Statistical Methods in Infectious Disease Modeling

Kaitlyn Stocker

2016-11-13

1 Introduction to Infectious Disease Modeling

Infectious diseases can be split up into various classifications. They can caused by micro-parasites or macro-parasites, and they can be directly or indirectly transmitted. For my thesis, I will be focusing on directly transmitted micro-parasites, meaning that I am looking at single-cell pathogens that are transmitted through direct contact with an infected individual. The micro-parasitic classification also means that the density of pathogens within an infected individual is of no concern - for modeling purposes, an individual is either infected or they are not.

When modeling infectious diseases, individuals are classified based on their ability to transmit or contract the disease. Susceptible individuals are those who have never contracted the disease and are therefore susceptible to contract it. Exposed individuals are those who have been exposed to an infection, but who are not yet contagious due to low levels of the pathogen. Infected individuals are those who are infected with the pathogen and who are able to transmit the disease. Recovered individuals are those who were once infected, but who have recovered with complete immunity. The proportion of susceptible, exposed, infected, and recovered individuals in a population are typically denoted by S, E, I, and R respectively.

The purpose of modeling is to learn about how diseases spread and operate, and to inform prevention methods to control infectious diseases. Mathematical models of infectious diseases are conceptual tools that attempt to explain how an infectious disease will be have in a population.

Models are specific to the characteristics of a given disease. For instance, not all diseases have an exposed state, and some do not have a recovered state. Some diseases cause mortality in infected individuals, while others do not. The length of the infected state varies between diseases, and the rate of transmission depends on both the infectiousness of the disease and the rate of contact between susceptible and infected individuals within a given population.

2 Deterministic SIR Model

I began my study of infectious disease modeling by working with a deterministic, continuous time model of an SIR disease. For the sake of simplicity, I began by looking at a closed population in which there was no effect of demographics or migration on the population.

When modeling the spread of an infectious disease through a population, there are two key parameters of interest: the transmission rate β, and the recovery rate γ. The transmission rate is the product of the rate of contact between susceptibles and infecteds in a given population, and the duration of the infection is the average length of time an individual will remain in the infected class. The duration of the infection is then given by the reciprocal of the recovery rate, $frac1\gamma$. These parameters, together with the initial values of S, I, and R, are the necessary pieces of information required to simulate the

spread of the infection through a population.

The proportion of susceptibles, infecteds, and recovered individuals over time is represented by a series of differential equations given by the following:

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

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

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

In this model, βSI is the transmission term, and represents the number of individuals flowing from the susceptible class into the infected class, and γI represents the flow of individuals from the infected class into the recovered class.

Simulation To produce an example of what this model looks like in action, I simulated the spread of influenza through a boy's boarding school. I took the values of the parameters and the initial conditions from the book **INSERT CITATION**. In this example, β is 1.66, and γ is $\frac{1}{2.2}$. In a population of 763 boys, at the start of the epidemic 3 were infected and the rest were in the susceptible class.

As the differential equations defining the model are not possible to solve explicitly, I used Euler's method to solve the system. In R, I ran the system of equations through Euler's method with a 0.01 time step for a period of 15 days. I outputted a data frame that included the value of S, I, and R for each time step through completion. Figure **??** shows a plot of the proportion of each infection class over time.

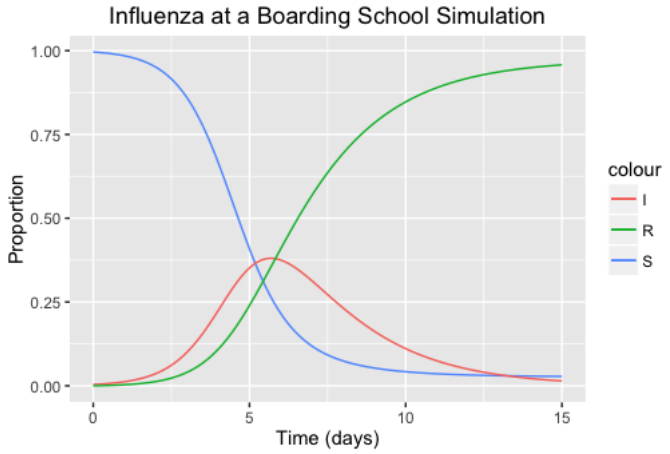

Figure 1: Deterministic SIR Simulation

Inference I then ran inference on the simulated data to retrieve back the value of β. To do this, I first created a function that simulated data for a series of β values ranging from 0 to 3 with a step of 0.01, maintaining the same initial conditions and γ value as the initial simulation. This function created a data frame of results for each value of β. I then ran a sum of squares function and took the squared difference between the results of my original simulation and the results of my estimation function. A plot of the estimated β values against the resulting sum of squares output is presented in figure **??**. It is visually evident that the minimum of the function occurs at 1.66, the actual value of β that I used to simulate my data. Running the optim function in R to minimize the sum of squares function returned the expected β of 1.66.

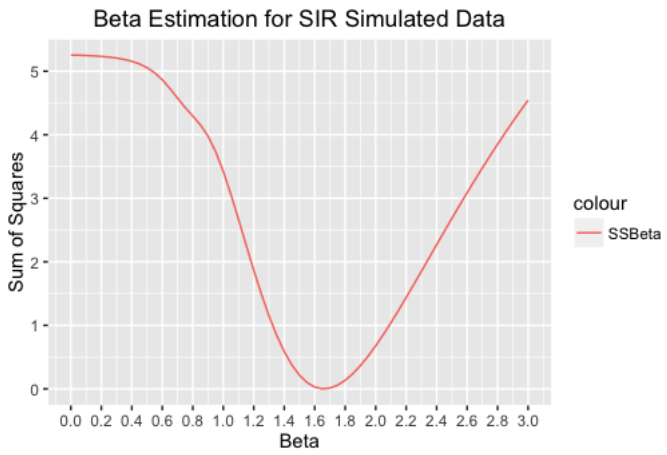

Figure 2: Plot of estimated β values against their sum of squares output

3 Adding Stochasticity

Deterministic models are useful tools for understanding the interplay of various parameters on the spread of a disease through a population, but they do not account for the inherent randomness of the real world. In reality, the amount of individuals who become infected during a given time interval is not fixed, but is rather randomly distributed. Therefore, the next step in modeling infectious diseases is to add stochasticity to the model.

3.1 The Chain Binomial Model

One way of adding stochasticity is to use a chain binomial model. In this type of model, we allow the the number of infecteds at each time step to follow a random binomial distribution in which the number of trials is equal to the number of susceptibles at the previous time step (the pool of individuals who

could potentially be infected). If we consider p to be the probability that contact occurs between a susceptible and a single infected individual and that the contact results in the infection, then the probability of an individual escaping infection from one infected is given by (1-p). In order for a given susceptible to escape infection entirely, they must avoid infection from all infecteds in the population at that time. In this way, the probability of a susceptible escaping infection entirely is given by $(1-p)^I$. In this case, the probability that a given susceptible will become infected in a given time step is given by $1-(1-p)^I$. It follows that the number of infecteds in a given time step follows a binomial distribution with S_{t-1} as the number of trials and $1-(1-p)^I$ as the probability of success.

The next logical step is to define p in terms of our parameters, β and γ. We can do this by defining the probability that a given susceptible will become infected in a given time step (with I_t infecteds) as $1-exp(-\beta I_t/N)$, where N is the total population size. I divided the number of infecteds by the population size to obtain the proportion of infecteds in the population. This was necessary because β is derived for use with population proportions, and in this model S, I, and R will refer to the number of individuals in each class. We allow length of the infectious period, γ, to be equal to the time step. In this way, at the end of each time step the infecteds from the previous time step all move into the recovered class.

Simulation I simulated an SIR infection using the chain binomial model. For consistency and comparison, I used the parameters and starting conditions from the Influenza at a Boarding School example, the same example that I used to simulate the deterministic example, taken from [**?**].

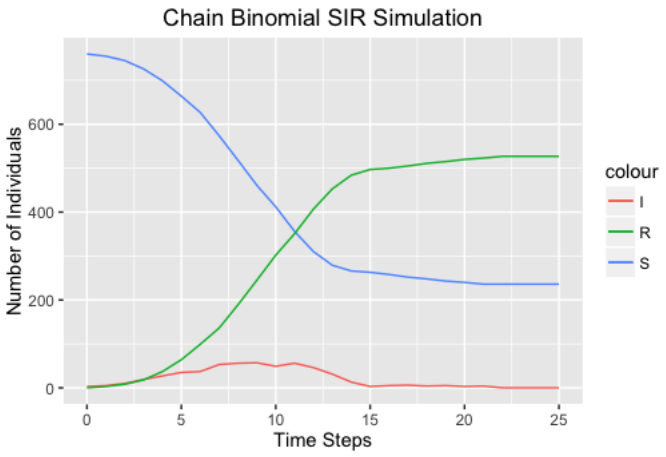

Figure 3: Chain Binomial Simulation of Influenza at a Boarding School Example

To achieve the output in figure **??**, I ran equations (4) through (6) through a

loop for 25 time steps. I chose 25 time steps based on the length of the epidemic observed from running the example using a deterministic model.

$$I_{t+1} \sim Binomial(S_t, 1 - exp(\frac{-\beta I}{N})) \tag{4}$$

$$S_{t+1} = S_t - I_{t+1} \tag{5}$$

$$R_{t+1} = R_t + I_t \tag{6}$$

It is clear from figure **??** and figure **??** that the chain binomial model does not have as dramatic of a peak as the deterministic model.

MLE Inference: To run inference on the chain binomial model, I used a maximum likelihood estimate (MLE) approach. To do this, I first found the likelihood function for β, which is given by equations (7) and (8) below.

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \binom{I_{i-1}}{S_{I-1}} p^{I_i}(1-p)^{S_i} \ , \ where \ p = 1 - exp(\frac{-\beta I_{i-1}}{N}) \tag{7}$$

This can be simplified to:

$$\mathcal{L}(\beta) \propto p^{I_{[i]}}(1-p)^{S_{[i]}} \tag{8}$$

I then used the optimize function in R to compute the value of β that maximized the log likelihood function, and received a value of 1.73, which is reasonably close to the value I simulated with, $\beta = 1.66$. A plot of the likelihood function can be found in figure **??**.

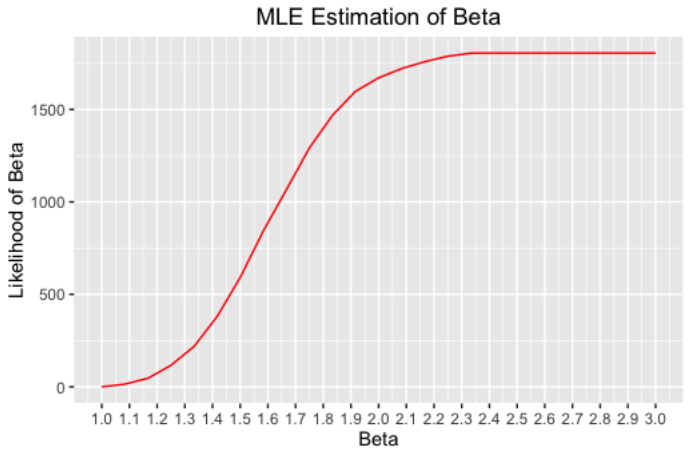

Figure 4: Plot of the likelihood function for β for the chain binomial model

Bayesian Inference: Unlike the classical approach, which treats model parameters as fixed values, Bayesian inference treats model parameters as random variables. The distribution of the parameters is calculated via Bayes' Theorem based on information given via a prior distribution and a likelihood computed based on the data. This final distribution is called the posterior distribution, and it gives all relevant information about the parameters, including point and interval estimates. The posterior distribution is defined more precisely in the equation below:

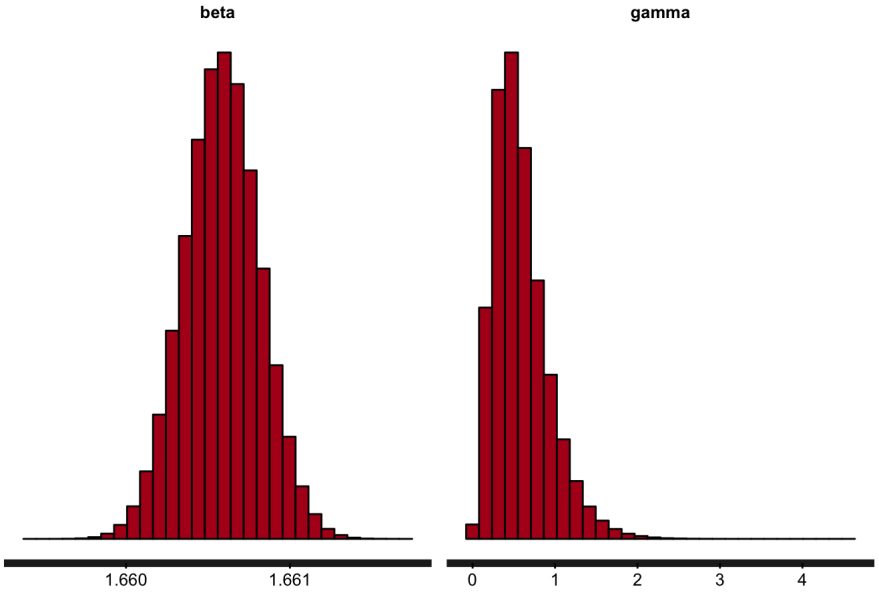

Figure 5: Plot of the posterior density functions for β and γ respectively

$$P(\theta \mid y) = \frac{P(\theta)P(y \mid \theta)}{\int P(\theta)P(y \mid \theta)d\theta} \tag{9}$$

Where $P(\theta)$ is the prior distribution, y is the data, and $P(y \mid \theta)$ is the likelihood function.

To run inference on my simulated epidemic data, I used Bayesian inference with the Markov chain Monte Carlo (MCMC) method. The MCMC method allows samples to be drawn from the target distribution - in this case, the samples are drawn from the joint posterior distribution of the model parameter.

I started by choosing a prior distribution for β and γ. I chose a gamma distribution with shape parameter equal to 3 and scale parameter equal to 1. This distribution has the majority of its density between 0.5 and 5, which is a reasonable range within which to expect β. For the prior on γ I chose a gamma

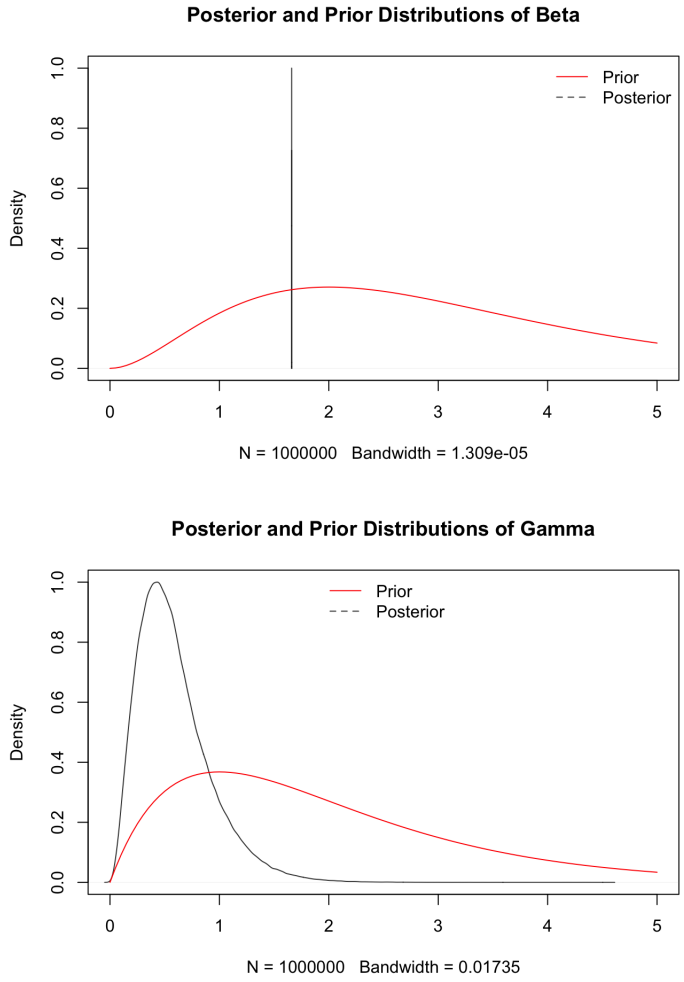

Figure 6: Plot of the posterior density function plotted over the prior density function for β and γ respectively

distribution with shape parameter equal to 2 and a scale parameter equal to 1. I chose this prior for γ because it has a strong right skew, and γ is typically less than one, as it is the reciprocal of the infectious period.

I used the package RStan to run Bayesian inference with the MCMC method on my simulated chain binomial model, with the aforementioned priors. Figure **??** displays the posterior density plots of β and γ.

The outputted estimate for β was 1.66, with a standard error of 0. Recall that the true value of β is 1.66. This level of accuracy is possible only with data simulated without any noise. Figure **??** also shows the lack of error in the posterior density of β. The outputted estimate for γ was 0.76 with a standard error of 0.33. The 95% credible interval for γ was (0.13, 1.40), which contains the true γ of $\frac{1}{2.2}$ or 0.455.

Based on simulated examples, recovery of β appears to be more precise than that of γ. This may, however, be a result of simulation mechanics. For small population sizes, recovery of γ was unreliable. However, population size did not affect the ability to accurately recover β.

3.2 The TSIR Model

The TSIR model is a time series, discrete time, stochastic model based on the basic SIR model. It is similar to the chain binomial, although a notable advantage of the TSIR model is that it can be used with reconstructed data.

Similar to the chain binomial, the time step of the model is set equal to the infectious period ($\frac{1}{\gamma}$).

The model is defined by the following equations:

$$E(I_{t+1}) = \beta I_t S_t N^{-1} \tag{10}$$

Where N is the population size, and I_t and S_t are the number of infected and susceptible individuals at time t, respectively.

$$I_{t+1} \sim NB(E(I_{t+1}), I_t) \tag{11}$$

Where NB(a, b) indicates the negative binomial distribution with expected value a and clumping parameter b. These equations follow the assumption of mass-action transmission with no demographics, as do the previous models specified.

Simulating I simulated epidemic data fit to a TSIR model using the initial conditions and parameters of the influenza example used with previously discussed models.

In order to allow for estimation of γ, I made the time step of the simulation an exponentially distributed random variable with $k = \frac{1}{2.2}$. In this way, the expectation of the time step would be equal to 2.2 weeks, the infectious period of influenza. Figure **??**.

Bayesian Inference

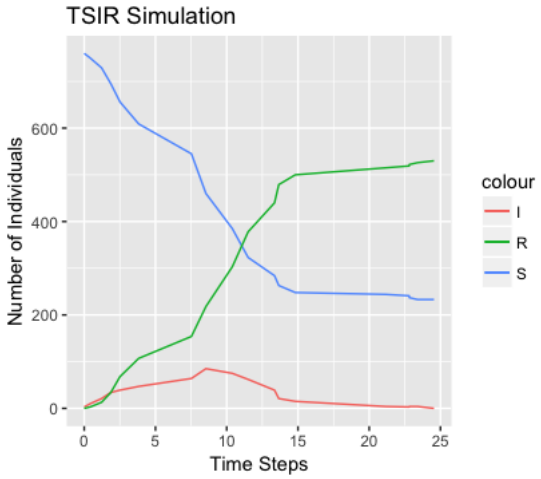

Figure 7: TSIR simulation of the Influenza at a Boarding School Example

3.3 Susceptible Reconstruction

In previous exercises, I had access to perfect and complete simulated data. While having such complete data is convenient, it is not realistic. Data gathered in real-world situations is much less inference-ready than the simulated data I have been using thus far.

Real data deviates from simulated data in a number of significant ways. For one thing, real data is incomplete. Not all cases of a disease are reported, so the number of infecteds at any given time point must be estimated using the number of reported cases multiplied by the reporting rate. There is typically no information about the true number of susceptible or recovered individuals, as collecting this information would be extremely impractical and cost-prohibitive.

In order to run any kind of meaningful inference on epidemic data, it is necessary to have at minimum the infected and susceptible dynamics over time. Using the reported cases and the rate at which cases are reported, it is easy enough to construct the infected class dynamics. However, reconstructing the susceptible class dynamics is not so straight-forward.

In order to reconstruct the susceptible class dynamics, let's first define the model. I will continue with the basic SIR model, but this time I am going to add in birth dynamics. The addition of birth dynamics into the susceptible class are crucial to the susceptible reconstruction process. To do this, I will define B_{t-d} as the number of births at time $t-d$. Since infants are born with natural immunity from their mothers, there is a time delay (denoted by d) between when a baby is born and when it enters the susceptible class. The length of this delay is dependent on the disease. As before, I define the size of the infected class at a given time point t to be $I_t, \in \{1, ..., T\}$. Similarly, I define the size of the susceptible class at a given time point t to be $S_t \in \{1, ..., T\}$. Equations 12

and 13 give the model specifications.

$$I_t = \beta S_{t-1} I_{t-1} \tag{12}$$

$$S_t = B_{t-d} + S_{t-1} - I_t \tag{13}$$

In equation 14 I allow I_t to be a product of the number of reported cases, C_t and ρ_t, the reporting rate at time t. I define ρ such that when $\rho_t = 1$, the number of true cases has been fully reported. When $\rho_t > 1$, the number of true cases has been underreported. Additionally, I assume that ρ_t follows a probability distribution with $E(\rho_t) = \rho$.

$$I_t = \rho_t C_t \tag{14}$$

Substituting equation 14 into equation 13, we get:

$$S_t = B_{t-d} + S_{t-1} - \rho_t C_t \tag{15}$$

If we define $E(S_t) = \bar{S}$, then we can define a new variable Z_t such that $S_t = \bar{S} + Z_t$, with $E(Z_t) = 0$. In this way, Z_t is the deviations from the mean of S_t. Z_t therefore follows the same recursive relationship as S_t, and can be defined as follows:

$$Z_t = B_{t-d} + Z_{t-1} - \rho_t C_t \tag{16}$$

If we allow Z_0 to be the initial value of Z, we can rewrite the previous equation to look like the following:

$$Z_t = Z_0 + \sum_{i=1}^{t} B_{i-d} - \sum_{i=1}^{t} \rho_i C_i \tag{17}$$

To de-clutter this notation, allow $Y_t = \sum_{i=1}^{t} B_{i-d}$ and $X_t = \sum_{i=1}^{t} C_i$. Additionally, we will assume a constant reporting rate. Now we can rewrite equation 17 as a simple linear regression equation:

$$Y_t = -Z_0 + Z_t + \rho X_t \tag{18}$$

Thus we have a linear regression equation relating cumulative births (Y_t) to cumulative reported cases (X_t). The susceptible dynamics Z_t are the regression remainder to equation 18, and can thus be fully reconstructed.

Simulated Susceptible Reconstruction Example In order to validate the accuracy of the susceptible reconstruction method taken from {citation}, I first used the method on simulated data. This way, I was able to compare the reconstructed susceptible dynamics to the true, simulated dynamics.

I first simulated an infection with transmission rate $\beta = 1.66$ and recovery rate $\gamma = 1/2.2$ using the previously defined TSIR model, with one modification. I added birth dynamics to allow the use of susceptible reconstruction. I simulated births using a birth rate of 12 births per 1,000 people annually {citation}.

Following the given method for reconstructing susceptible dynamics, I took the regression remainder from equation 18 and obtained Z_t. Figure **??** shows the reconstructed dynamics plotted alongside the true susceptible dynamics. The curves are equivalent, but as per definition Z_t is centered around zero.

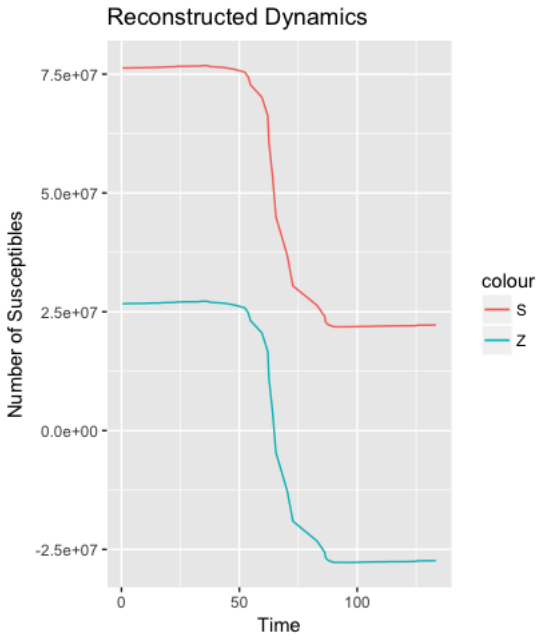

Figure 8: Comparison of reconstructed susceptible dynamics, Z, and the true susceptible dynamics S from the original simulation.

In order to fully reconstruct the susceptible dynamics, it would be necessary to know $\bar{S}$. However, it is not possible to compute the mean number of susceptibles directly, as we are assuming we do not have any data about the true susceptible dynamics. Therefore, I had to instead infer $\bar{S}$ along with the other model parameters (β and γ) when I performed Bayesian Inference.

I defined the prior on $\bar{S}$ to be normal with a mean of one fifth the initial population size and a standard deviation that was half of the mean. In other words, if we allow N_0 to be the initial population size, then I defined the prior on $\bar{S}$ as follows: $\bar{S} \sim Normal(\frac{N_0}{5}, \frac{N_0}{10})$. I then proceeded to replace any instance of S_t in the model definition with the equivalent expression $\bar{S} + Z_t$.

After running inference on my reconstructed susceptible dynamics, I took the upper and lower bounds from the 95% Credible Interval of $\bar{S}$ and used those values to obtain upper and lower bounds on the true susceptible dynamics. **??** shows a plot of the true susceptible dynamics (from the simulation) contained within the upper and lower bounds of the 95% credible interval for the reconstructed dynamics. The true values of β and γ were also contained within their respective 95% credible intervals.

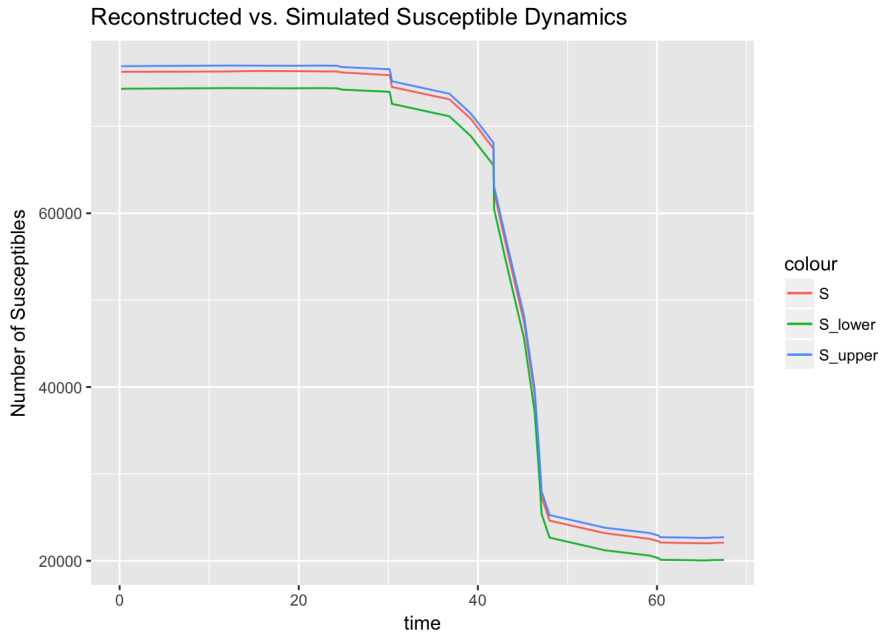

Figure 9: Plot of the true, simulated susceptible dynamics with the upper and lower estimated susceptible dynamics. The upper and lower bounds were obtained by taking the upper and lower limits of the 95% credible interval for the mean number of susceptibles, and adding each value to Z_t. The true number of susceptibles (in red) is clearly contained within the upper and lower bounds of the reconstructed susceptible dynamics.