Statistical Methods in Infectious Disease Modeling

Kaitlyn Stocker

4/20/2017

Contents

1 Introduction

2 Modeling and Inference for Infectious Diseases

2.1 Deterministic

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.

2.1.0.1 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 (Matt J. Keeling 2011). 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 1 shows a plot of the proportion of each infection class over time.

2.1.0.2 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 2. 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.

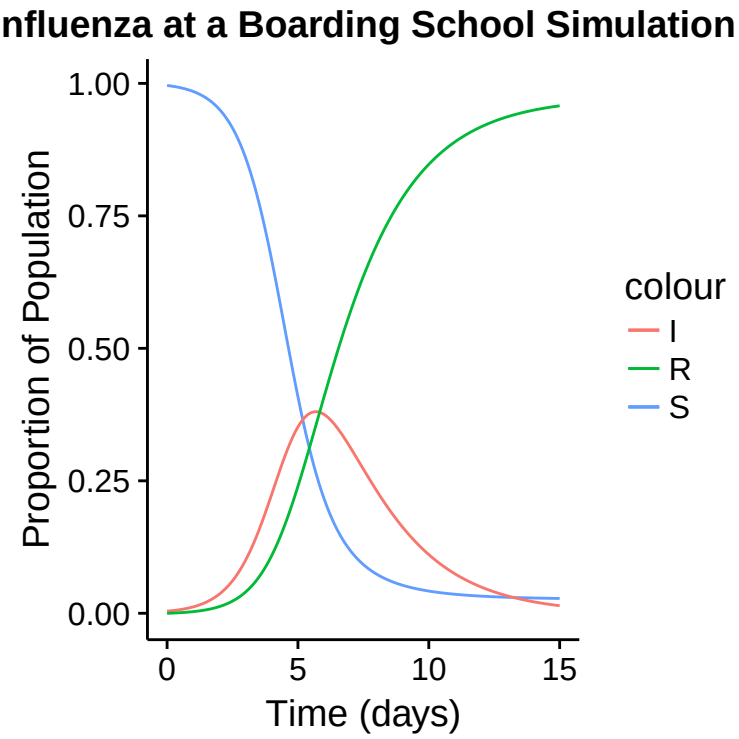

Figure 1: Deterministic SIR Simulation

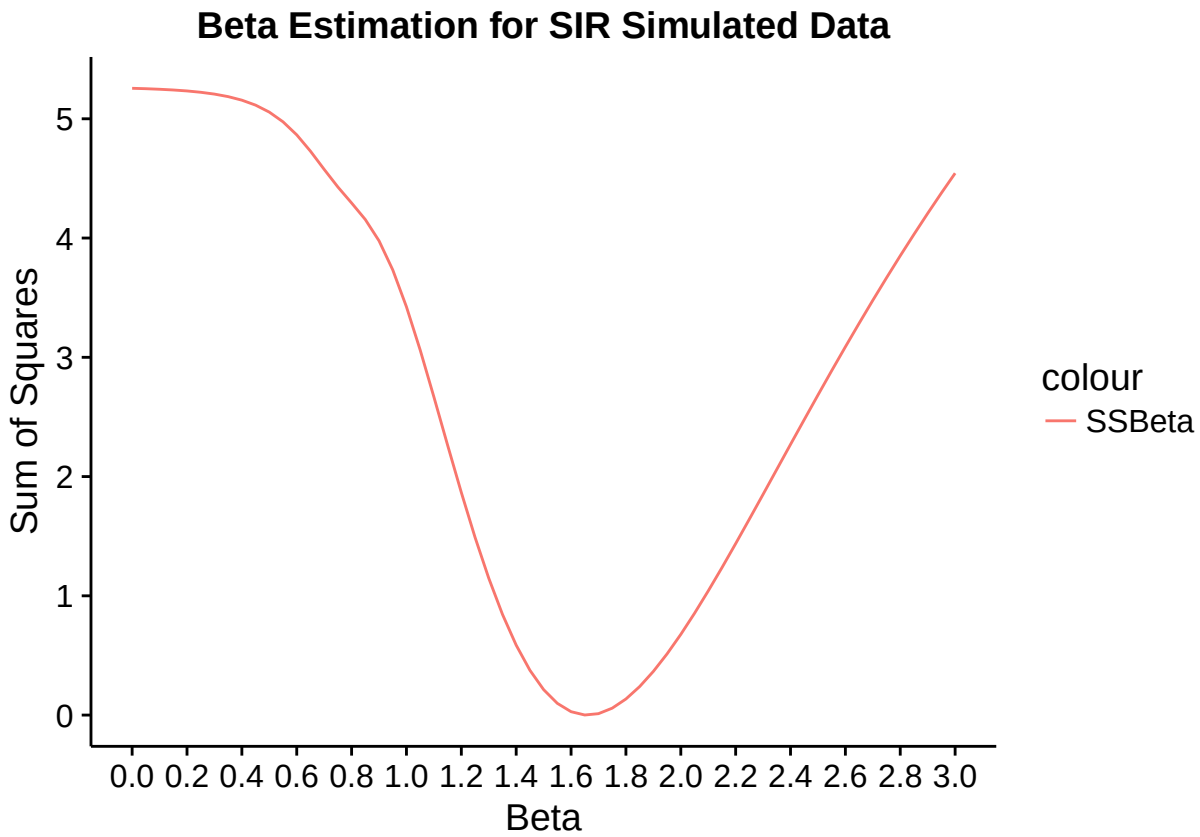

Figure 2: Plot of estimated beta values against their sum of squares output

2.2 Stochastic

2.2.1 Chain Binomial

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.

2.2.1.1 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 (Matt J. Keeling 2011).

To achieve the output in figure 3, 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 3 and figure **??** that the chain binomial model does not have as dramatic of a peak as the deterministic model.

2.2.1.2 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:

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

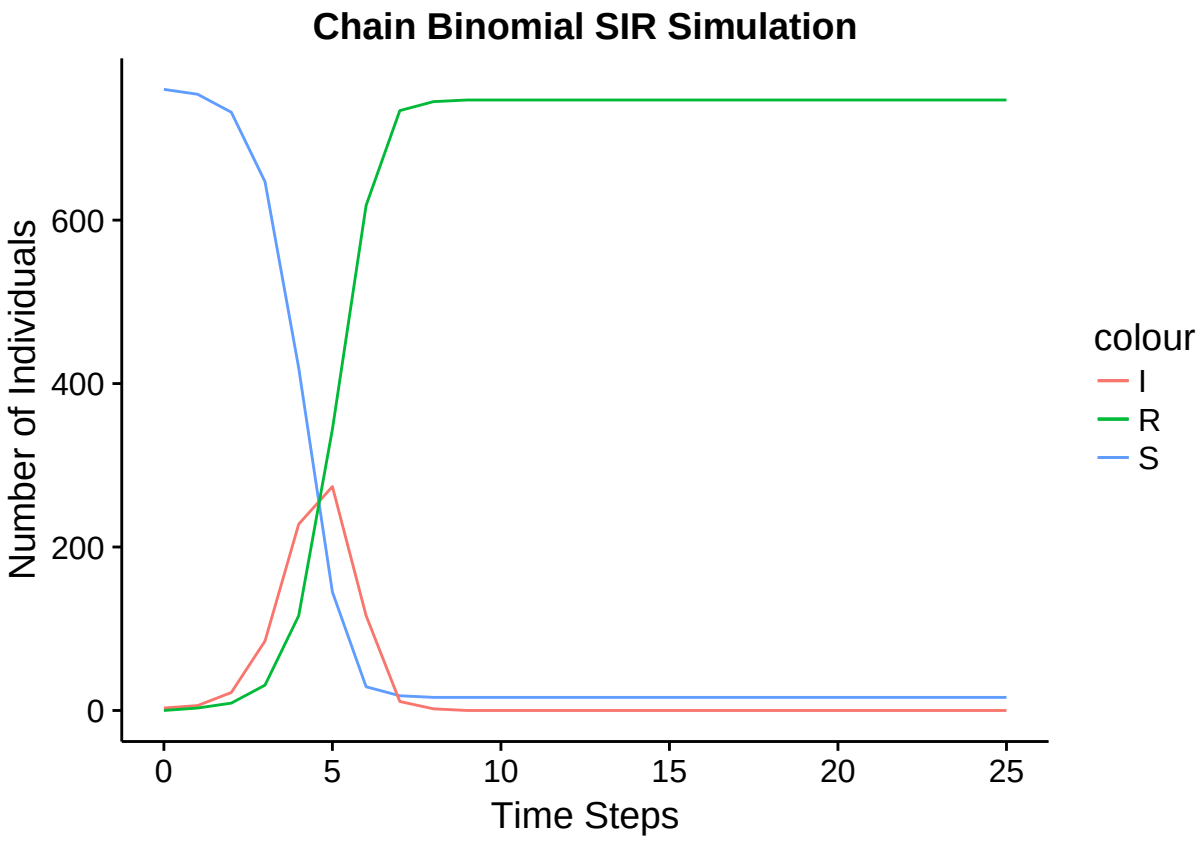

Figure 3: Chain Binomial Simulation

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 3.65, which is equal to the value I simulated with, $\beta = 1.66 * 2.2 = 3.65$. A plot of the likelihood function can be found in figure **??**.

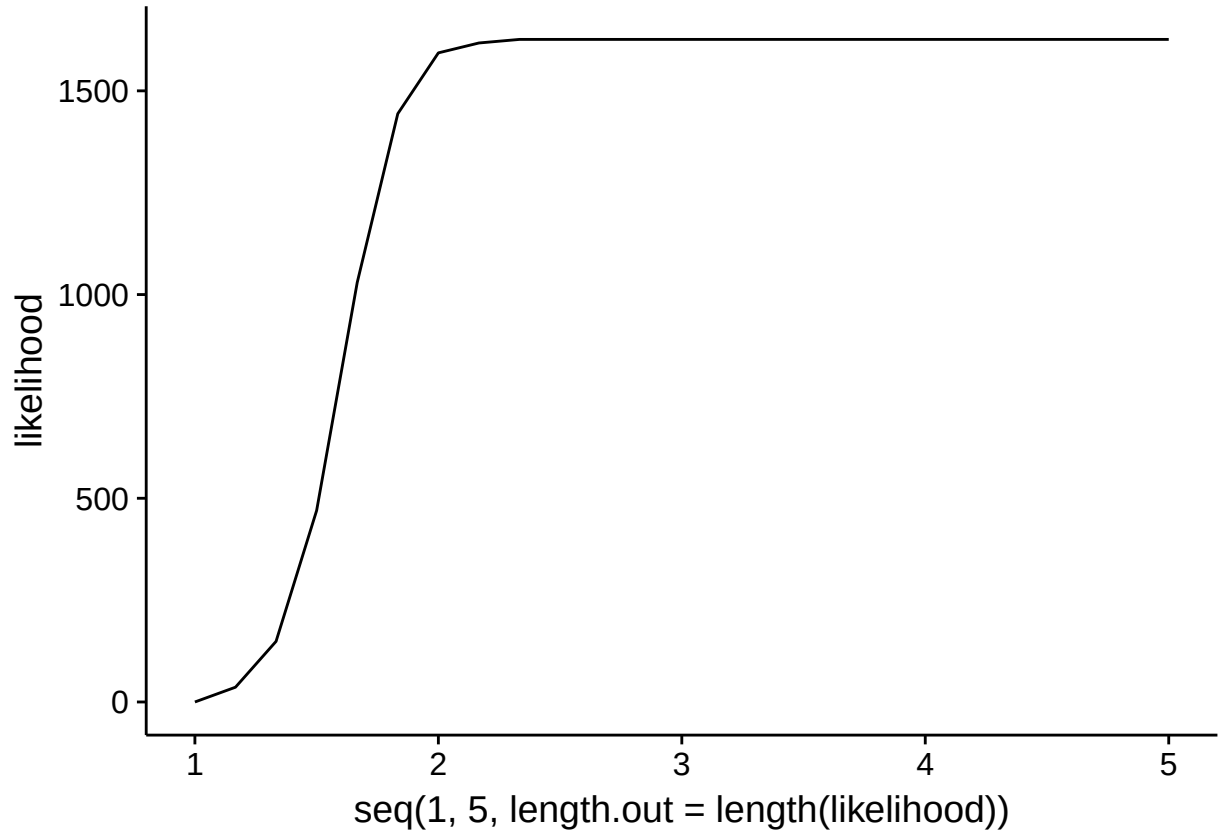

Figure 4: Plot of the likelihood function for beta for the chain binomial model

2.2.1.3 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:

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

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 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.

2.2.2 TSIR

2.2.2.1 Temporally Varying Transmission Rates

Up to this point, all of my examples have worked under the assumption that β, the transmission rate of the infection, is constant across time. However, this is really not the case! Many diseases display temporal shifts in transmission rate. For instance, most childhood diseases (such as measles or chickenpox) have higher transmission rates during school terms, and low transmission rates during school breaks. This makes intuitive sense - I would expect an infected child to come in contact with more susceptible children during school terms, when they are in constant contact with other children, than during school breaks, when they may be more isolated from their susceptible peers.

Measles during the pre-vaccination era displays temporally varying transmission rates, and I will be focusing from here on out on the dynamics of the measles virus. Measles has an infectious period of 2 weeks, and it is commonly modeled as having 26 time-varying transmission rates (one for each bi-week of the year).

It follows that the transmission rate can be expressed as a function of time, $\beta(t)$. Researcher Bailey (1975) found the transmission rate to be the sinusoid function given in equation 19 below:

$$\beta(t) = \beta_0(1 + \beta_1 cos(\omega t)) \tag{10}$$

The parameter β_0 in equation 19 gives the baseline transmission rate. The parameter ω gives the period of seasonal forcing, and is therefore equal to $\frac{\pi}{T}$ where T is the number of β's in a single calendar year. The parameter β_1 determines the amplitude of the seasonality and is therefore bounded between 0 and 1. For measles, β_0 is around 17 new cases per biweek, and $\omega = \frac{\pi}{26}$. In the case of seasonal transmission rates, $R_0 = \frac{\beta_0}{\gamma}$.

2.2.2.2 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.

$$I_t = \beta S_{t-1} I_{t-1} \tag{11}$$

$$S_t = B_{t-d} + S_{t-1} - I_t \tag{12}$$

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 $E(\rho_t) = \rho$.

$$I_t = \rho_t C_t \tag{13}$$

Substituting equation 14 into equation 13, we get:

$$S_t = B_{t-d} + S_{t-1} - \rho_t C_t \tag{14}$$

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:

$$Z_t = B_{t-d} + Z_{t-1} - \rho_t C_t \tag{15}$$

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

$$Z_t = Z_0 + \sum_{i=1}^{t} B_{i-d} - \sum_{i=1}^{t} \rho_i C_i \tag{16}$$

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:

$$Y_t = -Z_0 + Z_t + \rho X_t \tag{17}$$

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.

The infected dynamics are reconstructed by multiplying the reported cases at each time step by ρ, which is obtained as the slope of the linear regression equation.

2.2.2.3 Non-Constant Reporting Rate

In my previous explanation of susceptible reconstruction, I made a crucial assumption about the rate at which infected individuals are reported: I assumed that the reporting rate was constant across time. This assumption simplified the process of reconstructing susceptible and infected dynamics, but it rarely holds true in the world of real data.

As an example, let's look at measles data from pre-vaccination era New York City, from 1920-1940 (Willem G. van Panhuis 2013). In figure 5, plot (A) is a graph of the reconstructed susceptible dynamics, obtained as previously described using global linear regression. It is evident from plot (A) in figure 5 that the Z dynamics suffer from local shifts away from the mean of zero. This indicates that the previously held assumption of constant reporting rate, ρ, has been violated.

When local shifts in the mean of the Z dynamics are observed, such as those in plot (A) from Figure 5, it indicates a nonconstant reporting rate. Following the work of Finkenstadt and Grenfell (FinkenstaÈdt and Grenfell 2000), I addressed the issue of time-varying reporting rate by performing local linear regression with gaussian smoothers. To give an overview before I break down the process in detail, I split the data into overlapping chunks (or neighborhoods) centered around each $\{X_i, ..., X_N\}$, then assigned a gaussian weight to each observation in each neighborhood. Then I performed N weighted linear regressions using the previously defined gaussian weights, and derived the value of ρ for each time point by pulling the slope from each of these weighted linear regressions.

To break it down, I began as I did in my previous Susceptible Reconstruction example by computing the cumulative cases, $X_t = \sum_{i=1}^{t} C_i$, and the cumulative births, $Y_t = \sum_{i=1}^{t} B_{i-d}$. I then split the data into neighborhoods. To do this, I defined a bandwidth, h and a neighborhood size $m = T * h$ (where T is the number of observations) such that for each $\{x_t, ...x_T\}$, a neighborhood was constructed consisting of the m closest datapoints to the value of x_t. In this way, I constructed a matrix X_n with T rows and m columns. I additionally constructed a corresponding Y_n matrix.

I chose the bandwidth, h, in accordance with the method outlined by Finkenstadt and Grenfell (FinkenstaÈdt and Grenfell 2000). As Finkenstadt and Grenfell argued, automatic selection processes that minimize the residual to white noise are not suitable, as the residuals in the regression are significant. The method they proposed was to choose the bandwidth that minimized the difference between two sums of squares: the sum of square errors ($SSE_1(h)$), and the sum of squares of the deviations of the local estimator from the linear estimator ($SSE_2(h)$), where h is the bandwidth using a guassian kernel. Essentially, their method chooses a bandwidth that both minimizes the SSE while preventing over smoothing by tethering the local regression to the estimator obtained from the global linear regression. In the equations that follow, $\hat{m}_{h,t}(x_t)$ is the local estimator at point x with smothing parameter h.

$$SSE_1(h) = \sum_{t=1}^{T} \{Y_t - \hat{m}_{h,t}(x_t)\}^2 \tag{18}$$

$$SSE_2(h) = \sum_{t=1}^{T} \{\hat{Y}_t - \hat{m}_{h,t}(x_t)\}^2 \tag{19}$$

In figure 5 I reference data from the Tycho database (Willem G. van Panhuis 2013) for pre-vaccination measeles in New York City. The bandwidth that minimized the difference between SSE_1 and SSE_2 for this data was $h = 0.28$.

The next step was to apply a weight function. Following the work of Finkenstadt and Grenfell (FinkenstaÈdt and Grenfell 2000), I used a gaussian weight function, defined by the following:

$$K(x) = \frac{1}{\sqrt{2 * pi}} e^{\frac{-x^2}{2}} \tag{20}$$

$$w_i(x_t) = \frac{K(\frac{x_0 - x_i}{h})}{\sum_{i=1}^{m} K(\frac{x_0 - x_i}{h})} \tag{21}$$

In which $K(x)$ is the gaussian kernel function, and $w_i(x)$ gives the weight function. I applied the weight function to each row of the X_n matrix, for which x_0 is the focal, or central, x value.

To pull this together with an example, if I were to look at the datapoint x_{10} (the focal x_t of the 10^{th} row of the X_n matrix), I would first create a vector of length m of the x_t's nearest to x_{10}. I would then apply the weight function $w_i(x_{10}) = \frac{K(\frac{x_{10}-x_i}{h})}{\sum_{i=1}^{m} K(\frac{x_{10}-x_i}{h})}$ to each of the m x_t's in the neighborhood of x_{10}. I repeated this for each row of the X_n matrix, resulting in a weight matrix with the same dimensions of X_n.

Once I computed the gaussian weight for each datapoint in the X_n matrix, I ran a weighted linear regression on each of the T rows of the X_n matrix, using the gaussian weights from the previously computed weight matrix. The result of this was T simple linear regression outputs. It follows that ρ_t is a vector of the slopes of these linear regressions. I obtained fitted values by imputting the focal x (x_0) into the linear model for each time point. I then took the residual of the regression and obtained the Z dynamics from the local regression.

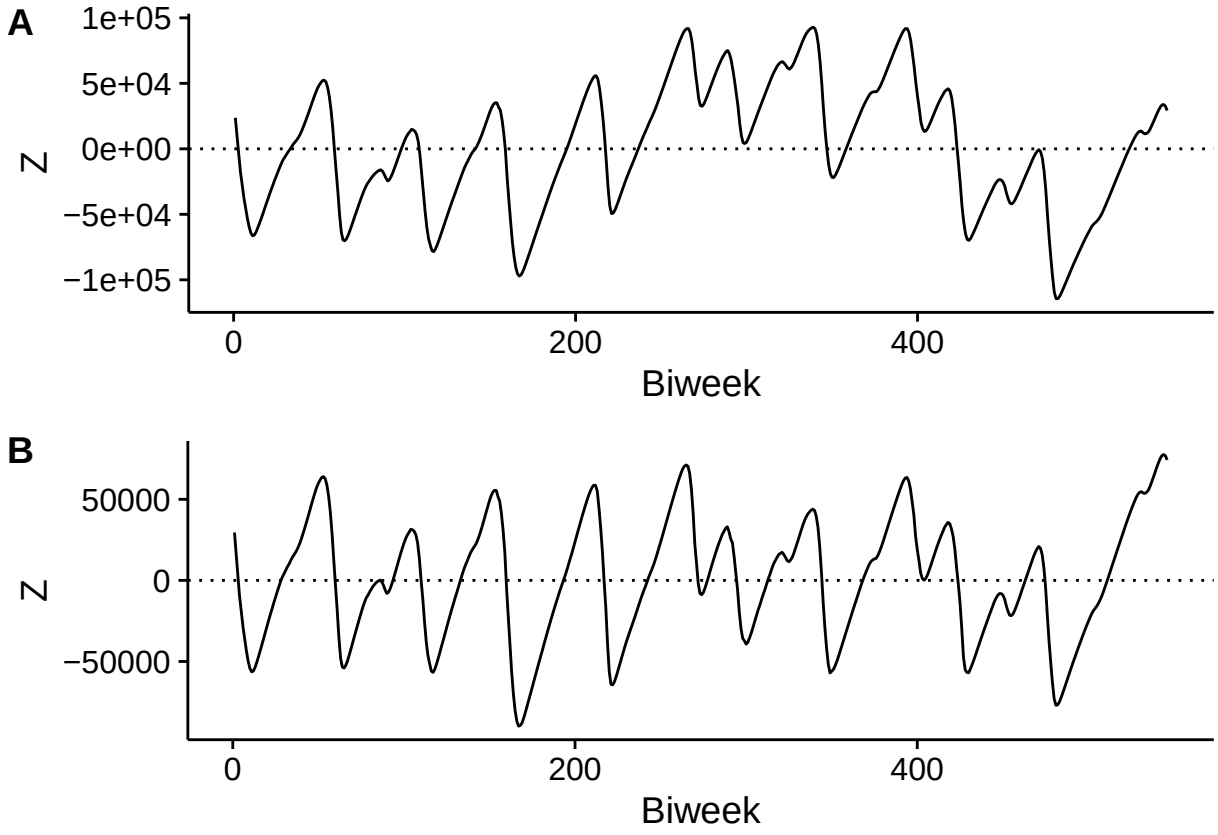

Figure 5: Comparison of reconstructed Z dynamics from the New York data. Plot (A) shows Z dynamics obtained via global linear regression. Plot (B) shows Z dynamics obtained via local linear regression.

Figure 5 demonstrates graphically that the Z dynamics obtained through local linear regression do not display the local shifts away from the mean seen in the Z dynamics obtained via global linear regression.

To complete the reconstruction process, the infected dynamics are obtained by multiplying ρ_t and the reported cases.

At this point, the data is fully reconstructed and can be passed through Bayes without any further modification. As previously discussed, the Bayes output will include posterior distributions from all 26 β_t's as well as for $\bar{S}$, the mean number of susceptibles.

2.2.2.4 Validating the Model

Once the data has been reconstructed and Bayesian inference has been performed, the model is complete. It is important, at this point, to make sure that the model obtained from the previous steps is actually able to accurately simulate the epidemic dynamics.

To this end, I simulated from my model for each of the cities I studied. However, before I could do that I needed to select point estimates for each β_t and for the initial proportion of susceptibles and infecteds (S_0 and I_0). I began by using the median from the posterior distrubution of each β_t, and taking the initial proportion of susceptible and infected individuals from the first time point of my reconstructed susceptible and infected dynamics. However, the simulations are highly influenced by small deviations in the initial proportion of susceptibles and infecteds. For this reason, I decided to choose my initial proportion of susceptibles and infecteds by simulating from a range of proportions and choosing the initial conditions that yielded the lowest mean square error when compared with the incidence data from (Willem G. van Panhuis 2013). I simulated with initial susceptible proportions between 0.025 and 0.05 and initial infected proportions between 0 and 0.001. I chose these values based on the work of Bjornstad et al (Benjamin D. Dalziel and Grenfell 2016). Figure 6 shows the successful forward simulation using the initial conditions and transmission rates taken from this procedure.

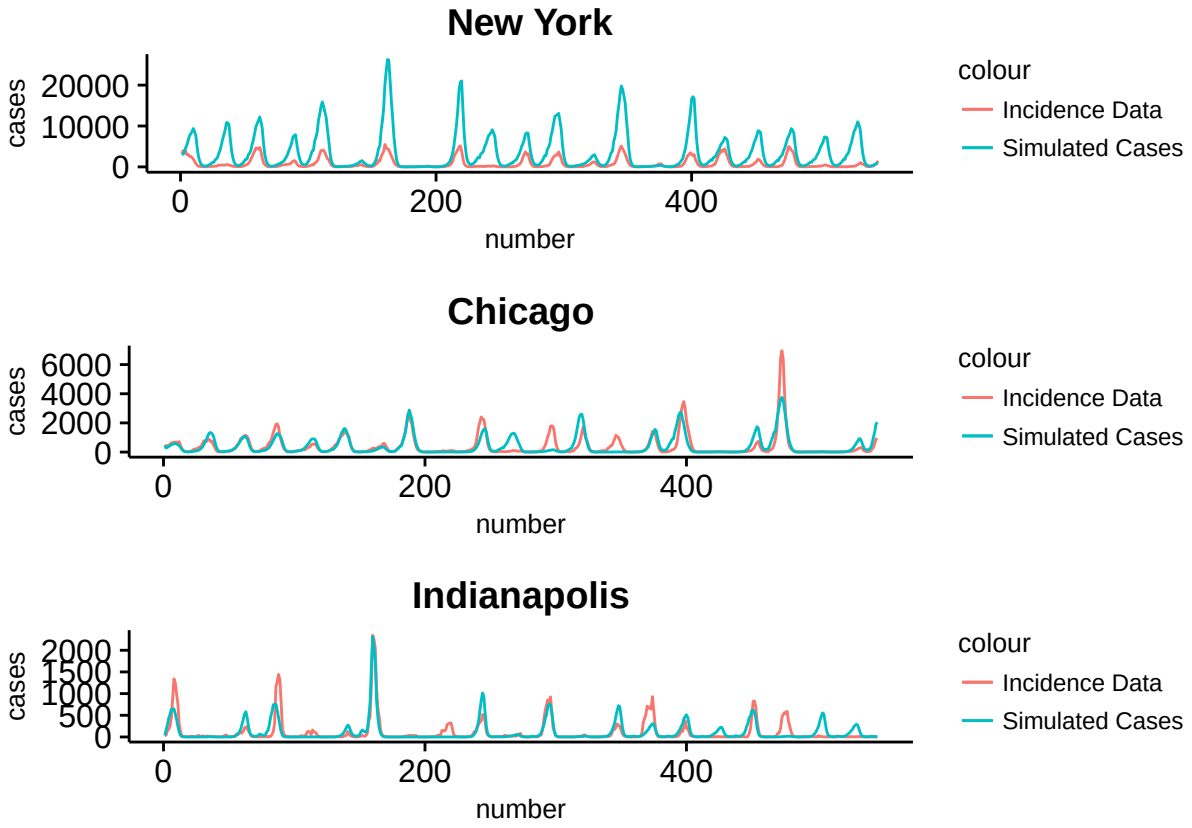

Figure 6: Plots show the successful forward simulation from the deterministic skeleton of the TSIR model parameterized with city-level data (in blue) versus the incidence data taken from tycho (in red). Simulations were started from the initial conditions obtained by minimizing the mean square error and were not fed any additional information from the incidence data once the simulation begun. Simulations were carried out for 20 years.

3 References

(Matt J. Keeling 2011) (Willem G. van Panhuis 2013) (FinkenstaÈdt and Grenfell 2000) (Benjamin D. Dalziel and Grenfell 2016)

Benjamin D. Dalziel, Willem G. van Panhuis, Ottar N. Bjørnstad, and Bryan T. Grenfell. 2016. "Persistent Chaos of Measles Epidemics in the Prevaccination United States Caused by a Small Change in Seasonal Transmission Patterns." Edited by Neil M. Ferguson. *PLoS Comput Biol* 12 (2).

FinkenstaÈdt, BaÈrbel F., and Bryan T. Grenfell. 2000. "Time Series Modelling of Childhood Diseases: A Dynamical Systems Approach." *Applied Statistics* 49 (2): 187–205.

Matt J. Keeling, Pejman Rohani. 2011. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press.

Willem G. van Panhuis, Su Yon Jung, John Grefenstette. 2013. "Contagious Diseases in the United States from 1888 to the Present." *NEJM* 369 (22): 2152–8.