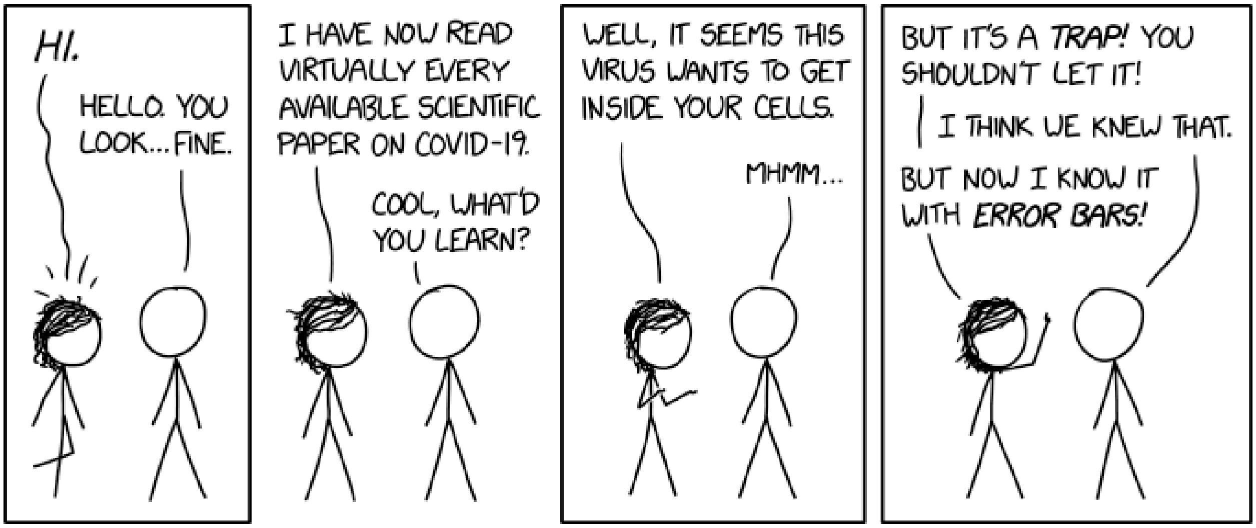
DRAFT Lab 3: A Regression Study of COVID-19

W203 - Statistics for Data Science

Harvi Singh, Christian Kutscherauer, Omar Kapur


HI.
HELLO. YOU LOOK... FINE.
I HAVE NOW READ VIRTUALLY EVERY AVAILABLE SCIENTIFIC PAPER ON COVID-19.
COOL, WHAT'D YOU LEARN?
WELL, IT SEEMS THIS VIRUS WANTS TO GET INSIDE YOUR CELLS.
MHMM...
BUT IT'S A TRAP! YOU SHOULDN'T LET IT!
I THINK WE KNEW THAT.
BUT NOW I KNOW IT WITH ERROR BARS!

(https://xkcd.com/2281/)

Contents

1. Introduction

As the COVID-19 virus has spread throughout the United States and the rest of the world, public health researchers have scrambled to understand the virus. The Center for Disease Control (CDC) has determined that there are categories of individuals at higher risk of severe illness and death from COVID-19[1], which includes older adults and people with certain medical conditions such as diabetes or respiratory illness. While a great deal has been learned, there is a lot of uncertainty given that this is a novel virus and many aspects of it are still unknown.

In this research paper we explore COVID-19 confirmed case and death data for the United States, reported as of July 6, 2020 and aggregated at the state level, along with variables on population demographics and public health measures to fight the virus, in an attempt to understand the relationships between explanatory variables and the **death rate**, which we define as the **ratio of the number of COVID-19-caused deaths per 1,000 confirmed cases of COVID-19**. We have chosen this variable for several reasons:

1) To better understand the devastation caused by COVID-19, which is ultimately borne out through the number of deaths caused by the virus.
2) To understand the impacts that affect the success of patients that are struggling with COVID-19, in order to potentially avoid or mitigate preventable deaths.

To calculate our dependent variable, we divide the count of deaths in each state by the number of confirmed cases, so as to remove the influence of larger population states impacting the model by virtue of simply having more people and higher numbers. We also multiple the death rate by 1,000 in order to make the variable easier to interpret.

Our research team weighed a primary concern when performing this analysis: the impact of COVID-19 has not been equal in all states. There are many factors for this, and one key factor is that not all states have been equally as exposed to the virus. States like Washington, New York, and California, with large cities that are major international destinations, were among the first to see infections spike and the death rate rise considerably as healthcare systems became overwhelmed and officials scrambled to respond to the virus. Other states,such as Florida, Arizona, and Texas are now starting to see significant increases in infection rates at the time the data was collected.

Therefore, because we are not performing this analysis at the end of the pandemic when all states have had sufficiently equal exposure, we will need to consider this aspect as we consider which variables fit the variation in our dependent variable better and whether it makes intuitive sense given the domain knowledge on COVID-19. In fact, we found that population density and length of time of public facemask mandates had the strongest associations to death rate - we believe this relationship exists for several reasons:

1) The states with large cities were the first to contract the virus and see spikes in transmission that caused their healthcare systems to be overwhelmed.
2) These states then had to implement response measures such as requiring face masks and closing businesses.
3) A number of states with lower population density were slower to see cases spread to their states, but now are seeing impacts rise. If we performed the same analysis at the end of the pandemic, it would be interesting to see if variables such as the population at risk would have a stronger relationship to the death rate.

We are aware of other studies that are considering broader datasets and controlling for more factors, which are finding that the density of people in an area is actually negatively correlated with death rate[2].

We are also aware of concerns that exist around the accuracy and completeness of COVID-19 data; counts of confirmed cases are only accurate if the population has been adequately tested, and counts of deaths are only accurate if COVID-19 deaths have been reported. We must accept these qualifications and assume that the

[1] https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/index.html

[2] *Does Density Aggravate the COVID-19 Pandemic?*, Hamidi, Sabouri, Ewing, Published June 18, 2020. Accessed at https://doi.org/10.1080/01944363.2020.1777891 on August 2, 2020.

data we have represents an accurate representation of the cases and deaths to analyze the associations between variables.

Data Cleaning

Our data comes from a variety of sources, which have been collected together in an Excel file. Our first step is to import this data into R, convert column names into code-friendly strings, and ensure that numbers and dates are formatted appropriately.

Our data source provides 52 rows - however upon inspection we found that one state - Arizona - appears twice (once as 'Arizona' and again as 'arizona'). All columns for these two rows are the same with the exception of total cases and total deaths. In order to assess which row is correct, we checked the source of those columns: the Center for Disease Control (CDC). We found that the correct values as of July 6, 2020 for Arizona's cases and deaths were much closer to the sum of these columns as opposed to either record individually. Therefore, we added the values of both rows together for both the total cases and total deaths columns for Arizona.

Additionally, the dataset is entirely comprised of states with one exception: the District of Columbia, which does not share many of the same characteristics as a state. For variables such as population density and demographics, this creates a clear outlier that would not be a fair comparison to the other data points, as can be seen from the plot in Figure 1 (the red point is the District of Columbia and blue points are the states):

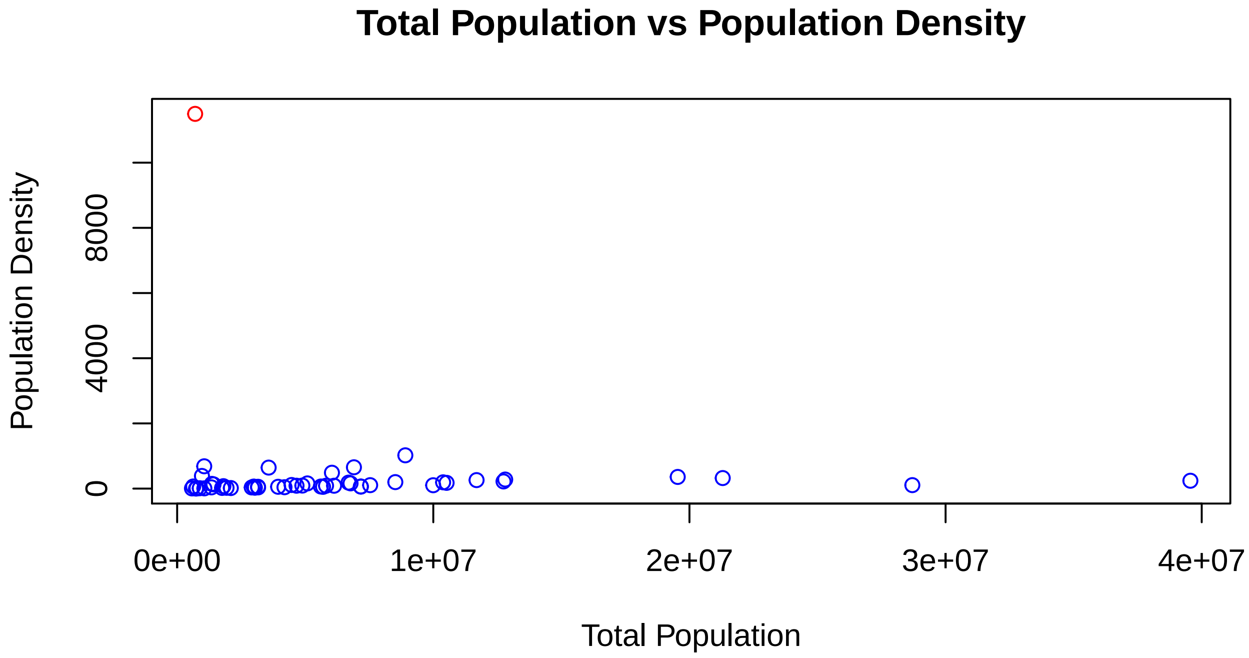


Figure 1: Relationship between total population and population density for all samples; blue dots show values for states, red dot is Washington, D.C.

Since the District of Columbia is not a state, and we want the dataset to be as close to iid as possible, we will not include the data for DC in this study. Note that to be iid, the data should be from the same distribution, and all the other datapoints are from the other (50) states.

After cleaning our data and filtering it to the key variables and covariates we are going to use in our model, we now have a total of 50 rows.

2. Data Exploration and the Model Building Process

After this step, we're ready to select which variables we used in the models. The dataset contains a number of variables that appear interesting, but on further inspection would actually be outcome variables. For example, we explored including the number of days from when each state instituted a stay-at-home or shelter-in-place policy to when they lifted that policy (if such as policy was instituted at all), however this would likely represent an outcome variable rather than an explanatory variable. The following is a walkthrough of the variables we decided to select for our models:

Dependent Variable:

- **Death Rate**
  As our dependent variable, the death rate is calculated as the number of total deaths divided by the number of total cases in a state, which is multiplied by 1000 to result in the death rate per 1,000 confirmed cases. We will refer to this variable simply as "death rate" throughout this paper. We should note that the death rate is dependent on a state's ability to test for confirmed cases and diagnose deaths due to COVID-19 - as well as their transparency in reporting those numbers. Figure 2 provides a histogram of this variable.

  The calculated death rate variable has a minimum value of 7.25 and a maximum value of 92.79, as well as a mean of 35.65 and a standard deviation of 22.41.

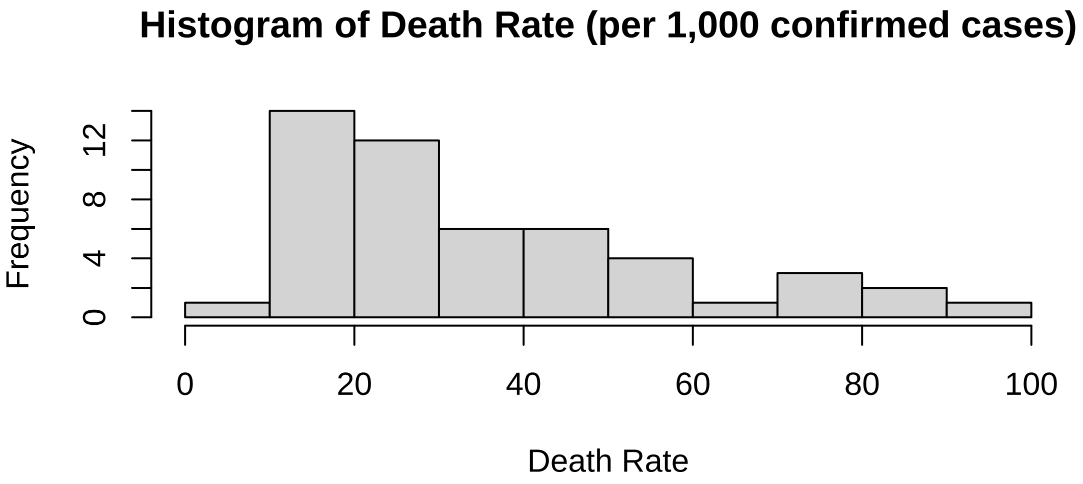


Figure 2: Distribution of dependent variable (Death Rate)

Explanatory Variables:

- **Population Density**
  As discussed previously, given that the data is from the middle of the pandemic, what we know about the places that have been hit the hardest so far is that they have dense areas with major cities. States such as New York, Washington, and California are some of the more populous areas and were the first to see their healthcare systems become overwhelmed. Therefore, Population Density is likely a strong predictor of death rate *at this stage* of the pandemic. Figure 3 provides a histogram of this variable to illustrate the distribution. The population density variable has a minimum value of 1.11 and a maximum value of 1021.27, as well as a mean of 170.56 and a standard deviation of 208.24.

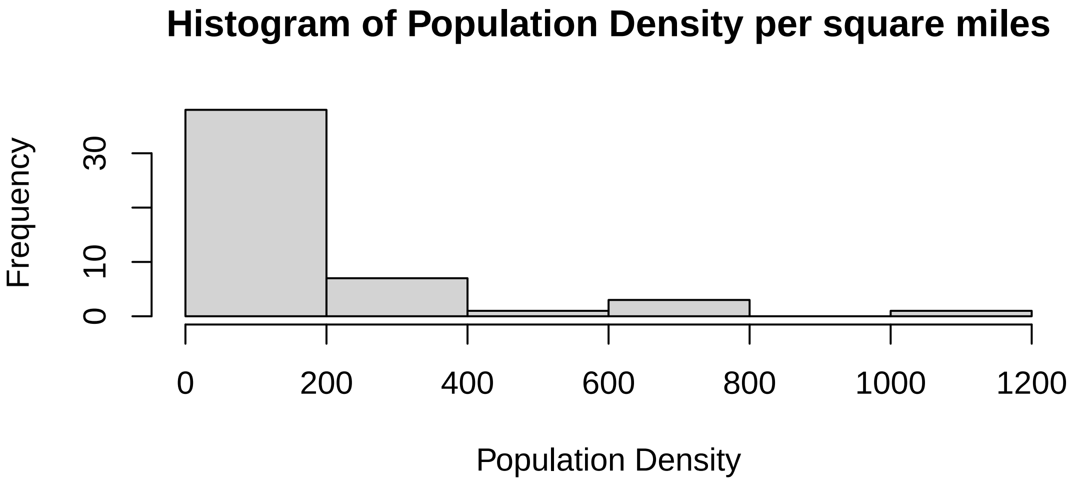


Figure 3: Distribution of population density per square mile

- **Number of Days in which facemasks have been required by employees in public-facing businesses** In order to represent actions taken by public health officials, we are including an explanatory variable that is the number of days in which a state has mandated the use of facemasks in public-facing businesses. While we expect that this variable should have a negative correlation to the dependent variable - that is, more days of mandates that protect the public from the virus will correlate with a lower death rate as the public and high risk populations are more informed and protected, and the healthcare system sees less of a surge.

  However, given how the virus has played out and the reactionary stance that governments have taken due to the lack of knowledge about the virus, we may also see a positive correlation - that is, states that were the first to be impacted by COVID-19 had to then implement public health restrictions (and therefore will have a greater value for this variable), while also seeing the virus spread and death rate rise.

  This variable is calculated by calculating the difference between the day on which a face mask mandate was implemented at the state level and the day on which the COVID-19 data for this study was collected (July 6, 2020). If a state had not implemented a face mask mandate, a value of 0 days is recorded. The days of facemask mandate variable has a minimum value of 0 and a maximum value of 94, as well as a mean of 52.16 and a standard deviation of 29.50. Figure 4 provides a histogram of this variable.

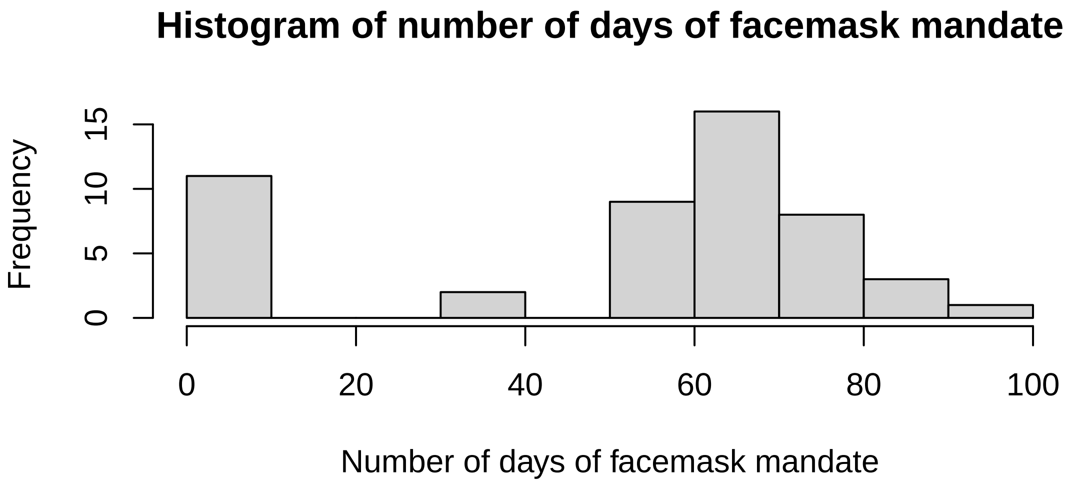


Figure 4: Distribution of the number of days that states have required facemask mandates.

- **Unemployment Claims** Our research team debated whether to include the number of weekly unemployment claims in a state - in part because we were unsure if this would be an explanatory variable or an outcome variable. That is to say - if the number of unemployment claims goes up, is that because businesses are seeing signs of the virus and shutting down to avoid the coming surge of cases, or because the surge - and increase in death rate - has already happened and businesses are closing as a result? While we determined other variables to be the outcome of a surge in infections, in a number of states businesses began feeling the economic hit of lower business as a result of consumers proactively not wanting to go into public places, and so we felt this would likely be a predictor rather than an outcome variable. Figure 5 provides a histogram of this variable. The unemployment claims variable has a minimum value of 190 and a maximum value of 823, as well as a mean of 460.40 and a standard deviation of 139.45.

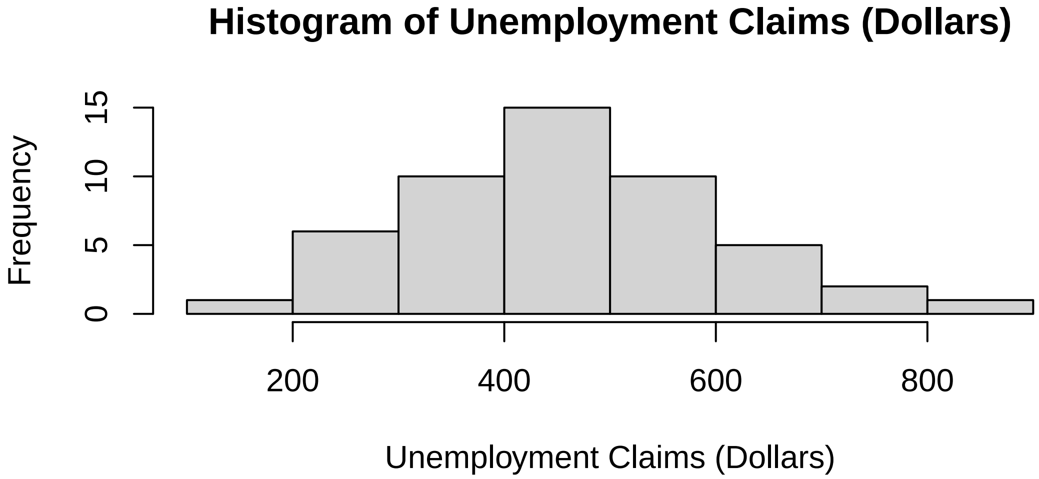


Figure 5: Distribution of state unemployment claims (in dollars).

- **Percent living under the federal poverty line** Based on reporting and indications from public health groups, we know that people who have less access to healthcare are more susceptible to the worst effects of the virus. Therefore, we decided to include poverty as an explanatory variable. Figure 6 provides a histogram of this variable. The poverty variable has a minimum value of 7.60 and a maximum value of 19.70, as well as a mean of 12.85 and a standard deviation of 2.83.

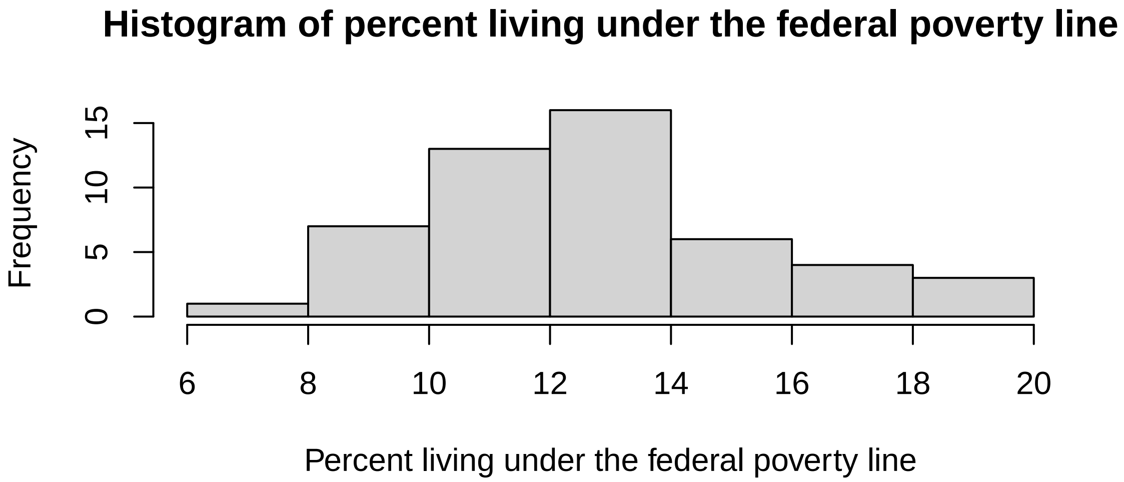


Figure 6: Distribution of state population that is living under the federal poverty line.

- **Percent of population that is 65 or older** Based on guidance from the CDC, we know that the elderly are more susceptible dying from the coronavirus. Figure 7 provides a histogram of this variable. The percent of elderly population variable has a minimum value of 0.11 and a maximum value of 0.21, as well as a mean of 0.17 and a standard deviation of 0.02.

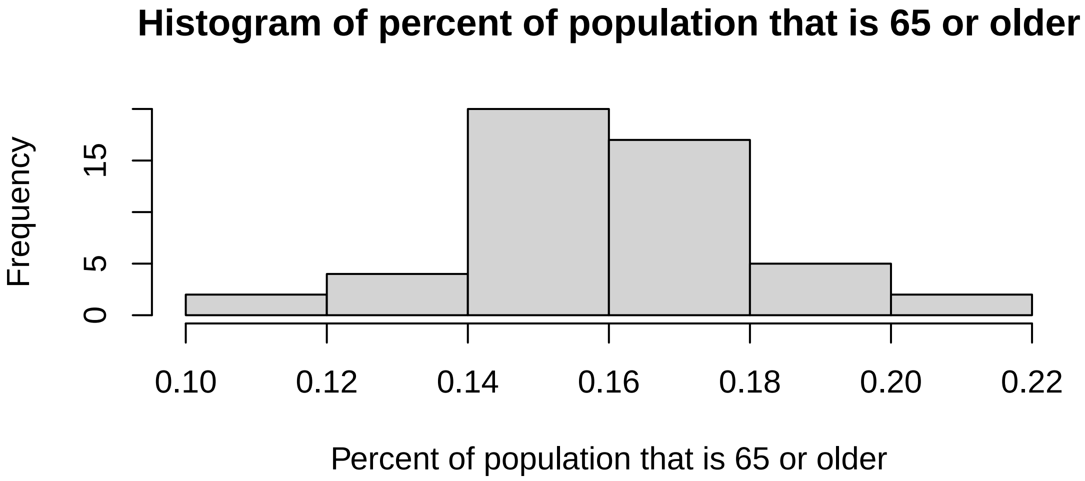


Figure 7: Distribution of the percent of state population that is 65 or older.

- **Percent at risk for serious illness due to COVID-19 (high risk population)** Also based on guidance from the CDC, we know that there is a portion of the population that has a higher risk of serious illness (and death) from COVID-19. Figure 8 provides a histogram of this variable. The percent of elderly population variable has a minimum value of 30 and a maximum value of 49.30, as well as a mean of 38.27 and a standard deviation of 3.61.

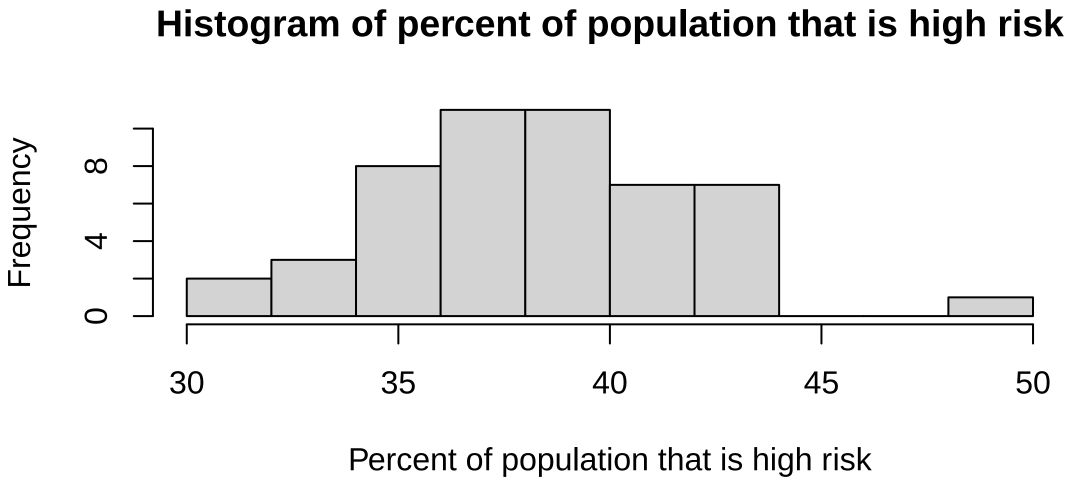


Figure 8: Distribution of state population that is considered high risk per CDC guidance.

In addition to considering summary statistics and distributions of these explanatory variables, we checked their correlations so as to consider multicollinearity, an important consideration when selecting which variables to use in each model. From the correlation heatmap shown in Figure 9, we can see that population density does indeed have a strong correlation to the dependent variable, as does the number of days of facemask mandate (indicating we are in fact seeing the positive correlation between days of facemask mandate and death rate, a sign that this variable has a reactive rather than proactive relationship with our dependent variable). We also see some evidence of potentially strong (but not likely perfect) multicollinearity between the high risk population and those in poverty and the elderly.

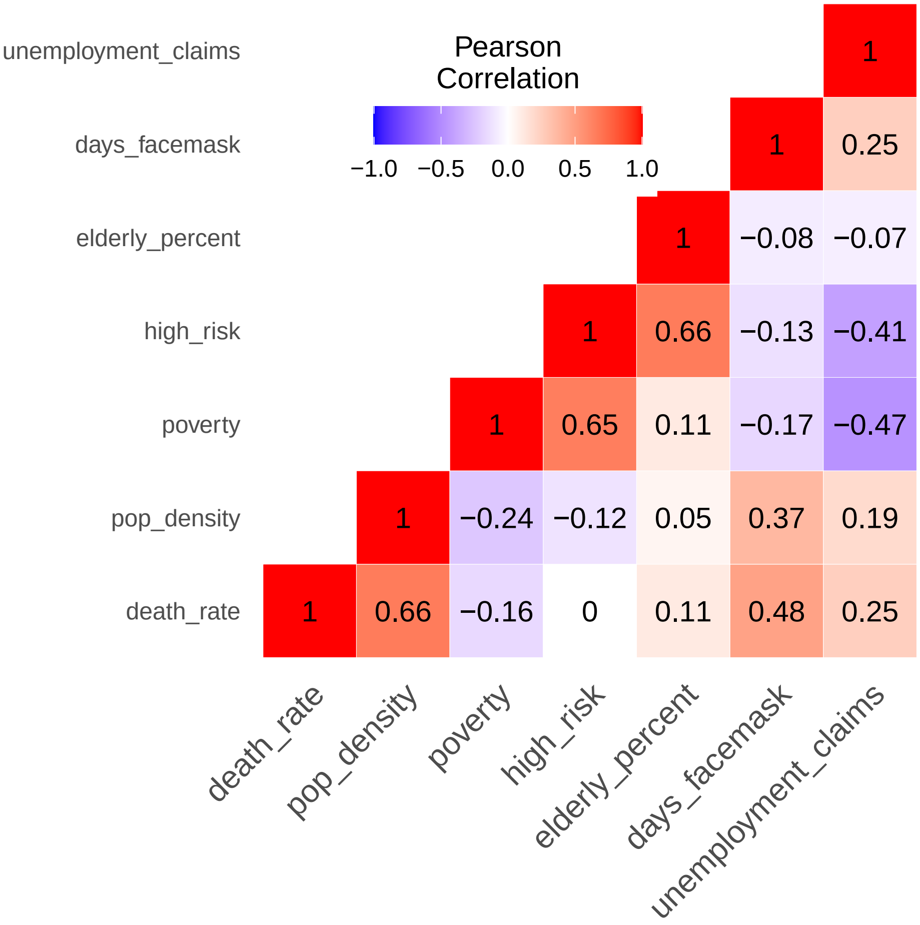


Figure 9: Pearson correlation heatmap for variables (includes dependent variable and explanatory variables)

We can also view variable distributions and relationships with the scatter plot matrix shown in Figure 10.

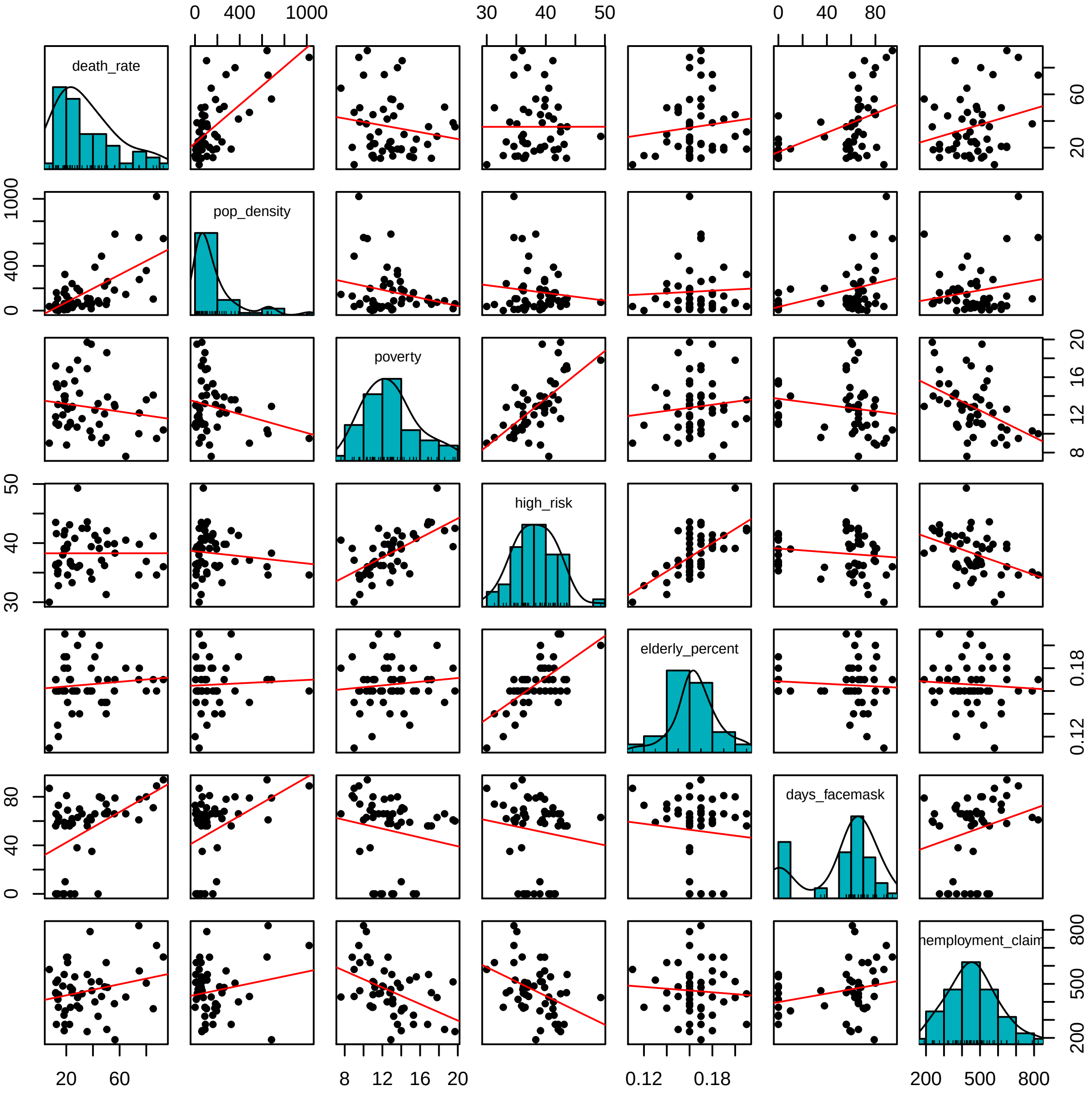


Figure 10: Variable distributions and relationships between each variable.

Given the conditions and variables described above, we developed the following series of linear regression models:

Baseline Model - Version 1
The strongest relationship between the explanatory variables and the death rate is with **population density** - therefore we will consider that our baseline model (version 1). While this model is not a multivariate regression, we feel that this gives us the best baseline from which to begin adding variables. Therefore our baseline model is:

$$deathrate = \beta_0 + \beta_1(PopulationDensity) + u$$

Improvement Model - Version 2
Our improvement on the baseline model involves adding two additional explanatory variables:

$$deathrate = \beta_0 + \beta_1 PopulationDensity + \beta_2 DaysOfFacemaskMandate + \beta_3 UnemploymentClaims + u$$

We chose to add the number of days of facemask mandate and unemployment claims because they are also variables that give us a stronger explanation of the death rate given our snapshot in the midst of the pandemic, not at the end of the pandemic.

Improvement Model - Version 3
Our second improvement model involves adding the remaining explanatory variables that we feel would be very relevant particularly at the end of the pandemic, but also should have some relevance to predicting the dependent variable given the :

$$\begin{aligned} deathrate = \beta_0 + \beta_1 PopulationDensity + \beta_2 DaysOfFacemaskMandate + \beta_3 UnemploymentClaims \\ + \beta_4 HighRisk + \beta_5 Elderly + \beta_6 Poverty + u \end{aligned}$$

3. Assessing the assumptions for CLM

We'll be explicitly assessing model2 for meeting the 6 assumptions of OLS.

CLM1. Linearity:

Since we are assuming a model that has the following structure, this assumption is met. It is always possible to fit a line through any data provided the error term has no limitations.

$$deathrate = \beta_0 + \beta_1 PopulationDensity + \beta_2 DaysOfFacemaskMandate + \beta_3 UnemploymentClaims + u$$

CLM2. Random Sampling (iid):

One of the primary assumptions for Classical Linear Models is that the data is a random sample of the population and is independent and identically distributed (iid). It is worth noting that the data we are using for this exercise consists of the entire population of the contiguous United States (plus Alaska and Hawaii). Therefore, this data does not represent a perfectly random sample of a population, but rather a snapshot of the population as a whole at a certain point in time. If we were to assume that the data represents a sample of the entire world's population, it would still not meet the classical requirements of iid, since the samples would be clustered within the USA, and would not take into account any cultural, demographic, or other differences that exist in other countries. Similarly it would be incorrect to say that all states are perfectly independent (would neighboring states not exhibit some clustering effect?).

However, from a practical standpoint, this sample represents the the best available snapshot of the country's population at the time that this sample was collected. Also, the data from different states may not be perfectly independent, but we also know that as far as Covid-19 is concerned, most neighboring states have had varying experiences (California's experience has been very different from Nevada for example). Therefore it would be reasonable to assume that the state data is nearly iid.

Further, it is important to note that even if this data-set was not perfectly iid and would not allow us to build a causal model that is BLUE, we can build an associative model that tracks the best fit line in the population.

CLM3. Multicollinearity:

First we will observe the correlation between our IVs in the model. Although this will not be sufficient to check for the presence of collinearity, it provides a good understanding on the level of correlation between IV pairs, as shown in Figure 11.

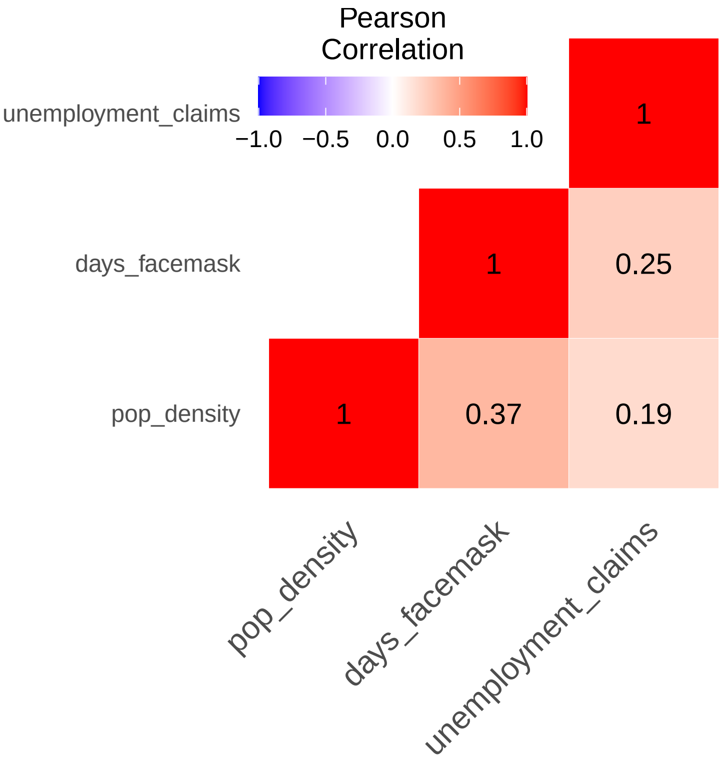


Figure 11: Pearson correlations for model2 explanatory variables only.

From the above correlation plot, we can see that there is relatively little correlation between all factors. To further check for collinearity we will obtain the Variance Inflation Factor (VIF), and if a VIF of 10 or more is found, it will indicate the presence of multicollinearity.

```
##         pop_density       days_facemask unemployment_claims 
##            1.177170            1.209521            1.080644
```

Since the VIF values are very low, we can assume there is no significant presence of collinearity in our model 2.

CLM4. Zero Conditional Mean:

For model2, the residuals are well behaved in that that the mean residual value stays close to 0. We can therefore say that model2 approximately meets the zero conditional mean assumption. Figure 12 shows the Residuals vs Fitted plots for all three models.

Note that for all models 1, 2, and 3, the mean residual value deviates from 0 to varying degrees. We can therefore conclude that we seem to have a violation of zero-conditional mean for all of our models. Because we have a large sample (over 30 points), we can rely on OLS asymptotics. If we set aside causality and just look for the best fit line, exogeneity tells us that our estimates are consistent. In other words, we get exogeniety for free since we only looking to track the line that best fits the population.

CLM5. Homoscedasticity:

Homoscedasticity essentially implies that the Variance of the error term is Constant. Figure 13 contains 2 plots - Residuals vs Fitted and Scale-Location - showing the characteristics of the error term for model2. We will be using these to visually gauge the variance of the residuals.

Visually we can see the variance in the Residuals vs Fitted curve does not change much. It appears that the

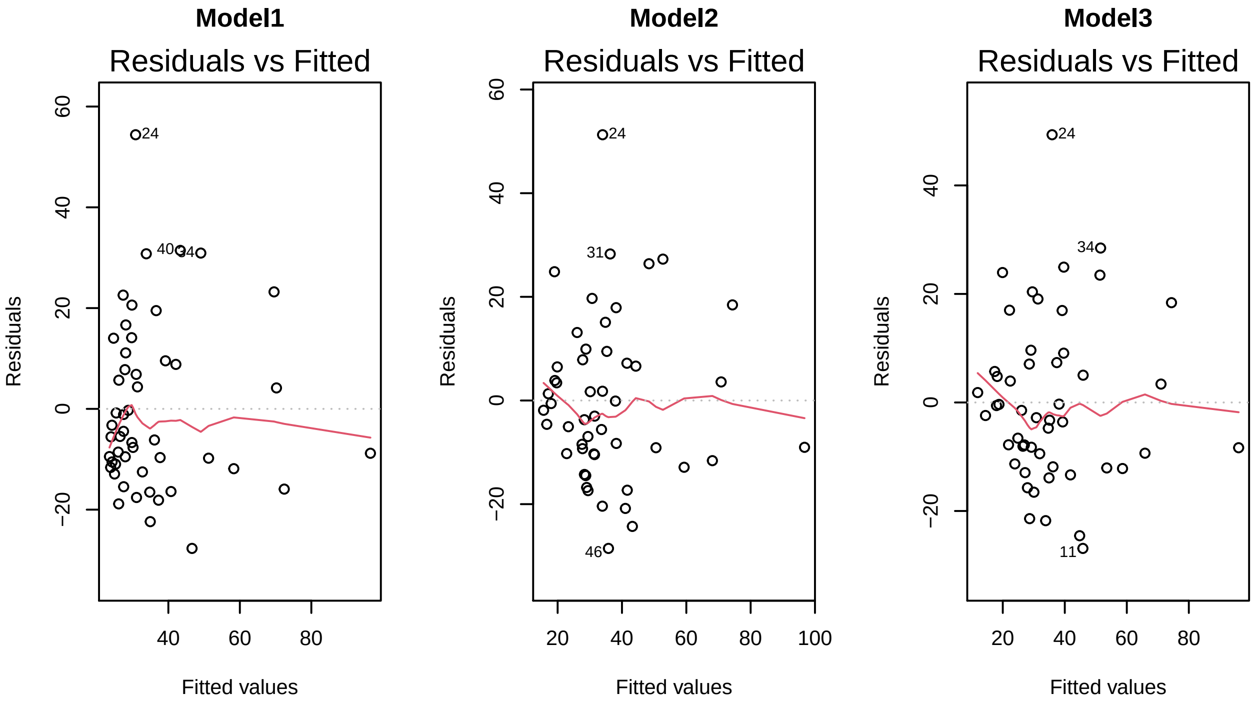


Figure 12: Residual vs Fitted plots for all three models.

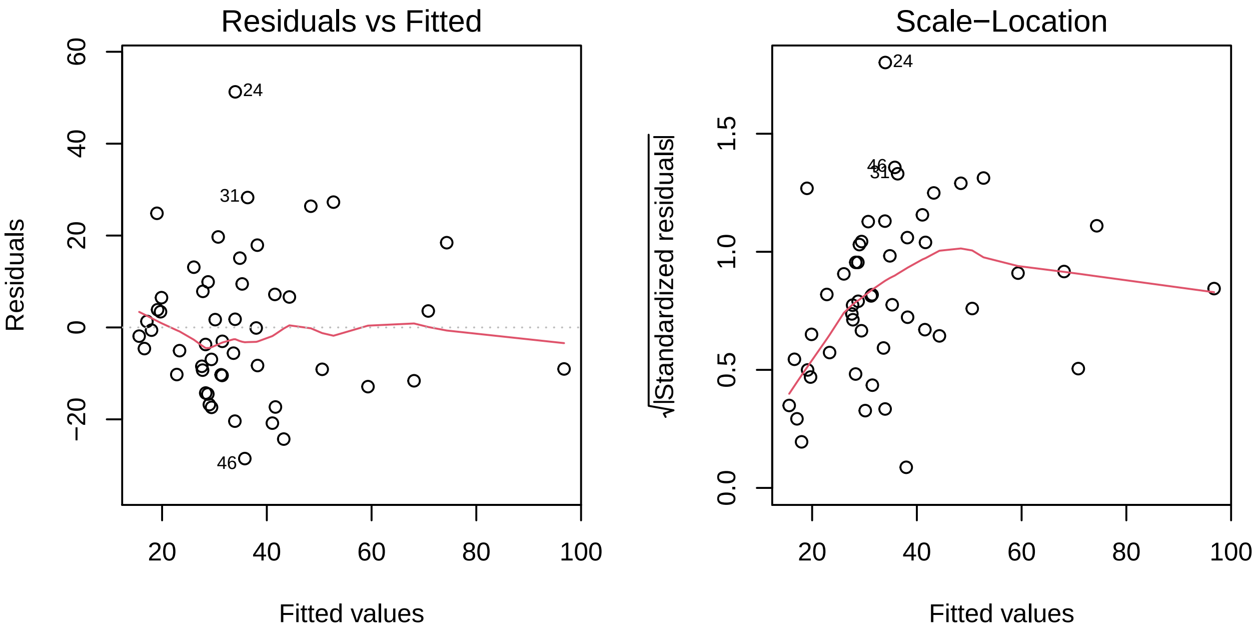


Figure 13: Residual vs Fitted and Scale-Location plots for model 2 only.

variability in the residuals decreases slightly as we increase the fitted values, but this could just be a result of fewer data points to the right. The average values of residual does appear flat-lined,which would indicate constant variance. Based on a visual inspection of the residual plots for model2, we can conclude that model2 exhibits approximate homoscedasticity.

In order to get further confirmation of our result, we will conduct the Breusch-Pogan Test to check for heteroskedasticity. Note the Null hypothesis is that homoscedasticity exists.

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 5.2768, df = 3, p-value = 0.1526
```

Since the p value is close to 0.15, we are unable to reject the null hypothesis. We therefore cannot support the alternate hypothesis that the model exhibits heteroskedasticity. This result is consistent with the result of our visual inspection of the residual plot above.

Note that for models 1 and 3, the variance appears to vary more in comparison with model2, but still remains relatively constant. This observation is consistent with the results of the Breusch-Pogan Test below.

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 0.21193, df = 1, p-value = 0.6453
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 6.6421, df = 6, p-value = 0.3552
```

Although, we do not seem to violate the homoscedasticity assumption for any of our models, we have the following comparison for model2 with and without robust standard errors. Note that the robust standard errors are almost identical to the (actually lower) than the standard errors. This further confirms our earlier argument that model2 exhibits homoscedasticity.

```
##
## t test of coefficients:
##
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         9.734575   7.924780  1.2284   0.2256
## pop_density         0.059862   0.011848  5.0524 7.37e-06 ***
## days_facemask       0.190125   0.079679  2.3861   0.0212 *
## unemployment_claims 0.012572   0.016107  0.7805   0.4391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = death_rate ~ pop_density + days_facemask + unemployment_claims,
##     data = data_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.541 -10.288  -2.455   7.687  51.275
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         9.73458    8.29943   1.173   0.2469
## pop_density         0.05986    0.01206   4.962 9.99e-06 ***
## days_facemask       0.19012    0.08633   2.202   0.0327 *
## unemployment_claims  0.01257    0.01726   0.728   0.4701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.21 on 46 degrees of freedom
## Multiple R-squared:  0.5089, Adjusted R-squared:  0.4769
## F-statistic: 15.89 on 3 and 46 DF,  p-value: 3.15e-07
```

CLM6. Normality of Error:

To determine whether the residuals follow a normal distribution, we will assess a normal probability plot for model2, as shown in Figure 14.

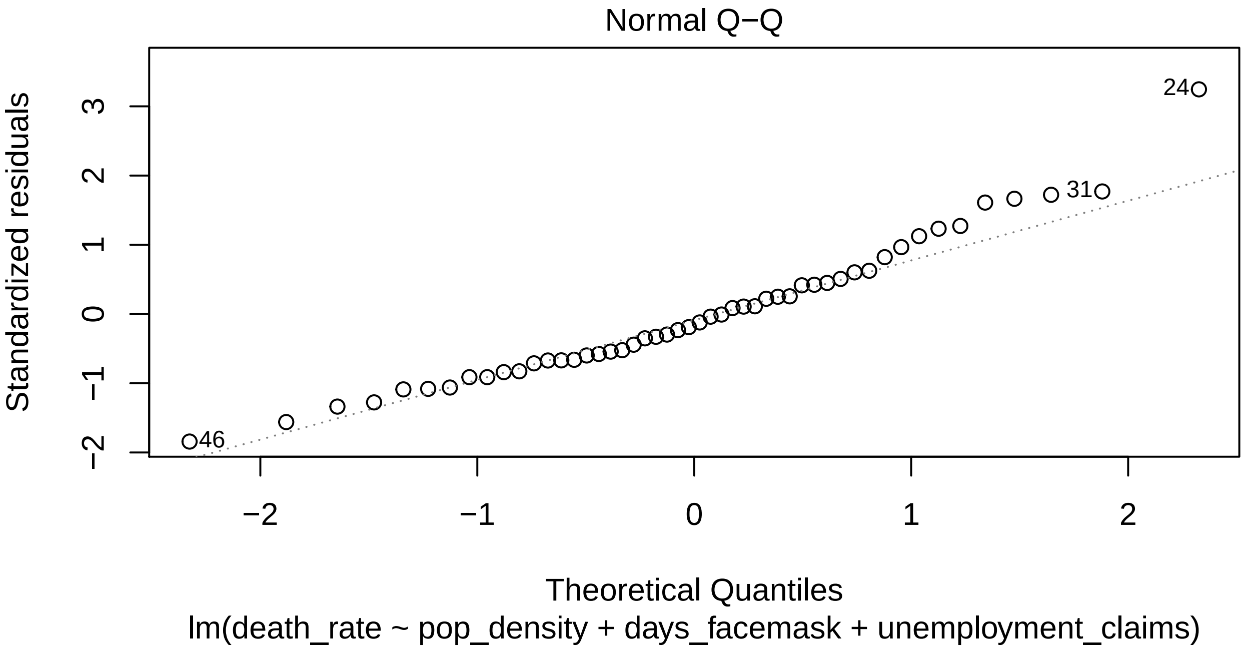


Figure 14: Normal Q-Q plot for model 2 only.

The normal probability plot of residuals shows that all points fall approximately along a standard line (with the exception of one outlier state), so we can assume normality distribution of error.

4. Regression Table

Table 1 provides the results for all three models. For model2, the population density and number of days of facemask mandate parameters were found to be statistically significant (with confidence levels of 99.9% and 95%, respectively). The practical significance of the population density parameter (0.06) is reasonably high - model2 indicates that an increase in 50 inhabitants per square mile is associated with an increase in the death rate by 3 deaths per 1,000 cases. Similarly, the practical significance of the face mask variable parameter is that as the days_facemask variable increases by 5 days, the death rate also increases by 1 death per 1,000 cases. This relationship is likely due to the fact that when states were hit very hard by the pandemic, they implemented face mask restrictions as a reaction and way to combat a virus that was already overwhelming the system, which is why there is a positive correlation between the two variables.

While model3 has the highest R-squared value, model2 has the highest adjusted R-squared value, indicating that the excessive amount of variables in model3 is likely causing overfitting and giving a false sense of efficacy. However for all three models, the adjusted R-squared value ranges from 0.431 to 0.477, indicating that there is still a large amount of variance in the death rate that is not explained by any of the models. All three models have an F-statistic that indicates that all of them fit the data better than an intercept-only model.

Table 1: Results

	Dependent variable:		
	death_rate		
	(1)	(2)	(3)
pop_density	0.072***	0.060***	0.060***
	(0.011)	(0.012)	(0.012)
days_facemask		0.190*	0.192*
		(0.080)	(0.079)
unemployment_claims		0.013	0.022
		(0.016)	(0.018)
high_risk			1.033
			(1.633)
elderly_percent			10.149
			(214.628)
poverty			−0.222
			(1.770)
Constant	23.444***	9.735	−32.868
	(2.788)	(7.925)	(32.366)
Observations	50	50	50
R^2	0.442	0.509	0.530
Adjusted R^2	0.431	0.477	0.465
Residual Std. Error	16.912 (df = 48)	16.209 (df = 46)	16.397 (df = 43)
F Statistic	38.049*** (df = 1; 48)	15.891*** (df = 3; 46)	8.089*** (df = 6; 43)
Note:			*p<0.05; **p<0.01; ***p<0.001

5. Discussion of Omitted Variables

There are a number of omitted variables that we expect are biasing the results of our model:

1) **The number of healthcare providers in a state**

- Context: If a state has more healthcare providers, they are better able to handle any surge in cases, and more patients are able to receive treatment and likely have a better chance of beating the virus. If a state has relatively few healthcare providers, the system is likely to be overwhelmed more quickly, and patients may not receive the care they need. With COVID-19, this is compounded by the fact that healthcare workers are on the front lines and, being exposed to the virus, may become sick themselves, further reducing the workforce capacity. We expect that more healthcare providers would likely result in a lower death rate.
- Related covariates: we believe this variable would have a positive correlation to population density, as more people tend to mean more providers. This variable likely would have relatively low correlation to the other variables included in our models.
- Bias direction: We expect this variable to be negatively correlated with death rate and positively correlated to the population density. Therefore the bias term is expected to be negative. Since population density has a positive coefficient for the regression model, the omitted variable has a bias towards 0.

2) **The number of ICU/hospital beds in a state**

- Context: If a state has more ICU beds and hospital beds, they will have more capacity to handle a surge in COVID-19 cases (while also being able to provide services for other types of illnesses). If a state becomes overwhelmed and runs low on hospital capacity, the probability of COVID-19 patients not receiving proper treatment increases (and many patients may be likely to try to stay home and "tough out" the virus instead of seeking treatment sooner).
- Related covariates: Similar to the number of healthcare providers in a state, we believe that this variable would be positively correlated with population density, and that more capacity would lower the death rate.
- Bias direction: We expect this variable to be negatively correlated with death rate and positively correlated to the population density. Therefore the bias term is expected to be negative. Since population density has a positive coefficient for the regression model, the omitted variable has a bias towards 0.

3) **PPE supply**

- Context: the supply of personal protective equipment (PPE) is critical for healthcare workers to be able to safely treat COVID-19 patients without becoming infected themselves. In some locations, healthcare workers have had less access to the amount of PPE supplies they need, and have had to resort to things like reusing face masks more than is recommended.
- Related covariates: Generally speaking, this variable would likely have a positive correlation to population density, as more dense areas will generally have more resources.
- Bias direction: We expect this variable to be negatively correlated with death rate and positively correlated to the population density. Therefore the bias term is expected to be negative. Since population density has a positive coefficient for the regression model, the omitted variable has a bias towards 0.

4) **Days since the start of the pandemic**

- Context: COVID-19 did not reach all states at the same time. While California, Washington and New York got to see the first wave of cases in the US, other states like Florida did not see the cases spike until much later. From modeling stand-point, As the states move along the curve, we would expect the death rate to start increasing because the health care system would start getting stressed over time.
- Related covariates: We would expect a strong positive correlation between this omitted variable and rate per 100,000 cases since the number of cases would increase over time.

- Bias direction: We expect this variable to be positively correlated with death rate and positively correlated to the rate per 100,000 cases. Therefore the bias term is expected to be positive. Since rate per 100,000 cases has a positive coefficient for the regression model, the omitted variable has a bias away from 0.

5) **Amount of population with disabilities**

- Context: Per CDC guidance, the subset of the population with disabilities is generally more susceptible to becoming infected by the virus, and will have a more difficult time getting the help they need.
- Related covariates: This variable would likely have a positive correlation to the high risk group identified by the CDC, although physical and mental disabilities are not indicated to be a direct illness that is considered high risk for COVID-19 (compared to illnesses such as respiratory issues).
- Bias direction: We expect this variable to be positively correlated with death rate and positively correlated to the high risk group. Therefore the bias term is expected to be positive. Since high risk group variable has a positive coefficient for the regression model, the omitted variable has a bias away from 0.

6) **Homelessness**

- Context: The homeless population is known to be very susceptible to COVID-19 (this is also listed in CDC guidance). This is due to their lack of resources or shelter, and therefore their inability to effectively quarantine or get proper medical treatment. Therefore we would expect that greater levels of homelessness to have a positive correlation with death rate.
- Related covariates: Just like the last variable, this variable would likely have a positive correlation to the high risk group identified by the CDC.
- Bias direction: We expect this variable to be positively correlated with death rate and positively correlated to the high risk group. Therefore the bias term is expected to be positive. Since high risk group variable has a positive coefficient for the regression model, the omitted variable has a bias away from 0.

7) **Nursing home population**

- Context: Since the beginning of the outbreak, nursing homes have been among the hardest hit by COVID-19. These locations contain groups of elderly people who often have high risk conditions that make them even more susceptible to COVID-19. Therefore we would expect areas with higher nursing home populations to have a higher death rate.
- Related covariates: We would expect a strong positive correlation between nursing home population and the percent of elderly population.
- Bias direction: We expect this variable to be positively correlated with death rate and positively correlated to the elderly population variable. Therefore the bias term is expected to be positive. Since elderly population variable has a negative coefficient for the regression model, the omitted variable has a bias towards 0.

6. Conclusion

Our assessment of the dataset used for this exercise to not be a random sample made a significant impact on how we could use linear regression to assess the data - we found we could not make any causal inference about the independent variable parameters that we calculated. Rather, we had a dataset that represented the entire population - all people within the USA, and all confirmed COVID-19 cases and deaths, among other data points.

Therefore we were only able to build associative models that found the line of best fit for the data. Further, as discussed in the introduction, the COVID-19 related data points were taken on July 6, 2020, about 4 months into the documented onset of the pandemic. At this point in the pandemic, some states such as New York, which was the global epicenter of the pandemic only several months before the point of data collection, are seeing reductions in cases and the virus becoming under control (at this time), while other states such as Florida, Arizona, and Texas are seeing massive surges in infections. States have not had an equal exposure to the pandemic. For that reason, the variables that describe the states that were the first to be hard hit by the pandemic (i.e., population density, number of days of face mask mandates) are the dominant variables in our models, while the variables that public health experts say will correlate to higher risk of severe symptoms or death from COVID-19, based on epidemiological domain knowledge, such as high risk populations and elderly populations, remain as weaker predictors of the death rate. We expect that this will not be the case by the end of the pandemic, which will be a very interesting research study to perform.

Additionally, as we developed our models we found that it would likely be more appropriate to perform this study at a level more granular than the state level. If, for example, one were to apply a multivariate regression at the county level, there would not only be more data to choose from (providing for more options to develop a random sample, given that there are more than 3,100 counties in the United States), but each row of data would better represent an area. For example, the state of New York consists of both an extremely dense city (New York City) as well as a great deal of very spread out land in much of the northern and central parts of the state. Fitting the data at the county level would therefore allow you to take those differences into consideration in the model.