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

w203: Statistics for Data Science

Haley Farber - Elizabeth Zhou - Francisco Medina

1. Introduction

The Coronavirus disease (COVID-19), an infectious disease caused by a newly discovered coronavirus, was first identified in the United States on January 20, 2020 and has fundamentally changed the way in which we live. In order to prevent the transmission of this deadly disease, it is imperative to stay well-informed and to protect ourselves and others from infection by practicing proper sanitation and respiratory etiquette such as washing our hands and coughing into a flexed elbow. As a group, we also believe that policies are crucial in preventing the spread of the disease. More specifically, we believe that a stay-at-home mandate by states is a necessary step to reduce both the case and death rates per state. According to data from the Centers for Disease Control and Prevention, "during March 1 – May 31, 42 states and territories issued mandatory stay-at-home orders, affecting 2,355 (73%) of 3,233 U.S. counties. The first territorial order was issued by Puerto Rico (March 15), and the first state order was issued by California (March 19). Throughout the pandemic states have had many obstacles to overcome and Thanksgiving weekend was a prime example of the challenge that states continue to face in reducing population movement and close person-to-person contact outside of peoples' respective households. Despite pleas from the CDC for Americans to stay put for Thanksgiving, there were nearly 5 million travelers that had already been screened at airports by Friday November 20th, the weekend before many people planned to visit their families and friends for Thanksgiving. Overall, our group is in alignment with the CDC's claim that stay-at-home orders are crucial in reducing activities associated with community spread of COVID-19. As a result, our group will focus on analyzing the following research question: What is the effect of a statewide stay-at-home/shelter-in-place mandate on the COVID-19 case rate per 100,000 people in each state. The dataset that we are using was compiled by Majid Maki-Nayeri and contains data from the Centers for Disease Control and Prevention, Google's Community Mobility Reports, the Kaiser Family Foundation, and the COVID-19 US state policy database (Raifman J, Nocka K, Jones D, Bor J, Lipson S, Jay J, and Chan P.).

We will operationalize the variables in our models as such:

- stay_home: a dummy variable indicating whether states implemented a stay-at-home order (1) or not (0) at any point during the data collection process
- ill-risk_rate: a float variable of the percent of state population at risk for serious illness due to covid
- poverty_rate: a float variable representing the percent of the state population living under the federal poverty line in 2018
- restriction: an indicator variable, 1 if the state had a mandate that gatherings must be less than 10 people and 0 if they did not have a mandate that gatherings must be less than 10 people
- response: the difference in days between when a state announced a state of emergency due to covid and when they enacted the stay-at-home policy
- duration: the difference between the days when a state enacted their stay at home policy and when they ended the policy; essentially how long the stay-at-home/shelter in place policy was enacted for
- older_age: the percent of people in state over 65 years old.
- case_rate: the number of covid cases per 100k for each state

Note: Some states like Arkansas never had a stay-at-home policy so their response and duration columns are 0; some states like California implemented the stay-at-home policy and never ended it so their duration will effectively be the difference in days between the day that they implemented it to the final day that the data

was collected which we proxy to be Oct 15th since the data was last updated in Oct when we received it.

2. A Model Building Process

Download the Data annd Clean It

```
#Read in the Data
covid <- read.xlsx("covid-19.xlsx",sheet =2, startRow = 2,sep.names = "_")
policies <- read.xlsx("COVID-19 US state policy database (CUSP).xlsx",sheet = 5, sep.names = "_")
```

Here, we rename some of the variables for ease of use later and we mutate some of the variables to create those we can utilize later. As mentioned above, we created a dummy variable for states that implemented a stay at home/shelter in place policy. We created a variable, older_age, that represents the percent of the population in each state that is over 55 years old in a state since it is known that people of this age are more vulnerable to COVID-19.

```
#Select key variables, filter out District of Columbia, rename variables so they are easier to
#use and call later, and mutate some variables to match operationalized definitions
vars <- c("State","Case_Rate_per_100000","Percent_at_risk_for_serious_illness_due_to_COVID",
          "No_legal_enforcement_of_face_mask_mandate","Median_Annual_Household_Income",
          "Percent_Unemployed_(2018)","Percent_living_under_the_federal_poverty_line_(2018)",
          "Workplaces","Adults_55-64","65+",
          "Medicaid_Expenditures_as_a_Percent_of_Total_State_Expenditures_by_Fund",
          "State_of_emergency","Stay_at_home/_shelter_in_place","End_stay_at_home/shelter_in_place")
d <- covid %>%
     select(vars) %>%
     filter(State != 'District of Columbia') %>%
     rename(
       state = State,
       case_rate = Case_Rate_per_100000,
       ill_risk_rate = Percent_at_risk_for_serious_illness_due_to_COVID,
       no_legal_enfor = No_legal_enforcement_of_face_mask_mandate,
       income = Median_Annual_Household_Income,
       unemployment_rate = `Percent_Unemployed_(2018)`,
       poverty_rate = `Percent_living_under_the_federal_poverty_line_(2018)`,
       work_mobility = Workplaces,
       medical_exp = Medicaid_Expenditures_as_a_Percent_of_Total_State_Expenditures_by_Fund,
       home = `Stay_at_home/_shelter_in_place`,
       announce = State_of_emergency,
       policy_end = `End_stay_at_home/shelter_in_place`
     ) %>%
     mutate(legal_enfor = (no_legal_enfor - 1) * -1)  %>%
     mutate(stay_home = ifelse(home == 0,0,1))  %>%
     mutate(
          older_age= `Adults_55-64` + `65+`) %>%
      mutate(
        medical_percent = as.numeric(medical_exp))
```

```
#set the last date
last_date <- as.Date(c("2020-10-15"))
#create function to convert string to date
date_conversion <- function(x){convertToDate(x, origin = "1900-01-01")}

# Create the response and duration variables
d[12:14] <- lapply(d[12:14], date_conversion)
```

```
d <- d %>%
  mutate(response = as.numeric(home - announce)) %>%
  mutate(duration = as.numeric(policy_end - home)) %>%
  mutate(response = replace(response, response < 0, 0))%>%
  mutate(duration = replace(duration, duration >40000, 0)) %>%
  mutate(duration = replace(duration, duration < -40000, last_date-home))
```

```
#Create the dataframe 'restrict' to collect the gathering_exempt and restriction variables
#from the policies excel sheet
restrict <- policies %>%
  rename(
       state=State,
       stay_home_shelter = 4,
       end_stay_home = 7,
       gathering_exempt = 5,
       restriction = 6) %>%
       select(c(state,gathering_exempt,restriction)) %>%
       mutate(restriction = case_when(
                            restriction !=0 ~0,
                            restriction ==0 ~1)) %>%
  filter(state != 'District of Columbia') %>%
  slice(-n())
```

```
#Join 'd' and 'restrict' dataframe by the variable 'state' to have all variables in one dataframe
d3 <-left_join(d,restrict,by="state")
```

EDA

Here we will explore our main variable of interest, stay_home, which represents if a state implemented a stay at home/shelter in place policy.

```
sum(d$stay_home == 0)
```

```
## [1] 11
```

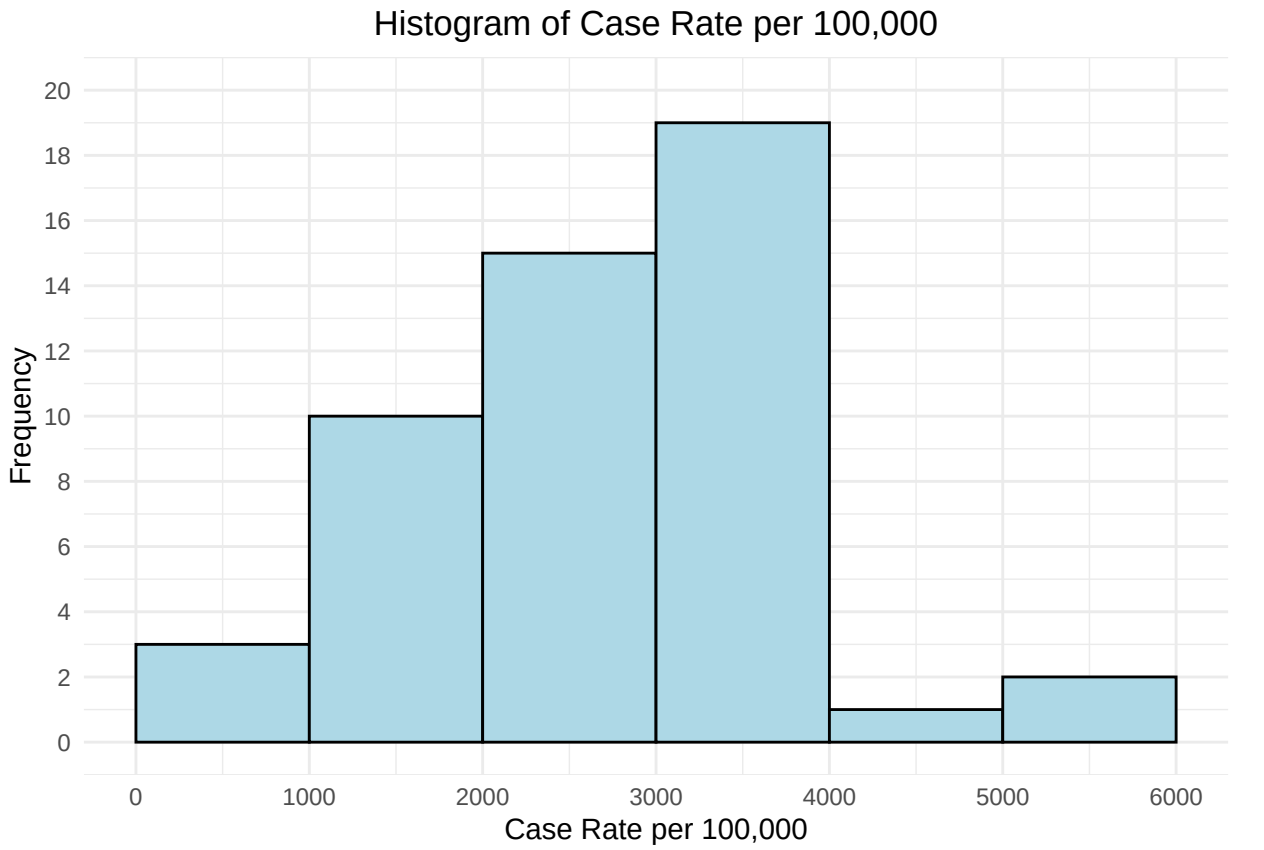
There are 11 states that never announced a stay at home policy and thus 39 states that have announced a stay at home policy before. Thus, a majority of states have implemented a stay at home or shelter in place policy before. We will look at the distribution of COVID-19 cases rates per 100,000 people and plot the case rates on a U.S. map to see if there is a geographical component to the case rates. Are states with the highest case rates clustered in one area of the United States or are they fairly distributed across the US? We will also plot a US map that shows if each state has implemented a stay at home or shelter in place policy to see if the distribution of case rates among the states matches the pattern of having a stay at home or shelter in place policy.

```
#Summary of Case Rate per 100,000
summary(d$case_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     344    2006    2726    2756    3524    5589
```

```
d %>% ggplot() +
       aes(x=case_rate)+
       geom_histogram(binwidth = 1000, boundary = 0, closed = "left",
                      color = 'black', fill = 'lightblue') +
       labs(title ='Histogram of Case Rate per 100,000',
            x ='Case Rate per 100,000', y="Frequency") +
```

```
        scale_x_continuous(breaks =  seq(0,7000,1000)) +
        scale_y_continuous(limits = c(0,20),breaks =  seq(0,22,2)) +
        theme_minimal() +
        theme(plot.title = element_text(hjust = 0.5))
```



Per 100,000 people, 10 states have between 1,000 and 2,000 cases, 16 states have between 2,000 and 3,000 cases, and 19 states have in between 3,000 and 4,000 cases per 100,000 people. There are only 3 states that have more than 4,000 cases per 100,000 people, one with cases between 4,000 and 5,000 and another with cases greater than 5,000.

```
#Looking at states with lowest and highest number of cases per 100,000
low_states = d[d$case_rate < 1000,]  %>% select(state,stay_home)
high_states = d[d$case_rate >= 4000,]  %>% select(state,stay_home)
low_states
```

```
##            state stay_home
## 19         Maine         1
## 29 New Hampshire         1
## 45       Vermont         1
```

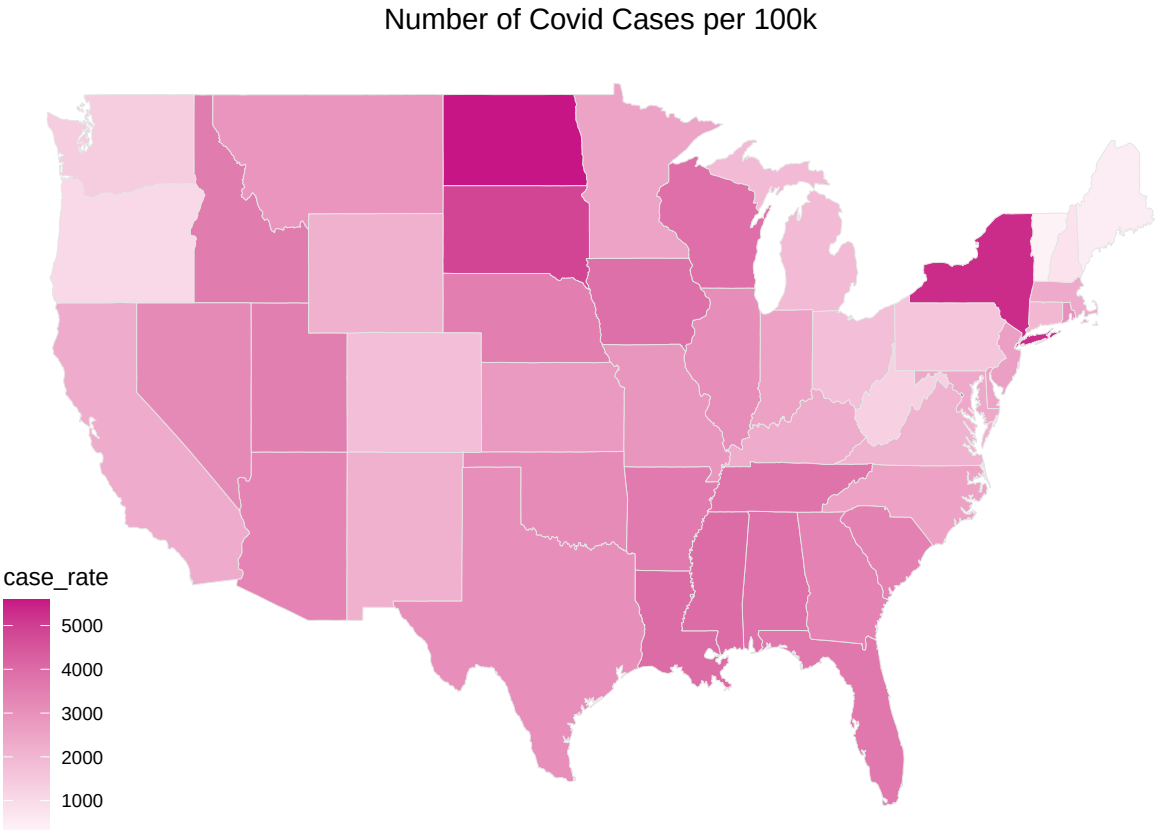
```
high_states
```

```
##           state stay_home
## 32     New York         1
## 34 North Dakota         0
## 41 South Dakota         0
```

It appears that the three states with the lowest number of cases per 100,000 people, Maine, New Hampshire, and Vermont, have all announced stay at home policies. Out of the three states with the highest number of

cases per 100,000, two did not announce stay at home policies. This supports our initial thoughts that states who have implemented stay at home policies will have lower rates of coronavirus.

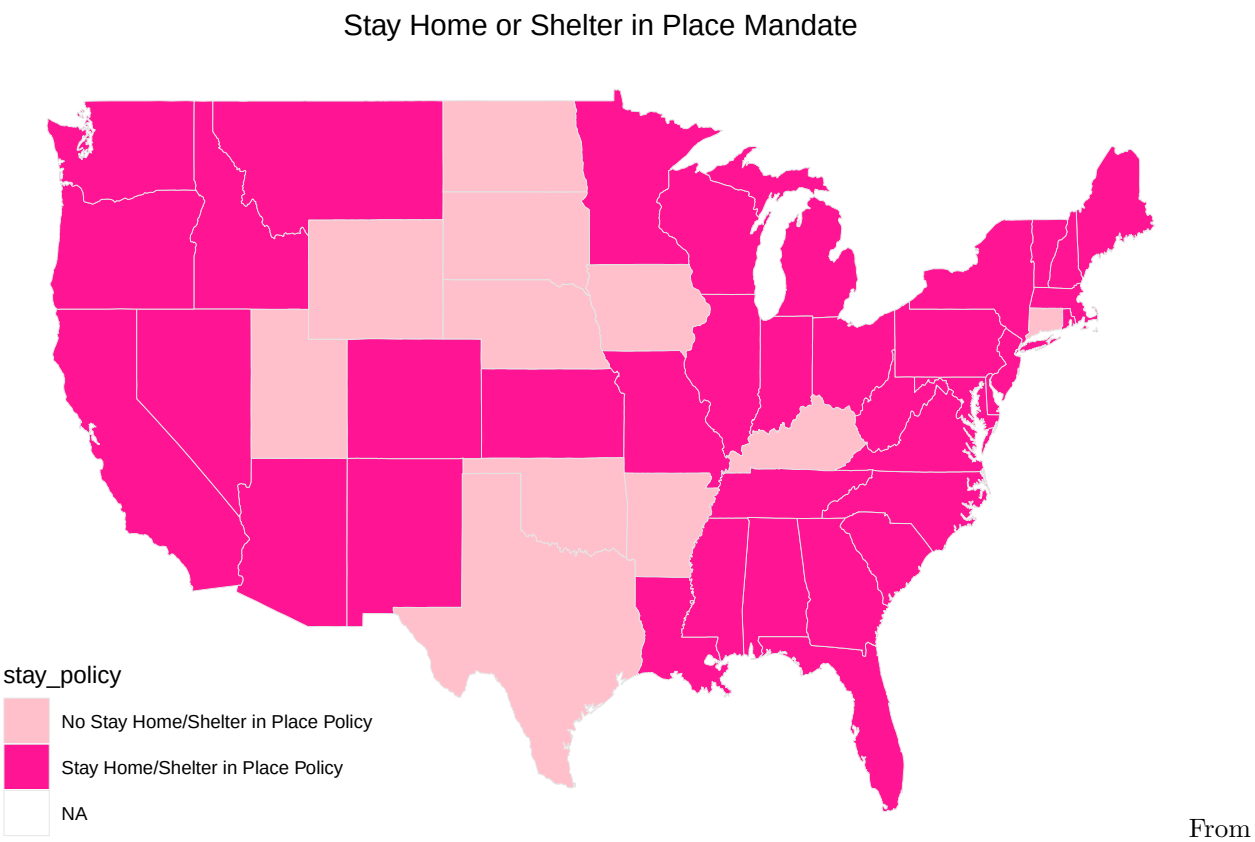
```
#Plot map of US with distribution of covid cases per 100k
us_states <- map_data("state") %>%
  rename(state = region)
d$state <- tolower(d$state)
cov_map <- left_join(us_states, d, by = "state")
p1 <- ggplot(data = cov_map,
             mapping = aes(x = long, y = lat, group = group, fill = case_rate)) +
             geom_polygon(color = "gray90", size = 0.1) +
             scale_fill_gradient2(low = "pink" , high = "mediumvioletred") +
             labs(title = "Number of Covid Cases per 100k") +
             theme_map() +
             theme(plot.title = element_text(hjust = 0.5))
p1
```



From the plot above, it does appear that there is some geographical clustering for the number of covid cases per 100k among states. Maine, New Hampshire, and Vermont have the lowest case rates and they are near one another. The west coast (California, Oregon, and Washington) all have lower case rates whereas the southern states such as Georgia and Florida appear to have higher case rates. This is clear evidence that there is some grouping of case rates geographically.

```
#Plot map of US with stay at home/shelter in place policy for each state
new_d <- d %>%
         mutate(stay_policy = case_when(
           stay_home == 1 ~ "Stay Home/Shelter in Place Policy",
           stay_home == 0 ~ "No Stay Home/Shelter in Place Policy")
         )
```

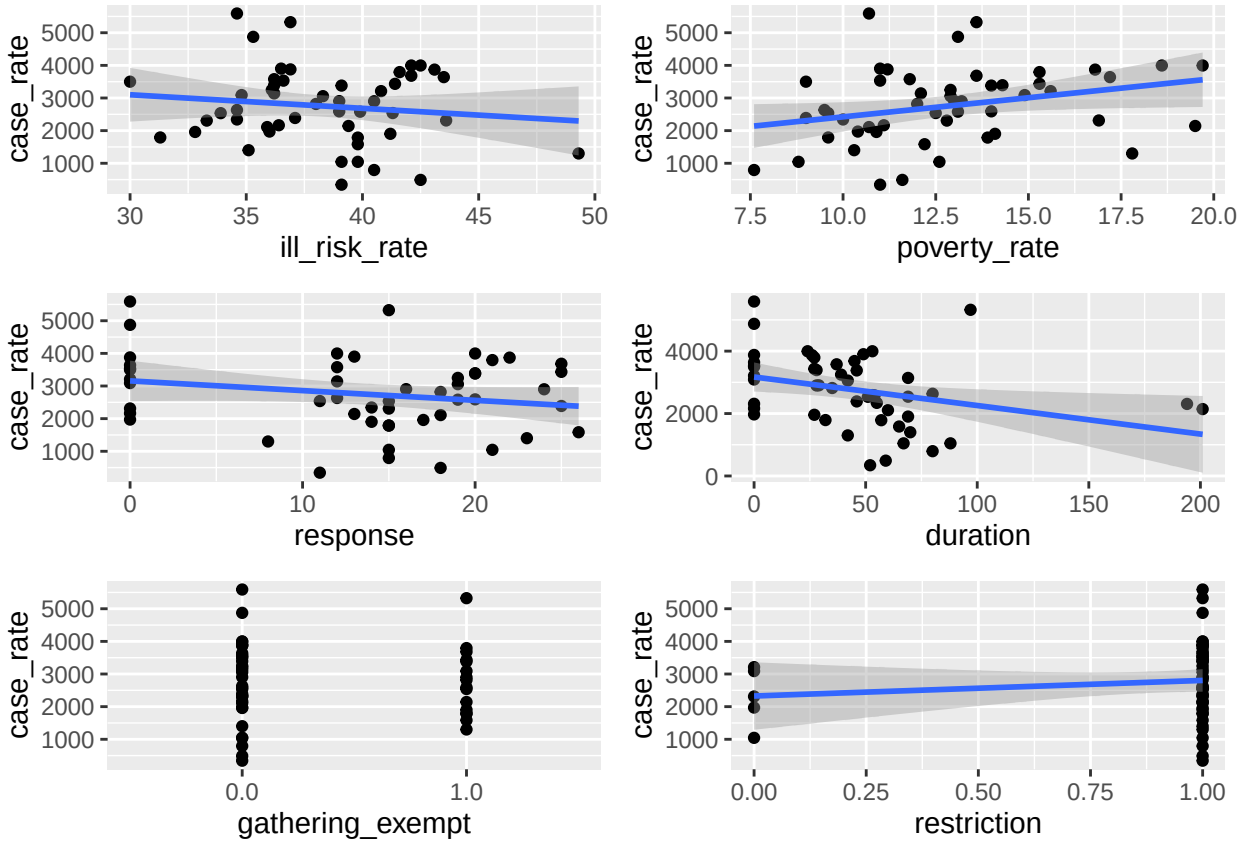
```
cov_map2 <- left_join(us_states, new_d, by = "state")
p2 <- ggplot(data = cov_map2,
             mapping = aes(x = long, y = lat,group = group, fill = stay_policy)) +
             geom_polygon(color = "gray90", size = 0.1) +
             scale_fill_manual(values = c("pink","deeppink")) +
             labs(title = "Stay Home or Shelter in Place Mandate") +
             theme_map() +
             theme(plot.title = element_text(hjust = 0.5))
p2
```



From this plot, we can see that most states have implemented a stay at home/shelter in place policy at some time. Those that haven't implemented a stay at home/shelter in place policy are geographically clustered leading us to believe there is some geographical influence to the adoption of such a policy. In order to determine variables that could affect the case rate and thus interfere with the direct causual intepretation of the stay at home/shelter in place policy on the case rate, we look at plots for covariates we plan to use in our regressions.

```
#Plot case rate vs each covariate
p1 <- ggplot(d3, aes(y=case_rate, x = ill_risk_rate)) +
      geom_point() +
      geom_smooth(method=lm)
p2 <- ggplot(d3, aes(y=case_rate, x = poverty_rate)) +
      geom_point() +
      geom_smooth(method=lm)
p3 <- ggplot(d3, aes(y=case_rate, x = response)) +
      geom_point() +
      geom_smooth(method=lm)
p4 <- ggplot(d3, aes(y=case_rate, x = duration)) +
      geom_point() +
```

```
      geom_smooth(method=lm)
p5 <- ggplot(d3, aes(y=case_rate, x = gathering_exempt)) +
      geom_point() +
      geom_smooth(method=lm)
p6 <- ggplot(d3, aes(y=case_rate, x = restriction)) +
      geom_point() +
      geom_smooth(method=lm)
grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 3, ncol = 2)
```

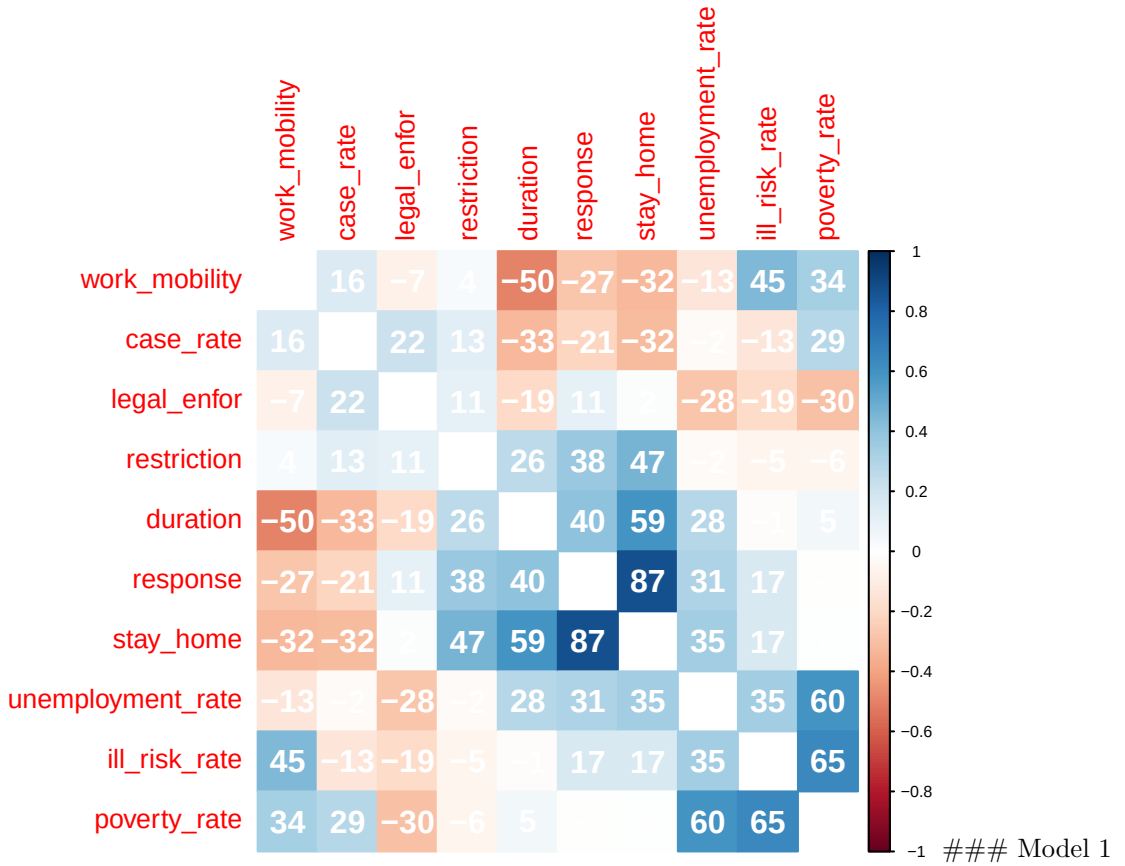


The duration of the stay at home/shelter in place policy appears to be negatively correlated with the number of cases per 100,000. This is because the longer a state implements the shelter in place policy, the the more people will stay at home, reducing the rate of cases. Additionally, poverty rate is positively related to case rate. The higher the poverty rate, the higher the case rate is because the poorer a population is, the higher the chance that they are living in places that have high population density and poor hygine conditions. Last but not least, the percent of those at risk for coronavirus, ill_risk_rate, is negatively correlated to case rate. This could be because people who are most at risk for being infected by the coronavirus will be more careful and thus stay at home or use more preventative measures, decreasing the rate of coronavirus. Overall, none of these variables completely explain the number of COVID-19 cases. With this, we will run 3 OLS models in which we regress the case rate on our key variable and other important variables.

Models

```
# Visualizing a correlation matrix
corr <- d3 %>%
  select(case_rate,ill_risk_rate,unemployment_rate,poverty_rate,response,duration,restriction,work_mobi
  drop_na()
```

```
corrplot(cor(corr),method = "color",order="AOE", tl.cex = 0.8, cl.cex = 0.5,diag=FALSE, addCoef.col = "
```



Model 1

```
# One-equation structural model

model_1 <- lm(case_rate ~ stay_home, data=d3)
coeftest(model_1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  3432.91     348.50  9.8506 4.136e-13 ***
## stay_home    -868.09     391.43 -2.2178   0.03134 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In our first model we regress our outcome variable, case_rate, on the explanatory variable, stay_home, to evaluate the effect of a state stay-at-home policy on the rate of COVID-19 per 100,000 people in each state. Our results show that if a state has a stay at home policy, it will reduce COVID-19 rates in that state by approximately 868 cases per 100,00 people. Given the small p-value of 0.03134, this result is significant at the 1% level.

Model 2

```
#Second Model

model_2 <- lm(case_rate ~ stay_home + restriction + poverty_rate + ill_risk_rate, data = d3)
```

```
coeftest(model_2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4604.844   1668.844  2.7593 0.008347 **
## stay_home     -1090.391    435.240 -2.5053 0.015923 *
## restriction    1216.833    469.691  2.5907 0.012864 *
## poverty_rate    235.289     83.345  2.8231 0.007060 **
## ill_risk_rate  -133.680     60.585 -2.2065 0.032497 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_2, model_1, test = 'F')
```

```
## Analysis of Variance Table
##
## Model 1: case_rate ~ stay_home + restriction + poverty_rate + ill_risk_rate
## Model 2: case_rate ~ stay_home
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     45 39367321
## 2     48 58013453 -3 -18646132 7.1047 0.0005233 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In our second model, we include the restriction, poverty rate, and ill risk rate variables. The restriction variable shows a positive relationship with our outcome variable. This means that states that had a mandate that gatherings must be less than 10 people have on average, approximately 1216 more cases per 100,000 people than states who don't have this restriction. This seems counterintuitive since you may expect that states that don't have a restriction on the size of gatherings to have a higher case rate, but some of the states that don't have gathering restrictions like Hawaii and Connecticut had relatively low case rates at the time that the data was collected. This result is statistically significant at the 1% level. Furthermore, the poverty rate coefficient is approximately 235 which means that an increase in a state's poverty rate by 1% will increase COVID-19 rates per 100,000 people in a state by an average of 235 cases. This is aligned with our intuition that poverty rates in each state play a role in the spread of COVID-19 because people who live in poverty may not have the luxury to stay home from work and are more likely to be exposed to the disease as a result. We find this result to be statistically significant at the 0.1% level. Finally, we include the ill risk rate because the percent of people considered high risk for COVID-19 in a state impacts case rates in any state in various ways. It is our intuition that people who are at higher risk for a serious illness due to COVID-19 will be more likely to take care of themselves and follow state guidelines such as staying home. Our results show that increasing the ill risk rate by 1% in each state will decrease COVID-19 case rates per 100,000 by approximately 133 cases. This result is statistically significant at the 1% level.

Our F-Test shows that adding these three variables decreases the sum of squared residuals, and the high statistical significance given by the p-value of the test of 0.0005233 jusifies adding them. As a result, the inclusion of these variables together improves our model.

Model 3

```
#Third Model

model_3 <- lm(case_rate~ stay_home + restriction + poverty_rate + ill_risk_rate + legal_enfor, data=d3)
coeftest(model_3, vcov = vcovHC)
```

```
##
```

```
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3685.879   1571.909  2.3448 0.023611 *
## stay_home     -1057.922    428.471 -2.4691 0.017500 *
## restriction    1088.997    462.642  2.3539 0.023110 *
## poverty_rate    274.152     73.133  3.7487 0.000515 ***
## ill_risk_rate  -136.013     54.733 -2.4850 0.016829 *
## legal_enfor     809.071    302.328  2.6761 0.010422 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_3, model_2, test = 'F')
```

```
## Analysis of Variance Table
##
## Model 1: case_rate ~ stay_home + restriction + poverty_rate + ill_risk_rate +
##     legal_enfor
## Model 2: case_rate ~ stay_home + restriction + poverty_rate + ill_risk_rate
##   Res.Df      RSS Df Sum of Sq      F   Pr(>F)
## 1     44 33686643
## 2     45 39367321 -1  -5680678 7.4198 0.009217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We proceed with our third and final model by adding the indicator variable for the legal enforcement of a mask because face masks are a preventive measure that affects the COVID-19 rates in each state. The coefficient for this variable tells us that states who have a legal enforcement of a face-mask have on average have 809 more cases than states who didn't have a face mask enforcement. We believe that this is because states who enforced the legal enforcement of a mask already had a high number of COVID-19 cases/deaths. Looking at our table, we can see that all of our results for each variable coefficient are at least statistically significant at the 1% level.

Once again, our F-Test shows that adding this variable decreases the sum of squared residuals, and the high statistical significance given by the small p-value of 0.009217 jusifies adding it to the model.

3.Limitations of Our Model

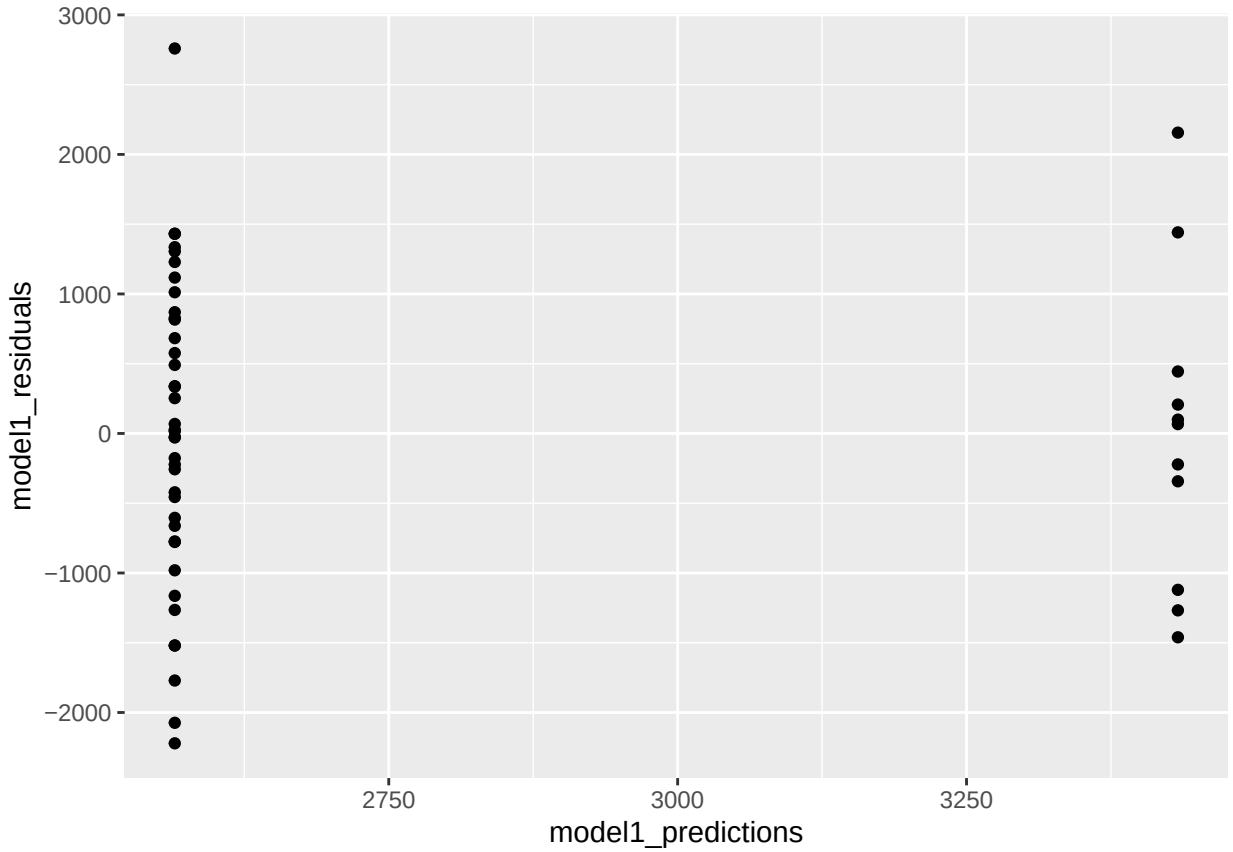
The first assumption of the classical linear model is that our data is independent and identically distributed. We believe that this assumption is not met for any of the models because of geographical clustering. The proximity of states influences our analysis because the closer states are to each other the more likely it is that they can influence each other's COVID-19 case rates by people traveling between states. Although this assumption is not met, we could potentially use more advanced statistical techniques to analyze the extent to which this is a problem for the interpretation of our results.

Another requirement of our data for the CLM to function is that there is no perfect collinearity between features. The main idea is to see if some combination of variables describes another variable. Perfect collinearity is when you can perfectly describe one input variable as a simple linear transformation of other input variables. Modeling a single piece of informtion with two variables gives rise to a problem. Usually, this is not an issue because our statistical software will automatically drop one of the variables in our model if we do have perfect multicollinearity. On the other hand, near-perfect collinearity is harder to spot because we do not get an automatic flag that a variable had been dropped. This requirement is met for all three of our models.

Assessing the rest of the CLM assumptions for Model 1

The next assumption in the classical linear model that we look at is that we have a linear conditional expectation after modeling. It will be easiest to see if our model meets this condition by plotting the model residuals vs model predictions. By doing this, we're able to pull in all the information about where we are on the input space and we're able to assess if our model is systematically predicting high or low. There might be evidence from this data that we have a non-linear relationship.

```
model1_linear_test = data.frame(model1_residuals = resid(model_1), model1_predictions = predict(model_1)
plot <- model1_linear_test %>%
  ggplot(aes(x = model1_predictions, y = model1_residuals)) +
  geom_point() + stat_smooth(se = TRUE)
plot
```



Note: Because this variable is binary, our interpretation of the plot is not very clear. Models two and three will hopefully give us a better visual to see if the assumption is met.

Next we can assess the fourth assumption of the classical linear model which is that we have constant error variance across the entire range of the x's or that we have homoskedastic errors across the range of the x's. We will do this by running a Breusch-Pagan test to see if we have any heteroskedasticity in the errors of our regression where the null hypothesis is that the error variances are all zero. This is a non-dispositive test in that if we fail to reject the null hypothesis it doesn't mean that we are good to go but it means that we don't have evidence for that problem even though that problem may exist.

```
lmtest::bptest(model_1)
```

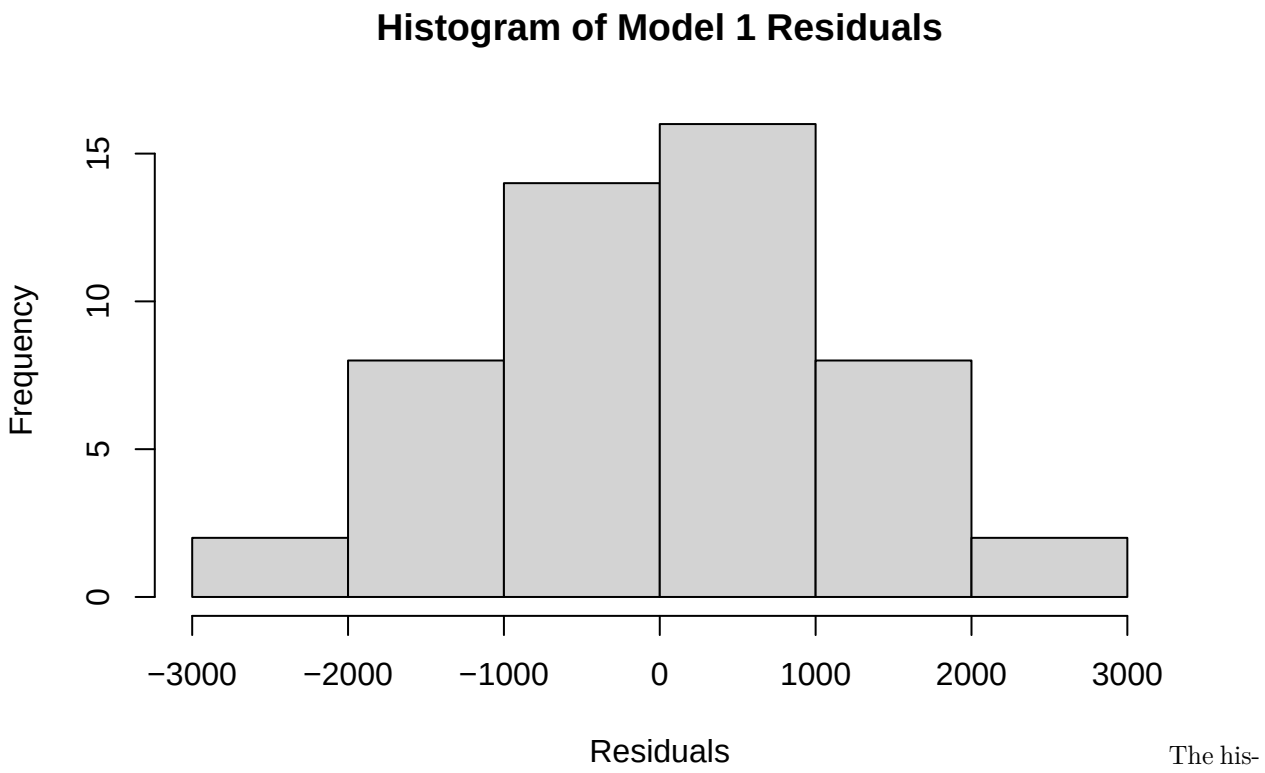
```
##
##  studentized Breusch-Pagan test
##
## data:  model_1
```

```
## BP = 0.01912, df = 1, p-value = 0.89
```

After running the test, we see that we obtain a p-value of 0.89 for our first model which means that we fail to reject the null hypothesis that the error variances are all zero. We cannot rule out the possibility that there could be heteroskedastic errors, but for now we will assume that we have homoskedastic errors.

The final assumption is that we have normally distributed errors. It will be easiest to see if we meet this assumption by displaying a histogram of the model residuals.

```
hist(model_1$residuals, main = 'Histogram of Model 1 Residuals', xlab = 'Residuals')
```

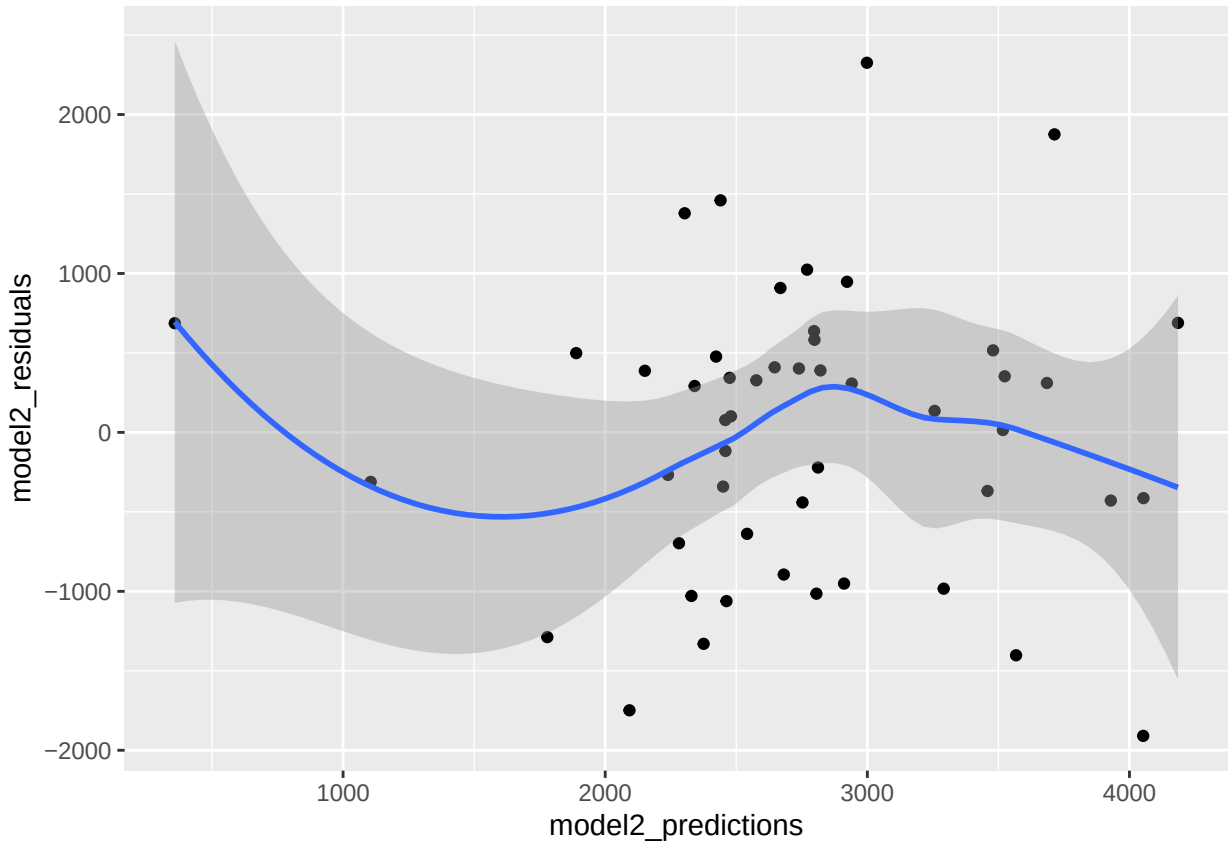


The histogram depicts that the model residuals are normally distributed.

Assessing the rest of the CLM assumptions for Model 2

Now let's assess the linear conditional expectation in a higher dimensional space by looking at the model residuals vs predicted values plot of our second model. Just as a note, our model predictions come from input values multiplied by coefficients, so these predictions have all the input data and all the dimensions that it's represented in which is boiled down to a single dimensional representation on the x-axis.

```
model2_linear_test = data.frame(model2_residuals = model_2$residuals, model2_predictions = predict(model
plot_2 <- model2_linear_test %>%
  ggplot(aes(x = model2_predictions, y = model2_residuals)) +
  geom_point() + stat_smooth(se = TRUE)
plot_2
```



After looking at our plot, it seems that we might satisfy the linear conditional expectation. Although we do not have a super clean average zero residual across the predicted values as we would expect if we had linear data, the model does not seem to be systematically under or over predicting. The plot points to something about our model that is not reflecting reality, but we believe that our model still captures a lot of that reality. As a result, we say that our model satisfies the linear conditional expectation in this case.

Next we can assess if our second model meets the assumption that our data has constant error variance across the entire range of the x's using the BP test.

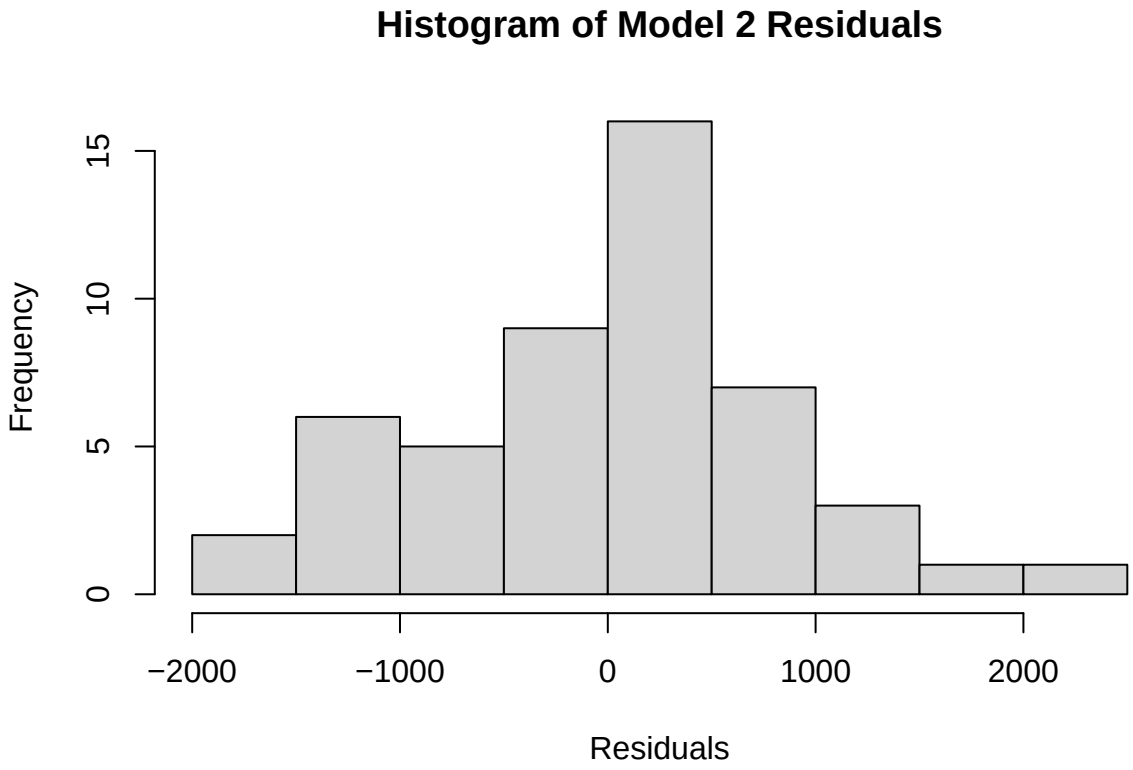
```
lmtest::bptest(model_2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_2
## BP = 2.1095, df = 4, p-value = 0.7156
```

After running the test, we see that we obtain a p-value of 0.7156 for our first model which means that we fail to reject the null hypothesis that the error variances are all zero. We cannot rule out the possibility that there could be heteroskedastic errors, but for now we will assume that we have homoskedastic errors.

We check again for normally distributed errors by looking at a histogram of model 2's residuals.

```
hist(model_2$residuals, main = 'Histogram of Model 2 Residuals', xlab = 'Residuals')
```



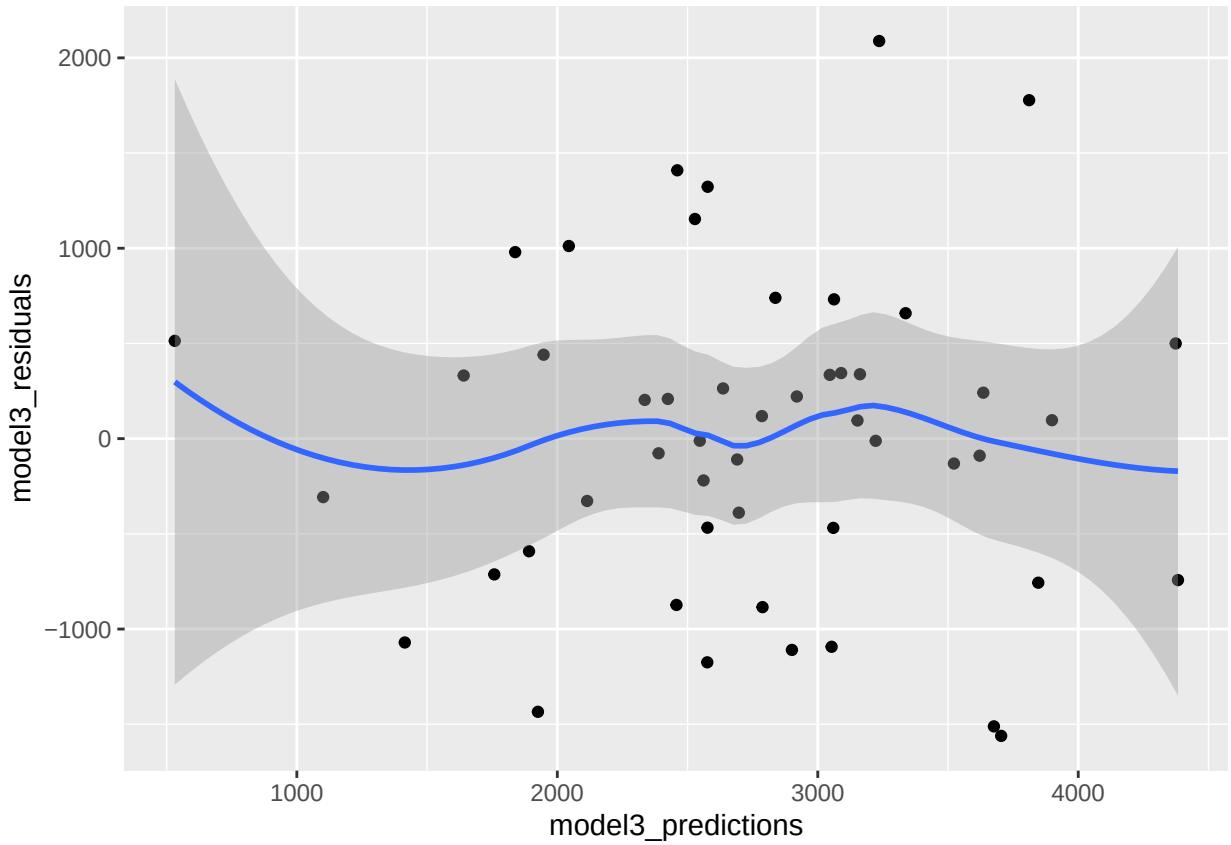
From the plot, we can see that the model residuals are normally distributed.

Assessing the rest of the CLM assumptions for Model 3

Finally, let's assess the linear conditional expectation for our last model by looking at the model residuals vs predicted values plot once again.

```
model3_linear_test = data.frame(model3_residuals = model_3$residuals, model3_predictions = predict(model

plot_3 <- model3_linear_test %>%
  ggplot(aes(x = model3_predictions, y = model3_residuals)) +
  geom_point() + stat_smooth(se = TRUE)
plot_3
```



The plot for model 3 shows us a smoother average residual that is closer to an average zero residual across the predicted values. This is further evidence that this model meets the linear conditional expectation since the model does not seem to be systematically under or over predicting.

Next we can assess that our third model meets the assumption that our data has constant error variance across the entire range of the x's using the BP test.

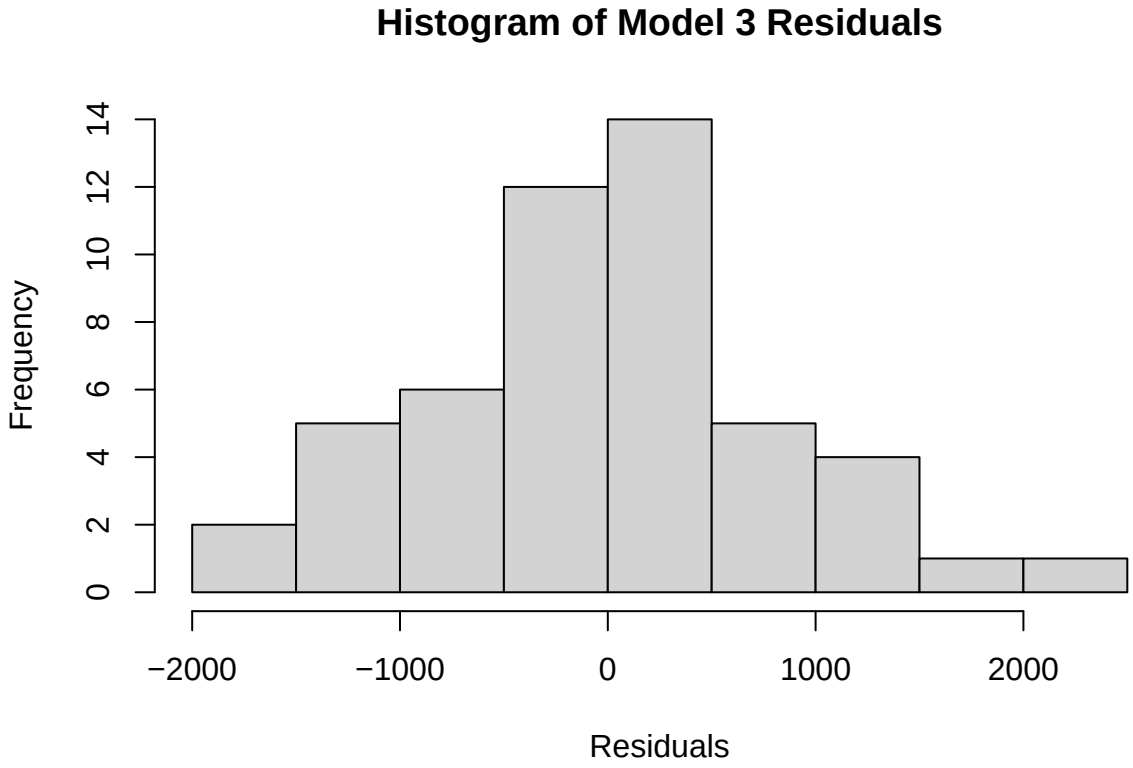
```
lmtest::bptest(model_3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_3
## BP = 2.3906, df = 5, p-value = 0.7929
```

After running the test, we see that we obtain a p-value of 0.7929 for our last model which means that we fail to reject the null hypothesis that the error variances are all zero. Again, we cannot rule out the possibility that there could be heteroskedastic errors, but for now we will assume that we have homoskedastic errors.

Finally, we check for normally distributed errors in our third model by looking at a histogram of model residuals.

```
hist(model_3$residuals, main = 'Histogram of Model 3 Residuals', xlab = 'Residuals')
```



Although the residuals are not perfectly normally distributed, they seem to be distributed enough to meet the condition for our last model.

4. A Regression Table

```
stargazer(model_1,model_2, model_3,
          type = "latex",
          title = "Regression Results",
          se = list( sqrt(diag(vcovHC(model_1))),sqrt(diag(vcovHC(model_2))) ,sqrt(diag(vcovHC(model_3)))
          column.labels = c("stay at home","stay at home + more","overfit"),
          table.placement = "H")
```

% 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 - 11:21:05 PM

Table 1: Regression Results

	Dependent variable:		
	case_rate		
	stay at home	stay at home + more	overfit
	(1)	(2)	(3)
stay_home	−868.089**	−1,090.391**	−1,057.922**
	(391.427)	(435.240)	(428.471)
restriction		1,216.833***	1,088.997**
		(469.691)	(462.642)
poverty_rate		235.289***	274.152***
		(83.345)	(73.133)
ill_risk_rate		−133.680**	−136.013**
		(60.585)	(54.733)
legal_enfor			809.071***
			(302.328)
Constant	3,432.909***	4,604.844***	3,685.879**
	(348.497)	(1,668.844)	(1,571.909)
Observations	50	50	50
R^2	0.100	0.389	0.478
Adjusted R^2	0.082	0.335	0.418
Residual Std. Error	1,099.370 (df = 48)	935.323 (df = 45)	874.989 (df = 44)
F Statistic	5.350** (df = 1; 48)	7.176*** (df = 4; 45)	8.044*** (df = 5; 44)
Note:			*p<0.1; **p<0.05; ***p<0.01

5. Omitted Variable Bias

The Effects of Omitted Variable Bias

Suppose we have two explanatory variables, A and B, and one response variable, Y. Suppose we fit a simple linear regression model with A as the only explanatory variable and we leave B out of the model.

If B is correlated with A and correlated with Y, then it will cause the coefficient estimate of A to be biased. The following diagram shows how the coefficient estimate of A will be biased, depending on the nature of the relationship with B:

Included Variable	Omitted Variable	Estimator Bias Sign	Toward or Away from Zero
stay_home	unemployment_rate	Negative Bias	Away from zero
ill_risk_rate	older_age	Positive Bias	Toward Zero
stay_home	duration	Negative Bias	Away from zero
stay_home	response	Positive Bias	Toward Zero
stay_home	medical_percent	Negative Bias	Away from Zero

	A and B are positively correlated	A and B are negatively correlated
B is positively correlated with Y	Positive Bias	Negative Bias
B is negatively correlated with Y	Negative Bias	Positive Bias

Figure 1: alt text here

OVB 1

Included variable: stay_home

Omitted variable: unemployment_rate

Estimator bias sign: negative bias

Toward or away from zero: away from zero

Explanation: The first OV we have is "Unemployment Rate" which indicates the unemployment rate for each state. If we implement an OLS model with unemployment rate and case rate per 100,000, we will find out the statistical significance and it negatively correlates with case rate per 100,000. People without a job are more likely to stay home and look for a job rather than actively leave their house for recreational activities. Thus, there could be less people leaving their houses which could help decrease the spread of covid. Not only correlated with case rate, unemployment rate is also highly correlated with the stay_home variable. With the same reasoning, higher unemployment of the state will make more people stay at home instead of wandering around because being stationary will save unnecessary spending such as transportation and dining. Because unemployment rate(B) is negatively correlated with case rate(Y), and unemployment is positively correlated with stay_home(A), we have a negative bias. Additionally, the sign of stay_home(A) is negative, so the direction of estimator sign is away from zero.

```
sprintf("a1 is %d", (a1 <-sign(cor(d$stay_home,d$unemployment_rate))))
```

```
## [1] "a1 is 1"
```

```
sprintf("a2 is %d", (a2 <-sign(cor(d$unemployment_rate,d$case_rate))))
```

```
## [1] "a2 is -1"
```

```
sprintf("Estimator bias sign is %d",(adir <- a1*a2))
```

```
## [1] "Estimator bias sign is -1"
```

```
m2 <- lm(case_rate~stay_home+unemployment_rate,data=d)
m1 <- lm(case_rate~unemployment_rate,data=d)
m3 <- lm(case_rate~stay_home,data=d)
stargazer(m1, m2,m3, type = 'text', no.space = T,
                     star.cutoffs = c(.05, .01, .001),
                     keep.stat = c("n","rsq"))
```

```
##
```

```
## ========================================================
##                            Dependent variable:
##                    ------------------------------------
##                                    case_rate
##                        (1)           (2)           (3)
## --------------------------------------------------------
## stay_home                          -963.894*     -868.089*
##                                    (402.533)     (375.319)
## unemployment_rate    -25.020        115.580
##                     (165.947)     (168.860)
## Constant           2,873.192*** 2,965.334*** 3,432.909***
##                     (795.679)     (760.100)     (331.472)
## --------------------------------------------------------
## Observations           50            50            50
## R2                   0.0005        0.109         0.100
## ========================================================
## Note:                        *p<0.05; **p<0.01; ***p<0.001
```

```
cor(d$stay_home,d$unemployment_rate)*cor(d$stay_home,d$case_rate)
```

```
## [1] -0.1101105
```

OVB 2

Included variable: ill_risk_rate

Omitted variable: older_age

Estimator bias sign: negative bias

Toward or away from zero: toward zero

Explanation: The second OV we discuss here is "older_age" which represents the percent population over 65 in each state. If we implement an OLS model with older_age and case rate per 100,000, we will find out the statistical significance and it is negatively correlated with case rate per 100,000. People who are over 65 may be more cautious and thus reduce their chance of being exposed to the public as much as possible. Therefore, older people could lead to a decrease in a state's case rate. Older_age variable is positively correlated with ill_risk_rate since older people are inherently at a greater risk for covid and thus most likely part of the group included in the ill_risk_rate group. Because older_age(B) is negatively correlated with case rate(Y), and is positively correlated with ill_risk_rate(A), we have a negative bias. Additionally, the the sign of ill_risk_rate(A) is negative, so the direction of estimator sign is away from zero.

```
sprintf("a1 is %d", (a1 <-sign(cor(d$older_age,d$ill_risk_rate))))
```

```
## [1] "a1 is 1"
```

```
sprintf("a2 is %d", (a2 <-sign(cor(d$older_age,d$case_rate))))
```

```
## [1] "a2 is -1"
```

```
sprintf("Estimator bias sign is is %d",(adir <- a1*a2))
```

```
## [1] "Estimator bias sign is is -1"
```

```
m2 <- lm(case_rate~older_age+ill_risk_rate,data=d)
m1 <- lm(case_rate~older_age,data=d)
m3 <- lm(case_rate~ill_risk_rate,data=d)
stargazer(m1, m2,m3, type = 'text', no.space = T,
                     star.cutoffs = c(.05, .01, .001),
                     keep.stat = c("n","rsq"))
```

```
##
## ================================================
##                         Dependent variable:
##                -----------------------------------
##                              case_rate
##                    (1)           (2)          (3)
## ------------------------------------------------
## older_age     -15,307.460** -19,771.210**
##                (5,244.184)   (6,696.908)
## ill_risk_rate                  57.651       -41.583
##                               (53.919)     (45.422)
## Constant      7,357.222***  6,492.612***  4,347.274*
##                (1,583.622)   (1,776.034)  (1,745.972)
## ------------------------------------------------
## Observations       50            50           50
## R2                0.151         0.171        0.017
## ================================================
## Note:                   *p<0.05; **p<0.01; ***p<0.001
```

```
cor(d$older_age,d$ill_risk_rate)*cor(d$older_age,d$case_rate)
```

```
## [1] -0.2420384
```

OVB 3

Included variable: stay_home

Omitted variable: duration

Estimator bias sign: negative bias

Toward or away from zero: away from zero

Explanation: The third OV is "duration" which represents the time between the start of a state's shelter-in-place policy and the end of a state's shelter-in-place policy. States that never implemented a policy have a duration of 0 days. If we implement an OLS model with duration and case rate per 100,000, we find that duration is negatively correlated with case rate. The longer the policy is in place, the less a state's case rate will increase because less people will go outside, reducing the risk of getting the virus. Additionally, duration is positively correlated with stay_at_home. The longer the duration is, the longer people will stay at home. Because duration (B) is negatively correlated with case rate (Y), and duration is positively correlated with stay_at_home (A), we have a negative bias. Since the sign of stay_at_home (A) is negative, the direction of estimator sign is away from zero.

```
sprintf("a1 is %d", (a1 <-sign(cor(d3$stay_home,d3$duration))))
```

```
## [1] "a1 is 1"
```

```
sprintf("a2 is %d", (a2 <-sign(cor(d3$duration,d3$case_rate))))
```

```
## [1] "a2 is -1"
```

```
sprintf("Estimator bias sign is %d",(adir <- a1*a2))
```

```
## [1] "Estimator bias sign is -1"
```

```
m2 <- lm(case_rate~stay_home+duration,data=d3)
m1 <- lm(case_rate~duration,data=d3)
m3 <- lm(case_rate~stay_home,data=d3)
```

```
stargazer(m1, m2,m3, type = 'text', no.space = T,
                     star.cutoffs = c(.05, .01, .001),
                     keep.stat = c("n","rsq"))
```

```
##
## ================================================
##                       Dependent variable:
##              -----------------------------------
##                             case_rate
##                  (1)          (2)          (3)
## ------------------------------------------------
## stay_home                    -516.792    -868.089*
##                             (462.711)    (375.319)
## duration       -9.142*       -6.030
##                (3.800)      (4.704)
## Constant     3,171.206*** 3,432.909*** 3,432.909***
##               (231.932)    (329.273)    (331.472)
## ------------------------------------------------
## Observations      50           50           50
## R2               0.108        0.131        0.100
## ================================================
## Note:                  *p<0.05; **p<0.01; ***p<0.001
```

```
cor(d3$stay_home,d3$duration)*cor(d3$stay_home,d3$case_rate)
```

```
## [1] -0.1875465
```

OVB 4

Included variable: stay_home

Omitted variable: response

Estimator bias sign: posotive bias

Toward or away from zero:towards to zero

Explanation: The next OV is "response" which represents the time between a state's state of emergency announcement and a state's implementation of a stay at home or shelter-in-place policy. In our case, we excluded the states with zero response days. If we implement the OLS model with response and case rate per 100,000, we find that response is positively correlated with case rate. The longer a state waited to announce a state of emergency, the higher its case rate became because people in the state were not made aware by their state government how dangerous the situation truly was. The earlier a state enacted the stay at home policy, the earlier people began staying in home, reducing the spread of the virus. Additionally, response is positively correlated with stay_at_home. The longer the response is, the longer people will stay at home. In our understanding, because states that react later usually have worse performance controlling COVID cases, they will stay home longer for flatten the curve. Because response (B) is positively correlated with both case rate (Y) and stay_at_home (A), we have a positive bias. And because the sign of stay_at_home is negative, the direction of the estimator is toward zero.

```
sprintf("a1 is %d", (a1 <-sign(cor(d3$stay_home,d3$response))))
```

```
## [1] "a1 is 1"
```

```
sprintf("a2 is 1", (a2 <-sign(cor(d3$response,d3$case_rate))))
```

```
## [1] "a2 is 1"
```

```
sprintf("Estimator bias sign is 1",(adir <- a1*a2))
```

```
## [1] "Estimator bias sign is 1"
```

```
m2 <- lm(case_rate~stay_home+response,data=d3)
m1 <- lm(case_rate~response,data=d3)
m3 <- lm(case_rate~stay_home,data=d3)
stargazer(m1, m2,m3, type = 'text', no.space = T,
                     star.cutoffs = c(.05, .01, .001),
                     keep.stat = c("n","rsq"))
```

```
##
## ================================================
##                       Dependent variable:
##              -----------------------------------
##                           case_rate
##                  (1)          (2)          (3)
## ------------------------------------------------
## stay_home                  -1,488.671   -868.089*
##                            (771.496)    (375.319)
## response       -29.824       35.962
##                (19.555)     (39.042)
## Constant     3,157.233*** 3,432.909*** 3,432.909***
##               (308.068)    (331.997)    (331.472)
## ------------------------------------------------
## Observations      50           50           50
## R2              0.046        0.116        0.100
## ================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

```
cor(d3$stay_home,d3$response)*cor(d3$stay_home,d3$case_rate)
```

```
## [1] -0.2765302
```

OVB 5

Included variable: stay_home

Omitted variable: medical_percent

Estimator bias sign: negative bias

Toward or away from zero: away from zero

Explanation: The last OV is "medical percent" which is the medicaid expenditures as a percent of total state expenditures by fund. We implement the OLS model with medical_percent and case rate, and we find out that medical_percent is negatively correlated with case rate. With the increase in the spending of medical percent, states will increase the effort in controlling the COVID situation, therefore there will be less case rate. Additionally, medical percent is positively correlated with stay_at_home. Higher medical percent means that the state puts more effort in controlling the cases, so they will encourage people to stay in home. Because medical_percent (B) is negatively correlated with case rate (Y) but positively correlated with stay_at_home (A), we have a negative bias. Since the sign of stay_at_home is negative, the direction of the estimator is away from zero.

```
sprintf("a1 is %d", (a1 <-sign(cor(d3$stay_home,d3$medical_percent))))
```

```
## [1] "a1 is 1"
```

```
sprintf("a2 is %d", (a2 <-sign(cor(d3$medical_percent,d3$case_rate))))
```

```
## [1] "a2 is -1"
```

```
sprintf("Estimator bias sign is %d",(adir <- a1*a2))
```

```
## [1] "Estimator bias sign is -1"
```

```
m2 <- lm(case_rate~stay_home+medical_percent,data=d3)
m1 <- lm(case_rate~medical_percent,data=d3)
m3 <- lm(case_rate~stay_home,data=d3)
stargazer(m1, m2,m3, type = 'text', no.space = T,
                     star.cutoffs = c(.05, .01, .001),
                     keep.stat = c("n","rsq"))
```

```
##
## ===================================================
##                          Dependent variable:
##                  ---------------------------------
##                                 case_rate
##                      (1)          (2)          (3)
## ---------------------------------------------------
## stay_home                       -934.473*    -868.089*
##                                 (405.298)    (375.319)
## medical_percent  -1,021.303   1,271.379
##                 (2,708.297)  (2,778.227)
## Constant        3,029.509*** 3,143.959*** 3,432.909***
##                  (744.048)    (714.422)    (331.472)
## ---------------------------------------------------
## Observations         50           50           50
## R2                 0.003        0.104        0.100
## ===================================================
## Note:                     *p<0.05; **p<0.01; ***p<0.001
```

```
cor(d3$stay_home,d3$medical_percent)*cor(d3$stay_home,d3$case_rate)
```

```
## [1] -0.1133397
```

6. Conclusion

In conclusion, despite the outbreak of the COVID-19 pandemic, there has been an ongoing debate regarding the effectiveness of the stay-at-home orders implemented in many states. In this study, we use 11 states that did not implement the order as a control group and 40 states that implemented the order as a treatment group to estimate the effect of the stay-at-home order on the number of covid cases in each state per 100,000 people. We find that, though residents in both groups have already voluntarily stayed at home, the order reduces the number of new COVID-19 cases by 832 cases per 100,000.

Our regression results back up our initial thoughts and are evidence that implementing a stay at home/shelter in place policy will decrease a state's rate of COVID-19. Variables such as the percent of those at risk for covid, the poverty rate, the unemployment rate of a state all affected the rate of covid and holding these variables constant did not diminish the effect of a stay at home policy on a state's covid rate.

As COVID-19 case rates continue to rise and without a clear date of when the distribution of vaccines will start, it is imperative that all states implement a stay at home/shelter in place policy to reduce covid rates and continue to implement them should cases continue to rise.