
# Investigating changes in the SEIRMD model applied to COVID-19

**Anne Cohen**

Jo Hardin, Advisor

Christina Edholm, Reader

Submitted to Scripps College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 10, 2021

**Department of Mathematics**

# Abstract

This thesis is an exploratory analysis of an SEIRMD epidemic model created by professors at the Claremont Colleges and applied to Los Angeles county COVID-19 data. The paper investigates the accuracy of predictions given different techniques for parameter optimization and decisions throughout the modeling process. The research also explores how noise affects the estimations of COVID-19 cases, deaths, and hospitalizations through simulation. Visualizations are used to compare the differential equation solutions with actual data, as well as plot the parameter values over time. The research finds that predictions are most accurate when the time interval of the predicted data is close to the data with which the parameters were fit, likely due to the variability of the COVID-19 data causing parameter values to change over time. However, while larger trends affect the predictions, random noise in the data does not appear to have a large effect on the model predictions.

# Contents

# List of Figures

# List of Tables

# Acknowledgments

I would like to thank Professor Jo Hardin for her mentorship and support throughout my senior thesis and entire college career – I feel extremely lucky to have had a professor who believes in me as much as you do.

I also want to acknowledge my thesis reader and academic advisor, Professor Christina Edholm, for her guidance and comments throughout this process, as well as Professors Maryann Hohn and Ami Radunskaya, for lending me their model to learn from and investigate.

Furthermore, I would like to express my gratitude to my family and friends for their love and encouragement, especially during this past year. In particular, thank you to my mother for being my role model, my father for his unwavering support, and Jimmy Melican for his advice and suggestions.

# Chapter 1

# Building the model: the mathematics

## 1.1 Background on COVID-19

The novel coronavirus, now known as COVID-19, was first discovered in the city of Wuhan, Hubei Province, China in December 2019. By January 20, 2020, the United States identified its first COVID-19 case in Washington State (Stone, 2021). Six days later, on January 26, Los Angeles County saw their first case, from a traveller coming from Wuhan (Staff, 2020). On March 11, the World Health Organization (WHO) declared COVID-19 a pandemic, citing their concern with the rate of transmission, severity of symptoms, and the lack of public response (WHO, 2020). Due to the novelty of COVID-19, the trajectory and effects of the disease were and continue to be difficult to predict with limited data. For this reason, a susceptible-infected-removed (SIR) model of differential equations, which relies on a theoretical understanding of the disease and minimal data, can allow us to make predictions and understand the nature of COVID-19 in order to prevent further spread.

## 1.2 Systems of differential equations

Systems of differential equations are sets of equations that describe the rates of change for different processes. They are often used in physics and engineering to describe the relationships between objects, chemistry to understand mixing properties, and within biological fields to model population growth or decay, species interactions, disease spread, and many

other biological relationships. Regardless of the application, there are three main components of a differential equation model, which are laid out by Blanchard et al. (2006). The first component is the independent variable, which often (and in the case of our model) is time ($t$). The second component is the set of dependent variables, which are a function of the independent variable, in our case $t$. The final component is the set of parameters, which are fixed values that do not change with respect to the independent variable (Blanchard et al., 2006). Depending on the application, the solutions to differential equations take on different meaning. Let's look at an example to understand how to interpret a differential equation and solution:

Blanchard et al. (2006) provide an example of a simple differential equation that models population growth. The rate of increase, in people per day, is defined by the following equation: $\frac{dP}{dt} = kP$, where $k$ is a proportionality constant and $t$ is days. We can solve for this equation with the separation of variables such that $\frac{dP}{P} = kdt$, which we can solve for $P$ to get $P(t) = e^{kt}$. This solution gives us the population value as a function of $t$. Using the value given by any particular solution, such as information about the population at $t = a$ or $P(a)$, we can solve for $k$. The population example only uses one equation to predict population growth. However, more complicated systems incorporate multiple equations to model related dependent variables. Systems of equations can be more difficult to find closed form solutions for and may require numerical solutions. The system of differential equations modeling COVID-19 that is described in this paper has 6 equations (see Equation 1.3) and is solved using computational methods in both R and MATLAB programming languages.

## 1.3 SIR models

The SIR model is a three compartment model that describes the spread of disease by rates of transition between different variables. In the most simple version of the model, the variables are susceptible people (S), infected people (I), and removed or recovered people (R). At all times, $S + I + R =$ population.

Kermack and McKendrick (1927) first describe this model in 1927 in *A Contribution to the Mathematical Theory of Epidemics* as a version of a more complicated model, where the parameters are constant rates. The biochemist and physician describe the rates as:

$$\begin{cases} \frac{dx}{dt} = -\kappa xy \\ \frac{dy}{dt} = \kappa xy - \ell y \\ \frac{dz}{dt} = \ell y \end{cases} \tag{1.1}$$

where $x, y, z$ are dependent variables, $t$ is the independent variable, and $\kappa$ and $\ell$ are constant parameters that correspond to rates. Specifically, the first equation describes the rate of change of $x$, which decreases by a constant rate $\kappa$ multiplied by the interaction of $x$ and $y$. The second equation describes the rate of change of $y$, where it increases by $\kappa xy$ and decreases by the constant rate $\ell$ times $y$.

Kernack and McKendrick's framework has been used in many models for infectious diseases since 1927, such as Cholera, SARS, Ebola, Plague, MERS, Influenza, and now COVID-19 (Yadav, 2020). In a simple example of the model, the COVID-19 model in Cooper et al. (2020) uses different notation, but essentially the same equations as in Kermack and McKendrick (1927).

A more complicated example of an SIR model is a SEIRS model described in Oluyori and Adebayo (2020), where variables included in the model are susceptible (S), exposed (E), infected (I), and recovered or removed (R). This model accounts for reinfection by reintroducing recovered and previously quarantined individuals back into the susceptible group, by adding a term to the differential equation corresponding to $\frac{dS}{dt}$, the rate of change in the susceptible variable. The equations in Oluyori and Adebayo (2020) are as follows:

$$\begin{cases} \frac{dS}{dt} = A - \mu S - \frac{\beta SI}{1+\alpha I} + (\rho + \epsilon) \\ \frac{dE}{dt} = \frac{\beta SI}{1+\alpha I} - (\gamma + \mu)E \\ \frac{dI}{dt} = \gamma E - (\sigma + \mu + \phi)I - T(I) \\ \frac{dR}{dt} = -(\rho + \epsilon + \mu + \omega)R + T(I) \end{cases} \tag{1.2}$$

where $A$ corresponds to the population growth rate, $\mu$ is the natural death rate, $\frac{\beta SI}{1+\alpha I}$ is the rate at which people become exposed from the population of susceptible individuals, $\rho$ is the recovery rate, $\epsilon$ is the rate of release from quarantine for the removed population, $\gamma$ is the rate of infection, $\phi$ is the disease death rate, $\omega$ is the rate of the removed population, and $T(I)$ is a piece-wise function that determines the rate of medical treatment response.

This SEIRS model builds upon the simple SIR model by taking into account aspects such as exposure rate, treatment response rate, and the possibility of reinfection. In particular, the exposure rate $\frac{\beta SI}{1+\alpha I}$ is based on a

model of saturation incidence developed by Esteva and Matias (2001) where the exposure rate accounts for behavioral changes based on the number of infected people and the saturation parameter corresponding to the rate at which people adhere to prevention tactics. Furthermore, Oluyori and Adebayo (2020) include the rate of treatment response, which is dependent on the number of infected individuals and therefore medical center capacities. These additions to their model allow for increased accuracy and attention to nuances within the model. While the SEIRMD model explored in this paper assumes a closed population (no new births or natural deaths), does not incorporate the saturation incidence described above, nor acknowledge cases of reinfection, it does account for medical capacities. Furthermore, the added variables – hospitalizations (M) and deaths (D) – allow for the inclusion of different death and recovery rates for those within hospitals versus the rest of the infected population in the model.

## 1.4 Our model

The model was created by Professors Christina Edholm, Maryann Hohn, and Ami Radunskaya in order to understand how the five Claremont colleges would be affected by reopening the school in the Fall of 2020 during the global COVID-19 pandemic. I will explore a version of this model that does not take into account the 5Cs specifically, but instead models the general trajectory of COVID-19 fit with LA County data. (Edholm et al., 2021)

### 1.4.1 Flow equations

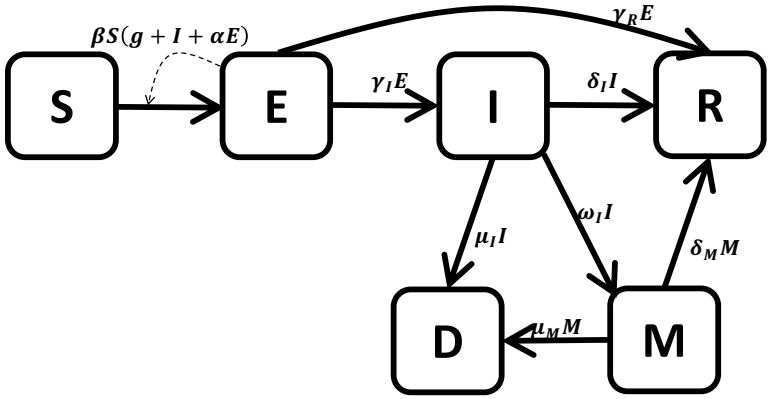


**Figure 1.1** Flow diagram between variables in SEIRMD model.

| Variable | Description |
|---|---|
| $S$ | Number of Susceptible individuals |
| $E$ | Number of Exposed individuals |
| $I$ | Number of Infected individuals |
| $R$ | Number of Recovered individuals |
| $M$ | Number of quarantined including Medical center individuals |
| $D$ | Number of Deceased individuals |

**Table 1.1** Description of the dependent variables from Figure 1.1.

The diagram in Figure 1.1 describes the rates of flow between each variable. The six variables, described in Table 1.1 are Susceptible ($S$), Exposed ($E$), Infected ($I$), Recovered ($R$), Medical center individuals ($M$), and Dead bodies ($D$). The parameters next to the arrows describe the rates of people moving from one variable to the next.

| Parameter | Initial value | Description |
|---|---|---|
| $\alpha$ | $6.7817 \times 10^{-4}$ | fraction reduction in transmission during the asymptomatic period |
| $\beta$ | $5.2257 \times 10^{-10}$ | transmission rate |
| $\delta_I$ | 0.08196 days | recovery rate for Infected individuals |
| $\delta_M$ | 0.0129 days | recovery rate for Medical center individuals |
| $g$ | 1 | greater community interaction probability. |
| $\gamma_I$ | 0.01332 days | transfer rate of Exposed to Infected individuals |
| $\gamma_R$ | 0.03964 days | transfer rate of Exposed to Recovered individuals |
| $M_{max}$ | 1967 | parameter representing the capacity of the hospital (approximately 2/3 of actual capacity) |
| $\mu_I$ | $6.3811 \times 10^{-6}$ | death rate of Infected Individuals |
| $\mu_M$ | 0.01964 days | death rate of Medical center Individuals |
| $\omega_I$ | 0.03413 | transfer rate of Infected individuals to Medical center |

**Table 1.2** Initial parameter values fit to LA County Data by Edholm et al., April 20-May 16, 2020.

The parameters in Table 1.2 have been approximated and fit to data from April 20 to May 16, 2020 by Edholm et al. (2021). These values are used as starting points to determine the optimal parameters to fit more recent data. I delve further into this process in Chapter 2.

The following bullet points will detail the parameters featured in the flow diagram (Figure 1.1) and the relationships between each of the variables:

- The transition rate between **Susceptible (S) and Exposed (E)** is $\beta S(g + I + \alpha E)$ people per day. This equation is the sum of the greater community interaction probability ($g$), the number of infected people ($I$), and the number of exposed people ($E$) scaled by their reduced transmission rate from the asymptomatic period ($\alpha$), all multiplied by the number of susceptible individuals ($S$) and the general transmission rate ($\beta$).

- For **Exposed (E) to Infected (I)**, the rate is $\gamma_I E$ people per day, which is the product of the number of exposed people ($E$) and the transfer rate from exposed to infected ($\gamma_I$).

- Similarly, the rate for **Exposed (E) to Recovered (R)** is $\gamma_R E$ per day, which is the product of the number of exposed people ($E$) and the transfer rate of exposed to recovered individuals ($\gamma_R$).

- Between **Infected (I) and Recovered (R)** individuals, the transition rate is $\delta_I I$ people per day, which is the product of the number of infected people ($I$) and the recovery rate for infected individuals ($\delta_I$).
- The transition rate between **Infected (I) and Medical center (M)** individuals is $\omega_I I$ people per day, which is the product of the number of people infected ($I$) and the transfer rate of infected individuals to medical centers ($\omega_I$).
- The mortality rate, or transition rate between **Infected (I) and Dead bodies (D)** is $\mu_I I$ people per day, which is the product of the number of infected people ($I$) and the death rate of infected individuals ($\mu_I$).
- The rate from **Medical center (M) to Recovered (R)** individuals is $\delta_M M$ people per day, the product of the number of people in medical centers ($M$) and the recovery rate for individuals in medical centers ($\delta_M$).
- Finally, the death rate in medical centers, or the rate of transition from **Medical centers (M) to Deceased (D)** individuals per day, is $\mu_M M$. This rate is the product of the number of people in medical facilities and the death rate of individuals in those medical centers ($\mu_M$).

### 1.4.2 Differential equations

The relationships defined in Figure 1.1 lead to the construction of a series of equations describing the rates of change within each state. The conservation principle defines the rate of change $\frac{dY}{dt}$ for a state $Y$ as the increase in $Y$ ($Y_+$) subtracted by the decrease in Y ($Y_-$). With this principle, we can create a system of differential equations describing the rates of change within each state.

$$\begin{cases} \frac{dS}{dt} = -\beta S(g + I + \alpha E) \\ \frac{dE}{dt} = \beta S(g + I + \alpha E) - (\gamma_I + \gamma_R)E \\ \frac{dI}{dt} = \gamma_I E - (\delta_I + \mu_I)I - \omega_I e^{-(M/M_{max})^2} I \\ \frac{dR}{dt} = \delta_I I + \delta_M M + \gamma_R E \\ \frac{dM}{dt} = \omega_I e^{-(M/M_{max})^2} I - (\delta_M + \mu_M)M \\ \frac{dD}{dt} = \mu_I I + \mu_M M \end{cases} \tag{1.3}$$

The following bullet points explain the equations that describe the change in dependent variables over time in Equation 1.3:

- The change in number of **susceptible** people ($\frac{dS}{dt}$) over time is equal to the negative of the number of new people becoming exposed over time ($\beta S(g + I + \alpha E)$). This is a variation of the exposure rate in Equation 1.2, which was $\frac{\beta SI}{1+\alpha I}$.

- The change in number of **exposed** people ($\frac{dE}{dt}$) is equal to the number of new people becoming exposed ($\beta S(g + I + \alpha E)$) subtracted by the summation of the number of people becoming infected from exposure and the number of exposed people who recover over time ($E(\gamma_I + \gamma_R)$). Building off of Equation 1.2, the only differences in Equation 1.3 for $\frac{dE}{dt}$ are that we do not include the decrease due to natural deaths, but we do account for those who recover directly after exposure, instead of becoming infected ($\gamma_R E$).

- The change in number of **infected** people ($\frac{dI}{dt}$) is equal to the number of exposed people becoming infected ($\gamma_I E$) subtracted by the summation of recovered and dead people who were previously infected and infected people who are moved to medical facilities over time ($I(\delta_I + \mu_I + \omega_I e^{-(M/M_{max})^2})$). Aside from the omission of natural deaths, the only difference between Equation 1.3 and 1.2 is the term that describes the medical facility capabilities. In Equation 1.3, the term is $\omega_I e^{(M/Mmax)^2} I$ while for Equation 1.2, the term is a piece-wise function.

- The change in number of **recovered** individuals is equal to the sum of the recovery rate from infected, exposed and those in medical facilities, which is a term that differs from Equation 1.2, as they use T(I) again, but we use $\delta_M M$.

- The change in number of people in **medical centers** ($\frac{dM}{dt}$) is equal to the number of new medical center patients over time ($\omega_I e^{-(M/M_{max})^2} I$) subtracted by the summation of the number of dead ($\mu_M$) and recovered ($\delta_M$) hospital patients over time. Both M and D are not included in the Oluyori and Adebayo (2020) model, but they are helpful because we have Los Angeles county data for these variables.

- Finally, the change in number of **deceased** people ($\frac{dD}{dt}$) is equal to the summation of people who were infected who died ($\mu_I I$) and dead people who were in medical facilities over time ($\mu_M M$).

These equations can be solved using computational methods, such as `deSolve` in R, to find counts of people in each state after a specified amount

of time.

## 1.5 Paper outline

In Chapter 1, I laid out the basic structure of the (Edholm et al., 2021) model. The following chapters will explore the SERIMD model in further detail. In Chapter 2, I will describe the process of the running the model, from the data selection and assumptions made about the initial values, to the parameter optimization and the ODE solver 'lsode'. In Chapter 3, I will describe the analyses that I conducted, including the creation of a Shiny app, exploration of changing parameter and data variability over time, and comparisons of accuracy with Marin county. Finally, Chapter 4 concludes the paper with a review of the main takeaways from my research into the Edholm et al. (2021) SEIRMD model and future directions.

# Chapter 2

# Running the model

## 2.1 The data

The majority of data used to fit and assess the Edholm et al. (2021) model comes from the Los Angeles Times Data and Graphics Department's GitHub repository, which provides daily-updated COVID-19 data (Los Angeles Times Staff, 2021). The GitHub repository contains state-wide data from the California Department of Public Health, as well as data compiled by the Times from all 58 counties' public health agencies and other local organizations. This data primarily counts California residents, excluding visitors and non-residents who test positive in California, beginning with the earliest case data reported in January 2020 and hospitalization data starting in March 2020.

### 2.1.1 Cases and deaths

The `latimes-county-totals.csv` file contains columns specifying the date, county, fips (a unique code for each county), number of confirmed cases, number of deaths, number of new confirmed cases since the previous day, and number of new deaths since the previous day (Los Angeles Times Staff, 2021). Each row is a vector of data corresponding to a unique date and county. The data begins on January 26, 2020 and is updated daily, with one row for each of the 58 counties in California. Counties where cases are not reported until March or April show `NA` values until the first cases appear. The Edholm et al. COVID-19 model was intended to be used for predicting Los Angeles county cases and other outcomes, so I filtered the data to only contain rows with Los Angeles county COVID-19 data. For example, Table

2.1 is a snapshot of seven rows of data for Los Angeles county between March 29 and April 4, 2020.

| Date | County | FIPS | Confirmed cases | Deaths | New confirmed cases | New deaths |
|---|---|---|---|---|---|---|
| 2020-03-29 | Los Angeles | 37 | 2147 | 37 | 329 | 5 |
| 2020-03-30 | Los Angeles | 37 | 2505 | 44 | 358 | 7 |
| 2020-03-31 | Los Angeles | 37 | 3037 | 54 | 532 | 10 |
| 2020-04-01 | Los Angeles | 37 | 3528 | 66 | 491 | 12 |
| 2020-04-02 | Los Angeles | 37 | 4071 | 80 | 543 | 14 |
| 2020-04-03 | Los Angeles | 37 | 4605 | 93 | 534 | 13 |
| 2020-04-04 | Los Angeles | 37 | 5325 | 119 | 720 | 26 |

**Table 2.1** Los Angeles Times case and death data between March 29 and April 4, 2020 in Los Angeles county.

### 2.1.2 Hospitalizations

The `cdph-hospital-patient-county-totals.csv` file is provided by the California Department of Public Health and contains county-wide data about hospitalized patients (Los Angeles Times Staff, 2021). The columns specify the date, county, Federal Information Processing Standards (FIPS), number of positive patients (including in ICUs), number of suspected patients (including in ICUs), number of positive patients specifically in the ICU, and the number of available beds specifically in the ICU. Similar to the case data in Section 2.1.1, each row contains data corresponding to a day and one of the 58 counties. Thus, there are 58 rows of data per day, beginning March 29, 2020 and continually updating as the pandemic continues to fill hospitals with COVID-19 patients. As seen in the previous section, I filtered the data to only include Los Angeles county data. I also created a new column called `Hpatient` that adds the number of positive patients

and the number of suspected patients to give the total number of COVID-19 hospitalized patients (including those in ICUs) in Los Angeles county per day. Table 2.2 captures the first seven rows of Los Angeles county hospital data, from March 29 to April 4, 2020.

| Date | County | FIPS | Positive patients | Suspected patients | ICU positive patients | ICU suspected patients | ICU available beds | Hpatient |
|---|---|---|---|---|---|---|---|---|
| 2020-03-29 | Los Angeles | 37 | 489 | 1132 | 191 | 182 | 345 | 1621 |
| 2020-03-30 | Los Angeles | 37 | 601 | 1277 | 245 | 244 | 456 | 1878 |
| 2020-03-31 | Los Angeles | 37 | 713 | 1239 | 315 | 239 | 445 | 1952 |
| 2020-04-01 | Los Angeles | 37 | 739 | 1332 | 335 | 220 | 492 | 2071 |
| 2020-04-02 | Los Angeles | 37 | 818 | 1270 | 346 | 193 | 488 | 2088 |
| 2020-04-03 | Los Angeles | 37 | 962 | 1239 | 422 | 209 | 473 | 2201 |
| 2020-04-04 | Los Angeles | 37 | 1007 | 1190 | 449 | 181 | 537 | 2197 |

**Table 2.2** California Department of Public Health hospitalization data between March 29 and April 4, 2020 in Los Angeles county. Note that the final column `Hpatient` is created by adding the positive and suspected patients.

### 2.1.3 Initial values

In order to find values for the dependent variables ($S, E, I, R, M,$ and $D$) in the Edholm et al. (2021) model at different times $t$, we must begin with initial values for these six variables. While these values rely on data, they are also based on assumptions and decisions made by Edholm et al. (2021). The following bullet points explain how the initial values of the dependent variables listed above are calculated and the assumptions on which they rely:

- For the **susceptible** state, the initial value $S_0$ is the total population in Los Angeles county ($9,651,332$) subtracted by the summation of the initial values for all five other states:

  $$S_0 = 9651332 - (E_0 + I_0 + R_0 + M_0 + D_0)$$

  The population data is given by the Los Angeles County Department of Public Health's 2018 census data (Department, 2021). The rest of the data—COVID-19 cases, hospitalizations, and death data used for remaining states—comes from the Los Angeles Times GitHub coronavirus repository (Los Angeles Times Staff, 2021).

- The initial value $E_0$ for the **exposed** state is equal to the summation of new cases in the ten days including and leading up to the model's start date multiplied by twenty:

$$E_0 = 20 \times \sum_{t=-9}^{0} C_t$$

where $t = 0$ is the first day of data with which the model is fit and $C_i$ is the number of new cases on the $i^{th}$ day. For example, $C_{-1}$ would be one day before the model start date. This calculation assumes that there are twenty times more people exposed than there are confirmed cases, implying that each person infects an average of 20 others.

- Thus it follows that the initial value of **infected** individuals $I_0$ is equal to the summation of new cases in the ten days including and leading up to the start date of the model:

$$I_0 = \sum_{t=-9}^{0} C_t$$

where $C_t$ follows the same notation as the previous bullet point. Using this calculation as the initial value assumes that no cases prior to ten days before $t = 0$ would still be within the infected group. In other words, the average duration of infection is ten days; nine days after testing positive, an individual is still infected, but ten days after a positive test they are either recovered, in a medical facility, or deceased.

- The initial value of **recovered** individuals $R_0$ is equal to the cumulative case count before the ten days prior to the model's start date, multiplied by twenty and 0.95:

$$R_0 = .95 \times 20 \times T_{-10}$$

where $T_{-10}$ corresponds to the cumulative number of cases 10 days prior to the model start at $t = 0$. This assumes that 95 percent of those who were exposed, from the beginning of the pandemic up until ten days before the model's start, have recovered.

- For the individuals in **medical centers**, the initial value is equal to the number of people in hospitals for suspected and confirmed COVID-19 cases on the model's start date:

$$M_0 = \text{Hpatient}_0$$

This assumes that all suspected COVID-19 cases in the hospital, where patients present symptoms and are waiting on COVID-19 tests, will have positive results.

- Similarly, the initial value for the **deceased** state is equal to the total number of people who have died from COVID-19 by the model's start date.

$$D_0 = \text{Total deaths}_0$$

### 2.1.4 Parameter Optimization

The parameters described in Table 1.2 were fit by Edholm at al. based on calculations and assumptions. Using these values as a baseline, I implemented my own optimization method with the help of the `optim` function from the R `stats` package. The `optim` function is a generic optimization function that, by default, attempts to minimize a value output by a function of the user's choice. The optimization process loops through the differential equation which produces a value for the specified optimization function with different parameter values until the optimization criterion does not decrease by a certain tolerance level, which by default is 1.490116e-08, based on the square-root of the smallest number that my computer identifies for 1+x ≠ 1.

The original optimization criterion was the summation of: the sum of squares error (SSE) between the predicted and actual case numbers, the SSE of the predicted and actual death numbers, and the SSE of the predicted and actual hospitalizations. The statistic can be written as:

$$\theta_1 = \sum_{t=1}^{n}((\hat{C}_t - C_t)^2 + (\hat{D}_t - D_t)^2 + (\hat{H}_t - H_t)^2) \tag{2.1}$$

where the $\hat{hat}$ marks the predicted values, $C$ denotes cases, $D$ is deaths, and $H$ represents hospitalizations.

This function was input into the `optim` function, with the original parameters and the data and initial states, as described in the previous sections. The function `optim` loops through the model a number of times that is specified by `maxit`, running the optimization function and determining the combination of parameters that outputs the smallest sum of SSE values.

However, Equation 2.1 does not take into account the differences in magnitude between the numbers of cases, deaths, and hospitalizations.

Most days, the number of total cases is over 40 times the number of total deaths and over 100 times the number of hospitalized individuals. Naturally, this means that the sum of squared errors for the total predicted versus actual cases will be larger than that of the death or hospital data, simply because of the difference in magnitude between the data. Thus, because the optimization process minimizes the raw sum of SSE values, the optimization will have the largest effect on the cumulative case predictions, providing the best predictions for the parameters associated with rates of infection and potentially neglecting the parameters for death and hospitalization rates. In Figure 2.1, I fit the parameters for each month separately, modeled the data for the respective month, and then concatenated the estimated data to get the full time series. This figure depicts the differential performance of the optimization criterion when the results are split up into modeled and actual values for each of the three data sets – cumulative cases, total deaths, and number hospitalized. The (Edholm et al., 2021) model with the $\theta_1$ optimization criterion described by 2.1 predicts cumulative cases much better than hospitalizations and deaths.

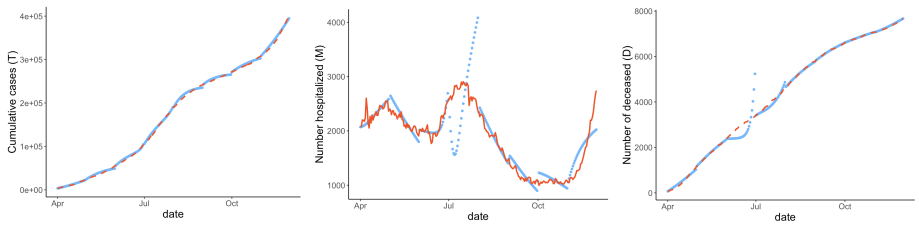


**Figure 2.1** Predicted (blue) and actual (red) COVID-19 cases, hospitalizations, and deaths using Equation 2.1 to optimize parameters.

After discovering this issue, I decided to change the optimization criterion. Adjusting onto Equation 2.1, I scaled each of the three summations by the mean of their actual values to get an error value that is not skewed by a difference in magnitude.

$$\theta_2 = \sum_{t=1}^{n} \frac{(\hat{C}_t - C_t)^2}{mean(C_t)} + \frac{(\hat{D}_t - D_t)^2}{mean(D_t)} + \frac{(\hat{H}_t - H_t)^2}{mean(H_t)} \tag{2.2}$$

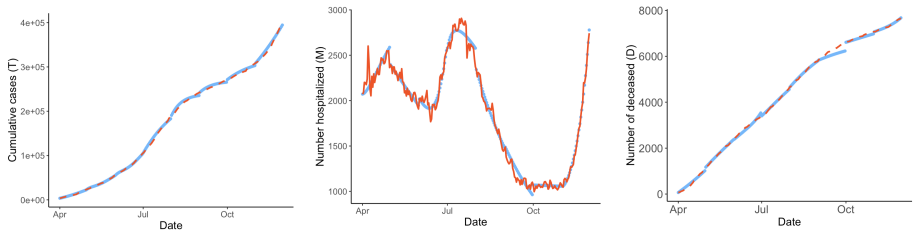


**Figure 2.2** Predicted (blue) and actual (red) COVID-19 cases, hospitalizations, and deaths using Equation 2.2 to optimize parameters.

Figure 2.2 shows that the new optimization statistic provides more equally accurate predictions across the three variables, while not substantially compromising the accuracy of the cumulative case predictions. Equation 2.2 seems to improve the model predictions for all three of the data sets (cases, deaths, and hospitalizations), thus it will be used for all further analyses.

Regardless of the criteria, the parameter optimization relies on COVID-19 case, death, and hospital data, which is ever-changing over time. Therefore, the optimal parameter values change depending on the time-frame used to optimize the model and there does not exist a single best value for each of the parameters. In Chapter 3, I will discuss the importance of this variability and what it says about the model's stability.

## 2.2 Solving the equations with `lsode` in R

The final step of running the model is solving it. To do so, I use a function in R called `ode`, which is a wrapper for different ODE solvers, meaning it calls other functions to solve the specified system of differential equations. After some exploration, I chose the `lsode` solver for its processing speed. This ODE solver which was originally developed in Fortran by Alan C. Hindmarsh and adapted to R. The function solves both stiff and non-stiff systems of differential equations (Hindmarsh). The default is to compute the ODE as a stiff system, meaning the solution is unstable unless the step size is very small (Xu). `lsode` solves stiff systems of equations using a Jacobian matrix. In Radhakrishnan and Hindmarsh (1993), the authors describe how the function solves the ODEs using the initial values. The solver finds the solution for each time point by determining an initial estimate and then improving on it until it no longer changes (Radhakrishnan and Hindmarsh, 1993). The output (shown below in Table 2.3) of the `lsode` function, and therefore, the wrapper `ode`, is a matrix with columns as the dependent

variables or states, rows as each time point $t$ (days, in our case), and cells as predicted values for each variable at each time point $t$.

| time | S | E | I | R | M | D | T |
|---|---|---|---|---|---|---|---|
| 1 | 9572401 | 62340.00 | 3117.000 | 7809.00 | 2071.000 | 66.0000 | 3528.000 |
| 2 | 9568638 | 63422.94 | 3261.243 | 10311.20 | 2075.331 | 95.4040 | 4037.325 |
| 3 | 9564804 | 64529.59 | 3398.648 | 12863.48 | 2082.994 | 124.8953 | 4555.517 |
| 4 | 9560900 | 65659.69 | 3530.368 | 15466.20 | 2093.441 | 154.5172 | 5082.769 |
| 5 | 9556923 | 66813.10 | 3657.409 | 18119.79 | 2106.193 | 184.3057 | 5619.269 |
| 6 | 9552874 | 67989.71 | 3780.661 | 20824.75 | 2120.831 | 214.2906 | 6165.205 |
| 7 | 9548750 | 69189.49 | 3900.886 | 23581.64 | 2137.000 | 244.4958 | 6720.766 |

**Table 2.3** First 7 rows of predictions for April 2020 where parameters are fit with April 2020 data

# Chapter 3

# Assessing the model

## 3.1 Parameter change over time

Theoretically, most of the parameters listed in Table 1.2 should be constant values. However, as mentioned in Chapter 2, because parameter optimization minimizes the error between predictions and real COVID-19 data, and COVID-19 data has been highly variable over the past year, the parameter values also vary. The plot in Figure 3.1 below depicts these changes. For each month from April to December 2020, I optimized the parameters using that month's data. Then I plotted the anomalies of the parameters, calculated by subtracting the mean of the values from each actual value to get a standardized version of the value that highlights any differences in the parameter values relative to the average value for that parameter. A value above zero indicates a higher value than the average parameter value while a value below zero indicates a lower value.

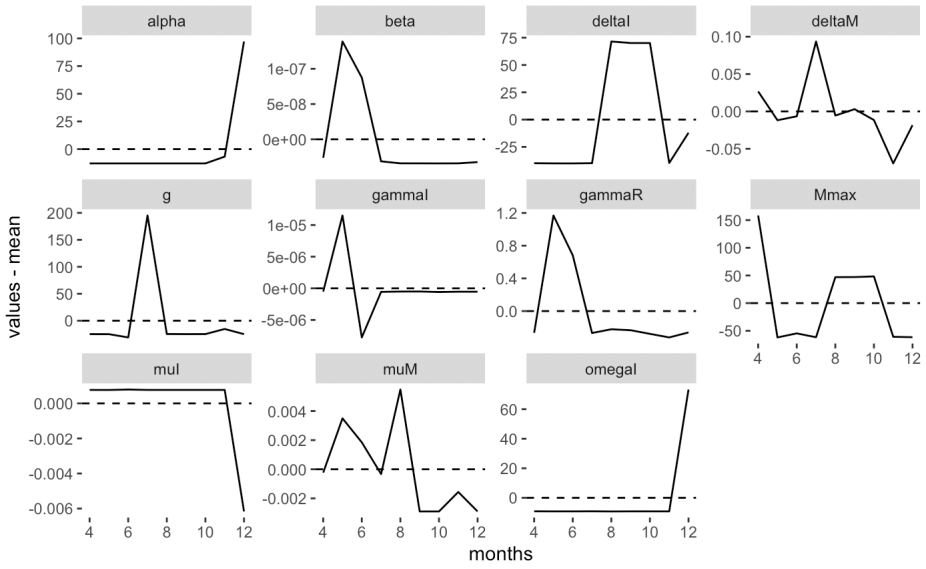


**Figure 3.1** Anomaly change in optimal parameters for each month's data

These plots allow us to see that there is variability in all of the parameters over time and there is not one single time interval that appears to cause the changes. There are a few patterns to notice:

- Parameters $\gamma_I$, the transfer rate of exposed to infected individuals, $\beta$, the transmission rate, and $\gamma_R$, the transfer rate from exposed to recovered, all have significant increase in May. These parameters are all associated with the first few dependent variables and the transitions from Susceptible to Infected or Recovered. In the context of Los Angeles county, the Department of Health created a five-stage plan and transitioned from Stage 1 (stay-at-home orders) to Stage 2 (limited reopening) at the beginning of May (Department of Public Health). The parameter change could suggest that this transition caused an uptick in transmission and exposure to COVID-19.

- Parameters $\delta_I$, recover rate for infected individuals, and $M_{max}$, the capacity of hospitals, have a similar increase, plateau, and then decrease from August through October. In fact, most of the parameters show plateaus during this time interval. This implies that the data (and therefore, model) was relatively constant for a few months in fall 2020.

- Parameters $\alpha$, the fraction reduction in transmission during the asymptomatic period, $\mu_I$, the death rate of infected individuals, and $\omega_I$, the transfer rate of infected individuals to medical centers all have relatively extreme changes in their values in December. In December 2020, COVID-19 cases, deaths, and hospitalizations skyrocketed, which likely explains the sharp changes in parameter value. In fact, I am surprised that more parameters did not drastically change for these later months.

## 3.2 Visualizing model accuracy

### 3.2.1 Shiny

Because the parameters vary depending on the data provided for parameter optimization, I wanted to try as many different analyses and visualizations as possible. Thus, I decided to create an interactive Shiny app that allows the user to determine the months used in the parameter estimation and plot the predictions from this model applied to the following month. For example, if I selected 7 (July) on the slider, the app would optimize the parameters to fit the COVID-19 data from April through July, predicting and plotting the values for August. The Shiny app is a helpful exploratory tool because it allows for repeated runs of the same analysis with certain tweaks in the parameter estimation code such as the months to fit and selection of variables to plot.

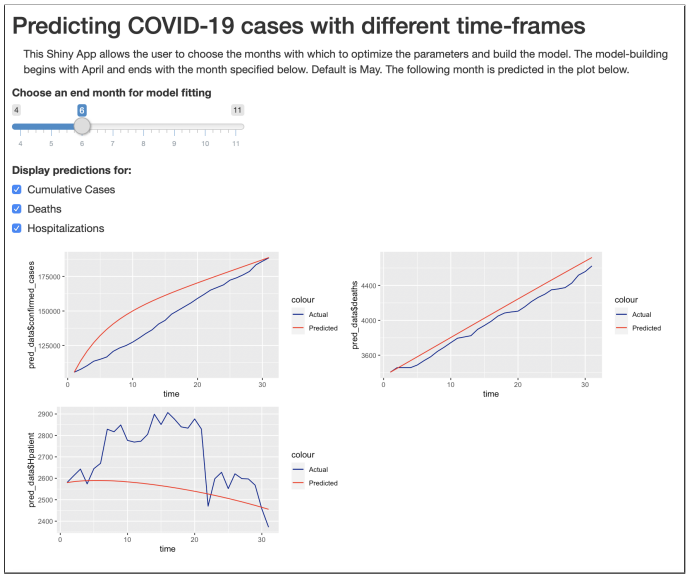


**Figure 3.2** Exploratory Shiny app for visualizing model-building and parameter optimization with different combinations of months. Please email me if you would like access to the app at anniecohen041@gmail.com.

Through the Shiny app, I discovered that optimizing parameters based on longer time frames (from April through November and predicting for December, for example) does not make for more accurate predictions. This may seem counter-intuitive, but the COVID-19 pandemic in Los Angeles county has experienced multiple surges of the virus and endured many phases of stay-at-home restrictions and re-opening policies. These changes and decisions may cause variation in the data and make predictions less accurate.

### 3.2.2 Predicting months out

Based on my conclusions from the Shiny app, I decided to use only one month at a time to optimize the parameters and build the model. First, I used the `optim` function (as described in Chapter 2) to estimate these parameters based on each month's data separately. I ended up with 8 vectors of parameters, one for each month April through November. Next, I used each month's parameters to run the model and predict outcomes for the same month, the following month, and the month after. I combined the predictions for each of these together with the other months of the same

prediction type (same month, month after, 2 months after) to visualize the prediction accuracy over the whole time series.

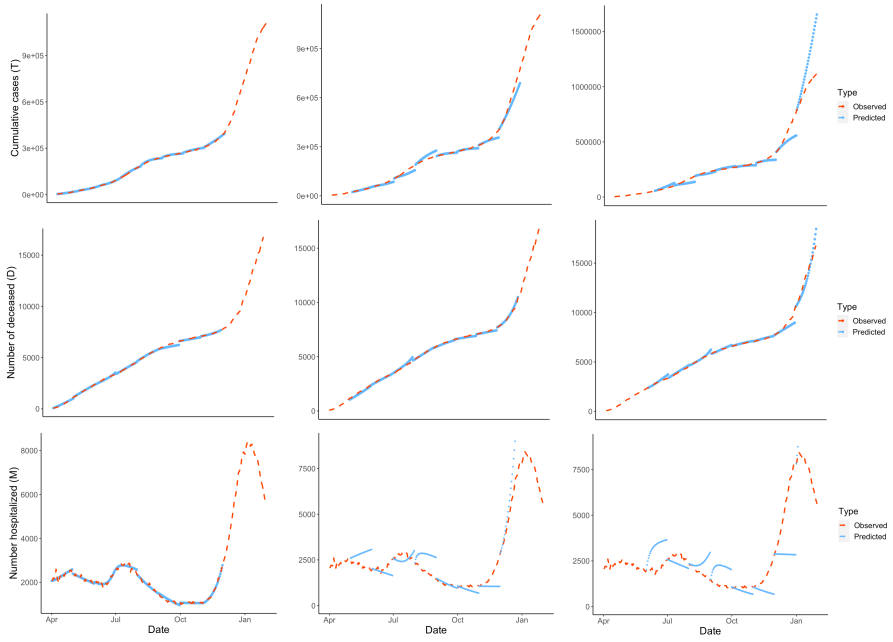


**Figure 3.3** Actual and predicted cumulative cases, deaths, and hospitalizations for the same month that model is fit, one month out, and two months out

Figure 3.3 displays plots of predictions, with the x-axes displaying the month predicted. The first row of plots are the number of cumulative cases, the second are deaths, and the third are hospitalizations. The first column depicts the predictions for the month with which the parameters are fit, the second are predictions for the month after, and the third are two months after.

In Figure 3.3, the prediction accuracy clearly gets worse the further out the predictions are. However, predictions for cumulative cases and deaths are much more accurate than that of hospitalizations. All of the predicted month segments begin on the observed line because their initial value matches that of the data, but as the month goes on, especially for the hospitalization predictions one and two months out, the predictions stray from the actual data. Both deaths and cases are monotonic functions because they are cumulative, while hospitalizations are individual daily measurements. This

allows for more variability in the number of hospitalized individuals at a given time than the number of cumulative cases at a given time, therefore it makes sense that the estimated numbers of hospitalizations are less accurate when predicting for further out months.

## 3.3 Measuring model accuracy

In order to quantify the accuracy of the predictions, I calculated the root mean squared error (RMSE) of the predictions for the same month as the model fitting, one month out, and two months out. Figure 3.4 displays the RMSE values, with the x-axis as the month for which the data was predicted, the y-axis as the RMSE value with a log scale, and each line as a different number of months prior that the parameters were fit: 0, 1, or 2. For example, this means that the point on each of the red, green, and blue lines where $x$ = Jul correspond to July predictions from parameters fit using July, June, and May data, respectively.

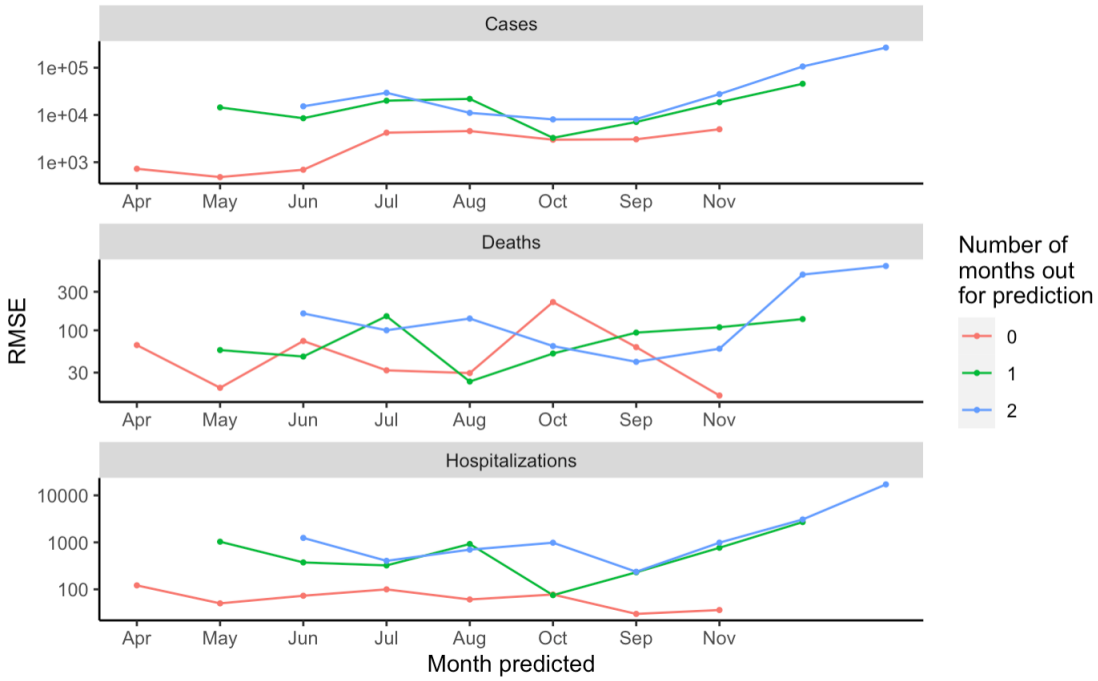


**Figure 3.4** Monthly fit model prediction accuracy (measured by RMSE) for same month, one month out, and two months out

For the most part, the plots in 3.4 demonstrate what we saw in the last section: farther out predictions tend to be less accurate. If we weigh cases, deaths, and hospitalization RMSEs equally, we run into a similar problem as in the Chapter 2 parameter optimization challenge, where the difference in magnitude between numbers of cases, deaths, and hospitalizations impacts our conclusions. The RMSE values for deaths are much lower than that of cases or hospitalizations. Also interestingly, we see death predictions that are one or two months out do not appear to be much worse than the estimations from the parameter-fitted month. Another pattern we can see is that after beginning in October, most of the RMSE values increase, inflating the difference between results from data for different numbers of months out. We can interpret this to mean that sharp changes (specifically, increases) in cases, deaths, and hospitalizations magnify the inaccuracies in the predictions.

### 3.3.1 Marin county accuracy

To understand the effect of the parameter values on the predictions and model accuracy, I ran the model with Los Angeles county parameters to predict Marin county cases, deaths, and hospitalizations. The only information that I changed was the initial values to reflect the current population and previous case, death, and hospitalization counts in Marin county.

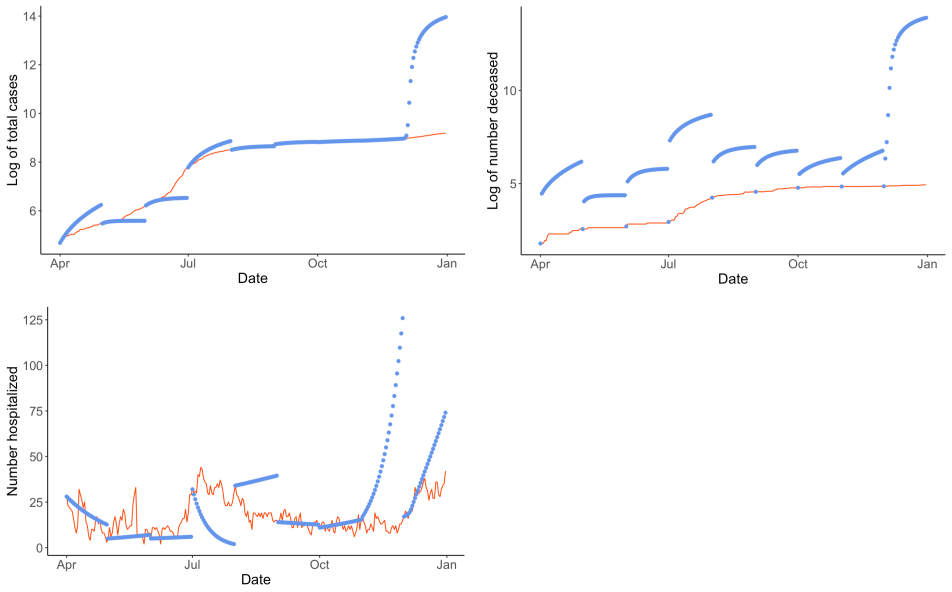


**Figure 3.5** Marin county data modeled with Los Angeles county parameters

Figure 3.5 displays the predictions for Marin county in May, based on the parameters fit for Los Angeles county in May, as well as the initial values calculated from Marin county's April data. It is important to note that the first "predicted" value for each month is the exact same as the data because it is the input initial value for the modeling process. This is especially clear in the middle plot of the logarithm of the number of deceased individuals, where the first value appears on the line with the actual values and then the predictions jump up at least 3 (back-transformed, this becomes $e^3 = 20.1$) and plateau until the next month. For the logarithm of the total cases and the number of hospitalized, the plateaus in September and October are predicted very accurately with the Los Angeles county parameters, suggesting the potential trend that both Los Angeles and Marin county had similar plateaus during the fall. Furthermore, we can see in all three of the plots that around November and December, the predicted cases, deaths, and hospitalizations increase sharply. This suggests that Los Angeles county had a much greater surge in these winter holiday months than Marin county.

## 3.4 Data inaccuracies

### 3.4.1 Exploration and intuition

Throughout my analysis, one of my main goals was to try to understand the variability in the data in order to make better predictions. COVID-19 data collection does not provide completely accurate daily statistics on COVID-19 cases because the majority of testing facilities are closed on Sundays, which means there appears to be less cases on Sundays. Furthermore, around holidays there can be spikes due to community and family contact and also drops on the day of the holiday due to less testing.

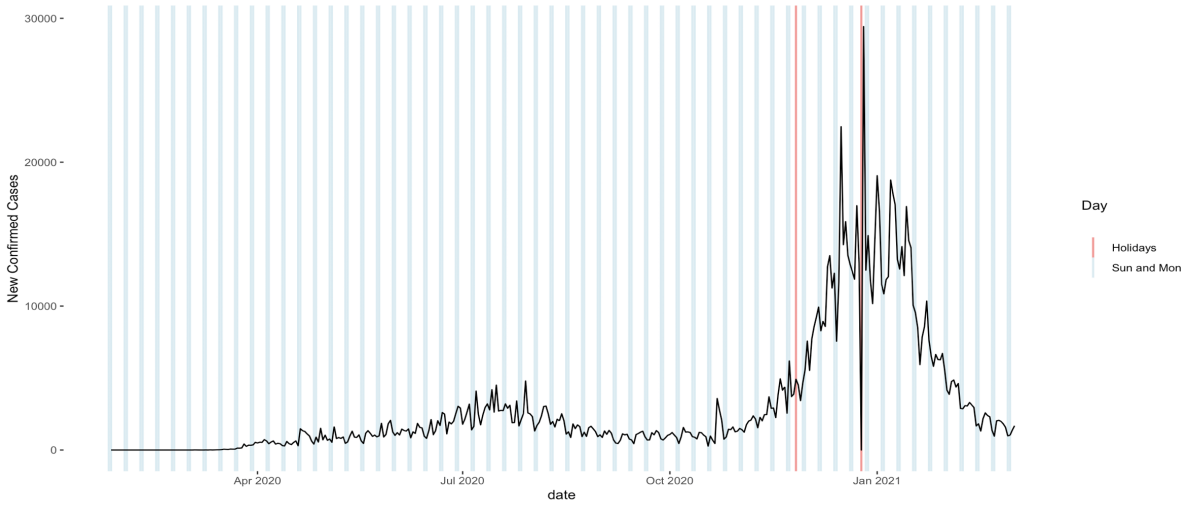


**Figure 3.6** New cases in Los Angeles county over time, with Sundays and Mondays in blue and Thanksgiving and Christmas in red

In Figure 3.6, the blue lines highlight Sundays and Mondays, which in most weeks happen to overlap with the dips in the data. Potentially, the systems are testing the data from Saturday and the past week and continue to catch up on testing on Monday, leading to fewer cases recorded both Sunday and Monday. During the holidays, there is a clear drop on Christmas day and a spike right after. For Thanksgiving, there is a small spike before the holiday, potentially people getting tested before visiting family. Thanksgiving also marks the beginning of the huge wave from the beginning of December through January in Los Angeles county.

### 3.4.2 Simulating noise

In order to understand how potential noise affects the parameter optimization and model prediction accuracy, I simulated new values for cases, deaths, and hospitalizations, taking 100 different samples from a multinomial distribution where the set proportions were the proportions of cases, deaths, and hospitalizations, independently, on each day. After this step, I had three 31x100 matrices (one for each data set: cases, deaths, and hospitalizations) simulating 100 values for each day of the month of May. I fed these matrices into the `optim` function to optimize the parameter with respect to this simulated data and obtained an 11x100 matrix, where each row is a different estimated parameter ($\alpha, \beta, \delta_I, \delta_M, \gamma_E, \gamma_R, M_{max}, \mu_I, \mu_M, \omega_I, g$) and the columns are the different simulated samples. With these values, I created the boxplots in Figure 3.7 to examine the distribution of each parameter.

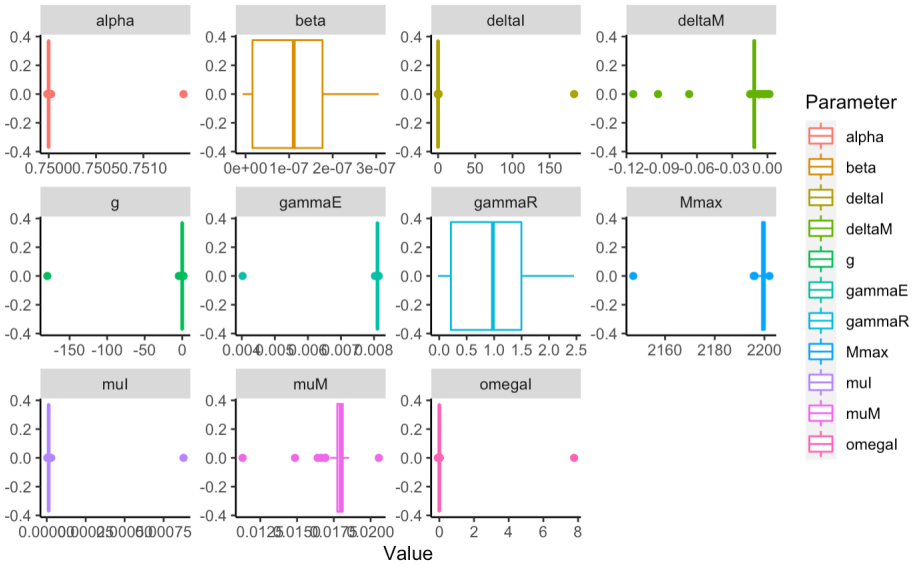


**Figure 3.7** Parameter values fit from bootstrapped May Los Angeles county data infused with random noise

With the exception of a few outliers, most notably in the distributions of $\delta_I, g,$ and $\omega_I$, the majority of the parameters do not appear to vary much with the infusion of random noise.

I used the 11x100 matrix of parameter values fit with bootstrapped data infused with random noise to estimate May cases, deaths, and hospitalizations

and plotted the results in 3.8, where the observed data line is in dark blue and the simulations are in orange.

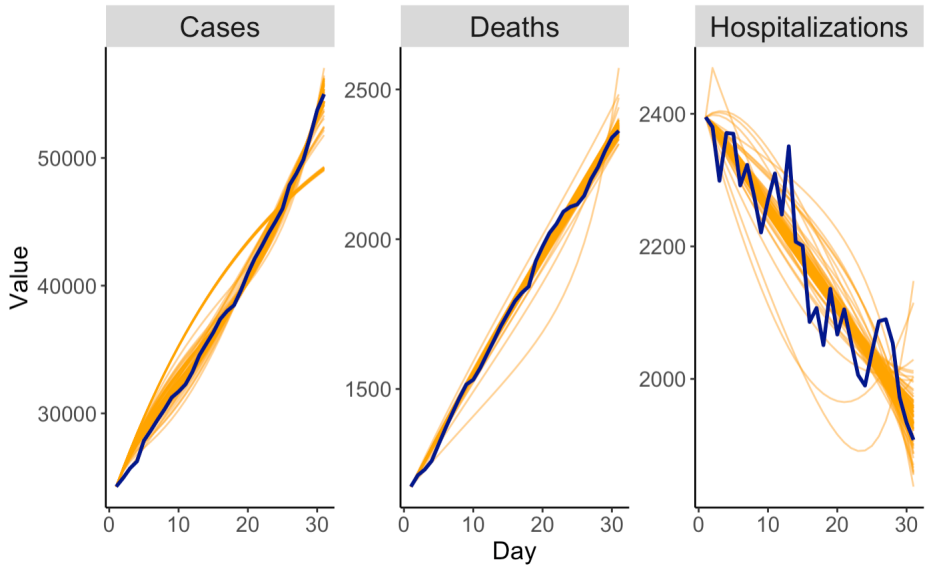


**Figure 3.8** Estimates of numbers of cases, deaths, and hospitalizations modeled with parameters fit using bootstrapped May Los Angeles county data infused with random noise

Similar to the results from Figure 3.7, most of the simulated predictions are accurate estimates of the observed data. There are a few outliers, especially for the hospitalization data. Overall, this analysis suggests that our model is not overly sensitive to noise.

# Chapter 4

# Conclusions and Future Work

## 4.1 Conclusion

Many of the aspects of the Edholm et al. COVID-19 model that I addressed in my research have to do with variability: variability of the data and parameters, and how to make the best decisions to optimize the prediction accuracy. The following categories lay out my conclusions.

### 4.1.1 COVID-19 data variability

Owing to many external variables including state-level restrictions on mask-wearing and school and business re-openings, COVID-19 data has surged, dipped, plateaued, and surged again over the past year. It is further varied by data collection practices such as closed testing facilities and labs on Sundays and events such as holidays that introduce more social interaction and potential for infection.

Through bootstrap simulations that added random noise into the COVID-19 case, death, and hospitalization data, I determined that noise does not appear to have a large effect on parameter optimization and model estimates.

### 4.1.2 Parameter variability

Due in part to the waves of COVID-19 cases throughout the pandemic, the optimal parameters change depending on the time interval of the data used to fit the values. By identifying the parameters that are variable at specific time intervals, it could be possible to implement changes to the model to decrease the parameter variation and therefore prediction accuracy.

Furthermore, when applying parameters fit from Los Angeles county data to model Marin county COVID-19 outcomes, the estimations overpredict the number of deaths and certain months of hospitalization and case data. Therefore, parameter estimation with this system of differential equations should be done independently for different locations and time frames.

### 4.1.3 Optimization and prediction options

There are *many* options for ways to optimize the parameters and predict the data. By developing a Shiny app and using other visualization and statistical tools, I was able to explore many of these combinations, determining that shorter intervals for model building proved to be more effective in capturing trends that would continue into the following month (or two).

### 4.1.4 Assessment of predictions

The inconstant parameter values across time also made it difficult to predict accurately for months farther out than one or two. Especially when comparing actual and predicted hospitalization data, the predictions become much worse the farther from the month with which the parameters were fit.

## 4.2 Future Work

Because this model was created early on during the pandemic, many of the initial values and assumptions could be updated based on more recent knowledge of the virus. There is far more data to accurately estimate parameters such as transmission and recovery rates than there were in June, when the model was developed. Furthermore, relationships that were defined based on assumptions, for example, the ratio of exposed individuals to infected individuals, have likely been studied since the creation of the model and could be updated. Specifically for the Exposed variable, Professor Edholm and I discussed the possibility of optimizing the initial Exposed value in order to move away from the simple assumption that 20 times more individuals are exposed than infected. Additionally, I would be interested in including a saturation incidence term into the rate between Susceptible and Exposed as developed by Esteva and Matias (2001) and seen in the Oluyori and Adebayo (2020) SEIRS model, as well as accounting for

waning immunity with the inclusion of recovered individuals back into the Susceptible population.

# Bibliography

2020. Who director-general's opening remarks at the media briefing on covid-19 - 11 march 2020. URL https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

2021. Compartmental models in epidemiology. URL https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology#The_SIR_model.

Blanchard, P., R.L. Devaney, and G.R. Hall. 2006. *Differential Equations*. Thomson Brooks/Cole. URL https://books.google.com/books?id=mwxX2pv9UvYC.

Cooper, Ian, Argha Mondal, and Chris G. Antonopoulos. 2020. A sir model assumption for the spread of covid-19 in different communities. *Chaos, Solitons Fractals* 139:110,057. doi:https://doi.org/10.1016/j.chaos.2020.110057. URL http://www.sciencedirect.com/science/article/pii/S0960077920304549.

Department, LA County Public Health. 2021. URL http://publichealth.lacounty.gov/media/coronavirus/locations.htm.

Department of Public Health. ???? Los angeles county announces 51 new deaths related to 2019 novel coronavirus (covid-19) - 815 new cases of confirmed covid-19 in los angeles county. URL http://publichealth.lacounty.gov/phcommon/public/media/mediapubdetail.cfm?unit=media&prog=media&ou=ph&prid=2362&keywords=stage&row=25&start=26.

Edholm, Christina J., Maryann E. Hohn, and Ami Radunskaya. 2021. Mathematical model of covid-19 at the claremont colleges. *in progress* .

Esteva, Lourdes, and Mariano Matias. 2001. A model for vector transmitted diseases with saturation incidence. *Journal of Biological Systems - JBS* 09:235–245. doi:10.1142/S0218339001000414.

Hindmarsh, Alan. ???? Odepack: Fortran ode solvers. URL https://computing.llnl.gov/projects/odepack.

Kermack, W. O., and A. G. McKendrick. 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character* 115(772):700–721. URL http://www.jstor.org/stable/94815.

Los Angeles Times Staff. 2021. california-coronavirus-data. https://github.com/datadesk/california-coronavirus-data.

Ng, Kok Yew, and Meei Mei Gui. 2020. Covid-19: Development of a robust mathematical model and simulation package with consideration for ageing population and time delay for control action and resusceptibility. *Physica D: Nonlinear Phenomena* 411:132,599. doi:https://doi.org/10.1016/j.physd.2020.132599. URL http://www.sciencedirect.com/science/article/pii/S0167278920302700.

Oluyori, David. A., and Helen. O. Adebayo. 2020. Global analysis of an seirs model for covid-19 capturing saturated incidence with treatment response doi:10.1101/2020.05.15.20103630.

Radhakrishnan, K, and A C Hindmarsh. 1993. Description and use of lsode, the livemore solver for ordinary differential equations doi:10.2172/15013302. URL https://computing.llnl.gov/casc/nsde/pubs/u113855.pdf.

Roda, Weston C., Marie B. Varughese, Donglin Han, and Michael Y. Li. 2020. Why is it difficult to accurately predict the covid-19 epidemic? *Infectious Disease Modelling* 5:271 – 281. doi:https://doi.org/10.1016/j.idm.2020.03.001. URL http://www.sciencedirect.com/science/article/pii/S2468042720300075.

Ross, R., and W. Hamer. 2013. The sir model and the foundations of public health.

Staff, LAist. 2020. Los angeles and orange counties confirm first cases of new coronavirus. URL https://laist.com/2020/01/26/coronavirus-orange-county-first-case-confirmed.php.

Stone, Will. 2021. It's been a year since 1st coronavirus case was reported in u.s. URL https://www.npr.org/2021/01/19/958472416/its-been-a-year-since-1st-coronavirus-case-was-reported-in-u-s.

Xu, Zhiliang. ???? Stiff differential equation. URL https://www3.nd.edu/~zxu2/acms40390F15/Lec-5.11.pdf.

Yadav, Ramjeet Singh. 2020. Mathematical modeling and simulation of sir model for covid-2019 epidemic outbreak: A case study of india doi:10.1101/2020.05.15.20103077.