Can Mask Mandates Help Reduce COVID-19 Deaths?

Srishti Mehra, Andi Morey Peterson, and David Djambazov

11/14/2020

Introduction

A common recommendation from doctors, scientists, and politicians is to wear face masks or other facial coverings to combat the spread of COVID-19. The World Health Organization states: "Masks are a key measure to suppress transmission and save lives. Masks reduce potential exposure risk from an infected person whether they have symptoms or not. People wearing masks are protected from getting infected. Masks also prevent onward transmission when worn by a person who is infected." We now have been in this pandemic for nearly 9 months, so with the data currently available, we want to build a causal model with data available about the current pandemic.

Many states have issued mandates for their citizens to wear masks in public and additionally, many states have issued mandates for employees who interact with the public to wear masks. Now that cases have surged in the United States (When this was written in December of 2020), we ask: *Do state-wide face mask mandates cause a reduction in the amount of deaths in that state due to COVID-19?*

Because these mandates and other variables occur at different times during the nine-month period, and the states were affected by the virus in different times as well, we will operationalize "deaths" as the *Death Rate per 100,000 in the past 7 days* (ending Oct. 30th). The *total deaths* variable is inappropriate, as the population per state varies widely, skewing the data in largely populated states. The total death rate per 100,000 is also inappropriate, as many states had large first waves in the beginning of the year and instituted measures only in response to an already high death toll. Also the large first wave numbers will skew the analysis of the current state of the pandemic. Using the current data (past 7 days) should provide basis for analysis if mandates are working *currently.*

Another important consideration for our choice of outcome variable is that states that got hit at an early stage (such as NY, NJ and WA) have learned from the experience and imposed a number of policy measures. It is an open question if those are successful and that's why it would be interesting to see in our EDA whether the states most impacted by the current record wave of infections were indeed spared by the first wave and perhaps less stringent in their measures.

We will operationalize "mandates" as TRUE/FALSE indicators.

As we move through the model, we will do some causal analysis on other variables that may interact with the main question. These variables include: percent of people over 65, racial diversity, political leanings, mobility, etc. We will analyze each of these variables to come up with a final model to determine if and how much face masks are currently reducing population death rates due to COVID-19.

Importing the data

One thing to note is that some values are marked as `NR` - "Not Reported". We will encode them to NA for consistency. Essentially, there is no meaningful difference between NA and NR other than the fact that NR value reflects that a given state specifically does not record a certain variable as opposed to some other possible reason. For our purposes we can treat them as the same.

```
covid_raw_data<-read.csv("covid-19.csv",skip=1, na.strings=c("NA", "NR"))
covid_masks_policies_data<-read.csv("covid policies masks.csv")

covid_data<-left_join(
  covid_raw_data,
  covid_masks_policies_data)
```

Rename variables and define our operationalized mandate variable

For better coding practices we are going to rename all of the variables we intend to look at. We also add an indicator variable whether the governor of a state is a Republican. The logic behind that is that mask mandates have become a politicized issue, one that mainly divides along partisan lines with Republicans opposed and Democrats in favor of the measure.

There are several different mask mandate variables. We are focusing on what we consider the most potentially impactful one - "statewide mask use by individuals for all in public spaces". The reason we want to avoid including other such variables is that there's probably a very high correlation with other mask mandates (e.g. mandates for state employees) and because those target a much smaller section of the population.

To keep our analysis clean we're not considering other factors that may affect the rate of adoption of actual mask wearing, such as legal enforcement. The reason is that it would introduce a host of potential omitted variable relationships. Ultimately, issuing a mask mandate is also a question of leadership. By issuing it, state leaders are communicating that COVID-19 is a real problem and should be taken seriously. This action might plausibly increase mask wearing and that could be a causal pathway to reducing the impact of COVID-19, if indeed mask wearing reduces infection rates. Or it could impact citizens' behaviors in other ways by simply highlighting the seriousness of the crisis. Either way, our question is whether there is evidence that this act of government has lead to a reduction in population death rates at the beginning of the fall wave of contagion.

That is why our operationalization is an indicator variable showing if at the time of measurement of the last 7 day death rate per 100K (October 30th) there was an enacted mask mandate for all in the given state or not.

```
covid_data <- covid_data  %>%
  rename(
      case_rate = "Case.Rate.per.100000",
      case_rate_in_last7 = "Case.Rate.per.100000.in.Last.7.Days",
      death_rate = "Death.Rate.per.100000",
      death_rate_in_last7 = "Death.Rate.per.100K.in.Last.7.Days",
      mask_for_all_mandated_on = 'Mandate.face.mask.use.by.all.individuals.in.public.spaces',
      mask_for_all_end = 'State.ended.statewide.mask.use.by.individuals.in.public.spaces',
      mask_enforced_by_fines = 'Face.mask.mandate.enforced.by.fines',
      mask_enforced_by_charge = 'Face.mask.mandate.enforced.by.criminal.charge.citation',
      no_legal_mask_enforcement = 'No.legal.enforcement.of.face.mask.mandate',
      population_density = 'Population.density.per.square.miles',
      stay_at_home_begin = 'Stay.at.home..shelter.in.place',
      stay_at_home_end = 'End.stay.at.home.shelter.in.place',
      retail_mobility_change ='Retail...recreation',
      grocery_pharm_mobility_change='Grocery...pharmacy',
      parks_mobility_change='Parks',
      transit_mobility_change = 'Transit.stations',
      workplaces_mobility_change = 'Workplaces',
      residential_mobility_change = 'Residential',
      white_percent = 'White...of.Total.Population',
```

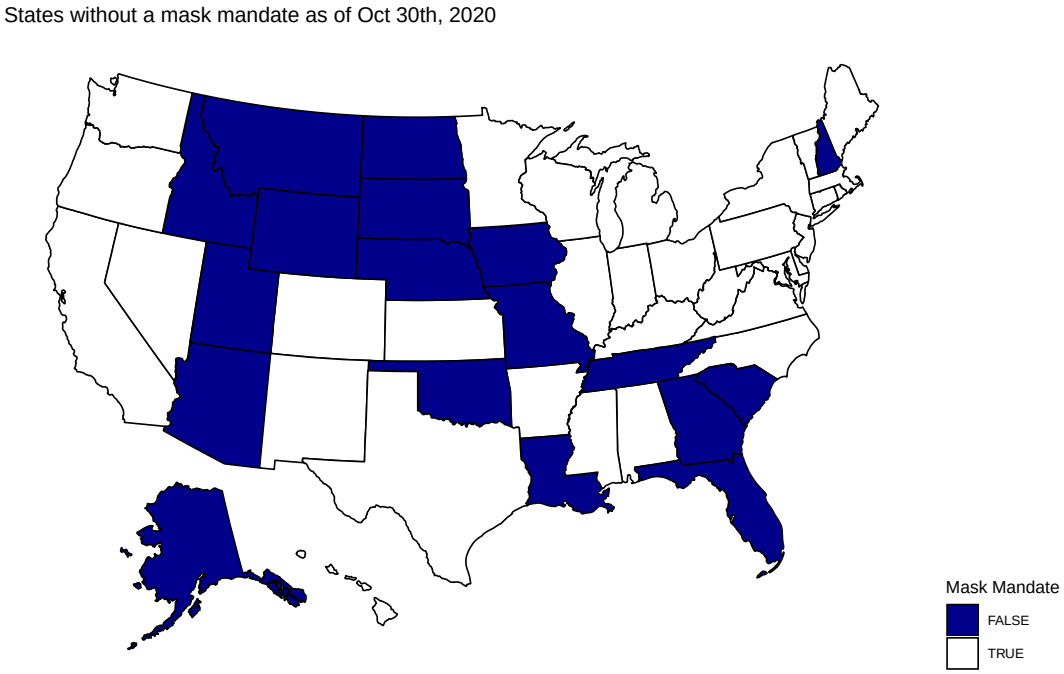
```
      percent_over_65='X65.',
      percent_at_risk='Percent.at.risk.for.serious.illness.due.to.COVID',
      fips = "State.FIPS.Code"
      )

covid_data$repgov <- grepl("(R)",covid_data$Governor)
covid_data$mask_mandate_all <- ifelse(or(covid_data$mask_for_all_mandated_on == 0,
```

Initial Exploratory Data Analysis (EDA)

Layout of the land

Here is the mask mandate status of all 50 states. We're omitting an NA generated by the `usmap` for the District of Columbia.

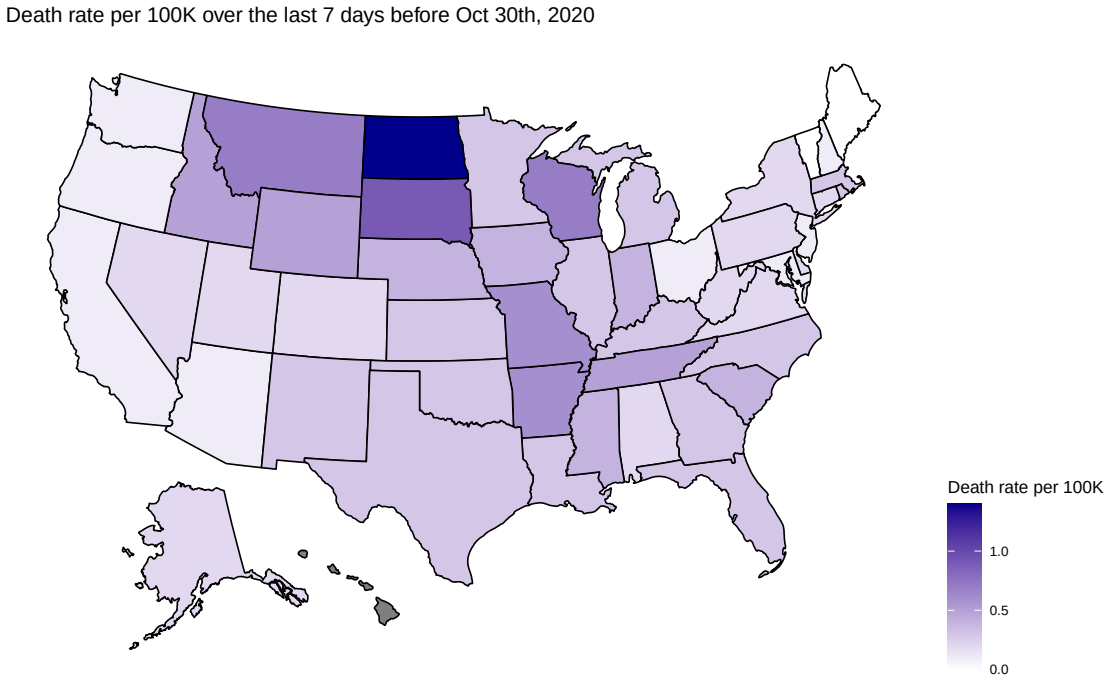
```
plot1 = plot_usmap(data = covid_data, values = "mask_mandate_all") +
        scale_fill_manual(values = c("TRUE" = "white", "FALSE" = "dark blue"),
                          na.translate = F,
                          name = "Mask Mandate") +
        theme(legend.position = "right") +
        labs(title = "States without a mask mandate as of Oct 30th, 2020")

plot1
```



Now let's look at the population death rate in the past 7 days before October 30th.

```
plot2 = plot_usmap(data = covid_data, values = "death_rate_in_last7") +
        scale_fill_continuous(low = "white", high = "dark blue",
                              name = "Death rate per 100K") +
```

```
        theme(legend.position = "right") +
        labs(title = "Death rate per 100K over the last 7 days before Oct 30th, 2020")

plot2
```



Potential data issues

Let's look to the missing values. Here they are by state.

```
subset(data.frame("state" = covid_data$State,
                  "NA_count" = apply(is.na(covid_data),1,sum)),NA_count > 0)
```

```
##            state NA_count
## 12        Hawaii        7
## 19     Louisiana        1
## 27       Montana        3
## 32    New Mexico        2
## 33      New York        4
## 35  North Dakota        5
## 46       Vermont        1
## 49 West Virginia        2
```

And here they are by variable.

```
colnames(covid_data)[colSums(is.na(covid_data)) > 0]
```

```
##  [1] "White...of.Cases"
##  [2] "Black...of.Cases"
```

```
##  [3] "Hispanic...of.Cases"
##  [4] "Other...of.Cases"
##  [5] "White...of.Deaths"
##  [6] "Black...of.Deaths"
##  [7] "Hispanic...of.Deaths"
##  [8] "Other...of.Deaths"
##  [9] "State.Abbreviation"
## [10] "fips"
## [11] "mask_enforced_by_fines"
## [12] "mask_enforced_by_charge"
## [13] "Attempt.by.state.government.to.prevent.local.governments.from.implementing.face.mask.orders"
```

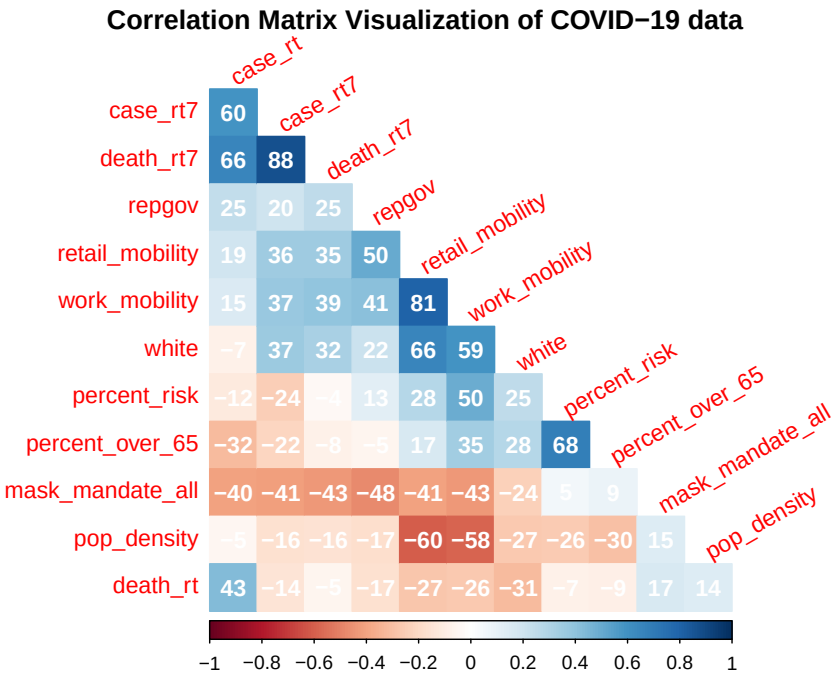
We are not looking into any of these variables and they mostly seems related to racial reporting of cases and deaths as some states do not report those numbers. But it is not a big concern.

Plot 2 above has already shown that our outcome variable has a reasonable range of values.

Variables

Let us take a look at a correlation table of a number of interesting variables.

```
covid_corr <- covid_data[,c("State","case_rate", "case_rate_in_last7", "death_rate",
                            "death_rate_in_last7", "white_percent", "percent_over_65",
                            "percent_at_risk", "population_density",
                            "mask_mandate_all", "workplaces_mobility_change",
                            "retail_mobility_change", "repgov")]

covid_corr <- covid_corr  %>%
  rename(
          case_rt = "case_rate",
          case_rt7 = "case_rate_in_last7",
          death_rt = "death_rate",
          death_rt7 = "death_rate_in_last7",
          white = "white_percent",
          pop_density = "population_density",
          work_mobility = "workplaces_mobility_change",
          retail_mobility = "retail_mobility_change",
          percent_risk = "percent_at_risk"
  )

cor_mat <- cor(covid_corr[,c("case_rt", "case_rt7", "death_rt", "death_rt7", "white",
                             "percent_over_65","percent_risk", "pop_density",
                             "mask_mandate_all", "work_mobility",
                             "retail_mobility", "repgov")])

corrplot(cor_mat, method = "color", order = "AOE", type = "lower",
         diag=FALSE, addCoef.col = "white", tl.srt = 30,
         addCoefasPercent = TRUE, mar = c(0, 0, 1, 0),
         title = "Correlation Matrix Visualization of COVID-19 data")
```

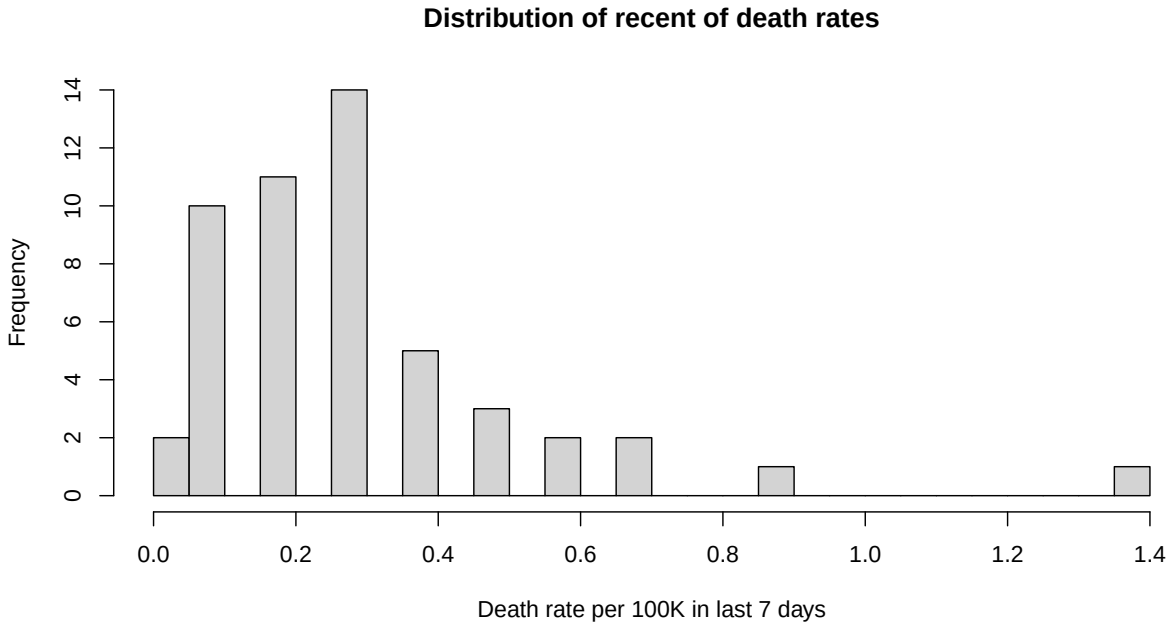



```
#mtext("Correlation Matrix Visualization of Mask Mandates", side=3, adj=0, cex=1.0)
```

An interesting pattern that emerges from the above correlation table is that the overall death rate and the death rate over the last 7 days have opposing relationships with a number of variables such as mobility, mask mandates and white population percentage. That seems to be indicative of something we've suspected in approaching the research question. Perhaps the states that suffered the worst of the first wave are not the states that are suffering now. And while in the first wave measures would have come too late to save the victims of the initial outburst of COVID-19, now the picture is changed.

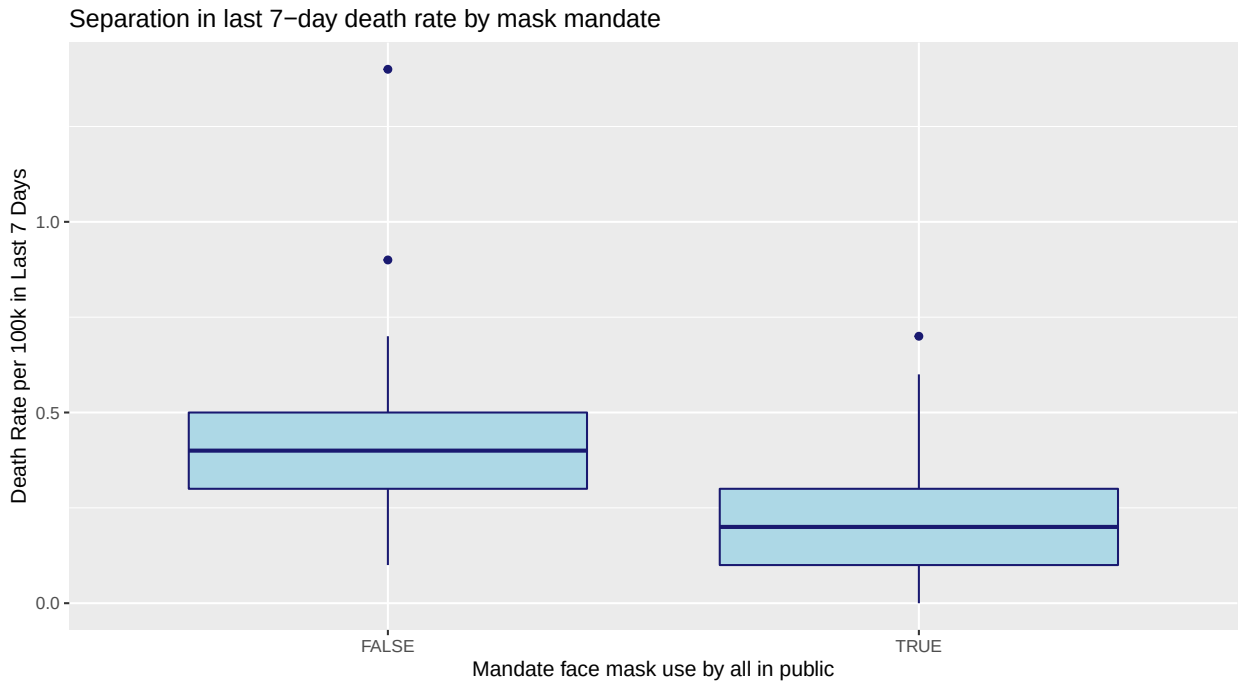
So let's focus on the variable *Death Rate per 100K in the Last 7 Days*.

```
hist(covid_data$death_rate_in_last7, breaks = 20,
     main = "Distribution of recent of death rates",
     xlab = "Death rate per 100K in last 7 days")
```



There is some skew, but perhaps we can argue that it's not too severe. Let's see how having it as an output variable looks against the mask mandate variable.

```
covid_data %>%
  ggplot(aes(x = mask_mandate_all, y = death_rate_in_last7)) +
  geom_boxplot(fill = 'lightblue', color = 'midnightblue') +
  labs(
    title = 'Separation in last 7-day death rate by mask mandate',
    x = 'Mandate face mask use by all in public',
    y = 'Death Rate per 100k in Last 7 Days'
  )
```



There seems to be a separation that would make using this relationship for our first model interesting.

Model 1

Our first model will only look at the death rate in past 7 days per 100k people and a boolean indicator on whether or not mandates for face masks are in place:

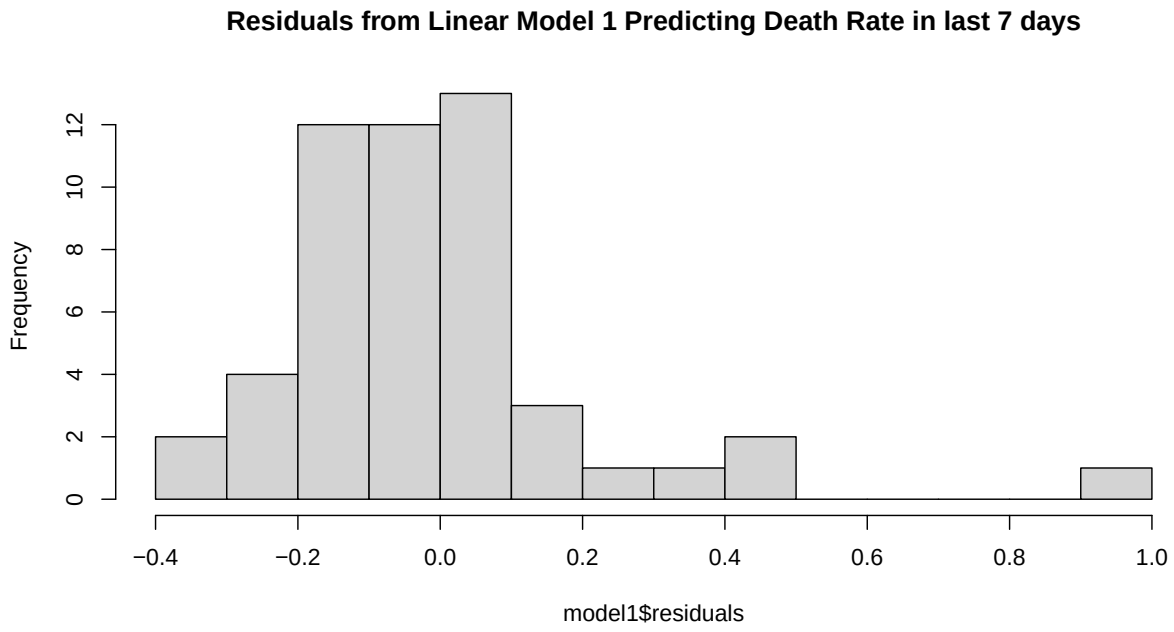
$$death\_rate_7 = \beta_0 + \beta_1 mandate + w \qquad \text{(Model 1)}$$

```
model1 <- lm(death_rate_in_last7 ~ mask_mandate_all, data = covid_data)
summary(model1)
```

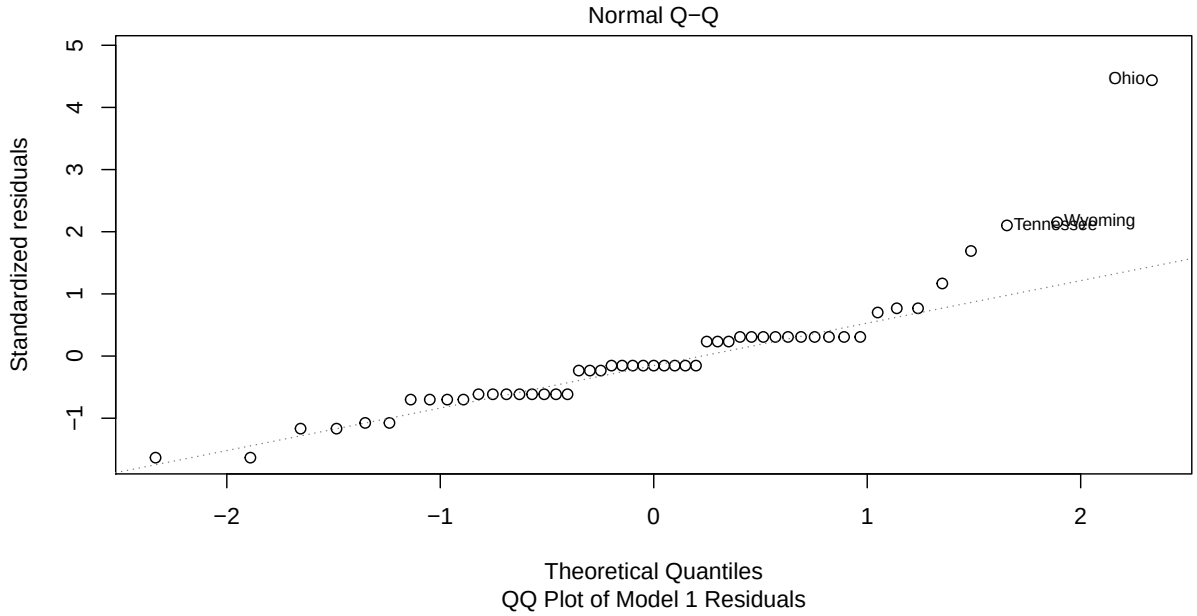
```
##
## Call:
## lm(formula = death_rate_in_last7 ~ mask_mandate_all, data = covid_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35000 -0.13333 -0.03333  0.06667  0.95000
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.45000    0.05193   8.666 1.87e-11 ***
## mask_mandate_allTRUE -0.21667    0.06456  -3.356  0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2203 on 49 degrees of freedom
## Multiple R-squared:  0.1869, Adjusted R-squared:  0.1703
## F-statistic: 11.26 on 1 and 49 DF,  p-value: 0.001533
```

Here we can see the coefficient of the mask_mandates is statistically significant with a p-value less than 0.01. Practically speaking, it says that having a face mask mandate for all is associated with a reduction of 0.21 deaths per 100k people per 7 days, or ~2 deaths per 1 million people per 7 days.

```
hist(model1$residuals, breaks = 10,
     main = "Residuals from Linear Model 1 Predicting Death Rate in last 7 days")
```




```
plot(model1, which=2,
     sub.caption = "QQ Plot of Model 1 Residuals",
     labels.id = state.name) # QQ Plot of Residual
```



QQ Plot of Model 1 Residuals

When we plot the residuals of our first model, we can see that it isn't quite normal, with one particular state (Ohio) having a residual above 1, which is quite high given that the median is below 0.5. Then again we don't see clear evidence of violations of assumptions 2 or 4 (see assumptions discussion section).

Model 2

Since we had a human controlled variable in the first model, we will add a variable that is a population parameter, not human controllable in model 2. We do this because we think all human controlled variables might influence one another and create instability in the model. Moreover, we want to be able to control for the mask mandate and describe our outcome variable, this will not be possible if another variable that is highly correlated with the mask mandate is also in the model. We will therefore add a variable that will help us to control for both the mask mandate and the population parameter variable.

Variables being considered: *percent_over_65* (indicates the percent of people over 65 years of age in the population), *white* (indicates the percent of people in the population who identify themselves as white, non-hispanic), and *percent_at_risk* (indicates the percent of people in the population who are at risk of serious illness if infected by the Coronavirus).

Our reasons for considering these three variables are as follows:

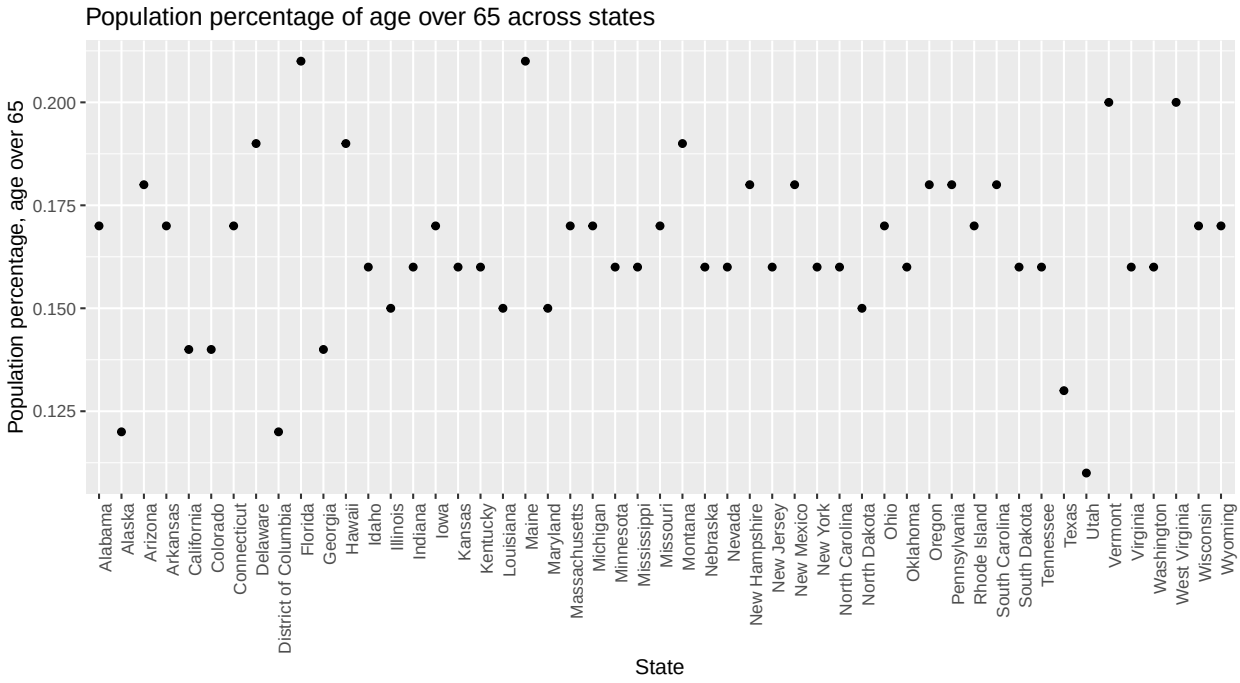
percent_over_65 - Being older has historically weakened people's immune systems and ability to fight an infection. These have resulted in higher death rate in age groups over 65 years all over the world. This could inform our death rate over the last 7 days.

white_percent - We have read from other studies that have controlled variables like income and access to insurance to show that some races have had higher death rates due to COVID-19 than others.

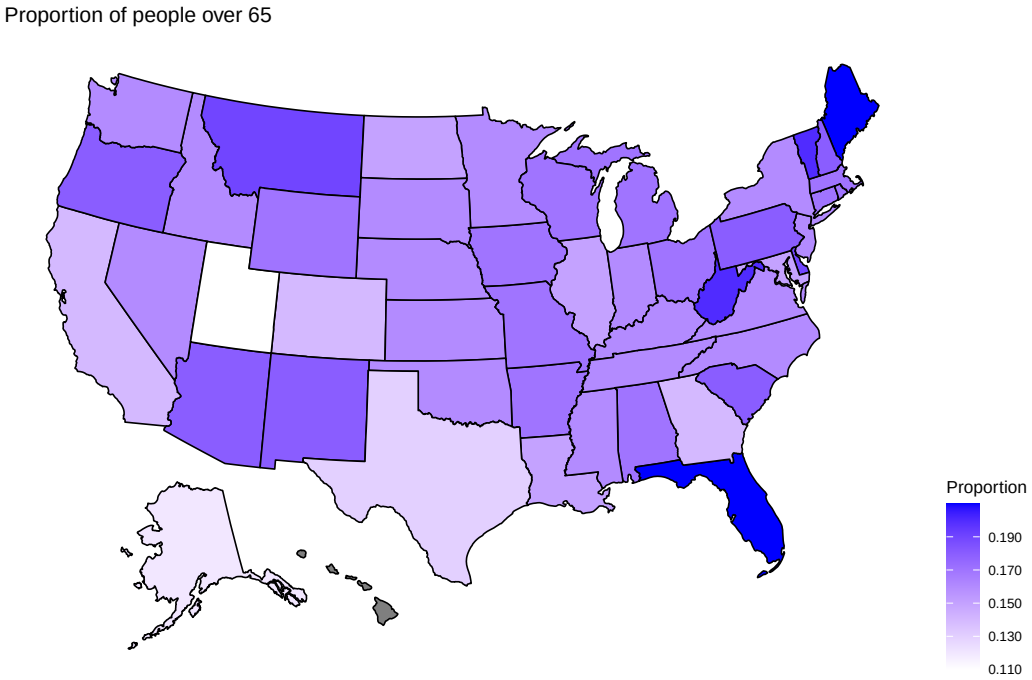
percent_at_risk - People who have an existing conditions like diabetes or heart disease have seen higher death rates due to either weaker immunity or to the additional damage to blood vessels and organs caused by COVID-19.

Distribution of *percent_over_65* across states:

```
covid_data %>%
  ggplot(aes(x = State, y = percent_over_65)) + geom_point() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(
    title = 'Population percentage of age over 65 across states',
    x = 'State',
    y = 'Population percentage, age over 65'
  )
```

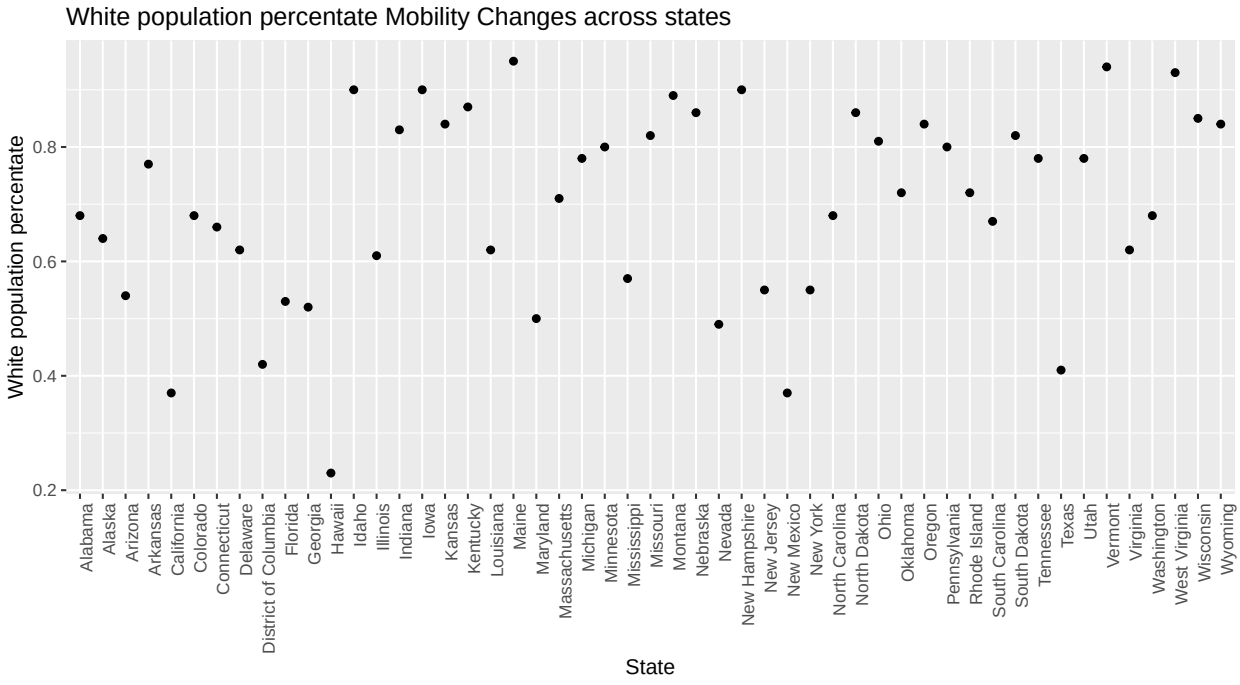



```
plot1_var2 = plot_usmap(data = covid_data, values = "percent_over_65") +
        scale_fill_continuous(low = "white", high = "blue",
                              name = "Proportion", label = scales::comma) +
        theme(legend.position = "right") +
        labs(title = "Proportion of people over 65")

plot1_var2
```

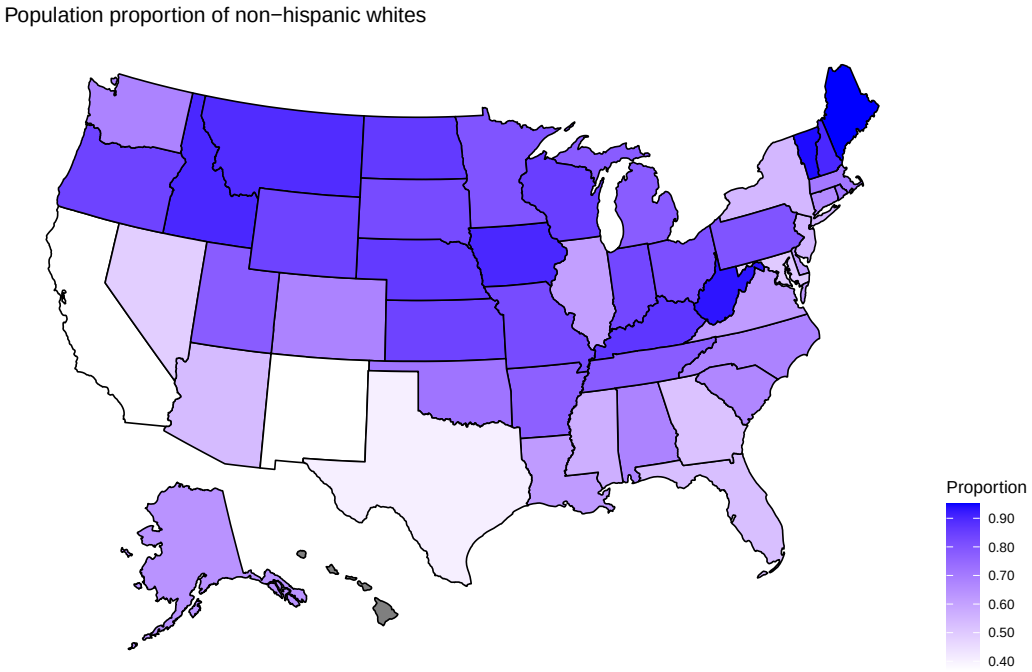


Distribution of white population percentage across states:

```
covid_data %>%
  ggplot(aes(x = State, y = white_percent)) + geom_point() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(
    title = 'White population percentate Mobility Changes across states',
    x = 'State',
    y = 'White population percentate'
  )
```

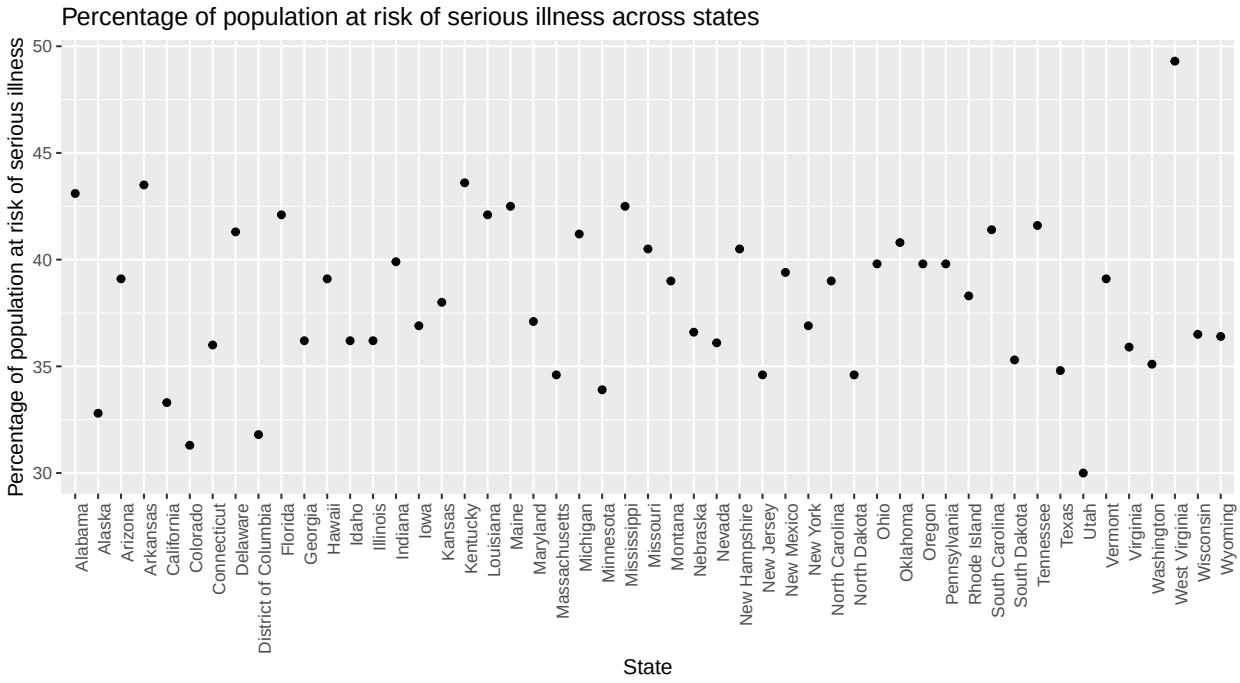



```
plot2_var2 = plot_usmap(data = covid_data, values = "white_percent") +
        scale_fill_continuous(low = "white", high = "blue",
                              name = "Proportion", label = scales::comma) +
        theme(legend.position = "right") +
        labs(title = "Population proportion of non-hispanic whites")

plot2_var2
```

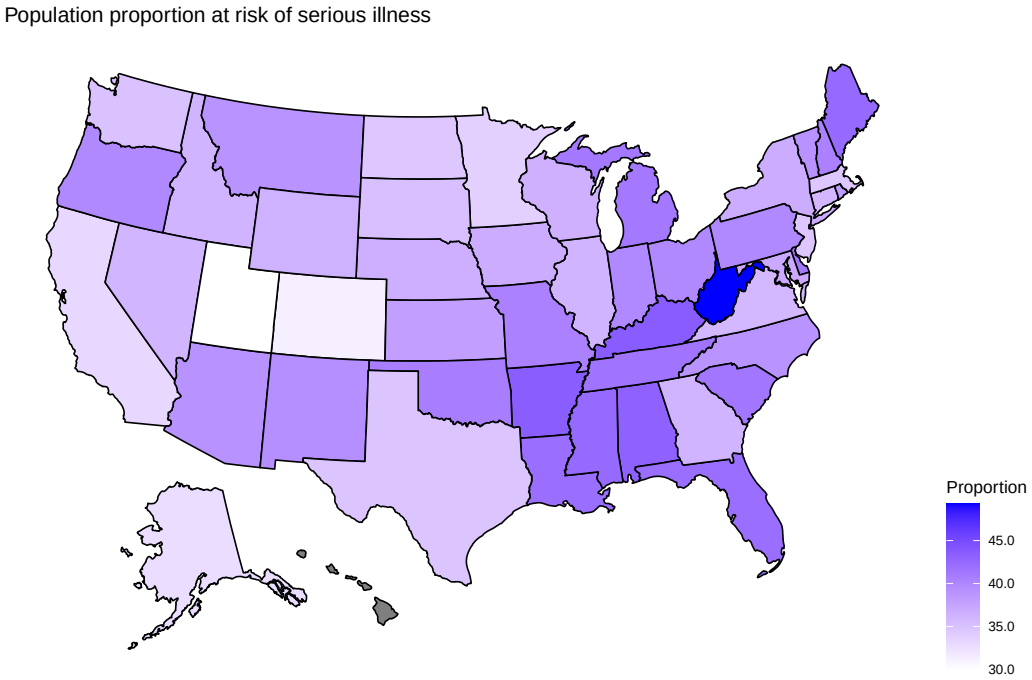


Distribution for percentage of population at risk of serious illness across states:

```r
covid_data %>%
  ggplot(aes(x = State, y = percent_at_risk)) + geom_point() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(
    title = 'Percentage of population at risk of serious illness across states',
    x = 'State',
    y = 'Percentage of population at risk of serious illness'
  )
```




```r
plot3_var2 = plot_usmap(data = covid_data, values = "percent_at_risk") +
        scale_fill_continuous(low = "white", high = "blue",
                              name = "Proportion", label = scales::comma) +
        theme(legend.position = "right") +
        labs(title = "Population proportion at risk of serious illness")

plot3_var2
```



We will use white population percentage as our additional variable in model 2 because we see variance in it over states so a change in it could indicate differences in our outcome variable.

$$death\_rate_7 = \beta_0 + \beta_1 mask\_mandate + \beta_2 white\_percent + w \qquad \text{(Model 2)}$$

```
model2 <- lm(death_rate_in_last7 ~ mask_mandate_all + white_percent, data = covid_data)
summary(model2)
```

```
##
## Call:
## lm(formula = death_rate_in_last7 ~ mask_mandate_all + white_percent,
##     data = covid_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39781 -0.09569 -0.01665  0.05203  0.91538
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.2011     0.1475   1.363  0.17927
## mask_mandate_allTRUE  -0.1888     0.0650  -2.905  0.00554 **
## white_percent          0.3297     0.1835   1.797  0.07858 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2155 on 48 degrees of freedom
## Multiple R-squared:  0.2382, Adjusted R-squared:  0.2064
## F-statistic: 7.504 on 2 and 48 DF,  p-value: 0.00146
```

Here we can see the coefficient of the mask_mandates is still statistically significant with a p-value less than 0.05. Practically speaking, it says that having a face mask mandate for all, while controlling the white

population percentage, is associated with a reduction of ~0.19 deaths per 100k people per 7 days, or ~1.9 deaths per 1 million people per 7 days. The *white_percent* is not statistically significant with a p-value greater than 0.05. Practically, its effect is that a 1% increase in the percentage of white population would result in 0.3 additional deaths per 100K people per 7 days, or additional 3 deaths per 1 million people per 7 days.

The coefficient for *white_percent* here might not be answering what we want to observe. The data is at state level, not patient, to be able to see the effect of a certain race (or any one different from it) on the death rate. We picked this variable because of background knowledge we gathered about people of color and people in minorities being hit worse by the pandemic than others. It could be reflecting another aspect, may be that more diverse states have more robust health-care systems and service providers which have helped to lower death rate. So we will be cautious about how we use this coefficient to understand our question.

Models 3, 4

Model 3.

For model 3 we will follow the approach of keeping one human controllable variable and multiple population parameter variables. This way, we will be able to control the population parameter variables and get a description of *mask_mandate_all* for the outcome variable and also get the additional information that the population parameter variables could provide in describing the death rate in last 7 days.

We will add *percent_over_65* (indicates the percent of people over 65 years of age in the population) and *percent_at_risk* (indicates the percent of people in the population who are at risk of serious illness if infected by the Coronavirus). Both are population parameters that are not controllable by humans.

$$death\_rate_7 = \beta_0 + \beta_1 mask\_mandate + \beta_2 white\_percent + \beta_3\ percent\_over\_65 + \beta_4\ percent\_at\_risk + u$$
(Model 3)

```
model3 <- lm(death_rate_in_last7 ~ mask_mandate_all + white_percent +
               percent_over_65 + percent_at_risk, data = covid_data)
summary(model3)
```

```
##
## Call:
## lm(formula = death_rate_in_last7 ~ mask_mandate_all + white_percent +
##     percent_over_65 + percent_at_risk, data = covid_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38039 -0.09654 -0.00989  0.05498  0.89110
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.3998729  0.3270654   1.223   0.2277
## mask_mandate_allTRUE -0.1787083  0.0667688  -2.677   0.0103 *
## white_percent          0.3856435  0.1963000   1.965   0.0555 .
## percent_over_65       -1.3524669  2.0807045  -0.650   0.5189
## percent_at_risk       -0.0005709  0.0115145  -0.050   0.9607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.2182 on 46 degrees of freedom
## Multiple R-squared:  0.2515, Adjusted R-squared:  0.1864
## F-statistic: 3.864 on 4 and 46 DF,  p-value: 0.008672
```

The mask mandate for all still remains statistically significant with a slight increase in p-value but still under 0.05. It's practical value in this model is that having a face mask mandate for all, while controlling all the other population parameter variables, is associated with a reduction of ~0.18 deaths per 100k people per 7 days, or ~1.8 deaths per 1 million people per 7 days. The population parameter variables in the model are all statistically not significant with surprising (still not significant) descriptive effects of percentage of population with age over 65 and percentage of population at risk reducing deaths per 100K in 7 days. The lack of significance in *percent_over_65* and *percent_at_risk* could be because of high correlation between them creating instability and not informing us well of either one's effect on the outcome variable. In the next model, we will drop one of these two variables while adding others that help us describe death rate in last 7 days better.

Surprisingly, the coefficient for *percent_over_65* is negative indicating that more people over 65 in the population would mean lesser death rate in the last 7 dats due to Covid-19. While this could be true, we must again view this with caution because it could not be answering what we are trying to understand. The reason for this discrepancy would be the same as the one we observed in the coefficient for *white_percent* variable, the data is not granular enough to see at patient level and the state level data might be reflecting an effect that is indirectly causing higher population of people over 65 to reduce death rate in the last 7 days.

Model 4.

For model 4 we will add an additional human controllable variable since we already saw the low/none significance that population parameter variables held while describing the death rate in last 7 days. We will add *workplaces_mobility_change*.

We choose work mobility here because we think that is one that can be acted upon that could have an impact. On reading about factors of social closeness that matter for COVID-19 transmission or the intensity of the infection caught by someone, a major one is the viral load. This is higher when people are spend long periods of time in closed areas with those carrying the coronavirus. In parks, we are not in closed areas, in retail, there isn't an extended period of time spent since people are coming in and going out more frequently than at work. Grocery and pharmacies are necessities so they cannot be changed much and transit and residential mobility changes are very particular to each city so coming up with a blanket policy or plan to tackle those might be difficult.

$$death\_rate_7 = \beta_0 + \beta_1 mask\_mandate + \beta_2 white\_percent + \beta_3\ percent\_over\_65 + \beta_4 workplaces\_mobility\_change + v \quad \text{(Model 4)}$$

```
model4 <- lm(death_rate_in_last7 ~ mask_mandate_all + white_percent  + percent_over_65
             + workplaces_mobility_change, data = covid_data)
summary(model4)
```

```
##
## Call:
## lm(formula = death_rate_in_last7 ~ mask_mandate_all + white_percent +
##     percent_over_65 + workplaces_mobility_change, data = covid_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35160 -0.09742 -0.00505  0.05174  0.93665
##
```

```
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.814215   0.407060   2.000   0.0514 .
## mask_mandate_allTRUE        -0.137230   0.072593  -1.890   0.0650 .
## white_percent                0.239060   0.221018   1.082   0.2851
## percent_over_65             -2.125157   1.632855  -1.301   0.1996
## workplaces_mobility_change  0.008489   0.006390   1.328   0.1906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2141 on 46 degrees of freedom
## Multiple R-squared:  0.2791, Adjusted R-squared:  0.2164
## F-statistic: 4.452 on 4 and 46 DF,  p-value: 0.003995
```

We lose statistical significance for *mask_mandate_all* indicating that some of it's effect may be shared with the newly added variable *workplaces_mobility_change*. We can also see that the significance of the *percent_over_65* variable has increase, still not to a statistically significant level. The correlation between *workplaces_mobility_change* and *white_percent* also seems to have affected the significance that white_percent's coefficient held in the previous model. It has introduced another instability to the model coefficients. Practically, we see a small increase in death rate for every unit of increase in *workplaces_mobility_change* which aligns with our background information - the more people will spend time with others in closed spaces, the more their chances of contracting the virus and that at a higher viral load leading to higher severity of the disease and higher chances of death due to it.

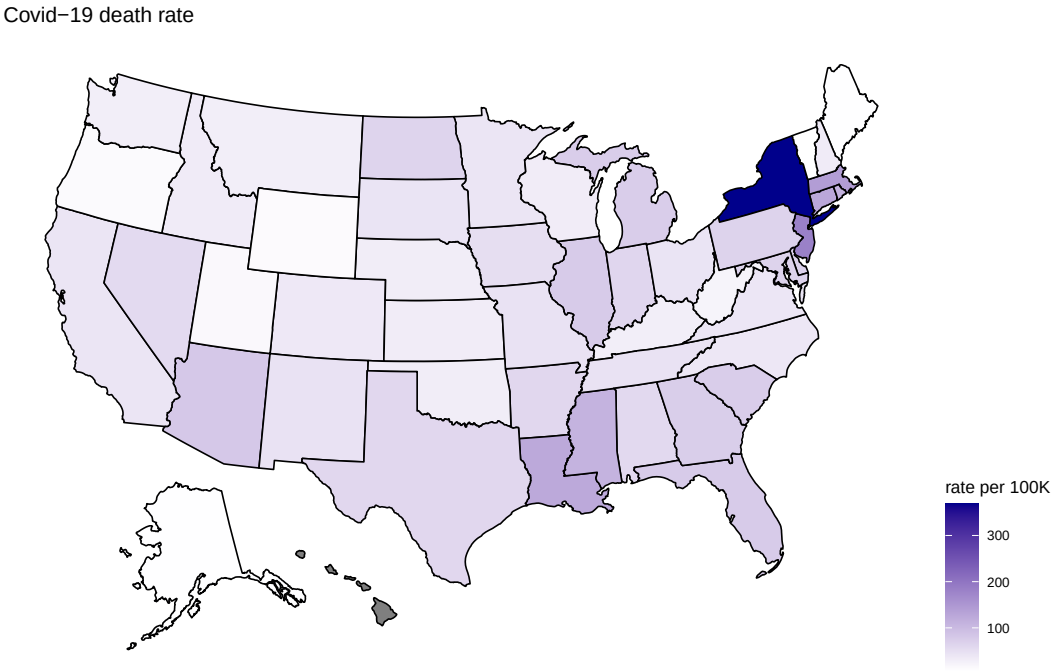
All the coefficients are statistically insignificant, so we're failing to reject the null hypothesis that all these coefficients are 0. That puts us in a difficult situation as it is on the one hand possible that COVID-19 is driven by randomness (such as occurrence of superspreading events). But it is also very possible that there is an impact on the coefficients and their significance due to correlations between the explanatory variables. Let's see if keeping only one of the population attributes: percentage of population over 65 years of age can give us a better description of death rate in 100K in the last 7 days. In terms of the practical significance if we can speak of such thing given the p-values, only *percent_over_65* seems to have a more sizeable coefficient.

Limitations Our Model

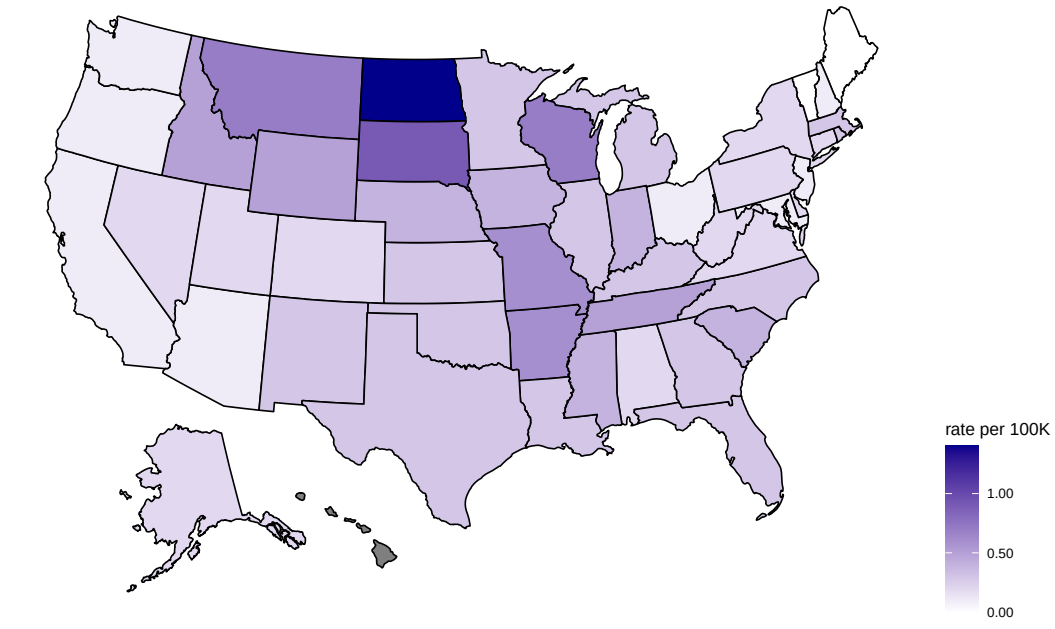
1. IID The question here is whether the observations for each different state are IID. One way that assumption might be challenged is because of some type of clustering of conditions in neighboring states for example. While that is a danger and of course some similarities do exist, it seems valid to argue that the connections between states are not such that one state can truly or easily influence the course of the epidemic in another state, either directly through policy (since all state policy is the domain of the local state government) or through demographic factors. That said perhaps travel between well-connected neighboring states can be a source of contagion (e.g. NY/NJ), so we need to be wary, as this impacts the "independent" portion of IID.

```
plot3 = plot_usmap(data = covid_data, values = "death_rate") +
        scale_fill_continuous(low = "white", high = "dark blue",
                              name = "rate per 100K", label = scales::comma) +
        theme(legend.position = "right") +
        labs(title = "Covid-19 death rate")

plot4 = plot_usmap(data = covid_data, values = "death_rate_in_last7") +
        scale_fill_continuous(low = "white", high = "dark blue",
                              name = "rate per 100K", label = scales::comma) +
```

```
        theme(legend.position = "right") +
        labs(title = "Covid-19 death rate in the last 7 days")

plot3
```




```
plot4
```



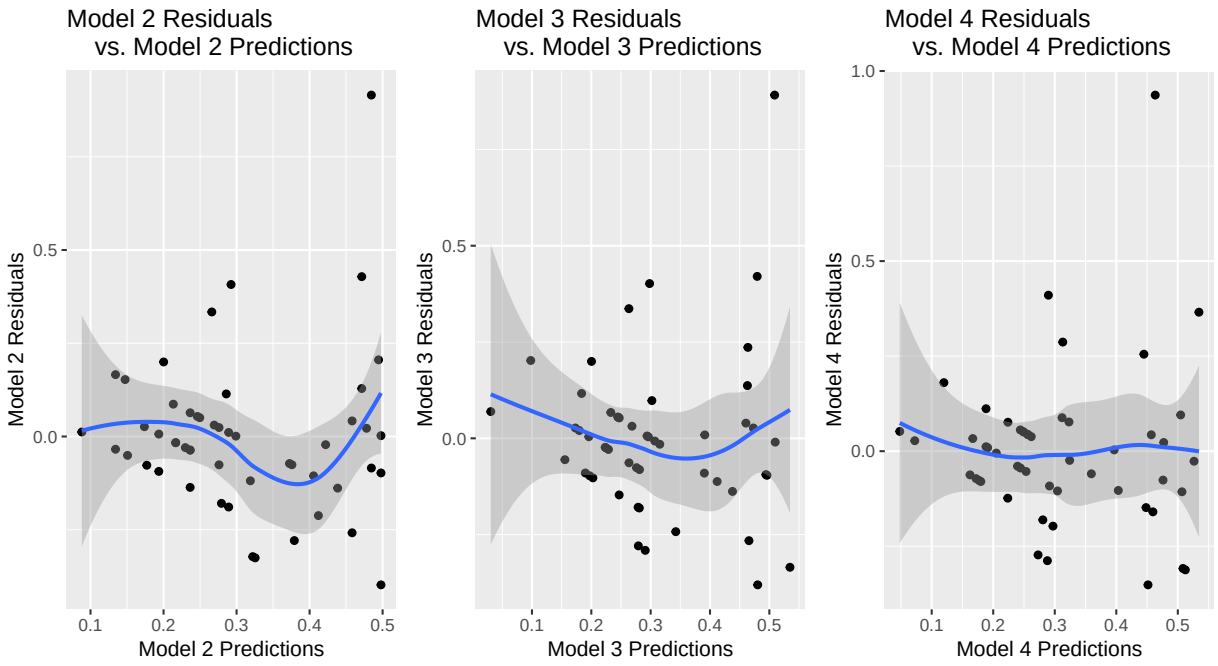
2. Linear Conditional Expectation This is an issue that is very specific to the variables that we're choosing.

Ideally, we'd be looking at the residuals of our models. One potential factor to worry about is the exponential nature of epidemics. That might mean that since the timing of the epidemic might be different in different states, we might be looking at different points on an exponential curve.

Judging by the stargazer table, we conclude that model 4 is our most robust model. Because our first model only has one TRUE/FALSE variable, the residuals versus prediction plot isn't useful.

```
covid_data <- covid_data %>%  mutate(model2_residuals = resid(model2))
covid_data <- covid_data %>%  mutate(model2_predictions = predict(model2))
covid_data <- covid_data %>%  mutate(model3_residuals = resid(model3))
covid_data <- covid_data %>%  mutate(model3_predictions = predict(model3))
covid_data <- covid_data %>%  mutate(model4_residuals = resid(model4))
covid_data <- covid_data %>%  mutate(model4_predictions = predict(model4))

plot_model_2 <- covid_data %>%
  ggplot(aes(x = model2_predictions, y = model2_residuals)) +
  geom_point() + stat_smooth(se=TRUE) +
  labs(
    title = 'Model 2 Residuals
    vs. Model 2 Predictions',
    x = 'Model 2 Predictions',
    y = 'Model 2 Residuals'
  )

plot_model_3 <- covid_data %>%
  ggplot(aes(x = model3_predictions, y = model3_residuals)) +
  geom_point() + stat_smooth(se=TRUE) +
  labs(
    title = 'Model 3 Residuals
    vs. Model 3 Predictions',
    x = 'Model 3 Predictions',
    y = 'Model 3 Residuals'
  )

plot_model_4 <- covid_data %>%
  ggplot(aes(x = model4_predictions, y = model4_residuals)) +
  geom_point() + stat_smooth(se=TRUE) +
  labs(
    title = 'Model 4 Residuals
    vs. Model 4 Predictions',
    x = 'Model 4 Predictions',
    y = 'Model 4 Residuals'
  )


plot_model_2 | plot_model_3 | plot_model_4
```

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



By the occular test, we see that model 4 is the most linear between the predictions and residuals. We will need to look at the other variables in the models versus the residuals to find where the most noise in our chosen model lies.

```
plot_1 <- covid_data %>%
  ggplot(aes(x = workplaces_mobility_change, y = model4_residuals)) +
  geom_point() + stat_smooth(se=TRUE) +
  labs(
    title = 'Model Residuals vs. Workplace Mobility',
    x = 'Workplace Mobility',
    y = 'Model Residuals'
  )

plot_2 <- covid_data %>%
  ggplot(aes(x = percent_over_65, y = model4_residuals)) +
  geom_point() + stat_smooth(se=TRUE) +
  labs(
    title = 'Model Residuals vs. Percent over 65',
    x = 'Percent over 65',
    y = 'Model Residuals'
  )

plot_3 <- covid_data %>%
  ggplot(aes(x = white_percent, y = model4_residuals)) +
  geom_point() + stat_smooth(se=TRUE) +
  labs(
    title = 'Model Residuals vs. White Percent',
    x = 'White Percent',
    y = 'Model Residuals'
  )

plot_4 <- covid_data %>%
```

```
  ggplot(aes(x = model4_predictions, y = model4_residuals)) +
  geom_point() + stat_smooth(se=TRUE) +
  labs(
    title = 'Model Residuals vs. Model Predictions',
    x = 'Model Predictions',
    y = 'Model Residuals'
  )

(plot_1 | plot_2) /
(plot_3 | plot_4)
```

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

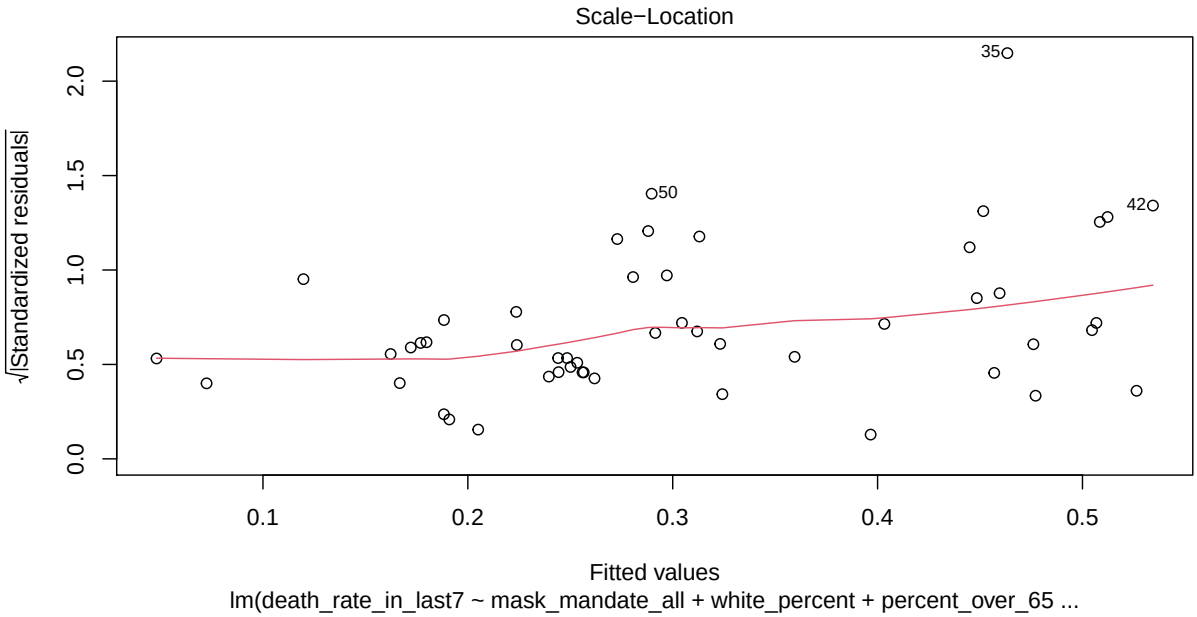
3. No perfect Collinearity The correlation table in our EDA gives us a clear indication that we're safe in that regard. While groups of variables have clear relationships (such as the different mobility categories), none of them are perfectly colinear.

4. Homoskedastic Conditional Errors This assumption can be affected by issues with assumption 2. If there is a problem with 2, it is likely to show up here as well. Another issue could be skewness of the data. We see in the EDA that there is a bit of skew in the distribution of our outcome variable. Here it seems to not be too bad and we are somewhat safer as we are looking at the death rate over 7 days. Never-the-less, if as we mentioned in 2, there are major discrepancies between states in terms of the exponential growth of the epidemic, there could be orders of magnitude differences in rates and that could be a significant issue. Helpful tools could be transforming the variable (taking a log for example) as well as using robust errors.

```
lmtest::bptest(model4)
```

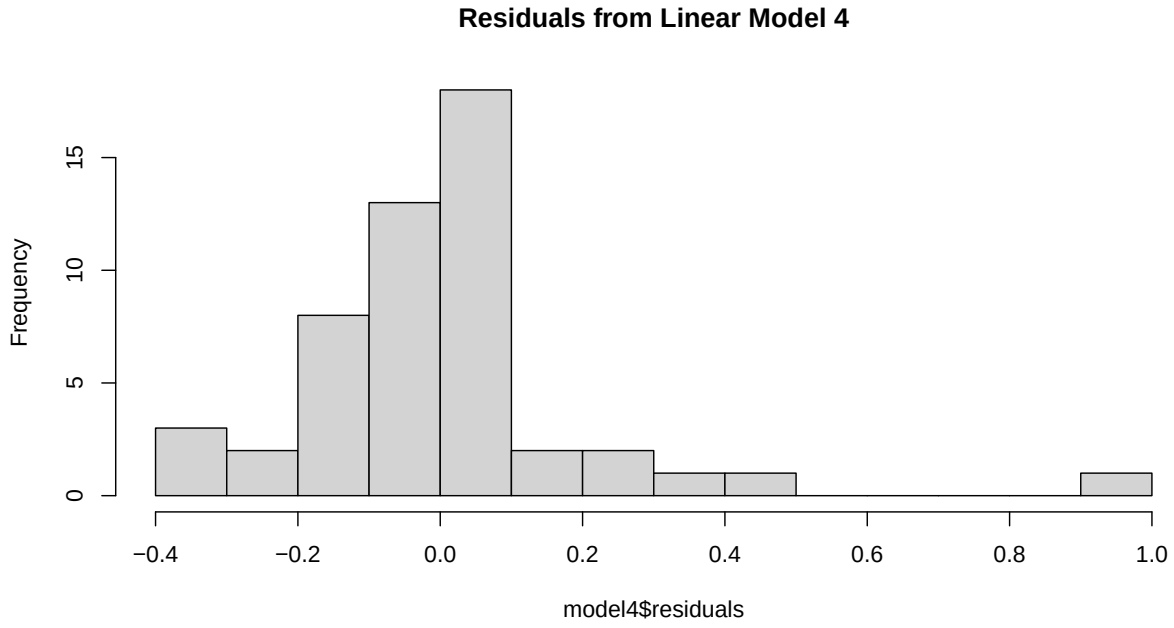
```
##
##  studentized Breusch-Pagan test
##
## data:  model4
## BP = 6.4808, df = 4, p-value = 0.166
```

```
plot(model4, which=3)
```

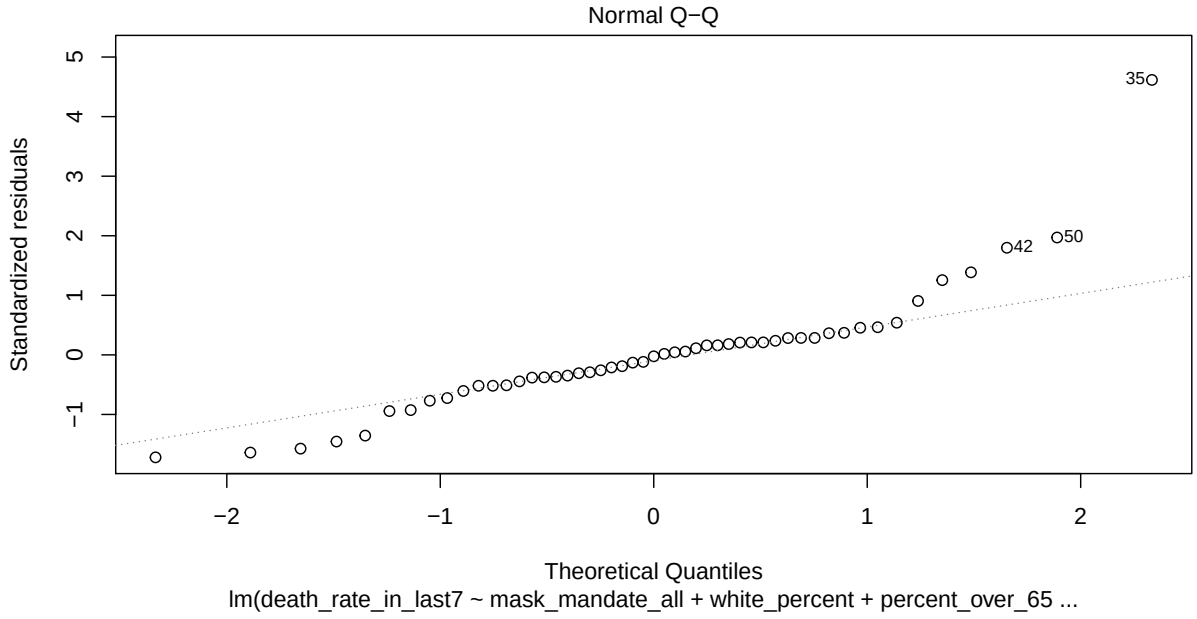


5. Normality of the errors According to the Gauss-Markov Theorem under assumptions 1-4 we still have OLS as the best linear unbiased estimator. How much should we worry about assumption number 5. Well, we will be updating this section as soon as the async catches up with the lesson on assumption 5. Meanwhile we are running some Q-Q plots of our residuals and if they seem sufficiently particularly away from normality, we'll have to consider the implications in terms of choosing our variables.

```
hist(model4$residuals, breaks = 10,
     main = "Residuals from Linear Model 4") # Histogram of the Residuals
```




```
plot(model4, which=2) # QQ Plot of Residuals
```



Regression Table

```
stargazer(model1,model2, model3, model4,
          type="latex",
          se = list( sqrt(diag(vcovHC(model1))),sqrt(diag(vcovHC(model2))) ,sqrt(diag(vcovHC(model3)))),
          column.labels = c("model1","model2","model3","model 4"))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Dec 07, 2020 - 05:24:52 PM

Table 1:

	Dependent variable:			
	death_rate_in_last7			
	model1	model2	model3	model 4
	(1)	(2)	(3)	(4)
mask_mandate_all	−0.217***	−0.189**	−0.179**	−0.137*
	(0.080)	(0.076)	(0.072)	(0.073)
white_percent		0.330*	0.386*	0.239
		(0.175)	(0.200)	(0.221)
percent_over_65			−1.352	−2.125
			(2.019)	(1.633)
percent_at_risk			−0.001	
			(0.011)	
workplaces_mobility_change				0.008
				(0.006)
Constant	0.450***	0.201	0.400	0.814*
	(0.076)	(0.124)	(0.384)	(0.407)
Observations	51	51	51	51
R^2	0.187	0.238	0.251	0.279
Adjusted R^2	0.170	0.206	0.186	0.216
Residual Std. Error	0.220 (df = 49)	0.215 (df = 48)	0.218 (df = 46)	0.214 (df = 4
F Statistic	11.265*** (df = 1; 49)	7.504*** (df = 2; 48)	3.864*** (df = 4; 46)	4.452*** (df = 4
Note:				*p<0.1; **p<0.05; ***p<

We see adding more variables decreases the residual standard errors, but not by a large amount. We also see that Adjusted R2 increases a bit from model1 to model2, decreases from model2 to model3, and then increases again in model4. So it indicates that the additional variables in model3 did not help much with the increase fit of the model to the data but those of model4 were helpful. The residuals convey that there are explanations we have not covered in our models that can describe the death rate in 100K in the last 7 days better. We will evaluate some omitted variables that might inform and improve the model.

Discussion of Omitted Variables

Mask Adoption

1. The first variable we want to consider as one that could be introducing omitted variable bias is *mask_adoption*. Mask Adoption will be a variable that reflects what percentage of the population is wearing masks and to what extent or in what capacities (all day, when they go to populous places, when they are around anybody, etc.). Mask Adoption is correlated with Mask Mandates as those states with mandates will expect to see higher mask adoption than states without mask mandates. Mask Adoption can also be a determinant of the death rate in 100K in the last 7 days as masks reduce potential exposure risk from an infected person whether they have symptoms or not and lower exposure, lower case rate would have an impact on the death rate.

Estimated model:

$$death\_rate_7 = \tilde{\beta}_0 + \tilde{\beta}_1 mask\_mandate\_all + w$$

Actual model:

$$death\_rate_7 = \beta_0 + \beta_1 mask\_mandate\_all + \beta_2 mask\_adoption + \omega$$

Secondary model:

$$mask\_mandate\_all = \delta_0 + \delta_1 mask\_adoption + a$$

Direction of Bias: With β_2 expected to be negative (i.e. higher adoption would be associated with a decrease in death rate in 100K in the last 7 days) and *mask_adoption* and *mask_mandate_all* to be positively correlated, the actual coefficient of mandate will be lesser negative than expected and the direction of bias will be away from zero.

$$\tilde{\beta}_1[-ve] = \beta_1 + \delta_1[+ve] * \beta_2[-ve]$$

Work from Home Availablility

2. The second variable we want to consider as one that could be introducing omitted variable bias is *work_from_home_availability*. This variable will reflect what percentage of the population has the option or has a mandate to work from home. 'Work from home availability' is correlated with *workplaces_mobility_change* and can also be a determinant of the death rate in 100K in the last 7 days as it would reduce the number of people who could possibly be spending extended hours in a close setting with the virus carriers, increasing viral load, and therefore severity.

Estimated model:

$$death\_rate_7 = \tilde{\beta}_0 + \tilde{\beta}_1 workplace\_mobility\_changes + w$$

Actual model:

$$death\_rate_7 = \beta_0 + \beta_1 workplace\_mobility\_changes + \beta_2 work\_from\_home\_availability + \omega$$

Secondary model:

$$workplaces\_mobility\_change = \delta_0 + \delta_1 work\_from\_home\_availability + b$$

Direction of Bias: With β_2 negative (i.e. increasing work from home option will be associated with decrease in death rate) and *work_from_home_availability* to be negatively correlated with *workplaces_mobility_change* (i.e. increasing work from home options will decrease workplace mobility), the actual coefficient of mandate will be lesser positive than expected, possibly even negative since the coefficient of *workplaces_mobility_change* is very close to 0 in model4. This indicates that the the direction of bias will be away zero.

$$\tilde{\beta}_1[+ve] = \beta_1 + \delta_1[-ve] * \beta_2[-ve]$$

Diabetic Population

3. The third variable we want to consider as one that could be introducing omitted variable bias is *diabetic_percent*. This variable will reflect what percentage of the population has diabetes. 'Type 2 Diabetic population percentage' is correlated with *white_percent* as type 2 diabetes is a condition more often observed common in non-white populations and it can also be a determinant of the death rate in 100K in the last 7 days as it has known to introduce co-morbidities in patients of COVID-19.

Estimated model:

$$death\_rate_7 = \tilde{\beta}_0 + \tilde{\beta}_1 white\_percent + w$$

Actual model:

$$death\_rate_7 = \beta_0 + \beta_1 white\_percent + \beta_2 diabetes\_percent + \omega$$

Secondary model:

$$white\_percent = \delta_0 + \delta_1 diabetes\_percent$$

Direction of Bias: With β_2 positive (i.e. higher population of people with diabetes associated with higher death rate) and *diabetic_percent* negatively correlated with *white_percent*, the actual coefficient of mandate will be more positive than expected and the direction of bias will be toward zero.

$$\tilde{\beta}_1[+ve] = \beta_1 + \delta_1[-ve] * \beta_2[+ve]$$

Donald Trump Supporters

4. The fourth variable we want to consider as one that could be introducing omitted variable bias is *percent_supporting_Donald_Trump*. This variable will be the percentage of the population that support Donald Trump. This will be correlated with the *workplaces_mobility_change* as reluctance on Donald Trump's part to pay heed to COVID-19 is reflected in his supporters' actions as well. Also, states with population that support the republican party took COVID-19 more lightly from the beginning and their lower concern or guard could heighten death rates.

Estimated model:

$$death\_rate_7 = \tilde{\beta}_0 + \tilde{\beta}_1 workplaces\_mobility\_changes + w$$

Actual model:

$$death\_rate_7 = \beta_0 + workplaces\_mobility\_changes + \beta_2 percent\_supporting\_Donald\_Trump + \omega$$

Secondary model:

$$workplaces\_mobility\_change = \delta_0 + percent\_supporting\_Donald\_Trump$$

Direction of Bias: With β_2 positive (i.e. higher percentage of Donald Trump supporters associated with higher death rates) and positive correlation of *percent_supporting_Donald_Trump* with *workplaces_mobility_change* (higher percentage of Donald Trump supporters associated with lesser negative changes in work mobility), the actual coefficient of mandate will be lesser positive than expected and the direction of bias will be away from zero.

$$\tilde{\beta}_1[+ve] = \beta_1 + \delta_1[+ve] * \beta_2[+ve]$$

Population with Heart Disease

5. The fifth variable we want to consider as one that could be introducing omitted variable bias is *heart_disease_percent*. This variable will reflect what percentage of the population has heart disease. 'Percentage of population with heart disease' is correlated with *white_percent* as it is more common in non-white minority populations and it can also be a determinant of the death rate in 100K in the last 7 days as it has known to introduce co-morbidities in patients of COVID-19.

Estimated model:

$$death\_rate_7 = \tilde{\beta}_0 + \tilde{\beta}_1 white\_percent + w$$

Actual model:

$$death\_rate_7 = \beta_0 + \beta_1 white\_percent + \beta_2 population\_with\_heart\_disease + \omega$$

Secondary model:

$$white\_percent = \delta_0 + \delta_1 heart\_disease\_percent$$

Direction of Bias: With β_2 positive (i.e. higher population of people with heart disease associated with higher death rate) and *heart_disease_percent* negatively correlated with *white_percent*, the actual coefficient of mandate will be more positive than expected and the direction of bias will be toward zero.

$$\tilde{\beta}_1[+ve] = \beta_1 + \delta_1[-ve] * \beta_2[+ve]$$

Conclusion

The question of the impact of state-level policy choices on the mortality rates from COVID-19 is of utmost societal and even political importance in the United States. Our analysis approaches the question by focusing on one of the least expensive policy approaches - issuing mask mandates at the state level and its effect on the 7-day death rate per 100K people in the US at the beginning of the fall wave (October 30th, 2020).

We believe that the implications of our chosen analysis are important as by the start of the fall of 2020, it was clear that some factors, among them perhaps policy measures, do work as evidenced by the decline of infection in the worst affected states from the spring wave. It was also clear that as of October 30th a new wave of infections was starting. Thus evidence in favor of the effectiveness of mask mandates would be useful for policy-makers as they try to stem the pressure from the fall wave of COVID-19 cases.

We built a few models using OLS regression. The strategy was to operationalize the presence of mask mandates at the state level at the time of the data report as an indicator variable, then look at its relationship to the 7-day death outcome rate variable and finally add more immutable population and demographic variables that permit controlling for structural differences between the states. In one model we also added some more policy- and behavior-impacted variables, such as workplace mobility, to see how robust the mask mandate coefficient and significance are in their presence.

The limitations we faced in doing the analysis were plentiful. The exponential nature of the contagion process makes it extremely dynamic. Small random differences at an early stage can lead to vastly different outcomes down the line even if all else is equal. Looking at single period 7-day numbers across many states is particularly vulnerable to that effect. Early in the first wave one way to mitigate would have been to compare periods equally offset from some threshold (e.g. first 25 deaths), but at this point it's much harder to resolve a clear "start" to the wave.

This poses a challenge also in terms of measuring effect size. Death rates per 100K people in late October are completely different from what they are now. Any coefficient we get from our model would only be relevant for that point in time.

Another limitation is that we're working on the state level. That means very few observations coupled with serious challenges to the IID assumption. States themselves can be very large with sometimes massive divergence in regional geography, culture and demographics. As COVID-19 is not yet spread among a huge percentage of the population, random or local events could strongly influence a state's overall numbers, and in the aggregation causal information could be lost.

Finally, we're trying to build a causal model for a highly unstable dynamic pandemic with many potential omitted variables, with few observations and without any possibility for setting up a true experiment or for finding a natural one That is a huge ask. That is why we have focused on a relatively simple setup with somewhat reliably measured variables such as the presence of a mask mandate in a state and the 7-day death rate from COVID-19. Of course, measuring the latter can have its issues. Precisely determining the cause of the death is a complex undertaking in the best of times. Under conditions of global pandemic caused by a novel virus, it is flat out daunting. At the same time, unlike infections, large scale deaths are impossible to hide or ignore. This gives us confidence that the uncertainty and reporting differences between states on that measure are much lower than for other COVID-19-related variables.

If we were to take another stab at building this specific causal model, more granular data (perhaps at the county level) would help. Another idea would be to look not at a single snapshot of death rates, but at some measure of the steepness of the death curve over time. If mask mandates are truly helping to somewhat flatten that curve, their impact might be very significant in the long run.

With these limitations in mind, we find that as of October 30th, at the beginning of the fall COVID-19 wave, there is evidence that having a mask mandate in place would reduce the death toll on the order of 1-2 per week per million people. The result is statistically significant and the effect size is robust across the four models we looked at.

In a state like California with a population of about 40 million, that would be 40-80 fewer deaths per week. Is that practically significant? To the families and communities of those people, it's hugely significant. For changing the overall direction and impact of the pandemic, it's a lot more difficult to say. Yet given the low practical, if not political, cost to the measure, and the fact that at a minimum wearing a mask is completely innocuous to the wearer, the answer to the question whether state officials should impose mask mandates is yes, absolutely.