Lab 2

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." Because we have not actually performed a controlled study/experiment, we want to build an descriptive 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, we ask: *Do state-wide facemask mandates correlate with the reduction in the amount of deaths in that state?*

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). Total deaths 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 descriptive 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

```r
library(magrittr)
library(tidyverse)
library(ggplot2)
library(readxl)
library(openxlsx)
library(stargazer)
library(lmtest)
library(sandwich)
library(patchwork)
library(corrplot)
```

```r
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)
```

```
## Joining, by = c("State", "Mandate.face.mask.use.by.all.individuals.in.public.spaces", "No.legal.enfo
```

```r
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',
      mask_for_public_facing_employee_mandated_on = 'Mandate.face.mask.use.by.employees.in.public.facin
      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',
      x65='X65.'
      )

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)

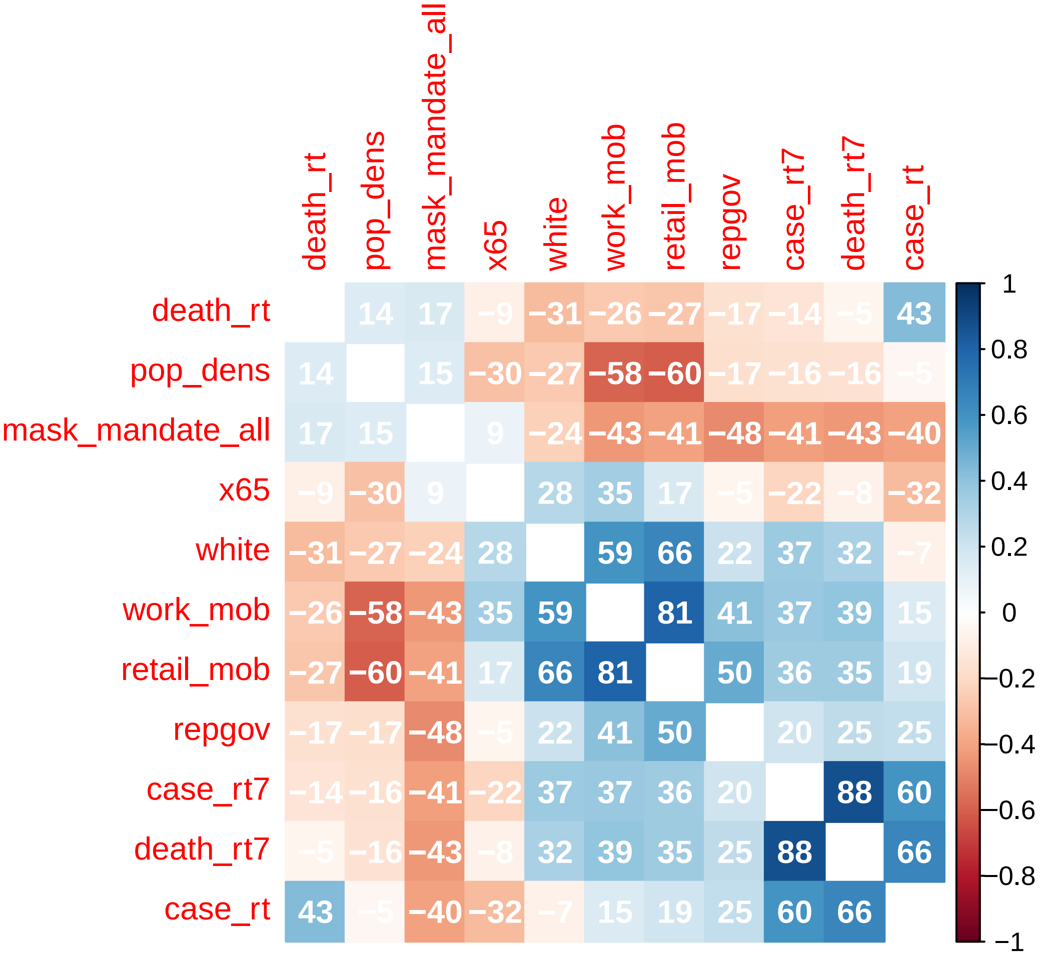
Let us take a look at a correlation table of a number of interesting variables.

```r
covid_corr <- covid_data[,c("State","case_rate", "case_rate_in_last7", "death_rate", "death_rate_in_las

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"
  )
```

```
cor_mat <- cor(covid_corr[,c("case_rt", "case_rt7", "death_rt", "death_rt7", "white", "x65", "pop_dens"

corrplot(cor_mat,method = "color", order = "AOE",
        diag=FALSE, addCoef.col = "white", addCoefasPercent = TRUE)
```

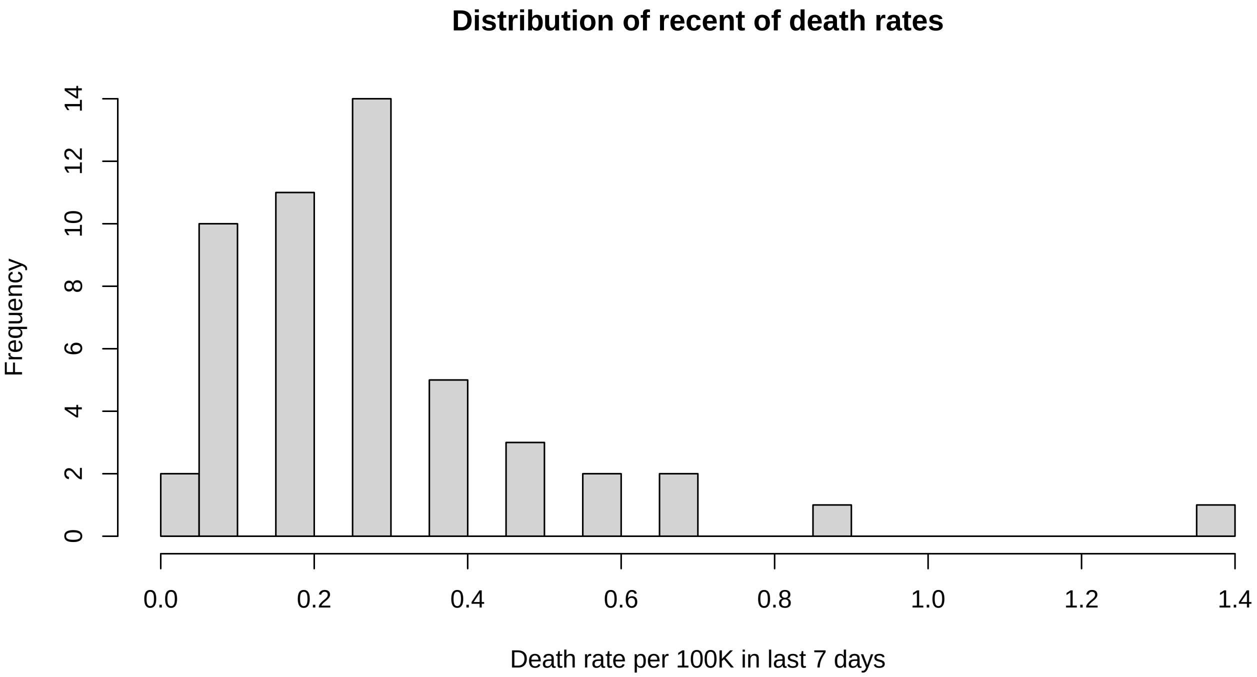


	death_rt	pop_dens	mask_mandate_all	x65	white	work_mob	retail_mob	repgov	case_rt7	death_rt7	case_rt
death_rt		14	17	−9	−31	−26	−27	−17	−14	−5	43
pop_dens	14		15	−30	−27	−58	−60	−17	−16	−16	−5
mask_mandate_all	17	15		9	−24	−43	−41	−48	−41	−43	−40
x65	−9	−30	9		28	35	17	−5	−22	−8	−32
white	−31	−27	−24	28		59	66	22	37	32	−7
work_mob	−26	−58	−43	35	59		81	41	37	39	15
retail_mob	−27	−60	−41	17	66	81		50	36	35	19
repgov	−17	−17	−48	−5	22	41	50		20	25	25
case_rt7	−14	−16	−41	−22	37	37	36	20		88	60
death_rt7	−5	−16	−43	−8	32	39	35	25	88		66
case_rt	43	−5	−40	−32	−7	15	19	25	60	66	

An intersting 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.

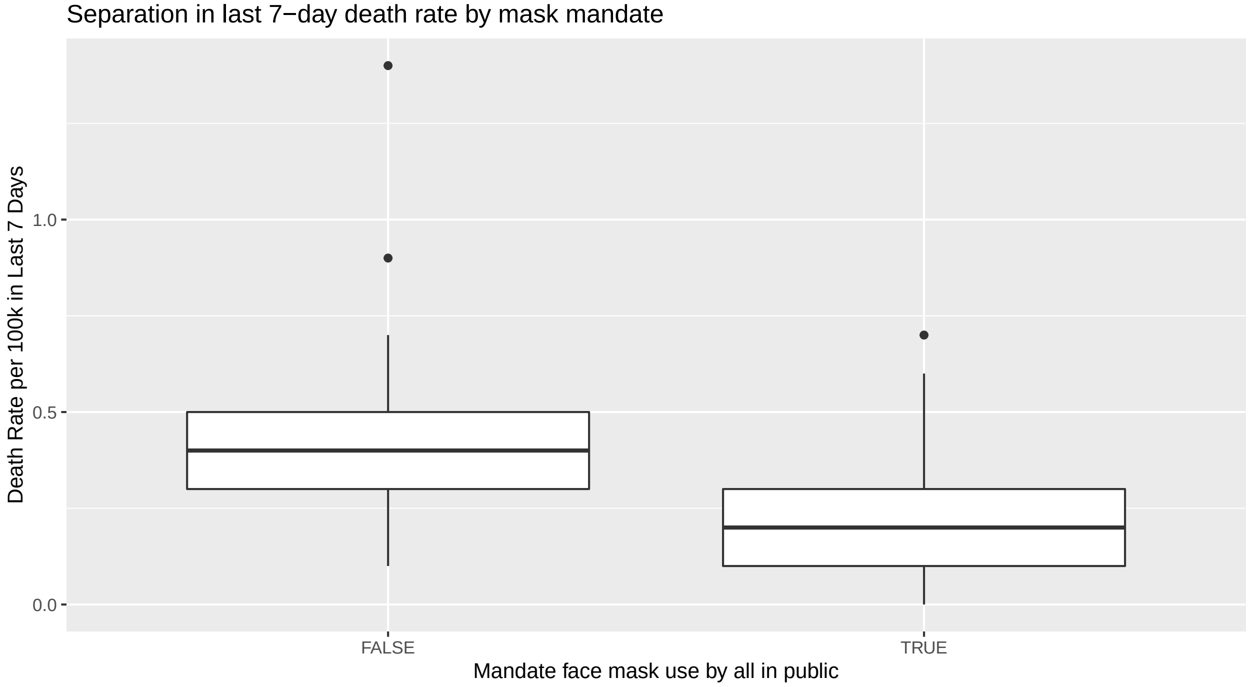
So let's focus on the variable *Death Rate per 100K in the Last 7 Days*.

```
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() +
  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

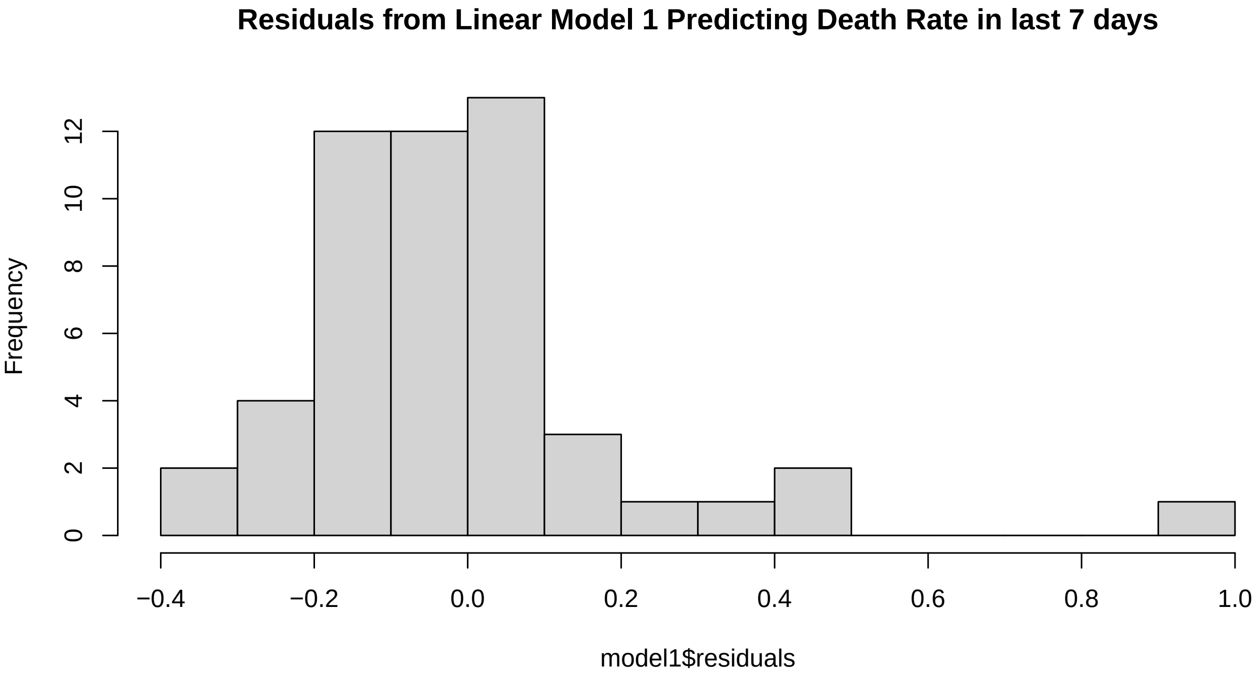
$$death\_rate_7 = \beta_0 + \beta_1 mandate + w \tag{1}$$

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

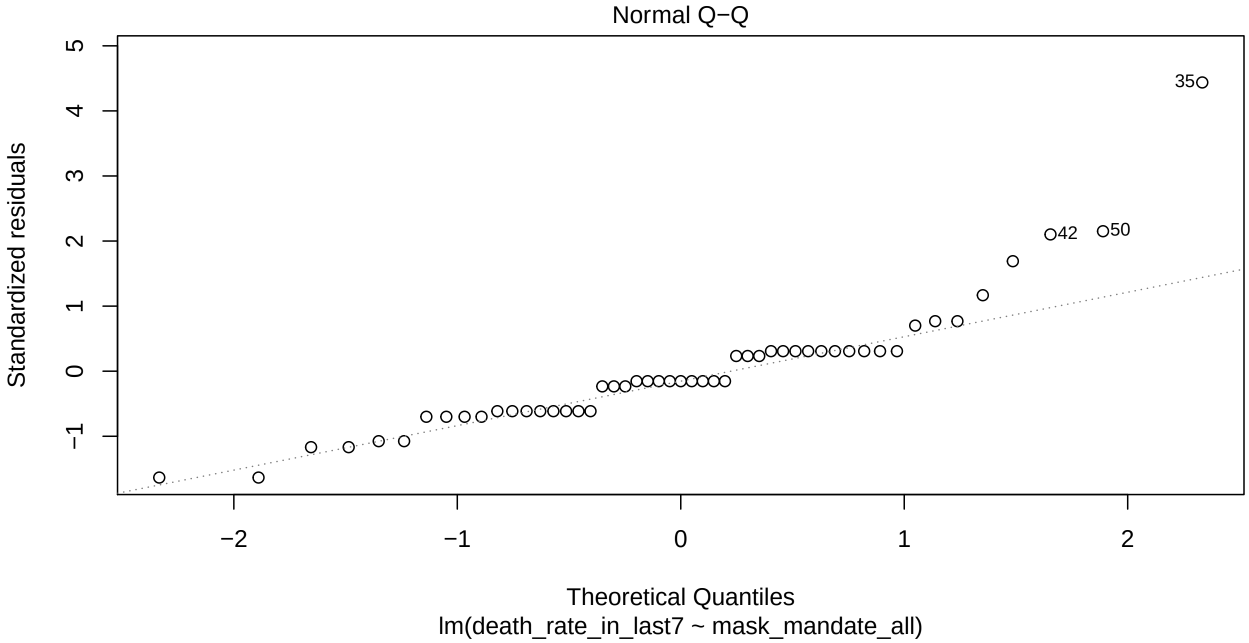
```
##
## t test of coefficients:
##
##                       Estimate Std. Error t value  Pr(>|t|)
## (Intercept)           0.450000   0.075903  5.9286 3.014e-07 ***
## mask_mandate_allTRUE -0.216667   0.080361 -2.6962  0.009587 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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 2.1 deaths per 1 million people per 7 days.

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



```
plot(model1, which=2) # QQ Plot of Residuals
```



```
#plot(model1, which=3) # Heteroskedasticity (looking for a straight line)
#plot(model1, which=5) # Cook's distance
```

When we plot the residuals of our first model, we can see that it isn't quite normal, with one particular state (North Dakota) having a residual above 1. Then again we don't see clear evidence of violations of assumptions 2 or 4.

Model 2.

We will add workplaces_mobility_change in model 2. From the correlations table, workplaces_mobility_change and mask_mandate_all seem not highly correlated so there shouldn't be any multi-collinearlity.

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

```
##
## t test of coefficients:
##
##                              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                 0.6629892  0.1097662  6.0400 2.178e-07 ***
## mask_mandate_allTRUE       -0.1622320  0.0966754 -1.6781   0.09983 .
## workplaces_mobility_change  0.0090420  0.0048311  1.8716   0.06736 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interestingly, we lose statistical significance on both coefficients. For the mask mandates we have a p-value of 0.09983 and for change in workplace mobility we have a p-value of 0.06736. This could indicate that underneath may be an omitted variable that is affecting both (such as not believing in Covid-19 or its health consequences, resulting in both lack of mask mandates and changes in workplace mobility).

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.

Now that we have observed two human actions that could have affected the death rate in 100K in last 7 days, we will add population factors in our model 3, particularly: population density, percentage of population that is white, and percentage of population over 65 years of age.

There is no perfect multi-collinearilty in the variables used as explanatory variables in this model but there is some correlation between population_density and white_percent and workplaces_mobility_change and white_percent. This might affect the weight on the coefficients as they might either share the weight or it might weigh heavily on one over the other. We need to keep this instability in mind while reading the results from model 3.

```r
model3 <- lm(death_rate_in_last7 ~ mask_mandate_all + workplaces_mobility_change + population_density +
coeftest(model3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                               Estimate  Std. Error t value Pr(>|t|)
## (Intercept)                 8.2298e-01  4.3826e-01  1.8778  0.06689 .
## mask_mandate_allTRUE       -1.3635e-01  9.6998e-02 -1.4057  0.16668
## workplaces_mobility_change  8.8873e-03  9.3192e-03  0.9537  0.34535
## population_density          2.4559e-06  6.8689e-05  0.0358  0.97164
## white_percent               2.3591e-01  2.7679e-01  0.8523  0.39856
## x65                        -2.1079e+00  1.8949e+00 -1.1124  0.27186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All the coefficents 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 occurance 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 x65 seems to have a more sizeable coefficient.

```r
model4 <- lm(death_rate_in_last7 ~ mask_mandate_all + workplaces_mobility_change + x65, data = covid_da
coeftest(model4, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.0483595  0.3428569  3.0577 0.003674 **
## mask_mandate_allTRUE       -0.1373952  0.0907953 -1.5132 0.136915
## workplaces_mobility_change  0.0119193  0.0045044  2.6462 0.011040 *
## x65                        -1.9577639  1.6744514 -1.1692 0.248220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient of the mask_mandates in this model is not significant, with a p-value at 0.14. Practically speaking, it says that having a face mask mandate for the public, keeping workplace mobility changes constant, is associated with a reduction of ~0.14 deaths per 100k people, or ~1.4 deaths per 1 million people.

The coefficient of the workplaces_mobility_change in this model is statistically significant, with a p-value at 0.01104. Practically speaking, it says that having a one unit decrease in workplace mobility, keeping all other explanatory variables constant, is associated with an decrease of ~0.01 deaths per 100k people, or ~10 deaths per 100 million people.

The coefficient of the x65 in this model is not significant either with a p-value at 0.25. Practically speaking, it says that every unit increase in population percentage of people over 65 years of age, all other explanatory variables constant, is associated with a decrease of ~2 deaths per 100k people, or ~20 deaths per 1 million people, which seems odd given the increased risk of dying associated with older age.

3. Limitations of your 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.

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.

3. No perfect Colinearity 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.

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.

4. A Regression Table

```r
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: Wed, Nov 18, 2020 - 12:47:55 AM

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.

Table 1:

	Dependent variable:			
	death_rate_in_last7			
	model1	model2	model3	model 4
	(1)	(2)	(3)	(4)
mask_mandate_all	-0.217^{***}	-0.162^{*}	-0.136	-0.137^{*}
	(0.080)	(0.097)	(0.097)	(0.073)
workplaces_mobility_change		0.009^{*}	0.009	0.012^{**}
		(0.005)	(0.009)	(0.006)
population_density			0.00000	
			(0.0001)	
white_percent			0.236	
			(0.277)	
x65			-2.108	-1.958
			(1.895)	(1.628)
Constant	0.450^{***}	0.663^{***}	0.823^{*}	1.048^{***}
	(0.076)	(0.110)	(0.438)	(0.345)
Observations	51	51	51	51
R^2	0.187	0.238	0.279	0.261
Adjusted R^2	0.170	0.206	0.199	0.214
Residual Std. Error	0.220 (df = 49)	0.215 (df = 48)	0.216 (df = 45)	0.214 (df = 4
F Statistic	11.265^{***} (df = 1; 49)	7.497^{***} (df = 2; 48)	3.487^{***} (df = 5; 45)	5.526^{***} (df = 3
Note:			*p<0.1; **p<0.05; ***p<	

5. Discussion of Omitted Variables

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 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 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.

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 β_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.

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 β_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.

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 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 β_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.

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 β_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.

6. 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.