Lab 2: Regression to Study the Spread of Covid-19

Ziwei Zhao, Jonathan Moges, and Brendan Mattina

12/8/2020

Introduction

The United States is in a crippling Third Wave of the COVID-19 Pandemic. States with seemingly catastrophic summertime infection rates have exceeded those rates by nearly and uniformly 100% with the onset of early winter. Tragically, the death rate has risen in concert with infection rate, setting the stage for a grim winter. States have responded with disparate strategies to mitigate COVID's impact, but, only some state responses seem grounded in science, while others seem grounded in public opinion. Despite political realities, nearly every state must critically explore the best and most palatable strategies to keep their populations safe over these next, critical months.

To address that concern, our team developed models that explored the association between the death rate in each state, and that state's implementation of CDC-recommended COVID mitigation strategies. Our central model will attempt to answer the following question:

Model 1: How are state-imposed mask mandates associated with total COVID deaths (per million people) in each state?

We will also develop secondary models that attempt to answer the following ancillary questions:

Model 2: How closely is COVID public opinion in each state associated with its total COVID deaths (per million people)?

Model 3: How closely is the imposition of 'secondary' state-wide COVID mitigation strategies and mobility data associated with total COVID deaths (per million people) in each state?

Most states have imposed state-wide mask mandates at some point since April. Our central model incorporates this strategies for two reasons. Mask mandates are core components of the United States' Center for Disease Control (CDC) suggested response and are widely acclaimed by experts as one of the most effective strategies for arresting COVID's impact. As such, a majority of states employed this mandate.

However, America is politically polarized like almost no other time in its history. Public opinion about COVID swings wildly from state to state. We felt it necessary to capture this dynamic and assess the association between public regard for COVID and death rates because state-wide mandates will only work well with buy-in from local populations . Through our research, we identified the 2020 American National Election Studies (ANES) as a viable quantification of Americans' personal concerns about COVID. A seemingly disenfranchised population, unconcerned with getting COVID, will not abide by any mandates. This situation could yield a death rate that belies a cautious approach by that state's government.

Our tertiary model will incorporate popular 'secondary' mitigation strategies, such as: state-wide stay at home orders, bar, gym, and school closures; inside dining restrictions; and, personal mobility data to capture any pertinent associations we might have missed in our first two models. Mobility data, which is drawn from Google location services databases, will describe how people in each state changed their travel habits during the pandemic. This data is limited in that it only captures movement for people with cell phones with Google Location Services activated, but it does afford us some sense of the pandemic movement trends.

Our state-wise data is pulled from local state databases and is a snapshot current as of 10/30/2020. For most of our modeled variables, we must regard associations with care, as states may have suspended mandates observed in our models months before the data capture. We will pull public opinion data from the American National Election Studies - 2020 survey, and will transform survey responses, collected as ordinal data, to make it more appropriate for descriptive analysis. Furthermore, we have 51 outcomes to observe (one for each state, plus Washington DC). Considering the inherent limitations and the reality that the granularity of the data will evolve with the pandemic, we consider this data appropriate for gleaning initial insights into the association of mitigation strategies and state by state death rates. We will discuss data transformation and limitations at length later in this report.

Model Building

General Discussion

All of our models are descriptive models that attempt to describe the association between an expanding number of explanatory variables and the outcome variable, which is the ratio of COVID deaths per million people. This ratio, calculated as total deaths as of 30 October 2020, divided by total population divided by 1 million, gives us a normalized statistic that we can assess across states.

In conceptualizing Model 1, we believed that COVID deaths would be most strongly associated with state-wide mask mandates, and would particularly yield a negative association. While a strong association held, we actually discovered a a positive vice negative association, which we will discuss below. Furthermore, a visual association between percent of ANES respondents in a given state concerned about COVID (further discussion below) and its total deaths, persuaded us to build a second model to study how public opinion could reinforce or detract from state-wide mandates. Virtually every variable we added to our Models 2 and 3 increased the risk of collinearity, as states with prolonged mask mandates seemed much more likely to impose secondary strategies. Additionally, state governments actions were likely influenced at least partially by public opinion. Finally, both public opinion and state-wide mandates would presumably impact mobility, which suggests a collinear relationship. A more detailed discussion on collinearity between explanatory variables can be found later in our report.

Data Processing

```r
install.packages("readxl")
install.packages("patchwork")
install.packages("sandwich")
install.packages("lmtest")
install.packages("stargazer")
library(dplyr)
library("readxl")
library(data.table)
library(tidyverse)
library(magrittr)
library(ggplot2)
library(patchwork)
library(sandwich)
library(stargazer)
library(lmtest)
```

In the following section, we read in four different raw data sets used in our analysis, from which we create new variables from existing ones such as number of mask mandate days, number of bar closure days, and aggregated variables related to mobility and public sentiment (worried). We merge these data sets together, conduct basic checks such as tabulating values for outliers, and rename relevant variables for easier usage.

```r
### Read in Covid 19 data
covid <- read.csv("covid-19.csv")

# Calculate number of mask mandate days
```

```
setnames(covid, old = c("Mandate.face.mask.use.by.all.individuals.in.public.spaces",
                        "State.ended.statewide.mask.use.by.individuals.in.public.spaces"),
                new = c("mask_start", "mask_end"))
covid$mask_start <- as.Date(covid$mask_start, "%m/%d/%y")
covid$mask_end[covid$mask_end=="0"] <- "10/30/2020"
covid$mask_end <- as.Date(covid$mask_end, "%m/%d/%y")
covid$mask_days <- as.numeric(covid$mask_end-covid$mask_start)
covid$mask_days[is.na(covid$mask_days)] <- 0

# Calculate number of shelter at home days
setnames(covid, old = c("End.stay.at.home.shelter.in.place",
                        "Stay.at.home..shelter.in.place"),
                new = c("shelter_end", "shelter_start"))
covid$shelter_start <- as.Date(covid$shelter_start, "%m/%d/%y")
covid$shelter_end[covid$shelter_end=="0"] <- "10/30/2020"
covid$shelter_end <- as.Date(covid$shelter_end, "%m/%d/%y")
covid$stayathome_days <- as.numeric(covid$shelter_end-covid$shelter_start)
covid$stayathome_days[is.na(covid$stayathome_days)] <- 0
covid$stayathome_days
```

```
##  [1]  26  27  46   0 225  32   0  69  58  45  28  67  37  69  54   0  35   0  53
## [20]  59  46  55  69  51  24  28  29   0  39  80  80 220  97  53   0  57   0  88
## [39]  65  42  27   0  27   0   0  52  60  70  42  49   0
```

```
# Create indicator variable for stay at home and mask mandates
covid <- covid %>%
  mutate(ind_mask = as.numeric(covid$mask_days>0),
         ind_stay = as.numeric(covid$stayathome_days>0))

table(covid$ind_stay)
```

```
## 
##  0  1 
## 11 40
```

```
table(covid$ind_mask)
```

```
## 
##  0  1 
## 17 34
```

```
# Make dependent variable: deaths per 1,000,000
setnames(covid, old = c('Total.Deaths','Population.2018'), new = c('tot_dth','st_pop'))
covid <- covid %>%
  mutate(dth_per_mil = tot_dth/(st_pop/1000000))


### Read in mobility data
data_mob <- read.csv('2020_US_Region_Mobility_Report.csv')

# Detect the code for state wide data
data_mob_states <- data_mob[str_detect(data_mob$iso_3166_2_code,'US'),]

# Caluclate mean for mobility for each state in each category (work, home, retail)
mob_means_wk <- aggregate(workplaces_percent_change_from_baseline
                          ~sub_region_1,data_mob_states, mean)
```

```r
mob_means_hm <- aggregate(residential_percent_change_from_baseline
                          ~sub_region_1,data_mob_states, mean)
mob_means_ret <- aggregate(retail_and_recreation_percent_change_from_baseline
                           ~sub_region_1,data_mob_states, mean)

# Combine with main covid data
covid$mob_data_wk <- mob_means_wk$workplaces_percent_change_from_baseline
covid$mob_data_hm <- mob_means_hm$residential_percent_change_from_baseline
covid$mob_data_ret <- mob_means_ret$retail_and_recreation_percent_change_from_baseline


### Read in 2020 ANES data
A <- read.csv("anes_pilot_2020.csv")
A <- na.omit(A)

#creating worried variable to reflect those worried about catching covid
A$worried <- (A$covid1 < 4)

#grouping responses and creating ratio of those worried about
#getting covid and those not worried - lower ratio, higher concern.

A1 <- aggregate(x = A$covid1,
                by = list(A$state),
                FUN = sum)

A2 <- aggregate(x = A$covid1 < 4,
                by = list(A$state),
                FUN = sum)


A3 <- inner_join(A1, A2, by = "Group.1")

setnames(A3, old =c('Group.1', 'x.x','x.y'), new = c('State','worried','covid1'))

A3$worried <- round((A3$covid1/A3$worried),2)

#create new dataframe and group by state
B <- A3 %>% select(worried, State)

B <- aggregate(x = B$worried,
                by = list(B$State),
                FUN = sum)

setnames(B, old =c('Group.1', 'x'), new = c('State','worried'))

#add worried data to our data frame
covid <- inner_join(covid, B, by = "State")

### Read in closure data
closure <- read_csv("CUSP_closures.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   `Initially reopen restaurants for outdoor dining only` = col_double()
## )
## i Use `spec()` for the full column specifications.
```

```r
closure$`Reopen Childcare`[closure$`Reopen Childcare`=="0"] <- "10/30/2020"

closure <- closure %>% mutate_at(vars(`Date closed K-12 public schools`,
                                      `Closed day cares`, `Closed Bars`,
                                      `Closed restaurants except take out`,
                 `Reopen Childcare`, `Reopen restaurants`, `Reopen bars`,
                 `Re-Close Bars (statewide)`), as.Date, format="%m/%d/%y")

# Daycare closure days
closure$daycare <- as.numeric(closure$`Reopen Childcare`-closure$`Closed day cares`)
closure$daycare[is.na(closure$daycare)] <- 0
closure$daycare
```

```
##  [1] 64  0  0  0  0  0  0 70  0  0  0  0  0 67  0  0  0 77  0  0 81 77 76  0  0
## [26]  0  0  0  0  0 89  0  0  0  0 66  0 52 80 77  0  0  0  0  0 75  0  0 35  0
## [51] 43
```

```r
# Bar closure days
closure$bar <- as.numeric(closure$`Reopen bars`-closure$`Closed Bars`)
closure$bar[is.na(closure$bar)] <- 0
closure$bar
```

```
##  [1]  53  51  56  60   0  93   0  91  74  80  69 101  66 102  88  72  70 105  80
## [20]  91  88   0  84  85  59  42  45  80  69  91  91   0   0   0  42  61  44  94
## [39]  79  76  47   0  60  69  43  66  65 108  69  57  57
```

```r
# Restaurant closure days
closure$restaurant <- as.numeric(closure$`Reopen restaurants`
                                 -closure$`Closed restaurants except take out`)
closure$restaurant[is.na(closure$restaurant)] <- 0
closure$restaurant
```

```
##  [1]  53  37  51  52  63  71  65  77  74  45  24  80  52  74  63  59  35  67  45
## [20]  75  74  83  84  76  34  42  45  59  49  63  91  68  97  65  41  60  30  93
## [39]  79  61  46   0  34  41   0  65  64 109  46  56  57
```

```r
# Combine with main covid data
closure <- closure %>% select(State, daycare, restaurant, bar)
covid <- full_join(covid, closure, by = "State", suffix = c("", ""))
```

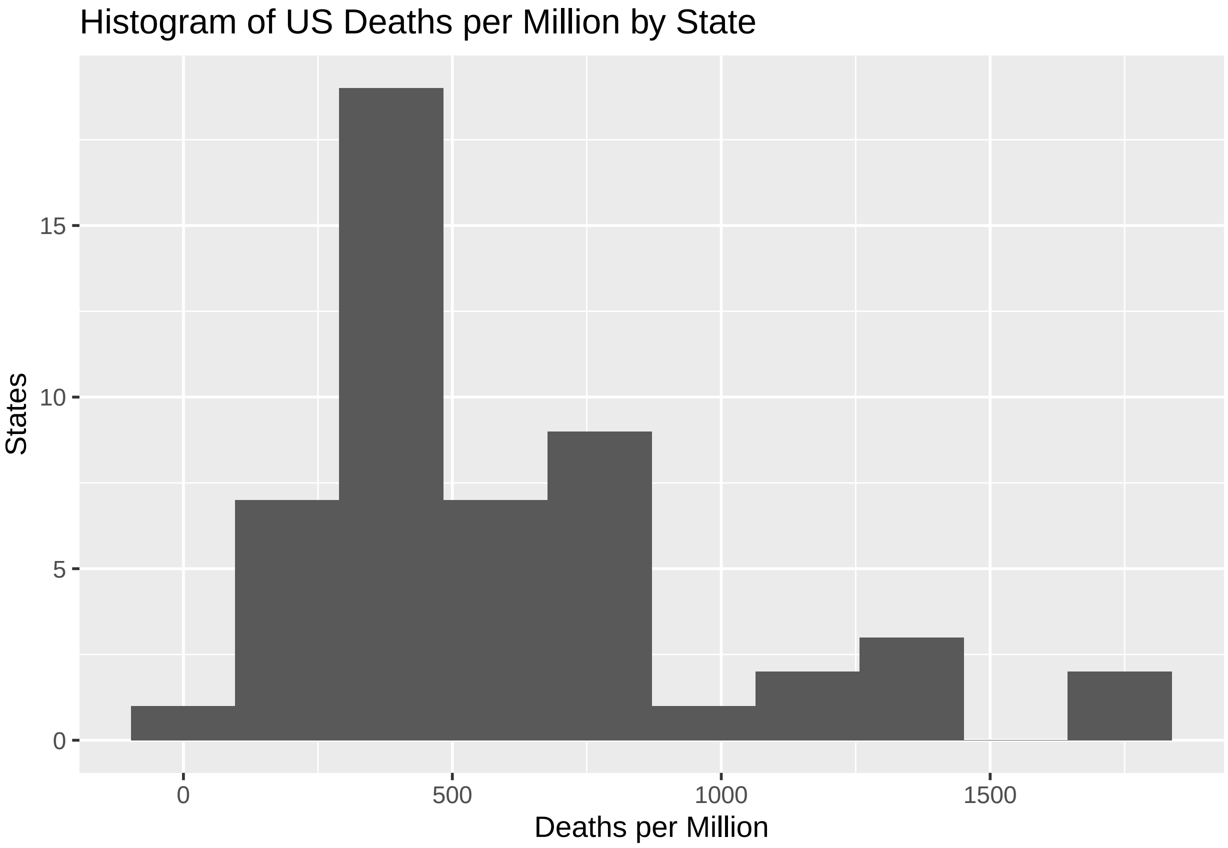
EDA for Model 1

$\hat{Death} = \beta_o + \beta_1 MaskInd + \beta_2 MaskDays$

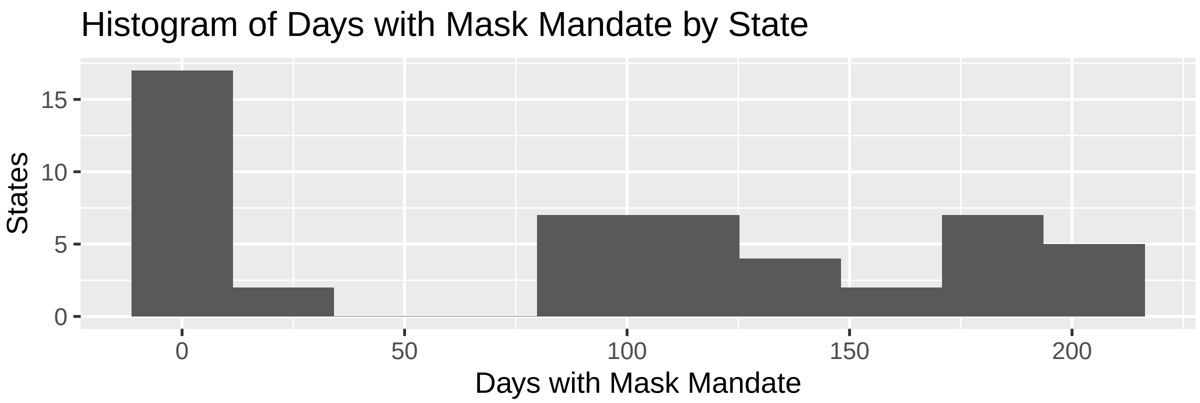
Our central model, Model 1 explores the association between the length of a state's mask mandate (in days) and a state's total number of deaths (per million people). Originally we sought to include length of stay at home mandates, as well as length of mask mandates in Model 1, believing both would have statistically and meaningfully strong associations with a state's death total. EDA suggested otherwise.

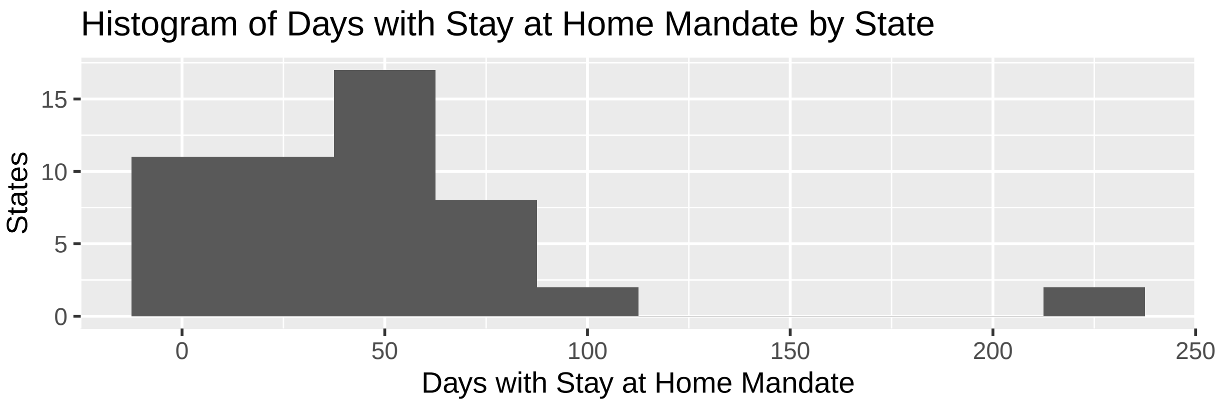
First, histograms of state-wise deaths per million and mask mandate length demonstrate those data are distributed without as much skew as is present within stay at home mandate length. Too much skew in stay at home data could suggest heteroskedasticity or residual non-normality.

```r
# Histograms of death rate, mask days, and stay at home days.
hist_dth <- covid %>%
  ggplot(aes(x=dth_per_mil)) +
  geom_histogram(bins = 10) +
  labs(title = "Histogram of US Deaths per Million by State",
       x = "Deaths per Million", y = "States")

hist_mm <- covid %>%
  ggplot(aes(x=mask_days)) +
  geom_histogram(bins = 10) +
  labs(title = "Histogram of Days with Mask Mandate by State",
       x = "Days with Mask Mandate", y = "States")

hist_stay <- covid %>%
  ggplot(aes(x=stayathome_days)) +
  geom_histogram(bins = 10) +
  labs(title = "Histogram of Days with Stay at Home Mandate by State",
       x = "Days with Stay at Home Mandate", y = "States")

hist_dth
```



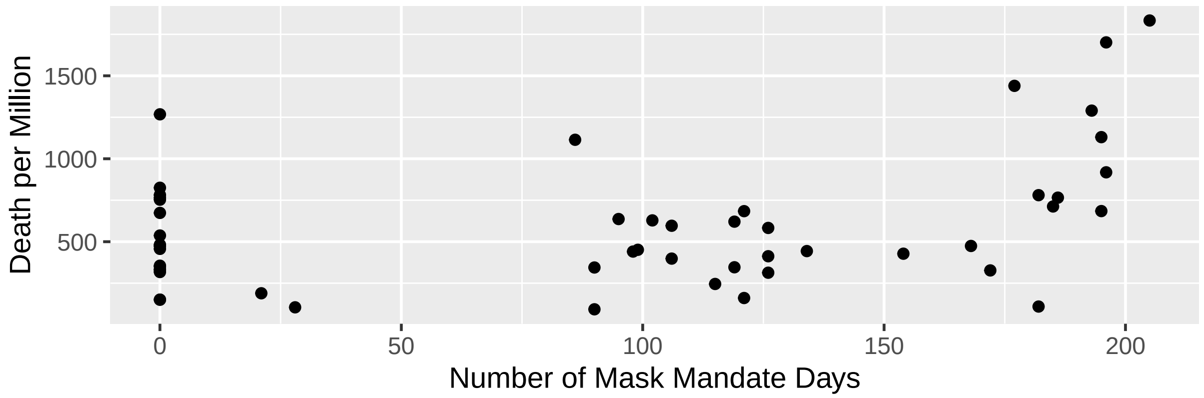
```r
hist_mm/hist_stay
```

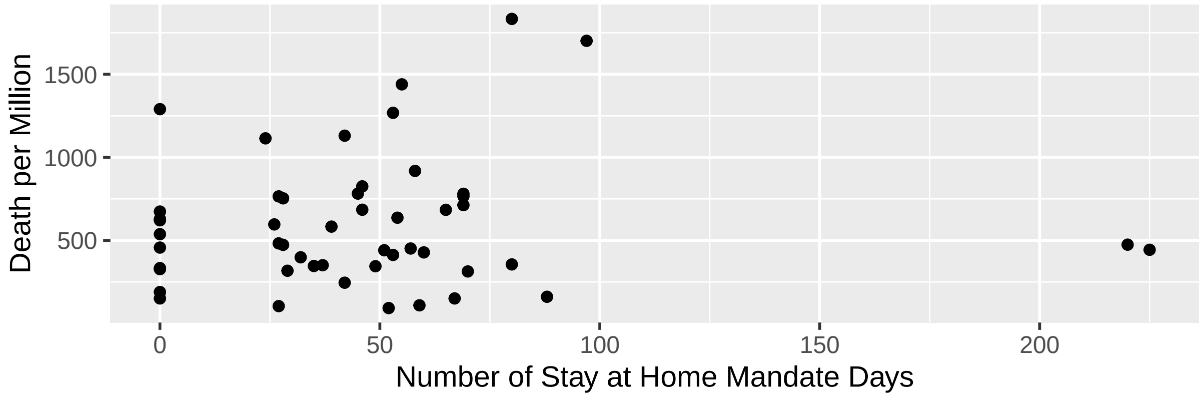




Second, deaths per million plotted against length of mask mandate revealed a much stronger visual linear relationship than that between deaths per million and stay at home length, which is corroborated by a much stronger correlation (.38 v .09). Interestingly, there is positive correlation between Mask Days and Deaths Per Million. This seems to contradict the popular intuition that mask mandates depress COVID's spread. If we were to offer a theoretical explanation for this potentially surprising correlation, it would be that states with the highest cumulative death totals suffered terribly at the outset of the outbreak. Consequently, these affected states were among the first to adopt mask mandates. Other states had yet to experience the full impact of the pandemic and only felt masks were necessary once their infection numbers crossed an unacceptable threshold, which was sometime after the initial outbreak. Therefore, we feel the positive correlation may suggest that mask mandates were a reaction to the local intensity of the disease. Further analysis is required to determine COVID's local trajectory after local governments imposed mask mandates.

```r
# Mask and stay at home mandate days against death per million
plot_mask_td <- covid %>%
  ggplot(aes(x = mask_days, y = dth_per_mil))+
  geom_point() +
  labs(x = "Number of Mask Mandate Days", y = "Death per Million")

plot_stay_td <- covid %>%
  ggplot(aes(x = stayathome_days, y = dth_per_mil))+
  geom_point() +
  labs(x = "Number of Stay at Home Mandate Days", y = "Death per Million")

plot_mask_td/plot_stay_td
```





```r
# Correlation b/w mask days and death rate, and stay at home days and death rate
cor(covid$dth_per_mil,covid$mask_days)
```

```
## [1] 0.3792884
```

```r
cor(covid$dth_per_mil, covid$stayathome_days)
```

```
## [1] 0.08537629
```

Thirdly, a summary of several sample linear models suggest that stay at home length offers little to no value, and actually increases the p-value of the model's F-statistic. For all these reasons we chose to drop stay at home length from Model 1 and include it in Model 3.

We used a dummy or indicator variable in Model 1 to account for the fact that some states have not imposed a mask mandate. We found that incorporating said indicator variable permitted the data and our model to most closely meet all Classic Linear Model assumptions (more discussion later in the report). Furthermore, incorporating the indicator variable led to a stronger model (F-statistic p-value of .0002 versus .007, r^2 of .299 versus .144).

```r
# Model for deaths per million with indicator mask mandate
model_1 = lm(dth_per_mil~ind_mask+mask_days,covid)

#without indicator and using stay at home days
model_1.1 = lm(dth_per_mil~mask_days,covid)
model_1.2 = lm(dth_per_mil~ind_stay+stayathome_days, covid)

#second model with indicator variables for mask and stay at home
model_dpm_3 <- lm(dth_per_mil~ind_mask+mask_days+ind_stay+stayathome_days, covid)

#summary of all models
```

```r
stargazer(model_1, model_1.1,
          type = "text", report=('vc*p'),
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Mask and Stay at Home Models")
```

```
##
## Mask and Stay at Home Models
## ===============================================================
##                                  Dependent variable:
##                      -------------------------------------------
##                                      dth_per_mil
##                               (1)                   (2)
## ---------------------------------------------------------------
## ind_mask                  -628.746**
##                            p = 0.003
##
## mask_days                  5.368***              1.978**
##                            p = 0.0001            p = 0.007
##
## Constant                  529.603***            417.206***
##                           p = 0.00000           p = 0.00001
##
## ---------------------------------------------------------------
## Observations                  51                    51
## R2                          0.299                 0.144
## Adjusted R2                 0.270                 0.126
## Residual Std. Error   336.527 (df = 48)     368.045 (df = 49)
## F Statistic         10.228*** (df = 2; 48) 8.234** (df = 1; 49)
## ===============================================================
## Note:                             *p<0.05; **p<0.01; ***p<0.001
```

```r
stargazer(model_1.2, model_dpm_3,
          type = "text", report=('vc*p'),
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Mask and Stay at Home Models")
```

```
##
## Mask and Stay at Home Models
## ===========================================================
##                                Dependent variable:
##                      ---------------------------------------
##                                    dth_per_mil
##                             (1)                 (2)
## -----------------------------------------------------------
## ind_mask                                    -651.647**
##                                             p = 0.002
##
## mask_days                                    5.658***
##                                             p = 0.0001
##
## ind_stay                  105.936            128.402
##                          p = 0.519          p = 0.362
##
## stayathome_days            0.207              -1.487
##                          p = 0.893           p = 0.281
##
## Constant                503.364***          487.361***
##                          p = 0.0002          p = 0.0001
##
## -----------------------------------------------------------
## Observations                51                  51
## R2                        0.016               0.319
## Adjusted R2               -0.025              0.260
## Residual Std. Error 398.672 (df = 48)   338.770 (df = 46)
## F Statistic         0.389 (df = 2; 48) 5.388** (df = 4; 46)
## ===========================================================
## Note:                         *p<0.05; **p<0.01; ***p<0.001
```

EDA for Model 2

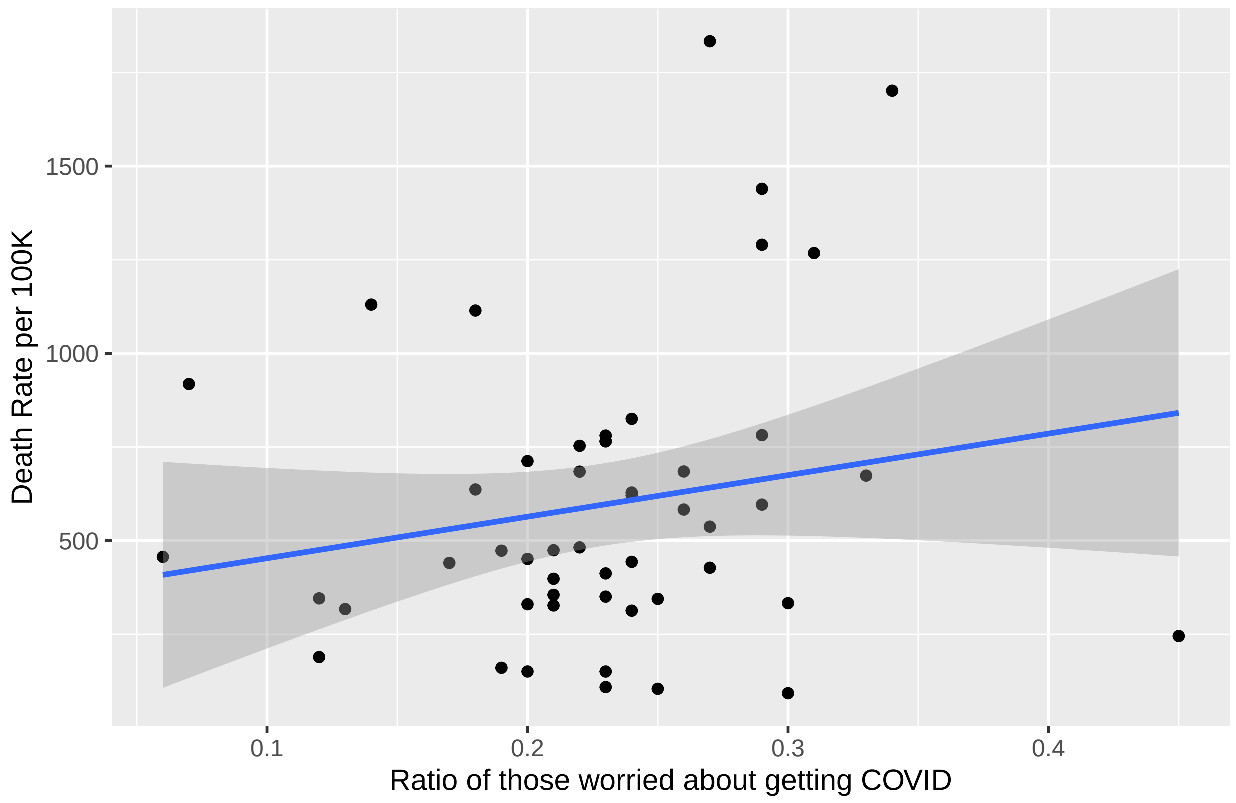
$\hat{Death} = \beta_o + \beta_1 MaskInd + \beta_2 MaskDays + \beta_3 PubOpinion$

Model 2 incorporated a PubOpinion variable derived from ANES 2020 survey data. To derive said variable, we used responses to the question “How worried are you personally about getting the Coronavirus?”, which was captured by the [COVID-1] survey variable. Survey participants answered the question using a 5 step likert scale from “Extremely Worried” to “Not Very Worried”. To transform [COVID 1] from ordinal to ratio data, we compiled responses by state, and then divided responses on the “Worried” side of the likert scale by total responses. This gave us a ratio of respondents, bucketed by states who were worried. In an exploratory scatter plot, we see a slight positive relationship between the ratio of people being worried about COVID and death rate per million, by state. It is important to note that the ANES survey data was “collected between April 10, 2020 and April 18, 2020.” This represents approximately six months between survey responses and COVID data used to support our analysis. We acknowledge the potential that some opinions have changed within that timeframe. However, there is not more reliable data available that provides insight into peoples’ concerns about getting Coronavirus.

Some concern was raised by colleagues about the reverse causality and biases stemming from including peoples’ opinion. This is a concern that we acknowledge. The question was opinion based and subjective. There is some chance for varying degree of interpretation and variability of responses in the data. Responses can be influenced by government intervention, or lack thereof, misinformation, genuine concern, predisposition and exposure risk to name a few. Any one of these factors could affect the response provided, and these factors are also influenced by the length of time between the ANES survey and release of the COVID data. However, there is no more reliable metric that proved to be reputable or easy to obtain compared to the ANES data. The data is not without its flaws, and despite the small rˆ2 movement and performance in the model - discussed more below, including the insight from the survey was helpful in providing a comprehensive understanding to support a response to our research question.

```r
# relationship between those worried about getting COVID and deaths per million
plot_dth_worried <- ggplot(data = covid, aes(x = worried, y = dth_per_mil)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  annotate("text", x=.15, y=1500, label= " ") +
  labs(title = "", x = "Ratio of those worried about getting COVID",
       y = "Death Rate per 100K")
plot_dth_worried
```

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



```
#model 2
model_dpm_worried = lm(dth_per_mil ~ mask_days + ind_mask + worried, covid)
summary(model_dpm_worried)
```

```
##
## Call:
## lm(formula = dth_per_mil ~ mask_days + ind_mask + worried, data = covid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -763.56 -183.53  -28.97  182.90  805.58
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  335.392    180.848   1.855  0.06994 .
## mask_days      5.255      1.215   4.325 7.89e-05 ***
## ind_mask    -615.855    192.455  -3.200  0.00246 **
## worried      855.334    711.585   1.202  0.23538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 335 on 47 degrees of freedom
## Multiple R-squared:  0.3197, Adjusted R-squared:  0.2763
## F-statistic: 7.364 on 3 and 47 DF,  p-value: 0.0003828
```

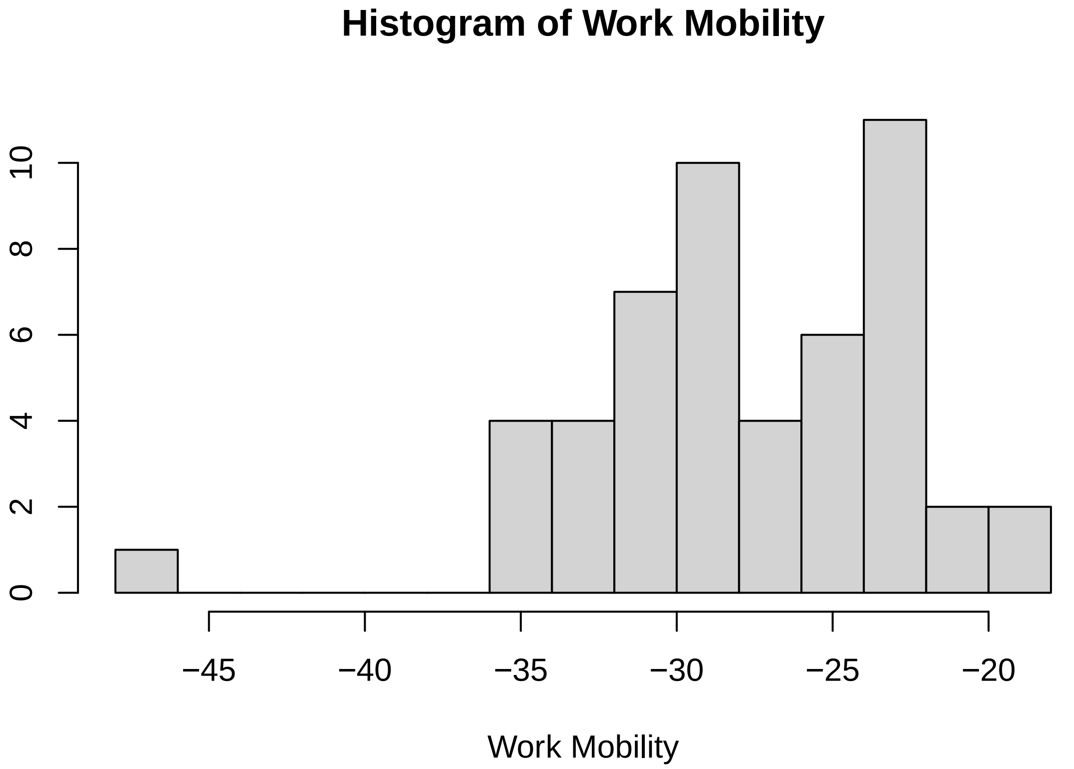
EDA for Model 3

$$\hat{Death} = \beta_o + \beta_1 MaskInd + \beta_2 MaskDays + \beta_3 StayHomeDays + \beta_4 StayHomeInd + \beta_5 PubOpinion + \beta_6 MobilityHome + \beta_7 DaycareClosure + \beta_8 RestaurantClosure + \beta_9 BarClosure$$

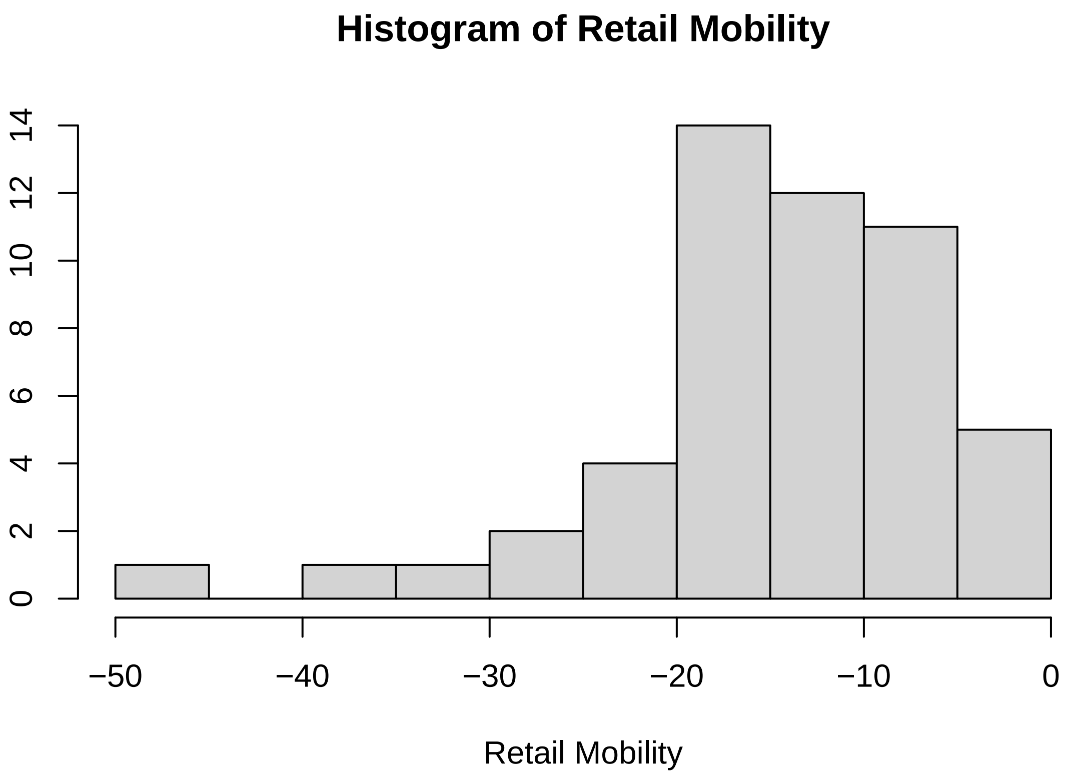
In Model 3, we incorporated all variables from Model 2 as well as additional secondary mitigation strategies including day care closures (public school closure could have offered similar or better insights, but we were unable to find reopening dates in our data), bar closures, restaurant closures, as well as mobility data. Secondary mitigation strategies were derived identically to mask and stay at home mandates discussed for Model 1. The mobility variables are the percentage change in statewide visitors to workplaces and retail services; and, percentage change of individuals' time spent within the home, all compared to a baseline period of January 3rd to February 6, 2020, as provided by Google's Location Services database.

First, we explored the mobility data set. By plotting a simple histogram for each variable, we see that retail mobility data are skewed slightly to the left. They are both always negative. Meanwhile, home mobility data skews slightly to the right and is always positive.

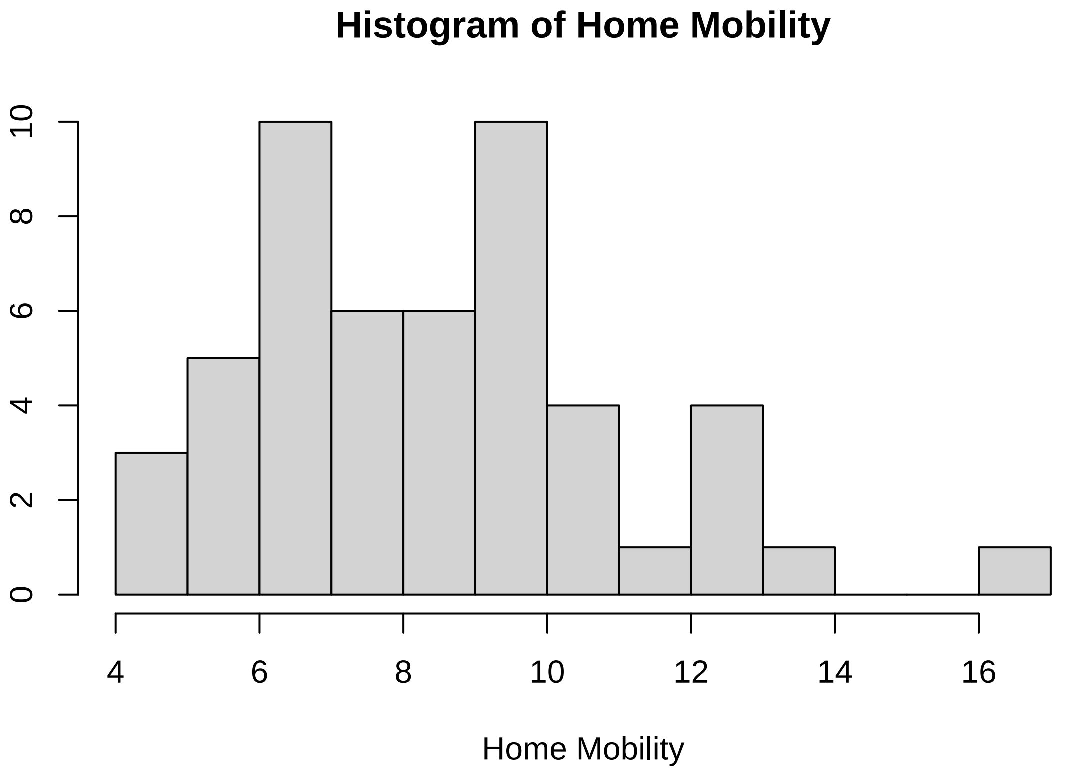
```
hist(covid$mob_data_wk, breaks = 10,
     xlab = "Work Mobility", ylab = "",
     main = "Histogram of Work Mobility")
```



```
hist(covid$mob_data_ret, breaks = 10,
     xlab = "Retail Mobility", ylab = "",
     main = "Histogram of Retail Mobility")
```

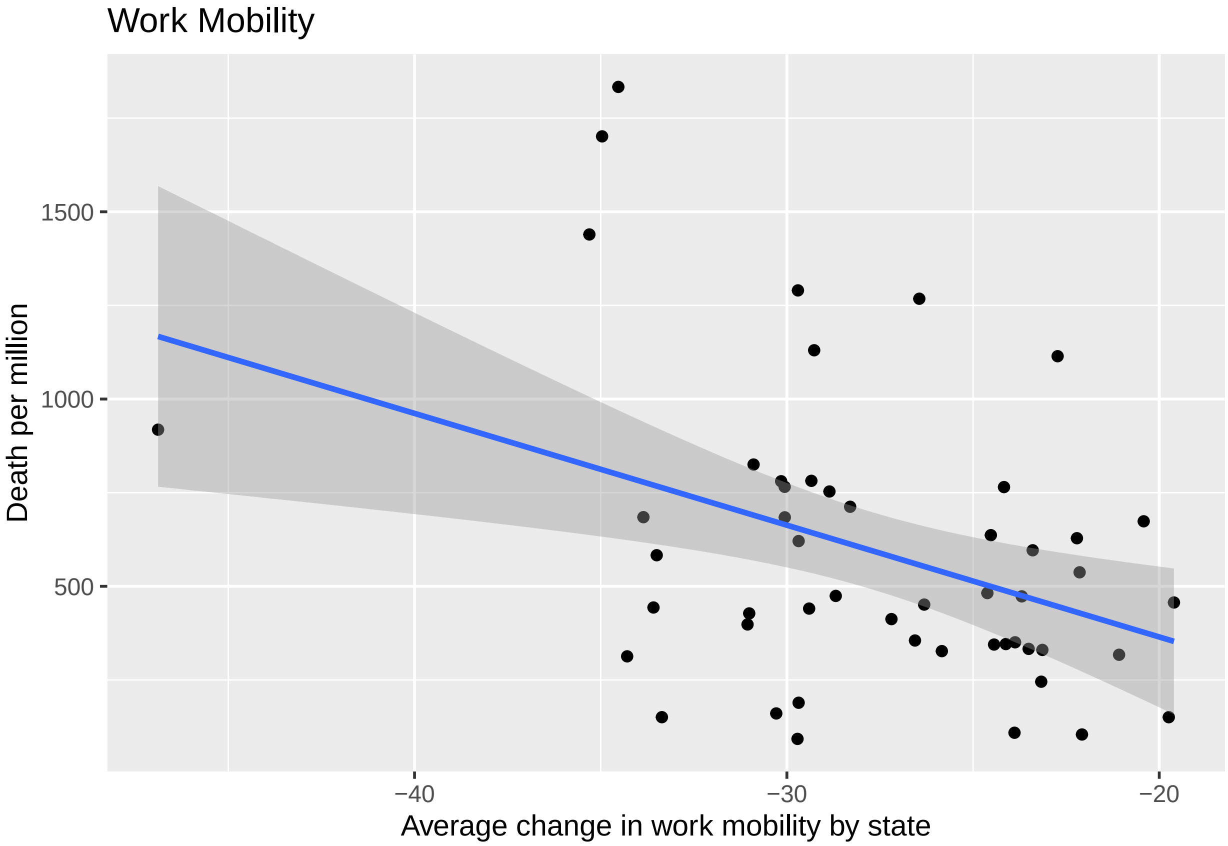


```r
hist(covid$mob_data_hm, breaks = 10,
     xlab = "Home Mobility", ylab = "",
     main = "Histogram of Home Mobility")
```



Next we plotted mean change in work, retail and home mobility data against death per million by state, respectively, and calculated correlation between mobility data and death rate. Home mobility data shows the strongest visual correlation, which is confirmed with a correlation of 0.43. The direction of correlation indicates that the more people stayed away from work and retail locations compared to the baseline period in a state, the higher the death rate; and the more people stayed at home compared to the base line period, the higher the death rate. This is possibly due to the fact that states with the most cases and deaths at the beginning of the pandemic mandated or encouraged working from home and and the shutdown of non-essential retails shops.

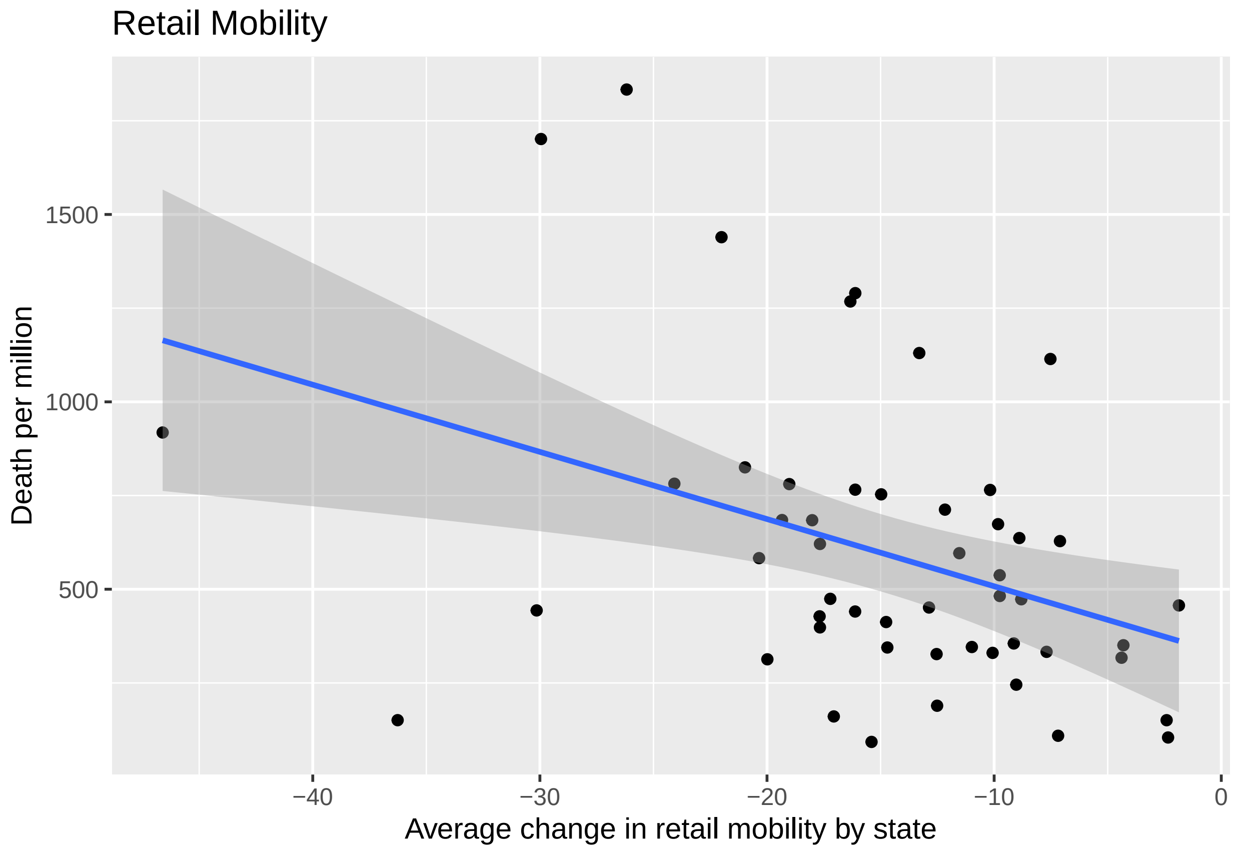
```
# Work
plot_mob_wk <- ggplot(data = covid, aes(x = mob_data_wk, y = dth_per_mil)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(title = "Work Mobility", x = "Average change in work mobility by state",
       y = "Death per million")
plot_mob_wk
```



```
cor(covid$mob_data_wk,covid$dth_per_mil)
```

```
## [1] -0.389334
```

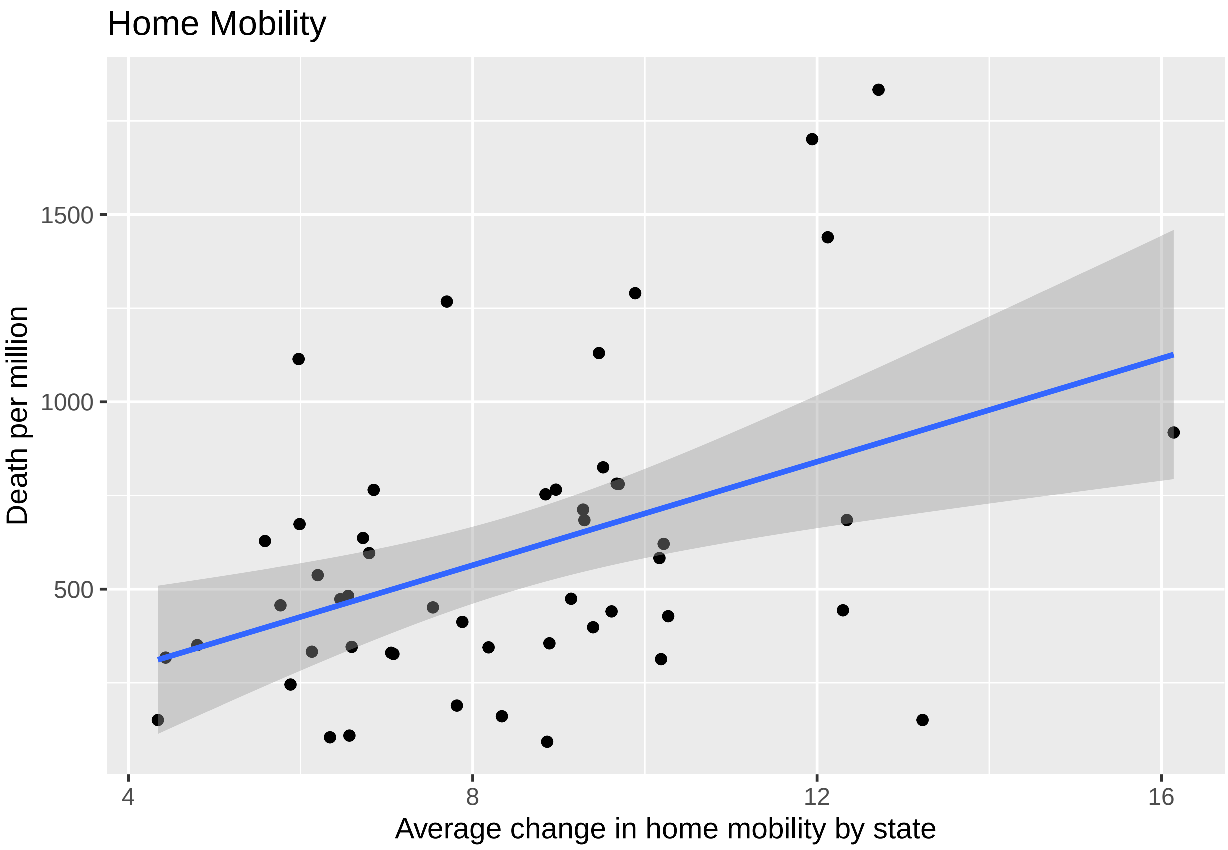
```
# Retail
plot_mob_ret <- ggplot(data = covid, aes(x = mob_data_ret, y = dth_per_mil)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(title = "Retail Mobility", x = "Average change in retail mobility by state",
       y = "Death per million")
plot_mob_ret
```



```
cor(covid$mob_data_ret,covid$dth_per_mil)
```

```
## [1] -0.3869987
```

```
# Home
plot_mob_hm <- ggplot(data = covid, aes(x = mob_data_hm, y = dth_per_mil)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(title = "Home Mobility", x = "Average change in home mobility by state",
       y = "Death per million")
plot_mob_hm
```



```
cor(covid$mob_data_hm,covid$dth_per_mil)
```

```
## [1] 0.4329346
```

As the final step of exploring the mobility data, we ran a linear regression model of each mobility data again death rate. In each of the three models, the p-value for the mobility variable is statistically significant at the 0.01 level. However, we then ran a model that includes both retail and work mobility, and a model that includes all three mobility data. In these models, none of the mobility variables is statistically significant, with decreases or no improvement in adjusted r-squared. This indicates some strong collinearity between the mobility variables, which makes intuitive sense - as people go to work less, they spend more time at home. From the models that only include one mobility variable each, home mobility has the smallest p-value and the largest adjusted r-squared. Therefore, we will only include home mobility as a secondary variable in Model 3.

```
model_mob_wk <- lm(dth_per_mil~mob_data_wk,covid)
model_mob_ret <- lm(dth_per_mil~mob_data_ret,covid)
model_mob_hm <- lm(dth_per_mil~mob_data_hm,covid)
model_mob_ret_wk <- lm(dth_per_mil~mob_data_ret+mob_data_wk,covid)
model_mob_ret_wk_hm <- lm(dth_per_mil~mob_data_ret+mob_data_wk+mob_data_hm,covid)

stargazer(model_mob_wk, model_mob_ret, model_mob_hm,
          type = "text", omit.stat = "f", report=('vc*p'),
          star.cutoffs = c(0.05, 0.01, 0.001), title = "Mobility Models - Single Variable")
```

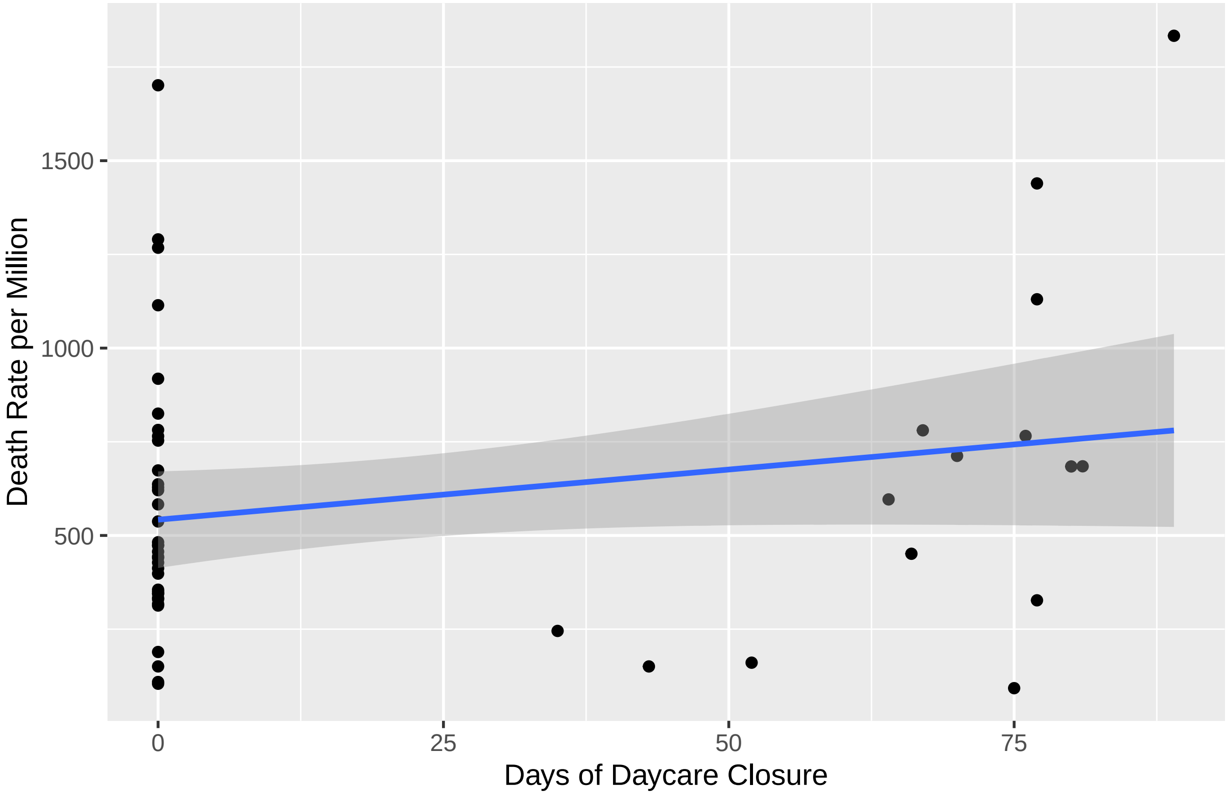
```
##
## Mobility Models - Single Variable
## ================================================================
##                                       Dependent variable:
##                                 --------------------------------
##                                          dth_per_mil
##                                   (1)        (2)        (3)
## -------------------------------------------------------------
## mob_data_wk                   -29.844**
##                               p = 0.005
##
## mob_data_ret                             -17.929**
##                                          p = 0.006
##
## mob_data_hm                                         69.056**
##                                                     p = 0.002
##
## Constant                       -232.073  328.636**   11.488
##                               p = 0.419  p = 0.003  p = 0.950
##
## -------------------------------------------------------------
## Observations                      51         51         51
## R2                              0.152      0.150      0.187
## Adjusted R2                     0.134      0.132      0.171
## Residual Std. Error (df = 49)  366.382    366.773    358.557
## =============================================================
## Note:                           *p<0.05; **p<0.01; ***p<0.001
```

```
stargazer(model_mob_ret_wk,model_mob_ret_wk_hm,
          type = "text", omit.stat = "f", report=('vc*p'),
          star.cutoffs = c(0.05, 0.01, 0.001), title = "Mobility Models - Multiple Variables")
```

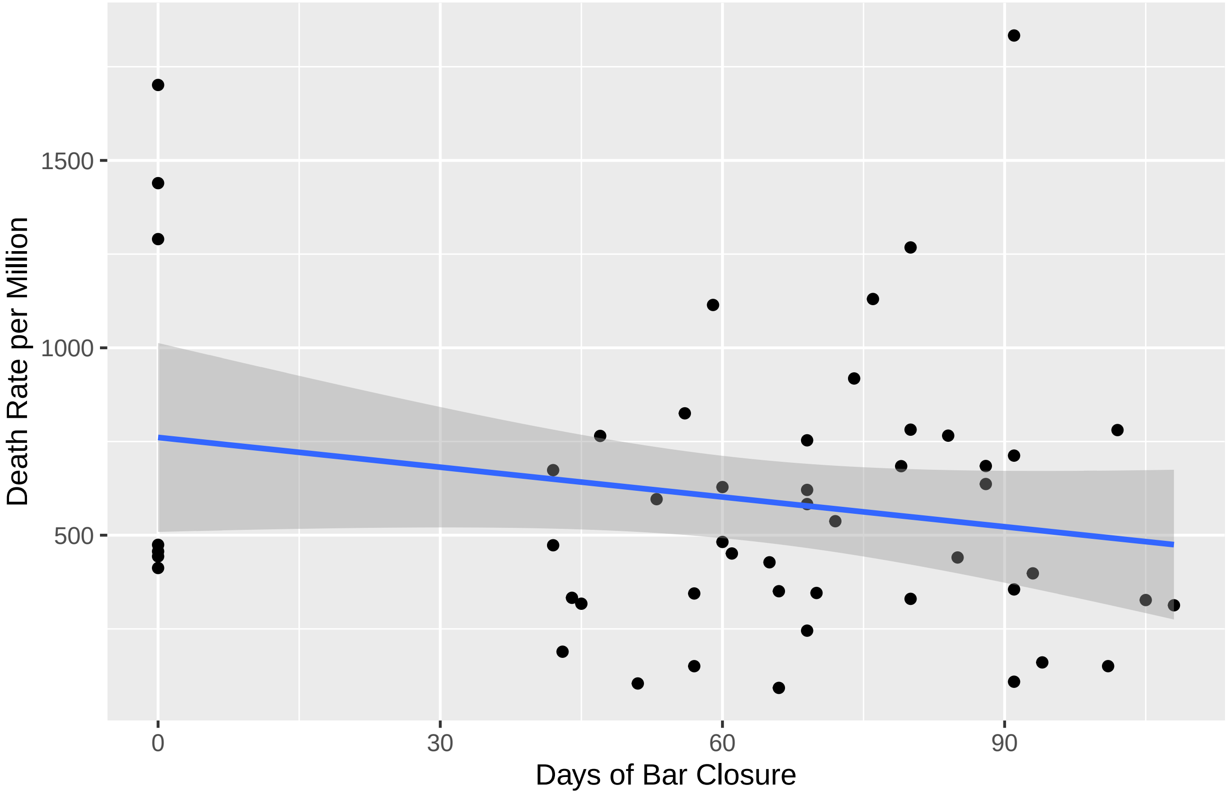
```
##
## Mobility Models - Multiple Variables
## =======================================================
##                              Dependent variable:
##                      -----------------------------------
##                                  dth_per_mil
##                            (1)               (2)
## -------------------------------------------------------
## mob_data_ret             -8.789             2.892
##                         p = 0.560         p = 0.866
##
## mob_data_wk              -16.579           14.442
##                         p = 0.507         p = 0.663
##
## mob_data_hm                                106.785
##                                           p = 0.162
##
## Constant                  4.902            135.984
##                         p = 0.993         p = 0.787
##
## -------------------------------------------------------
## Observations               51                51
## R2                        0.158             0.192
## Adjusted R2               0.123             0.141
## Residual Std. Error 368.857 (df = 48) 365.002 (df = 47)
## =======================================================
## Note:                     *p<0.05; **p<0.01; ***p<0.001
```

Finally, we explored daycare, bar, and restaurant closure variables. For each variable, we created a scatterplot against death per million. We see that many states have not had a mandated daycare closure. There's a slight positive relationship between days of daycare closure and death per million. The number of days of restaurant closure also has a similar relationship with death per million, which seems counter-intuitive. However, the number of days of bar closure does have a negative relationship with death per million - longer bar closures in a state is associated with less deaths. We conducted linear regression analysis for each of these three variables separately, as well as altogether. While only the bar closure variable has a statistically significant p-value as a part of the model where all three variables are included, r-squared is improved quite a bit from the model with only bar closure included (0.06) to the model with all three closure variables (0.17). Therefore, for the final model, we will include all three closure variables and see how they are impacted by the other variables in our model.

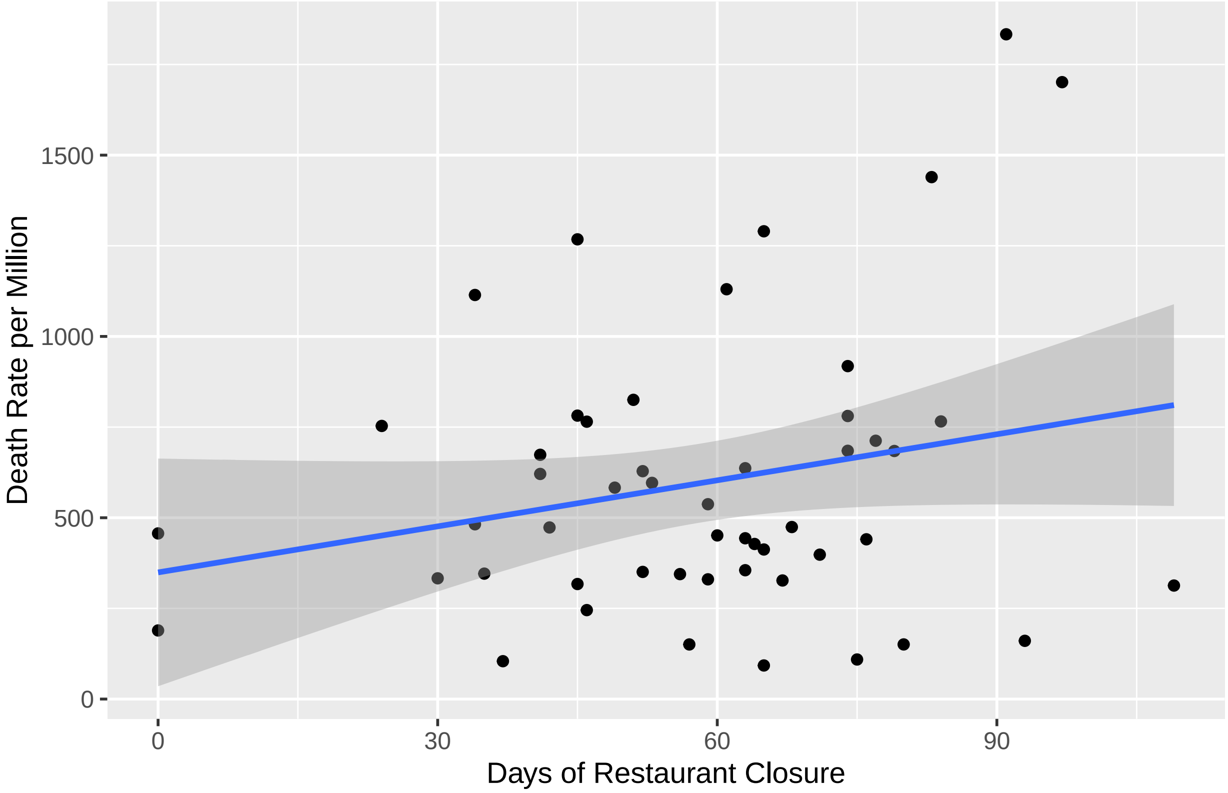
```r
# plot relationships between closure days and death per million
plot_daycare <- ggplot(data = covid, aes(x = daycare, y = dth_per_mil)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(title = "", x = "Days of Daycare Closure", y = "Death Rate per Million")

plot_bar <- ggplot(data = covid, aes(x = bar, y = dth_per_mil)) +
  geom_point() +
  geom_smooth(method = 'lm')+
  labs(title = "", x = "Days of Bar Closure", y = "Death Rate per Million")

plot_restaurant <- ggplot(data = covid, aes(x = restaurant, y = dth_per_mil)) +
  geom_point() +
  geom_smooth(method = 'lm')+
  labs(title = "", x = "Days of Restaurant Closure", y = "Death Rate per Million")

plot_daycare
```



```
plot_bar
```



```
plot_restaurant
```



```r
model_daycare <- lm(dth_per_mil~daycare, covid)
model_bar <- lm(dth_per_mil~bar, covid)
model_restaurant <- lm(dth_per_mil~restaurant, covid)
model_dbr <- lm(dth_per_mil~daycare + bar + restaurant, covid)

stargazer(model_daycare, model_bar, model_restaurant, model_dbr,
          type = "text", report=('vc*p'),
          star.cutoffs = c(0.05, 0.01, 0.001), title = "Daycare, bar, restaurant closure models")
```

```
##
## Daycare, bar, restaurant closure models
## ================================================================================================
##                                                  Dependent variable:
##                      ----------------------------------------------------------------------------
##                                                      dth_per_mil
##                            (1)                (2)                (3)                 (4)
## ------------------------------------------------------------------------------------------------
## daycare                   2.676                                                     2.522
##                         p = 0.119                                                 p = 0.165
##
## bar                                          -2.648                               -4.335*
##                                            p = 0.152                              p = 0.023
##
## restaurant                                                      4.233               4.495
##                                                               p = 0.099           p = 0.103
##
## Constant                542.128***         760.852***          349.257*           552.782**
##                         p = 0.000         p = 0.00000         p = 0.030           p = 0.002
##
## ------------------------------------------------------------------------------------------------
## Observations                51                 51                 51                  51
## R2                        0.049              0.042              0.055               0.172
## Adjusted R2               0.029              0.022              0.035               0.119
## Residual Std. Error 387.923 (df = 49)  389.403 (df = 49)  386.730 (df = 49)   369.499 (df = 47)
## F Statistic         2.518 (df = 1; 49) 2.127 (df = 1; 49) 2.837 (df = 1; 49) 3.261* (df = 3; 47)
## ================================================================================================
## Note:                                                              *p<0.05; **p<0.01; ***p<0.001
```

Finalized Models and Results

```r
model1 = lm(dth_per_mil ~ mask_days + ind_mask, covid)
model2 = lm(dth_per_mil ~ mask_days + ind_mask + worried, covid)
model3 = lm(dth_per_mil ~ mask_days + ind_mask + worried + stayathome_days + ind_stay +
              mob_data_hm + daycare  + restaurant + bar, covid)

se.model1 = coeftest(model1, vcov = vcovHC)[ , "Std. Error"]
se.model2 = coeftest(model2, vcov = vcovHC)[ , "Std. Error"]
se.model3 = coeftest(model3, vcov = vcovHC)[ , "Std. Error"]

stargazer(model1, model2, model3, type = "text",
          se = list(se.model3),
          star.cutoffs = c(0.05, 0.01, 0.001), title = "Final Models")
```

```
##
```

```
## Final Models
## ======================================================================================
##                                            Dependent variable:
##                      ------------------------------------------------------------------
##                                                dth_per_mil
##                              (1)                    (2)                   (3)
## --------------------------------------------------------------------------------------
## mask_days                  5.368**               5.255***               4.633**
##                            (1.778)                (1.215)               (1.547)
##
## ind_mask                  -628.746**            -615.855**            -571.045**
##                           (191.308)              (192.455)             (192.116)
##
## worried                                           855.334               888.237
##                                                  (711.585)             (695.437)
##
## stayathome_days                                                         -3.167*
##                                                                         (1.450)
##
## ind_stay                                                                216.408
##                                                                        (136.035)
##
## mob_data_hm                                                             56.928*
##                                                                        (23.825)
##
## daycare                                                                  0.854
##                                                                         (1.685)
##
## restaurant                                                              -1.823
##                                                                         (3.323)
##
## bar                                                                     -4.097*
##                                                                         (1.811)
##
## Constant                   529.603                335.392               194.338
##                           (283.311)              (180.848)             (255.303)
##
## --------------------------------------------------------------------------------------
## Observations                  51                     51                    51
## R2                           0.299                  0.320                 0.491
## Adjusted R2                  0.270                  0.276                 0.379
## Residual Std. Error   336.527 (df = 48)      334.979 (df = 47)     310.292 (df = 41)
## F Statistic         10.228*** (df = 2; 48) 7.364*** (df = 3; 47) 4.391*** (df = 9; 41)
## ======================================================================================
## Note:                                                    *p<0.05; **p<0.01; ***p<0.001
```

Our final version of Model 1 is as follows:

$\hat{Death} = 530 - 629 * I(Mask) + 5.37 * MaskDays$

The model indicates that no days of mask mandates would result in 530 deaths per million, simply having a mask mandate is associated with 630 less deaths per million, while one day of mandated mask usage is associated with 5.37 more deaths per million. All coefficients are statistically significant at at least the 0.01 level, with the coefficient for mask indicator being significant at the 0.001 level. Readers are referred to the "EDA: Model 1" paragraph for a discussion on the somewhat surprising positive correlation between mask mandate length and deaths per million.

Our final version of Model 2 is as follows:

$\hat{Death} = 335.39 - 615.86 MaskInd + 5.26 * MaskDays + 855.334 PubOpinion$

Model 2 incorporated a PubOpinion variable derived from ANES 2020 survey data. To derive said variable, we used responses to the question "How worried are you personally about getting the Coronavirus?", which was captured by the [COVID-1] survey variable. Survey participants answered the question using a 5 step likert scale from "Extremely Worried" to "Not Very Worried". To transform [COVID-1] from ordinal to ratio data, we compiled responses by state, and then divided responses on the "Worried" side of the likert scale by total responses. This gave us a ratio of respondents, bucketed by states who were worried. We use β_3 to quantify the association between ratio of worried respondents and death rate, by state.

In Model 2, the coefficients for mask indicator and days with mandated mask usage have similar magnitude as in model 1 and continue to be statistically significant at a 0.01 level. PubOpinion variable has a coefficient of 855.33, meaning an increase in the ratio of those worried about getting Coronavirus by .01 is associated with an increase of 855 deaths per million. However, this coefficient is not statistically significant.

Our final version of Model 3 is as follows:

$\hat{Death} = 194.34 - 571.05 MaskInd + 4.63 MaskDays - 3.17 StayHomeDays + 216.41 StayHomeInd + 888.24 PubOpinion + 56.93 MobilityHome + .85 DaycareClosure - 1.82 RestaurantClosure - 4.10 BarClosure$

In this model, the coefficients for mask indicator and days with mandated mask usage have around the same magnitude as in model 1 and continue to be statistically significant at a 0.01 level. Shelter at home indicator variable has a coefficient of 216.41, meaning that having a shelter at home mandate is associated with an increase of 216.41 deaths per million. However, this coefficient is not significant at the 0.05 level. An increase in one day of the shelter at home mandate is associated with a decrease of 2.16 deaths per million. Home mobility data measures the percentage change in duration spent in residential areas compared to a baseline period. According to our model, an increase of one percent of time spent at home is associate with an incraese in death by 56.93 per million. This coefficient is statistically significant at the 0.05 level. An increase in one day of restaurant closure, and bar closure are associated with decreases in 0.64 and 4.10 deaths per million, respectively, while an increase in one day of daycare closure is associated with an increase of .85 deaths per million. The coefficient for bar closures is statistically significant at the 0.05 level, while the coefficients for daycare closure and restaurant closure are not.

Omitted Variables

Although our team's model focused on a descriptive question, we believe a discussion of omitted variables is helpful to further the understanding of the question the group is seeking to answer. There are several variables that would bias the coefficients, for the purposes of our discussion, we will focus on the top five omitted variables that affect our models.

Omitted Variables - Model 1

In model one, we would consider a variable introducing omitted variable bias to be "Personal Compliance". This refers to the choice citizens make to personally comply with state ordered mask mandates. Due to the wildly divisive stances against mask mandates throughout and even within states, it is important to acknowledge peoples' willingness to comply with stated mask mandates. Generally, we would expect to see a positive relationship between mask mandates and compliance with those mandates as a consequence of other businesses being forced to adopt the states' mask mandates. However, we acknowledge that people may not agree with the mandates or rationale behind it and choose to ignore or even rebel against the mandate. For the purposes of this analysis, we will assume that a large portion of the population supports the mask mandates and personal compliance is high. We note that there is no proxy of this variable as the neither the ANES data nor the data provided explicitly asks about how well someone complies with a mask order.

Should this data be available in the ANES data, we would maintain another level of omitted variable bias (see Model 2).

Estimated Model: $\hat{Death} = \tilde{\beta}_o + \tilde{\beta}_1 MaskInd + \tilde{\beta}_2 MaskDays$

True Model: $\hat{Death} = \beta_o^* + \beta_1 MaskInd^* + \beta_2 MaskDays^* \cdot \beta_3 PersonalCompliance$

Direction of bias: Given the thought above, we can safely reason if most of the population complies with mask mandates, we would anticipate a negative direction of the bias. Meaning, as we observe mask mandates, we would anticipate high levels of personal compliance, which would drive the bias toward zero. Practically speaking, higher personal compliance would lead to fewer deaths per million than already taken computed by model 1. We would reason that given the bias toward zero and the potential for a strong negative relationship with the number of deaths.

Omitted Variables - Model 2

The second variable that could be causing omitted variable bias would be a variable called "Change in Opinion". What this variable refers to is the fact that the ANES survey was "collected between April 10, 2020 and April 18, 2020." This is in contrast to covid data set, which was compiled on October 30, 2020. Given the nearly six month discrepancy between the ANES survey and the COVID data coupled with the fact we are looking at a variable that measures public opinion in the form of "How worried are you personally about getting the Coronavirus (COVID-19)?" we would want to measure any changes in that opinion from April to October. This variable would effectively be a restatement of the question in ANES survey but asked in late October. There is no proxy variable readily available from either the data provided or the ANES data.

Estimated Model: $\hat{Death} = \tilde{\beta}_o + \tilde{\beta}_1 MaskInd + \tilde{\beta}_2 MaskDays + \tilde{\beta}_3 PubOpinion$

True Model: $\hat{Death} = \beta_o^* + \beta_1 MaskInd^* + \beta_2 MaskDays^* + \beta_3 PubOpinion^* \cdot \beta_4 ChangeInOpinion$

Direction of bias: By having an updated variable that reflect changes in opinion and it were to follow the same trend as the death rate, it would be safe to estimate a negative direction bias direction toward zero. If we had a variable that could reflect changes in opinion since April, and those changes reinforced what was observed in the ANES data, we would reason that if people still felt worried about personally getting COVID, there would be fewer deaths per million. Bias size would be fairly small even if there were strong changes in opinion compared to the ANES survey. This is due to the fact that the public opinion variable is not statistically significant. Practically speaking, we observed that most people are at least moderately worried, any changes of that opinion will move our bias toward zero but have a small effect due to the lack of statistical significance of the public opinion variable in Model 2.

Omitted Variables - Model 3

The third variable that could be causing omitted variable bias would be a variable called "Business Mask Compliance". This refers to the choice businesses make to comply with state ordered mask mandates. Just as with personal compliance, we have businesses stand with or in defiance of state mandated mask orders.

Generally, we would expect to see a positive relationship between mask mandates and businesses' compliance with those mandates as a consequence of wanting to reopen and remain open. Despite the clear relationship between following mask mandates and a business staying open, there is much division between business owners and state governments. For the purposes of this analysis, we will assume that a large portion of business owners support the mask mandates and business compliance is high. We note that there is no proxy of this variable as the neither the ANES data nor the data provided explicitly ask about how well a business complies with a mask order. Should this data be available in the ANES data, we would maintain another level of omitted variable bias (see model 2). We would reason that given the bias toward zero and the potential for a strong negative relationship with the number of deaths.

Estimated Model: $\hat{Death} = \tilde{\beta}_o + \tilde{\beta}_1 MaskInd + \tilde{\beta}_2 MaskDays + \tilde{\beta}_3 StayHomeDays + \tilde{\beta}_4 StayHomeInd + \tilde{\beta}_5 PubOpinion + \tilde{\beta}_6 MobilityHome + \tilde{\beta}_7 DaycareClosure + \tilde{\beta}_8 RestaurantClosure + \tilde{\beta}_9 BarClosure$

True Model: $\hat{Death} = \beta_o^* + \beta_1 MaskInd^* + \beta_2 MaskDays^* + \beta_3 StayHomeDays^* + \beta_4 StayHomeInd^* + \beta_5 PubOpinion^* + \beta_6 MobilityHome^* + \beta_7 DaycareClosure^* + \beta_8 RestaurantClosure^* + \beta_9 BarClosure^* \cdot \beta_{10} BusinessMaskCompliance$

Direction of bias: Given the thought above, we can safely reason if most businesses comply with mask mandates, we would anticipate a negative direction of the bias. Meaning, as we observe mask mandates, we would anticipate high levels of business compliance, which would drive the bias toward zero. Practically speaking, higher business compliance would lead to fewer deaths per million by virtue of forcing those who want to patronize businesses to wear a mask and follow state established guidelines. We would reason that given the bias toward zero and the potential for a strong negative relationship with the number of deaths.

The fourth variable that could be causing omitted variable bias would be a variable called "Business Closure Compliance". This refers to the choice businesses make to comply with state ordered closures. Just as with personal compliance and business mask compliance, there are businesses that stand with or in defiance of state mandated closures.

As with other compliance related metrics, we would expect to see a positive relationship between state ordered closures and compliance with these orders to avoid punishment by the state. As with other compliance related variables, division between business owners and the state is apparent and contributes to the omitted variable discussion. Unlike previous assumptions, we will assume that a some portion of business owners do not follow mandatory closures and business compliance is somewhat low. We note that there is no proxy of this variable as the neither the ANES data nor the data provided explicitly ask about how well a business complies with a closure order. Should this data be available in the ANES data, we would maintain another level of omitted variable bias (see Model 2).

Estimated Model: $\hat{Death} = \tilde{\beta}_o + \tilde{\beta}_1 MaskInd + \tilde{\beta}_2 MaskDays + \tilde{\beta}_3 StayHomeDays + \tilde{\beta}_4 StayHomeInd + \tilde{\beta}_5 PubOpinion + \tilde{\beta}_6 MobilityHome + \tilde{\beta}_7 DaycareClosure + \tilde{\beta}_8 RestaurantClosure + \tilde{\beta}_9 BarClosure$

True Model: $\hat{Death} = \beta_o^* + \beta_1 MaskInd^* + \beta_2 MaskDays^* + \beta_3 StayHomeDays^* + \beta_4 StayHomeInd^* + \beta_5 PubOpinion^* + \beta_6 MobilityHome^* + \beta_7 DaycareClosure^* + \beta_8 RestaurantClosure^* + \beta_9 BarClosure^* \cdot \beta_{10} BusinessClosureCompliance$

Direction of bias: Given the thought above, we can safely reason if some businesses do not comply with closure mandates, we would anticipate a negative direction of the bias. Meaning, as we observe mask mandates, we would anticipate high levels of business compliance, which would drive the bias away from zero. Practically speaking, lower business compliance would lead to more deaths per million by not following state mandates and closing when asked. We would reason that given the bias away from zero and the potential for a strong positive relationship with the number of deaths.

Classic Linear Model (CLM) Assumptions

CLM - Model 1

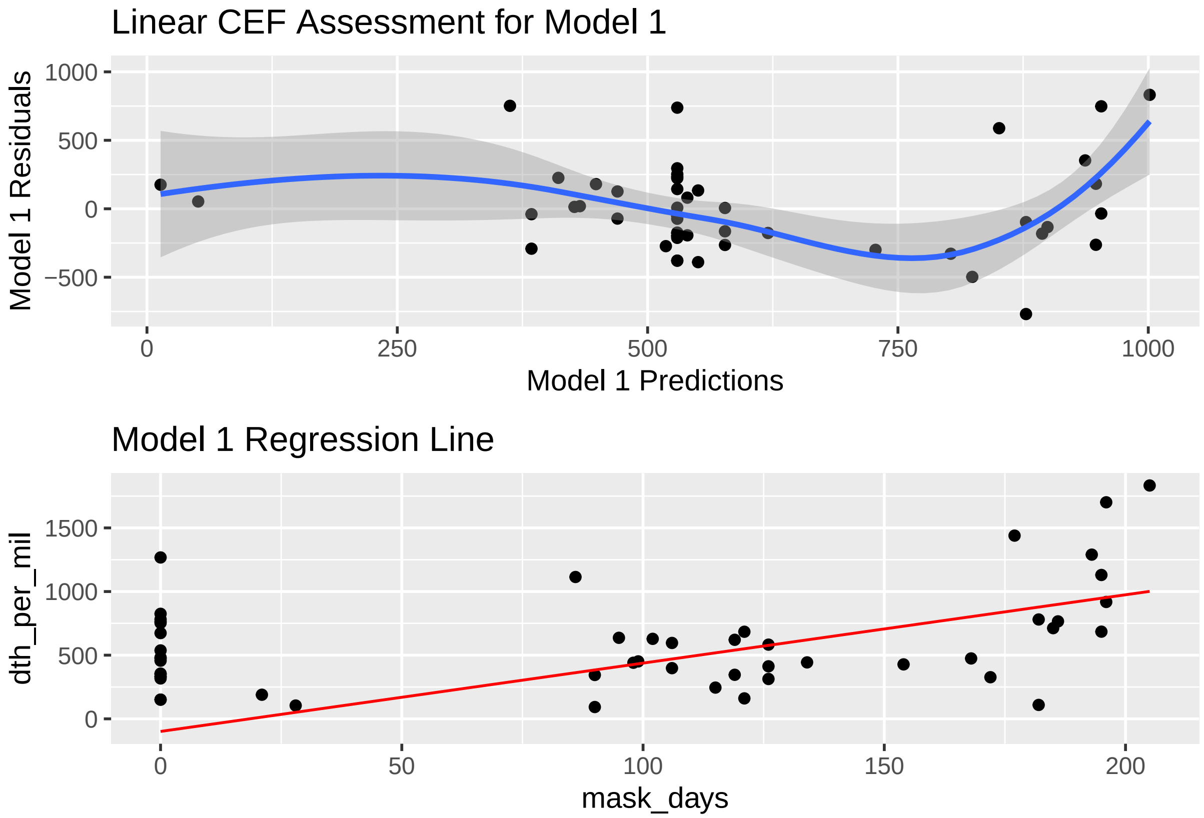
Assumption 1 - IID

Our sample consists of all possible states. We have no distribution concerns; however, independence is more complicated. While each state ostensibly would implement precautions exclusively based on the circumstances within the given state, we may have clustering within each RV as the disease seemed to have different regional impacts, which would instigate reactive responses within that 'cluster' of states. Subsequent studies should attempt to address the clustering presumable present in each RV.

Assumption 2 - Linear Conditional Expectation Function (CEF)

```r
#use these strategies for models 2 and 3 to ensure that meet CLM

#Linear Conditional Expectation - assess whether the data is truly linear.
#so first create new variables for the models predictions and residuals
covid <- covid %>%
  mutate(
    model1_pred = predict(model_1),
    model1_res = resid(model_1)
    )

# Plot the residuals against the prediction, along with a regression.
plot_mod1_lin <- covid %>%
  ggplot(aes(x = model1_pred , y = model1_res)) +
  geom_point() + stat_smooth() +
  labs(title = 'Linear CEF Assessment for Model 1',
       x = 'Model 1 Predictions',
       y = 'Model 1 Residuals')

#attempting to plot regression line on the mask days scatter plot
eqn1 = function(x){coef(model_1)[1]+coef(model_1)[2]+x*coef(model_1)[3]}

plot_mask_td2 <- covid %>%
  ggplot(aes(x = mask_days, y = dth_per_mil)) +
  geom_point()+
  stat_function(fun = eqn1, geom = "line", color = "red") +
  labs(title = 'Model 1 Regression Line')

plot_mod1_lin/plot_mask_td2
```

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



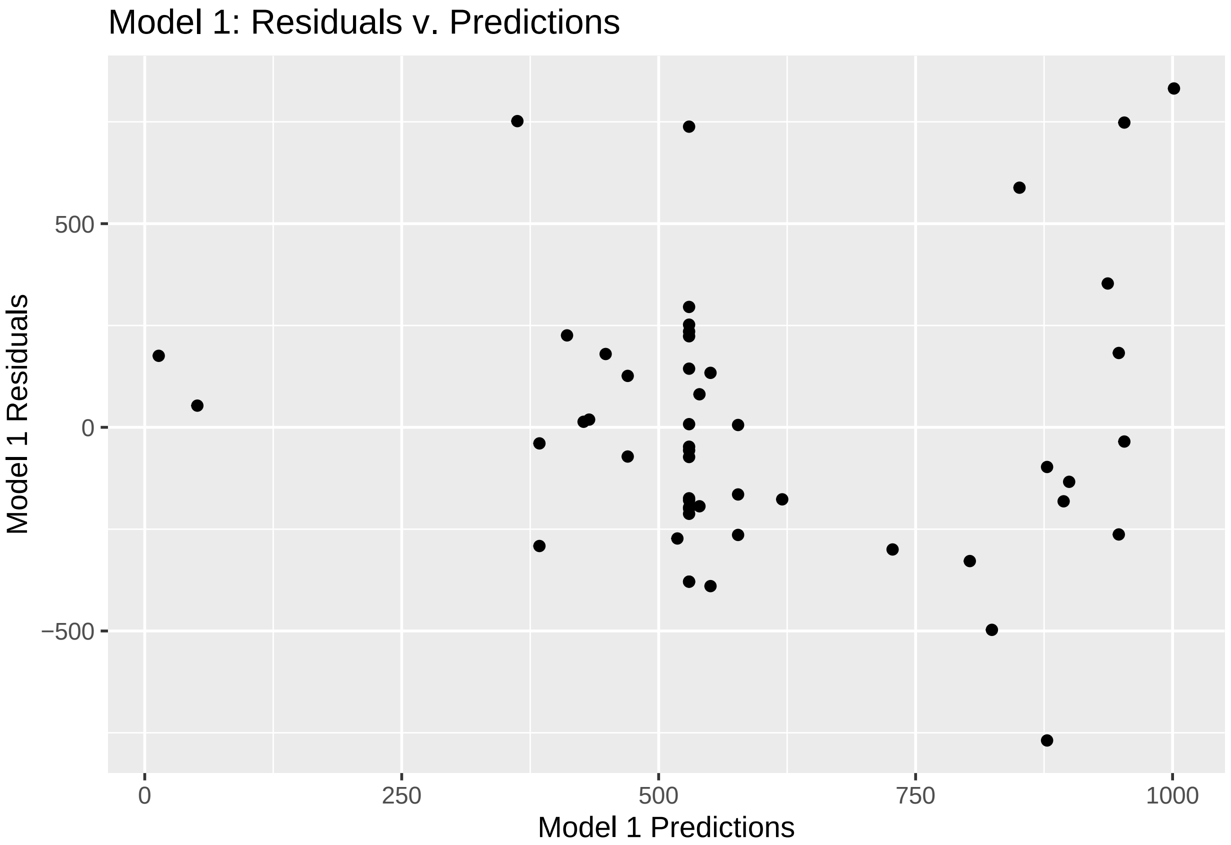
The above Linear CEF assessment plot suggests that Model 1 as constructed does not meet the Linearity assumptions necessary to validate the Model. Especially when observed against the plot of the regression line against total mask days for each state, it's clear that the model performs well for states with less than 125 days of mandated mask time, but the model should be adjusted to account for severe under-estimates as mandated mask time approaches 200 days.

Assumption 3 - No Perfect Collinearity

Model 1 only describes the association between deaths per million and one variable, mandated mask days. Therefore there isn't opportunity for collinearity.

Assumption 4 - Homskedasticity

```
#first plot resids against pred and look for fanning
plot_mod1_lin2 <- covid %>%
  ggplot(aes(x = model1_pred , y = model1_res)) +
  geom_point() +
  labs (title = "Model 1: Residuals v. Predictions",
      x = 'Model 1 Predictions',
      y = 'Model 1 Residuals')

plot_mod1_lin2
```



```
# There is a little fanning there
lmtest::bptest(model_1)
```

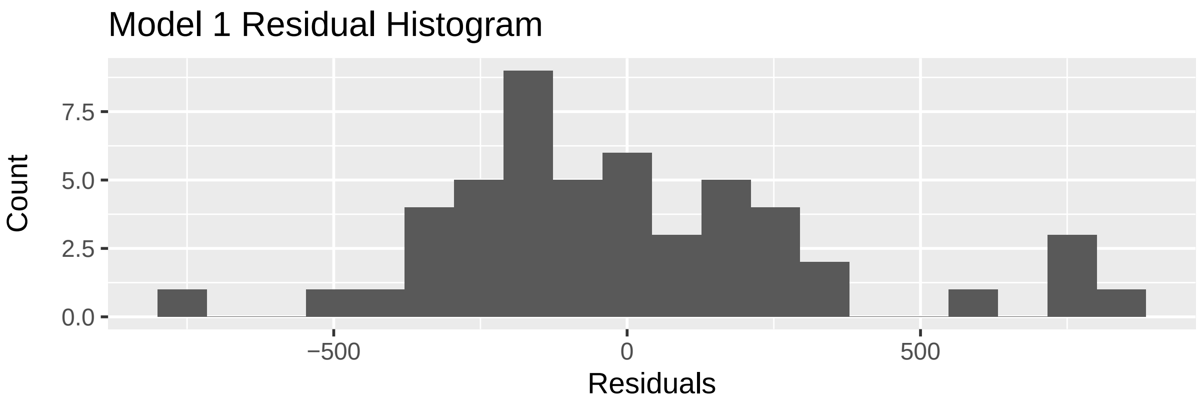
```
##
##  studentized Breusch-Pagan test
##
## data:  model_1
## BP = 5.9101, df = 2, p-value = 0.05208
```

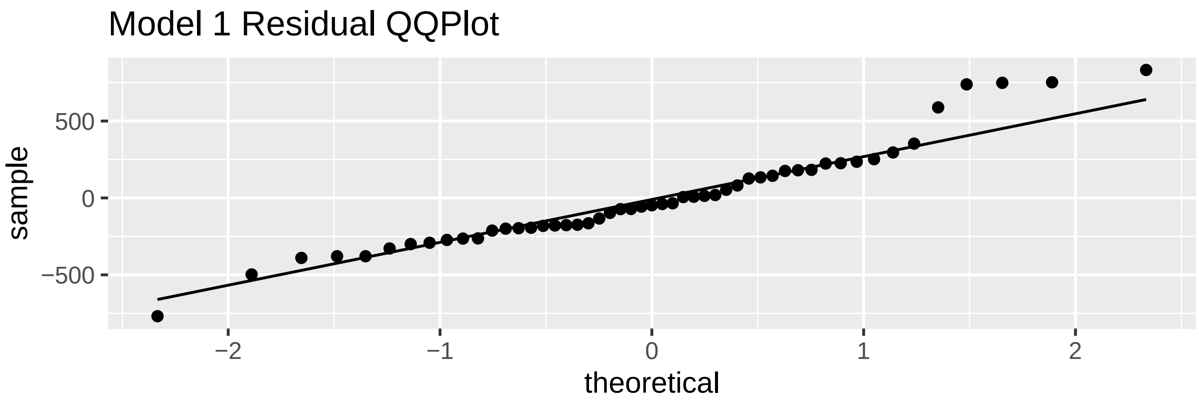
```r
# The model returns a p value above .05, we fail to reject the null hypothesis.
# We feel confident enough that our data is homsekdastic.
```

Visually assessing Model 1 residuals versus predictions, some slight fanning appears to be present in the residuals, suggesting heteroskedasticity. Furthermore, the Breusch-Pagan test does yield a p-value that does not allow us to reject the null hypothesis, suggesting that heteroskedasticity may not be present in this sample; however, the p-value of .052 is not convincing. A correction to Model 1 may lead to a safer assumption of homoskedasticity, please see the discussion below. We account for potential heteroskedasticity by using robust standard errors to asses our model's coefficients.

Assumption 5 - Normal Error Distribution

```r
#finally normality of residuals
hist_mod1.resid <- covid %>%
  ggplot(aes(x = model1_res)) +
  geom_histogram(bins = 20) +
  labs(title = "Model 1 Residual Histogram",
       x = "Residuals",
       y = "Count")

qqplot_mod1.resid <- covid %>%
  ggplot(aes(sample = model1_res)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "Model 1 Residual QQPlot")

#looks good to me - need to figure out the error on the QQline function
hist_mod1.resid/qqplot_mod1.resid
```


Model 1 Residual Histogram


Model 1 Residual QQPlot

Both the Model 1 residual histogram and qq plot suggest relatively normal residuals, so it seems like Model 1 satisfies this assumption.

Model 1 Adjustment

```
#without indicator and using stay at home days to capture potentially
#exponential relationship between deaths per million and mask mandate days.

model2 = lm(dth_per_mil ~ mask_days + ind_mask + worried, covid)

model_1.1.2 = lm(dth_per_mil~ind_mask+I(mask_days^14),covid)

summary(model_1.1.2)
```

```
##
## Call:
## lm(formula = dth_per_mil ~ ind_mask + I(mask_days^14), data = covid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -592.21 -195.37  -36.62  169.07  822.81
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.296e+02  7.104e+01   7.455 1.48e-09 ***
## ind_mask        -9.058e+01  9.194e+01  -0.985    0.329
## I(mask_days^14)  5.995e-30  9.357e-31   6.406 5.98e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 292.9 on 48 degrees of freedom
## Multiple R-squared:  0.4688, Adjusted R-squared:  0.4466
## F-statistic: 21.18 on 2 and 48 DF,  p-value: 2.551e-07
```

```r
coeftest(model_1.1.2,vcov = vcovHC)
```

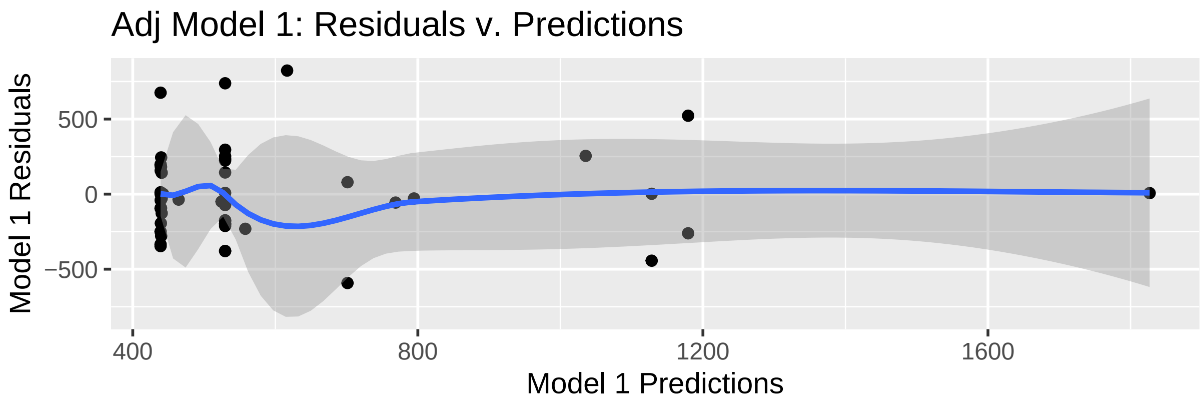
```
##
## t test of coefficients:
##
##                     Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)      5.2960e+02  7.1424e+01  7.4149 1.705e-09 ***
## ind_mask        -9.0584e+01  8.9338e+01 -1.0139    0.3157
## I(mask_days^14)  5.9947e-30  8.5276e-31  7.0298 6.620e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

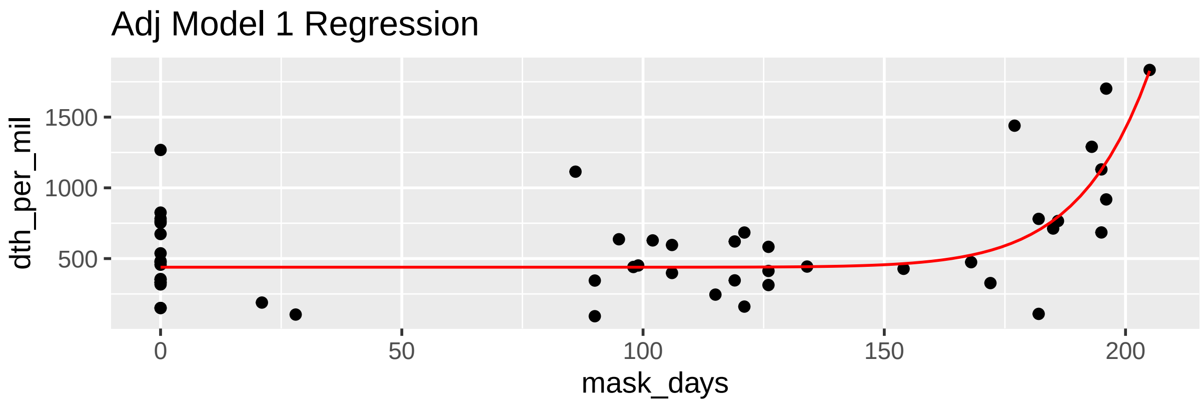
The Model 1 Linear CEF assessment suggests that Model 1 should capture the apparently exponential relationship between deaths per million and mandated mask days, especially for states that mandated masks for more than 150 days. To account for this relationship we provide a new Model 1 below:

$$\hat{Death} = 529.6 - 90.6 * I(\text{Mask}) + 5.99x10^{-30} * \text{MaskDays}^{14}$$

The Adjusted Model summary suggests that this adjusted model performs slightly better than Model 1 (comparing F-statistics), with significant intercept and Mask Days coefficients (using robust standard errors).

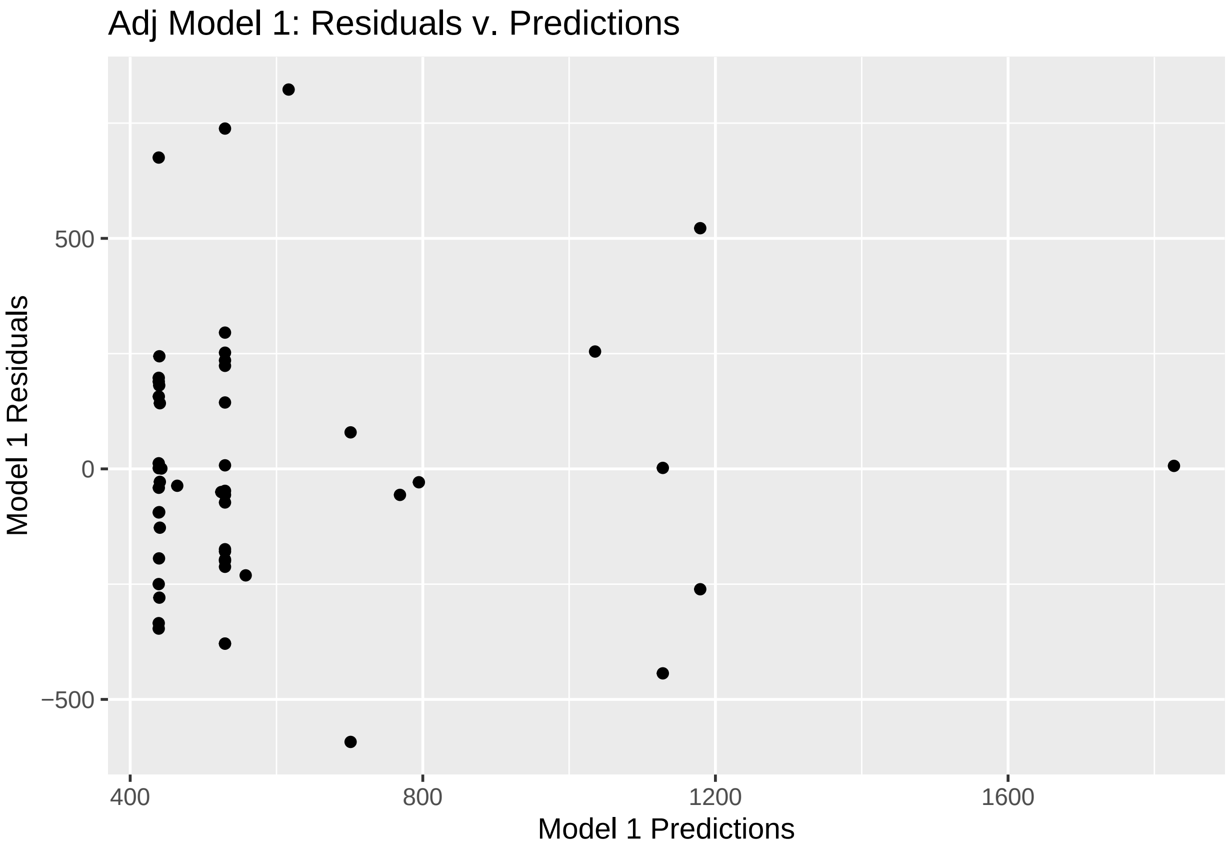
```r
#so first create new variables for the models predictions and residuals
covid <- covid %>%
  mutate(
    model1.2_pred = predict(model_1.1.2),
    model1.2_res = resid(model_1.1.2)
    )

#then plot the residuals against the prediction, along with a regression.
plot_mod1.2_lin <- covid %>%
  ggplot(aes(x = model1.2_pred , y = model1.2_res)) +
  geom_point() + stat_smooth() +
  labs (title = "Adj Model 1: Residuals v. Predictions",
      x = 'Model 1 Predictions',
      y = 'Model 1 Residuals')

#attempting to plot regression line on the mask days scatter plot
eqn2 = function(x){coef(model_1.1.2)[1]+coef(model_1.1.2)[2]+(x^14)*coef(model_1.1.2)[3]}

plot_mask_td3 <- covid %>%
  ggplot(aes(x = mask_days, y = dth_per_mil)) +
  geom_point() +
  stat_function(fun = eqn2, geom = "line", color = "red") +
  labs(title = "Adj Model 1 Regression")

plot_mod1.2_lin/plot_mask_td3
```





The Adjusted Model 1 residuals v. predictions and regression line plots visually demonstrate linearity and a closer fit than before. Additionally, our adjustments to Model 1 yields a residual plot that demonstrates much more homoskedasticity and much more convincing Breusch-Pagan test results without compromising residual normality. Unfortunately while our Adjusted Model 1 seems to satisfy all CLM assumptions, it is significantly more complicated than our original Model 1 and much more difficult to explain. Our models are intended to be descriptive, as such, we prefer to use our original Model 1, and present the potential CLM limitations for our audience to consider when assessing our results.

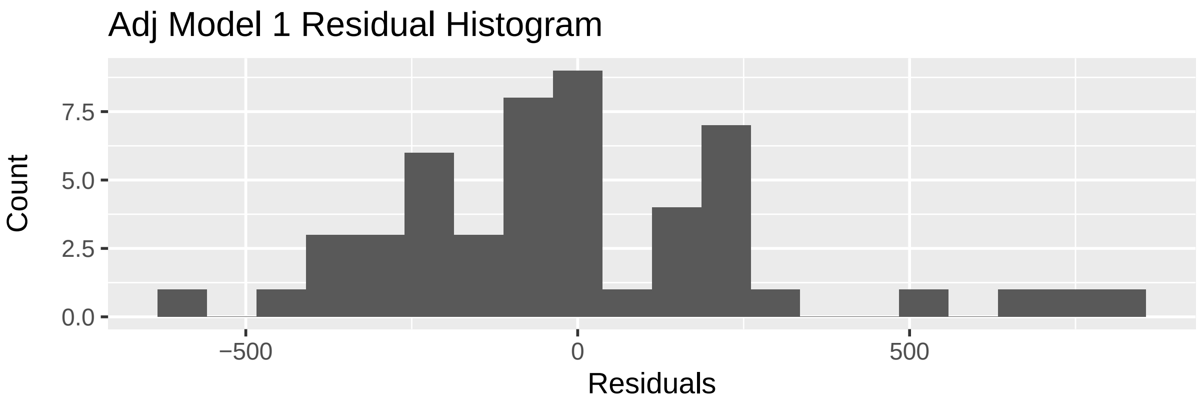
```
plot_mod1.2_error <- covid %>%
  ggplot(aes(x = model1.2_pred , y = model1.2_res)) +
  geom_point() +
  labs (title = "Adj Model 1: Residuals v. Predictions",
        x = 'Model 1 Predictions',
        y = 'Model 1 Residuals')

plot_mod1.2_error
```

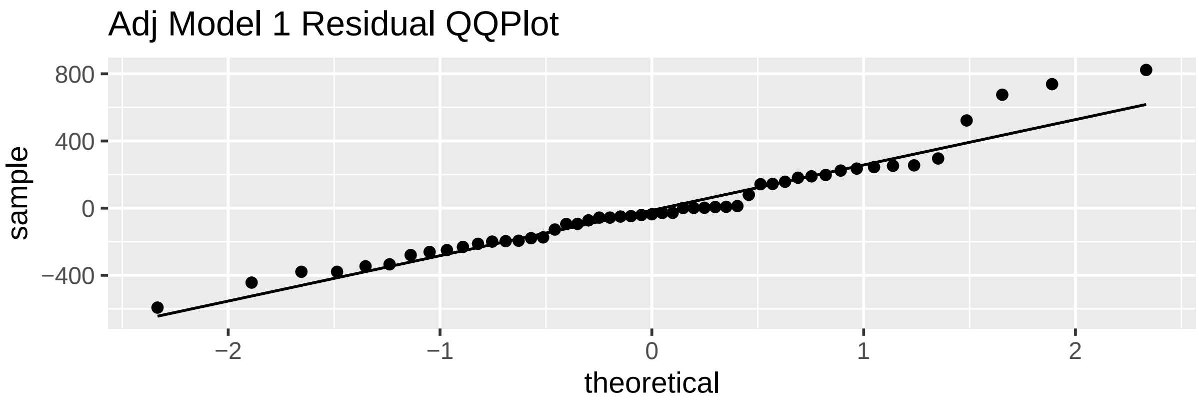


```
lmtest::bptest(model_1.1.2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_1.1.2
## BP = 0.20695, df = 2, p-value = 0.9017
```

```
#finally normality of residuals
hist_mod1.2_resid <- covid %>%
  ggplot(aes(x = model1.2_res)) +
  geom_histogram(bins = 20) +
  labs(title = "Adj Model 1 Residual Histogram",
       x = "Residuals",
       y = "Count")

qqplot_mod1.2_resid <- covid %>%
  ggplot(aes(sample = model1.2_res)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "Adj Model 1 Residual QQPlot")

#looks good to me - need to figure out the error on the QQline function
hist_mod1.2_resid/qqplot_mod1.2_resid
```





CLM - Model 2

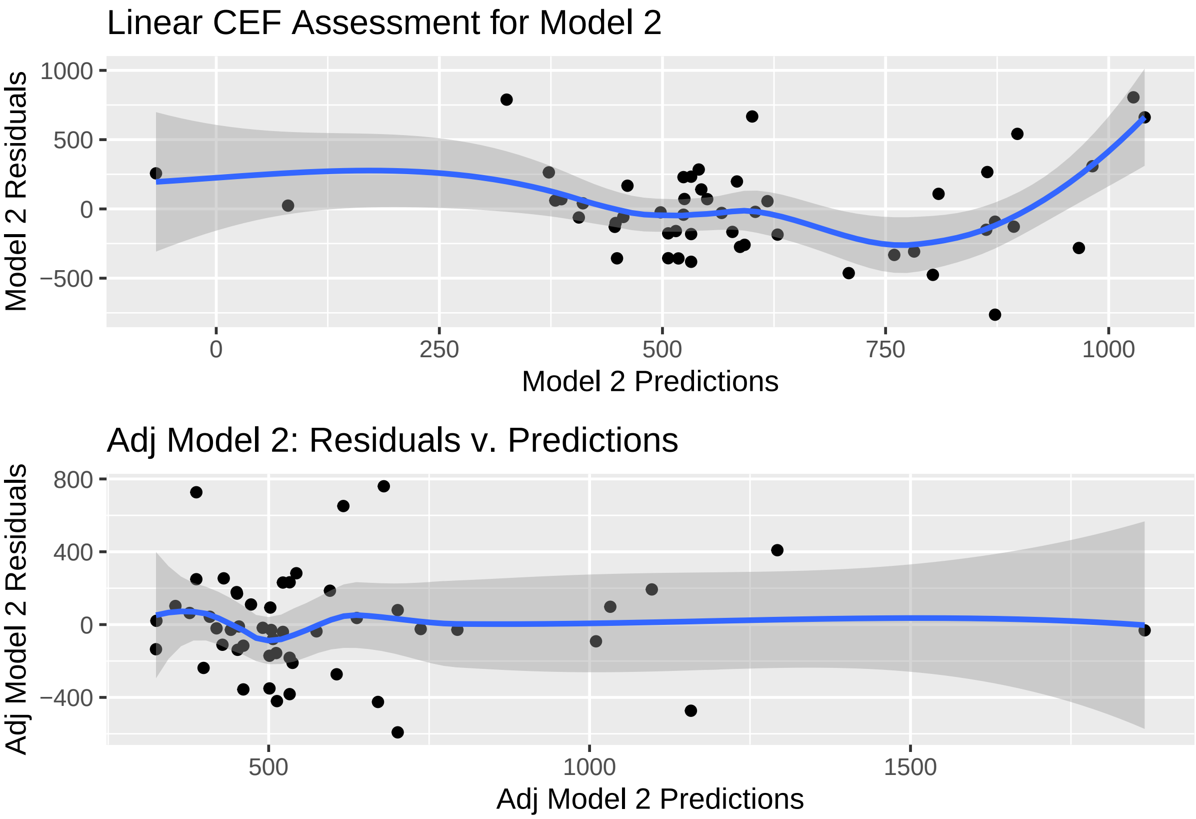
Assumption 1 - IID

Please see our IID discussion for Model 1.

Assumption 2 - Linear CEF

```
#Linear Conditional Expectation - assess whether the data is truly linear.
#so first create new variables for the models predictions and residuals
covid <- covid %>%
  mutate(
    model2_pred = predict(model2),
    model2_res = resid(model2)
    )

#then plot the residuals against the prediction, along with a regression
plot_mod2_lin <- covid %>%
  ggplot(aes(x = model2_pred , y = model2_res)) +
  geom_point() + stat_smooth() +
  labs(title = 'Linear CEF Assessment for Model 2',
       x = 'Model 2 Predictions',
       y = 'Model 2 Residuals')

#adjust model two and create a linearity plot

adj_model2 = lm(dth_per_mil ~ ind_mask + I(mask_days^14) + worried, covid)

#so first create new variables for the models predictions and residuals
```

```
covid <- covid %>%
  mutate(
    adj_model2_pred = predict(adj_model2),
    adj_model2_res = resid(adj_model2)
    )

#then plot the residuals against the prediction, along with a regression
plot_adj_mod2_lin <- covid %>%
  ggplot(aes(x = adj_model2_pred , y = adj_model2_res)) +
  geom_point() + stat_smooth() +
  labs (title = "Adj Model 2: Residuals v. Predictions",
      x = 'Adj Model 2 Predictions',
      y = 'Adj Model 2 Residuals')

plot_mod2_lin/plot_adj_mod2_lin
```



Unsurprisingly the Linear CEF assessment plot for Model 2 demonstrates the same linearity issues as Model 1, with especially non-linear performance towards the models 'later' predictions. If we adjust Model 2 to incorporate our changes to Model 1, we see much stronger evidence of linearity.

Assumption 3 - No Perfect Collinearity

```
#Collinearity: we don't have perfect collinearity
#nothing was dropped when we built the model.
#Additionally we don't have near perfect collinearity.

model2.1 <- lm(dth_per_mil~worried, covid)
```

```r
summary(model2.1)
```

```
##
## Call:
## lm(formula = dth_per_mil ~ worried, data = covid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -595.87 -240.19  -90.12  157.76 1191.74
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    342.1      197.2   1.735    0.089 .
## worried       1109.0      827.2   1.341    0.186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 390.7 on 49 degrees of freedom
## Multiple R-squared:  0.03538,    Adjusted R-squared:  0.0157
## F-statistic: 1.797 on 1 and 49 DF,  p-value: 0.1862
```

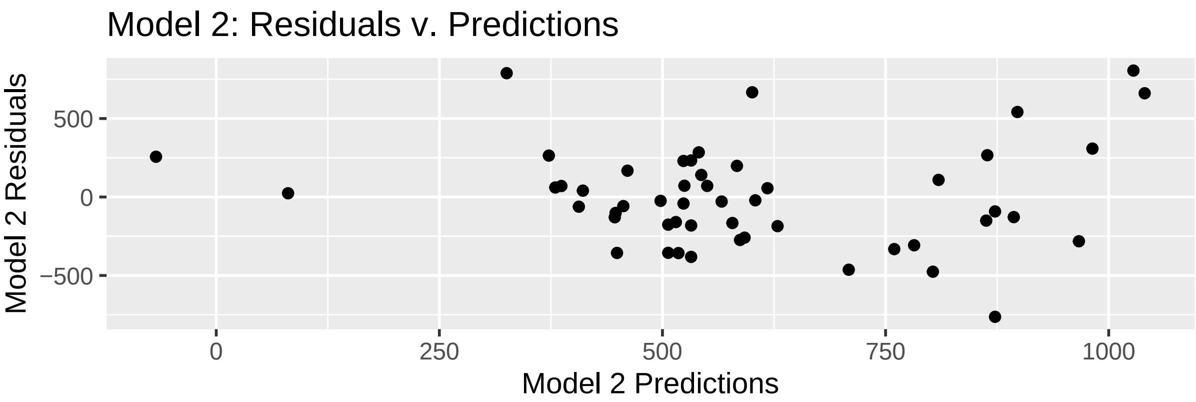
```r
#compute correlation between the variables
cor(covid$worried,covid$mask_days)
```

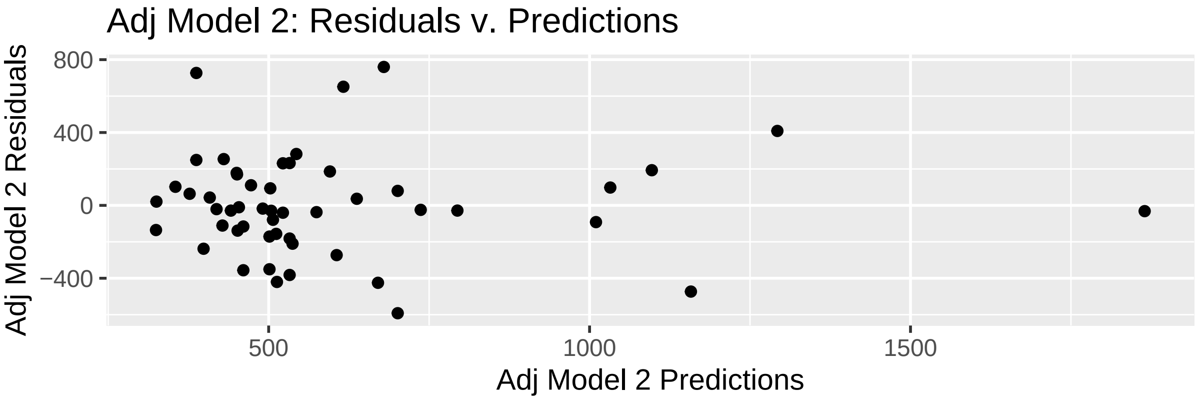
```
## [1] 0.05817922
```

There does not appear to be perfect or near-perfect collinearity present in Model 2. First, we do not see wild swings in coefficient value or significance when we add the RV ‘worried’ to Model 2. Second, the correlation between ‘worried’ and ‘MaskDays’ is relatively weak (~.06). Consequently, neither variable was dropped during Model 2 derivation.

Assumption 4 - Homoskedasticity

```r
#heteroskedasticity check

#first plot resids against pred and look for fanning
plot_mod2_lin2 <- covid %>%
  ggplot(aes(x = model2_pred , y = model2_res)) +
  geom_point() +
  labs (title = "Model 2: Residuals v. Predictions",
      x = 'Model 2 Predictions',
      y = 'Model 2 Residuals')

#Now adjust
#first plot resids against pred and look for fanning
plot_adj_mod2_lin2 <- covid %>%
  ggplot(aes(x = adj_model2_pred , y = adj_model2_res)) +
  geom_point() +
  labs (title = "Adj Model 2: Residuals v. Predictions",
      x = 'Adj Model 2 Predictions',
      y = 'Adj Model 2 Residuals')

plot_mod2_lin2/plot_adj_mod2_lin2
```





```
#pre-adjustment
lmtest::bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 7.6625, df = 3, p-value = 0.05353
```

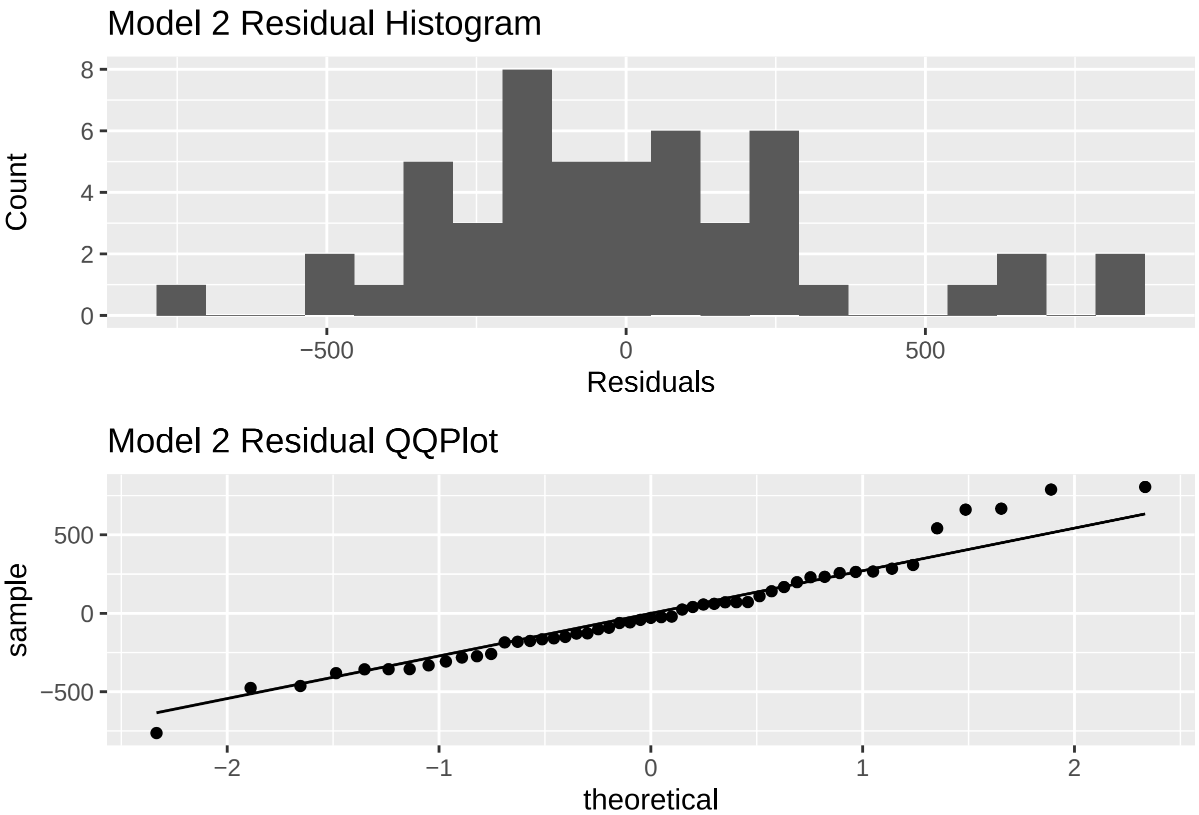
```
#after adjustment
lmtest::bptest(adj_model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  adj_model2
## BP = 3.8997, df = 3, p-value = 0.2725
```

Very similar to Model 1, our un-adjusted Model 2 offers scant evidence to suggest homoskedasticity. We account for potential heteroskedasticity by using robust standard errors to asses our model's coefficients. Once we adjust Model 2, we see less fanning in the residual v. prediction plot and a higher Breusch-Pagan test p-value, again allowing us to fail to reject the null hypothesis that our residual variance is homoskedastic.

Assumption 5 - Normal Error Distribution

```
#finally normality of residuals
hist_mod2.resid <- covid %>%
  ggplot(aes(x = model2_res)) +
  geom_histogram(bins = 20) +
  labs(title = "Model 2 Residual Histogram",
       x = "Residuals",
       y = "Count")

qqplot_mod2.resid <- covid %>%
  ggplot(aes(sample = model2_res)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "Model 2 Residual QQPlot")

#looks good to me - need to figure out the error on the QQline function
hist_mod2.resid/qqplot_mod2.resid
```



Both the Model 2 residual histogram and qq plot suggest relatively normal residuals, so it seems like Model 2 satisfies this assumption.

CLM - Model 3

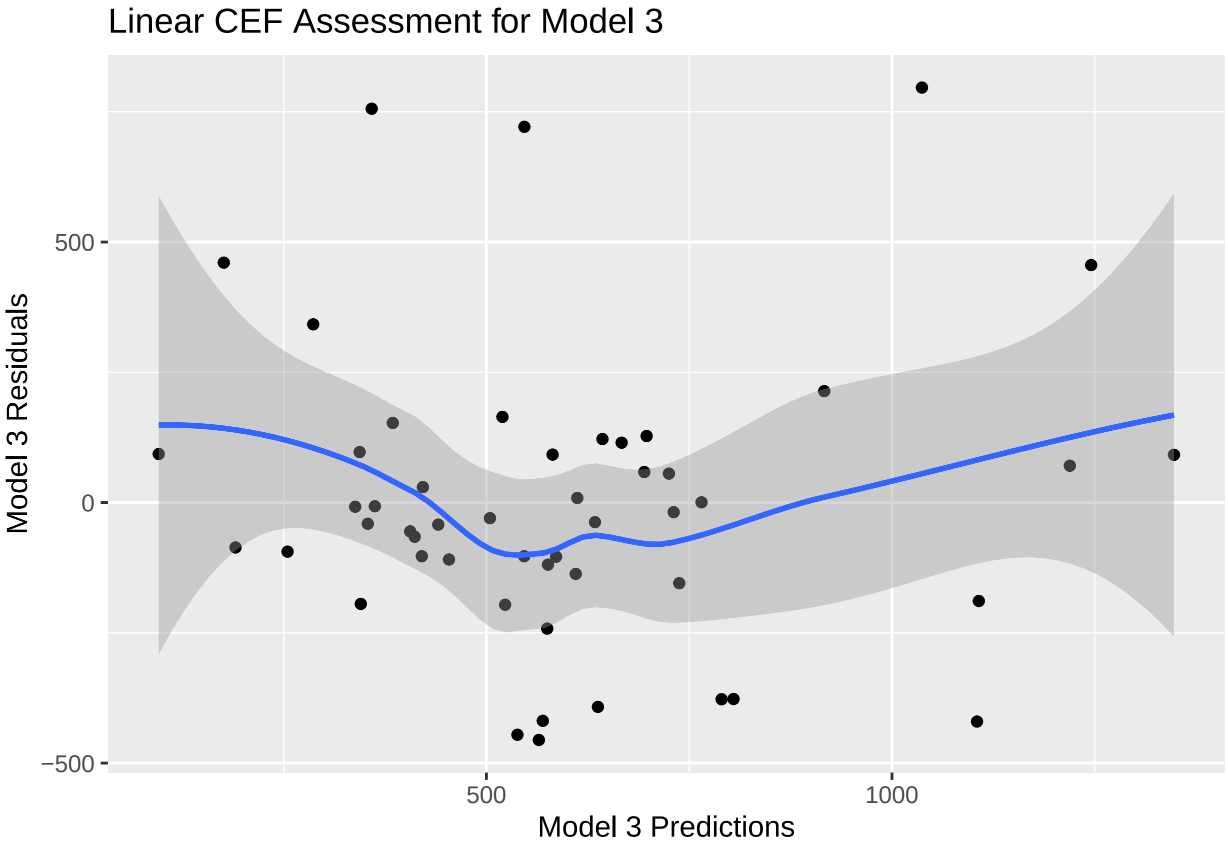
Assumption 1 - IID

Please see our IID discussion for Model 1.

Assumption 2 - Linear CEF

```
#Linear Conditional Expectation - assess whether the data is truly linear.
covid <- covid %>%
  mutate(
    model3_pred = predict(model3),
    model3_res = resid(model3)
    )
```

```
#then plot the residuals against the prediction, along with a regression.
plot_mod3_lin <- covid %>%
  ggplot(aes(x = model3_pred , y = model3_res)) +
  geom_point() + stat_smooth() +
  labs(title = 'Linear CEF Assessment for Model 3',
       x = 'Model 3 Predictions',
       y = 'Model 3 Residuals')

plot_mod3_lin
```



Model 3 stays much more linear through out all predictions than does either Model 1 or Model 2 and seems to meet the Linearity assumption. Presumably adding variables reduced abated the impact of MaskDays on Model 3 residuals.

Assumption 3 - No Perfect Collinearity

```
#Overall summary
summary(model3)
```

```
##
## Call:
## lm(formula = dth_per_mil ~ mask_days + ind_mask + worried + stayathome_days +
##     ind_stay + mob_data_hm + daycare + restaurant + bar, data = covid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -455.55 -128.02  -29.98   95.07  796.35
##
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)   
## (Intercept)      194.3376   255.3025   0.761  0.45089   
## mask_days          4.6326     1.5466   2.995  0.00464 **
## ind_mask        -571.0453   192.1162  -2.972  0.00493 **
## worried          888.2369   695.4367   1.277  0.20870   
## stayathome_days   -3.1667     1.4503  -2.184  0.03477 * 
## ind_stay         216.4077   136.0352   1.591  0.11933   
## mob_data_hm       56.9275    23.8247   2.389  0.02156 * 
## daycare            0.8539     1.6854   0.507  0.61512   
## restaurant        -1.8231     3.3234  -0.549  0.58629   
## bar               -4.0967     1.8113  -2.262  0.02908 * 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 310.3 on 41 degrees of freedom
## Multiple R-squared:  0.4908, Adjusted R-squared:  0.379 
## F-statistic: 4.391 on 9 and 41 DF,  p-value: 0.0004644
```

```r
#individual RV model sumaries
model3.1 <- lm(dth_per_mil ~ stayathome_days + ind_stay, covid)
summary(model3.1)
```

```
##
## Call:
## lm(formula = dth_per_mil ~ stayathome_days + ind_stay, data = covid)
##
## Residuals:
##    Min     1Q Median     3Q    Max 
## -527.5 -268.5 -132.9  146.0 1207.4 
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     503.3642   120.2040   4.188  0.00012 ***
## stayathome_days   0.2071     1.5178   0.136  0.89201    
## ind_stay        105.9359   163.0270   0.650  0.51892    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.7 on 48 degrees of freedom
## Multiple R-squared:  0.01595,    Adjusted R-squared:  -0.02506 
## F-statistic: 0.3889 on 2 and 48 DF,  p-value: 0.6799
```

```r
model3.2 <- lm(dth_per_mil ~ mob_data_hm, covid)
summary(model3.2)
```

```
##
## Call:
## lm(formula = dth_per_mil ~ mob_data_hm, data = covid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -774.14 -220.08  -80.74  132.89  943.81 
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.49     181.00   0.063  0.94965
## mob_data_hm    69.06      20.54   3.362  0.00151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 358.6 on 49 degrees of freedom
## Multiple R-squared:  0.1874, Adjusted R-squared:  0.1708
## F-statistic:  11.3 on 1 and 49 DF,  p-value: 0.001508
```

```
model3.3 <- lm(dth_per_mil ~ daycare, covid)
summary(model3.3)
```

```
##
## Call:
## lm(formula = dth_per_mil ~ daycare, data = covid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -650.21 -210.55  -74.25  112.98 1159.16
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  542.128     64.095   8.458 3.84e-11 ***
## daycare        2.676      1.686   1.587    0.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 387.9 on 49 degrees of freedom
## Multiple R-squared:  0.04888,    Adjusted R-squared:  0.02947
## F-statistic: 2.518 on 1 and 49 DF,  p-value: 0.119
```

```
model3.4 <- lm(dth_per_mil ~ restaurant, covid)
summary(model3.4)
```

```
##
## Call:
## lm(formula = dth_per_mil ~ restaurant, data = covid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -582.36 -236.11  -53.86  134.41 1098.81
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  349.257    156.256   2.235   0.0300 *
## restaurant     4.233      2.514   1.684   0.0985 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 386.7 on 49 degrees of freedom
## Multiple R-squared:  0.05472,    Adjusted R-squared:  0.03543
## F-statistic: 2.837 on 1 and 49 DF,  p-value: 0.0985
```

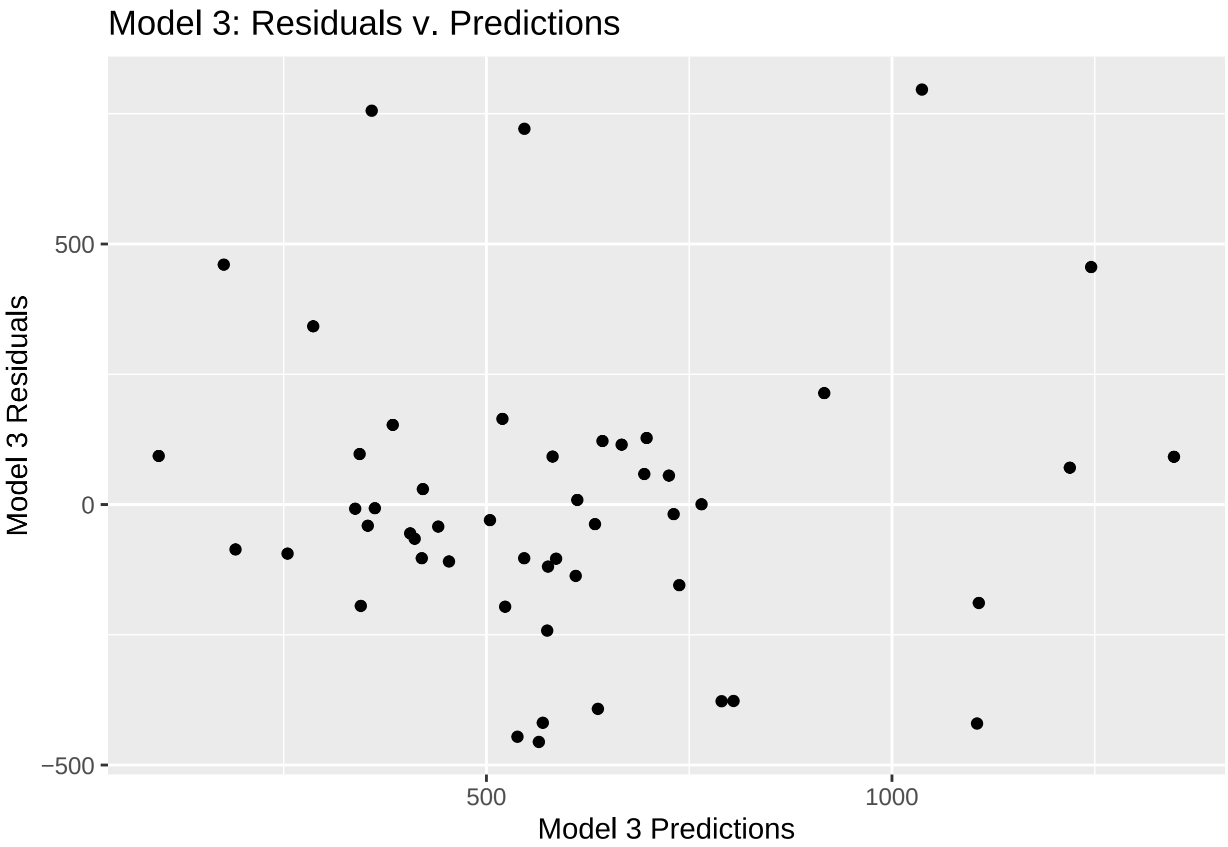
```
model3.5 <- lm(dth_per_mil ~ bar, covid)
summary(model3.5)
```

```
##
## Call:
## lm(formula = dth_per_mil ~ bar, data = covid)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -521.4 -295.3 -116.4  183.8 1313.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  760.852    125.417   6.067 1.85e-07 ***
## bar           -2.648      1.815  -1.459    0.151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 389.4 on 49 degrees of freedom
## Multiple R-squared:  0.04161,    Adjusted R-squared:  0.02205
## F-statistic: 2.127 on 1 and 49 DF,  p-value: 0.1511
```

We do not see any evidence of perfect or near-perfect collinearity within Model 3. Primarily, per the above summaries, we do not observe wild swings in coefficient values or significance. Consequently, none of the RVs were dropped during Model 3 derivation.

Assumption 4 - Homoskedasticity

```r
#plot to visually assess homskedasticity
plot_mod3_lin2 <- covid %>%
  ggplot(aes(x=model3_pred,y=model3_res)) +
  geom_point() +
  labs(title = "Model 3: Residuals v. Predictions",
      x = 'Model 3 Predictions',
      y = 'Model 3 Residuals')

plot_mod3_lin2
```



```
lmtest::bptest(model3)
```

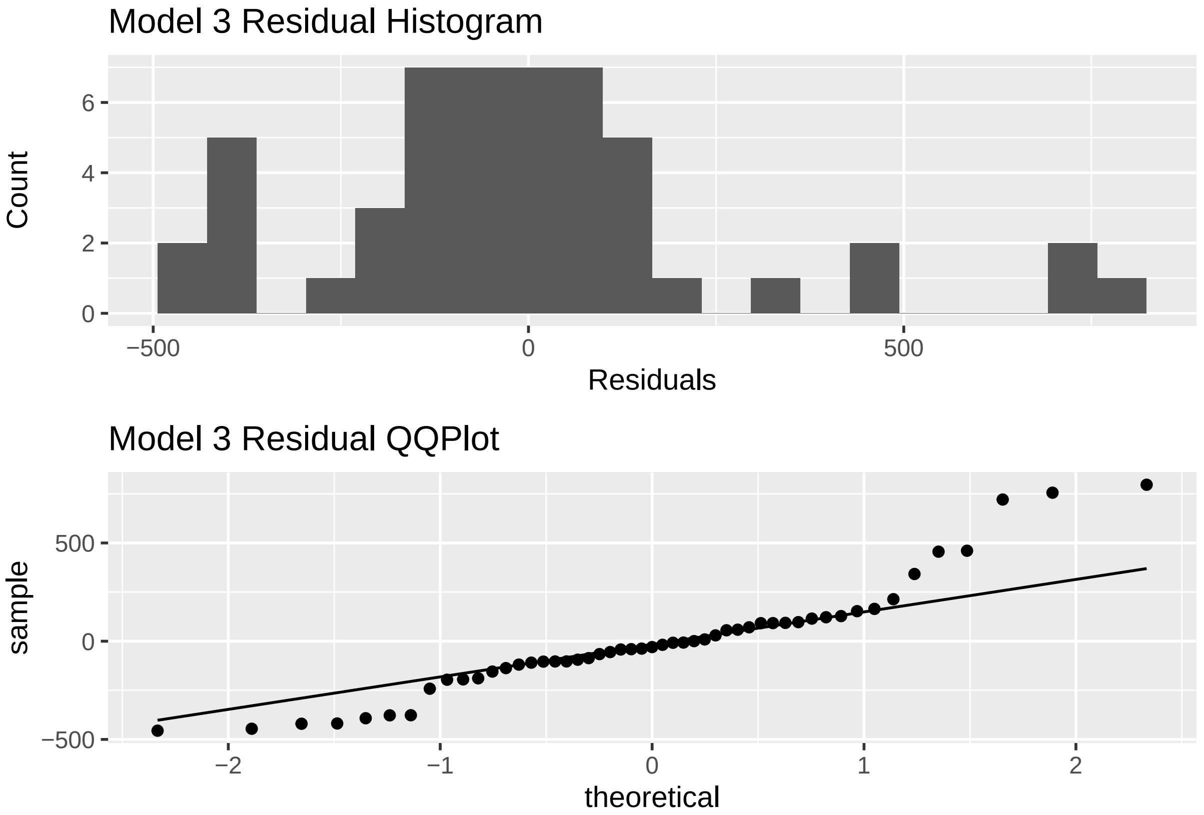
```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 6.0258, df = 9, p-value = 0.7373
```

The plot of Model 3 residuals v. predictions shows good spread in the residuals and the Breusch-Pagan test yielded a p value of .737, meaning that we fail to reject the null hypothesis and suggesting that Model 3 is homoskedastic.

Assumption 5 - Normal Error Distribution

```
#finally normality of residuals
hist_mod3.resid <- covid %>%
  ggplot(aes(x = model3_res)) +
  geom_histogram(bins = 20) +
  labs(title = "Model 3 Residual Histogram",
       x = "Residuals",
       y = "Count")

qqplot_mod3.resid <- covid %>%
  ggplot(aes(sample = model3_res)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "Model 3 Residual QQPlot")
```

```
#looks good to me - need to figure out the error on the QQline function
hist_mod3.resid/qqplot_mod3.resid
```



The above plots indicate some residual skew and do not present strong ocular evidence that Model 3 residuals depart from normality at the extremes of our sample. This does not overly concern us given the inclusive nature of Model 3 and how well it met Assumptions 1-4.

Conclusions

Our models suggested interesting associations between total deaths in a state and mitigation strategies imposed by each state, as well as the prevailing opinion of COVID in said state, at least according to how it was assessed in one nationwide opinion poll. While we chose to focus on exclusively descriptive models, in initially assessing our results, we could not help be surprised by positive associations with some mitigation strategies, and also with the worried percentage of a state population. However, we presumably were under the affect of causal thinking. It makes intuitive sense to consider that states with the longest mask mandates might have been forced into that reality because of their excessively high death numbers, especially at the outset of the pandemic. Additionally states with the highest worried percentage likely experienced the ferocity of the pandemic first-hand. Rather than considering how mandates or opinions might influence the course of the disease, if one considers how the diseases shaped different realities in different states, then positive associations begin to make much more sense. Importantly, these associations are just that, we cannot be sure that any one factor caused any other factor. More analysis is required to verify the causal impact of the disease on state-wise decision making and vice-versa.

Regardless, it remains interesting to observe the quantifiable differences in state-wise response to what could be the defining force of our time. In developing and responding to the research question, we observe that mask mandates are associated with total deaths per million. However, mitigation strategies were not imposed in a vacuum, they require compliance, by individuals and businesses. While our focus is not on a causal model, so drawing relationships between mask mandates and deaths is not our objective, we see evidence that supports our research question, and opportunity for further discussion and analysis around some of the omitted variables and opportunity to build on what has been developed and create an explanatory model. The associations described in the analysis exist. We observed statistically significant associations between mask mandates, mobility data of changes in time spent at home, length of stay at home orders and closures of bars. While not every variable we included was deemed statistically significant, we observed associations that would open doors to additional research questions that may be descriptive or explanatory in nature. One could wonder what affect these state ordered mandates had on the number of deaths, or measuring the affect of changes in home mobility on death. There are several potential research questions that could stem from the more statistically significant variables captured in this analysis. We see a complex web of interactions between mask mandates, closure orders, public opinion, mobility data and deaths due to Coronavirus, and it's within this web that we all live in work. We might think of these associations and coefficients or elements of model, but they are in fact, human behavior. A complex series of decisions made by individuals, businesses and governments that shape the outcomes of the world we interact in, the businesses we frequent and the masks we're mandated to wear.