Virus tracker
INFO248
By: Jingyue Zhang,Peilin Guo

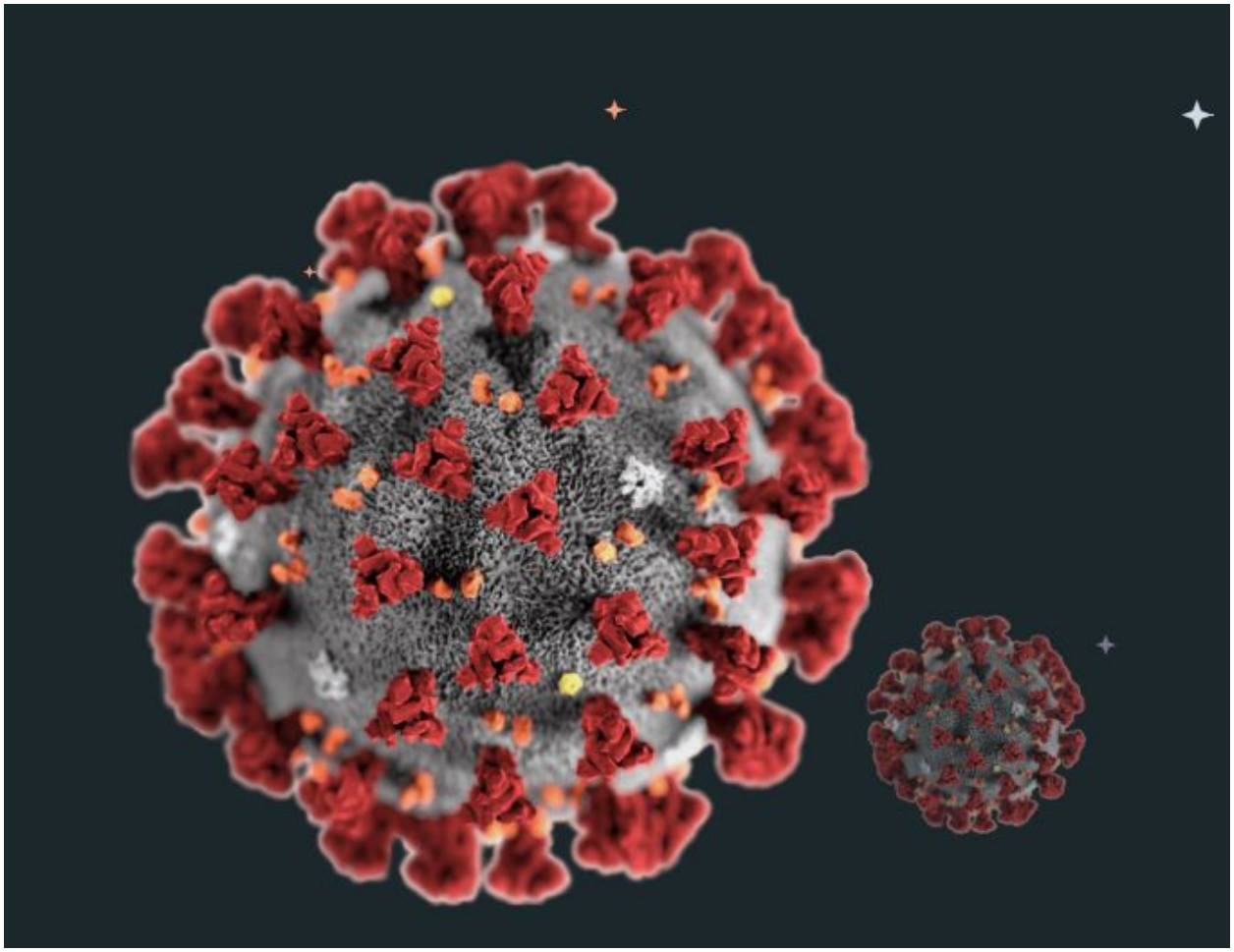


Table of Contents

Introduction

Project Description and research question:

An exploratory study in factors affecting the number of deaths from COIVD19

Our team project for this course is to construct a predictive model in analyzing Coronavirus Data in the United State. This dataset is released by New York Time with cumulative counts of coronavirus cases and deaths in the United State, at the state level [1]. The data was collected by NYT as new cases and deaths were being reported. However, it is to our understanding that due to different extents of shortages of testing kits, the data is likely to have a considerable amount of variance in depicting the situation in the country. Nonetheless, this data still serves as a reasonable source for model building as due to its cumulative effort.

In addition, the data is also useful when analyzed with the state of America public health system in different states, as described by three different measures - public health accessibility, public health quality and public health ranking - as reported by the U.S. News [2], and the population density of each state as of 2019 as reported by Statista [3].

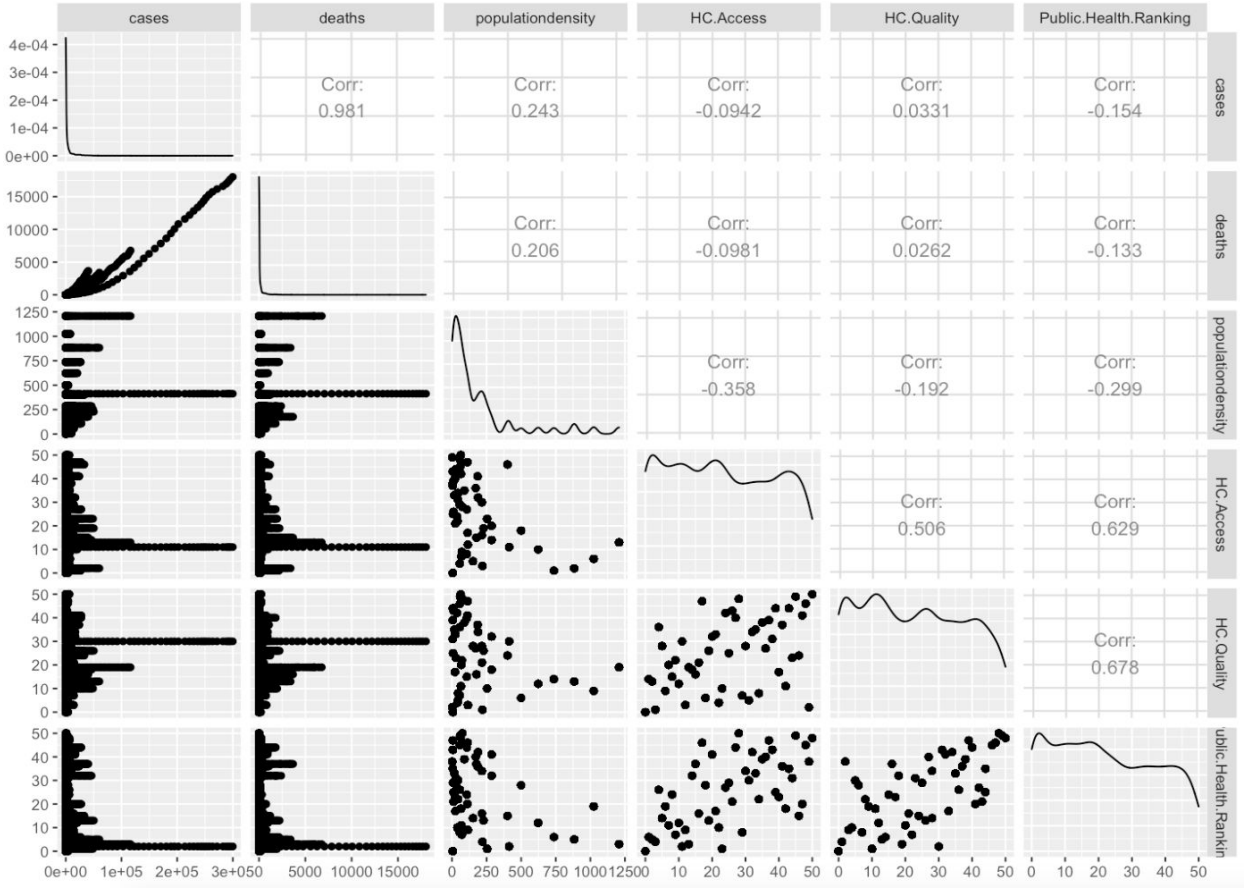
When combined with these two scopes, a fuller picture could be assessed in analyzing the damages of, effect from and effort in response to the COVID19 pandemic.

Data Preprocessing

```
import pandas as pd

countries = pd.read_csv("us-states.csv")
density = pd.read_excel("density.xlsx")
health = pd.read_excel("countryHealth.xls")
df = pd.read_csv("Countries_Merged.csv")

for i in range(len(countries['state'])):
    for j in range(len(density)):
        if countries['state'][i] == density['State'][j]:
            countries['Pop. Density'][i]= density['Density'][j]
            print(i)

for i in range(len(countries['state'])):
    for j in range(len(health)):
        if countries['state'][i].lower() == health['SATE'][j]:
            countries['HC Access'][i]  =health['HEALTH CARE ACCESS'][j]
            countries['HC Quality'][i] =health['HEALTH CARE QUALITY'][j]
            countries['Public Health Ranking'][i] =health['PUBLIC HEALTH'][j]
            print(i)

countries_clean = countries[countries['state'] != 'District of Columbia']


countries.to_csv('Countries_Merged.csv')
```

A combined master data frame is constructed using Python, where the main dataset of coronavirus cases in the United States is combined with population density and public health data, such as healthcare access, healthcare quality and public health ranking, by states. For consistency in state-level analysis, the District of Columbia has been removed from our dataset.

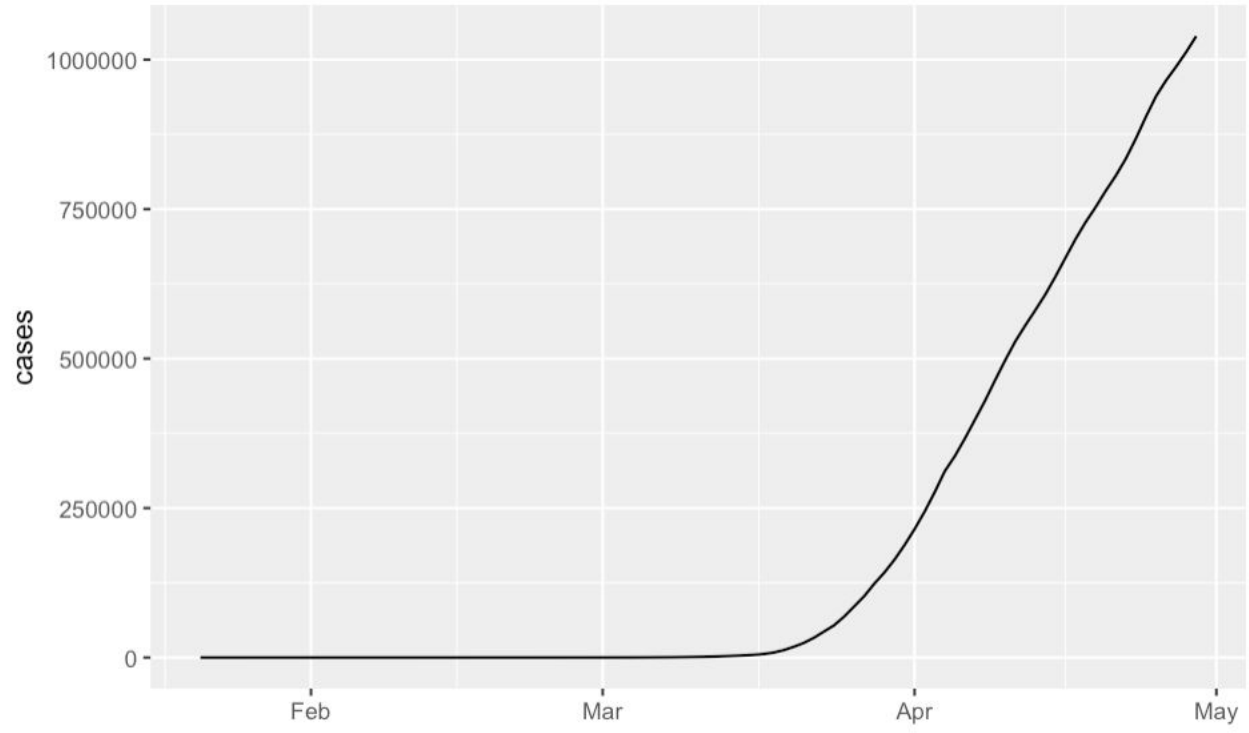
Correlation table

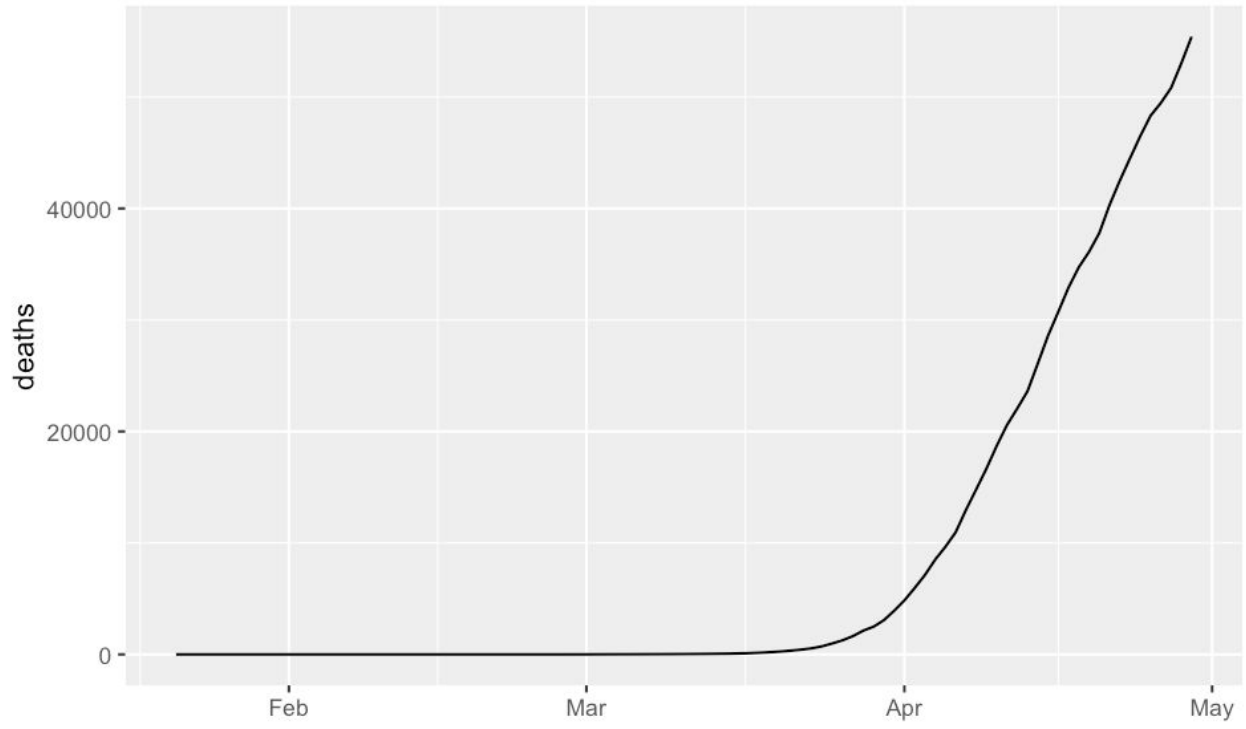


In assessing potential multicollinearity issues among the variables, a correlation matrix is constructed using R to assess the correlation coefficient between any two variables. The first shows both the correlation coefficients among the variables and the distribution of them. It is clear that the number of cases and deaths are strongly

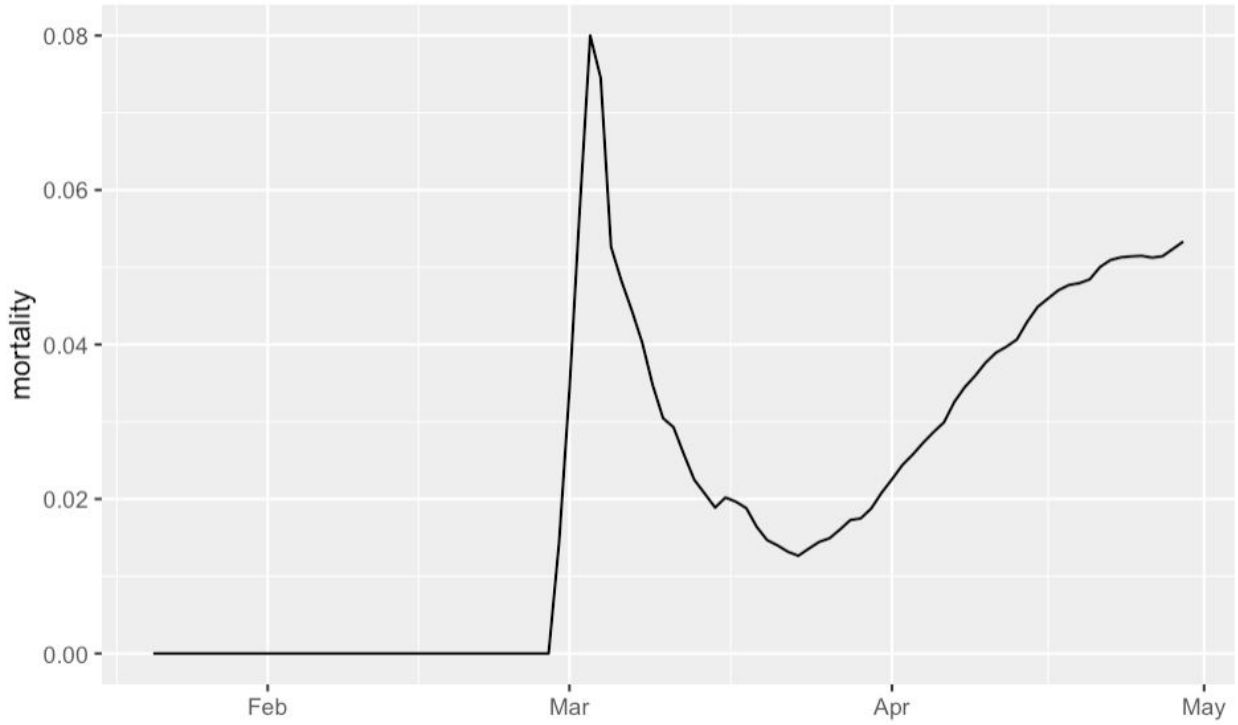
positively linearly correlated, as evidenced from the correlation coefficient and the straight line plot between them. Healthcare Access (HC Access) and Healthcare Quality (HC Quality) may seem to have a weak correlation from the scatter point but it is proven weak from both the correlation coefficient (0.506) and the line graph between. For a better visualization in correlation coefficient, a colored matrix is constructed as follows:

Time series


1000000
750000
500000
250000
0
cases
Feb
Mar
Apr
May


40000
20000
0
deaths
Feb
Mar
Apr
May

Both cases and deaths show a similar increasing trend with a rather flat straight line before mid-March followed by a straight line increase. However, it is noted that the increase in cases is much faster than deaths as expressed in the y-axis.



To assess the interaction between cases and deaths, a new derived variable mortality is constructed using deaths/cases. As shown above mortality has a different trend as compared to the other two. Before March, mortality stays relatively constant. It spikes to a peak in early March, followed by a parabolic curve and then trending to be leveled off into May. Possible explanations for this phenomenon could be due to the fact that early efforts in identifying infected individuals were much limited to both how sick an individual was and if the individual had been exposed to the virus. Starting March, efforts in increasing testing and identifying infected individuals took place resulting in an unproportionally increase in new confirmed cases than confirmed deaths. [4]With

more effective testing and identifying new cases, mortality started to drop significantly and possibly below the true mortality rate of the virus. However, as efforts in testing continue, mortality starts to approach its true value, especially as many states have started to test recently deceased via autopsies.

Multiple Linear Regression

```
fit1<-lm(deaths~day+cases+populationdensity+HC.Access+HC.Quality+Public.Health.Ranking,mergedat)
summary(fit1)
```

```

Call:
lm(formula = deaths ~ day + cases + populationdensity + HC.Access +
    HC.Quality + Public.Health.Ranking, data = mergedat)

Residuals:
     Min       1Q   Median       3Q      Max
-2535.06   -36.19    16.53    61.51  1836.34

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          72.698575  12.220681   5.949 2.99e-09 ***
day                  -0.047377   0.004648 -10.193  < 2e-16 ***
cases                 0.054896   0.000199 275.902  < 2e-16 ***
populationdensity    -0.203011   0.016954 -11.974  < 2e-16 ***
HC.Access            -3.212831   0.353311  -9.093  < 2e-16 ***
HC.Quality           -2.935454   0.375387  -7.820 7.13e-15 ***
Public.Health.Ranking  5.166439   0.422581  12.226  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 228.6 on 3197 degrees of freedom
Multiple R-squared:  0.9654,    Adjusted R-squared:  0.9654
F-statistic: 1.489e+04 on 6 and 3197 DF,  p-value: < 2.2e-16
```

The initial linear regression model obtained is:

Deaths = 72.699 - 0.047*Day + 0.055*Cases - 0.203*PopulationDensity
- 3.213*HC.Access - 2.935*HC.Quality + 5.166*Public.Health.Ranking

After running a linear regression model using death as the outcome, day, cases, population density, healthcare access, healthcare quality and public health ranking as predictors, it is established that all the predictors are statistically significant using Wald test. Among the variables:

Variables	Description	Regression effect
day	Time variable with unit of day	A negative coefficient suggests that as time goes by, ceteris paribus, the number of cases will decrease. However, as the magnitude of the coefficient for this variable is extremely small (less than 0.05), the effect of *day* on *deaths* is much limited for discussion based on the data.
cases	Number of confirmed cases	A positive coefficient suggests as more people are confirmed with infections, more people will die. However, the small magnitude (0.055) may also suggest that the true mortality rate of the virus is low.
population density	Number of residents per mile squared	A negative coefficient shows a higher population density in a state is generally associated with (slightly) lower number of deaths. This is surprising at a first glance, but could be reasonable as a more populous state may be associated with greater medical resources, and knowing the high population density may prompt officials to act more rapidly in response to the pandemic.
healthcare access	Proportion of adults and children with medical and dental care	A negative coefficient shows a higher proportion of healthcare access is generally associated with lower number of deaths. This could be explained as more people have access to medical aid when in need.
healthcare quality	Proportion of preventable hospital admission, medicare plan ratings and the quality of nursing homes and hospitals.	A negative coefficient shows a higher healthcare quality in the state is generally associated with a lower number of deaths. This could be so as medical resources are used more efficiently in these states.
public health ranking	Rates of obesity, smoking, suicide, mental health and mortality for adults and infants in the state	A positive coefficient shows a higher value of the ranking (the worse the ranking) is generally associated with a higher number of deaths. This could be explained as a healthy population as a whole has a strong community immunity against the virus as people practice healthy lifestyles and hygienic routines.

Predictor selection (step wise)

```r
fit2 <- lm(deaths ~1, data = mergedat)
fit2step <- step(fit2,scope =
deaths~day+cases+populationdensity+HC.Access+HC.Quality+Public.Health.Ranking, direction = "both")
```

```
Start:  AIC=45585.3
deaths ~ 1

                       Df  Sum of Sq        RSS   AIC
+ cases                 1 4649385548  185649345 35143
+ day                   1  280394702 4554640191 45396
+ populationdensity     1  204741646 4630293247 45449
+ Public.Health.Ranking 1   85066807 4749968086 45530
+ HC.Access             1   46486545 4788548348 45556
+ HC.Quality            1    3326248 4831708645 45585
<none>                               4835034893 45585

Step:  AIC=35142.96
deaths ~ cases

                       Df  Sum of Sq        RSS   AIC
+ populationdensity     1    5450928  180198417 35049
+ day                   1    2379196  183270149 35104
+ Public.Health.Ranking 1    1758911  183890434 35114
+ HC.Quality            1     187415  185461930 35142
+ HC.Access             1     159878  185489467 35142
<none>                               185649345 35143
- cases                 1 4649385548 4835034893 45585
```

```
Step:  AIC=35049.47
deaths ~ cases + populationdensity

                        Df  Sum of Sq        RSS   AIC
+ day                    1    3524380  176674037 34988
+ HC.Access              1    1668596  178529821 35022
+ HC.Quality             1     874146  179324271 35036
+ Public.Health.Ranking  1     513408  179685009 35042
<none>                                 180198417 35049
- populationdensity      1    5450928  185649345 35143
- cases                  1 4450094830 4630293247 45449

Step:  AIC=34988.19
deaths ~ cases + populationdensity + day

                        Df  Sum of Sq        RSS   AIC
+ HC.Access              1    1704117  174969920 34959
+ Public.Health.Ranking  1     940097  175733940 34973
+ HC.Quality             1     767532  175906505 34976
<none>                                 176674037 34988
- day                    1    3524380  180198417 35049
- populationdensity      1    6596113  183270149 35104
- cases                  1 4141475194 4318149230 45227

Step:  AIC=34959.13
deaths ~ cases + populationdensity + day + HC.Access
```

.....

```
Step:  AIC=34815.25
deaths ~ cases + populationdensity + day + HC.Access + Public.Health.Ranking +
    HC.Quality

                      Df  Sum of Sq        RSS   AIC
<none>                                 167077732 34815
- HC.Quality            1    3195717  170273449 34874
- HC.Access             1    4321531  171399263 34895
- day                   1    5429237  172506970 34916
- populationdensity     1    7493173  174570906 34954
- Public.Health.Ranking 1    7811570  174889302 34960
- cases                 1 3978199378 4145277110 45102
```

After adopting step-wise predictor selection by using AIC, all of the variables inherited from the initial model have been retained as important predictors in the linear regression for deaths. The regression model thus takes the form of

Deaths = 72.699 - 0.047*Day + 0.055*Cases - 0.203*PopulationDensity
- 3.213*HC.Access - 2.935*HC.Quality + 5.166*Public.Health.Ranking

SIR model :

An SIR model is an epidemiological model that computes the number of people infected in a closed population over time. This model involves several differential equations relating the number of susceptible people, number of people infected, and number of people who have recovered. Until recently, the SIR model is still one of the most commonly used models.

The equations of the model are:

$$\frac{dS}{dt} = -\beta \frac{SI}{N} \quad (1)$$

$$\frac{dI}{dt} = \beta \frac{SI}{N} - \gamma I \quad (2)$$

$$\frac{dR}{dt} = \gamma I \quad (3)$$

Susceptible refers to people who are not infected but easily infected after contact with infected people. S(t) is the number of susceptible persons at time t. I is an infective group and I(t) is the number of infective persons at time t. R(t) is the number of people recovered in time t. β is the contact rate. The infectious individuals can spread the disease and each contact β new person per day. is the average infectious period.

From (1)(2)(3), we obtain the total population N:

$$N = S + I + R$$

For this model, we estimate β and with the function ode from the deSolve package in R. Then we optimize the parameters with function optim by minimizing the sum of the squared differences between the number of infections I(t) and the corresponding predicted number of infections at time t.

The equation is:

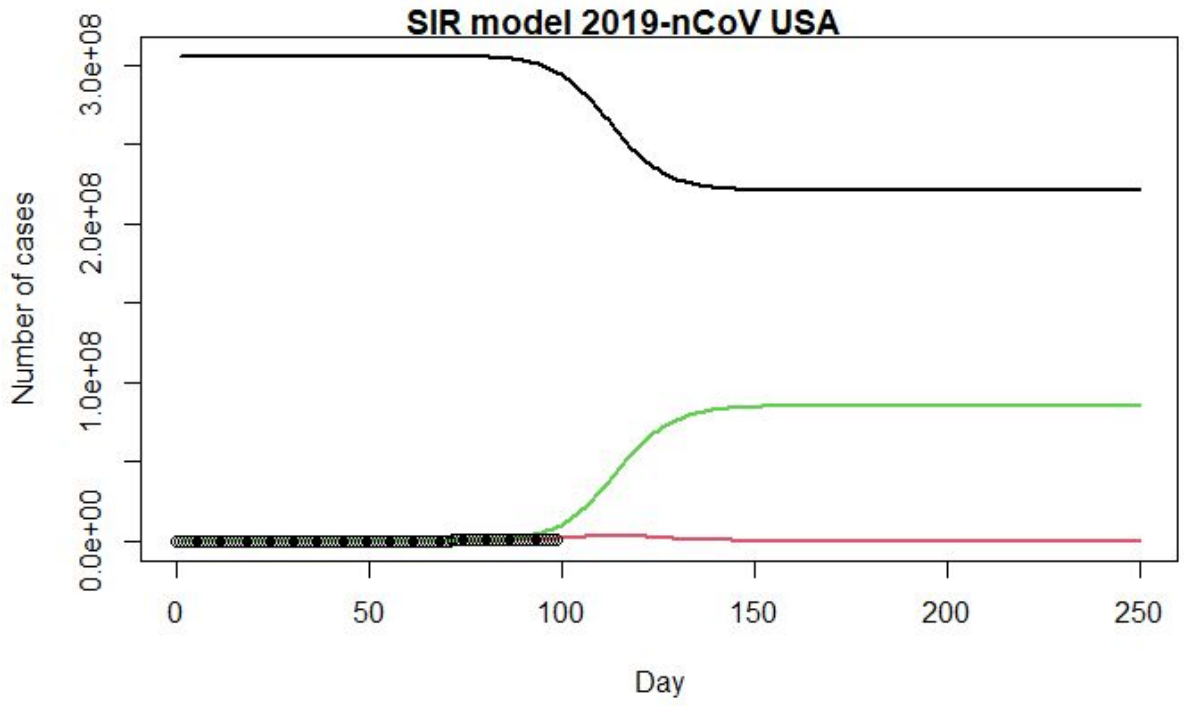
$$RSS(\beta,\gamma) = \sum_t (I(t) - \hat{I}(t))^2$$

R0 is the reproduction number. It describes the intensity of an infectious disease outbreak. R0 smaller than 1 indicates that the disease is declining and will disappear soon. The original R0 for SARS is 2.75 and it decreases to 1 after two months. Some of the research indicates that the R0 of H1N1 is between 1.4-1.6. For covid-19, researchers indicate it is between 2-3. We will examine R0 in our model to demonstrate the intensity of the outbreak.

$$R_0 = \frac{\beta}{\gamma}$$

For this SIR model, it is difficult to accurately estimate the number of the initial population size N. We simulate the result using N equals to the current population in continental United States and N equals the U.S. population.

This is the SIR model for N equals to the population:



β = 1 and γ = 0.8532

$$R_0 = \frac{\beta}{\gamma} = 1.17$$

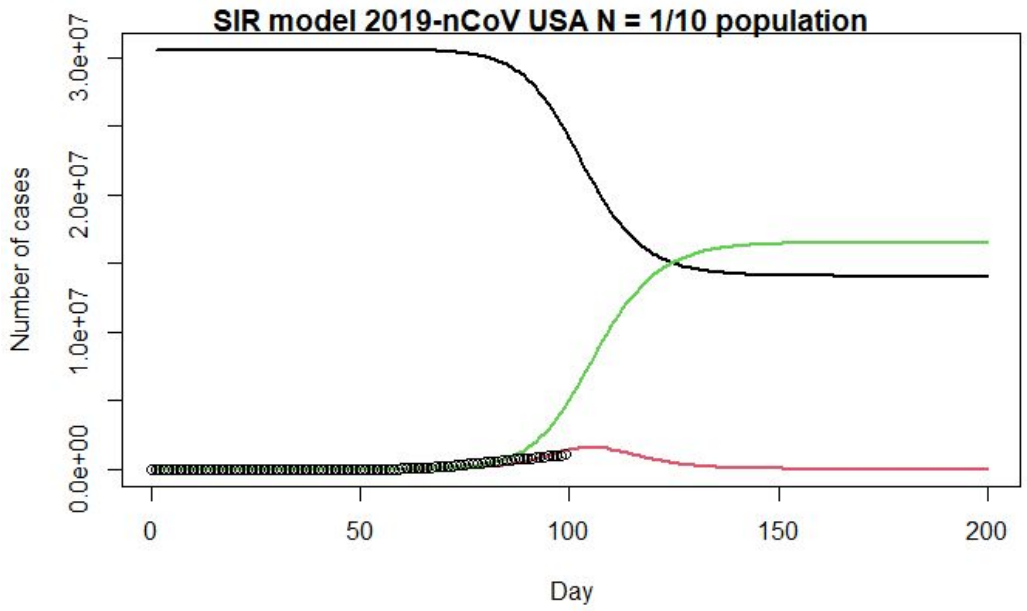
The prediction of climax:

	time <dbl>	S <dbl>	I <dbl>	R <dbl>
114	114	259803407	3466819	42729774

1 row

This prediction indicates that the climax of the outbreak is around 114 days after the first cases were detected. The date of the beginning of the pandemic is 2020-01-21. Therefore the climax is 2020-05-14 according to the model. The predicted number of infected cases is 3466819. This number is too large for the present situation. This error may relate to the large number of susceptible people since not all individuals in the U.S. are susceptible to the disease. R0 is 1.17 and this indicates that the spread of the epidemic is under control.

We create the model for susceptible people of the U.S. population at the beginning of the pandemic. The plot of SIR model for n = 1/10 N：



β = 0.5 and γ = 0.3476811

$$R_0 = \frac{\beta}{\gamma} = 1.44$$

The prediction of climax:

	time <dbl>	S <dbl>	I <dbl>	R <dbl>
105	105	21283535	1591113	7725352

1 row

After 240 days:

	time <dbl>	S <dbl>	I <dbl>	R <dbl>
240	240	14074128	0.6677864	16525871

1 row

1 row

This prediction indicates that the climax of the outbreak is around 105 days after the first cases were detected. The date of the beginning of the pandemic is 2020-01-21. Therefore the climax is 2020-05-05 according to the model. The predicted number of infected cases is 1591113. This number is higher than the recent data but it is closer to the actual data than the first model. According to this model, there will be less than 1 case after 240 days since the first cases were detected. Therefore, the disease will die out in the middle of September. R0 is 1.44 and this indicates that the spread of the epidemic is under control.

Conclusion

Based on empirical analysis into the possible variables, the number of death from COVID19 is determined by three aspects - the state of the population's health as measured by *Public Health Ranking*, the effectiveness of the health system as measured by *Healthcare Access* and *Healthcare Quality* and the characteristics of the population as measured by *Population Density* and *Cases*. Based on limited available data, a preliminary extrapolation of the data shows a mortality rate of below 0.06 with further increased number of cases and deaths in the United States into May. Based on the limited data released and available, it is to certain extent that these three aspects play a crucial role in determining how deadly COVID19 is. Whether there is any other variable

not present in the model also playing a deterministic role shall be studied once such data is made available.

The SIR model is a simple model in epidemiology and some factors like exposed cases and policies are not considered, so the prediction according to the model is not very close to the real-world data. However, this model can demonstrate the development of the pandemic and approximate the climax and end of the outbreak.

Reference

[1]"Coronavirus Data in the United States" New York Times
https://github.com/nytimes/covid-19-data(data source)

[2] "Health Care Rankings" U.S.News
https://www.usnews.com/news/best-states/rankings/health-care(data source)

[3]"population density in the U.S by federal states including the District of Columbia in 2019"
https://www.statista.com/statistics/183588/population-density-in-the-federal-states-of-the-us/(data source)

[4]"FDA Issues first Emergency Use Authorization for Point of Care Diagnostic" FDA
https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-issues-first-emergency-use-authorization-point-care-diagnostic

For SIR model:

https://labblog.uofmhealth.org/rounds/how-scientists-quantify-intensity-of-an-outbreak-like-covid-19

https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/

https://www.who.int/csr/sars/en/WHOconsensus.pdf

https://www.healthline.com/health/r-nought-reproduction-number