Population density and death rate prediction of Covid-19

Arsalan - EJ - Angel

7/19/2020

Introduction:

Over the past several months, the world has experienced much turmoil due to the novel coronavirus, Covid-19. Many researchers are trying to understand the primary reasons for the spread of this virus. When the virus first appeared in the United States, many large cities, such as Seattle and New York, experienced major outbreaks as the virus spread very quickly. Later, a similar phenomenon occurred in Miami, New Orleans, and Chicago. States with a very high population density, such as New Jersey, Connecticut, and Massachusetts, also had spikes in the number of cases. It is very likely that high population density may be a cause for the high number of coronavirus cases due to the close proximity of people in these locations.

Medical professionals have also stated that people may have contracted the coronavirus while still being asymptomatic, i.e. they do not show symptoms. It is likely that people who are in a good state of health have an immune system that is able to fight off the virus while experiencing mild or no symptoms. However, those people who are at high risk for infection or have severe comorbidities may experience very severe symptoms of the virus, such as respiratory problems, high fever, muscle or body aches, or persistent chest pain. When the virus first appeared in the US, the shortage of testing capacity led to only those people who showed symptoms being tested for the virus. It is possible that the presence of comorbidities and those at risk for illness due to Covid-19 may be a cause for a higher number of cases.

The purpose of this project is to investigate whether differences in population density, number of cases of Covid-19, and population of select age groups among states is causing a higher number of deaths due Covid-19 in each state, respectively. We plan to investigate this by using state-specific data and building several linear regression models. We have the total number of Covid-19 deaths by state, the population, population density and population of different age groups for each of the 50 states and the District of Columbia as of 2018. To investigate whether a higher population density, number of positive cases, and age groups are causing a higher number of Covid-19 deaths, we will build a regression model that predicts the the death rate based on these variables.

Other variables, such as state-ordered facemask mandates, state-of-emergency declarations, percentage of people over 65, and percentage of the population living in poverty, will also be considered as covariates that could further explain the causes for the number of cases by state.

Question:

Does population density and number of cases help predict the number of deaths of Covid-19 by state?

** Section 2.1 ** For the population density, we will use each state's population density as of 2018. For the number of cases, we'll use the percentage of positive cases in the state's population to measure this variable. We transform it this way because we want to normalize the counts by state. For the number of deaths (outcome), we'll also use the percentage of deaths in the state's population to measure this variable.

Section 2.2 We will include the age group percentages as additional covariates to help measure the causal effect. We do not expect multicollinearity because these new variables are not directly related to the main effect variables(percent cases and population density). However, there may be multicollinearity between different age groups.

Section 2.3 For the covariates, we will calculate the ratio of the number of positive cases to the population of that state. For the response, we need to check that the distribution is normal. We'll check the QQ-plot for the normality assumption. If this assumption is violated, then we may need to transform the response, possibly with a logarithm.

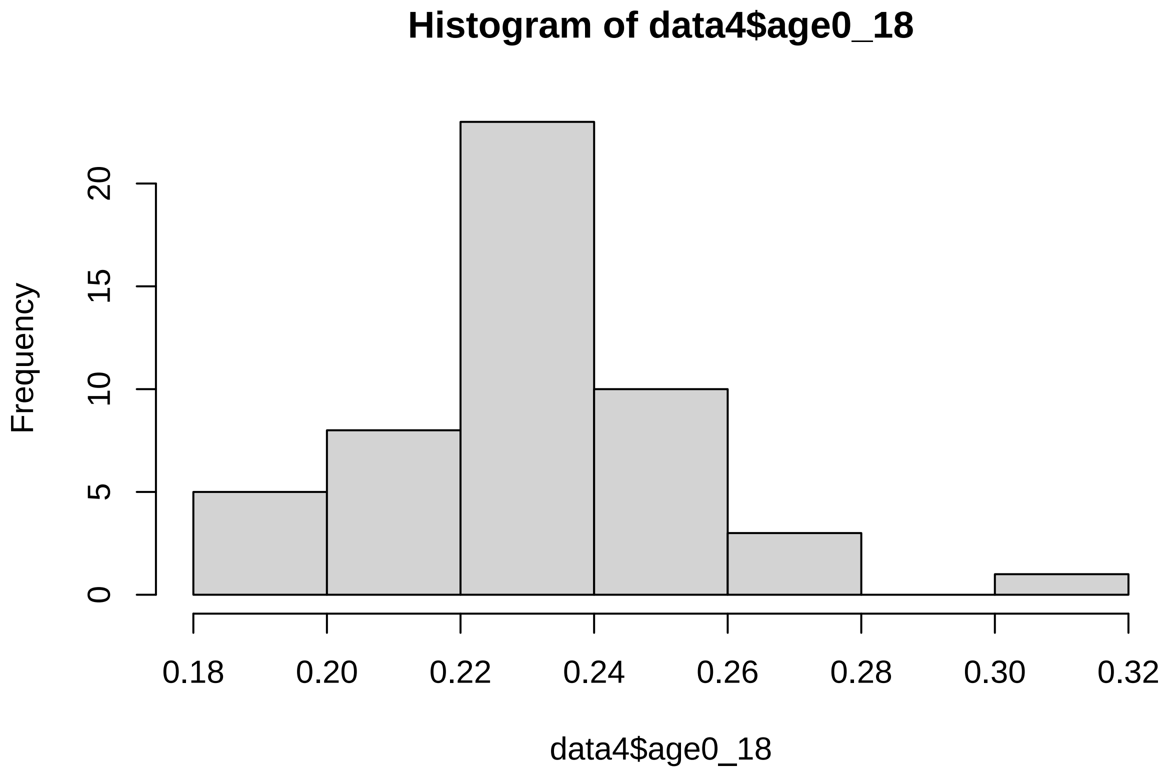
Section 2.4 We will check for duplicate data entries, such as multiple rows of the same state in the data. If we encounter multiple entries of the same state, we will combine these rows into a single row. For missing data, we will consider using the mean or median value as an imputation, or as a more sophisticated approach we could build a simple linear regression model with complete data where the response is the covariate with missing values. Then we will use this SLR to predict the missing entry.

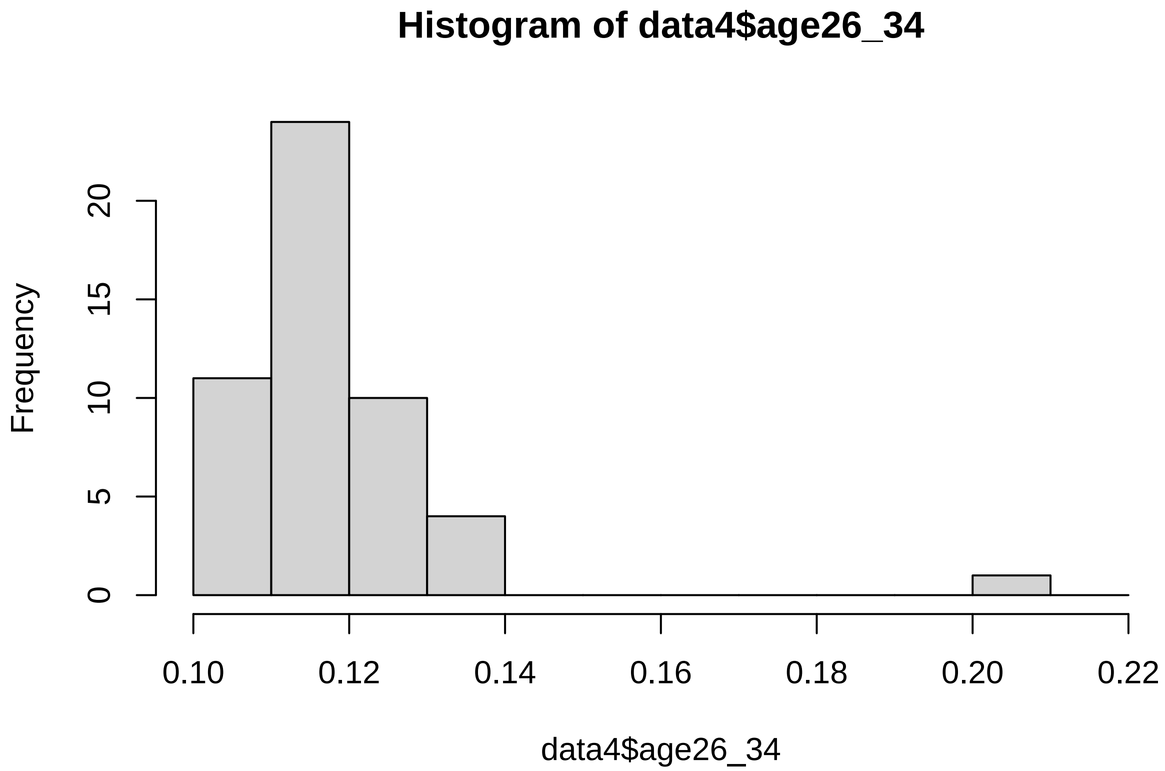
Our first observation is that Arizona occurs twice in the data set. We also notice all entries of these two rows are identical, except for the total number of cases and total number of deaths: 83,376 cases for the first Arizona row and 14,713 cases for the second Arizona row. If we add these together, we get 98,089 total cases. This value divided by the population 7,171,646, then multiplied by 100,000 gives us the 1,367.7. We perform a similar calculation for deaths. We add the total number of cases and deaths together, while keeping all other variables unchanged, to form one row for Arizona.

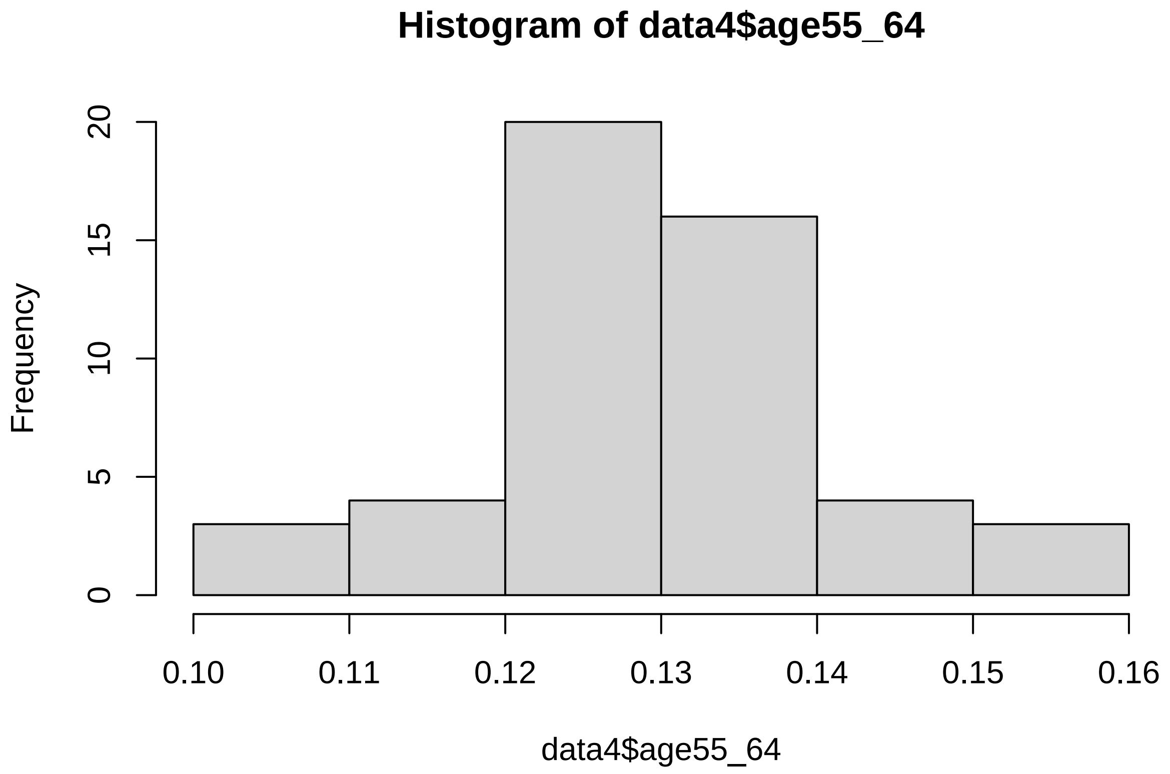
Next, we check if any of the variables have missing data. Based on our observations, totalTestResults, Children.0.18, Adults.26.34, Adults.55.64 each have 1 missing value.

Since we are not including total test results in the model, we can ignore the missingness of this variable. But if decide to include later, then we will impute the value based on an SLR on the population.

For the age groups, we check the histograms to see if they follow a normal distribution. We need to decide how we should impute the missing values.


Histogram of data4$age0_18
Frequency
0
5
10
15
20
0.18
0.20
0.22
0.24
0.26
0.28
0.30
0.32
data4$age0_18

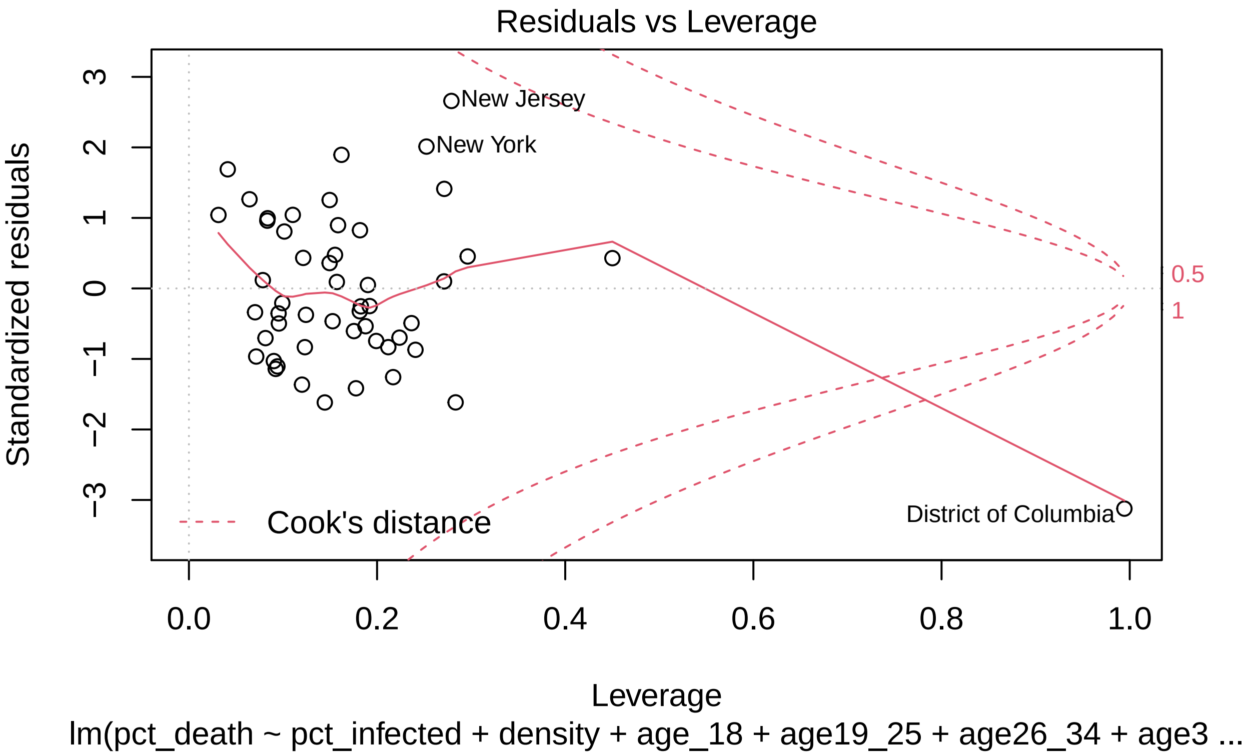

Histogram of data4$age26_34
Frequency
0
5
10
15
20
0.10
0.12
0.14
0.16
0.18
0.20
0.22
data4$age26_34



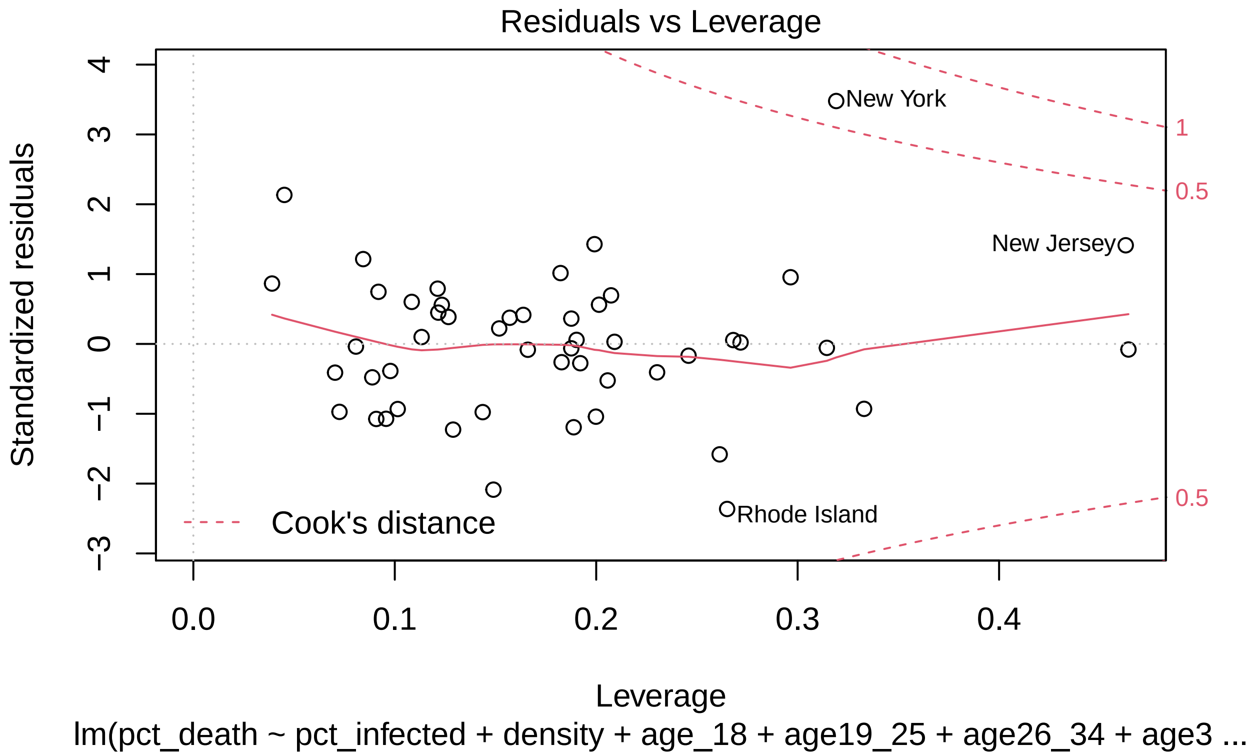
It appears there may be some outliers in the age percentage variables. Without these outliers, however, the data appears to be mostly normally distributed. We will impute the median instead of the mean for each of these missing values.

In the first step, we decided to include the percentage of Covid-19 infected to poulation, density, age 18, age 19-25, age 26-34, age 35-54, age55-64 and age65 in our model.

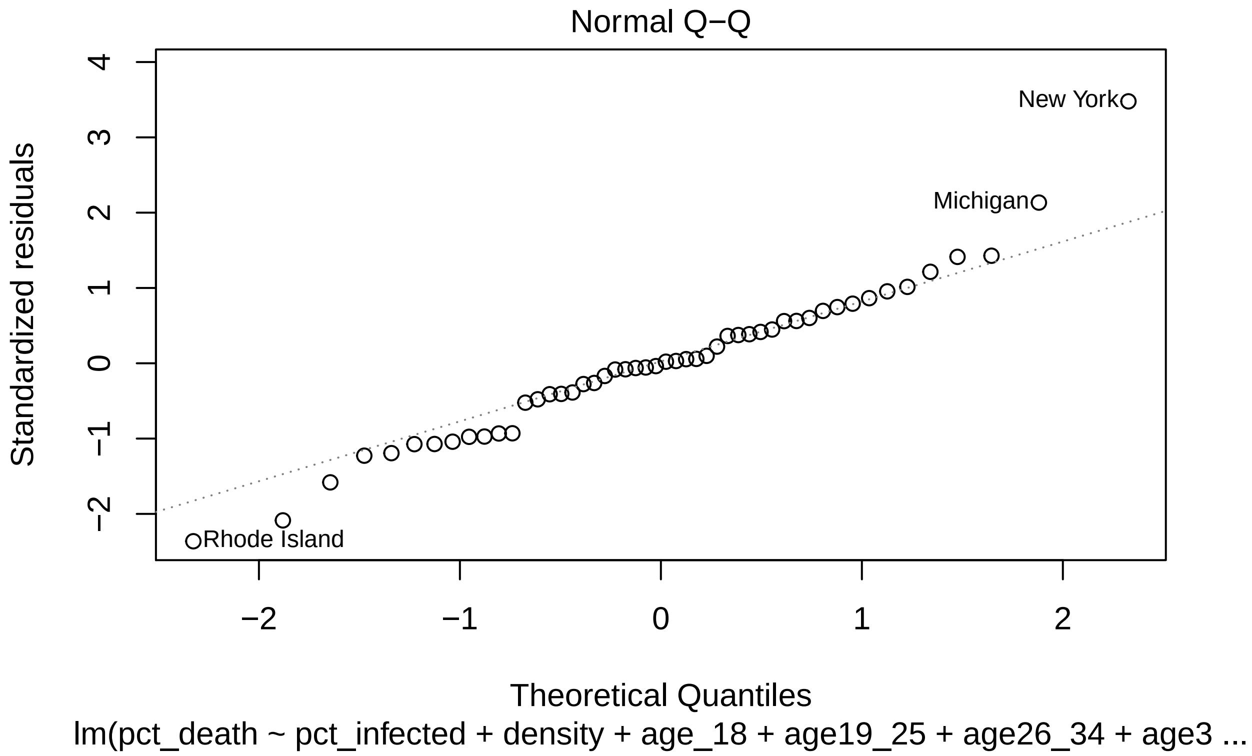
```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

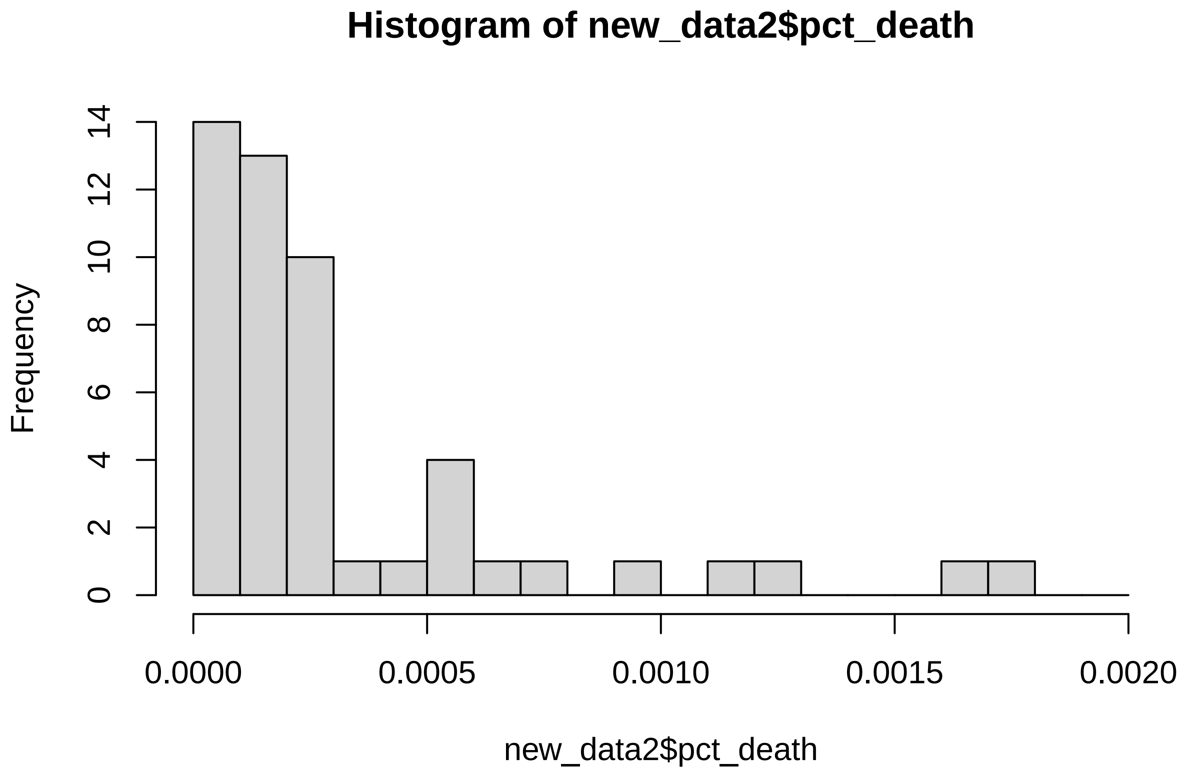


We check for influential observations using the Residuals vs Leverage plot. We find that District of Columbia is outside the acceptable region, and is therefore considered an influential observation. We recognize that DC is very different in terms of population density compared to the other states and does not follow the same pattern as the other states. Therefore, we decide to remove this observation.

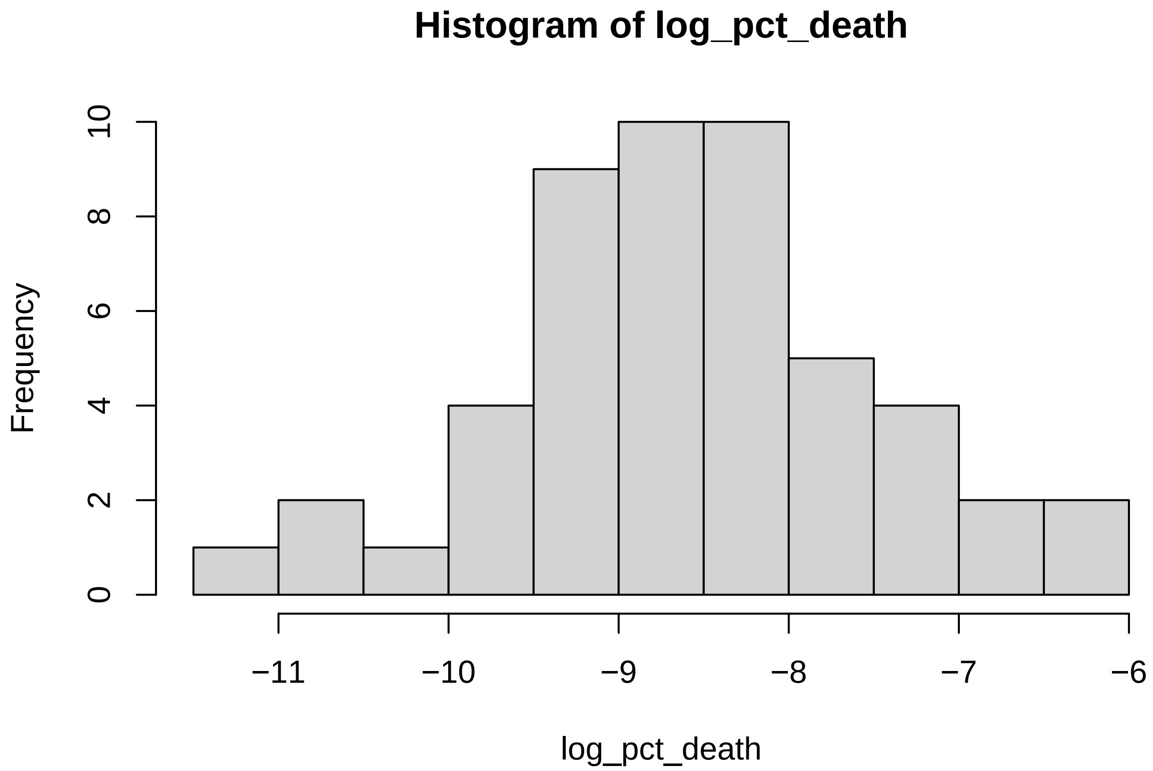


After removing DC, all states (except for New York), are within the acceptible region for leverage values. The leverage value for NY is much less than DC, so instead of removing New York, we will check if there are any violations to the other assumptions.


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
New York
Michigan
Rhode Island
−2
−1
0
1
2
Theoretical Quantiles
lm(pct_death ~ pct_infected + density + age_18 + age19_25 + age26_34 + age3 ...



The Normal QQ plot shows that the error variance is not normally distributed. Therefore, it may make the most sense to transform the response variable (death_rate) using a logarithm. The histogram of the death rate also shows that it is not normally distributed.



The histogram of the logarithm of the death rate now looks normally distributed. We can safely use this variable as the response in the linear regression.

After observing different models with different variables, we found the most proper model.

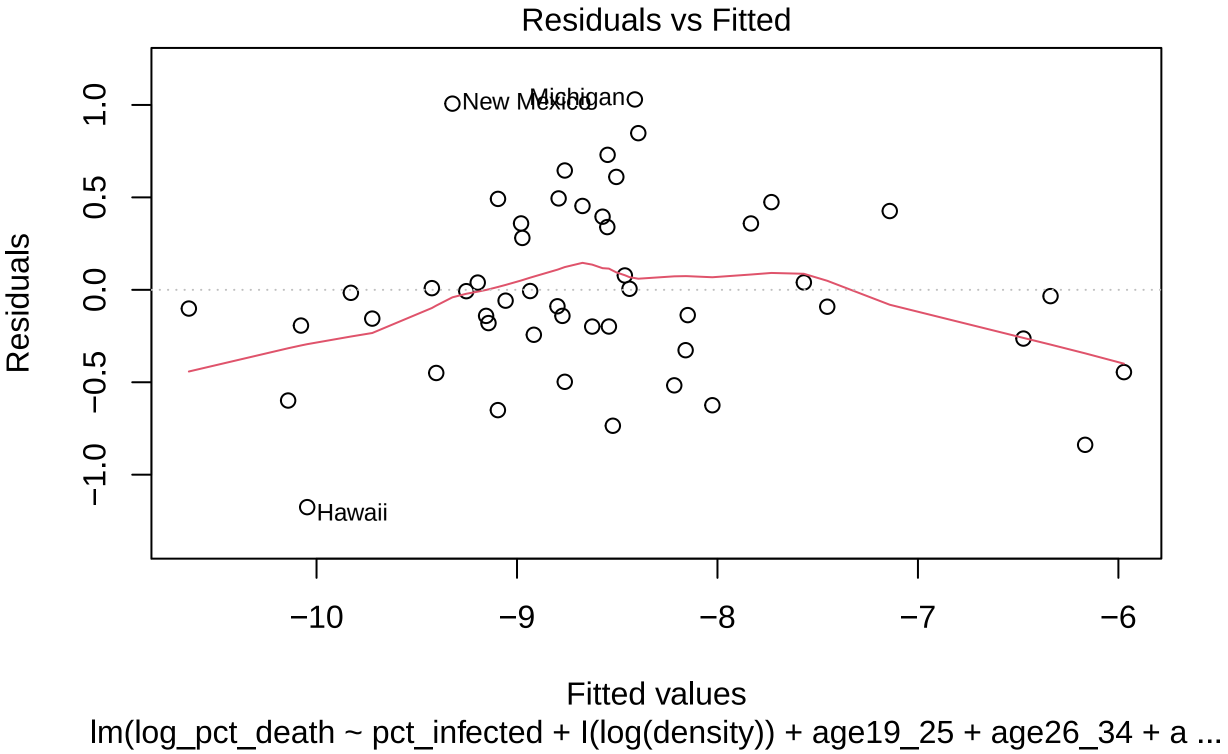
Linear model assumption:

We assume we have a linear population model, so no checks here.

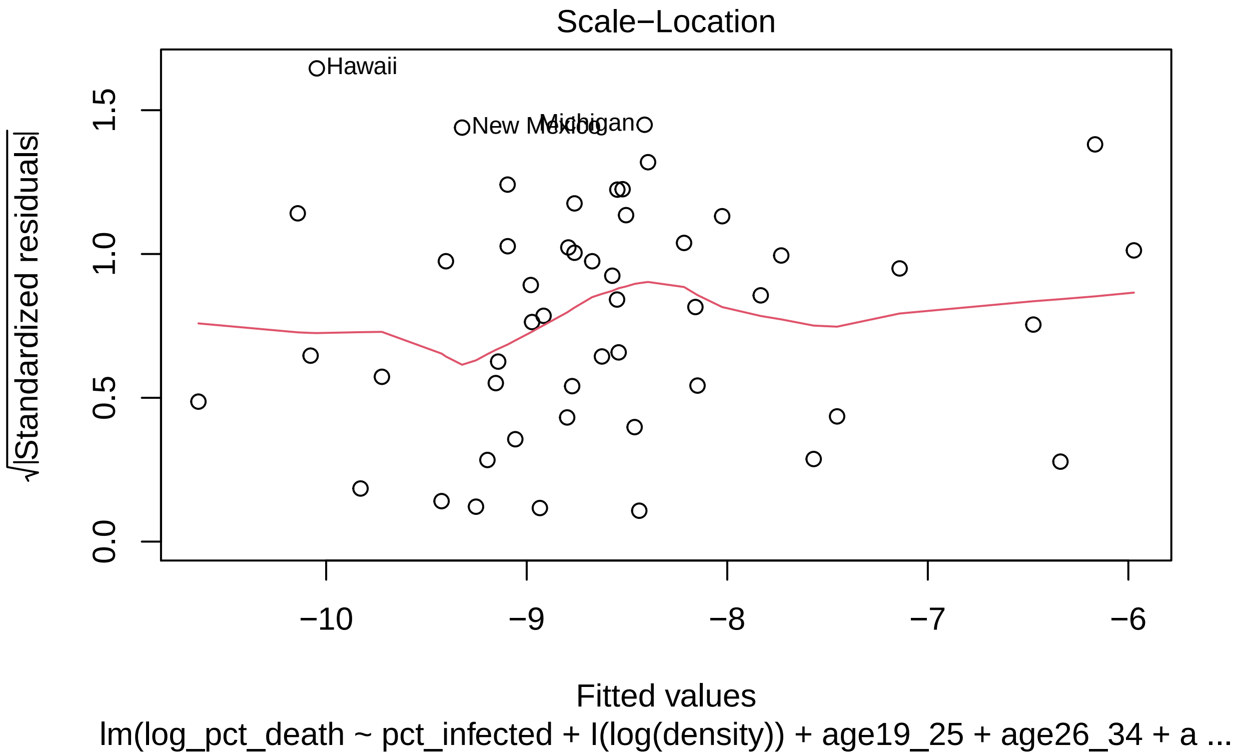
For the random sampling assumption, our study uses a population sample, so the idea of random sampling does not really apply to this study.

```
##    pct_infected I(log(density))        age19_25        age26_34        age55_64 
##        1.793567        1.798291        1.684986        1.364111        1.727971 
```

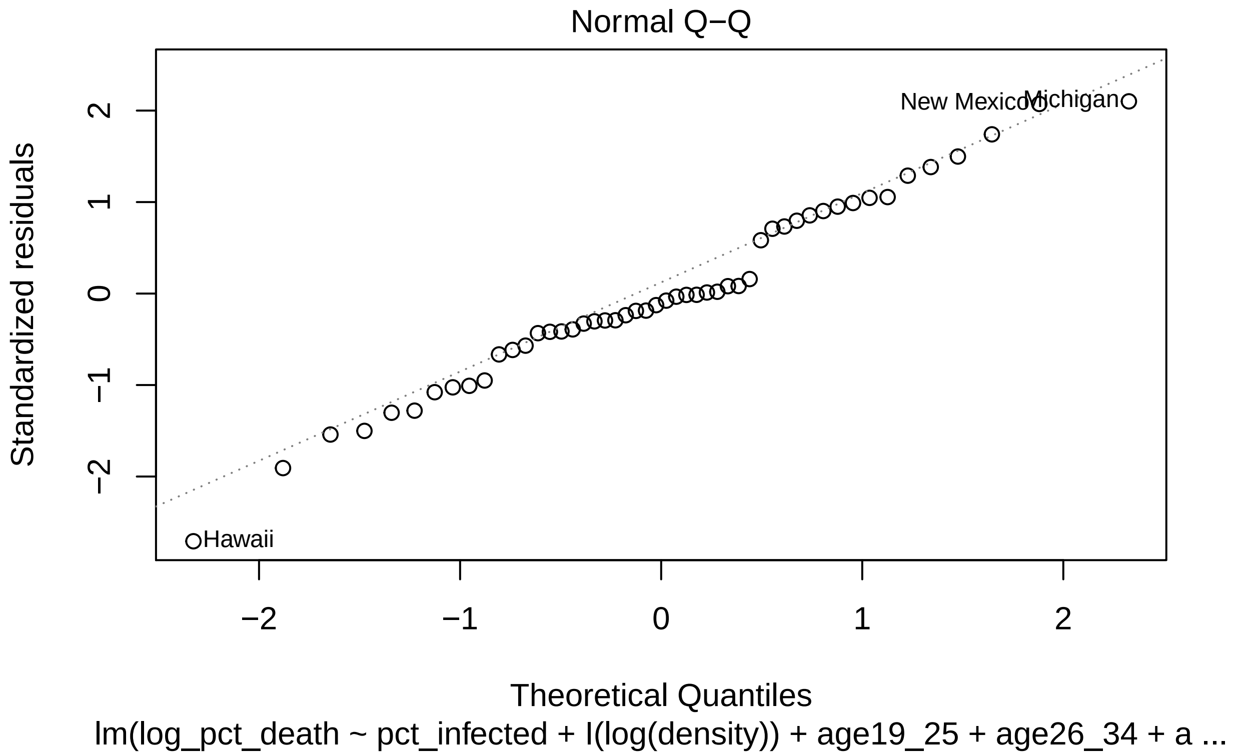
All variables in the final model have a VIF < 4, which indicates that there is not multicollinearity in the model.



The Residuals vs. Fitted plot does not show any major non-linearity in the model, so this satisifies the zero-conditional mean assumption.



Because the shape of the line of the Scale-Location plot is mostly flat, we assume that the homoskedasticity assumption is satisified.



For the normality assumption, all points lie on the straight diagonal line with very little variation, so the normality of errors assumption is satisfied.

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      -18.345254   3.380071 -5.4275 2.324e-06 ***
## pct_infected     148.871110  22.587272  6.5909 4.559e-08 ***
## I(log(density))    0.273709   0.068846  3.9756 0.0002577 ***
## age19_25          21.706120  17.591078  1.2339 0.2237803
## age26_34          14.775484  10.725129  1.3777 0.1752783
## age55_64          27.504563  11.280512  2.4382 0.0188658 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The percent infected and population density are very statistically significant, each with a p-value less than 0.01. The age group 55-64 is also statistically significant, with a p-value less than 0.05. For practical significance, we calculate Cohen's f2 statistic, which is the effect size for the model.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Jul 23, 2020 - 00:34:13

Cohen's f2 statistic is 4.3, which corresponds to a very large effect size. This indicates that our model is practically significant.

The model regression table shows the model specifications.

Table 1: Model Regression Table

	Dependent variable:
	log_pct_death
pct_infected	148.871^{***}
	(21.129)
I(log(density))	0.274^{***}
	(0.072)
age19_25	21.706^{*}
	(12.542)
age26_34	14.775
	(9.749)
age55_64	27.505^{***}
	(8.201)
Constant	-18.345^{***}
	(2.383)
Observations	50
R^2	0.811
Adjusted R^2	0.790
Residual Std. Error	0.501 (df = 44)
F Statistic	37.854^{***} (df = 5; 44)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

```
## [1] 4.301606
```

Omitted Variables

Many omitted variables may exist, such as Humidity, Pollution, Temperature, Gender, and Ethnicity

Humidity

We consider the omitted variable bias of excluding the variable Humidity from the model. This would be a quantitative variable, and we would use the average humidity from March 2020 to July 2020 for each state. Humidity ~ pct_infected + pop_density + age19_25 + age26_34 + age55_64. Because humidity is primarily affected by weather, we can safely say that no variable in our model will affect the average humidity of that state. Therefore, we have almost no omitted variable bias for humidity. We cannot determine the direction of omitted variable bias because we do not believe there is any practical relationship between humidity and the other variables in our model.

Pollution

We consider the omitted variable bias of excluding the variable Pollution from the model. This would be a quantitative variable, and we could use the average weight of the pollution. Polution ~ pct_infected + pop_density + age19_25 + age26_34 + age55_64. The only variable that may affect pollution is population density. As the population density increases, we might expect pollution to also increase. By omitting pollution from the model, we would expect to see an ommited variable bias on population density, and the coefficient of population density would decrease towards zero. This coefficient would become less significant.

However, we would not expect the significance of population density to change very much because the variable is highly significant. No other variable in the model has any practical relationship to pollution so we would not be able to assess the omitted variable bias for other variables.

Risk factors

For risk factors, we could measure this variable as a quantitat variable that represents the percentage of the population that has at least one adverse heath condition. risk_factors ~ pct_infected + pop_density + age19_25 + age26_34 + age55_64. The only variables that have any practical relevance to risk factors are population density and age55_64. The higher percentage of older age would imply that there are more health conditions among people in that state, which would increase the risk factors. For population density, people in a urban area may not be as healthy as those in a rural area due to food choices/diet/living conditions. This would also increase risk factors. Omitted variable bias for pop_density and age55_64 would cause both coefficients to go towards 0.

Temperature

We consider the omitted variable bias of excluding the variable Temperature from the model. This would be a quantitative variable, and we would use the average temperature from March 2020 to July 2020 for each state. Temperature ~ pct_infected + pop_density + age19_25 + age26_34 + age55_64. Because temperature is primarily affected by weather related phemonena, we can safely say that no variable in our model will affect the average temperature of that state. Therefore we have almost no omitted variable bias for temperature. We cannot determine the direction of omitted variable bias because we do not believe there is any practical relationship between temperature and the other variables in our model.

Conclusion

Based on the table of coefficients, we see that the number of deaths is predicted well by the percent infected and population density, which were the two most significant variables in the model. Although the R-squared is 81%, which does mean that our model explains quite a large fraction of the number of deaths, there are still omitted variables, such as temperature, humidity, risk factors, and pollution, that will help explain the remainder that are not included in this data set. Our model also is of high practical significance due to the high R-squared value, which means that it's of high accuracy in predicting the number of deaths due to Covid-19.