# Can Mask Mandates Help Lower the COVID-19 Death Rate?

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 pandamic 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 issues mandates for their citizens to wear masks in public and additionally, many states have issues 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 November of 2020), we ask: *Do state-wide facemask mandates correlate with the reduction in the amount of deaths in that state due to COVID-19?*

Because these mandates and other variables occur in a different times during the nine-month period, and the states were affected by the virus in different times as well, we will operationalize the variable "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 which will skew the current state of the pandemic. Using the current data (the past 7 days) will give a good analysis if the 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 they are successful, however 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 thus 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: population density, 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 death rates of citizens due to COVID-19.

## Import the data

```
library(magrittr)
library(tidyverse)
library(ggplot2)
```

```
library(readxl)
library(openxlsx)
library(stargazer)
library(lmtest)
library(sandwich)
library(patchwork)
library(corrplot)
```

```
covid_raw_data<-read.csv("covid-19.csv",skip=1)
covid_masks_policies_data<-read.csv("covid policies masks.csv")

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

```
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'
      )

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)

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_dens = "population_density",
        work_mob = "workplaces_mobility_change",
        retail_mob = "retail_mobility_change",
        per_risk = "percent_at_risk"
  )


cor_mat <- cor(covid_corr[,c("case_rt", "case_rt7", "death_rt", "death_rt7", "white",
                     "percent_over_65","per_risk", "pop_dens", "mask_mandate_all",
                     "work_mob", "retail_mob", "repgov")])

corrplot(cor_mat, method = "color", order = "AOE", diag=FALSE, addCoef.col = "white",
        addCoefasPercent = TRUE)
mtext("Correlation Matrix Visualization of Mask Mandates", side=3, adj=0, cex=1.0)
```
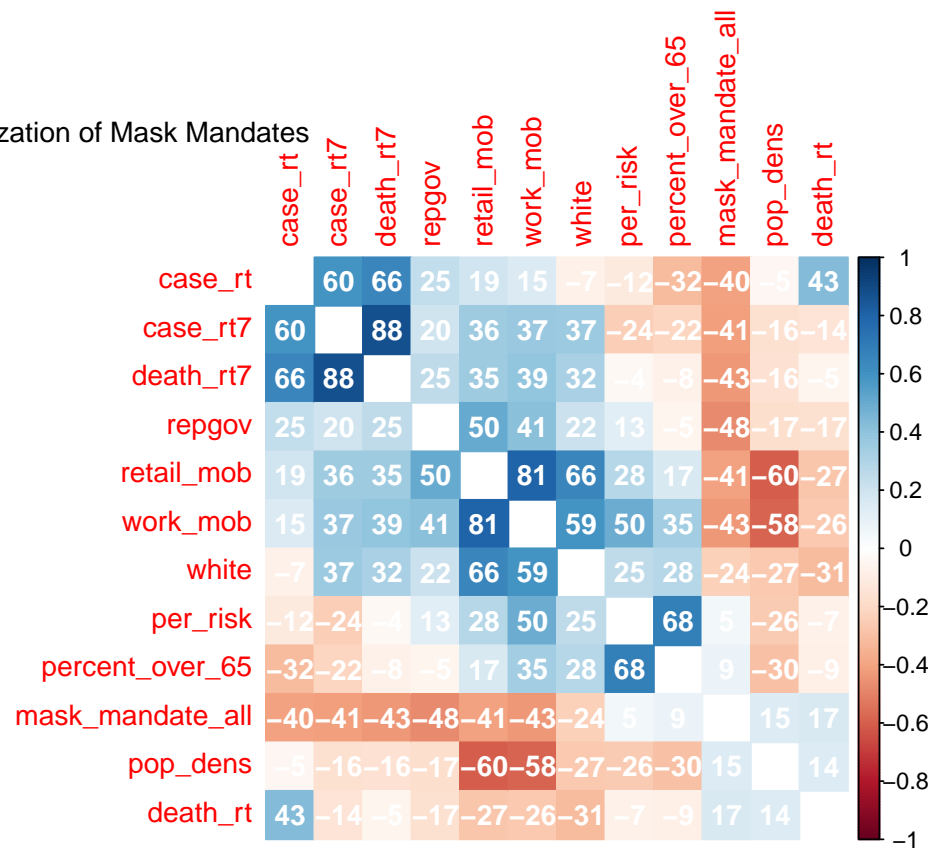


Correlation Matrix Visualization of Mask Mandates

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")
```

**Distribution of recent of death rates**
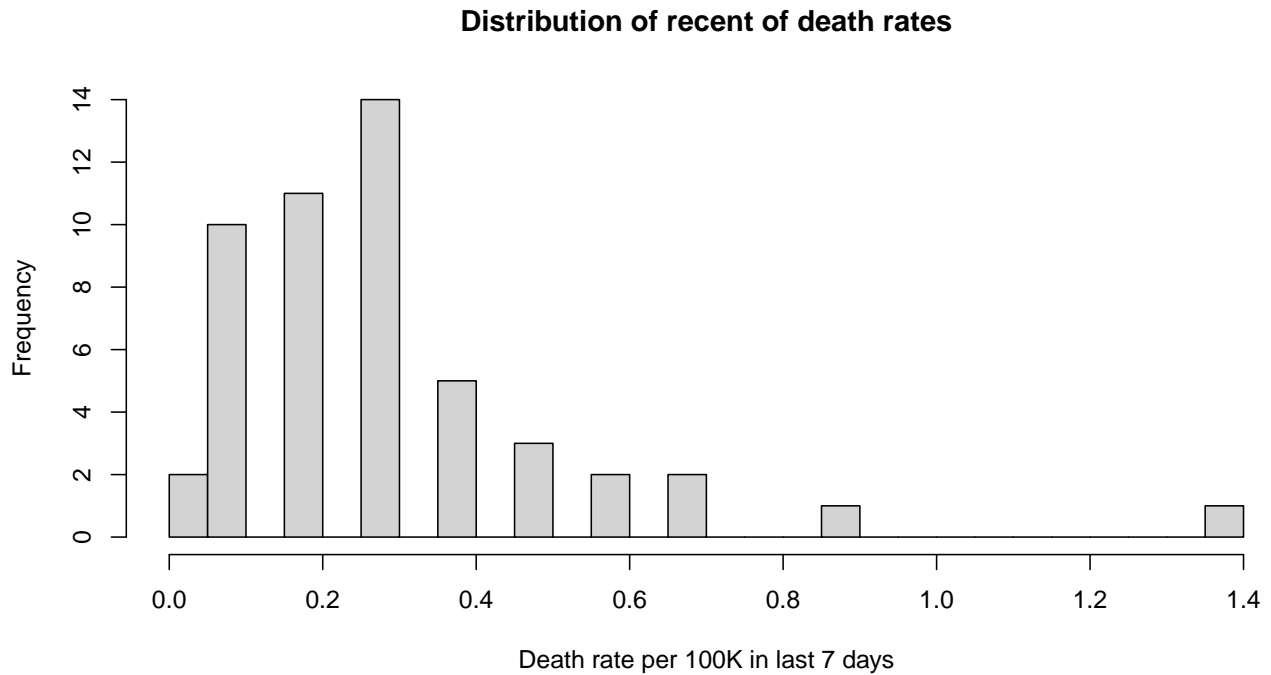


There is some skew, but perhaps we can argue that it's not too severe. Let's see how having it as an output variable looks against the mask mandate variable.

```
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'
  )
```

Separation in last 7–day death rate by mask mandate
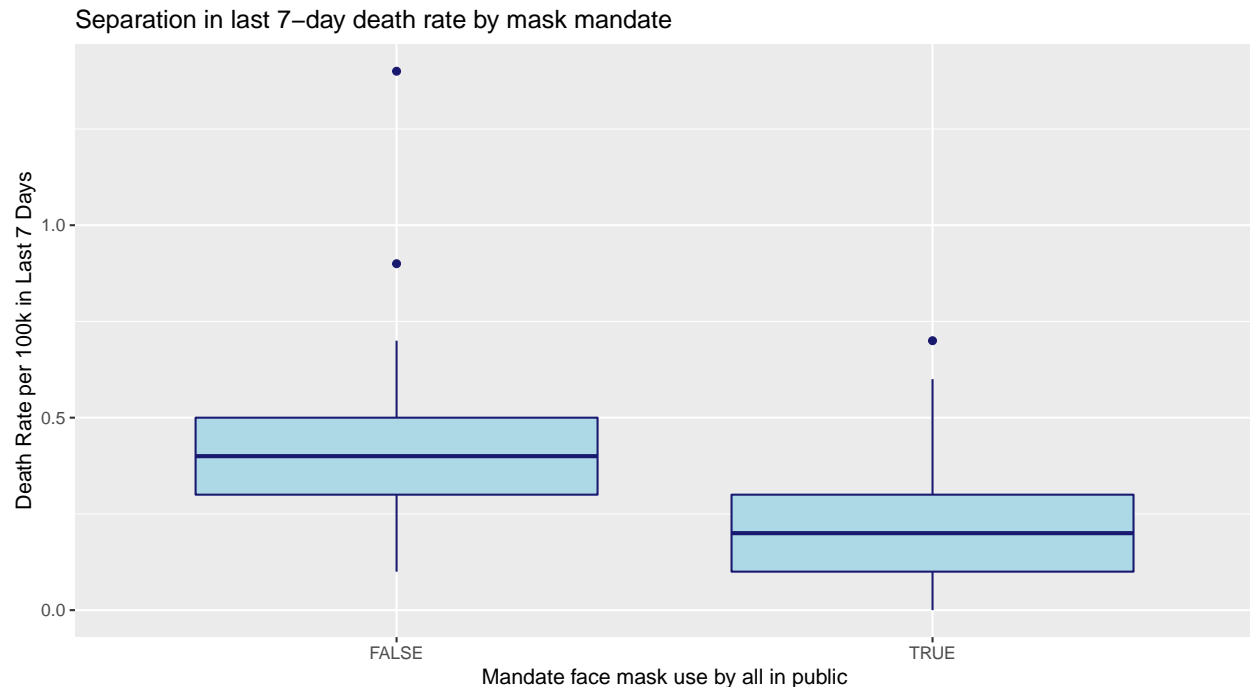
There seems to be a separation that would make using this relationship for our first model interesting.

## Model 1

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

$$death\_rate_7 = \beta_0 + \beta_1 mandate + w \tag{Model 1}$$

```
model1 <- lm(death_rate_in_last7 ~ mask_mandate_all, data = covid_data)
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                      Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          0.450000   0.075903  5.9286 3.014e-07 ***
## mask_mandate_allTRUE -0.216667   0.080361 -2.6962  0.009587 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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")
```

**Residuals from Linear Model 1 Predicting Death Rate in last 7 days**
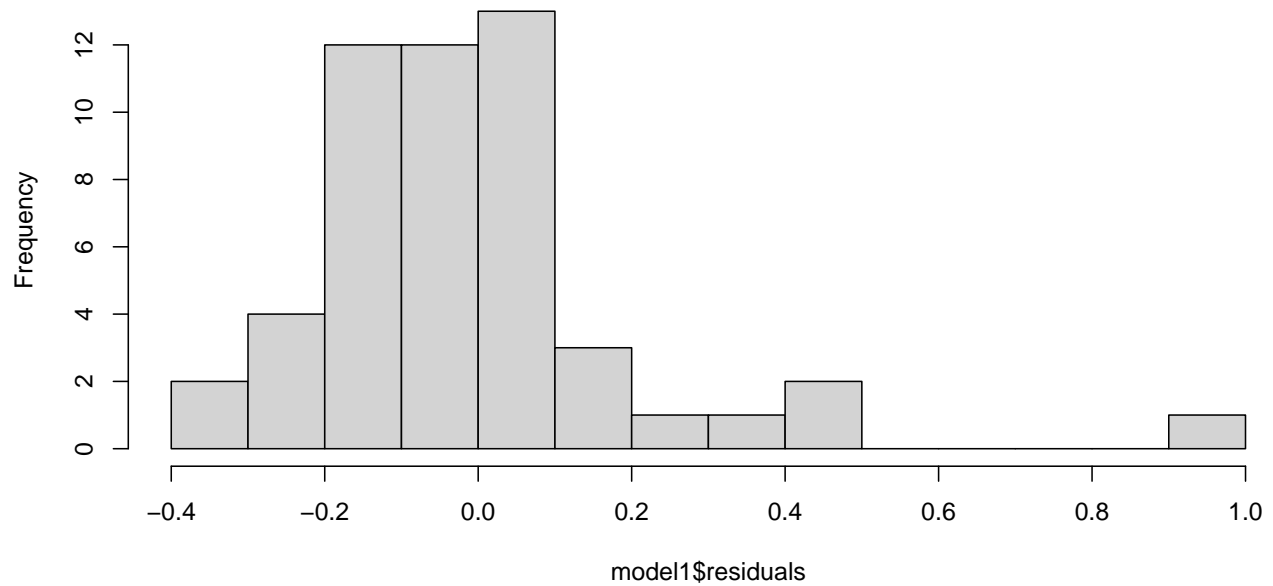


```
plot(model1, which=2,
     sub.caption = "QQ Plot of Model 1 Residuals",
     labels.id = state.name) # QQ Plot of Residual
```

Normal Q–Q



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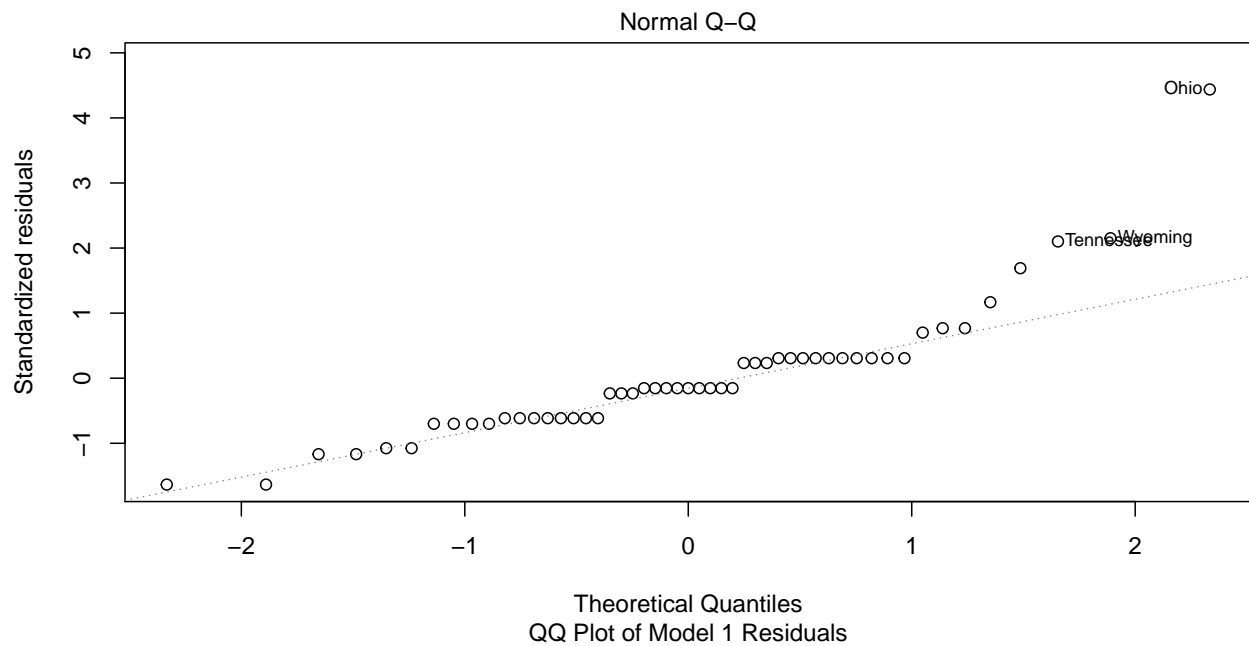
When we plot the residuals of our first model, we can see that it isn't quite normal, with one particular state (Ohio) having a residual above 1, which is quite high given that the median is below 0.5. Then again we don't see clear evidence of violations of assumptions 2 or 4.

# Model 2

Since we had a human controlled variable in the first model to describe, 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 will 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 add 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 are of the white race), 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 insurance to show that some races have had higher death rates due to COVID-19 than others.

*percent_at_risk* - Groups with people who have an existing illness like Diabetes or Heart Disease have seen higher death rates due to the lack of immunity and fighting resources available in the body to fight the coronavirus. COVID-19 also seems to worsen the existing co-morbities that people with such diseases have, making them susceptible to death by the virus.

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 with age over 65 across states',
    x = 'State',
    y = 'Population percentage with age over 65'
  )
```

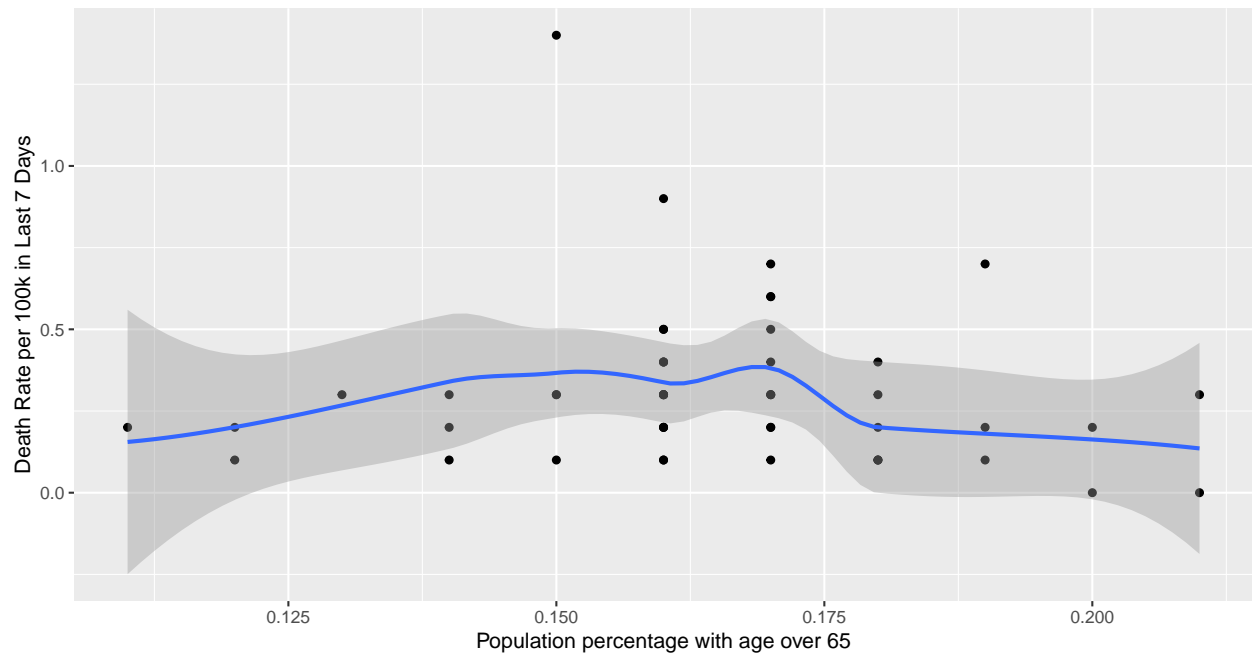**Population percentage with age over 65 across states**
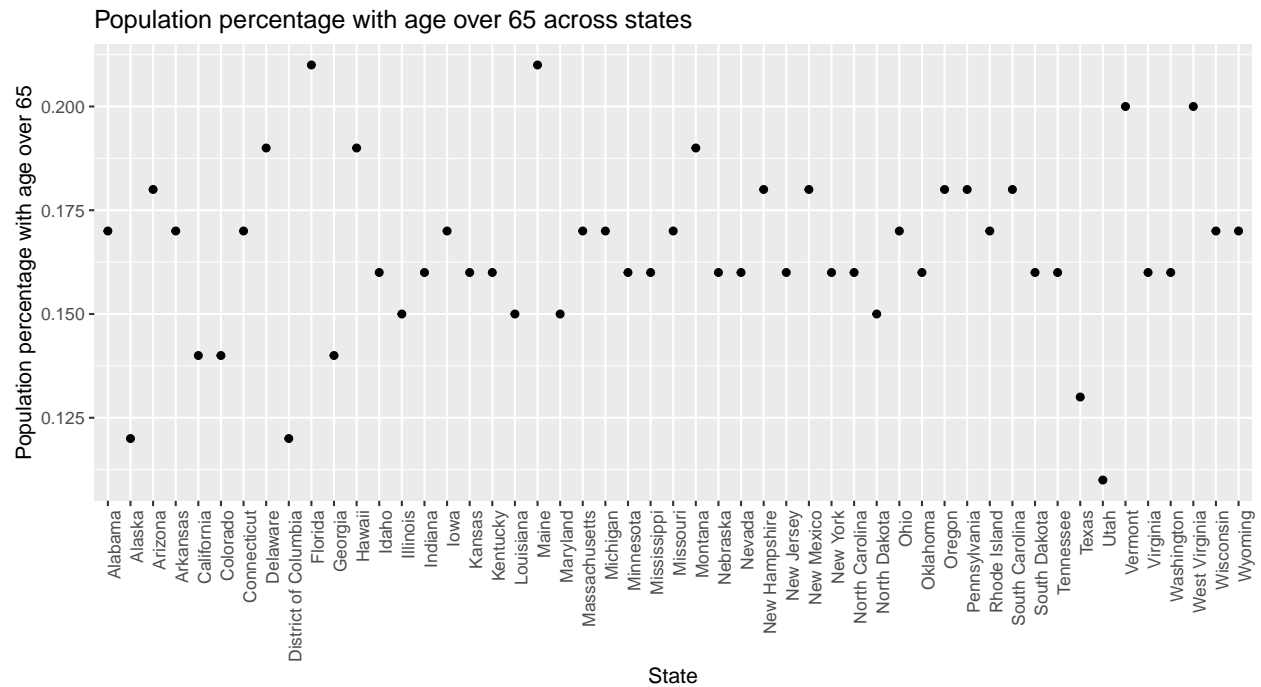


Distribution of last 7-day death rate by population percentage with age over 65:

```
covid_data %>%
  ggplot(aes(x = (percent_over_65), y = death_rate_in_last7)) +
  geom_point() +  geom_smooth() +
  labs(
    title = 'Distribution of last 7-day death rate by population percentage with age over 65',
    x = 'Population percentage with age over 65',
    y = 'Death Rate per 100k in Last 7 Days'
  )
```
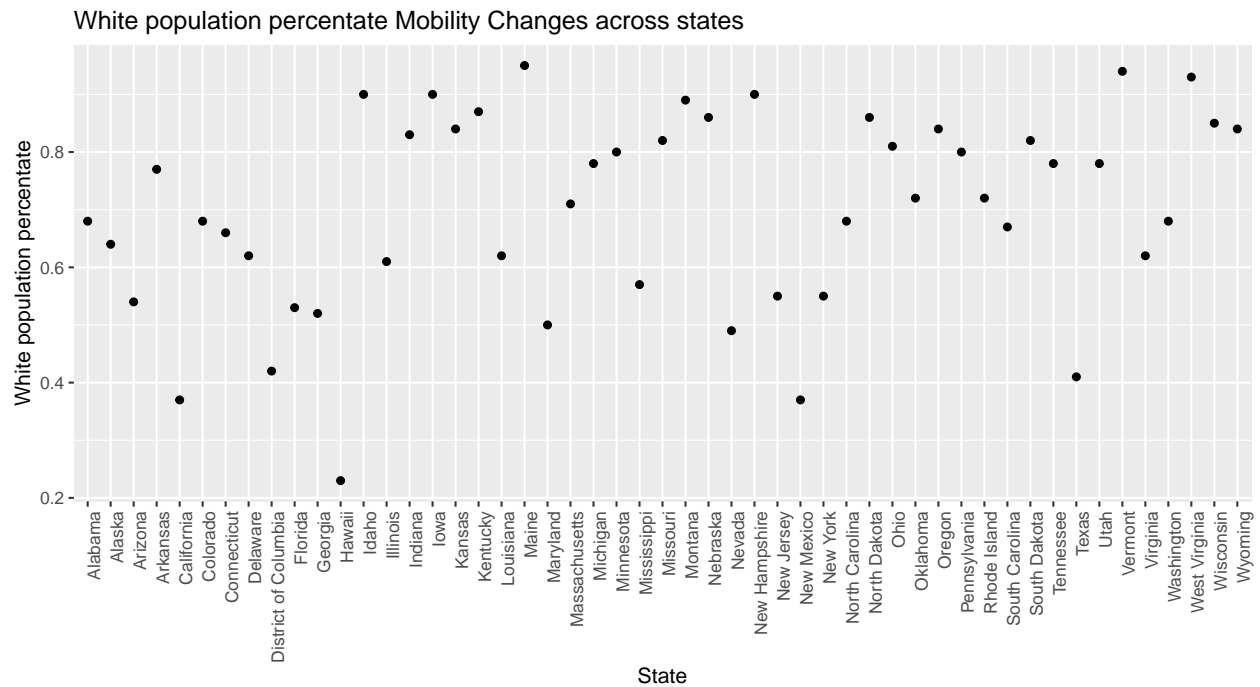
```
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
```

### Distribution of last 7–day death rate by population percentage with age over 65



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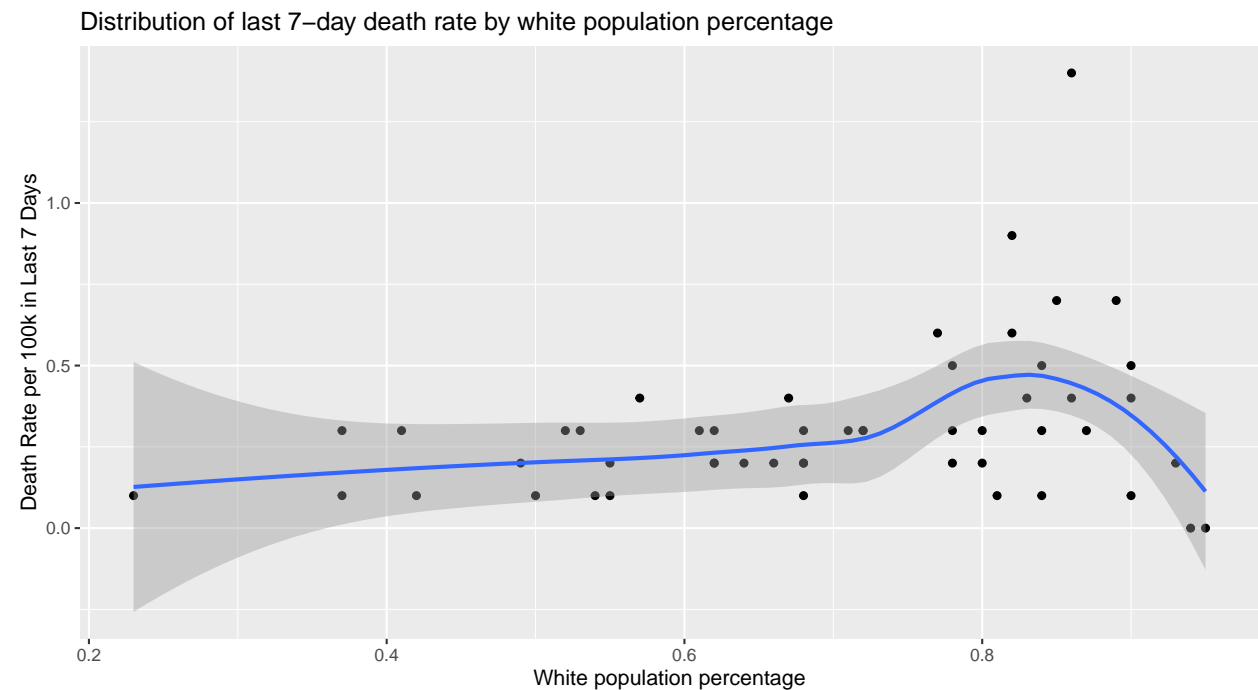
```
covid_data %>%
  ggplot(aes(x = State, y = white_percent)) + geom_point() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(
    title = 'White population percentate Mobility Changes across states',
    x = 'State',
    y = 'White population percentate'
  )
```

Distribution of last 7-day death rate by white population percentage:
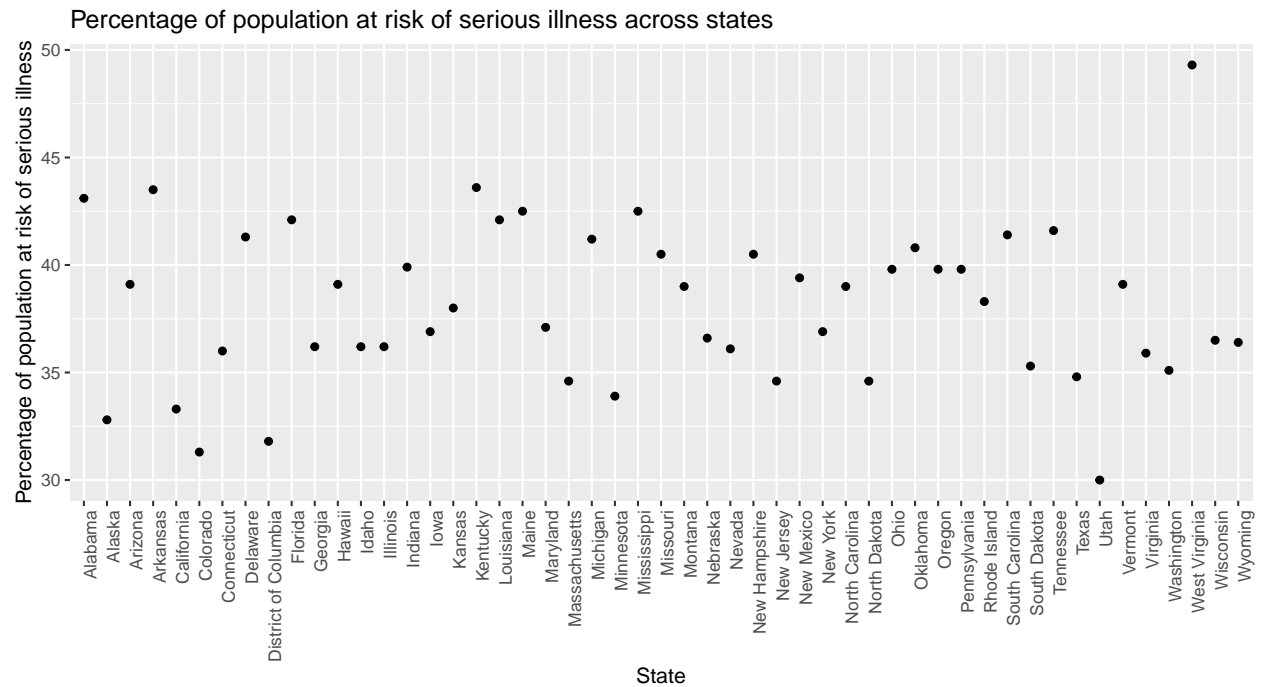
```
covid_data %>%
  ggplot(aes(x = (white_percent), y = death_rate_in_last7)) +
  geom_point() +  geom_smooth() +
  labs(
    title = 'Distribution of last 7-day death rate by white population percentage',
    x = 'White population percentage',
    y = 'Death Rate per 100k in Last 7 Days'
  )
```

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

Distribution of last 7–day death rate by white population percentage



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

```
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'
  )
```

**Percentage of population at risk of serious illness across states**



Distribution of last 7-day death rate by percentage of population at risk of serious illness:

```
covid_data %>%
  ggplot(aes(x = percent_at_risk, y = death_rate_in_last7)) +
  geom_point() +  geom_smooth() +
  labs(
    title = 'Distribution of last 7-day death rate by percentage
    of population at risk of serious illness',
    x = 'Percentage of population at risk of serious illness',
    y = 'Death Rate per 100k in Last 7 Days'
  )
```

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

Distribution of last 7–day death rate by percentage
of population at risk of serious illness
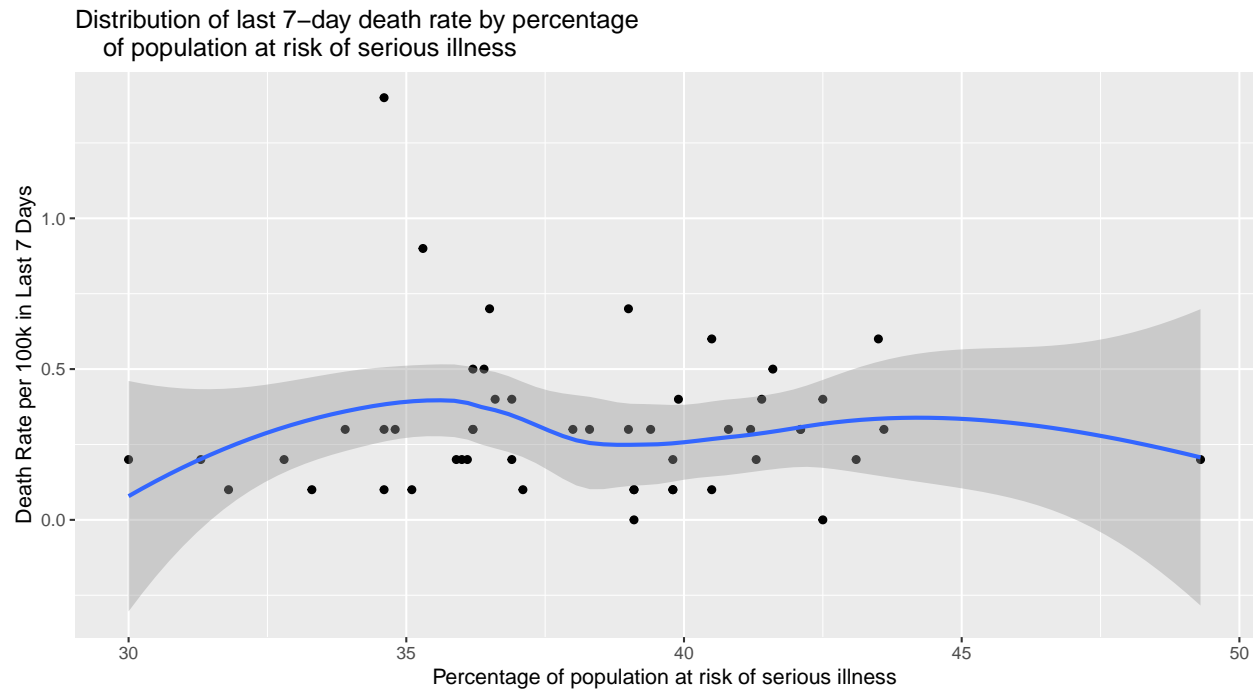


We will use white population percentage as our additional variable in model 2 because we see variance in it over states so a change in it could indicate differences in our outcome variable.

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

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

```
##
## t test of coefficients:
##
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.201059   0.123825  1.6237  0.11098
## mask_mandate_allTRUE -0.188840   0.076316 -2.4744  0.01693 *
## white_percent         0.329724   0.174842  1.8858  0.06538 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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, it's 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 healthcare 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)
coeftest(model3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                       Estimate  Std. Error t value Pr(>|t|)
## (Intercept)         0.39987293  0.38416978  1.0409  0.30337
## mask_mandate_allTRUE -0.17870825  0.07232444 -2.4709  0.01724 *
## white_percent        0.38564354  0.20030849  1.9252  0.06039 .
## percent_over_65     -1.35246686  2.01923246 -0.6698  0.50634
## percent_at_risk     -0.00057089  0.01091469 -0.0523  0.95851
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.

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.

$$death\_rate_7 = \beta_0 + \beta_1 mask\_mandate + \beta_2 white\_percent + \beta_3\ percent\_over\_65$$

$$+\beta_4 workplaces\_mobility\_change + v \hspace{3cm} \text{(Model 4)}$$

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

```
##
## t test of coefficients:
##
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.8142148  0.4197540  1.9397  0.05856 .
## mask_mandate_allTRUE      -0.1372304  0.0910756 -1.5068  0.13871
## white_percent              0.2390596  0.2578993  0.9269  0.35879
## percent_over_65           -2.1251574  1.8084567 -1.1751  0.24599
## workplaces_mobility_change 0.0084887  0.0066873  1.2694  0.21069
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We loose the 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.

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 "independant" portion of IID.

*DAVID TO ADD STATE CHART HERE*

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.

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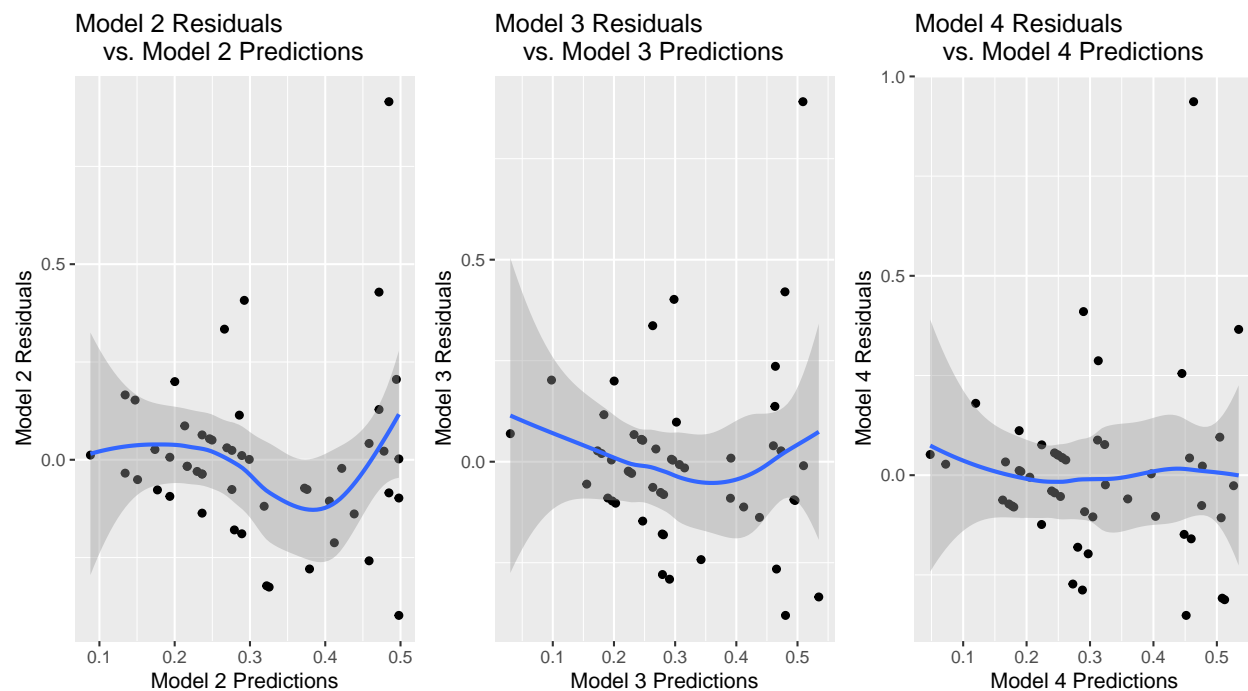
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

15

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.

```r
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 6',
    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 %>%
```

16

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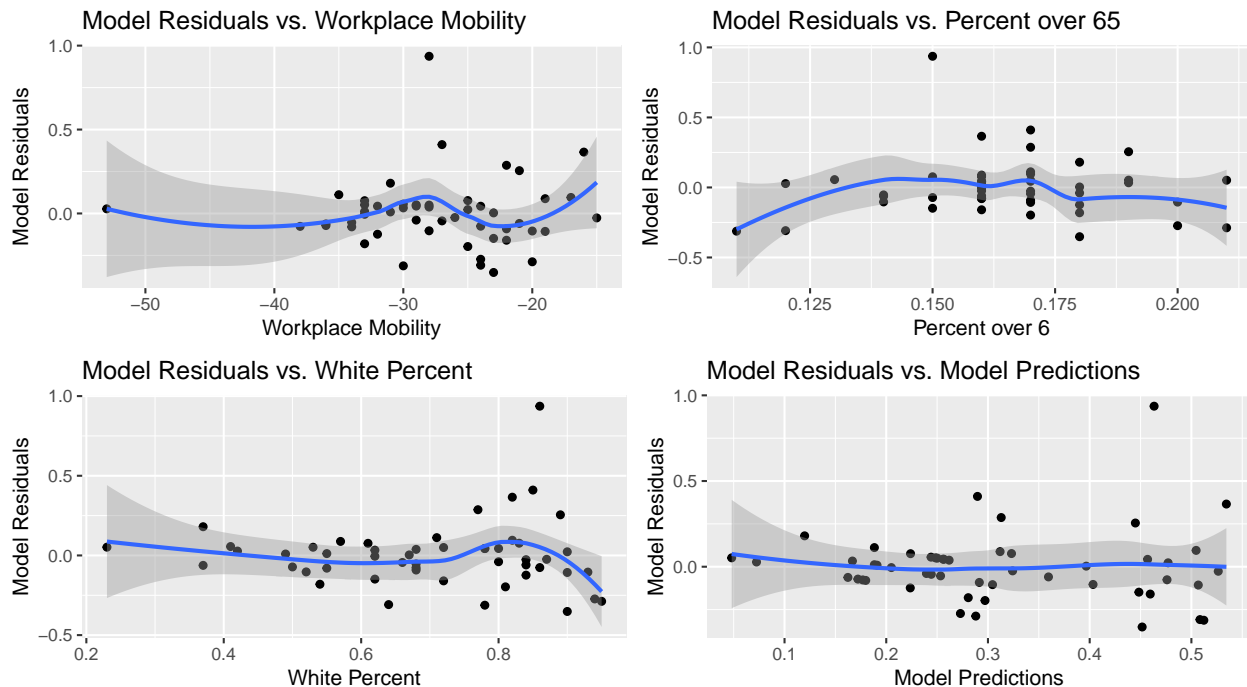
```
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
```
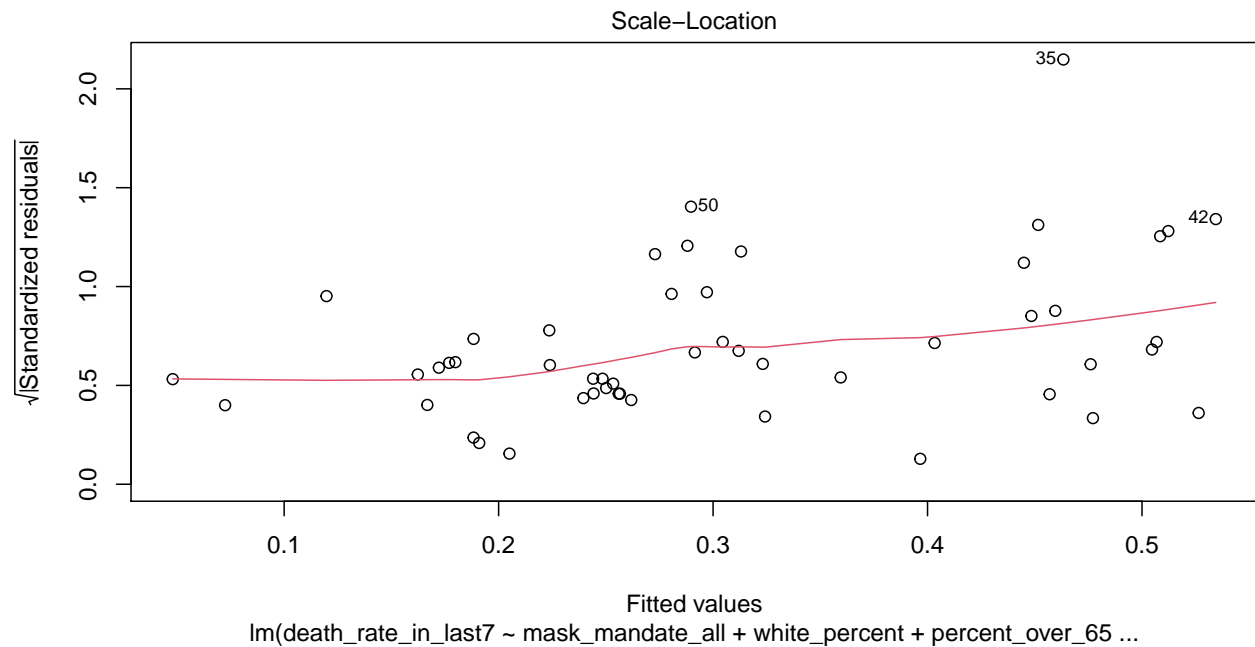


3. No perfect Collinearity The correlation table in our EDA gives us a clear indication that we're safe in that regard. While groups of variables have clear relationships (such as the different mobility categories), none of them are perfectly colinear.

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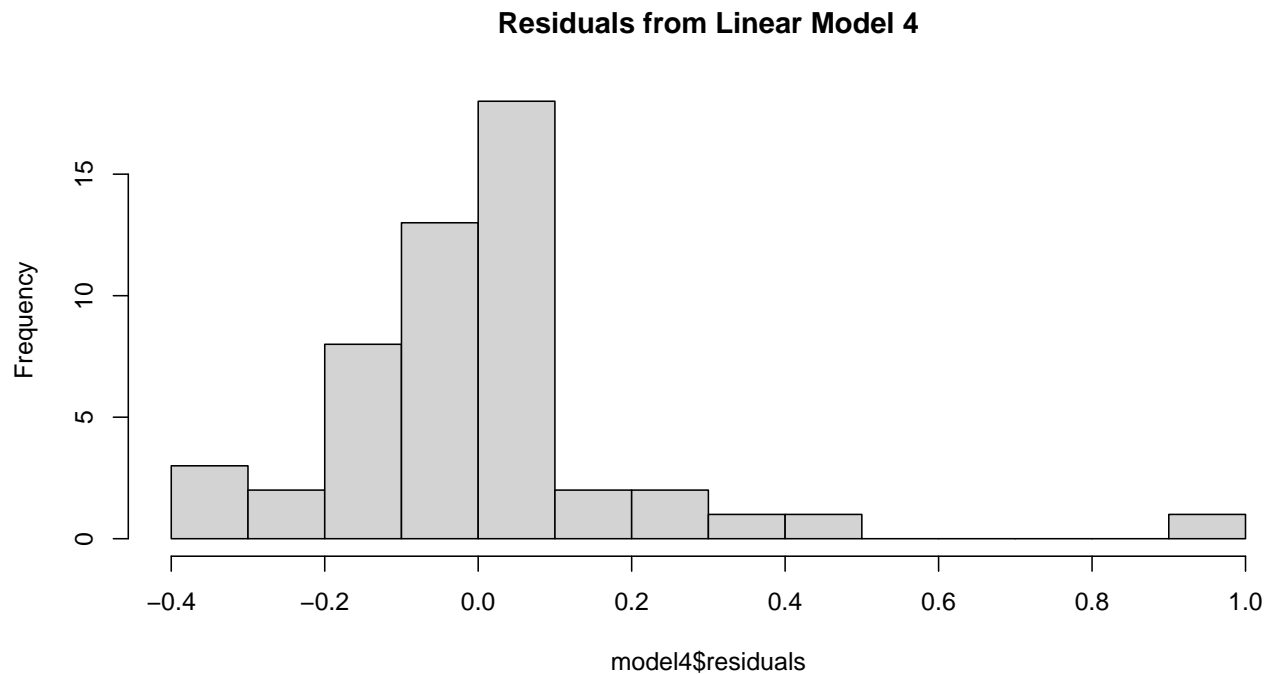
```
lmtest::bptest(model4)
```

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

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



Scale–Location

lm(death_rate_in_last7 ~ mask_mandate_all + white_percent + percent_over_65 ...
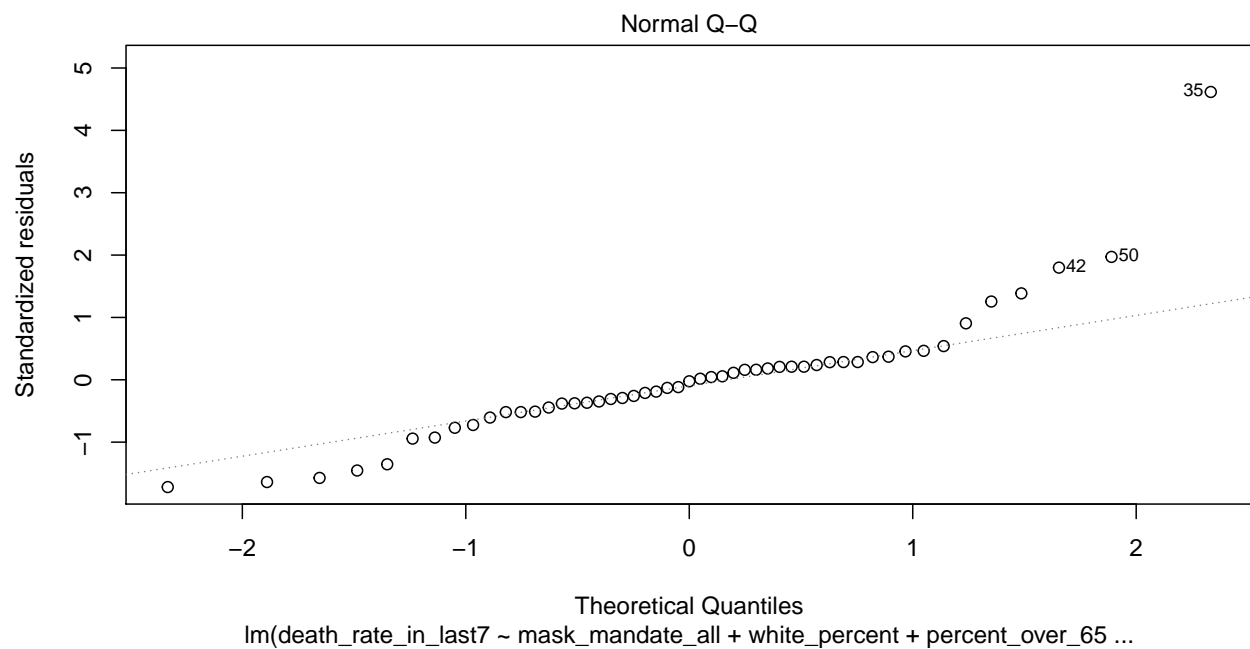
5. Normality of the errors According to the Gauss-Markov Theorem under assumptions 1-4 we still have OLS as the best linear unbiased estimator. How much should we worry about assumption number 5. Well, we will be updating this section as soon as the async catches up with the lesson on assumption 5. Meanwhile 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
```

18

**Residuals from Linear Model 4**



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

**Normal Q–Q**



Theoretical Quantiles
lm(death_rate_in_last7 ~ mask_mandate_all + white_percent + percent_over_65 ...

# 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 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 but not a whole lot from model2 to model3. 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:
$$death\_rate_7 = \tilde{\beta}_0 + \tilde{\beta}_1 mandate + w$$

Actual model:
$$death\_rate_7 = \beta_0 + \beta_1 mandate + \beta_2 adoption + \omega$$

Direction of Bias: With $\beta_2$ 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 expected 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.

### 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 'workplace mobility changes' 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 and increasing viral load and therefore severity.

Estimated model:
$$death\_rate_7 = \tilde{\beta}_0 + \tilde{\beta}_1 workplace\ mobility\ changes + w$$

Actual model:

$$death\_rate_7 = \beta_0 + \beta_1 workplace\ mobility\ changes + \beta_2 work\ from\ home\ option + \omega$$

Direction of Bias: With $\beta_2$ negative (i.e. increasing work from home option will be associated with decrease in death rate) and work from home options to be negatively correlated with workplace mobility changes (i.e. increasing work from home options will decrease workplace mobility), the actual coefficient of mandate will be more negative than expected and the direction of bias will be toward zero.

### Diabetic Population

3. The third variable we want to consider as one that could be introducing omitted variable bias is 'diabetic population'. This variable will reflect what percentage of the population has diabetes. 'Type 2 Diabetic population' is correlated with 'white population' 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:
$$death\_rate_7 = \beta_0 + \beta_1 white\ population + w$$

Actual model:
$$death\_rate_7 = \beta_0 + \beta_1 white\ population + \beta_2 diabetes\ population + \omega$$

Direction of Bias: With $\beta_2$ positive (i.e. higher population of people with diabetes associated with higher death rate) and diabetes population negatively correlated with white population, the actual coefficient of mandate will be more positive than expected and the direction of bias will be toward zero.

**Donald Trump Supporters**

4. The fourth variable we want to consider as one that could be introducing omitted variable bias is 'percent of Donald Trump supporters'. This variable will population of people that support Donald Trump. This will be correlated with the work mobility 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:
$$death\_rate_7 = \beta_0 + mask\ mandate + w$$

Actual model:

$$death\_rate_7 = \beta_0 + mask\ mandate + \beta_2 population\ supporting\ Donald\ Trump + \omega$$

Direction of Bias: With $\beta_2$ positive (i.e. higher percentage of Donald Trump supporters associated with higher death rates) and positive correlation with work mobility (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.

**Population with Heart Disease**

5. The fifth variable we want to consider as one that could be introducing omitted variable bias is 'population with heart disease'. This variable will reflect what percentage of the population has heart disease. 'Population with heart disease' is correlated with 'white population' 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:
$$death\_rate_7 = \beta_0 + \beta_1 white\ population + w$$

Actual model:

$$death\_rate_7 = \beta_0 + \beta_1 white\ population + \beta_2 population\ with\ heart\ disease + \omega$$

Direction of Bias: With $\beta_2$ positive (i.e. higher population of people with heart disease associated with higher death rate) and population with heart disease negatively correlated with white population, the actual coefficient of mandate will be more positive than expected and the direction of bias will be toward zero.

# 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 building models that incorporate both policy variables and more immutable population and demographic variables in an attempt to explain what might affect the most recent 7-day death rate by state.

Overall, our results are that we don't find evidence supporting the idea of a simple policy magic bullet or a population characteristic that can protect against the deadly virus. That said, we have identified a number of omitted variables, not available in the studied dataset that could shine a brighter light onto what could work in the short term before a long awaited vaccine solves the problem in a sustainable and permanent way.