An Initial Guess Free Method for Parameter Estimation in Compartmental Models of Epidemiology

Dhruv Sharma, Guanglu Zhang, and Jonathan Cagan

Dept. of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213

ABSTRACT

To understand the spread and control of diseases, it is imperative to fit epidemiological models to real-world data. The current methods employed for parameter estimation in epidemiology often require practitioners to guess the initial value. For newly discovered diseases, especially COVID-19, the pre-requisite knowledge to make an accurate initial guess may be lacking. In this paper, an initial guess free method is developed to evaluate the optimal parameter values, that subsequently minimize the squared error of the fit. The paper also studies the resultant predictive fit of COVID-19 data, to assess the effectiveness of this method.

Keywords: COVID-19, compartmental models, parameter estimation, initial guess

1 INTRODUCTION

Epidemiology, the study and analysis of the incidence, distribution and control of diseases, utilizes many complex mathematical models to accomplish this purpose. These 'compartmental' models involve parameterized differential equations that represent the spread of diseases through a population. The current methods to estimate these parameters, and fit the models to real data, often require an initial guess. The initial guesses, often require prior intuition and knowledge. In many cases, including that of newly discovered diseases, such knowledge may not exist, and it seems unreasonable to turn towards randomness.

This is certainly true for the case of the Coronavirus (COVID-19). Therefore, the purpose of the paper is to develop a method, to estimate the numerous parameters in the compartmental models of epidemiology, that does not require any initial guess [1]. Furthermore, this method will be used to fit the models to COVID-19 data of the New York City area, and evaluate it's effectiveness.

2 EPIDEMIOLOGICAL MODELS

Mathematical models that compartmentalize the disease-prone population, often make use of a system ordinary differential equations (ODEs). There can be a number of compartments, depending of the nature of the disease. The transfer of population between compartments is modelled by these parameterized ODEs with the number of parameters, again, depending on the nature of the spread and control of a disease.

SIR Model

The **S**usceptibles-**I**nfectious-**R**emoved (SIR) model is one of the simplest epidemiological model, consisting of 3 compartments. More complex, and arguably accurate, models can be generated by adding new or dividing existing compartments. The SEIR model for instance, features the **E**xposed compartment, that can replicate the time period, wherein a person might carry and spread a certain virus, while not explicitly exhibiting symptoms [3]. This is a distinctive characteristic of the COVID-19 disease. A lot of the contemporary research surrounding COVID-19, makes use of this base compartmental model. Thus we will also seek to compare our results, to those of an SEIR model analysis [3].

Within the SIR model, the susceptible population can be infected by the virus, and thus giving rise to the infectious group. This group is now responsible for further spread the disease to the susceptibles.

Those infected will eventually, either recover or pass away, thus resulting in the removed compartment. These transfers are modelled by parameterized differential equations.

$$\frac{dS}{dt} = -\frac{\beta}{N}SI \tag{1}$$

$$\frac{dI}{dt} = \frac{\beta}{N}SI - \gamma I \tag{2}$$

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

Here, β represents the infection or transmission rate, and γ represents the inverse of average recovery/death time. We assume that the global birth rate and death rate are negligible and thus do not change the total population (N). We have that, $S+I+R=N$, at any time.

COVID-19 Data - New York

The data available for New York City [4], is in the form of daily, novel cases (see Figure 1). The data ranges from March, whence the spread of coronavirus was unhindered. Beginning April, control policies such as quarantine and social distancing, contact tracing, testing, and closure of business and school, amongst others, were implemented. We see the daily case count peak around this time, and eventually dissipate. We can consider this to be the first 'wave' of COVID-19, which has now passed through New York. Thus we can evaluate the parameter values as 'average', eliminating the dynamic nature of the problem. Since the SIR model is a time-static model [2], this selection of data set should help provide the most accurate predictive model.

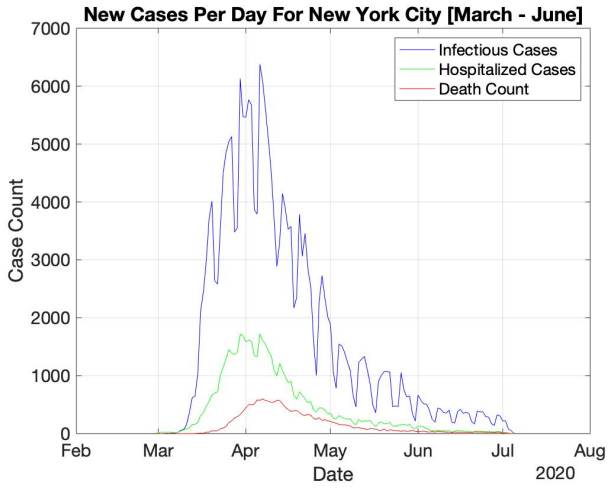

Figure 1. COVID-19 Data for New York City

We make 2 assumptions, that establish a mathematical relationships between our ODEs and the available NY data set. Since the change in susceptibles can only be brought about by the spread of the virus, S_t decreases each day, by the number of cases recorded on that particular day (C_t). This is our **first** conjecture:

$$S_t - S_{t+1} = C_t \tag{4}$$

The **second** assumption, as aforementioned, ensures that the sum of the compartments remains a known constant population value, N.

$$S_t + I_t + R_t = N \tag{5}$$

Equations (4) and (5) are crucial in fitting the model to the data set and solving for the parameter values.

3 THE INITIAL GUESS FREE METHOD

This section introduces an initial guess free method that utilizes equations (1)-(5) and the given New York data, to find the optimal parameter values that minimize the squared error of the predicted fit. This method can be generalized to other compartmental models (discussed further in section 5).

The aim of the section is to detail the important steps of this process, and explain its mathematical intuition. The lack of prior intuition about the problem in discussion, in this case the coronavirus pandemic, hinders practitioners from guessing an initial value of the parameters. We would seek to find these values from the available data itself. For distinct data points, we aim to solve for the parameter values, resulting in a solution interval. We define L_θ as the length of the solution interval, and θ_{mid} as the mid-point of the solution interval.

$$L_\theta = \theta_{max} - \theta_{min} \tag{6}$$

$$\theta_{mid} = \frac{\theta_{max} + \theta_{min}}{2} \tag{7}$$

Here, θ_{max} and θ_{min} are the maximum and minimum values of the solution interval. It is proven that when a nonlinear model meets certain conditions, the optimal estimator for each parameter that minimizes the squared error of the fit, belongs to interval $[\theta_{mid} - L_\theta, \theta_{mid} + L_\theta]$ [1]. We can iterate through the interval to find the optimal parameter values, however this can be very computationally expensive. An alternative, employed in this study, is to calculate the median of the solution interval, and supply that as the initial value to established least-squares methods of parameter estimation, e.g., Gauss-Newton or the Levenberg-Marquardt method.

The crucial step of this process is that to solve for the parameter values at each data point, it is required to solve a system of ODEs. To address this complication, we use Euler's Forward Approximation for ODEs, explored further in the following subsections.

Euler's Approximation

The forward difference Euler method, based on a truncated Taylor series expansion, allows us to express the values of a function after a certain time-step, in terms of it's initial value. As the step size, h, decreases to 0, it is proven that the forward Euler method, converges to the true solution [5]. Let m be the current data point, y, the variable whose recursive solution is required, and f the function to be evaluated. We define n, as $1/h$, or the number of discretizations of a time-step. Then, for $i \in \{1, 2, ...n\}$,

$$y_{m+i} = y_{m+(i-1)} + h * f\left(y_{m+(i-1)}, t_{m+(i-1)}\right) \tag{8}$$

For each of the differential relationship in the SIR system of ODEs, equations (1)-(3), represent the change in each compartment respectively. Using equation (8), we are able to rewrite each of these equations as a recursive approximation of their value after a time-step, $y_{m+n} \in [S_{m+n}, I_{m+n}, R_{m+n}]$, in terms of their initial value, $y_m \in [S_m, I_m, R_m]$.

$$S_{m+n} = S_m \prod_{i=1}^{n} \left(1 - h\beta I_{m+(i-1)}\right) \tag{9}$$

$$I_{m+n} = I_m \prod_{i=1}^{n} \left(1 + h\beta S_{m+(i-1)} - h\gamma\right) \tag{10}$$

$$R_{m+n} = R_m + h\gamma \sum_{i=1}^{n} \left(I_{m+(i-1)}\right) \tag{11}$$

The unknowns to be evaluated in equations (9)-(11) are: $\{\beta, \gamma, S_m, S_{m+n}, I_m, I_{m+n}, R_m, R_{m+n}\}$. However, not all unknowns must be solved for. We use the data, and equations (4), (5), to make simplifications to the aforementioned relationships.

Fitting to Data

The available data ranges over $M = 128$ days, March-June, and at each data point, m, where $m \in \{1,2,...,M\}$, the new coronavirus reported cases equal C_m. We initialise the value of $S_0 = N$, and therefore using equation (4), we can evaluate all 128 subsequent values of S_m, with respect to the data set. From Eq. (9) and (10), Eq. (4) reduces to:

$$S_m - S_m \prod_{i=1}^{n} \left(1 - h\beta I_m \prod_{j=1}^{i-1} \left(1 + h\beta S_{m+(j-1)} - h\gamma\right)\right) = C_m \tag{12}$$

Where the only unknowns are $\{\beta, \gamma, I_m\}$.

For two adjacent data points, Eq. (12) can be implemented twice, two obtain 2 equations with 4 unknowns (parameter values, β, γ remain constant). However, since we have picked adjacent points, we possess substantial computational power to express I_{m+2*n} in terms of I_m, thus reducing the number of unknowns to 3. The third equation can be obtained from Eq. (5), (10), and (11).

$$\left(I_{m+n} + R_{m+n}\right) - \left(I_m + R_m\right) = C_m \tag{13}$$

Here, I_{m+n} and R_{m+n} can be substituted by equations (10) and (11) respectively. Therefore, for a set of **2** adjacent data points (y_m, y_{m+n}), we have resolved **3** equations to be solved for **3** unknowns, namely $\{\beta, \gamma, I_m\}$. We solve for 127 parameter values each for β and γ. The histograms for these solutions ($n = 6$) are represented in Figure 2 below.

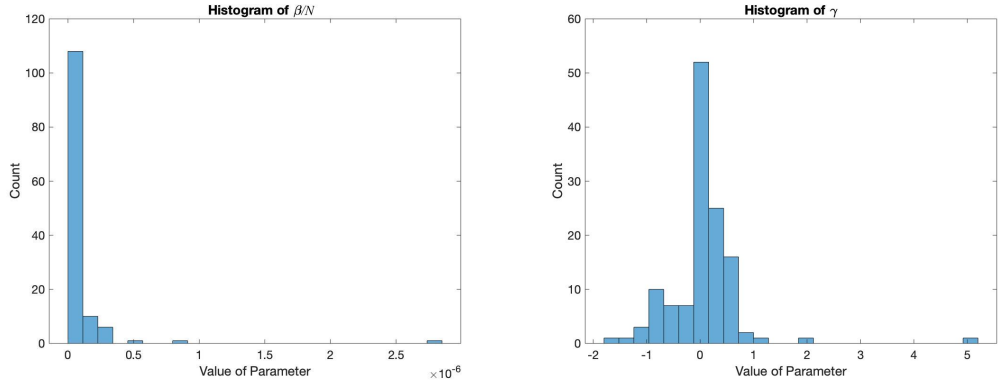

Figure 2. Histogram plots for β and γ values

It becomes increasingly computationally expensive to perform Euler's forward difference approximation, and solve for parameter solutions, as h approaches 0, i.e. for high values of n. We are interested in the median of the parameter solutions. The median values till and including $n = 6$ are furnished in Table 1 below.

Table 1. MEDIAN β AND γ VALUES FOR $n = 6$

No. of Discretizations (n)	Median β	Median γ
1	0.1114	0.0716
2	0.0113	0.0333
3	0.0145	0.0526
4	0.0096	0.0877
5	0.0097	0.0877
6	0.0097	0.0877

4 RESULTS

Using the median values from Table 1 in section 3, it noticeable that the difference in successive median values diminishes with increasing n. However, it is imperative to note that $n = 6$, which provides a step size of $h = 0.1667$ is not very close to 0, and thus not the best approximation of the true solution. This limitation of computational power is discussed further in section 5.

The median values can be used to as initial points for existing iterative least-squared optimization methods. We use the MATLAB function *lsqnonlin()* [6] to find the optimal parameter values. For $n = 6$, the obtained results are:

$$\beta_{opt} = 4.1875$$
$$\gamma_{opt} = 4.0085$$

The SIR model produced using these parameter values is displayed in Figure 3. The strength of the fit is represented by the R^2 or the coefficient of determination. We find that $R^2 = 0.7932$. It is claimed that this is reasonable fit, when further comparisons are made. However, this report accepts that there are errors in approximations, which are further discussed in section 5.

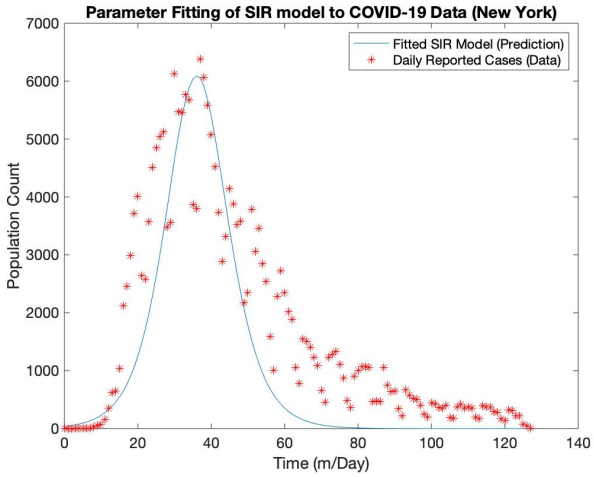

Figure 3. SIR Model with β_{opt} and γ_{opt}

The R_0 value, which possesses the physical description as the 'reproduction rate' of a pandemic, denotes the spread of the virus. For our model, this value is defined in terms of the parameters:

$$R_0 = \frac{\beta}{\gamma} \tag{14}$$

The solved value of $\mathbf{R_0 = 1.0447}$. It is important to note that R_t is a dynamic value. In other models, this value represents the average reproduction rate of a single infected person over their entire infected span, at a certain day. The value that we present in our analysis is the average of all R_t values over the 128 days time-span, indicated in the data.

We compare this value to that presented by Systrom and Vladeck in their dynamic SEIR model of New York City [3]. The R_t value on March 15 ≈ 1.44 and that on July 07 ≈ 0.94. An R value lower than 1 signifies the termination of the spread of the disease. We can explain the difference in these values owing to the control policies implemented by city's health department over the months beginning April. The average of these values is $R_{t,avg} = 1.19$, which is comparable to the R_0 presented in this report.

5 CONCLUSION AND FURTHER EXPLORATION

An initial guess free method is developed for least squares parameter estimation in compartmental models of epidemiology. This method allows practitioners to obtain optimal parameter values, without the need to guess their values prior to the analysis. Using Forward Euler's difference approximation, the method constructs a solution interval for the solved initial parameter values. Owing to restricted computational power, the median of this interval is supplied as initial parameter values to existing iterative optimization methods (here, Levenberg-Marquardt). Once found, the optimal parameter values are used to develop a predictive SIR model, to compare with other similar existing models.

As expressed in earlier sections in this report, the major limitation of the presented results is the limited computational power available. To this effect, this study expects that approximations with smaller step-sizes, and evaluations with random data points (as opposed to adjacent) would provide more accurate results. Furthermore, it is expected that other contemporary difference methods such as the Centered Euler method and the Runge-Kutta method may provide parameter values closer to the true solutions. Larger computational power and further mathematical exploration would allow this to occur. It is also expected that other models with additional compartments such as SEIR and/or SIRD are more representative of the spread of coronavirus. Therefore, extending the initial guess free method to generalize to larger compartmental models, and eventually, any system of ODEs, is the eventual aim of the study. Future research may seek to serve this purpose and develop a generalized initial guess free method for parameter estimation in nonlinear models characterized by system(s) of ODEs. Further research may also seek to evaluate the effectiveness of government policies. By considering the data until the known date of policy implementation, a model can be developed and compared to existing figures, to evaluate the effectiveness of a policy in stunting the spread of COVID-19.

REFERENCES

[1] Zhang, G., Allaire, D., and Cagan, J. (2020). An initial guess free method for least squares parameter estimation in nonlinear models.
[2] Tan, S. X.-D. and Chen, L. (2020). Real-time differential epidemic analysis and prediction for covid-19pandemic.arXiv: Populations and Evolution.
[3] Systrom, K., Vladeck, T., & Krieger, M. (2020). Rt COVID-19. Retrieved July 07, 2020, from https://rt.live/
[4] Wadhera, R. K., Wadhera, P., Gaba, P., Figueroa, J. F., Maddox, K. E. J., Yeh, R. W., Shen, C. (2020). Variation in COVID-19 hospitalizations and deaths across New York City boroughs. Jama.
[5] Zeltkevic, M. (n.d.). Forward and Backward Euler Methods. Retrieved August August 15, 2020, from https://web.mit.edu/10.001/Web/Course_Notes/Differential_Equations_Notes/node3.html
[6] Lsqnonlin. (n.d.). Retrieved August 15, 2020, from https://www.mathworks.com/help/optim/ug/lsqnonlin.html