COVID-19 US Prediction

MFR

6/5/2020

File Name: SIR-US-Prediction.Rmd

Executive Summary:

This project made use of two different datasets, one is a past dataset from Kaggle [1] and another one is an ongoing data from JHU CCSE [2]. Analyses revealed that two datasets contain different number of columns with different types of information. Kaggle data is a past COVID-19 data covering mostly the onset of pandemic in China on the other hand JHU CCSE is an ongoing data mainly to be used for "COVID-19 US Prediction" using SIR model. Maching learning algorithms are applied on Kaggle dataset that revealed some interesting facts related to mostly to gender and age of the pandemic patients. Kaggle dataset revealed that mean age of 48 years who survived, on the other hand the mean age of about 68.6 years and older were victims of death. After calculating the means, we see that men in this dataset have a death rate of 8.5% as opposed to 3.7% in women. The accuracy of this calculation is being further verified by calculating the Accuracy = 0.5524862, Sensitivity = 0.5194805, Specificity = 0.5769231 and with a kind of high Prevalence = 0.4254144.

Therefore, the main focus of this project is based on the study of the propagation pattern of the COVID19 pandemic in US using a widely used SIR model. A R-library is created that can be fetched from Github [3]. The prediction is made over a period of 150 days starting from the first US reported death on February 29 up to the projected end sometime at the end of July, 2020. In this analysis an initial population, N is considered to be 20.23M which is the combined population of New York City (8.399M), Washington (7.615M) and Oregon (4.218M). We calculated the optimized parameter β that controls the transition between S to I (susceptible (S) to infectious (I)) and γ that controls the transition between I to R (infectious (I) to recovered (R)) to be 0.4132663 and 0.2659657 respectively. Furthermore, we estimated the reproduction number of 1.55, and using the formula $1 - \frac{1}{R_0} = 1 - \frac{1}{1.44} = 0.35$, about 35.00% of the population of US which is about 115 million, need to be vaccinated to stop the pandemic in the future. Based on the updated US data for the period of February 29 to June 04, the model predicts what happens if the pandemic continues over a total time span of 150 days or longer. It is estimated from the analysis that peak infected case already happened around May 20, 2020, and the pandemic is expected to come back to the normal state sometime in the middle of July, 2020. The **RMSE** of the prediction comes out to be 695 which is very small with respect to the big number of infected cases in the US. This paper is based on the similar case study reported in [4] but adapted for US with modifications with our own library and methods.

Introduction:

To analyze the Kaggle dataset and for performing the **Machine Learning** the caret package used for generating the training and test sets is by splitting the data with specified proportions. The **confusion matrix**, which basically tabulates each combination of prediction and actual value were calculated along with t-test to verify the accuracy of the results.

The data from JHU CCSE was mainly used for up-to-date status of the COVID-19 in the US and the code can be easily modified to apply for other countries. Users can choose from a variety of common Ordinary Differential Equation (ODE) models (such as the SIR, SEIR, SIRS, and SIS models), or create on their own to study various kinds of epidemic and pandemic diseases. The objective of this study is to find ways to

adapt COVID-19 data to make predictions of the propagation of COVID-19 and take appropriate measures to slow down the spread of the disease to save human lives.

In this analysis we expect to estimate the expected numbers of individuals over time who will be infected, recovered, susceptible, or dead. Infected individuals first pass through an exposed/incubation phase where they are asymptomatic and not infectious, and then move into a symptomatic to infectious stage classified by the clinical status of infection (mild, severe, or critical). The initial population size, initial conditions, and parameter values are used to simulate the spread of infection. In SIR model we consider S susceptible, I infectious and R recovered cases to start with the simulation. Susceptible become infected at a rate equal to the product of an infectious contact rate β and the number of infectious I. Infectious people recover at a rate γ.

Timeline of COVID-19 Pandemic:

The first outbreak of novel corona virus was reported sometime in December 2019 from Wuhan, China. In the span of about four months this virus spread to almost all countries in the world. As of May 10, 2020, more than 4.10 million people got infected and about 0.28 million people lost their lives. Many data scientists are trying to predict the nature and timeline of its propagation to save human lives as many as possible by taking all practical measures including social distancing in their models. The first death case was reported by Chinese state media on January 11, 2020, of a 61-year-old man from Wuhan, the capital city of Hubei province. On January 21, 2020, United States declared its first infected case with a man in his 30s from Washington State, who traveled to Wuhan. Oregon, Washington and New York soon report their own cases of possible community transmission. On January 23, 2020, Chinese government imposed strict measures in Wuhan by shutting down almost all activities including travel restrictions in air, trains, subways, buses and put restrictions on public gathering, mass congregation, closing all schools, and colleges. On January 30, 2020, WHO declares global health emergency a **"public health emergency of international concern".** On February 11, 2020, novel coronavirus was named as **Corona Virus Disease 2019** with its abbreviation as **COVID-19**. On February 26, 2020, the Centers for Disease Control and Prevention (CDC) confirms the first case of COVID-19 infection in a patient in California with no travel history to an outbreak area. The first reported death in the USA was on February 29, 2020, however, a death in Santa Clara, California, on February 6 is considered to be the first COVID-19 death. In this analysis the data is considered to be from February 29, 2020, to June 04, 2020, but the prediction is made upto the end of July, 2020.

Loading the required libraries

Downloading dataset from Kaggle for initial analysis

https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/discussion/139288

```
## [1] 1085   27
```

Clening data by elimating NA's

```
## [1] 0.05806452
```

Age Analysis

After running the abovee code, we discover that the mean age of 48 years who survived, on the other hand the mean age of about 68.6 years and older who have died.

```
## [1] 68.58621
```

```
## [1] 48.07229
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  dead$age and alive$age
## t = 10.839, df = 72.234, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  16.74114 24.28669
## sample estimates:
## mean of x mean of y
##  68.58621  48.07229
```

We see from the above code that the death rate of our dataset, which is about 5.8%. Next we will try to analyze the age group of people who have died and of those who did not.

Gender Analysis

```
## [1] 0.08461538
```

```
## [1] 0.03664921
```

```
##
##  Welch Two Sample t-test
##
## data:  men$death_dummy and women$death_dummy
## t = 3.084, df = 894.06, p-value = 0.002105
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01744083 0.07849151
## sample estimates:
##  mean of x  mean of y
## 0.08461538 0.03664921
```

```
##      ï..id        case_in_country     reporting.date    X
##  Min.   :   1.0   Min.   :   1.00   1/22/2020: 61    Mode:logical
##  1st Qu.: 245.2   1st Qu.:  12.00   1/25/2020: 52    NA's:902
##  Median : 477.5   Median :  35.00   2/27/2020: 51
##  Mean   : 498.0   Mean   :  54.65   2/26/2020: 47
##  3rd Qu.: 762.5   3rd Qu.:  72.00   2/20/2020: 45
##  Max.   :1085.0   Max.   :1443.00   1/24/2020: 41
##                   NA's   :197       (Other)  :605
##
##  new COVID-19 patient confirmed in Bahrain: female, returning from Iran
##  new COVID-19 patient confirmed in Bahrain: female, Saudi Arabian, returning from Iran
##  new COVID-19 patient confirmed in Bahrain: male, returning from Iran
##  new confirmed COVID-19 patient in Germany: Baden-Wuerttemberg, female, attended business meeting in
##  new confirmed COVID-19 patient in Japan: Ishikawa Prefecture, female, 50s
##  (Other)
##  NA's
##          location          country       gender         age
##  Singapore   : 91   China      :197   female:382   Min.   : 0.50
##  South Korea : 90   Japan      :185   male  :520   1st Qu.:35.00
##  Hong Kong   : 80   Hong Kong  : 94               Median :51.00
##  Hokkaido    : 47   South Korea: 92               Mean   :49.76
##  Taiwan      : 34   Singapore  : 91               3rd Qu.:64.00
##  Wuhan, Hubei: 32   Taiwan     : 34               Max.   :96.00
```

```
##  (Other)     :528   (Other)    :209                NA's   :77
##    symptom_onset If_onset_approximated  hosp_visit_date   exposure_start
##  1/23/2020: 27   Min.   :0.0000        1/23/2020: 34    1/26/2020: 14
##  1/24/2020: 21   1st Qu.:0.0000        1/24/2020: 30    1/25/2020:  7
##  1/25/2020: 21   Median :0.0000        1/21/2020: 20    01/12/20 :  6
##  1/30/2020: 19   Mean   :0.0432        1/20/2020: 19    1/20/2020:  6
##  02/03/20 : 18   3rd Qu.:0.0000        2/17/2020: 19    1/22/2020:  6
##  (Other)  :452   Max.   :1.0000        (Other)  :384    (Other)  : 80
##  NA's     :344   NA's   :347           NA's     :396    NA's     :783
##     exposure_end visiting.Wuhan     from.Wuhan           death
##  1/22/2020: 34   Min.   :0.0000   Min.   :0.0000   0        :844
##  1/23/2020: 27   1st Qu.:0.0000   1st Qu.:0.0000   1        : 42
##  1/20/2020: 26   Median :0.0000   Median :0.0000   2/23/2020:  4
##  1/21/2020: 23   Mean   :0.1818   Mean   :0.1682   2/26/2020:  3
##  1/19/2020: 19   3rd Qu.:0.0000   3rd Qu.:0.0000   2/27/2020:  2
##  (Other)  :191   Max.   :1.0000   Max.   :1.0000   02/01/20 :  1
##  NA's     :582                    NA's   :4        (Other)  :  6
##       recovered                     symptom
##  0         :749                        :633
##  2/18/2020 : 13   fever                : 73
##  2/19/2020 : 12   fever, cough         : 36
##  02/12/20  : 11   cough                : 13
##  12/30/1899: 11   fever, malaise       :  7
##  2/21/2020 :  9   fever, cough, malaise:  6
##  (Other)   : 97   (Other)              :134
##                           source
##  Ministry of Health          :170
##  å¤®è§†æ-°é-»                :133
##  KCDC                        : 81
##  Ministry of Health Singapore: 67
##  Government HK               : 60
##  Channel News Asia           : 28
##  (Other)                     :363
##
##  https://www.mhlw.go.jp/stf/houdou/houdou_list_202002.html
##  https://www.mhlw.go.jp/stf/newpage_09713.html
##  https://www.moh.gov.sg/news-highlights/details/two-more-cases-discharged-two-new-cases-of-novel-cor
##  https://www.mhlw.go.jp/stf/newpage_09652.html
##  http://www.nhc.gov.cn/yjb/s3578/202001/5d19a4f6d3154b9fae328918ed2e3c8a.shtml
##  https://m.weibo.cn/status/4464497211305006?
##  (Other)
##    X.1             X.2             X.3             X.4             X.5
##  Mode:logical   Mode:logical   Mode:logical   Mode:logical   Mode:logical
##  NA's:902       NA's:902       NA's:902       NA's:902       NA's:902
##
##
##
##
##
##    X.6            death_dummy
##  Mode:logical   Min.   :0.0000
##  NA's:902       1st Qu.:0.0000
##                 Median :0.0000
##                 Mean   :0.0643
```

```
##                 3rd Qu.:0.0000
##                 Max.   :1.0000
##
```

```
## [1] 902  28
```

We subset our original data into two sets. After calculating the means, we see that men in this dataset have a death rate of 8.5% as opposed to 3.7% in women. Well, this is unexpected. Again, can we trust this data? Here is the t.test output:

Our confidence interval of 95% shows that man will have from 1.7% to 7.8% higher death rate than women. A p-value of 0.002 signifies that we can reject the null hypothesis that men and women have the same death rate, since $0.002 < 0.05$. There have been articles written that men indeed do have a higher coronavirus death rate. Here is one of them if you are interested.

Loading caret package for Machine Learning

```
## female   male
##    382    520
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.50   35.00   51.00   49.76   64.00   96.00      77
```

```
## [1] 0.5524862
```

```
##          actual
## predicted female male
##    female     40   44
##    male       37   60
```

```
## # A tibble: 2 x 2
##   gender accuracy
##   <fct>     <dbl>
## 1 female    0.519
## 2 male      0.577
```

```
## [1] 0.4235033
```

```
##  Accuracy
## 0.5524862
```

```
## Sensitivity Specificity  Prevalence
##   0.5194805   0.5769231   0.4254144
```

Loading covid19 Library that updates COVID-19 Data from JHU-CCSE:

The 'covid19' dataset contains the worldwide daily **confirmed**, **recovered**, and **death** cases of the COVID-19 (the 2019 Novel Coronavirus COVID-19). Let's load the dataset from the **covid19** package:

The dataset has the following fields:

- `date` - The date of the summary
- `Province.State` - The province or state, when applicable
- `Country.Region` - The country or region name
- `Lat` - Latitude point
- `Long`- Longitude point
- `cases` - the number of daily cases (corresponding to the case type)
- `type` - the type of case (i.e., confirmed, death)

We can use the `head`, 'tail' and `str` functions to see the structure of the dataset:

```
##   Province.State Country.Region Lat Long       date cases      type
## 1                    Afghanistan  33   65 2020-01-22     0 confirmed
## 2                    Afghanistan  33   65 2020-01-23     0 confirmed
## 3                    Afghanistan  33   65 2020-01-24     0 confirmed
## 4                    Afghanistan  33   65 2020-01-25     0 confirmed
## 5                    Afghanistan  33   65 2020-01-26     0 confirmed
## 6                    Afghanistan  33   65 2020-01-27     0 confirmed
```

```
##        Province.State Country.Region     Lat     Long       date cases
## 105970       Zhejiang          China 29.1832 120.0934 2020-05-30     0
## 105971       Zhejiang          China 29.1832 120.0934 2020-05-31     0
## 105972       Zhejiang          China 29.1832 120.0934 2020-06-01     0
## 105973       Zhejiang          China 29.1832 120.0934 2020-06-02     0
## 105974       Zhejiang          China 29.1832 120.0934 2020-06-03     0
## 105975       Zhejiang          China 29.1832 120.0934 2020-06-04     0
##             type
## 105970 recovered
## 105971 recovered
## 105972 recovered
## 105973 recovered
## 105974 recovered
## 105975 recovered
```

```
## 'data.frame':    105975 obs. of  7 variables:
##  $ Province.State: chr  "" "" "" "" ...
##  $ Country.Region: chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Lat           : num  33 33 33 33 33 33 33 33 33 33 ...
##  $ Long          : num  65 65 65 65 65 65 65 65 65 65 ...
##  $ date          : Date, format: "2020-01-22" "2020-01-23" ...
##  $ cases         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ type          : chr  "confirmed" "confirmed" "confirmed" "confirmed" ...
```

Countrywise Tabular Data

Show 10 entries Search:

	Confirmed	Death	Recovered	Death Rate
US	1872660	108211	485002	5.78%
Brazil	614941	34021	254963	5.53%
Russia	440538	5376	204197	1.22%
United Kingdom	283079	39987	1219	14.13%
Spain	240660	27133	150376	11.27%
Italy	234013	33689	161895	14.40%
India	226713	6363	108450	2.81%
France	189569	29068	70094	15.33%
Germany	184472	8635	167909	4.68%
Peru	183198	5031	76228	2.75%

Showing 1 to 10 of 174 entries Previous 1 2 3 4 5 … 18 Next

Coronavirus - Worldwide Cumulative Distribution as Bar Diagram

```
## # A tibble: 3 x 2
##   type        cases
##   <fct>       <int>
## 1 confirmed 6632985
## 2 death      391136
## 3 recovered 2869963
```

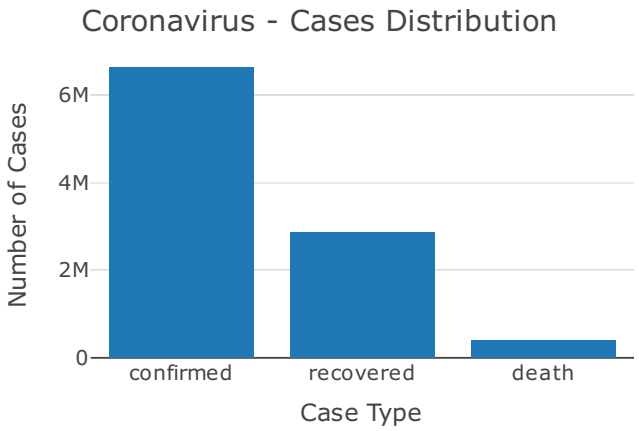

Coronavirus - Cases Distribution
Number of Cases
6M
4M
2M
0
confirmed
recovered
death
Case Type

Table of the top ten countries with the highest confirmed cases.

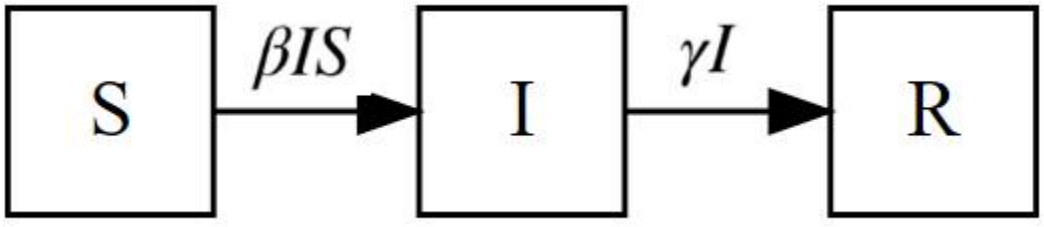
Show 10 entries Search:

Country	Cases	Perc of Total
US	1872660	28.23%
Brazil	614941	9.27%
Russia	440538	6.64%
United Kingdom	283079	4.27%
Spain	240660	3.63%
Italy	234013	3.53%
India	226713	3.42%
France	189569	2.86%
Germany	184472	2.78%
Peru	183198	2.76%

Showing 1 to 10 of 10 entries Previous 1 Next

The SIR Model for Spread of Disease:

In this study SIR model is used. A graphical representation of SIR Model is shown in the figure below:



SIR model has 3 variables S, I and R which are respectively the numbers of susceptibles, infectious and recovered, and has two parameters β and γ which are respectively the infection rate and the recovery rate.

The basic idea behind the SIR model is based on three groups of people with the following assumptions:

S: is assumed to be the entire population of New York City, Washington State and Oregon who were susceptible to the disease since no one was immued to the disease,

I: the infected population, and

R: individuals who were contaminated but who have either recovered or died.

The dependent variables represent the fraction of the total population in each of the three categories. In this analysis N is cosidered to be $20,232,000$ which is the combined population of New York City (8.399M), Washington (7.615) and Oregon (4.218) where the pandemic started in the USA on *February* 29, 2020.

$$\frac{dS}{dt} = -\beta S(t)I(t)/N \tag{1}$$

where β is the rate of infection, which controls the transition between S and I,

$$\frac{dI}{dt} = \beta S(t)I(t)/N - \gamma I(t) \tag{2}$$

and γ is the recovery rate, which controls the transition between I and R.

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

The first equation (Eq. 1) states that the number of susceptible individuals (S) decreases with the number of newly infected individuals, where new infected cases are the result of the infection rate β multiplied by the number of susceptible individuals (S) who had a contact with infected individuals (I).

The second equation (Eq. 2) states that the number of infectious individuals (I) increases with the newly infected individuals βIS, minus the previously infected people who recovered (i.e., γI which is the removal rate γI multiplied by the infectious individuals I).

Finally, the last equation (Eq. 3) states that the recovered group (R) increases with the number of individuals who were infectious and who either recovered or died (γI).

An epidemic develops as follows:

- Before the start of the disease outbreak, S equals the entire population as no one has anti-bodies.
- At the beginning of the outbreak, as soon as the first individual is infected, S decreases by 1 and I increases by 1 as well.
- This first infectious individual contaminates (before recovering or dying) other individuals who were susceptible.
- The dynamic continues, with recently contaminated individuals who in turn infect other susceptible people before they recover.

Solving Differential Equations of SIR Model:

Solving a system of differential equations means finding the values of the variables (here S, I and R) at a number of points in time. These values will depend on the parameters' values. We can numerically solve differential equations in R using **ode()** function of the **deSolve** package. If this package is not installed on your system, you need to install it.

The initial values of the variables need to be defined in a named vector:

Put the daily cumulative incidence numbers for US from February 29 to May 10 into a vector called **Infected**. Create an incrementing Day vector with the same length as our cases vector.

Training and testing using the residual mean squared error (RMSE):

The regression approaches for epidemic analysis are trained and tested on realtime data using the number of confirmed, recovered, and death cases as the label for the corresponding day. The residual sum of squares (RSS) is the most widely used objective function and root mean square error (RMSE) as a metric function for evaluating the regression models.

Passing in parameters beta and gamma that are to be optimized for the best fit to the incidence data. Fitting a SIR model to the US data we need the following two things:

1. a solver for solving the three differential equations

2. an optimizer to find the optimal values for our two unknown parameters, β and γ.

The function ode() (for ordinary differential equations) from the {deSolve} R package makes solving the system of equations easy, and to find the optimal values for the parameters we wish to estimate, we can just use the optim() function built into base R.

Specifically, what we need to do is to minimize the sum of the squared differences between I(t), which is the number of infected people at time t, and the corresponding number of cases as predicted by the model $\hat{I}$. The quantity to minimize is called residual mean squared error (RMSE). Passing in parameters beta and gamma that are to be optimized for the best fit to the incidence data. In order to fit a model to the incidence data for US, we need a value N for the initial uninfected population. To start the simulation we need to specify initial values for N, S, I and R. Here, N is assumed to be the population of New York City (8.399M), Washington (7.615) and Oregon (4.218) combined is N = 20232000.

Next, we need to create a vector with the daily cumulative incidence for US, from February 29 (when first daily incidence data starts in the US), through to June 04 (last available data). We will then compare the predicted incidence from the SIR model fitted to these data with the actual incidence since February 29. The daily cumulative incidence data for US is loaded from the {covid19} R-package developed in this work.

$$RMSE(\beta, \gamma) = \sqrt{\frac{1}{N}\sum_{t}(I(t) - \hat{I}(t))^2} \tag{4}$$

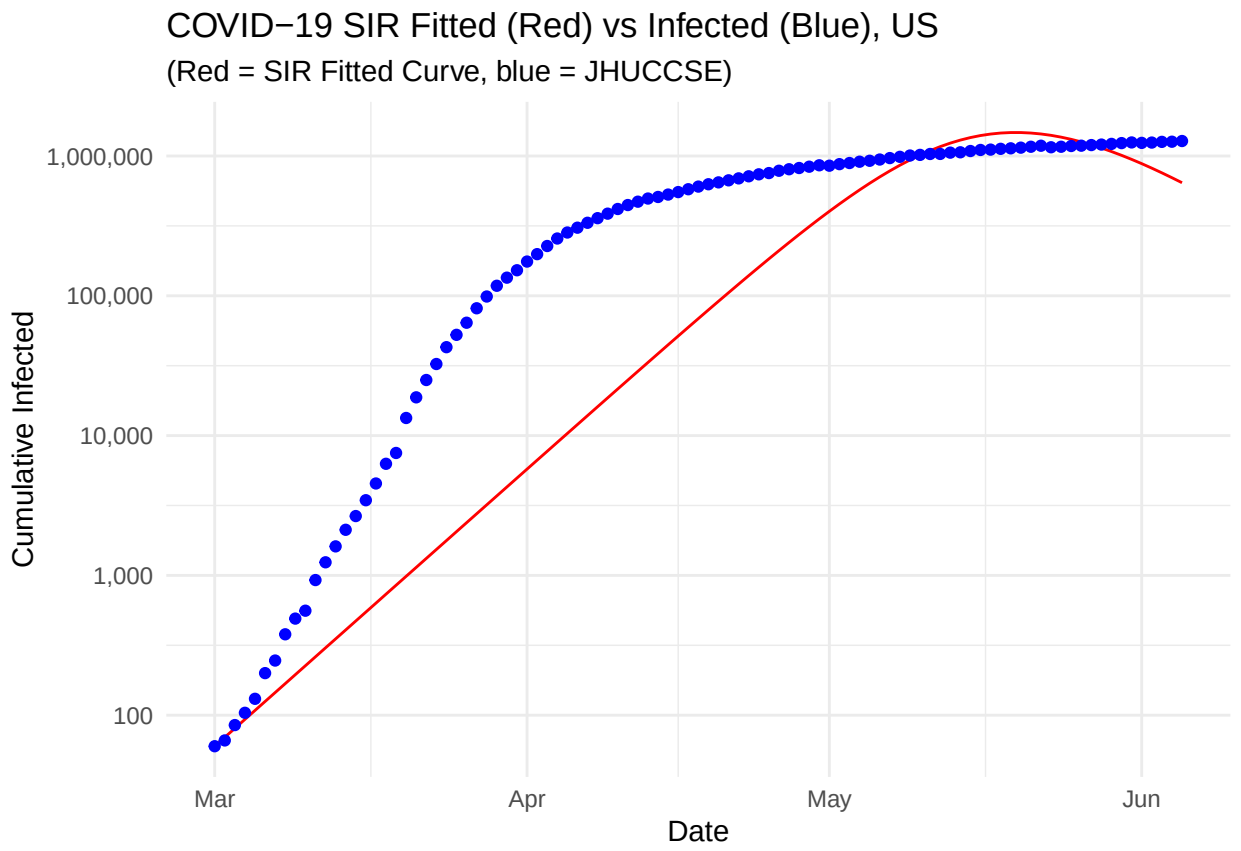
where I(t) indicates the real infected dat value, and $\hat{I}(t)$ indicates the predicted value, and N is the total number of days predicted.

Now we find the values of beta and gamma that give the smallest RMSE, which represents the best fit to the data. First we check for convergence and once it is confirmed then we get the fitted or optimized parameters β and γ. Once the convergence is confirmed then we can examine the fitted values for beta and gamma for further analysis and prediction of the spread of COVID19.

```
## [1] "ERROR: ABNORMAL_TERMINATION_IN_LNSRCH"
```

```
##      beta     gamma
## 0.4132637 0.2659637
```

With the optimized parameter β and γ that controls the transition between S to I (susceptible (S) to infectious (I)) and controls the transition between I to R (infectious (I) to recovered (R)) respectively. The plot for the fitted data with the observed data collected from JHU in the time span of February 29 to June 04 is shown in the figure below. The y-axis of the following plot is on a log10 scale versus time in days.



Reproduction number R_0

In this study SIR model with best fit to the observed cumulative incidence data in US, so as to get the correct reproduction number R_0, also referred to as basic reproduction ratio, and is given by:

$$R_0 = \frac{\beta}{\gamma} \tag{5}$$

In other words, the reproduction number R_0 refers to the number of healthy people that get infected per number of infected people.

If R_0 is greater than $R_0 > 1$, the infection will spread exponentially. If R_0 is less than $R_0 < 1$, the infection will spread very slowly, and it will eventually die down. The higher the value of R_0, the faster an epidemic or a pandemic will spread.

In this simulation R_0 comes out to be 1.55 which is relatively lower than reported by others for COVID-19 which is in the range 2.2–2.7 for Wuhan, China [5]. In the literature, it has been estimated that the reproduction number for COVID-19 is approximately 2.7 (with β close to 0.54 and γ close to 0.2). Our reproduction number being lower is mainly due to the fact that the number of confirmed cases stayed constant and equal to 1 at the beginning of the pandemic.

A R_0 of 1.55 means that, on average in US, 1.55 persons are infected for each infected person. For simple models, the proportion of the population that needs to be effectively immunized to prevent sustained spread of the disease, known as the "herd immunity threshold", has to be larger than $1 - \frac{1}{R_0}$.

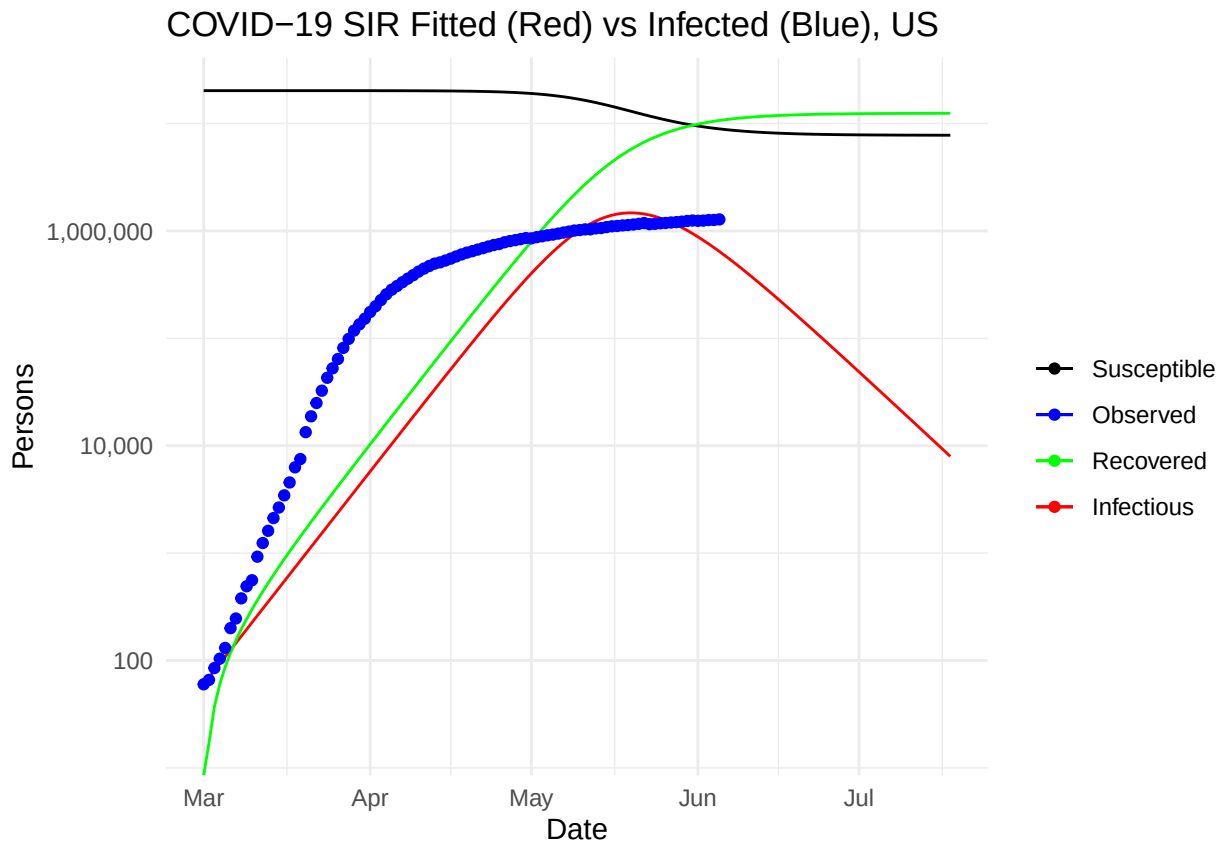
The reproduction number of 1.55 that we estimated suggests that about 35.00% of the population should be immunized to stop the spread of the infection. Considering the population in US of approximately 328.20 million, about 115 million people need to be vaccinated to stop the spread of this disease.

```
##      beta     gamma
## 0.4132637 0.2659637
```

```
## [1] 1.553835
```

Now we make time series predictions in days for the period of 150 days and plot them to see their pattern for finding the end date for this pandemic.

Based on the available US data for the period of February 29 to June 04, the model now preddicts what happens if the pandemic continues over a total time span of 150 days. The y-axis is in log10 scale in persons versus the time in days.



The date of peak of pandemic in the US with maximum infected and maximum death. Maximum Infected = confirmed - death - recovered.

```
##          Date       I
## 81 2020-05-20 1471342
```

```
## [1] 109997.6
```

RMSE with optimized parameter

```
## [1] 695.0369
```

Conclusions:

This paper is the outcome of a very simple SIR epidemiological model. The numbers we ended up with are very close to the actual data. A better model called SEIR could be used for accomodating the effect of social distancing. SEIR model is similar to the SIR model, where S stands for Susceptible and R stands for Recovered,

but the infected people are divided into two categories, E stands for the exposed/infected is asymptomatic and I for the infected and symptomatic. Both SIR and SEIR models belong to the continuous-time dynamic models that assume fixed transition rates. In our analysis we used a fixed reproduction number R_0 but it reality it is much more realistic to estimate the current effective reproduction number R_e on a day-by-day basis so as to track the effectiveness of public health interventions, such as social distancing and increased number of tests etc. This project resulted in a numerical comparison with visual presentation of the fitted and observed cumulative incidence in US. The graphs indicated the exponetial nature of the COVID-19 pandemic in US with peak reaching sometime in May 20 and is expected to fall and come to a normal state sometime in the end of July, 2020, provided public health intervention is in strict compliance.

We estimated the reproduction number of 1.55 using the optimized β and γ that suggests that about 35.00% of the population should be immunized to stop the spread of the infection. Considering the total population in US of approximately 328.2 million, about 115 million need to be vaccinated. Based on the available US data for the period of February 29 to June 04, the model now preddicts what happens if the pandemic continues over a total time span of 150 days. It is estimated from the analysis that peak infected case will happen around May 20, 2020, and is expected to come down to the normal state sometime at the end of July, 2020. Humans are expected to fight in against of the natural epidemic or pandemic with their all kinds of resources because life is more important than the material resources and wealth. However, we must also ensure that there are sufficient resources, medical supplies and PPEs are available for treating infected patients that will help them to survive or increase the chances for their survival rates.

Unlike the small **RMSE** value of Movilens project which was rated with maximum rating of 5.0. The estimated **RMSE** value in this project is about **695** with optical parameters which is very small with respect to US infected cases of the order of more than 100,000 and bigger.

References:

1. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/discussion/139288
2. https://systems.jhu.edu/research/public-health/ncov/
3. https://github.com/irays/covid19
4. https://www.statsandr.com/blog/covid-19-in-belgium/#more-sophisticated-projections.
5. High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus-2, Volume 26, Number 7—July 2020, Steven Sanche, Yen Ting Lin, Chonggang Xu, Ethan Romero-Severson, Nick Hengartner, and Ruian Ke.