THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF STATISTICS

STATISTICAL METHODS IN MODELING MEASLES DYNAMICS

Kaitlyn Stocker
Summer 2017

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Psychology
with honors in Statistics

Reviewed and approved* by the following:

Murali Haran
Professor of Statistics
Thesis Supervisor and Honors Adviser

Matthew Beckman
Professor of Statistics
Faculty Reader

*Signatures are on file in the Schreyer Honors College.

Abstract

In this thesis, I consider a few important examples of SIR models, study how these models behave, and consider approaches that have been used to fit these models. Based on several simulated examples, I examine how various model fitting approaches work in practice, and provide some practical recommendations. I apply the models and methods to measles data from multiple American cities. Based on the simulated data and real examples, I summarize some current challenges and open research questions.

Table of Contents

List of Figures

List of Tables

Acknowledgements

A great deal of work has gone into this thesis, and I would not have been able to do it without a strong support system.

First and foremost, I want to give my sincere thanks to my thesis and honors adviser, Dr. Murali Haran. One year ago, he agreed to take me on as a thesis advisee, and I have learned and grown more under his guidance than I have in any other year at Penn State. I want to thank my collaborator Ben Roberson, who worked with me in parallel on this project, for his help with some of the stickier problems we encountered, and for patiently helping me with coding in R when I was new to the program. I would also like to thank Dr. Matthew Beckman, who worked with me on a separate research project and who has also provided me with a great deal of guidance and support. Additional thanks are owed to Deb Rodgers for her help and patience as I navigated some logistical tangles, and, of course, to my friends and family who give me endless love and support. I am grateful to those friends and family members who humored me as I rambled about the highs and lows of this project, the problems that frustrated me to no end and the eventual solutions that I was bursting to share. As I have said, I could not have done it without all of you. My success is a reflection of the time, energy, and support that you have all given me, and I could not be more grateful.

Chapter 1

Introduction

Mathematical models of infectious disease dynamics have been used to both learn about the behavior of such diseases, as well as to make predictions about the most effective steps to handle an outbreak. An important class of models is the Susceptible-Infected-Recovered (SIR) , see for instance [1]. For such models, individuals in a population are first categorized based on their ability to either spread or contract the disease. In the case of the SIR model, these categories include susceptible (individuals who are not infected but who are susceptible to becoming infected), infected (individuals who are able to infect others), and recovered (individuals who once were infected, but who have since recovered with full immunity). The model is typically built by taking the proportion of individuals in each of those classes, combined with the length of the infection, to infer the rate of transmission and other features of the epidemic. Inference for model parameters may be carried out in a number of ways, including maximum likelihood estimation and Bayesian inference.

In this thesis, I consider a few important examples of SIR models, study how these models behave, and consider approaches that have been used to fit these models. Based on several simulated examples, I examine how various model fitting approaches work in practice, and provide some practical recommendations. I apply the models and methods to measles data from multiple American cities. Based on the simulated data and real examples, I summarize some current challenges and open research questions.

1.1 Overview of Infectious Disease Modeling

Diseases have altered the course of history. They have changed the course of wars, wiped out large portions of populations, and been a cause of much fear and strife. Therefore, humans have been studying infectious diseases and trying to find ways to prevent (and sometimes to weaponize) their spread throughout recorded history [2][3].

Models of infectious diseases increase understanding of the inner workings of epidemic dynamics, and allow for prediction of how an epidemic will behave under a given set of circumstances. A good model strikes a balance between accuracy,the ability of the model to match

real data, and transparency, the ability to use the model to learn about how different model parameters work to create the epidemic dynamics.

The models I will be working with look at population-level epidemic dynamics of directly transmitted diseases for which individuals who contract the disease recover with full immunity. Direct transmission refers to a property of an infection in which infection occurs only via direct contact with an infected individual. Models for such diseases are called SIR models [1].

The construction of an SIR model begins by dividing the individuals in a population into one of three classes: susceptible (a healthy individual who could become infected), infected (an individual who has contracted the disease and can infect others), and recovered (an individual who has previously been infected, but who has since recovered with full immunity to the diease). The proportion of a population belonging to each class is denoted by S, I, and R respectively. The SIR model derives its name from these classes. Figure 1.1 provides a visual representation the dynamics of an SIR model.

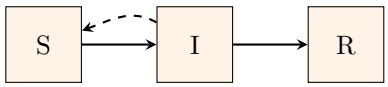

Figure 1.1: Diagram of the SIR model. The solid arrows represent the movementbetween the S and I classes and the I and R classes.The dashed arrow represents the fact that the rate at which susceptibles enter the infected class is affected by the amount of infected individuals in the population.

There are, of course, other types of diseases, and the model can be adjusted to reflect the properties of different infectious diseases. For instance, for diseases that are life-long and/or fatal, the recovered class is removed and the model is called an SI model. For diseases in which there is an incubation period between when an individual is exposed to the disease and when they become infectious, an SEIR model is used, in which the E represents the exposed class [1].

Throughout the course of my thesis, I will refer to the use of two methods of inference: maximum likelihood estimation and Bayesian inference. A brief introduction to these methods is included in the following sections.

1.2 Overview of Maximum Likelihood Estimation

Maximum likelihood estimation is an approach to estimating the value of a model parameter. It is based on the idea that values of the parameter that make the observed data appear relatively propbable are more likely to be the true parameter value than values that make the observed data appear relatively improbable [4]. Mathematically, maximum likelihood estimation is based on finding parameter values that maximize the likelihood function of the parameters for a given data set.

For a model with observed values $X = x_1...x_n$, and a probability density function given by $f(X \mid \theta)$, where θ is the model parameters, the goal of estimation is to find a value for θ based on the observed X.

In order to assess the likelihood of a given value of θ, we begin by constructing the likelihood function: $L(\theta \mid X)$. $L(\theta \mid X)$ gives the likelihood of the parameter θ given the observed data X. While $f(X \mid \theta)$ gives the probability of a given data value given the distribution of the data, $L(\theta \mid X)$ gives the probability of a given value of θ given the observed data X. It is typical to write the likelihood as $L(\theta) = f(X \mid \theta)$, as X is now fixed and the likelihood is a function of θ. For computational convenience it is most common to work with the log likeihood $l(\theta) = logL(\theta)$.

The maximum likelihood eestimate, or MLE, of θ is then given by:

$$\theta = \arg\max_{\theta} l(\theta) \tag{1.1}$$

1.3 Overview of the Bayesian approach

An alternative method for estimating the values of model parameters, θ, comes from the Bayesian approach. While maximum likelihood estimation treats the model parameters as fixed values, the Bayesian appraoch treats model parameters as random variables. Thus, the end result of Bayesian inference is a probability distribution on θ.

The Bayesian approach begins by defining a prior distribution on θ, which includes any information about θ known at the outset of the inference process. The goal of Bayesian inference is to update the prior distribution on θ in light of the observed data X. The result is the posterior density distribution, given in equations 1.2 and 1.3, in which $f(\theta)$ is the prior distribution on θ, and the likelihood $f(X \mid \theta)$ is specified by our model.

$$\pi(\theta) = \frac{f(\theta)f(X \mid \theta)}{\int f(\theta)f(X \mid \theta)d\theta} \tag{1.2}$$

$$f(X) = \int f(X \mid \theta)f(\theta)d\theta \tag{1.3}$$

The posterior distribution is rarely available in closed form, that is, for most models, it is impossible to compute the posterior distribution analytically. A common approach for approximating the posterior distribution $\pi(\theta|X)$ is to use Markov chain Monte Carlo. Here, a Markov chain is constructed such that the sequence of random variables in the chain is used as if they were approximately samples from the posterior distribution. Hence, it is possible to approximate various properties of the posterior distribution by simply using the Markov chain samples. This approach is called Markov chain Monte Carlo (MCMC). See, for instance, Gelfand and Smith [5] or Brooks et al. [6].

The Metropolis-Hastings algorithm provides a means of simulating a Markov chain with stationary distribution $f(\theta \mid X)$ by only requiring evaluation of $f(\theta, X)$.

$$f(\theta \mid X) \propto f(\theta, X) (= f(X \mid \theta) f(\theta)) \tag{1.4}$$

The MCMC methods simulate samples from the distribution of the model parameters given the data. Based on the parameter values at one step, a new set of paramter values is generated such that the stable distribution of the resulting Markov chain is the posterior distribution, $f(\theta \mid X)$.

Throughout this thesis, I use the statistical computing language R [7] to carry out simulations and model fitting. For the MCMC algorithms I use the package rstan [8].

The remainder of my thesis is organized as follows. In Chapter 2 I discuss several popular models used to study infectious disease dynamics. In Chapter 3 I explore the application of such models to the measles epidemic in pre-vaccination United States. I conclude in Chapter 4 with a summary of my work and some open research questions.

Chapter 2

Models of Infectious Disease Dynamics

Models for infectious disease dynamics fall into one of two categories: deterministic or stochastic. Deterministic models are often in continuous-time and consist of a set of differential equations which give the rates at which individuals enter and leave each of the classes (susceptible, infected, and recovered). I will consider discrete-time stochastic models, with each time step representing the completion of one infectious period. A popular model for measles research is the TSIR model [9], which utilizes a negative binomial distribution to capture the inherent stochasticity involved in the infection process. The TSIR model is also popular because it allows for the reconstruction of the susceptible class dynamics, which is a necessary step when processing incidence data for inference.

2.1 Deterministic Models

I began my study of infectious disease modeling by working with a deterministic, continuous time model of measles, 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.

Models for SIR diseases are built around two important parameters: the transmission rate, β, and the recovery rate γ. The transmission rate is the product of the rate of contact between infected and susceptible individuals in a population and the probability of transmission given contact. The recovery rate is the rate at which infected individuals recover and move to the recovered class. The inverse of the recovery rate, $\frac{1}{\gamma}$, denotes the infectious period, or the average length of the infection [1]. The recovery rate is typically learned via laboratory studies, and is therefore assumed to be known when working with population-level models. 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:

$$\frac{dS}{dt} = -\beta SI \tag{2.1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{2.2}$$

$$\frac{dR}{dt} = \gamma I \tag{2.3}$$

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.

The infectiousness of a particular disease is often characterized by the basic reproductive ratio, R_0, or the average number of secondary cases arising from an average primary case in an entirely susceptible population [1]. R_0 measures the maximum reproductive potential for an infectious disease in a particular host population, and is calculated by multiplying the transmission rate (β) by the average infectious period ($\frac{1}{\gamma}$) . In other words, $R_0 = \frac{\beta}{\gamma}$.

In order for an epidemic to take hold in a population, the initial proportion of susceptibles must be greater than $\frac{1}{R_0}$. If $S(0)$ is less than $\frac{1}{R_0}$, then $\frac{dI}{dt} < 0$ and the infection will die out [1]. In other words, in a fully susceptible population, an epidemic can only invade if $R_0 > 1$. In the case where an epidemic burns out, the chain of transmission eventually breaks due to a decline in infecteds and not due to a complete lack of susceptibles. This means there will always be some susceptibles in a population who avoid infection.

2.1.0.0.1 Simulated Example

To produce an example of what this model looks like in action, I simulated the spread of measles through a boarding school. Since a boarding school is a closed population, it is reasonable to assume that no one is being born, no one is dying, and no one is migrating into or out of the population during the epidemic.

I took the values of the parameters and the initial conditions from Keeling and Rohani (2011). Since measles has an infectious period of 2 weeks, γ is $\frac{1}{2}$. The transmission rate, β, in this example is 8 new cases per infected individual per week. I want to take a moment at this point to note that the transmission rate is scalable to a time period. Typically, it is scaled to the infectious period. However, it would be just as correct to say that the β for this example is 16 new cases per infected individual per biweek, or 1.14 new cases per infected individual per day. In a population of 763 students, at the start of the epidemic 3 were infected and the rest were assumed to be 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 weeks. I outputted a data frame that included the value of S, I, and R for each time step through completion. Figure 2.1 shows a plot of the proportion of each infection class over time.

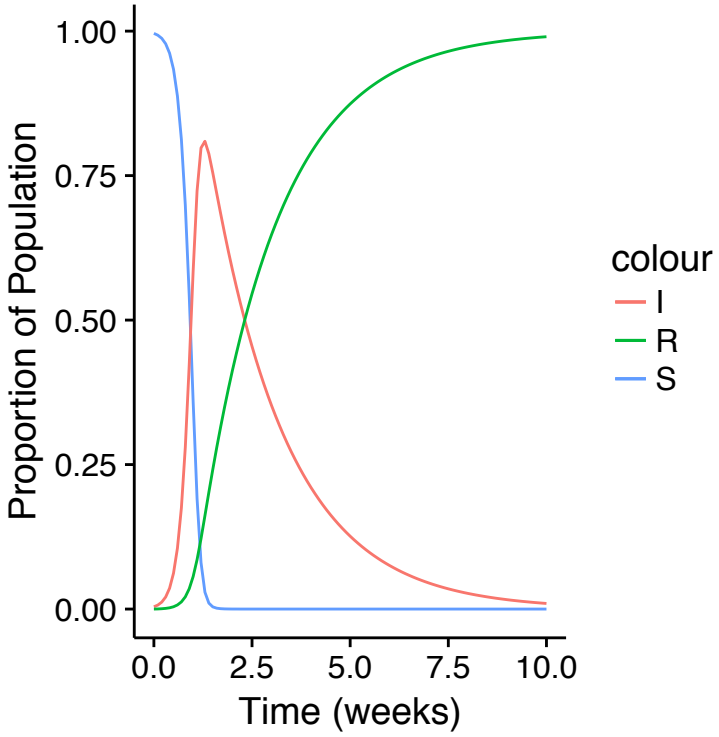

Figure 2.1: Deterministic SIR Simulation

2.1.0.0.2 Inference

I then proceeded to use my simulated data to retrieve the transmission rate. For this first example, I chose to minimize the sum of squared errors of the infected class dynamics generated by a given beta versus the infected dynamics from the simulated data. I did not retrieve the value of γ, the recovery rate, as this parameter is known for most diseases and is most accurately obtained via laboratory studies [10].

To do this, I first created a function that simulated data for a series of β values ranging from 5 to 15 with a step of 0.01, maintaining the same initial conditions and γ value as the initial simulation. I chose the value range for β based on the fact that the R_0 for measles is typically between 15 and 20. Since the infectious period is 2 weeks, that means that β is expected to be between 7.5 and 10.

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 2.2. It is visually evident that the minimum of the function occurs around 8, the 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 8.

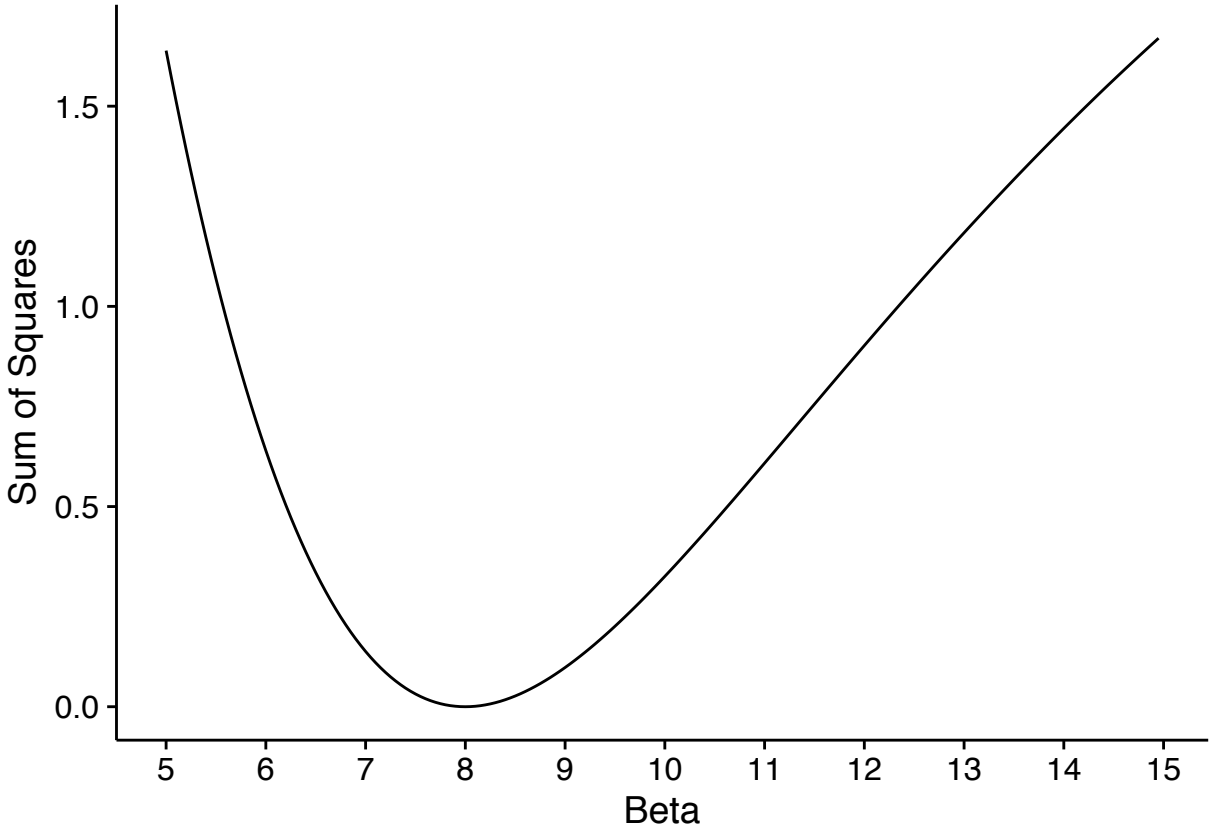

Figure 2.2: Plot of estimated beta values against their sum of squares output

2.1.1 Equilibrium States

Equilibrium occurs when $\frac{dS}{dt} = \frac{dI}{dt} = \frac{dR}{dt} = 0$. The only way for SIR diseases in closed populations to reach equilibrium is to have the epidemic die out, as occurred in the previous example. This is due to the fact that individuals leave the susceptible class as they become infected, but no individuals enter the susceptible class. This causes the number of susceptibles to drop too low for the epidemic to persist.

In order to have a sustained epidemic, a population must have an influx of new susceptible individuals. In an open population, the susceptible class is replenished through births. Deaths may also affect the number of individuals in each class. In order to include demographic information in the basic SIR model given in equations 2.1 through 2.3, allow the population birth rate to be given by μ. If we make the simplifying assumption that the population death rate is equal to the population birth rate, equations 2.1 through 2.3 can be modified to include demographics as follows:

$$\frac{dS}{dt} = \mu - \beta SI - \mu S \tag{2.4}$$

$$\frac{dI}{dt} = \beta SI - \gamma I - \mu I \tag{2.5}$$

$$\frac{dR}{dt} = \gamma I - \mu R \tag{2.6}$$

Note that all newborns are assumed to enter the susceptible class, and that it is assumed that the death rate is the same for the susceptible, infected, and recovered classes.

Now it is possible for an endemic equilibrium to occur. That is, it is now possible for $\frac{dS}{dt} = \frac{dI}{dt} = \frac{dR}{dt} = 0$ without the proportion of infecteds dropping to 0. At endemic equilibrium, the proportion of susceptibles in the population stabilizes at $\frac{1}{R_0}$ and the proportion of infecteds in the population stabilizes at $\frac{\mu}{\beta}(R_0 - 1)$. In other words, if we allow S^*, I^*, and R^* to be the proportion of susceptibles, infecteds, and recovered individuals at endemic equilibrium, then the state of endemic equilibrium can be given by:

$$(S^*, I^*, R^*) = (\frac{1}{R_0}, \frac{\mu}{\beta}(R_0 - 1), 1 - \frac{1}{R_0} - \frac{\mu}{\beta}(R_0 - 1)) \tag{2.7}$$

In the SIR model with demographics, the endemic equilibrium is stable if $R_0 > 1$ and the disease-free equilibrium is stable if $R_0 \leq 1$ [1].

Figure 2.3 shows an example of the infected class dynamics of a deterministic SIR simulation as it approaches a state of endemic equilibrium. For this simulation, $\beta = 8$, $\gamma = \frac{1}{2}$, and $\mu = 5.49 \times 10^{-4}$. The initial conditions were chosen using the definition of endemic equilibrium given in equation 2.7 and are as follows: $(S(0), I(0), R(0)) = (6.25 \times 10^{-2}, 1.03 \times 10^{-3}, 9.36 \times 10^{-1})$. It is graphically evident in figure 2.3 that as the epidemic progresses, the fraction of infectious individuals displays damped oscillatory behavior as it settles towards equilibrium.

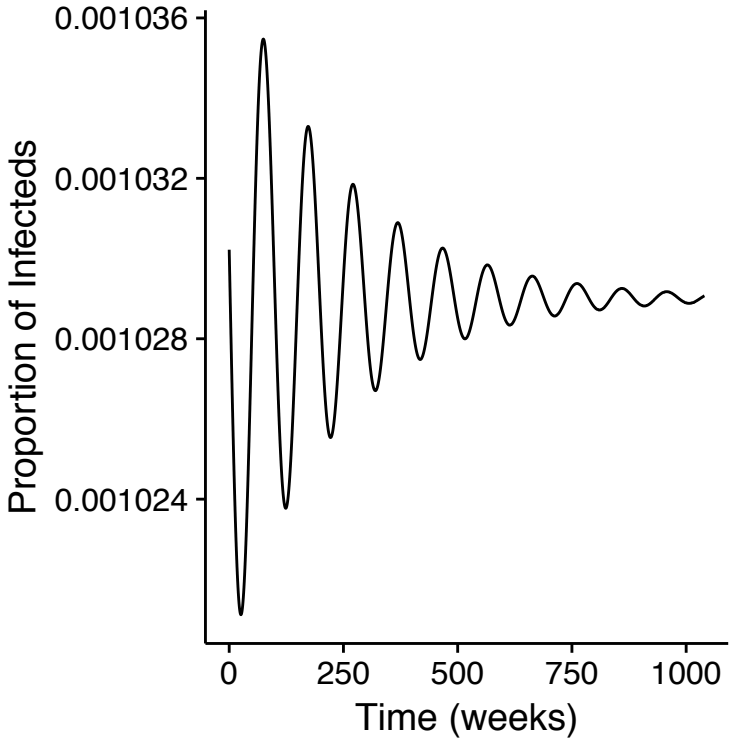

Figure 2.3: Example of the infected class dynamics of a deterministic SIR simulation at endemic equilibrium, simulated over a 20 year period.

2.2 Stochastic

While deterministic models provide a useful foundation for working with infectious disease dynamics, they fail to capture the inherent stochasticity in the infection process. The number of individuals who become infected at any given time is most accurately described as a random process. The probability that a given susceptible becomes infected depends on two random events: first, the susceptible individual must come in contact with an infected individual. Second, the susceptible individual must actually contract the disease, which is not guaranteed given contact with an infected individual. This inherent randomness in the infection process, in addition to natural variation between individuals in the infectious period, make the application of stochastic modeling to the study of infectious disease dynamics a natural choice.

I will discuss the chain binomial model and the TSIR model [9] in this chapter. However, other researchers have taken alternate approaches to adding stochasticity to models of infectious disease dynamics [11].

2.2.1 Chain Binomial

The chain binomial model is a discrete-time stochastic model in which the time step is equal to the infectious period of the disease. Since the time step is equal to the infectious period, it

is assumed that all infected individuals move into the recovered class at the end of each time step.The chain binomial model addresses the inherent randomness in the infection process by fitting the infection process to a binomial distribution.

In this model, the number of infected individuals at each time step follows a random binomial distribution with n equal to the the number of susceptibles in the previous time step.

In order to explore the logic behind the chain binomial model, allow p to be the probability that contact occurs between a susceptible and a single infected individual and that this contact leads to infection. It follows that the probability of a susceptible individual escaping infection from one infected is given by $(1-p)$. In order for a susceptible individual to escape infection entirely in a given time step, they must avoid infection from all infected individuals in the population during that time step. In this way, the probability of a susceptible escaping infection in a given time step is given by $(1-p)^{I_t}$. In this case, the probability that a given susceptible will become infected in a given time step is given by $1-(1-p)^{I_t}$. It follows that the total number of susceptible individuals to become infected during a given time step follows a binomial distribution with $n = S_t$ and a probability of success equal to $1-(1-p)^{I_t}$. Since the chain binomial model operates under the assumption that all infected individuals move to the recovered class at the end of each time step, it also follows that the number of infected individuals in a given time step follows a binomial distribution with $n = S_{t-1}$ and a probability of success equal to $1-(1-p)^{I_{t-1}}$. In other words, $I_t \sim Binomial(S_{t-1}, 1-(1-p)^{I_{t-1}})$.

This definition of the chain binomial model, while accurate, is not very useful if the goal is to learn about the transmission rate and R_0. Therefore, it is necessary to re-parameterize the probability of success to include β in some capacity. A more useful parameterization of the chain binomial is given by $I_t \sim Binomial(S_{t-1}, 1-exp(\frac{-\beta I_{t-1}}{N}))$, where N is the population size. The probability of success for the distribution was chosen such that the expected value of infected individuals, $E(I_t)$, is equal to $\beta I_{t-1} S_{t-1} N^{-1}$.

The full set of equations that define the chain binomial model (without demography) are as follows:

$$I_t \sim Binomial(S_{t-1}, 1-exp(\frac{-\beta I_{t-1}}{N})) \tag{2.8}$$

$$S_t = S_{t-1} - I_t \tag{2.9}$$

$$R_t = R_{t-1} + I_{t-1} \tag{2.10}$$

Adding demography to equations 2.8 through 2.10, and allowing B_t to represent the number of births in time step t, yields:

$$I_t \sim Binomial(S_{t-1}, 1-exp(\frac{-\beta I_{t-1}}{N})) \tag{2.11}$$

$$S_t = B_t + S_{t-1} - I_t \tag{2.12}$$

$$R_t = R_{t-1} + I_{t-1} \tag{2.13}$$

2.2.1.0.1 Simulated Example

I simulated an SIR infection using the chain binomial model defined in equations 2.8 through 2.10. For consistency and comparison, I used the parameters and starting conditions that I used to simulate the measles example from the section on deterministic models. That is, $\beta = 8$, $\gamma = \frac{1}{2}$, $N = 763$, $I(0) = 3$, $S(0) = 760$, and $R(0) = 0$.

The sharp points observed in figure 2.4 are due to the large time steps compared to the relatively small number of data points included in the simulation.

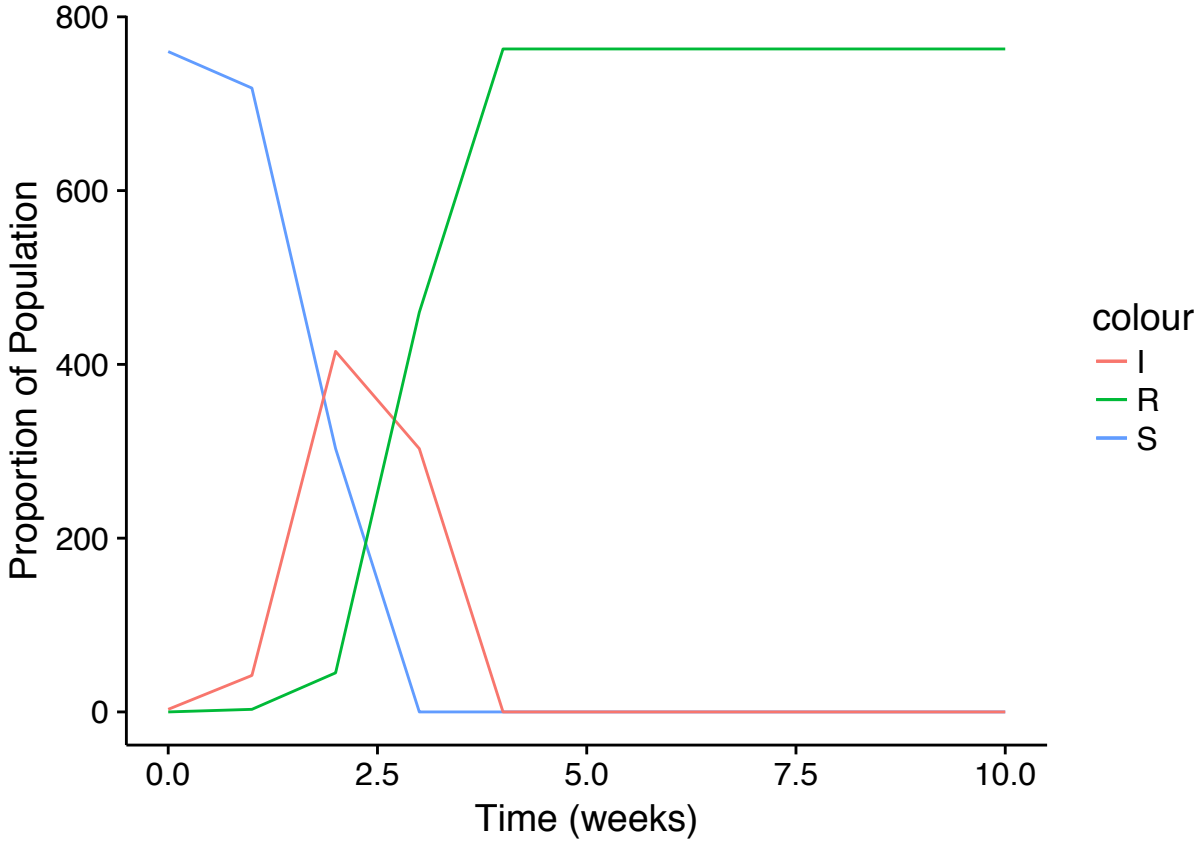

Figure 2.4: Chain Binomial Simulation

2.2.1.0.2 MLE Inference:

I used maximum likelihood estimation with the data simulated from the chain binomial model defined in equations 2.8 through 2.10 to retrieve the value of β. To do this, I first found the likelihood function for β, which is given by equations 2.14 and 2.15 below.

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \binom{I_{i-1}}{S_{I-1}} p^{I_i}(1-p)^{S_i} \ , \ where \ p = 1 - exp(\frac{-\beta I_{i-1}}{N}) \tag{2.14}$$

This can be simplified to:

$$\mathcal{L}(\beta) \propto p^{I_{[i]}}(1-p)^{S_{[i]}} \tag{2.15}$$

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 16.00, which is equal to the value I simulated with, $\beta = 8.00 * 2 = 16.00$.

2.2.1.0.3 Bayesian Inference:

Recall that the Bayesian approach involves defining a prior distribution, $f(\theta)$, and then using equation 1.4 and MCMC to obtain the posterior distrubtion of the model parameters.

I began by choosing a prior distribution for β. When choosing the prior distribution, I considered that for measles, R_0 is typically between 12 and 18 [1], which means that, since my model is scaled such that the time step is equal to one infectious period, I am expecting a β in approximately range. Therefore, I chose a gamma distribution with shape parameter of 10 and scale parameter of 1.5. This distribution has an expectation of 15, and the majority of its density lies between 5 and 25, which is a reasonably broad range within which to expect β.

I used the package RStan to run Bayesian inference on my simulated data with the MCMC method, with the aforementioned priors.

The outputted estimate for β was 16.0, with a standard error of 0. Recall that the true value of β was 16.0. This level of accuracy is possible only with data simulated without any noise.

2.2.2 TSIR

The TSIR (Time Series Susceptible-Infected Removed) model is a time series, discrete time, stochastic model based on the basic SIR model [9]. 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:

$$E(I_{t+1}) = \beta I_t S_t N^{-1} \tag{2.16}$$

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

$$I_{t+1} \sim NB(E(I_{t+1}), I_t) \tag{2.17}$$

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.

2.2.2.1 Temporally Varying Transmission Rates

Up to this point, all of the given 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 sinusoidal function given in equation 2.18 below:

$$\beta(t) = \beta_0(1 + \beta_1 cos(\omega t)) \tag{2.18}$$

The parameter β_0 in equation 2.18 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, I 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 2.19 and 2.20 give the model specifications.

$$I_t = \beta S_{t-1} I_{t-1} \tag{2.19}$$

$$S_t = B_{t-d} + S_{t-1} - I_t \tag{2.20}$$

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 under reported. Additionally, I assume that ρ_t follows a probability distribution with $E(\rho_t) = \rho$.

$$I_t = \rho_t C_t \tag{2.21}$$

Substituting equation 2.21 into equation 2.20, we get:

$$S_t = B_{t-d} + S_{t-1} - \rho_t C_t \tag{2.22}$$

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:

$$Z_t = B_{t-d} + Z_{t-1} - \rho_t C_t \tag{2.23}$$

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

$$Z_t = Z_0 + \sum_{i=1}^{t} B_{i-d} - \sum_{i=1}^{t} \rho_i C_i \tag{2.24}$$

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 as a simple linear regression equation:

$$Y_t = -Z_0 + Z_t + \rho X_t \tag{2.25}$$

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.

Chapter 3

Application to Measles

For my analysis, I chose to work with pre-vaccination measles data in major United State cites over a 20-year period from 1920 through 1940. I chose to study measles due to the wealth of data available on the measles epidemic, in addition to the impact of the measles virus on public health across the globe. During the pre-vaccination era (prior to the 1960's), measles was a major cause of infant and child mortality. Additionally, although measles incidence in developed countries was greatly reduced by the start of mass vaccination in the 1960s [12] [13], measles is still a significant cause of child mortality for developing countries [14]. Measles outbreaks in developed countries also remain a significant public health concern [15][16].

I obtained my incidence data from the Tycho database [17], and my population and birth data I obtained from the dataset appended to a 2016 paper by Dalziel et al [18]. For each city, I performed susceptible reconstruction and fit a model with 26 seasonal β's (to model the time-varying nature of the transmission rate). In this chapter, I will explain in depth the process of performing susceptible reconstruction when the assumption of a constant reporting rate is violated, in addition to discussing the characteristics of the measles epidemic in various US cities during the pre-vaccination era.

3.1 The Dataset

I used the dataset given in the supplementary materials supplied by Dalziel et al. from their February 2016 publication in PLOS Computational Biology [18] [19]. The dataset contains biweekly measles incidence, population totals, births, and spatial coordinates for 40 United States cities between 1920 and 1940. Dalziel et al. obtained the incidence data from the Tycho database [17]. The researchers obtained decennial population data from the U.S. census, and then used a spline function at each biweek to reconstruct the biweekly total and infant population.

I chose this dataset over the widely studied England and Wales dataset, which contains data for 952 cities in England and Wales on biweekly measles incidence, population demographics,

and spatial information [20]. The England and Wales dataset is worthy of mention due to its importance and prevalence in measles research.

There were some missing values in the dataset, as well as places where the reported number of cases was equal to zero (indicating a break in the chain of transmission). In the case of missing values, I extrapolated the missing incidence datapoint by taking the average of the reported cases immediatel before and after the missing datapoint. In the case of die outs, I re-coded all incidences of $C_t = 0$ with $C_t = 0.5$. It was necessary to re-code die outs as the TSIR model with no spatial compenent has no way of accounting for recruitment into the infected class via migration.

3.2 Methods

My goal was to fit a TSIR model for the dynamics of the measles epidemic in a sample of 5 of the cities contained in the Dalziel et al. dataset. I worked with New York City, Chicago, Indianapolis, New Orleans, and Spokane. I chose these cities based on their population sizes; New York City and Chicago are large cities (in the fourth quartile of population size), Indianapolis and New Orleans are medium cities (in the third quartile), and Spokane is a small city (in the first quartile).

An outline of the workflow for fitting a TSIR model to data is given in figure 3.1. The dataset contained biweekly data on the number of reported measles cases ($\{C_t\}$), births, and total population for each city. In order to fit the TSIR model, the susceptible class dynamics ($\{S_t\}$), the reporting rate (ρ), and the vector of transmission rates ($\{\beta_t\}$) must be obtained. The first step in this process is to complete susceptible reconstruction, through which ($\{Z_t\}$), the deviations from the mean number of susceptibles, and (ρ) can be obtained. The infected class dynamics are also obtained in this step, as they can be computed by dividing the number of reported cases by the reporting rate. The final step in the process of fitting the model is to pass the population data with the recovered susceptible and infected dynamics through Bayesian Inference, through which a probability distribution is obtained for each of the 26 $\{\beta_t\}$'s and $\bar{S}$. Once all of the above values are obtained, the model has been fit and is ready to be validated (or else rejected as a poor fit).

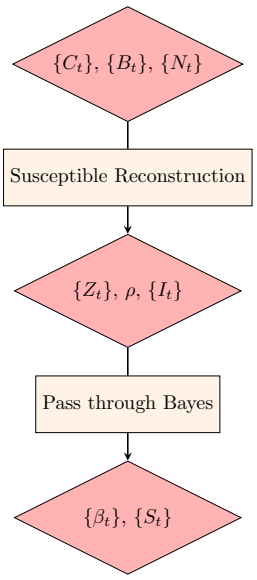

Figure 3.1: Workflow for fitting the TSIR model.

3.2.1 The Model

I fit a TSIR (time-series-susceptible-infected-recovered) model to the measles data. Measles follows a 2 week infectious period, so model had a 2-week time step. In this way, it can be assumed in biweek t that all infected individuals from biweek $t-1$ have recovered, and that all infected individuals in biweek t can be traced to a contact between an infected and a susceptible individual in the previous time step.

As previously discussed, the TSIR model is defined as follows:

$$I_{t+1} \sim NB(E(I_{t+1}), I_t) \tag{3.1}$$

$$S_{t+1} = S_t + B_t - It + 1 \tag{3.2}$$

In which $NB(a, b)$ indicates the negative binomial distribution with expected value a and clumping parameter b. Additionally, the expected value of the infected class can be expressed as:

$$E(I_{t+1}) = \beta I_t S_t N^{-1} \tag{3.3}$$

3.2.2 Susceptible Reconstruction

The next step in the process of fitting a model to the data is performing susceptible reconstruction to recover the susceptible and infected dynamics. Following the process described

in the Susceptible Reconstruction section of Chapter 2, I performed a simple linear regression on the cumulative births, $\sum_{i=1}^{t} B_i$ and the cumulative cases, $\sum_{i=1}^{t} C_i$, and fit it to the model in equation 3.4. For this regression, the response is the cumulative births and the predictor is the cumulative cases. Note that it is possible to switch the response and the predictor of this regression [18].

$$\sum_{i=1}^{t} B_i = -Z_0 + Z_t + \rho \sum_{i=1}^{t} C_i \tag{3.4}$$

However, when I looked at the regression residuals (the recovered $\{Z_t\}$ dynamics), I noticed places where the mean drifted away from zero. Plot (A) of figure 3.2 shows the output from the linear regression, in which the local shifts from the mean are visible. A potential problem pointed out by other researchers [21] was the strong assumption of a constant reporting rate. To handle situations where the reporting rate is not constant, a local regression must be used in place of a simple linear regression during susceptible reconstruction.

3.2.3 Handling Time-Varrying Reporting Rates

In real data, it is often the case the reporting rate is found to vary over time. As previously mentioned, when this occurs it is necessary to compensate by performing local linear regression in place of global linear regression during the susceptible reconstruction process.

As an example, let's look at measles data from pre-vaccination era New York City, from 1920-1940 [17]. In figure 3.2, 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 3.2 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 3.2, it indicates a non-constant reporting rate. Following the work of previous measles researchers [21], 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 data points 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 [21]. 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 Gaussian 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 smoothing parameter h.

$$SSE_1(h) = \sum_{t=1}^{T} \{Y_t - \hat{m}_{h,t}(x_t)\}^2 \tag{3.5}$$

$$SSE_2(h) = \sum_{t=1}^{T} \{\hat{Y}_t - \hat{m}_{h,t}(x_t)\}^2 \tag{3.6}$$

In figure 3.2 I reference data from the Tycho database [17] for pre-vaccination measles 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 [21], I used a Gaussian weight function, defined by the following:

$$K(x) = \frac{1}{\sqrt{2 * pi}} e^{\frac{-x^2}{2}} \tag{3.7}$$

$$w_i(x_t) = \frac{K(\frac{x_0 - x_i}{h})}{\sum_{i=1}^{m} K(\frac{x_0 - x_i}{h})} \tag{3.8}$$

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 data point 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 data point 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 inputting 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.

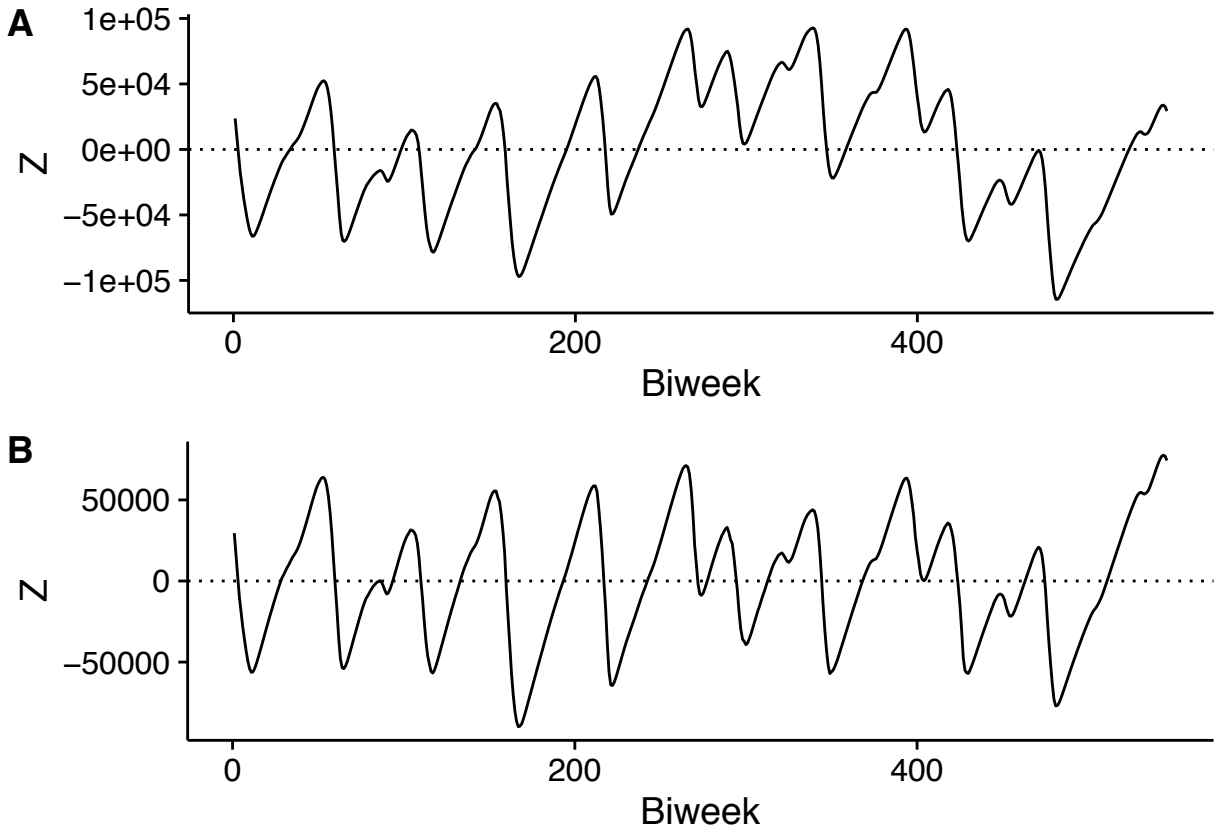

Figure 3.2: Comparison of reconstructed Z dynamics from the New York data. Plot (A) shows Z dynamics obtained via global linear regression. Plot (B) shows Z dynamics obtained via local linear regression.

Figure 3.2 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.

3.3 Model Fit

The final step in constructing a model to fit the epidemic is to use Bayesian inference to obtain estimates for the transmission rate and $\bar{S}$. The overall patterns of seasonality is qualitatively similar to those estimated by previous mealses researchers [18].

Estimates for the transmission rate of each city can be found in figure 3.3. The appendix also contains estimates of the transmission rate terms, in addition to $\bar{S}$.

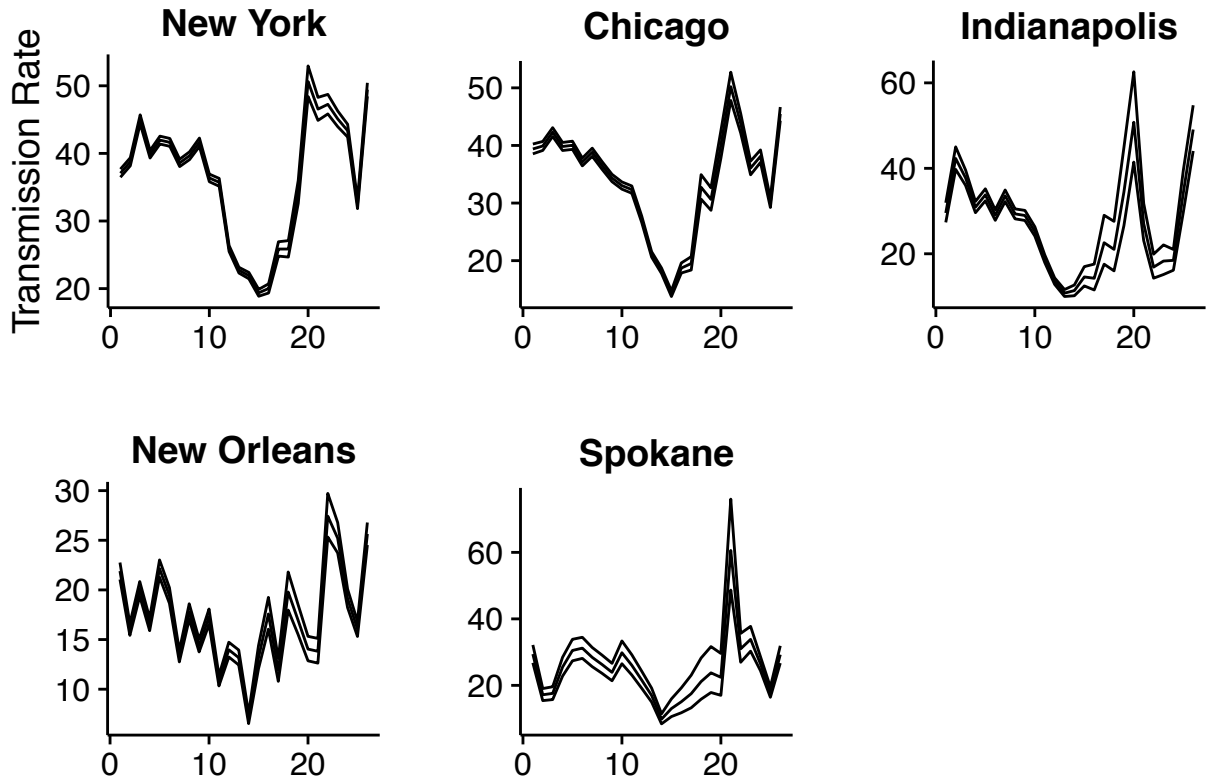

Figure 3.3: Estimates of the seasonal transmission rates for measles (median-centered) within a confidence band that delineates the 95% credible interval.

Now the model is complete. It is important, at this point, to ensure that the model is a

good fit for the data.

To this end, I simulated from my model for each of the cities I studied. In order to do this, I had to consider two important components of the model: first, I needed to select the initial proportion of susceptibles and infecteds. Second, I needed to take into account the dependencies of $\{\beta_{1...26}\}$. Table 4.3 gives the median of the posterior distribution of each β_t. This is a useful way to get information about the transmission rate of the epidemic, but for simulation it is necessary to consider that the β_t's are dependent on one another. For this reason, I simulated using the entire Markov chain for $\{\beta_{1...26}\}$ (producing thousands of curves) and then selected the median number of infecteds at each time point to obtain the epidemic curve displayed in figure 3.4.

I next had to consider how to chose the initial proportion of susceptibles and infecteds ($S(0)$ and $I(0)$) to use when simulating. While the obvious choice would be to use the values obtained from reconstruction, the simulations are highly influenced by small deviations in the initial conditions. As the simulations were sensitive to the error in my reconstructed data, I decided to choose $S(0)$ and $I(0)$ 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 [17]. 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 previous work on the same dataset [18]. For each city, the value of $S(0)$ and $I(0)$ that minimized the mean square error of the simulation was used in the epidemic curves produced in figure 3.4.

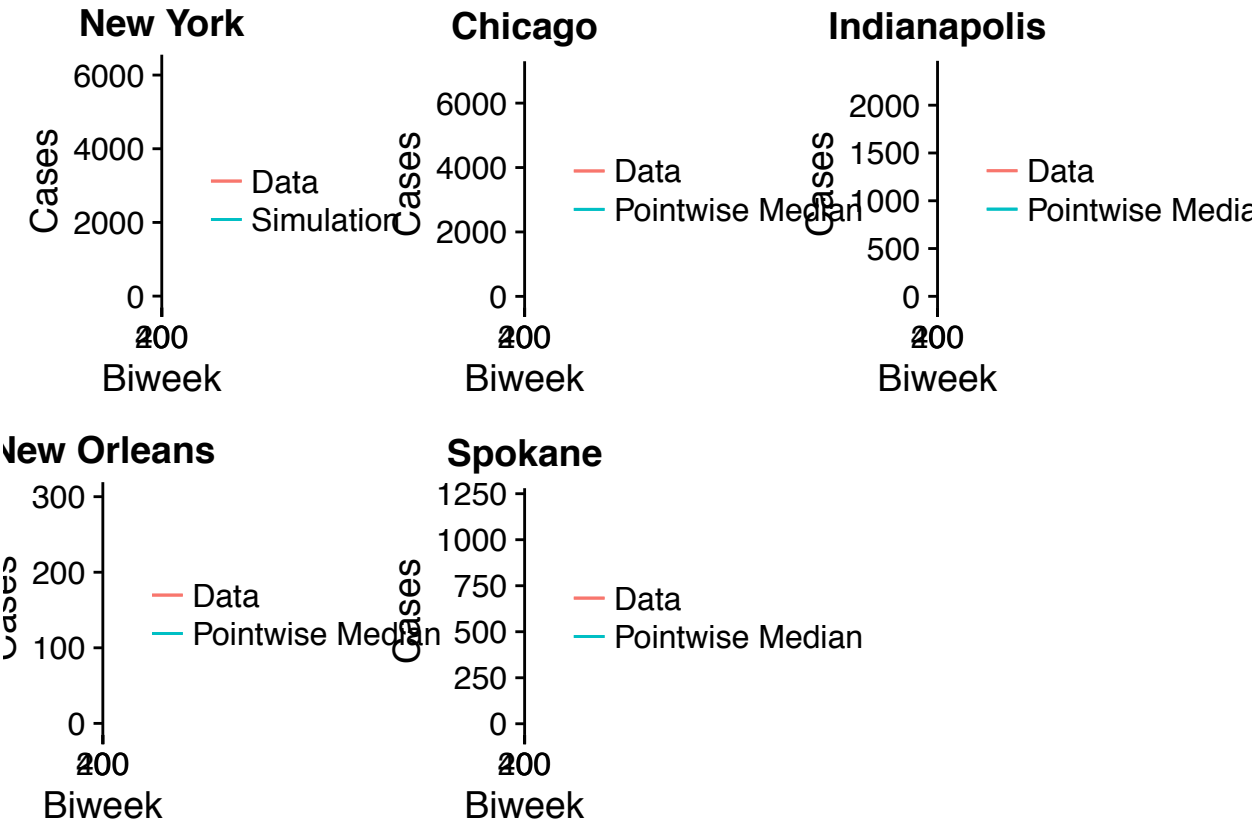

Figure 3.4: Plots show the successful forward simulation from the deterministic skeleton of the TSIR model parameterized with city-level data (in blue) versus the incidence data taken from tycho (in red). Simulations were started from the initial conditions obtained by minimizing the mean square error and were not fed any additional information from the incidence data once the simulation begun. Simulations were carried out for 20 years.

Chapter 4

Conclusions

There is still much to be learned and refined about constructing models for measles dynamics. Some key areas for further consideration include the validity of the reporting rate and the sensitivity of the TSIR model to small changes in the initial conditions. ##Caveats and Limitations

4.0.0.0.1 Reporting Rates While many measles researchers have written about the existence of time-varying reporting rates, there is very little information about how or why reporting rates vary [21] [18].

The evidence for non-constant reporting rates comes from the susceptible reconstruction process, in which the susceptible dynamics are derived from the regression in equation 2.25. In equation 2.25, the reporting rate, ρ, is the slope of the regression, and the susceptible dynamics, Z_t, are the regression residuals. When the reporting rate is not constant, the regression residuals display local shifts away from the mean of zero. Plot A of figure 3.2 shows an example of Z_t dynamics obtained from a dataset in which ρ is non-constant.

As previously discussed, non-constant reporting rates require the use of localized linear regression when reconstructing the susceptible class dynamics. The result of such measures is a vector of reporting rates, $\rho_t \in \{1, ..., T\}$. The reporting rates are then used to reconstruct the infected class dynamics.

In published measles research, very little mention is given to the reporting rate beyond its use for reconstruction. Most researchers state the average reporting rate for a dataset, and do not mention it further. It is possible that the reason for this is that the reporting rate is somewhat of a fudge factor, accounting for some inadequacies in the model.

Here, I will explore in more detail the dynamics of ρ_t that were obtained during the susceptible reconstruction process. Measles researchers who used the same dataset reported a mean reporting rate of 0.37 and standard deviation of 0.13 [18]. Note that these values are equivalent to $\frac{1}{\rho_t}$, and were obtained using a sample of 40 United States cities, whereas I worked with a subset of 5 cities.

For the 5 cities I worked with (New York, Chicago, Indianapolis, New Orleans, and Spokane), I obtained a mean reporting rate of 0.33 and a standard deviation of 0.24. Figure 4.1 shows a plot of the time-varying reporting rate for the cities I studied.

I note that the a major dip in the reporting rates occurs between the 150^{th} and the 250^{th} biweek, or somewhere between 1927 and 1930. This dip appears to coincide with the Great Depression. However, it is unclear whetheher reporting rates were truly much lower during that time interval, or if reporting rates are compensating for an inadquqacy in the model. Further research in this area is necessary to provide further insights.

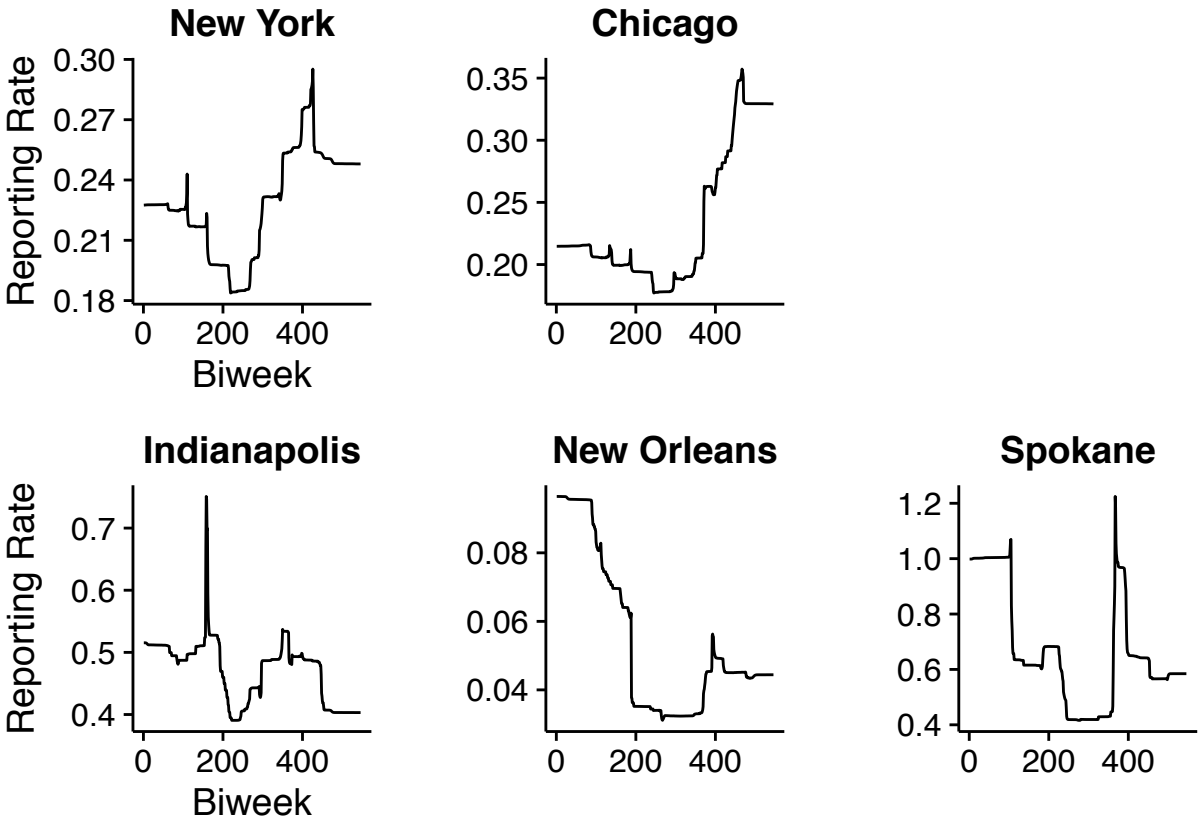

Figure 4.1: Plots of the reporting rates for New York, Chicago, and Indianapolis. A reporting rate of 1 indicates full reporting, a reporting rate below 1 indicates underreporting, and a reporting rate over 1 indicates overreporting.

4.0.0.0.2 Initial Condition Sensitivity

Forward simulations from the fitted model are highly sensitive to small changes in the initial values of $S(0)$ and $I(0)$. As there is high uncertainty in estimating these initial conditions, this sensitivity is a cause for concern in regards to the adequacy of th model.

Previous measles researchers studying the measles epidemic in pre-vaccination United States found that cities in the United States are much more sensitive to such small changes in initial conditions than are UK cities [18]. It was postulated that a systematic difference in

transmission rates between the United States and the UK may drive such differences between the epidemic dynamics of the two countires. However, the cause of systematic differences in transmission is unclear.

Further research into the factors driving country-level differences in epidemic patterns, as well as the cause of initial condition sensitivity in the populations that display them, is recommended.

4.0.0.0.3 Effect of Population Size

I chose to work with cities with a broad range of population sizes in order to assess the effect of population size on epidemic dynamics and model fit. A breakdown of the populations of the cities I studied, along with some additional information about the data, is given in table 4.1.

The two largest cities, New York and Chicago, had very similar reporting rates in addition to seasonal transmission rates. There was a strong correlation between the reporting rates and transmission terms of the two cities ($r = 0.744$ and $r = 0.917$ respectively). Other cities studied were not as strongly correlated. It is unclear what drives this relationship.

Smaller cities, such as Spokane, had many more missing values and die-outs than larger cities, such as New York or Chicago. As the model used did not have any spatial components, which would have allowed for recruitment into the infected class via migration, it was necessary to re-code die-outs as having a single infected. Missing values were estimated by taking the mean of the timepoints on either side of the missing biweek. In addtion to smaller population sizes, the data were less complete in smaller cities. The model fits for such cities were also, therefore, less accurate than those for larger cities. It is likely that better methods for handling missing values and die outs are available.

City	Population	Missing.Values	Die.Outs
New York	5,620,048	0	0
Chicago	2,701,705	1	1
New Orleans	387,219	27	122
Indianapolis	314,194	2	32
Spokane	104,437	29	95

Table 4.1: Breakdown of the dataset by city. The population given is the initial population. Missing values gives the number of biweeks for which incidence data was not available. The die outs columns gives number of biweeks in which the reported number of cases was zero.

Appendix

Parameters	Estimate	Standard.Error
β_1	37.07	0.31
β_2	38.75	0.30
β_3	45.04	0.33
β_4	39.88	0.29
β_5	41.97	0.30
β_6	41.63	0.29
β_7	38.59	0.28
β_8	39.66	0.29
β_9	41.64	0.32
β_{10}	36.37	0.29
β_{11}	35.72	0.29
β_{12}	25.93	0.23
β_{13}	22.72	0.22
β_{14}	21.92	0.24
β_{15}	19.38	0.26
β_{16}	20.01	0.35
β_{17}	25.85	0.54
β_{18}	25.85	0.62
β_{19}	34.25	0.87
β_{20}	50.62	1.16
β_{21}	46.55	0.87
β_{22}	47.26	0.74
β_{23}	45.09	0.59
β_{24}	43.36	0.48
β_{25}	32.48	0.33
β_{26}	49.43	0.51
$\bar{S}$	224642.73	1448.51

Table 4.2: Beta Estimates for the measles outbreak in New York City 1920-1940. The given estimates are the median of the posterior distribution obtained via Bayesian inference.

Parameters	Estimate	Standard.Error
β_1	39.39	0.43
β_2	39.91	0.40
β_3	42.32	0.39
β_4	39.82	0.36
β_5	40.03	0.36
β_6	37.13	0.34
β_7	38.82	0.36
β_8	36.51	0.34
β_9	34.35	0.33
β_{10}	33.04	0.32
β_{11}	32.33	0.32
β_{12}	27.19	0.28
β_{13}	21.08	0.24
β_{14}	18.24	0.24
β_{15}	14.27	0.25
β_{16}	18.71	0.44
β_{17}	19.50	0.59
β_{18}	32.74	1.08
β_{19}	30.64	1.00
β_{20}	40.09	1.21
β_{21}	50.20	1.25
β_{22}	43.81	0.86
β_{23}	36.07	0.61
β_{24}	38.12	0.55
β_{25}	30.02	0.40
β_{26}	45.47	0.60
$\bar{S}$	115302.72	879.52

Table 4.3: Beta Estimates for the measles outbreak in Chicago 1920-1940. The given estimates are the median of the posterior distribution obtained via Bayesian inference.

Parameters	Estimate	Standard.Error
β_1	21.90	0.43
β_2	16.01	0.30
β_3	20.10	0.38
β_4	16.52	0.32
β_5	22.14	0.44
β_6	19.43	0.40
β_7	13.31	0.29
β_8	17.81	0.40
β_9	14.38	0.32
β_{10}	17.29	0.40
β_{11}	10.85	0.27
β_{12}	13.97	0.38
β_{13}	13.18	0.38
β_{14}	7.03	0.26
β_{15}	13.18	0.59
β_{16}	17.57	0.82
β_{17}	11.89	0.59
β_{18}	19.79	0.98
β_{19}	16.95	0.77
β_{20}	14.04	0.63
β_{21}	13.82	0.63
β_{22}	27.43	1.12
β_{23}	25.21	0.78
β_{24}	19.14	0.48
β_{25}	16.05	0.39
β_{26}	25.64	0.58
$\bar{S}$	30690.66	495.37

Table 4.4: Beta Estimates for the measles outbreak in New Orleans 1920-1940. The given estimates are the median of the posterior distribution obtained via Bayesian inference.

Parameters	Estimate	Standard.Error
β_1	29.55	1.17
β_2	42.31	1.34
β_3	37.77	0.92
β_4	30.92	0.67
β_5	33.79	0.71
β_6	29.04	0.63
β_7	33.56	0.69
β_8	29.34	0.59
β_9	28.96	0.60
β_{10}	25.22	0.56
β_{11}	18.85	0.46
β_{12}	13.59	0.39
β_{13}	10.79	0.41
β_{14}	11.39	0.64
β_{15}	14.58	1.15
β_{16}	14.27	1.54
β_{17}	22.59	2.93
β_{18}	21.03	2.96
β_{19}	34.44	4.68
β_{20}	50.78	5.40
β_{21}	27.08	2.21
β_{22}	16.88	1.44
β_{23}	18.30	1.77
β_{24}	18.46	1.24
β_{25}	34.22	2.50
β_{26}	49.08	2.73
$\bar{S}$	15640.30	253.74

Table 4.5: Beta Estimates for the measles outbreak in Indianapolis 1920-1940. The given estimates are the median of the posterior distribution obtained via Bayesian inference.

Parameters	Estimate	Standard.Error
β_1	21.90	0.43
β_2	16.01	0.30
β_3	20.10	0.38
β_4	16.52	0.32
β_5	22.14	0.44
β_6	19.43	0.40
β_7	13.31	0.29
β_8	17.81	0.40
β_9	14.38	0.32
β_{10}	17.29	0.40
β_{11}	10.85	0.27
β_{12}	13.97	0.38
β_{13}	13.18	0.38
β_{14}	7.03	0.26
β_{15}	13.18	0.59
β_{16}	17.57	0.82
β_{17}	11.89	0.59
β_{18}	19.79	0.98
β_{19}	16.95	0.77
β_{20}	14.04	0.63
β_{21}	13.82	0.63
β_{22}	27.43	1.12
β_{23}	25.21	0.78
β_{24}	19.14	0.48
β_{25}	16.05	0.39
β_{26}	25.64	0.58
$\bar{S}$	5188.95	207.95

Table 4.6: Beta Estimates for the measles outbreak in Spokane 1920-1940. The given estimates are the median of the posterior distribution obtained via Bayesian inference.

References

[1] Keeling MJ, Rohani P. Modeling infectious diseases in humans and animals. Princeton University Press; 2011.

[2] Dobson AP, Carper ER. Infectious diseases and human population history. BioScience. 1996;46:115–126.

[3] Burnet SM, White DO. Natural history of infectious disease. Fourth. Cambridge University Press; 1972.

[4] Wood SN. Core statistics. Cambridge University Press; 2015.

[5] Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association. 1990;85:398–409.

[6] Brooks S, Gelman A, Jones G, et al. Handbook of markov chain monte carlo [Internet]. CRC press; 2011. Available from: https://scholar.google.de/scholar.bib?q=info:A4xlUax02S0J:scholar.google.com/&output=citation&hl=de&ct=citation&cd=0.

[7] R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013. Available from: http://www.R-project.org/.

[8] Stan Development Team. RStan: The R interface to Stan [Internet]. 2016. Available from: http://mc-stan.org/.

[9] Bjørnstad ON, Finkenstädt B, Grenfell BT. Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series sir model. Ecological Monographs. 2002;72:169–184.

[10] Centers for Disease Control and Prevention. Epidemiology and prevention of vaccine-preventable diseases. 13th ed. Hamborsky J, Kroger A, Wolfe C, editors. Washington D.C.: Public Health Foundation; 2015.

[11] Abbey H. An examination of the reed frost theory of epidemics. Human Biology. 1952;24:201–233.

[12] Anderson R, May R. Infectious diseases of humans: Dynamics and control. Oxford: Oxford University Press; 1991.

[13] Bolker BM, Grenfell BT. Impact of vaccination on the spatial correlation and persistence

of measles dynamics. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 1996;93:12648–12653. Available from: http://www.jstor.org/stable/40673.

[14] Morse D, Oshea M, Hamilton G, et al. Outbreak of measles in a teenage school population - the need to immunize susceptible adolescents. Epidemiology and Infection. 1994;113:355–365.

[15] McLean AR, Anderson RM. Measles in developing countries. part i. Epidemiological Parameters and Patterns. 1988;100:111–133.

[16] McLean A. R., Anderson RM. Measles in developing countries. part ii. the predicted impact of mass vaccination. Epidemiology and Infection. 1988;100:419–442.

[17] Panhuis WG van, Grefenstette J, Jung SY, et al. Contagious diseases in the united states from 1888 to the present. NEJM. 2013;369:2152–2158.

[18] Dalziel BD, Bjørnstad ON, Panhuis WG van, et al. Persistent chaos of measles epidemics in the prevaccination united states caused by a small change in seasonal transmission patterns. Ferguson NM, editor. PLoS Comput Biol. 2016;12.

[19] Dalziel B, Bjornstad O, van W Panhuis, et al. Data from: Persistent chaos of measles epidemics in the prevaccination united states caused by a small change in seasonal transmission patterns. PLOS Computational Biology. Dryad Digital Repository; 2016.

[20] Fine PEM, Clarkson JA. Measles in england and wales: I, an alaysis of factors underlying seasonal patterns. Int. J. Epidem. 1982;11:5–14.

[21] Finkenstädt B, Grenfell BT. Time series modelling of childhood diseases: A dynamical systems approach. Applied Statistics. 2000;49:187–205.

Kaitlyn Stocker

Education:

The Pennsylvania State University, University Park, PA Aug 2017

- B.S. in Psychology, Biological-Evolutionary Sciences Option
- Minor: Statistics
- *Schreyer Honors College* – Top 5% of PSU undergraduates; requires separate admissions process

Related Experience:

Thesis: Infectious Disease Modeling Sept 2016 – Aug 2017

Penn State University Department of Statistics: Dr. Murali Haran

- Simulated the trajectory of infectious diseases in R using models prevalent in current research
- Used methods such as Bayesian inference and maximum likelihood estimation to run inference and retrieve model parameters
- Ran simulations of infectious disease dynamics and performed inference in R

Independent Contractor Aug 2016 – Aug 2017

myON

- Worked with Dr. Matthew Beckman to provide insights about the usage and efficacy of the client's educational software
- Performed data analysis in R to connect student's use of the client's product with increased literacy scores

Math Tutor May 2012 – Present

Independent Tutor

- Worked with students of various ages ranging from 8 to 30, in various math-related topics
- Specialize in elementary math (4^{th} grade level), algebra (pre-algebra through algebra II), and statistics
- Experience totals approximately 1,000 hours of tutoring

Study Abroad Jun. 2015 – Jul. 2015

India: Delhi, Pune, Jaipur, Dahanu

- Spent 1 month abroad in India, spread between time in Delhi, Pune, Jaipur, and a rural village in Dehanu, with a focus on cultural immersion

Intern Sept. 2014 – May 2015

Penn State Center for Public Diplomacy, World in Conversation

- Led an archiving project in which I reorganized and transferred the Center's records from one application to another, using information gathered from surveys and interviews I conducted with the faculty and staff to assess the relevancy of various files
- Worked in a professional office environment completing tasks including data entry, managing the Center's social media, and sending emails on behalf of the Center

Marketing Intern Sept. 2014 – May 2015

Penn State LGBTQA Student Resource Center

- Responsible for running the Center's social media accounts to increase community awareness and outreach
- Responsible for providing Center tours, greeting incoming speakers, and helping to facilitate discussion groups

Related Skills & Coursework:

Programming

R; Markdown; LaTeX; SAS; SPSS; Excel; PowerPoint

Relevant Coursework

Probability; Analysis of Variance; Applied Regression Analysis; Child Psychopathology; Developmental Psychology; Evolutionary Psychology; Genetics; NeuroPsychoogy; Spanish