Modeling Covid-19 with the SIR(D) Model

Peter Gartin

May 13, 2020

1 Introduction

With so much uncertainty during a pandemic like the one currently going on all over the world with Covid-19, there is much information to be gained from data representing spread of the disease. This kind of information is useful for policymakers and medical workers to understand the effects of they're decisions and help efforts to mitigate the impact, but also to the individual who just wants to know what to expect. The spread of an infectious disease is highly dependent on various factors; however, a global pandemic produces enough data to show general trends. A simple model for understanding this trend that is commonly used is the SIR-model. In this paper, I implement a trust region method to fit an SIR(D) Model to current data for the Covid-19 case counts. The reproduction number, R_0, can then be computed and compared to current value estimates such as the values ranging from $2.0 - 3.5$ from [1] and [2].

2 SIR Model

The model I will be using to track the disease spread is what is know as the Susceptible-Infected-Recovered (or SIR) Model. The model divides the population into the three bins listed in the name. The model is a system of 1d ODEs which relate the rates at which people move between the three bins. The solution to the model is the time evolution of the population of each bin. Consider first the Susceptible bin. A person will remain in this bin until they come in contact with an infected person. Let β represent the number of people that move from the S bin to the I bin per person in S per person in I per day. Then we can write,

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta S(t)I(t) \tag{1}$$

Assuming that people will not remain infected indefinitely (which would be both boring to model and unfortunate to experience), then we can define some rate at which people recover, γ, per infected person per day. Then we can write,

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I(t) \tag{2}$$

If the number of people is to be conserved, then we certainly want that if there is flow into one bin, then we want a flow of equal magnitude from the previous bin. By that logic then,

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta S(t)I(t) - \gamma I(t) \tag{3}$$

We also want to maintain the condition that,

$$S(t) + I(t) + R(t) = N \tag{4}$$

for all t where N is the total population. To check that our rate conditions satisfy this condition, we can differentiate both sides:

$$\frac{\mathrm{d}}{\mathrm{d}t}(S(t) + I(t) + R(t)) = 0$$

$$(-\beta S(t)I(t)) + (\beta S(t)I(t) - \gamma I(t)) + (\gamma I(t)) = 0$$

$$0 = 0$$

Some useful relations to consider,

$$\frac{\mathrm{d}S}{\mathrm{d}t} \Big/ \frac{\mathrm{d}I}{\mathrm{d}t} = \frac{-\beta SI}{(\beta S - \gamma)I}$$
$$\frac{\mathrm{d}S}{\mathrm{d}I} = \frac{-\beta S}{(\beta S - \gamma)}$$

We can then apply the method of separation of variables, to solve,

$$dI = (\frac{\gamma}{\beta S} - 1)dS$$
$$I(t) = I(0) + \frac{\gamma}{\beta}\ln\left(\frac{S(t)}{S(0)}\right) - S(t) + S(0)$$
$$S(t) + I(t) = S(0) + I(0) + \frac{\gamma}{\beta}\ln\left(\frac{S(t)}{S(0)}\right)$$

And using our normalization condition ($S + I + R = N$) produces the following two relations,

$$I(t) = N - R(0) + \frac{\gamma}{\beta}\ln\left(\frac{S(t)}{S(0)}\right) - S(t)$$
$$R(0) - R(t) = \frac{\gamma}{\beta}\ln\left(\frac{S(t)}{S(0)}\right)$$

If we assume that $R(0) = 0$, then

$$I(t) = N + \frac{\gamma}{\beta}\ln\left(\frac{S(t)}{S(0)}\right) - S(t) \tag{5}$$
$$S(t) = S(0)e^{-\frac{\beta}{\gamma}R(t)} \tag{6}$$

This is useful to show that the information stored in $S(0)$, β, γ, and $R(t)$ is entirely sufficient for determining the $S(t)$. This is good since apart from assuming that the entire uninfected population is susceptible, there is no way to provide data on how the susceptible population changes in time. So optimizing with respect to these other parameters will be sufficient.

For the algorithm, we want the parameters of our objective function to be roughly the same scale. β and γ are of order 1 [2], and if I want to optimize initial population counts (of order 10^8), we will need to scale the problem so that these parameters are closer in proportion. A convenient way to scale the population numbers down is to define the following densities: $s(t) = S(t)/N, i(t) = I(t)/N$ .... where N is the total population. Then $s(t)$, $i(t)$, and $r(t)$ then represent densities of the population. Our rate equations, (with the appropriate redefinition of β) then are represented as:

$$\begin{cases} \frac{\mathrm{d}s}{\mathrm{d}t} = -\beta s(t)i(t) \\ \frac{\mathrm{d}i}{\mathrm{d}t} = \beta s(t)i(t) - \gamma i(t) \\ \frac{\mathrm{d}r}{\mathrm{d}t} = \gamma i(t) \end{cases} \tag{7}$$

We can also choose to allow people to die in our model. To do this, we include a fourth bin d and an rate ominous Greek letter such as δ to represent the rate at which people move from the "infected" bin to the "dead" bid. And so we get:

$$\begin{cases} \frac{\mathrm{d}s}{\mathrm{d}t} = -\beta s(t)i(t) \\ \frac{\mathrm{d}i}{\mathrm{d}t} = \beta s(t)i(t) - \gamma i(t) - \delta i(t) \\ \frac{\mathrm{d}r}{\mathrm{d}t} = \gamma i(t) \\ \frac{\mathrm{d}d}{\mathrm{d}t} = \delta i(t) \end{cases} \tag{8}$$

To understand the severity of a disease, we define what is call the reproduction number of the disease. Consider the our rate equation for the number of infected individuals:

$$\frac{\mathrm{d}i}{\mathrm{d}t} = (\beta s(t) - \gamma - \delta)i(t) \tag{9}$$

So note that the number of infected people is increasing when $(\beta s(t) - \gamma - \delta) > 0$. So we define the reproduction number as $R_0 = \frac{\beta}{\gamma+\delta}$, or more generally, the ratio of incoming rates to out coming rates. Note then that for $s(0) \approx 1$, then $R_0 < 1$ the number of infected people is decreasing, and for $R_0 > 1$, the infected cases is increasing.

3 The Objective Function

Our model has quite a few parameters which we can optimize over. Primarily, we want to find the optimal rate relations β, γ, δ. However if we optimize these alone, we may have a difficulty fitting the data since a virus like this depends heavily on the initial conditions. Instead of just setting the initial conditions to match the data, I chose to include them in my parameters for optimization. Since we are doing a data fit, we want to minimize the error in the model from the know data, which is updated daily. Since we know that the $S(t)$ can be fully determined by the other parameters, we can do this by minimizing the error of the infected count, the recovered count, and the death count. So the objective function can be stated as:

$$f(x) = \sigma_i(x) + \sigma_r(x) + \sigma_d(x) \qquad \text{where} \qquad x = \begin{bmatrix} \beta \\ \gamma \\ \delta \\ i(0) \\ r(0) \\ d(0) \end{bmatrix} \tag{10}$$

Under the constraints that:

$$\beta, \gamma, \delta, i(0), r(0), d(0) \geq 0$$

*It may seem silly to optimize the initial dead count density since the initially dead should hopefully be 0. But I eventually end up running the algorithm for various time intervals, where the initial condition of the dead population may be greater than 0.

The data itself is from John's Hopkins Centers for Systems Science and Engineering online data base [3].

4 Algorithm

Evaluating the function requires solving the system of differential equations (7) or (8) with numpy's *odeint* function to output numerical solutions for $s(t)$, $i(t)$, $r(t)$, $d(t)$. The residual vector is then computed as outlined in Numerical Optimization by Jorge Nocedal and Stephen Wright in the chapter on Least-Squares Problems [4], calculating the difference between the model and data for each day. The residual vector is then used to compute the Jacobian with finite derivatives, and from that, an approximation of the hessian. A diagonal modification is then used to guarantee that the approximated hessian is symmetric positive definite.

For this project, I used a trust region method algorithm similar as Algorithm 4.1 in [4]. A quadratic model is needed which is built from the jacobian and hessian already calculated. The model minimization is done by scipy's minimize function. (I did also use a pre-built trust-region solver to compare, and got very similar results. Using my own algorithm allowed me to collect the trajectory in the solution space which was nice to have for plotting).

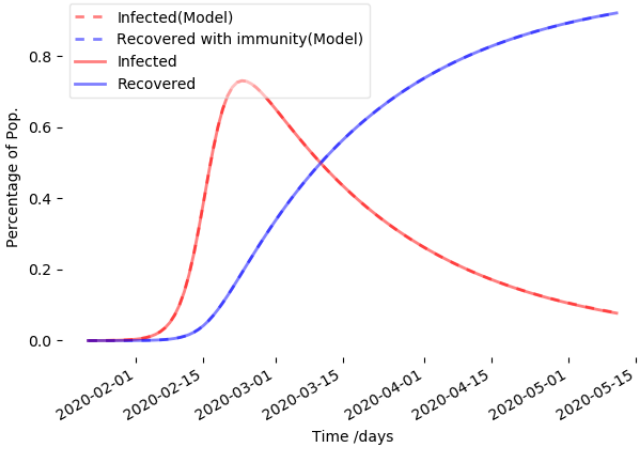


Figure 1: Result running model on SIR-model solution.

To be sure that the algorithm was working properly before running it on the real data, I wanted to test it on an exact solution of the model. The algorithm converged to x to where:
$x - x^* = [-0.00127, 0.0000132694, 0.0000780, 0.00155, 1.54717 \times 10^{-6}, -0.00155368]$ with the definition of x from section 3 in 87 iterations. The ending objective function value of 2.686×10^{-5} (See Figure 1).

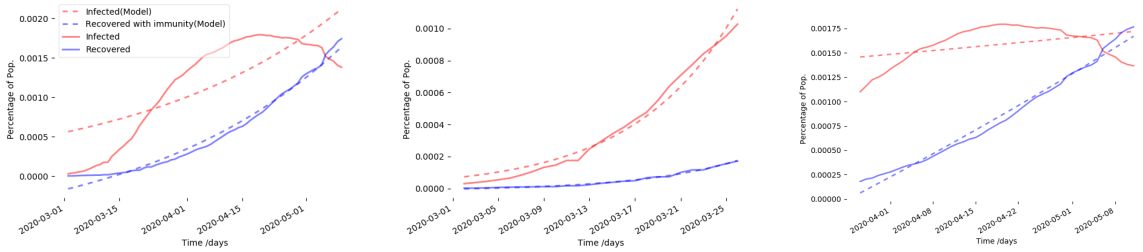
5 Results

The result were all tested on multiple countries; however, I will only discuss the results on Italy, since Italy had a very intense and uncontrolled spread early on. (Note: I feel it is important to say that though the traits of the data are convenient for modeling purposes, the numbers themselves represent the misfortunes of many, and should be taken seriously.)

The first runs were done using the SIR model (no D bin). The the output which the algorithm converged to did not appear to fit the data well (See Fig 2). The model seemed to have little problem imitating the exponential behavior of the $r(t)$ function, yet was unable to get the curvature necessary to imitate the $i(t)$ curve. Note that the only option for removing someone from the infected bin, is to move them to the recovered. So to match the provided data, another exit bin is necessary to more quickly drop the number of people in the infected bin.

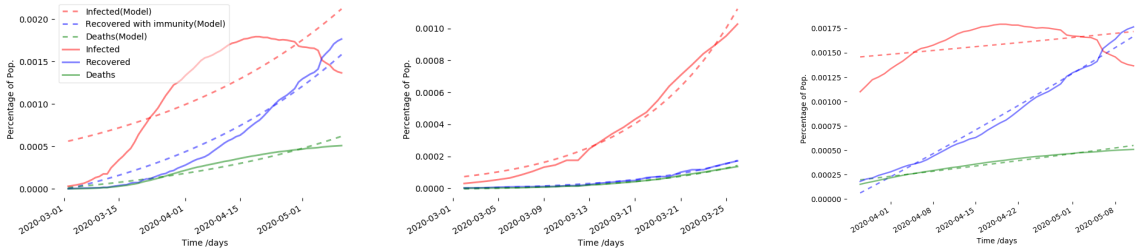
It was at this point the SIRD model, was considered to provide another exit flow from the Infected category. Switching from using (7) to (8) and adding the contribution of the error in between the model $d(t)$ and the data death count, the fit did not seem to improve much (See Fig 3). Note that it is difficult to make this statement quantitative since the objective functions are different between the SIR and SIRD models.

In both of the models it is apparent that there is not enough outflow from the I bin. To confirm that this was the primary issue with the data fit, I removed σ_d from the objective function. This caused the algorithm to make no effort fit to the death count data, which allows the algorithm to "kill-off" people to allow for a better fit to the $i(t)$ curve. These results are shown in Fig. 4. A much better fir was achieved and the R_0 value is 1.0149.



(a) (b) (c)

Figure 2: Result running the SIRD model on Italy. Plot (a) shows the result of running the algorithm for days since 3/01/20. Plot (b) shows 3/01/20 - 3/16/20. Plot (b) shows 3/01/20 - 5/11/20. The respective reproduction numbers are $R_0 = .29194$, .9619, and .1900.



(a) (b) (c)

Figure 3: Result running the SIR model on Italy. Plot (a) shows the result of running the algorithm for days since 3/01/20. Plot (b) shows 3/01/20 - 3/16/20. Plot (b) shows 3/01/20 - 5/11/20. The respective reproduction numbers are $R_0 = 1.71755$, 4.28919, and 1.13599.

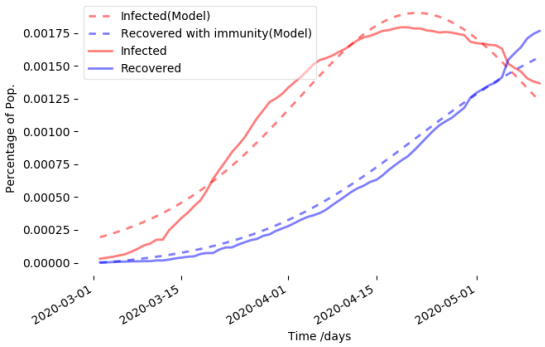


Figure 4: Result running the SIR model on Italy. $R_0 = 1.0149$.

6 Conclusion

It is clear from these results that the SIR(D) model is not sufficient for fitting the public data. When considering what could have gone wrong, and the results of Figure (4), either a more complex model is needed, or the quality of data might be insufficient. Others have used more complex models utilizing 8 bins, or even more such as work done by Giordano [1], reporting reproduction numbers ranging between 2.0 - 3.5. Adding more bins requires more data categories to fit the data to. By contrast, the SIR model is very simple, it only attempts to put people in one of three bins. Though a course grained approach like this can only tell you basic information about the virus, it also quite fool proof and relies minimally on published data.

It is also worth considering that the Model's inability to fit the data may be saying more about how well the data itself represents the disease spread. With asymptomatic reactions from the virus, and an insufficient number of testing kits, it is difficult to report accurately the number of people who have the virus. There has also been recent criticism that the number of deaths due to the virus has been over counted, counting individuals of other causes of death.

Some other improvements to the SIR(D) model would be to not assume constant values for β, γ, or δ. In the real world the rate of spread has been shown to be considerably dependent on the public response. With the introduction of the social distancing protocol, the reproduction number is likely to have been affected.

References

[1] Raffaele Bruno Patrizio Colaneri Alessandro Di Filippo Angela Di Matteo Marta Colaneri Giulia Giordano, Franco Blanchini. Modelling the covid-19 epidemic and implementation of population-wide interventions in italy. *Nature Medicine*, (https://doi.org/10.1038/s41591-020-0883-7), April 2020.

[2] Mohamed Hamidouche. Estimating covid-19 outbreak in algeria: A mathematical model to predict cumulative cases. *Bull World Health Organ*, (256065), March 2020.

[3] John's Hopkins Center for Systems Science and Engineering. Novel coronavirus (covid-19) cases. *Github*, (httpshttps://github.com/CSSEGISandData?tab=repositories), May 2020.

[4] Stephen J. Wright Jorge Nocedal. Numerical optimization. *MedRxiv*, (20039388), April 2020.