Indian Institute of Technology Goa

Thinking like Epidemiologists!

Authors

- Sanjay Marreddi 1904119 EE
- Rishabh Tripathi 1904129 EE

Affiliated to Dr. Sreenath Balakrishnan,
Instructor, IIT GOA.

Abstract: Our main aim is to understand the basics of *Epidemiology* and appreciate its complexity. With the help of mathematical modelling and coding we fitted the *COVID-19* data into two compartment models *SIR* and *SEIR* to understand, analyse, estimate as well as forecast currently prevailing Biological phenomena, *COVID-19 Epidemic*.

*Level-**1***: **SIR** model for COVID-19 data of **Georgia**, US.

Introduction: *SIR* model is a simple mathematical description of the spreading of a disease in a population, which divides the population of N individuals into three "compartments" which may vary with a function of time.

- S(t)→(Susceptible) are those people who can affected by disease in future but not yet infected with the disease.
- I(t)→(Infected) Is the number of infectious individuals.
- R(t)→(Recovered) are those individuals who have recovered from the disease and now have immunity to fight with disease.

The SIR model describes the change in the population of each of these compartments in terms of two parameters, β and γ.

- ❖ β is expected amount of people that an infected person infects per day. γ is the mean recovery rate.

The differential equations describing this model:-

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}t} &= -\frac{\beta SI}{N},\\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \frac{\beta SI}{N} - \gamma I,\\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I.\end{aligned}$$

Method:

- *Data Source*:

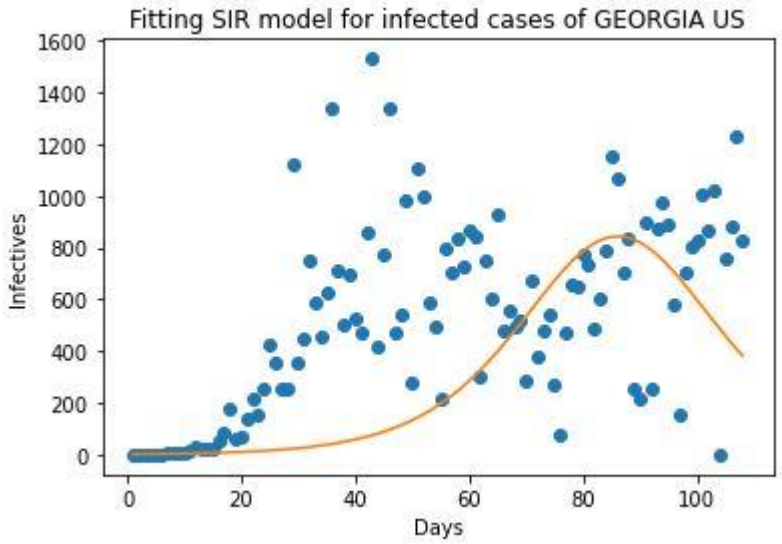
  We collected the data of Confirmed Cases and Fatalities of Georgia, US during the interval from 22-01-2020 to 18-06-2020 from the GitHub data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Centre for Systems Science and Engineering (JHU CSSE) which can be found using the below link: *https://github.com/CSSEGISandData/COVID-19*

- *Details of SIR Model:*

  First, we defined the equations governing the SIR model using functions. Then using the Python Modules like scipy and sklearn we imported functions like integrate, optimize and odeint for solving the Ordinary Differential Equations. Then Initialization of variables is done. Now using the pandas library, we took input csv files and used the data for Fitting Procedure. Once the data is fitted, we obtained the Optimized parameters of the model. From these model parameters, we predicted the rest of the quantities like Transmissibility, Recovery rate, R_o, maximum number of infectives, the time location of the maximum number of infectives, total number of recovered and duration of the epidemic.

- *Fitting Procedure:*

  We used the Infectives data from input csv files and tried to fit a curve through the data points using the optimization function 'curve_fit' which is imported from scipy.optimize. So as a result of fitting curve into data , we estimated the model parameters β and γ. The plot of curve fitting is as shown below.



- *Note:* Since we used a simple Optimization function which fits a Quadratic polynomial into the given data points ,we can see that some points are missed during the fitting procedure. However, we can reduce this error by either Data enrichment or using some complex functions and models which fits a higher degree polynomial into the data points like Logistic Regression.

Results obtained from Coding:

- ✓ Beta(β) = *6.944192705488434*
- ✓ Transmissibility(r)=β/N = *6.540374915352279X10^{-7}*
- ✓ Recovery Rate(γ=a) = *6.8569660606622875*
- ✓ Relative Removal Rate (ρ=a/r) = *10484056.57077375*
- ✓ Contact Rate(σ=r/a) = *9.538292675639363X10^{-8}*
- ✓ Basic Reproductive Rate (R_o=rS_o/a) = *1.0127206895847956*

- ✓ Maximum number of Infectives in a single day:
  (I_{max} = N- ρ+ ρln(ρ/S_o)=*843*

- ✓ Time location of maximum number of infectives is 85th day from day of initial infection i.e. *26thMay 2020*

- ✓ *Total number of recovered is around 2,65,687*

- ✓ Duration of Epidemic in Georgia,US is *approximately 179 days from the day of initial infection.*

Numerical Verification of our SIR model

- ➢ For US, we calculated the average death rate by :-
  davg = (Number of deaths)/(number of cured + number of deaths).
  Now we assumed that for Georgia, the average death rate is the national average and it remains constant with time and estimated the number of recovered each day by R(t) = Number of deaths/davg

- ✓ Deaths in US = *118432 (up to 18 June 2020)*
- ✓ Recovered in US = 599115 *(up to 18 June 2020)*
- ✓ Death average of Georgia = *0.16505120919*
- ✓ Total Death in Georgia = *2564 (up to 18 June 2020)*
- ✓ Recovered in Georgia = *15535 (up to 18 June 2020)*
- ✓ *Total number of Recovered by Coding = 18113 (up to18 June 2020)*

- ➢ Since the number of recovered in both the cases is of same *order* , we can verify that our SIR model is working in a better way only amidst the limitations and errors associated with this model.

Analysis of Model Results and its Predictions:

Now we will compare our results obtained from coding with actual scenario regarding COVID-19 in *Georgia, US*.

- We got Basic Reproductive Rate (R_o) as *1.0127206895847956* from coding which is approximately matching with practical result, *1.41*. This value of R_o justifies that currently the COVID-19 is pandemic in nature in Georgia, US because Basic Reproductive Rate (R_o) *>1*.

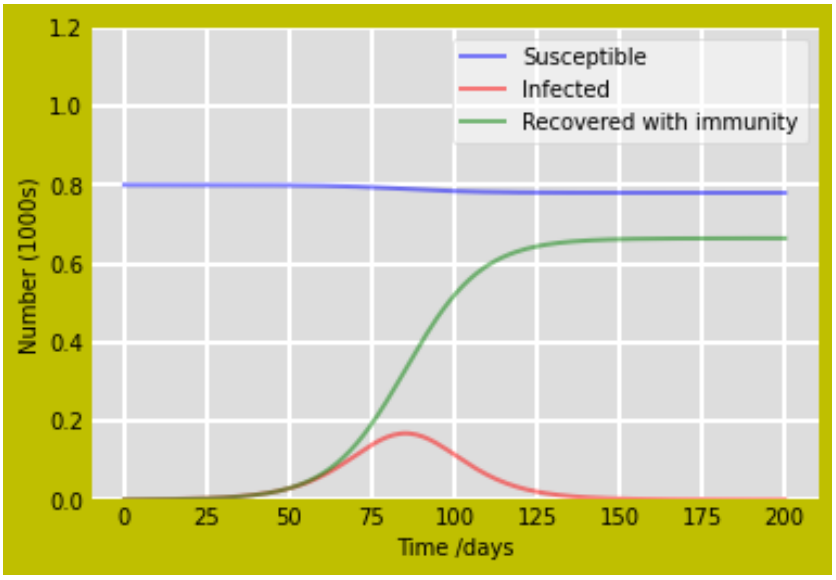
- From our data modelling to Georgia, we got maximum number of Infectives equal to *843* and the time location of maximum number of Infectives as *26th May 2020*. Whereas practical results show that maximum number of Infectives is around *1530* and the time location as *14 April 2020*. This difference in results can be accounted to the assumption that every susceptible person has equal chance of infection from the disease which in general is not correct because Every person has their own Immunity and also every person maintains different ways of Hygiene in life.

- We got duration of the epidemic is around 179 days from the day of initial infection. It is also known to us that the first infection in Georgia US came out on 3rd March 2020. Upon adding 179 days to this day we get 18th august 2020. It means that our SIR Model predicts that the COVID-19 Epidemic in Georgia US *comes to an end by 18th August 2020.*

- Also from the obtained optimized parameters we could also predict how the number of Infectives , Susceptible, Recovered varies with respect to time which in turn helps Government and many organizations like WHO in taking decisions about when to impose or loosen Lockdowns and various restrictions to its citizens.

- Based on output of this SIR model, we can state that Georgia government succeeded in implementing their rules and regulations very well because of which pandemic duration is less when compared to other US states.

Plotting the values obtained from coding, we got the following curves for Susceptible, Infectives and Recovered.



- This plot proves the well-known fact that Corona disease is indeed a *pandemic* because number of susceptible persons initially became increasing or almost constant and after some days, they started decreasing which follows the exact pandemic disease curve shape given by standard authority.

- This graph also predicts that the Duration of epidemic is approximately *179* days, measured from initial day of infection to the day until which number of infected becomes less than 1.

- This graph also tells us that *entire population is not going to be affected from this disease* because susceptible graph is not coming down it, it always remains up to the infected persons and those infected from this will gain immunity.

- It also informs us that people in Georgia,US either have strong immunity or they strictly follow government rules like lock down and social distancing because not all susceptible people are infected by virus and rate of increase in number of infectives is less when compared to other states of US.

- Their recovery rate is also high which in turn reflects the availability of good medical facilities in Georgia.

Limitations of SIR model and its predictions:

Almost all the above variations between the results obtained from our modelling of data and that of practical results can be accounted to some of the following limitations of SIR model.

- This SIR model assumes that every person is moving and has equal chance of contact with each other person among the population instead of the space or distance between different people.

- In this model we assumed that the transmission rate remains constant throughout the period of pandemic which is not happening in our real world.

- We made an assumption that there is an Exponentially distributed duration of infection which in turn assumes that a person becomes infectious immediately upon being infected, and that the probability of recovery per unit time does not depend on the time that has passed since infection going on. But both the assumptions are unrealistic.

- Moreover, this model does not cater for those infected who have been diagnosed and are in quarantine. It treats them same as those who have not been quarantined. So, both are considered to have the same transmission rate.

- This SIR model also imposes further simplifications with respect to contact patterns, as it is not designed to capture details of individual connection patterns and networks patterns, and contact is assumed to be an instantaneous event which is not possible in real life situations.

- One most important simplification is that populations are viewed as continuous entities, and individuals are not considered which is not true with real world because everyone has own characteristics and immunity power.

- In some situations, SIR model is even criticized because it is failing to produce realistic and useful results, particularly for complex disease systems like corona disease. It does not give exact realistic information to real world.

Discussion

- We can clearly observe that lock down helped in controlling the spread of COVID-19 in Georgia because lock down was lifted in Georgia from 24 th April and after that, on average 750 cases are recorded per day and early before lock down opening it is on average 380 which is reflected from our analysis of data. This is because transmissibility will be very less during lock down and social distancing.

- Our Results of this SIR model suggests that government and citizens of Georgia,US should be alert at least until 18th august 2020 as epidemic is expected to end after that.

- But strict Lockdowns are not required in Georgia,US from now as already the time of maximum number of Infectives is passed away.

- Effective Social Distancing and Containment near new Infective case areas is a good option.

We can improve our SIR Model by implementing the following extensions:

- ✓ an "Exposed" state for individuals that have contracted the disease but are not yet infectious (this is known as the ***SEIR***-model)
- ✓ a "Dead" state for individuals that passed away from the disease
- ✓ Time-dependent R_0-values that will allow us to model quarantines, lockdowns,
- ✓ Resource- and age-dependent fatality rates that will enable us to model overcrowded hospitals, populations with lots of young people

So, let us take a step forward and model the COVID-19 data of Georgia, US using ***SEIR*** *model which gives results more accurately than the SIR model*

Level-**2: SEIR** model for COVID-19 data of **Georgia**, US.

Introduction:

SEIR is a standard model for the spreading of a viral disease which divides the population into Susceptible, exposed (infected, but not yet infectious), Infectious (they can infect others), Recovered (SEIR) model.

In the *SEIR* model during an epidemic, a node can change its status from Susceptible (S) to Exposed (E) to Infected (I), and then to Recovered (R). SEIR assumes that if, during a generic iteration, a susceptible node comes into contact with an infected one, it becomes infected after an exposition period with probability beta, than it can switch to recovered with probability gamma (the only transition allowed are S→E→I→R). This is a system of non-linear Ordinary Differential Equations (ODEs), which must be solved numerically.

- S(t): number of people susceptible on day t
- E(t): number of people exposed on day t
- I(t): number of people infected on day t
- R(t): number of people recovered on day t

- γ: the proportion of infected recovering per day
- β: expected amount of people an infected person infects per day
- δ: length of incubation period
- N: total population

where N = S + E + I + R is the total population

$$\frac{dS}{dt} = -\beta \cdot I \cdot \frac{S}{N}$$
$$\frac{dE}{dt} = \beta \cdot I \cdot \frac{S}{N} - \delta \cdot E$$
$$\frac{dI}{dt} = \delta \cdot E - \gamma \cdot I$$
$$\frac{dR}{dt} = \gamma \cdot I$$

Method:

- *Data Source*:

We collected the data of Confirmed Cases and Fatalities of Georgia, US during the interval from 22-01-2020 to 18-06-2020 from the GitHub data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Centre for Systems Science and Engineering (JHU CSSE) which can be found using the below link:
https://github.com/CSSEGISandData/COVID-19

***Note**:* We divided the entire data set into 2 parts:-

- ✓ *Train Data*:-Used to train our ML model.
- ✓ *Test Data* :-Used to test results obtained after training model

- *Details of SEIR Model:*

First, we defined the equations governing the SEIR model using functions. Then using the Python Modules like scipy and sklearn we imported functions like integrate, optimize, *solve_ivp mean_squared_log_error, mean_squared_error*, and odeint for solving the Ordinary Differential Equations. Then Initialization of variables is done. Now using the pandas library, we took *two* input csv files and used the data for Fitting Procedure. Once the data is fitted, we obtained the 4 Optimized parameters *Ro ,CFR, K and L* of *HILL DECAY* function used in this SEIR model. This optimized value of Ro is then fed into initially defined SEIR function to generate the plots of S(t),E(t),I(t),R(t).

- *Fitting Procedure:*

We used the data of *Confirmed Cases* and *Fatalities* from our 2 input csv files. We started with an initial guess for all the 4 parameters and fed them into a function named *fit_model_full* which reduces the error present in our initial guess and returns other 4 parameters and keeps on going till we get the mean value of parameters based on the number iterations specified by us. The obtained optimized Ro will be used as described earlier.

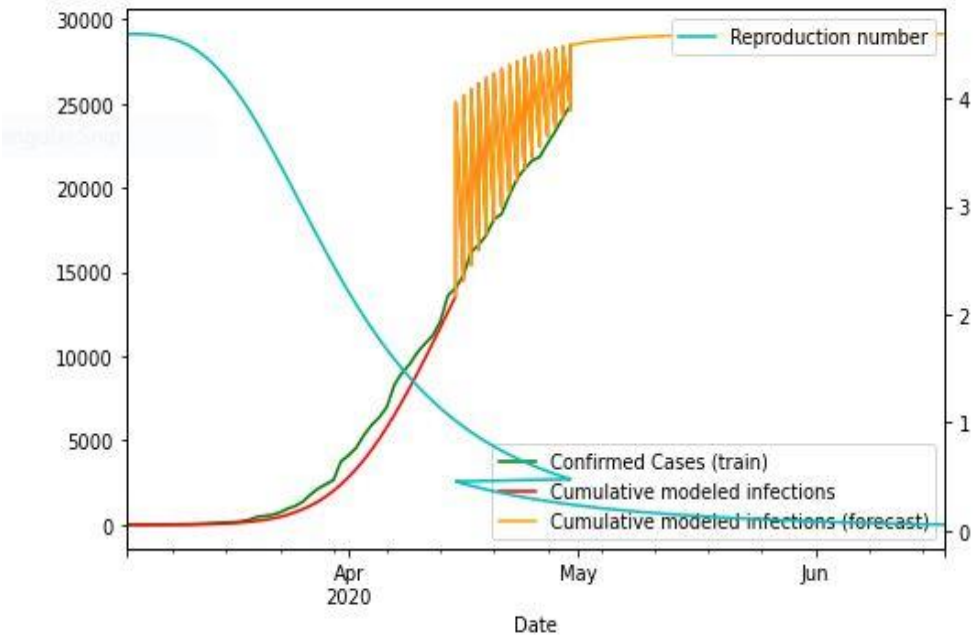
***Note**:* We usually assume these two values as constants.

- Average incubation period, T_inc = *5.2*
- Average infection period, T_inf = *2.9*

Results obtained from Coding:

- ✓ Reproduction number(R_o) = *4.59406765*
- ✓ Case Fatality Rate(CFR) = *0.004780403*
- ✓ Value of constant(K) = *2.99999753*
- ✓ Value of constant(L) = *28.25749439*

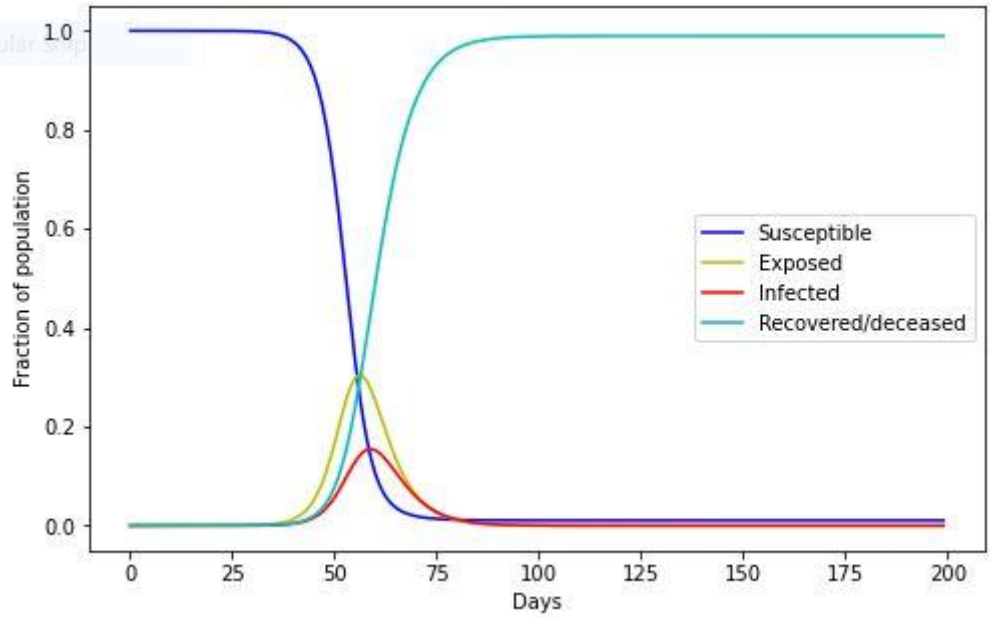
❖ *Upon fitting the data into our model, we yield the following curves which shows the time variance of an important quantity Reproduction Number(Ro)*



❖ Upon careful observation of this graph, we can see that reproductive rate in March is about 4.59 which is also matching with our result from coding and it is decreasing from then onwards.

Analysis of Model Results and its Predictions:

✓ At the end we get the following plots for Susceptible, Exposed, Infectives, Recovered as a function of time.



➢ This graph estimates that after around 100 days from initial day of infection, all the Susceptible are converted into Recovered .

➢ It predicts that almost entire population of Georgia,US will infected by the disease and at the end of the epidemic they gain *Immunity* against the disease or else they die.

➢ However, Infectives curve initially increases and reaches the maximum and then starts to decrease. So, shape of infectives we got as a result of modelling also represent the known fact that COVID-19 *is pandemic* in nature in Georgia,US because its shape matches with curve shape given by standard authority.

➢ From our Modelling we got reproductive number(Ro) = 4.59 whereas practical calculations estimate that Georgia has a reproductive number (Ro) = 4.30 up to 17th march 2020, which in turn implies that our SEIR model is *best fitted* into the data during the initial phase of infection than the later phase.

Limitations of SEIR model and its predictions:

- There are several parameters which should be fitted into the empirical data and this requires a lot of computational work.
- There is loss of biological meaning that is caused by going out to the absorbing state without going through the whole chain of incubation.
- However, our model could be applicable when sufficient empirical information of the incubation period present. For example, it might enable us to describe the SEIR model of a distribution, like bimodal, that is not expressed in a conventional way, such as Plasmodium vivax malaria in temperate regions.

Discussion

Differences between SIR and SEIR model:

- The key difference between SIR and SEIR model is that SIR is one of the simplest models of epidemiology which has three compartments as susceptible, infected, and recovered, while SEIR is a derivative of SIR which has four compartments as susceptible, exposed, infected and recovered.
- And in Addition of latency period present in SEIR model but it is not present in SIR model.
- In SIR model, total population is represented by $N = S + I + R$ while in the SEIR model, total population is represented by $N = S + E + I + R$.
- SIR is one of the simplest and basic models, and SEIR is an elaboration of it with complication form, so it has higher accuracy.

Similarities between SIR and SEIR model:

- ✓ SIR and SEIR are Both epidemiological models.
- ✓ Both models are in general applicable and applied to measles, mumps and rubella.
- ✓ SEIR model can be modified into SIR model by turning off the incubation period.

Our Experience with the SEIR model:

- ❖ If we can add some vaccination term then this SEIR model gives almost accurate results because due to vaccination, the epidemic will stop gradually. So, there is spontaneous decrement in number of infectives.
- ❖ Also, if we can modify this SEIR model in such a way that SEIR differential equation is independent of Incubation period than we can reduce its complication and we will be able to apply it easily everywhere.
- ❖ It is highly sensitive to the quality of data used to calibrate it.
- ❖ In our opinion, the SEIR model (and it's derivatives) is best suited for scenario testing using end-point assumptions rather than for generating accurate forecasts.

APPENDIX

The Code for the both the models can be found at the Kaggle Website in our notebook using the following links:-

SIR : https://www.kaggle.com/marreddisanjay/sir-model-for-covid-19-data-of-georgia-us

SEIR: https://www.kaggle.com/marreddisanjay/seir-model-for-covid-19-data-of-georgia-us