Random Sampling

Random Sampling means that the data we obtain are independent and identically distributed (iid) draws from the population distribution. For this specific lab, we know the data is primarily consensus based and we will proceed with our analysis.

3. No Perfect Collinearity

The only independent variable used in this model is days until mask mandated. Since there are no other independent variables involved, we can claim that this assumption has been met

4. Zero conditional Mean

From the Residuals Vs Fitted plot, we can see that Zero Conditional Mean assumption is violated. We can infer that the omitted variables potentially influencing this violation and we will further discuss in the next 2 improvements of this basic model.

5. Homoskedasticity

Looking at the scale-location plot, the homoskedasticity assumption does not hold. So, we use the Heteroskedasticity-consistent estimation of the covariance matrix of the coefficient estimates in regression models.

6. Normality

This basic model confirms to the normality principle as we have observed in the Normal Q-Q plot. There is a slight variation for the data points in the extreme ends of the line. But, given the sample size is >30 (we have 51 data points), we can invoke OLS Asymptotics and Central Limit Theorem and claim normality.

3.1.2: Regression Table

```
# run the coeftest
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value       Pr(>|t|)
## (Intercept) 4.4081858  0.5942514  7.4180 0.000000001497 ***
## mm          0.0157948  0.0061123  2.5841        0.01279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# get the Heteroskedastic Consistent variance-covariance vector
se_model <- sqrt(diag(vcovHC(model1)))
# stargazer output
stargazer(model1, type = "text", omit.stat = "f",
          se = se_model, #Using the robust standard error
          report=('vc*p'),
          title = "Basic Model: Assess mask mandate impact",
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Basic Model: Assess mask mandate impact
## ================================================
##                            Dependent variable:
##                        ----------------------------
##                                   log(tcd)
## ------------------------------------------------
## mm                                 0.016
##
##
## Constant                          4.408***
##                                   p = 0.000
##
## ------------------------------------------------
## Observations                        51
## R2                                 0.120
## Adjusted R2                        0.102
## Residual Std. Error          1.329 (df = 49)
## ================================================
## Note:                  *p<0.05; **p<0.01; ***p<0.001
```

3.1.3: Statistical Significance

Based on the coeftest results, `mm` is statistically significant with a p-value less than 0.05.

3.1.4: Practical Significance

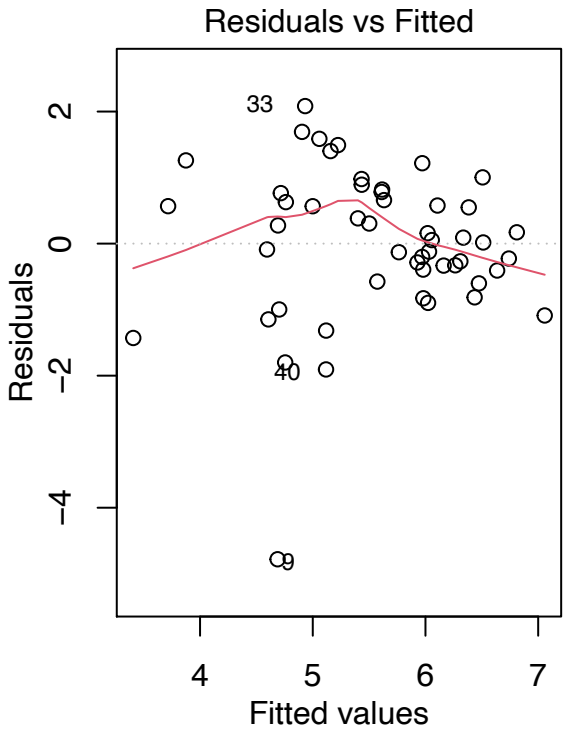
Based on the coef test, `mask mandate` is statistically significant. Based on β_1 value, we can also conclude that for everyday delay in mask mandate, there is a 1.6% increase in the normalized COVID-19 cases. From the Residuals Vs Leverage plot, we can see that data points that have leverage however, they do not significantly impact residuals.

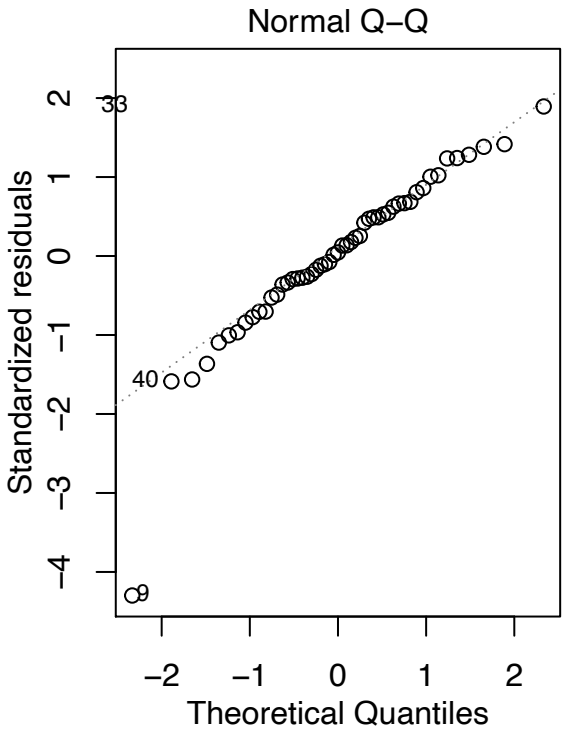
3.2: Model 2

Let us further enhance the regression model as:

$log(NormalizedTotalCases) = \beta_0 + \beta_1.DaysUntilMaskMandated + \beta_2.DaysUntilBusinessesReopened + \beta_3.DaysUntilShelterInPlace + \beta_4.DaysUntilMaskMandated * DaysUntilBusinessesReopened + \beta_5.DaysUntilBusinessesReopened * DaysUntilShelterInPlace + \beta_6.DaysUntilMaskMandated * DaysUntilBusinessesReopened * DaysUntilShelterInPlace + u$

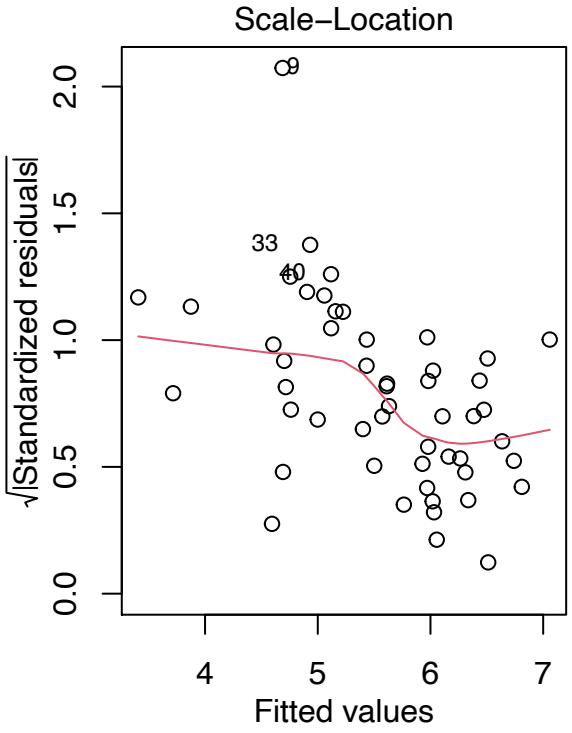
```
#model2 <- lm(log(tcd) ~ mm + drs + dsp + mm * drs + mm * dsp + drs * dsp)
model2 <- lm(log(tcd) ~ mm + drs + dsp + mm * drs + mm * dsp + mm * drs * dsp)

plot1 <- as.ggplot(~plot(model2, which = 1))
plot2 <- as.ggplot(~plot(model2, which = 2))
plot3 <- as.ggplot(~plot(model2, which = 3))
plot4 <- as.ggplot(~plot(model2, which = 5))

grid.arrange(plot1, plot2, layout_matrix = rbind(c(1,1,2,2),c(1,1,2,2)))
```

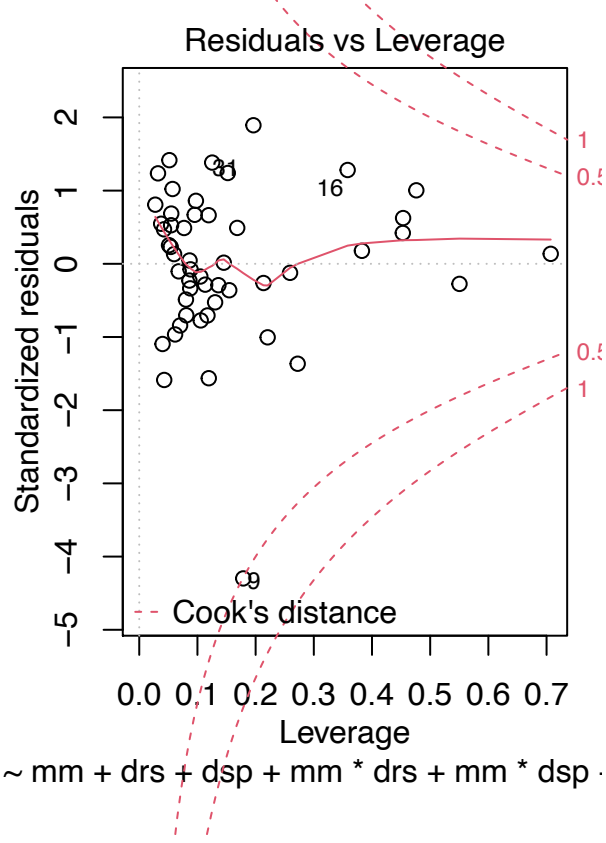




l) ~ mm + drs + dsp + mm * drs + mm * dsp l) ~ mm + drs + dsp + mm * drs + mm * dsp

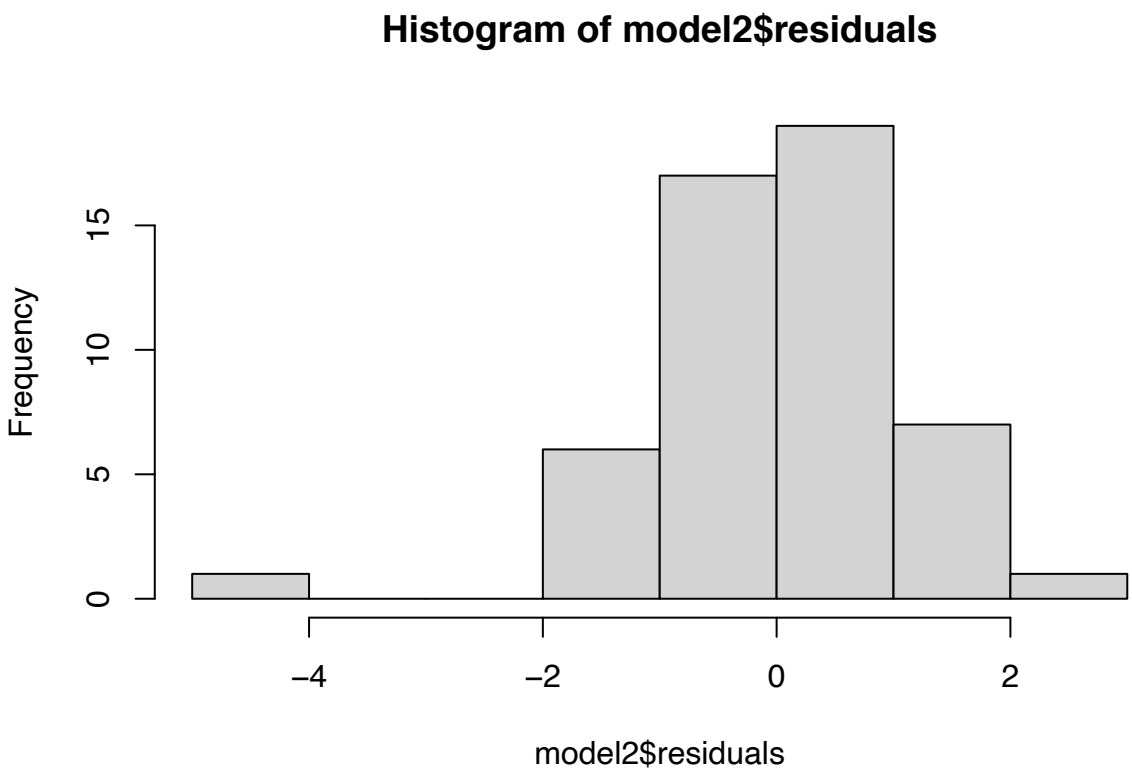
```
grid.arrange(plot3, plot4, layout_matrix = rbind(c(1,1,2,2),c(1,1,2,2)))
```





l) ~ mm + drs + dsp + mm * drs + mm * dsp l) ~ mm + drs + dsp + mm * drs + mm * dsp

```
hist(model2$residuals)
```



`drs` is the businesses reopen date - businesses closed date `dsp` is the shelter in place end date - shelter in place start date

3.2.1: CLM Assumptions

In this enhanced model, there is no change in the assumptions related to Linear in Parameters, Random Sampling, Normality and hence we do not cover them here. Let us examine the other 3 assumptions:

3. No Perfect Collinearity

The 3 variables involved in the model are: `DaysUntilMaskMandated`, `DaysUntilBusinessesReopened`, & `DaysUntilShelterInPlace`. Let's examine the correlation among these variables.

```
c1 <- as.numeric(mm)
c2 <- as.numeric(drs)
c3 <- as.numeric(dsp)
cols <- data.frame(mask_mandate = c1, business_reopen = c2, shelter_inplace = c3)
corrMatrix <- cor(cols)
corrMatrix
```

```
##                 mask_mandate business_reopen shelter_inplace
## mask_mandate      1.00000000      -0.3531170     -0.09096103
## business_reopen  -0.35311697       1.0000000      0.38303721
## shelter_inplace  -0.09096103       0.3830372      1.00000000
```

Based on the above, we see that the correlation among all three variables vary in the range -0.36 to 0.38. We can conclude that there is no perfect collinearity among the involved variables of the model.

4. Zero conditional Mean

From the Residuals Vs Fitted plot, we can see that Zero Conditional Mean assumption is violated. We can

infer that the omitted variables potentially influencing this violation and we will further discuss in the next improvement of this improved model.

5. Homoskedasticity

Looking at the scale-location plot, the homoskedasticity assumption does not hold. So, we use the Heteroskedasticity-consistent estimation of the covariance matrix of the coefficient estimates in regression models.

3.2.2: Regression Table

```
# run the coeftest
coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                 Estimate   Std. Error t value   Pr(>|t|)
## (Intercept) 14.928237265  2.851153314  5.2359 0.00000466 ***
## mm          -0.074722213  0.024547662 -3.0440   0.003974 **
## drs         -0.201177733  0.068364921 -2.9427   0.005226 **
## dsp         -0.203048009  0.083055701 -2.4447   0.018670 *
## mm:drs       0.001591024  0.000618282  2.5733   0.013606 *
## mm:dsp       0.001660641  0.000864441  1.9211   0.061370 .
## drs:dsp      0.003590787  0.001772160  2.0262   0.048974 *
## mm:drs:dsp  -0.000025843  0.000018102 -1.4276   0.160636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# get the Heteroskedastic Consistent variance-covariance vector
se_model <- sqrt(diag(vcovHC(model2)))
# stargazer output
stargazer(model2, type = "text", omit.stat = "f",
          se = se_model, #Using the robust standard error
          report=('vc*p'),
          title = "Improved Model#1: Assess mask mandate impact",
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## Improved Model#1: Assess mask mandate impact
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                log(tcd)
## ------------------------------------------------
## mm                              -0.075
##
##
## drs                             -0.201
##
##
## dsp                             -0.203
##
##
## mm:drs                           0.002
##
```

```
##
## mm:dsp                            0.002
##
##
## drs:dsp                           0.004
##
##
## mm:drs:dsp                      -0.00003
##
##
## Constant                         14.928***
##                                 p = 0.00000
##
## ------------------------------------------------
## Observations                       51
## R2                                0.342
## Adjusted R2                       0.235
## Residual Std. Error          1.227 (df = 43)
## ================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

3.2.3: Statistical Significance

Based on the coeftest results, the following variables are statistically significant with a p-value less than 0.05:

- `mm`, `drs`, `dsp`
- interaction between `mm` & `drs`; `drs` & `dsp`

3.2.4: Practical Significance

We introduced `DaysUntilBusinessesReopened` and `DaysUntilShelterInPlace` in the second model. We know that the two new variables have an impact on the normalized total cases. However, we also want to study the interaction between the days until mask was mandated with the days when business was shut down and the days when shelter in place was in effect.

Reviewing the coeftest output, we conclude the following about this model:

- An addition of a day to the mask mandate date will result in a 7.422% decrease in the normalized test cases
- An addition of a day to reopen businesses date will result in a 20.11% decrease in the normalized test cases
- An addition of a day to shelter in place order will result in a 20.30% decrease in the normalized test cases
- An addition of a day to both mask mandate date as well as reopen businesses date will result in a 0.16% increase in the normalized test cases
- An addition of a day to both mask mandate date as well as shelter in place order will result in a 0.17% increase in the normalized test cases
- An addition of a day to both reopen businesses date as well as shelter in place order will result in a 0.36% increase in the normalized test cases

We observed high standard errors in comparison to the coefficients. Also, the sign of mask mandate data changed across the two models. Therefore, we conclude this model is not practically significant.

3.3: Model 3

Through the above listed operationalized variables, let's enhance the regression model further as:

$log(NormalizedTotalCases) = \beta_0 + \beta_1.DaysUntilMaskMandated + \beta_2.DaysUntilBusinessesReopened + \beta_3.DaysUntilShelterInPlace + \beta_4.log(PeopleUnderPovertyLine) + \beta_5.log(PeopleAtRiskForSeriousIllness) + u$

Before we add these 2 new variables, let's check their correlation

```
# PeopleUnderPovertyLine
c4 <- as.numeric(pdp)
# PeopleAtRiskForSeriousIllness
c5 <- as.numeric(prs)
cols <- data.frame(PplUPoverty = c4, PplAtRisk = c5)
corrMatrix <- cor(cols)
corrMatrix
```

```
##             PplUPoverty PplAtRisk
## PplUPoverty   1.0000000 0.9892368
## PplAtRisk     0.9892368 1.0000000
```

As you can see from the above test results these two variables have a very high correlation. We have also observed that `PeopleAtRiskForSeriousIllness` has strong correlation with `X65.` (99.32% correlation) & `All.cause.deaths.2018` (99.26% correlation)

We decided to keep `PeopleAtRiskForSeriousIllness` and to add these other variables instead:

- `Weekly.unemployment.insurance.maximum.amount..dollars.`
- `CasesInLast7Days`

Before running the model, let's check the correlations again:

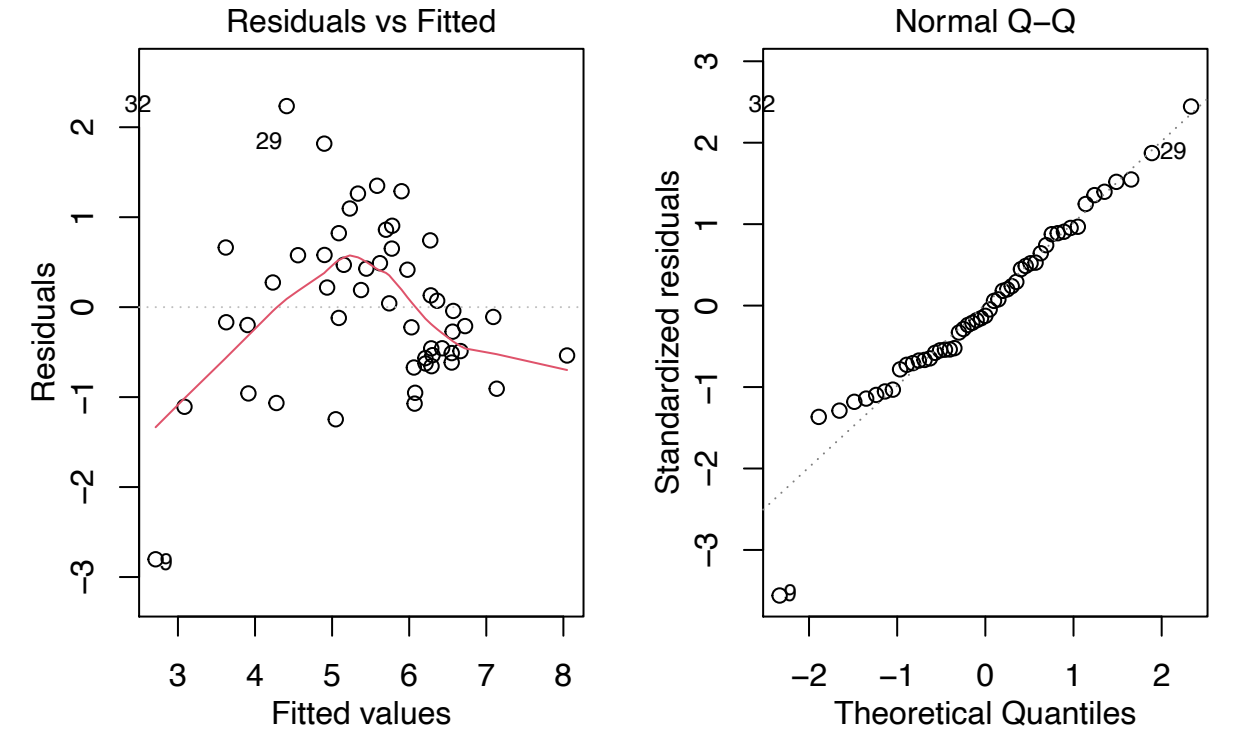
```
l7d <- ds$CasesInLast7Days
wui <- ds$Weekly.unemployment.insurance.maximum.amount..dollars.
# PeopleAtRiskForSeriousIllness, CasesInLast7Days
# Weekly.unemployment.insurance.maximum.amount..dollars.
cols <- data.frame(PplAtRisk = c5, l7d = l7d, wui = wui)
corrMatrix <- cor(cols)
corrMatrix
```

```
##             PplAtRisk        l7d         wui
## PplAtRisk  1.00000000  0.8258895 -0.05750598
## l7d        0.82588950  1.0000000 -0.25874661
## wui       -0.05750598 -0.2587466  1.00000000
```

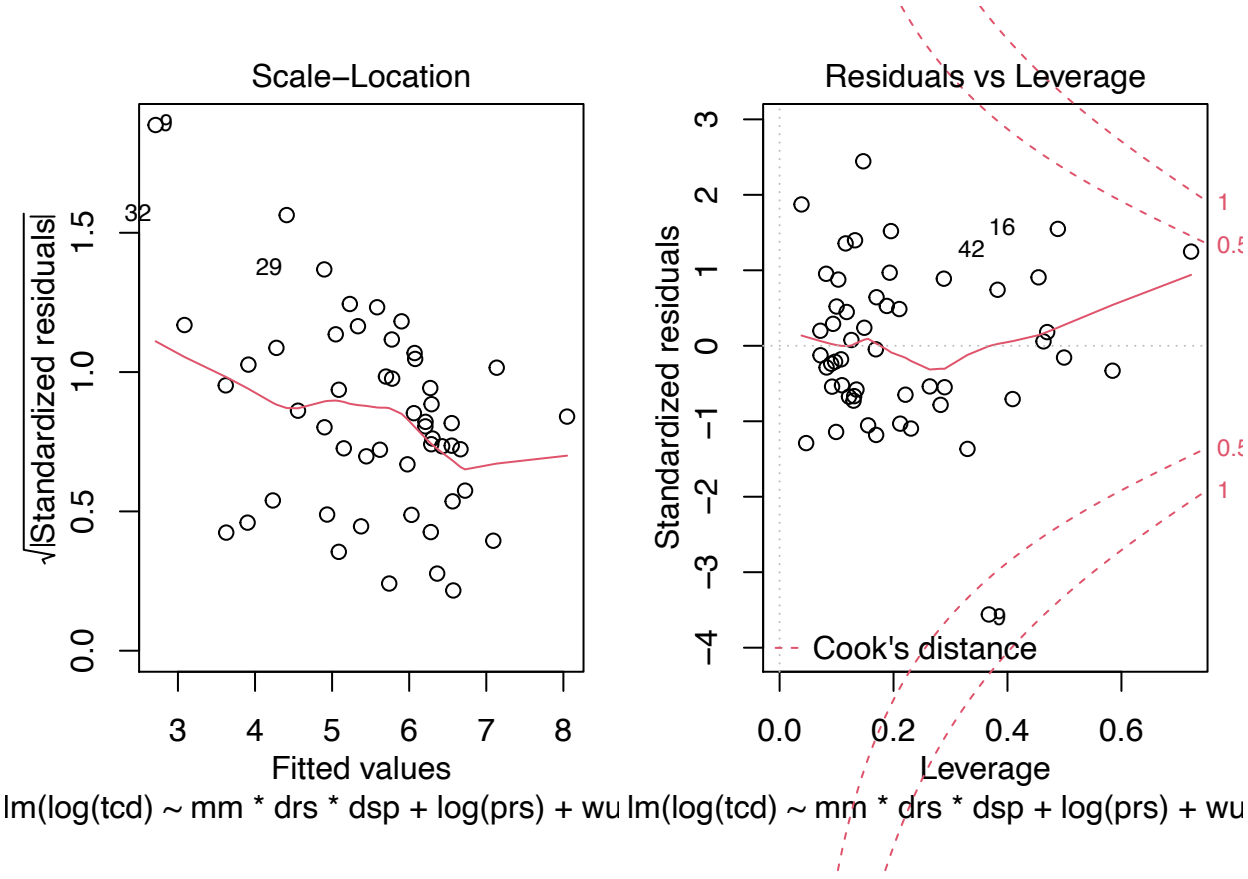
The correlation of these variables seems good for their inclusion in the second improvement of the model. The model is therefore:

$log(NormalizedTotalCases) = \beta_0 + \beta_1.DaysUntilMaskMandated * \beta_2.DaysUntilBusinessesReopened * \beta_3.DaysUntilShelterInPlace + \beta_4.log(PeopleAtRiskForSeriousIllness) + \beta_5.Weekly.unemployment.insurance.maximum. \beta_6.CasesInLast7Days + u$

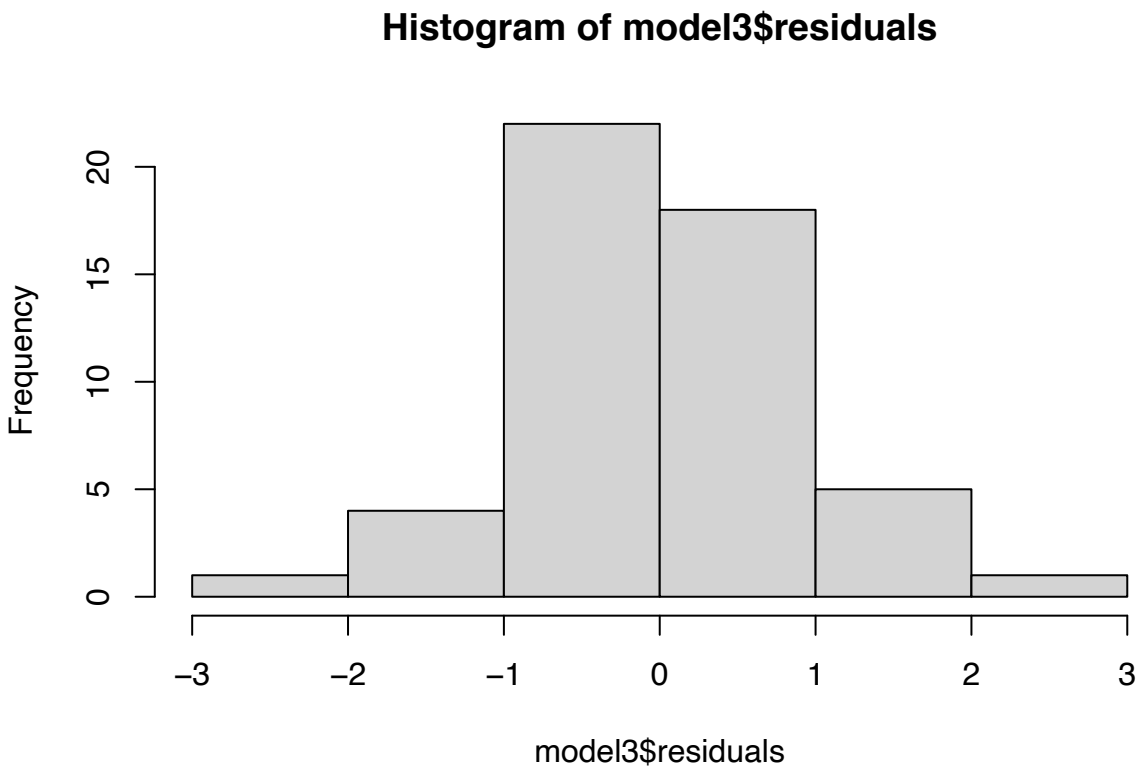
```
model3 <- lm(log(tcd) ~ mm * drs * dsp + log(prs) + wui + l7d)

plot1 <- as.ggplot(~plot(model3, which = 1))
plot2 <- as.ggplot(~plot(model3, which = 2))
plot3 <- as.ggplot(~plot(model3, which = 3))
plot4 <- as.ggplot(~plot(model3, which = 5))

grid.arrange(plot1, plot2, layout_matrix = rbind(c(1,1,2,2),c(1,1,2,2)))
```




```
grid.arrange(plot3, plot4, layout_matrix = rbind(c(1,1,2,2),c(1,1,2,2)))
```



```r
hist(model3$residuals)
```



`prs` is the percent people at risk for serious illness due to covid & 2018 population `wui` is the weekly unemployment insurance maximum amount dollars `l7d` is the CasesInLast7Days

3.3.1: CLM Assumptions

In this improved Model, there is no change in the assumptions related to Linear in Parameters, Random Sampling, Normality and hence we do not cover them here. Let us examine the other 3 assumptions:

3. No Perfect Collinearity

The additional variables involved in the model are: `PeopleAtRiskForSeriousIllness`, `Weekly.unemployment.insurance.max` `CasesInLast7Days`, `Death_100k`. Based on our correlation study, there does not exist any perfect collinearity among these variables.

4. Zero conditional Mean

From the Residuals Vs Fitted plot, we can see that Zero Conditional Mean assumption is violated. We can infer that the omitted variables potentially influencing this violation. We will now address the omitted variables in the next section.

5. Homoskedasticity

Looking at the scale-location plot, the homoskedasticity assumption does not hold. So, we use the Heteroskedasticity-consistent estimation of the covariance matrix of the coefficient estimates in regression models.

3.3.2: Regression Table

```r
# run the coeftest
coeftest(model3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                 Estimate   Std. Error t value Pr(>|t|)
## (Intercept) -0.163810339  6.860410488 -0.0239 0.981069
## mm          -0.084335224  0.034126797 -2.4712 0.017821 *
## drs         -0.235596153  0.072066715 -3.2691 0.002222 **
## dsp         -0.189329011  0.069135281 -2.7385 0.009170 **
## log(prs)     0.896825245  0.414404233  2.1641 0.036481 *
## wui          0.000408113  0.001216951  0.3354 0.739108
## l7d         -0.000013341  0.000018261 -0.7306 0.469295
## mm:drs       0.001749080  0.000802010  2.1809 0.035132 *
## mm:dsp       0.001443221  0.000860984  1.6762 0.101490
## drs:dsp      0.003466370  0.001390355  2.4932 0.016899 *
## mm:drs:dsp  -0.000024909  0.000017370 -1.4340 0.159345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# get the Heteroskedastic Consistent variance-covariance vector
se_model <- sqrt(diag(vcovHC(model3)))
# stargazer output
stargazer(model3, type = "text", omit.stat = "f",
          se = se_model, #Using the robust standard error
          report=('vc*p'),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                log(tcd)
## ------------------------------------------------
## mm                              -0.084
##
##
## drs                             -0.236
##
##
## dsp                             -0.189
##
##
## log(prs)                         0.897
##
##
## wui                             0.0004
##
##
## l7d                            -0.00001
##
##
## mm:drs                           0.002
##
##
## mm:dsp                           0.001
##
```

```
##
## drs:dsp                          0.003
##
##
## mm:drs:dsp                     -0.00002
##
##
## Constant                        -0.164
##                                p = 0.981
##
## ------------------------------------------------
## Observations                      51
## R2                               0.602
## Adjusted R2                      0.502
## Residual Std. Error        0.990 (df = 40)
## ================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

3.3.3: Statistical Significance

Based on the coeftest results, the following variables are statistically significant with a p-value less than 0.05:

- `mm`, `drs`, `dsp`, `log(prs)`
- interaction between `mm` & `drs`; `drs` & `dsp`

3.3.4: Practical Significance

We introduced `prs`, `wui`, and `l7d` in the third model. We introduced variables which are not high collinear with each other into the model. Reviewing the coeftest output, we conclude the following about this model:

- An addition of a day to the mask mandate date will result in a 8.43% decrease in the normalized test cases
- An addition of a day to reopen businesses date will result in a 23.56% decrease in the normalized test cases
- An addition of a day to shelter in place order will result in a 18.93% decrease in the normalized test cases
- An addition of 1% of people at risk for serious illness due to covid will results in a 0.9% of increase in the normalized test cases
- An addition of a day to both mask mandate date as well as reopen businesses date will result in a 0.17% increase in the normalized test cases
- An addition of a day to both reopen businesses date as well as shelter in place order will result in a 0.34% increase in the normalized test cases

We observed high standard errors in comparison to the coefficients. The new variables introduced did not impact the practical significance. Therefore, we conclude this model is not practically significant.

Chapter 4: Omitted Variables

We used model 1 to define the omitted variables' impact. To generalize an omitted variable, we can use below definitions:

$$log(tcd) = \beta_0 + \beta_1 mm + u$$

$$log(tcd) = \beta_0 + \beta_1 mm + \beta_2 OV + u$$

$$OV = \alpha_0 + \alpha_1 mm + v$$

where OV stands for Omitted Variable

$$log(tcd) = \beta_0 + \beta_1 mm + \beta_2(\alpha_0 + \alpha_1 mm + v)u$$

$$log(tcd) = (\beta_0 + \beta_2.\alpha_0) + (\beta_1 + \beta_2.\alpha_1)mm + (\beta_2.v + u)$$

We are assuming that $\beta_1 > 0$ as mask mandate implementation is delayed. Therefore, normalized test cases will increase.

The OV bias is estimated using $\beta_2.\alpha_1$. Let's assess bias for the following omitted variables.

- IsMale: There has been a general outcry form certain states in US to not adopt mask wearing. This clearly delayed the mask mandate date in those states. It was predominantly men who had this opinion. Using Gender as an omitted variable, let's assess the coefficients.

  The pushback from men imply mask mandate date is delayed. Therefore, $\alpha_1 > 0$. We also assume that men are susceptible to the virus than women which imply $\beta_2 > 0$. Therefore, the omitted variable bias in this case is > 0 and will result in the over-estimation of β_1.

- Mask Wearing: This is another omitted variable that indicates if mask wearing was prevalent given that mask mandate was implemented. Re-using above set of equations, we can argue that mask wearing will not be adopted if the mask mandate is delayed or not implemented. Therefore, $alpha_1 < 0$. We also assume that wearing mask will reduce the spread of the virus. Therefore, $\beta_2 < 0$. Therefore, the omitted variable bias in this case is > 0 and will result in the over-estimation of β_1.

- Number of Test Kits: Test kit availability will have an impact on numner of test cases as well as on the date of mask mandate implementation. Less number of test kits imply less confirmed cases and thereby authorities assuming fewer cases of covid causing a delay in the mask mandate implementation decision. Therefore, $\alpha_1 < 0$. More testing would confirm more cases, i.e., $\beta_2 > 0$. Therefore, the omitted variable bias in this case is < 0 and will result in the under-estimation of β_1.

- Emergency Funding: This omitted variable is similar in nature to the Number of Test Kits. More Emergency Funding implies, better infrastructure that can handle covid cases and can potentially cause the delay in the Mask Mandate implementation. Therefore, $\alpha_1 < 0$ and $\beta_2 > 0$. Therefore, the omitted variable bias in this case is < 0 and will result in the under-estimation of β_1.

- Access to Information (Internet, TV): This omitted variable is a form of Education. We assume that more the venues through which the people get awareness of the virus, the mask mandate date implementation will be sooner. Therefore, $alpha_1 > 0$. We also assume that access to information will drive the normalized test cases to go down i.e., $\beta_2 < 0$. Therefore, the omitted variable bias in this case is < 0 and will result in the under-estimation of β_1.

Chapter 5: Conclusion

```
(se.model1 = sqrt(diag(vcovHC(model1))))
```

```
## (Intercept)          mm
## 0.594251436 0.006112254
```

```
(se.model2 = sqrt(diag(vcovHC(model2))))
```

```
##   (Intercept)            mm           drs           dsp        mm:drs
## 2.85115331386 0.02454766237 0.06836492105 0.08305570098 0.00061828249
##        mm:dsp       drs:dsp    mm:drs:dsp
## 0.00086444129 0.00177216032 0.00001810236
```

```
(se.model3 = sqrt(diag(vcovHC(model3))))
```

```
##   (Intercept)            mm           drs           dsp      log(prs)
## 6.86041048825 0.03412679659 0.07206671461 0.06913528050 0.41440423271
##          wui           l7d        mm:drs        mm:dsp       drs:dsp
## 0.00121695118 0.00001826147 0.00080200976 0.00086098445 0.00139035467
##    mm:drs:dsp
## 0.00001737048
```

```
stargazer::stargazer(model1, model2, model3, type = "text", omit.stat = "f",
                     se = list(se.model1, se.model2, se.model3),
                     star.cutoffs = c(0.05, 0.01, 0.001),
                     title = "Three Models Comparison")
```

```
##
## Three Models Comparison
## ========================================================================
##                                  Dependent variable:
##                     ---------------------------------------------
##                                        log(tcd)
##                           (1)             (2)             (3)
## ------------------------------------------------------------------------
## mm                      0.016**        -0.075**         -0.084*
##                         (0.006)         (0.025)         (0.034)
##
## drs                                    -0.201**        -0.236**
##                                         (0.068)         (0.072)
##
## dsp                                     -0.203*         -0.189**
##                                         (0.083)         (0.069)
##
## log(prs)                                                 0.897*
##                                                          (0.414)
##
## wui                                                      0.0004
##                                                          (0.001)
##
## l7d                                                     -0.00001
##                                                         (0.00002)
##
## mm:drs                                   0.002*          0.002*
##                                         (0.001)         (0.001)
```

```
##
## mm:dsp                                  0.002           0.001
##                                        (0.001)         (0.001)
##
## drs:dsp                                 0.004*          0.003*
##                                        (0.002)         (0.001)
##
## mm:drs:dsp                            -0.00003        -0.00002
##                                       (0.00002)       (0.00002)
##
## Constant                  4.408***     14.928***        -0.164
##                            (0.594)      (2.851)         (6.860)
##
## ------------------------------------------------------------------------
## Observations                 51           51              51
## R2                          0.120        0.342           0.602
## Adjusted R2                 0.102        0.235           0.502
## Residual Std. Error 1.329 (df = 49) 1.227 (df = 43) 0.990 (df = 40)
## ========================================================================
## Note:                                  *p<0.05; **p<0.01; ***p<0.001
```

We set out to answer the research question: "Does wearing a mask help prevent the spread of COVID-19?". We created three models with some statistically significant variables. Through our EDA, we purposefully stayed away from variables such as `Total Death` & `Total Test Results`. These variables could have produced a much better fitting model. However, we would have deviated from the research question to assess the mask mandate impacts on the total test cases.

Model3, which had most of the relevant variables included, had an adjusted R^2 of around 50%. Finally, we conclude that we do not have enough information to create a model that explains enough of the variation in our dataset. Therefore, we cannot conclude that wearing a mask help prevent the spread of COVID-19. One potential reason for the unexplained variability is the omitted variables as addressed in the previous chapter. Availability of time-series data that highlights the test cases results before and after mask mandate implementation could have helped improved our final model.