Mathematical Statistics in Epidemiology:

A Discussion of the Mathematical Concepts Employed in Regression Analyses and Studies of Disease Spread

A Thesis Presented in Partial Fulfillment of the
Bachelor of Science in Mathematics (Dean's Scholars Honors)
Degree

Maike Lynn Morrison

Department of Mathematics
The University of Texas at Austin

May 2020

Note: This document has been abridged for online distribution.

Abstract

Mathematics is essential to understanding the spread and control of disease. This thesis presents three aspects of mathematical epidemiology that I pursued throughout my undergraduate career.

The first chapter introduces a continuous-time Markov model of an epidemic; this is a stochastic model, allowing for randomness as individuals become infected. We discuss the mathematical details of this model and derive its likelihood function–an equation used to estimate how well its parameters fit to the data. We then develop an algorithm to simulate the epidemic process and use COVID-19 case data from two rural counties to estimate maximum-likelihood parameter estimates. We find that the epidemics are spreading at very different rates in these counties.

In the second chapter we explore the consequences of ignoring terrain heterogeneity (i.e. patches of good and bad terrain) when modeling disease spread in a wild animal population. We model disease spread using a Susceptible-Infected (SI) epidemic model and evaluate the effect of simulated terrain patterns on disease dynamics. Our work demonstrates that ignoring terrain patterns causes us to overestimate model parameters critical to predicting a disease's spread. This result suggests that, when modeling epidemic spread through a real population, accounting for terrain heterogeneity is vital. In the mathematical supplement to this chapter, we discuss the history of compartmental epidemic models (such as the SI model) and derive some results from the historic framework.

Finally, in the third chapter we analyze the percentages of conscientious vaccination exemptions (CVE) in Texas elementary and middle schools. We conduct this analysis using beta regression, a method that is generally used to identify the relationship between an array of predictors and a percentage we seek to model. We find that CVE percentages across the state are positively associated with variables such as median income, college educational attainment, and the proportion of the population that reports as ethnically white. We also develop a metric for outbreak risk due to high vaccination exemption rates and identify three metropolitan areas–Austin, Dallas-Fort Worth, and Houston–as potential vaccination exemption "hot spots." In the mathematical supplement to this chapter, we expand upon the mathematics underpinning beta regression methods.

Contents

Background **1**

1 A Non-Linear Stochastic Epidemic **3**

1.1 Abstract . . . 3

1.2 Introduction . . . 3

1.3 Statistical Background . . . 4

1.3.1 Describing an Epidemic as a Continuous-Time Markov Process . . . 4

1.3.2 Computing General Transition Probabilities . . . 5

1.3.3 Computing the Likelihood of λ . . . 6

1.3.4 Alternatives to the Exponentiated Generator . . . 7

1.3.5 Simulating the Non-Linear Epidemic . . . 8

1.4 Results . . . 10

1.4.1 Maximum Likelihood Estimation of λ for COVID-19 Case Data from Two Counties . 10

1.4.2 Simulation Results . . . 11

1.5 Discussion . . . 11

1.6 Acknowledgments . . . 13

2 Distinguishing Resource Selection From Heavy-Tailed Dispersal in Spatial Epidemic Models **14**

2.1 Abstract . . . 14

2.2 Acknowledgements . . . 14

2.3 Mathematical Supplement . . . 15

2.3.1 SI Model . . . 15

2.3.2 SIR Model . . . 15

2.3.3 Kermack and McKendrick's Epidemic Model . . . 16

2.3.4 Connecting Kermack and McKendrick's Model and the SIR Model 18
2.3.5 Exploring the SIR Model 19

3 Conscientious vaccination exemptions in kindergarten to eighth-grade children across Texas schools from 2012-2018: A regression analysis 23

3.1 Abstract 23
3.2 Acknowledgments 24
3.3 Mathematical Supplement 24

Bibliography 30

Background

The influence of disease is woven throughout human history. The evolution and dispersal of early human populations were shaped by disease [1] and there have been innumerable plagues and pandemics before SARS-COV-2. Humans have long sought to understand, control, and prevent the spread of infectious disease, and mathematics has played an essential role in this endeavor.

The field of epidemiology—the study of the spread and control of disease—has roots some would trace to the ancient writings of Hippocrates [2]. However, it wasn't until the 18^{th} century that mathematics was used as a tool to understand disease spread; in 1760, the Swiss mathematician and physicist Daniel Bernoulli famously conducted a mathematical analysis of smallpox to support his argument for widespread inoculation [3]. Despite the novelty of Bernoulli's work, it was at least 100 years before the application of mathematics to infectious disease was again seriously considered.

In 1916, Sir Ronald Ross wrote, "It is somewhat surprising that so little mathematical work should have been done on the subject of epidemics, and, indeed, on the distribution of diseases in general. Not only is the theme of immediate importance to humanity, but it is one which is fundamentally connected with numbers, while vast masses of statistics have long been awaiting proper examination" [4]. Ross was a British doctor and Nobel laureate; his discovery that malaria was caused by a blood-borne parasite and was transmitted by mosquitoes revolutionized our understanding of the disease. His 1916 paper "An application of the theory of probabilities to the study of a priori pathometry," was the first of a three-part series by the same title; the latter two installations were co-authored with algebraic geometer Hilda P Hudson [4–6].

This overlooked series sparked interest in mathematical epidemiology more widely, inspiring the Scottish biochemist William Ogilvy Kermack and the Scottish physician and epidemiologist Anderson Gray McKendrick to pen their seminal 1927 work "A contribution to the mathematical theory of epidemics." This paper is often cited as the origin of compartmental epidemiological models–those which assign individuals into compartments based on their infection status (e.g. susceptible, infected, or recovered) and describe the rates at which individuals move between compartments [7].

The writings of these early 20^{th}-century physicians, epidemiologists, and mathematicians were remarkably

prescient. When they were written there was very little information available about the prevailing communicable diseases. The field of microbiology was yet nascent in 1890, just thirty years before the aforementioned work of Sir Robert Ross, when Koch published his famous criteria for determining whether a microbe caused a disease. In the absence of extensive mechanistic knowledge of disease processes, the capacity to understand and even predict infectious disease dynamics could have been revolutionary. But these ideas failed to find fertile soil until they were expanded upon by mathematicians and population biologists more than 20 years later.

In 1957, Norman Bailey published a book titled *The Mathematical Theory of Epidemics* [8] which led to a proliferation of disease modeling methods. In the 1970s and 80s Herbert Hethcote, Robert May, Roy Anderson, George MacDonald, and others published seminal papers which laid the foundations for modern mathematical epidemiology [9–13]. The tools they developed are at the core of our modern understanding of infectious disease dynamics, providing insights into threshold phenomena (how many susceptible individuals must there be to sustain an epidemic?), reproduction numbers (how many new cases will result from a single new infection?) [14], and other now-familiar results. These findings are at the core of the mathematics discussed in chapter 2.

The compartmental models of Kermack, McKendrick, and others are deterministic—that is, they contain no element of randomness. The need for models which could account for the randomness of disease spread was met by another class of models, called stochastic models, which incorporate random dynamics. Early stochastic epidemic models were presented in work by Metz et al. in 1978 [15] and Billard et al. in 1979 [16].

Over the years, such models have grown in scope and complexity. Today's contact-network models of COVID-19 transmission can account for variable contact rates between age groups, a delay between exposure and infection, asymptomatic transmission, hospitalization, and transmission along edges of a massive graph connecting populations [17]. But, beneath the big data, such advanced models still share fundamental similarities with the work of Kermack and McKendrick nearly 100 years ago—a reminder of the great progress that has been made in this field already and the great work left to do.

Chapter 1

A Non-Linear Stochastic Epidemic

1.1 Abstract

The spread of an infectious disease through a population can be modeled as a continuous-time Markov process, where the rate of new infections each day is proportional to the number of distinct healthy-infected pairs. Here, we discuss the non-linear epidemic model introduced in [18] and derive its likelihood function. We then use COVID-19 data from two counties to estimate the maximum-likelihood model parameter in a real epidemic and develop an algorithm to simulate the epidemic process.

1.2 Introduction

An alternative to the deterministic compartmental epidemiological models (which are discussed in chapter 2) is to think of disease spread as a non-linear stochastic process [18]. In this model, individuals are either healthy or infected, and we track the total number of individuals in each category (we do not track individuals). The model assumes that the epidemic occurs in a finite population and begins with a single infected individual. The rate at which individuals change from healthy to infected is proportional to the product of the number of infected individuals and the number of healthy individuals (that is, the number of possible encounters between healthy and infected individuals). The model assumes that infected individuals never recover and are able to continue transmitting the disease indefinitely. This assumption combined with the assumption that infected individuals never recover means that the entire population will eventually become infected.

Because the rate at which new infections occur is directly proportional to the number of possible encounters between healthy and infected individuals, this rate necessarily changes over the course of the epidemic.

While the model is defined in terms of its transition probability for single-individual steps (that is, the probability of a single new infection occurring), when not too close to the entire population being infected it can also be approximated as a nonhomogeneous Poisson process, with an intensity that is proportional to the number of possible encounters between healthy and infected individuals (as opposed to a traditional Poisson process, which has a constant intensity). More details on the relation between this epidemic model and a nonhomogeneous Poisson process can be found in section 1.3.5.

1.3 Statistical Background

1.3.1 Describing an Epidemic as a Continuous-Time Markov Process

The epidemic process we consider in this chapter is a continuous-time Markov process. The evolution of a continuous-time Markov process is characterized by its generator matrix, $G = [g_{ij}]$. If we denote the transition probability $p_{ij}(h) = P(X(t+h) = j | X(t) = i)$ (the probability of moving from state i to state j over a small time period of duration h), then each g_{ij} satisfies: $p_{ij}(h) \approx g_{ij}h$ if $i \neq j$ and $p_{ii} \approx 1 + g_{ii}h$. It is clear that $g_{ij} \geq 0$ for $i \neq j$ and $g_{ii} \leq 0$ for all i. For this scenario, (a) nothing happens on $(t, t+h)$ with probability $1 + g_{ii}h + o(h)$ and (b) the chain jumps to state $j(\neq i)$ with probability $g_{ij}h + o(h)$, where $o(h)$ is a first-order error term. Since $1 = \sum_j p_{ij}(h)$, it follows that $1 = 1 + h\sum_j g_{ij} \implies 0 = \sum_j g_{ij}$ for all i; that is, the rows of G sum to zero.

In our case, we treat the number of healthy individuals at time t, $X(t)$, as a continuous-time Markov process in a finite state space of size $N+1$ (the population size). We define the generator's entries g_{ij} so that they are directly proportional to the number of possible distinct encounters between the i healthy individuals and the $N + 1 - i$ infected individuals (there are $i \cdot (N + 1 - i)$ such combinations). Under this framework, the probability of a single new infection occurring in a time window of width h is: $p_{i,i\ \ 1}(h) = P(X(t+h) = i - 1 | X(t) = i) = \lambda_i \cdot h = \lambda \cdot i(N + 1 - i) \cdot h$, where λ is a parameter representing the probability of any healthy-infected interaction occurring times the probability of that interaction resulting in a new infection. Recall that $p_{ij}(h) \approx g_{ij}h$ if $i \neq j$ and $p_{ii} \approx 1 + g_{ii}h$. It follows that the generator $G = [g_{ij}]$ for this process is an $(N + 1) \times (N + 1)$ matrix

$$G = \begin{pmatrix} 0 & 0 & 0 & 0 & ... & 0 & 0 & 0 \\ \lambda_1 & -\lambda_1 & 0 & 0 & ... & 0 & 0 & 0 \\ 0 & \lambda_2 & -\lambda_2 & 0 & ... & 0 & 0 & 0 \\ 0 & 0 & \lambda_3 & -\lambda_3 & ... & 0 & 0 & 0 \\ & & & & \vdots & & & \\ 0 & 0 & 0 & 0 & ... & \lambda_{N\ 1} & -\lambda_{N\ 1} & 0 \\ 0 & 0 & 0 & 0 & ... & 0 & \lambda_N & -\lambda_N \end{pmatrix}, \text{ where } \lambda_i = \lambda \cdot i(N+1-i).$$

We suppose that the process begins with $X(0) = N$; that is, in our population of $N+1$ individuals, there are initially N healthy individuals and 1 infected individual. We can think of this as starting in the bottom row of the generator matrix G: the possible next steps are to either move to the state corresponding to $N-1$ healthy individuals (one individual gets infected, which occurs in the time interval $(0,h)$ with probability $g_{N+1,N}h = \lambda_N h$) or to stay at the state corresponding to N healthy individuals (there are no new infections, which occurs in the time interval $(0,h)$ with probability $1 + g_{N+1,N+1}h = 1 - \lambda_N h$). Note that the probabilities of these two possible scenarios sum to 1. In this way the number of healthy individuals decreases over time until the entire population has been infected, at which point $X(t) = 0$ for all subsequent values of t (this is why the first row of the generator is 0).

1.3.2 Computing General Transition Probabilities

Up to this point, we have only considered the probability of the number of healthy individuals decreasing by 1 in a time window of width h (i.e. $p_{i,i\ 1}(h) = P(X(t+h) = i-1|X(t) = i)$). What if we want to consider the more general case of going from n to m healthy individuals, with $m \leq n$, i.e., $p_{nm}(h) = P(X(t+h) = m \leq n|X(t) = n)$? To compute this value, we must return to concepts discussed in the first paragraph of this section. Let P_h be the matrix of transition probabilities $[p_{ij}(h)]$. Recall that $p_{ij}(h) = P(X(t+h) = j|X(t) = i)$ is the probability of moving from state i to state j over a small time period of duration h. One can prove that (a) $P_0 = [p_{ij}(0)] = I$ (the identity matrix), (b) P_h is stochastic (it has non-negative entries and its rows sum to one), and (c) P_h satisfies the Chapman-Kolmogorov equations (that is, $P_{s+t} = P_s P_t$ if $s, t \geq 0$). Our goal is to relate this matrix to G in a way that enables us to estimate $p_{nm}(h)$ for general m, n ($m \leq n$, m not necessarily equal to $n-1$).

Recall that we defined the entries of G, g_{ij}, as $p_{ij}(h) \approx g_{ij}h$ if $i \neq j$ and $p_{ii} \approx 1 + g_{ii}h$. It follows that $\lim_{h\#0} \frac{1}{h}(P_h - I) = G$, which implies that P_t is differentiable at $t = 0$. We will now compute the derivative

of P_t:

$$\begin{aligned} p_{ij}(t+h) &= \sum_k p_{ik}(t)p_{kj}(h) \text{ (by the Chapman-Kolmogorov condition)} \\ &\approx p_{ij}(t)(1+g_{ij}h) + \sum_{k:k\neq j} p_{ik}(t)g_{kj}h \text{ (by the definition of } G) \\ &= p_{ij}(t) + h\sum_k p_{ik}(t)g_{kj} \end{aligned}$$

Subtracting $p_{ij}(t)$ from both sides and dividing by h gives:

$$\frac{1}{h}[p_{ij}(t+h) - p_{ij}(t)] \approx \sum_k p_{ik}(t)g_{kj} = (P_tG)_{ij}$$

the left-hand side of which is the definition of a derivative with respect to t, $(P_t^{\prime})_{ij}$. Thus we have shown that $P_t^{\prime} = P_tG$.

In the same way that $p(t) = e^{tg} = \sum_{l=0}^{\infty} \frac{t^l g^l}{l!}$ solves the differential equation $p^{\prime}(t) = p(t)g$, the above differential equation is solved by the matrix

$$P_t = e^{tG} = \sum_{l=0}^{\infty} \frac{t^l G^l}{l!} \quad \text{where } G^0 = I.$$

The $(n,m)^{\text{th}}$ element of this matrix, $P_t(n,m) = [e^{tG}]_{n,m}$, is the probability of going from state n to state m over time t, given the generator G—this is what we sought to compute.

1.3.3 Computing the Likelihood of λ

We now have introduced the concepts necessary to compute the likelihood of λ given observed case data for a real epidemic. Recall that $G = \begin{pmatrix} 0 & 0 & 0 & 0 & ... & 0 & 0 \\ \lambda_1 & -\lambda_1 & 0 & 0 & ... & 0 & 0 \\ 0 & \lambda_2 & -\lambda_2 & 0 & ... & 0 & 0 \\ & & & \vdots & & & \\ 0 & 0 & 0 & 0 & ... & \lambda_N & -\lambda_N \end{pmatrix}$ for $\lambda_i = \lambda \cdot i(N+1-i)$, and that $P(X(t+h) = m \leq n | X(t) = n) = [e^{hG}]_{n,m}$.

Suppose I have the COVID-19 case counts for K time points, $t_1, t_2, ..., t_K$. Let $\mathbf{X} = \{X_k | k = 1, ..., K\}$ denote these observations. The likelihood for λ, the only parameter in this model, is:

$$L(\lambda|\mathbf{X}) = \prod_{k=1}^{K-1} P(X(t_{k+1}) = X_{k+1} | X(t_k) = X_k) = \prod_{k=1}^{K-1} [e^{(t_{k+1} - t_k)G}]_{X_k, X_{k+1}}$$

This can be evaluated over a range of λ values, and the λ at which $L(\lambda|\mathbf{X})$ obtains a maximum can be identified numerically.

I implemented this method in Algorithm 1. I first construct the generator matrix, G. I begin by creating an $N \times N$ matrix with diagonal elements $g_{ii} = \lambda_i = \lambda \cdot i \cdot (N + 1 - i)$ and off-diagonal elements $g_{ij} = 0$ for $i \neq j$ (lines 1-3). I then construct G by creating two matrices: the first with a row of zeros to the top and a column of zeros to the left of this diagonal matrix $[g_{ii}]$, and the second with with a row of zeros to the top and a column of zeros to the right of the diagonal matrix $[g_{ii}]$. Subtracting the second matrix from the first gives G (line 4). I next exponentiate this matrix (line 5) and identify all of the (X_k, X_{k+1}) coordinate pairs (line 6). Finally, I take the product of all of these elements of G in order to compute the likelihood (line 7).

Algorithm 1: Computation of the likelihood of lambda

```
   // Construct generator matrix, G, based on given lambda value
   // N+1 is the population size
   // X is a vector of observed healthy individuals (X[t] = # of healthy individuals
      on day t)
1  lambda.vec ← function(lambda, healthy.vec)lambda*healthy.vec*(N-healthy.vec) // Creates a
      vector of λ_i s for a vector of healthy individuals over time and a fixed lambda
      parameter
2  G ← matrix(rep(0,(N^2)), nrow = N, ncol = N)
3  diag(G) ← lambda.vec(lambda = lambda, healthy.vec = 1:N)
4  G ← cbind(rep(0,N+1),(rbind(rep(0,N),G))) + cbind(rbind(rep(0,N),-G),rep(0,N+1))
   // Compute exponential of G
5  exp.G ← expm::expm(G)
   // Identify which elements of G must be multiplied to compute the likelihood
6  choose ← matrix(c(X[-length(X)], X[-1]), ncol=2)
   // Return the likelihood, ∏_{k=1}^{K-1} [e^G]_{X_k, X_{k+1}}
7  return(prod(exp.G[choose]))
```

While this algorithm works, it has some major limitations. In particular, for large population sizes, the necessary matrices become very large–these matrices require a lot of memory and exponentiation becomes computationally infeasible. My computer cannot work with a matrix of dimension surpassing $40,000 \times 40,000$. While it is possible to compute the matrix exponential of such large matrices, for example using the *expm()* function in R [19, 20], this computation is very time consuming. I found that it can take upwards of 10-15 minutes for even a population of just 3,000 individuals. Thus, a less time consuming option would facilitate easier and more thorough analysis.

1.3.4 Alternatives to the Exponentiated Generator

One way to possibly reduce the time necessary to compute the exponential of a large matrix is to eliminate the need to exponentiate a large, non-diagonal matrix altogether. This can be done, for example, by substituting

$G = ADA^{-1}$, for some matrix A and diagonal matrix D.

Denote $H = e^G = \sum_{l=0}^{\infty} \frac{G^l}{l!}$, the exponentiation of the generator, G. It follows that

$$\begin{aligned} H = e^G = e^{ADA^{-1}} &= \sum_{l=0}^{\infty} \frac{(ADA^{-1})^l}{l!} \\ &= \sum_{l=0}^{\infty} \frac{(ADA^{-1})(ADA^{-1})...(ADA^{-1})}{l!} \\ &= \sum_{l=0}^{\infty} \frac{AD^lA^{-1}}{l!} \\ &= A\left(\sum_{l=0}^{\infty} \frac{D^l}{l!}\right)A^{-1} \\ &= Ae^DA^{-1} \end{aligned}$$

e^D is easily computed because D is diagonal: we simply calculate the exponential of the N diagonal entries ($(e^D)_{ii} = e^{d_i}$, where d_i for $i = 1, ..., N$ are the diagonal elements of D). We claim that there is an iterative method inspired by divided differences which can be used to identify matrices A and D satisfying this requirement.

We construct D as a diagonal matrix with the diagonal elements of G as its diagonal elements: $D_{ii} = G_{ii}$ for all $i = 1, ..., N$, $D_{ij} = 0$ for $i \neq j$. We construct A as a lower triangular matrix such that $A_{ii} = 1$ for all $i = 1, ..., N$, and $A_{ij} = 0$ for $i < j$. For $i > j$, we define A_{ij} as follows:

$$A_{ij} = A_{i\ 1,j} \times \frac{G_{ii}}{G_{ii} - G_{jj}}$$

Is it more time-efficient to iteratively fill the elements of A, compute A's inverse, and evaluate Ae^DA^{-1}, than to compute e^G using R's *expm* function [20]? Unfortunately, it is not. Both methods take more than ten minutes, on average, for my local machine to compute. Thus, we will continue using the *expm* function for our analysis.

1.3.5 Simulating the Non-Linear Epidemic

In order to better understand the dynamics of this stochastic process and compare this process to real data, I formulated an algorithm to simulate it (Algorithm 2).

I observed that in this epidemic model, when the number of infected individuals (denoted $I(t) = N - X(t)$) is not too close to the total population size (N), $I(t)$ can be approximated as a nonhomogeneous Poisson process with a dynamic parameter $\lambda_{X(t)} = \lambda \cdot X(t) \cdot (N + 1 - X(t))$, where, $X(t)$ is the number of healthy

individuals on day t, $N+1$ is the total population size, and λ is the probability of any healthy-infected pair occurring *and* that contact resulting in a new infection.

In general, a nonhomogeneous Poisson process is a process $N = \{N(t)|t \geq 0\}$ taking values in $S = \{0, 1, 2, ...\}$ such that:

1. $N(0) = 0$, $N(s) \leq N(t)$ if $s < t$
2. $P(N(t+h) - N(t) \geq 2) = o(h)$,
   $P(N(t+h) - N(t) = 1) = \lambda_t h + o(h)$,
   $P(N(t+h) - N(t) = 0) = 1 - \lambda_t h - o(h)$
3. The number of events in a given increment depends only on λ_t, and not on previous events.

It follows that $N(t+h) - N(t)$ is a Poisson random variable with mean $E[N(t+h) - N(t)] = \int_t^{t+h} \lambda_i di$.

For our model, we can see that $I(t) = N - X(t)$, the number of infected individuals on day t, nearly satisfies these conditions when $I(t)$ is far from N. This is because:

1. $I(0) = 0$ by assumption, $I(s) \leq I(t)$ if $s < t$ because infected individuals never recover, so I is strictly increasing.
2. The transition probability given in [18] for the number of healthy individuals $X(t)$ is: $P(X(t+h) - X(t) = -1) = \lambda_{X(t)} h$ where $\lambda_{X(t)} = \lambda X(t)(N+1-X(t))$. It follows that $P(I(t+h) - I(t) = 1) = \lambda_{X(t)} h$, since $I(t) = N - X(t) \implies X(t) = N - I(t) \implies X(t+h) - X(t) = -I(t+h) + I(t)$.
3. The number of events in a given increment depends only on $\lambda_{X(t)}$, and not on previous events.

This process breaks when $I(t)$ gets too close to N. This is because $I(t) \leq N$ for all t, but a Poisson process is by definition not bounded. For the purpose of the simulation, I account for this property by setting $I(t) = N$ and ending the process when the cumulative number of infections exceeds the population size.

Thus, until nearly the entire population is infected, I can simulate the number of new infections on day t, $I(t) - I(t-1)$, as a Poisson random variable with mean $\int_{t\ \ 1}^{t} \lambda_i di = \lambda_{X(t\ \ 1)} = \lambda \cdot X(t-1) \cdot (N+1-X(t-1))$. Algorithm 2 outlines this simulation. I begin with N healthy individuals (and, implicitly, $N + 1 - N = 1$ infected individual). For each subsequent day, I determine the number of new infections by drawing a random variable from a Poisson distribution with parameter proportional to the number of distinct healthy-infected pairs the previous day: $Pois(\lambda X[t-1](N+1-X[t-1]))$, where λ is fixed and $X[t-1]$ is the number of healthy individuals on the previous day. I repeat this process until nearly the entire population is infected. At some point t , the cumulative number of healthy individuals will equal 0 or go negative: $X(t\ ) \leq 0$. This occurs when the total number of infected individuals exceeds the population size, which obviously cannot

happen in real life. I resolve this issue by setting $X(t\ ) = 0$ and ending the simulation (Algorithm 2, lines 5-7).

Algorithm 2: Stochastic simulation of a non-linear epidemic through a population of size N+1

```
1 X← c() // initialize X as an empty vector to contain healthy individuals
2 X[1]←N
3 for day = 1 to 100000 do
4 |   X[day+1] ←X[day]-rpois(n = 1, lambda = lambda·X[day]·(N+1-X[day]))
  |   // rpois(n=1,lambda) draws a Poisson random variable with parameter lambda
5 |   if X[day+1]≤ 0 then
6 |   |   X[day+1] ← 0
7 |   |   break
```

1.4 Results

1.4.1 Maximum Likelihood Estimation of λ for COVID-19 Case Data from Two Counties

I aimed to use our formulation of the likelihood function (Algorithm 1) to fit the non-linear epidemic model to real COVID-19 case data. However, the population size limitations of this method posed a challenge. If a county's population was too large, we wouldn't be able to compute the likelihood due to computational restraints. But if a county was too small and remote, there would likely be too few test-verified COVID-19 cases to conduct meaningful analysis. Thus, I chose to analyze two of the smallest counties in America which had more than 10 COVID-19 cases as of April 18, 2020: Donley County, Texas, population 3,387, which lies east of Amarillo in the Texas panhandle, and Baker County, Georgia, population 3,038, which lies in the southwest corner of the state. As of April 20, 2020, Donley County had 23 test-verified cases of COVID-19 and Baker county had 19. I used data from the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering on COVID-19 cases between January and April 2020 [21], and data from the US Census on Texas County population sizes [22].

In order to estimate the λ parameter for the epidemic in these counties, I selected a range of plausible λ values between 10^{-6} and 10^{-4}, and evaluated the likelihood of each. I identified 10^{-5} as the most likely order of magnitude. I then evaluated the likelihood of 31 equally-spaced values between 2×10^{-5} and 8×10^{-5} (Figure 1.1). I found that the maximum likelihood estimator of λ is approximately 4.6×10^{-5} in Donley County and 6.4×10^{-5} in Baker County.

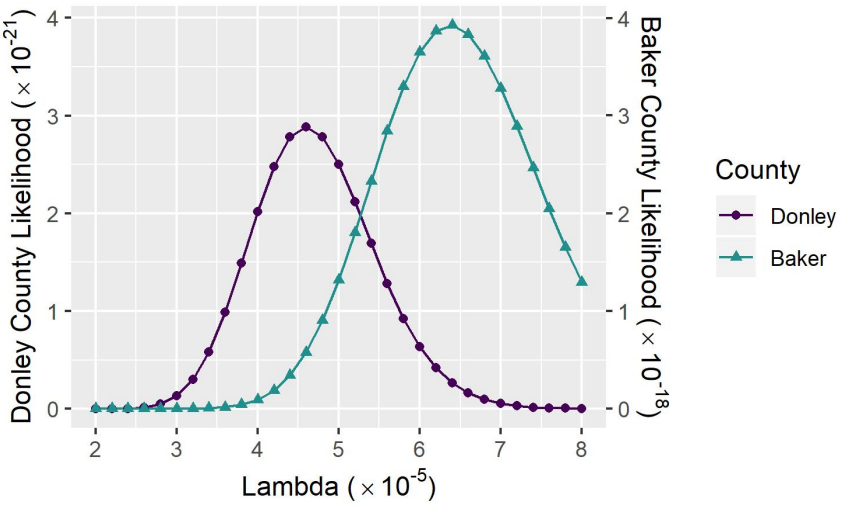

Figure 1.1: **Likelihood distribution for** λ . Estimated using COVID-19 data from Donley County, Texas (purple circles), between March 31 and April 17, 2020 and Baker County, Georgia (turquoise triangles) between March 24 and April 17, 2020 [21]. Maximum likelihood λ estimates for Donley and Baker Counties are 4.6×10^{-5} and 6.4×10^{-5} respectively.

1.4.2 Simulation Results

I simulated disease spread through a population the same size as Donley County (3,387 people) for nine values of λ between 2×10^{-5} and 8×10^{-5} (Figure 1.2). In order to compare these simulated epidemic trajectories to the true epidemic spread in Donley and Baker Counties, I plotted the official COVID-19 case counts in Donley County as black circles, and those in Baker County as black triangles (Figure 1.2).

Recall that λ represents the probability of any healthy-infected pair occurring *and* that contact resulting in a new infection. Thus, it is unsurprising that lower values of λ (purples and blues in Figure 1.2) result in an epidemic which travels more slowly through the population, while higher values of λ (yellow and green in Figure 1.2) result in an epidemic which infects the entire population in as few as 50 days.

For simulations run with the maximum likelihood estimate of λ for Donley County, 4.6×10^{-5}, on average the epidemic took 60 days to infect half of the population and 113 days to infect the entire population. For simulations run with the maximum likelihood estimate of λ for Baker County, 6.4×10^{-5}, on average the epidemic took 45 days to infect half of the population and 80 days to infect the entire population.

1.5 Discussion

Models of epidemic spread can play vital roles in understanding the future dynamics of infectious disease. Based on the maximum-likelihood estimate of $\lambda = 4.6 \times 10^{-5}$ for Donley County, our simulations suggest that the next two months will see approximately half of the population of the county infected. Based on the

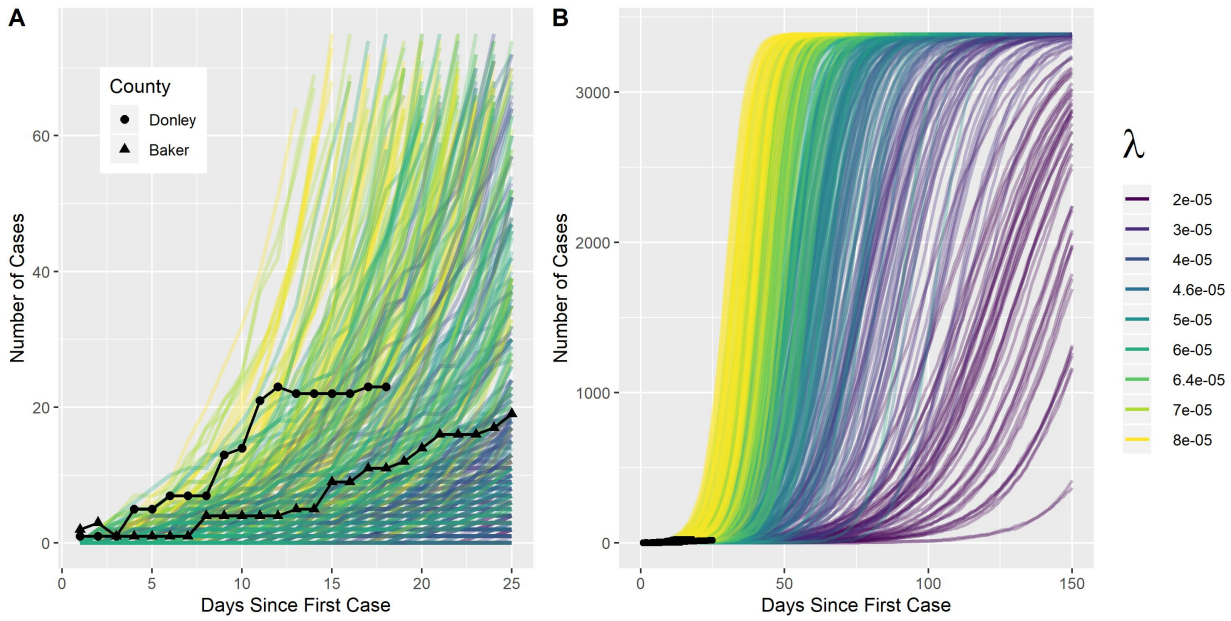

Figure 1.2: **Stochastic simulations of non-linear epidemics in a population of 3,387.** 50 simulations were run for each of the eight λ values considered. Each line represents a simulated epidemic. Circles represent cumulative COVID-19 cases observed in Donley County, TX (population 3,387) and triangles represent cumulative COVID-19 cases observed in Baker County, Georgia (population 3,038). **A** Zoomed-in view of the first 25 days of the epidemic simulations. **B** View of the first 150 days of the simulations.

maximum-likelihood estimate of $\lambda = 6.4 \times 10^{-5}$ for Baker County, our simulations suggest that approximately half of the population could be infected within the next month and a half.

However, this result relies on some very strong assumptions: namely that infected individuals remain infectious indefinitely and never recover. As a consequence of these assumptions, every individual will eventually be infected–something that does not occur in models which incorporate recovery, such as SIR models. Because COVID-19 patients do not stay infected indefinitely, our simulations likely overestimate the total number of infected individuals at any one time and the rate of disease spread through this community. However, they could provide meaningful insights on short time scales. A possible improvement on this model would be to incorporate a recovered category, converting this model from an SI to an SIR framework. In order to do this, we could incorporate a second process, this one with a rate proportional to the number of infected individuals, to simulate the movement of individuals from infected to recovered.

A second shortcoming of this model is its reliance on a matrix with dimensionality equal to the population size. This quality poses a strict limit on the utility of this model, as analyzing populations larger than a few thousand would be computationally infeasible. As we saw in section 1.3, this is a formidable challenge. A possible solution would need to avoid exponentiating the generator matrix, or creating such a large matrix at all. Possible approximations could be explored, potentially employing the fact that G is bidiagonal. However,

this could also be a fundamental shortcoming of working with Markov processes in very large state spaces.

1.6 Acknowledgments

This work was advised by Dr. Stephen Walker, who holds appointments in the UT Austin Mathematics Department and Statistics and Data Sciences Department.

Chapter 2

Distinguishing Resource Selection From Heavy-Tailed Dispersal in Spatial Epidemic Models

2.1 Abstract

The tail of the dispersal kernel of individuals plays a critical role in the spatial spread of infectious disease, invasive species, and other spreading phenomena. However, most studies where the dispersal kernel has been estimated from observed natural systems have assumed homogeneous dispersal in space, even though non-uniform use of space (i.e., resource selection) has long been recognized as important in many systems. In this study we explore the consequences of ignoring terrain heterogeneity when estimating parameters governing the tail of a dispersal kernel. We show that ignoring resource selection in general leads to estimates of dispersal kernels with heavier tails than the true kernels used for simulation. In addition, this often leads to predictions of the rate of spatial spread of infectious disease that are much faster than the true spread through a population that is moving across patchy terrain.

2.2 Acknowledgements

The work originally presented in this chapter was done at The Pennsylvania State University, under the supervision of Dr. Ephraim Hanks, with funding from the Mathematical Biosciences Institute Research Experiences for Undergraduates Program. It was done in collaboration with my research partner for the

summer, Emily Strong.

Because the work has not been published, I am here including only the mathematical supplement, which I wrote independently for the purpose of the Honors Math Thesis with reference to Keeling and Rohani's *Modeling Infectious Diseases in Humans and Animals* [23] and Kermack and McKendrick's 1927 paper *A Contribution to the Mathematical Theory of Epidemics* [7].

2.3 Mathematical Supplement

2.3.1 SI Model

In this project, we simulate disease transmission with a Susceptible-Infected (SI) model. With only two compartments, this is among the simplest forms of compartmental epidemiological models, which were first proposed by Kermack and McKendrick in 1927 [7]. Our formulation categorizes all individuals into one of two categories: susceptible, which contains individuals who have not yet contracted the disease, and infected, which contains individuals who have contracted the disease and are able to transmit it to those around them. This is a deterministic model and assumes that individuals never recover once they have been infected. In our simulation, the SI epidemic model at each location l is described by the following system of coupled ordinary differential equations:

$$\frac{dS_\ell}{dt} = -\frac{\beta S_\ell I_\ell}{N_\ell}$$
$$\frac{dI_\ell}{dt} = \frac{\beta S_\ell I_\ell}{N_\ell}$$

where S_ℓ is the total number of susceptible individuals, I_ℓ is the total number of infected individuals, the total population size is $N_\ell = S_\ell + I_\ell$, and β is an infectivity parameter which is the product of the contact rate and the transmission probability.

2.3.2 SIR Model

We used this model because our work was motivated by a disease, brucellosis, from which individuals never recover. However, a more common disease model, the SIR, adds a third compartment for individuals who are recovered and immune from the disease, and lets S, I, and R represent the *proportion*, not the number,

of individuals in each category:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{2.1}$$

In this formulation (and for the rest of this discussion), the fact that we are modeling population proportions means that we no longer need to divide each term by the total population size, N. β is still an infectivity parameter and the new parameter, γ, represents the removal rate. The removal rate is the sum of the mortality and recovery rates; $1/\gamma$ is accordingly the average infectious period.

2.3.3 Kermack and McKendrick's Epidemic Model

The original formulation of this model was in a seminal 1927 paper by Kermack and McKendrick [7]. When reading this paper, I noticed that Kermack and McKendrick assumed that an individual's infection and recovery rates depended on how long that individual had been infected. I observed that relaxing this assumption results in exactly the SIR model discussed earlier in equation 2.1.

In this section, I will first outline the work leading to Kermack and McKendrick's compartmental model. I will then outline the math necessary to connect this model with the SIR model of equation 2.1. Finally, I will discuss aspects of the SIR model which yield meaningful insights into the long-term dynamics of epidemics.

Kermack and McKendrick divide time into discrete intervals, with new infections occurring only at the boundary between intervals. They denote the number of individuals per unit area who at time t have been infected for a total of θ intervals as $v_{t,\theta}$ (these individuals became infected in the instant between interval $t-\theta-1$ and interval $t-\theta$). In general, they denote the number of individuals who became infected between $t-1$ and t as $v_t = v_{t,0}$ for $t \geq 1$. For $t = 0$, they denote

$$v_{0,0} = v_0 + y_0 \tag{2.2}$$

where y_0 is the number who were initially infected by some outside process.

They denote ψ_θ as the sum of the recovery and death rates, also known as the removal rate, for individuals who have been infected for θ time intervals. Thus we can express the number of individuals at time t who

have been infected for a total of θ intervals, $v_{t,\theta}$, as

$$\begin{aligned} v_{t,\theta} &= v_{t\ \ 1,\theta\ \ 1}(1-\psi_{\theta\ \ 1}) \\ &= v_{t\ \ 2,\theta\ \ 2}(1-\psi_{\theta\ \ 1})(1-\psi_{\theta\ \ 2}) \\ &= v_{t\ \ \theta,0}B_\theta \end{aligned} \tag{2.3}$$

where $B_\theta = (1-\psi_{\theta\ \ 1})(1-\psi_{\theta\ \ 2})...(1-\psi_0)$ is the probability that an individual does not recover over θ time intervals and $v_{t\ \ \theta,0}$ is the number of individuals who became infected between time interval $t-\theta-1$ and time interval $t-\theta$.

We can also express the number of individuals who became infected between time interval $t-1$ and time interval t, v_t, as

$$v_t = x_t \sum_{\theta=1}^{t} \phi_\theta v_{t,\theta} \tag{2.4}$$

where x_t is the number of susceptible individuals at time t and ϕ_θ is the infectivity rate of individuals who have been infected for θ time intervals (we assume that ϕ_0 is 0).

If z_t is the total number of individuals who have been removed due to recovery or death, then

$$N = x_t + y_t + z_t \tag{2.5}$$

where N is the total population size.

It follows that:

$$\begin{aligned} v_t &= x_t \sum_{\theta=1}^{t} \phi_\theta v_{t,\theta} \text{ (by equation 2.4)} \\ &= x_t \sum_{\theta=1}^{t} \phi_\theta B_\theta v_{t\ \ \theta,0} \text{ (by equation 2.3)} \\ &= x_t \left(\phi_t B_t y_0 + \sum_{\theta=1}^{t} \phi_\theta B_\theta v_{t\ \ \theta} \right) \text{ (by equation 2.2)} \\ &= x_t \left(A_t y_0 + \sum_{\theta=1}^{t} A_\theta v_{t\ \ \theta} \right) \text{ (where } A_\theta = \phi_\theta B_\theta) \\ \implies x_t - x_{t+1} &= x_t \left(A_t y_0 + \sum_{\theta=1}^{t} A_\theta v_{t\ \ \theta} \right) \text{ (since } v_t = x_t - x_{t+1} \text{ by definition)} \\ \implies x_{t+1} - x_t &= -x_t \left(A_t y_0 + \sum_{\theta=1}^{t} A_\theta v_{t\ \ \theta} \right) \end{aligned} \tag{2.6}$$

We also know that the total number of people who recover or die at the end of time interval t is

$$\begin{aligned} z_{t+1} - z_t &= \sum_{\theta=1}^{t} \psi_\theta B_\theta v_{t\ \theta,0} \\ &= \psi_t B_t y_0 + \sum_{\theta=1}^{t} \psi_\theta B_\theta v_{t\ \theta} \text{ (by equation 2.2)} \\ &= C_t y_0 + \sum_{\theta=1}^{t} C\theta v_{t\ \theta} \text{ (where } C_\theta = \psi_\theta B_\theta) \end{aligned} \tag{2.7}$$

In the limit as the width of each time interval goes to zero, the result of equation 2.6 becomes

$$\frac{dx_t}{dt} = -x_t \left(A_t y_0 + \int_0^t A_\theta v_{t\ \theta} d\theta \right) \tag{2.8}$$

and the result of equation 2.7 becomes

$$\frac{dz_t}{dt} = C_t y_0 + \int_0^t C_\theta v_{t\ \theta} d\theta \tag{2.9}$$

Since $x_t + y_t + z_t = N$, a constant (eq. 2.5), $\frac{dx_t}{dt} + \frac{dy_t}{dt} + \frac{dz_t}{dt} = 0$. It follows that

$$\frac{dy_t}{dt} = \frac{-dx_t}{dt} - \frac{dz_t}{dt} = x_t \left(A_t y_0 + \int_0^t A_\theta v_{t\ \theta} d\theta \right) - C_t y_0 + \int_0^t C_0 v_{t\ \theta} \tag{2.10}$$

2.3.4 Connecting Kermack and McKendrick's Model and the SIR Model

Kermack and McKendrick's 1927 paper explores the implications of equations 2.8, 2.9, and 2.10 while maintaining the assumption that an individual's removal (through recovery or death) and infectivity rates depend on how long that individual has been infected (θ). Because the SIR model of disease spread we introduced did not make this distinction, I will relax this assumption.

So, for all θ, suppose that $\psi_\theta = \psi$ and $\phi_\theta = \phi$. It follows that $B_\theta = (1 - \psi_{\theta\ 1})(1 - \psi_{\theta\ 2})...(1 - \psi_0) = (1 - \psi)^\theta$, and thus that $A_\theta = \phi_\theta B_\theta = \phi(1 - \psi)^\theta$ and $C_\theta = \psi_\theta B_\theta = \psi(1 - \psi)^\theta$. Note also that $(1 - \psi)^t y_0$ is the number of initially-infected individuals who are still infected at the current time t, and $\int_0^t (1 - \psi)^\theta v_{t\ \theta} d\theta$ is the number of individuals who were infected at subsequent time points and are still infected, and so the total number of infected individuals at time t can be expressed as $y_t = (1 - \psi)^t y_0 + \int_0^t (1 - \psi)^\theta v_{t\ \theta} d\theta$.

It follows that

$$\begin{aligned}\frac{dx_t}{dt} &= -x_t\left(\phi(1-\psi)^t y_0 + \int_0^t \phi(1-\psi)^\theta v_t \ _\theta d\theta\right) \\ &= -\phi x_t\left((1-\psi)^t y_0 + \int_0^t (1-\psi)^\theta v_t \ _\theta d\theta\right) \\ &= -\phi x_t y_t\end{aligned} \tag{2.11}$$

$$\begin{aligned}\frac{dz_t}{dt} &= \psi(1-\psi)^t y_0 + \int_0^t \psi(1-\psi)^\theta v_t \ _\theta d\theta \\ &= \psi\left((1-\psi)^t y_0 + \int_0^t (1-\psi)^\theta v_t \ _\theta d\theta\right) \\ &= \psi y_t\end{aligned} \tag{2.12}$$

$$\begin{aligned}\frac{dy_t}{dt} &= -\frac{dx_t}{dt} - \frac{dz_t}{dt} \\ &= \phi x_t y_t - \psi y_t\end{aligned} \tag{2.13}$$

Simple notation substitutions give us the SIR model from equation 2.1. Letting $S = x_t/N$, $I = y_t/N$, $R = Z_t/N$, $\beta = \phi$, and $\gamma = \psi$, we arrive at:

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

Note that S, I, R are proportions, so now $S + I + R = 1$, not N.

2.3.5 Exploring the SIR Model

This simplified parameterization of Kermack and Mckendrick's classic model reveals some behaviors which yield valuable insights into epidemic dynamics.

The Basic Reproductive Number, R_0

For example, the average number of new infections generated by a single infected individual in an entirely susceptible population, a parameter known in the epidemiology community as the basic reproductive number or R_0, can be computed as the infectivity rate times the infectious period: $R_0 = \beta/\gamma$ (since γ is the removal–

via recovery or mortality–rate, its inverse is the average time period before removal). This parameter measures the maximum potential of an infectious disease to spread through a population. R_0 is critical in understanding epidemic behavior because a pathogen can only invade a susceptible population if $R_0 > 1$. This makes sense intuitively: if each case results in fewer than 1 new cases, then on average each recovered individual will have fewer than one replacement in the infected population and the outbreak will eventually die out. We will show this more rigorously in the Equilibrium Conditions subsection.

It is also true that, in a closed population such as ours, an infectious disease can only invade the population if the fraction of susceptible individuals in the population exceeds $1/R_0$. This is because, per our earlier formulation, $S > 1/R_0 \implies S > \gamma/\beta$. Plugging this into the equation for $\frac{dI}{dt}$ gives $\frac{dI}{dt} > \beta(\gamma/\beta)I - \gamma I \implies \frac{dI}{dt} > 0$. This phenomenon is the origin of the concept of herd immunity: through natural immunity or vaccination, reducing the fraction of the population that is susceptible to below $1/R_0$ can prevent $\frac{dI}{dt}$ from being positive, thus preventing an infectious disease outbreak.

Equilibrium Conditions

Additionally, we can explore the equilibrium conditions of this model by setting each equation in 2.1 equal to 0:

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI = 0\\ \frac{dI}{dt} &= \beta SI - \gamma I = 0\\ \frac{dR}{dt} &= \gamma I = 0\end{aligned}$$

It is clear that if $I = 0$ these equations hold for all values of β and γ. This gives the disease-free equilibrium $(S, I, R) = (S(\infty), 0, R(\infty))$, where $S(\infty) = \lim_{t \to \infty} S(t)$ and $R(\infty) = \lim_{t \to \infty} R(t)$. We can calculate $S(\infty)$ and $R(\infty)$ by dividing the equation for $\frac{dS}{dt}$ by that for $\frac{dR}{dt}$:

$$\frac{dS}{dR} = \frac{-\beta SI}{\gamma I} = -\frac{\beta}{\gamma}S = -R_0 S$$

This equation can be solved using the separation of variables method:

$$\frac{dS}{dR} = -R_0 S \implies \frac{1}{S}dS = -R_0 dR \implies \int \frac{1}{S}dS = \int -R_0 dR \implies \ln S(t) = -R_0 \cdot R(t) + c$$

Plugging in $t = 0$ and solving for c gives that $c = \ln S(0)$, since $R(0) = 0$. So

$$\ln S(t) - \ln S(0) = \ln \frac{S(t)}{S(0)} = -R_0 \cdot R(t) \implies \frac{S(t)}{S(0)} = e^{-R_0 R(t)} \implies S(t) = S(0)e^{-R_0 R(t)}$$

To understand the long-term behavior of this equation, we can take the limit as $t \to \infty$:

$$\lim_{t \to \infty} S(t) = \lim_{t \to \infty} S(0)e^{-R_0 R(t)} \implies S(\infty) = e^{-R_0 R(\infty)}$$

Because we have already established that this is a disease-free equilibrium (that is, $I(\infty) = 0$), we know that $S(\infty) + R(\infty) = 1$. So

$$\begin{aligned} 0 &= 1 - R(\infty) - S(\infty) \\ &= 1 - R(\infty) - S(0)e^{-R_0 R(\infty)} \end{aligned} \tag{2.14}$$

This equation is transcendental and cannot be solved analytically for $R(\infty)$. However, note that when $R(\infty) = 1$ the expression is less than zero, and that when $R(\infty) = 0$ the expression is greater than zero: this suggests that a solution exists and can be estimated numerically. The result of such a numerical method for a range of R_0 values is plotted in figure 2.1.

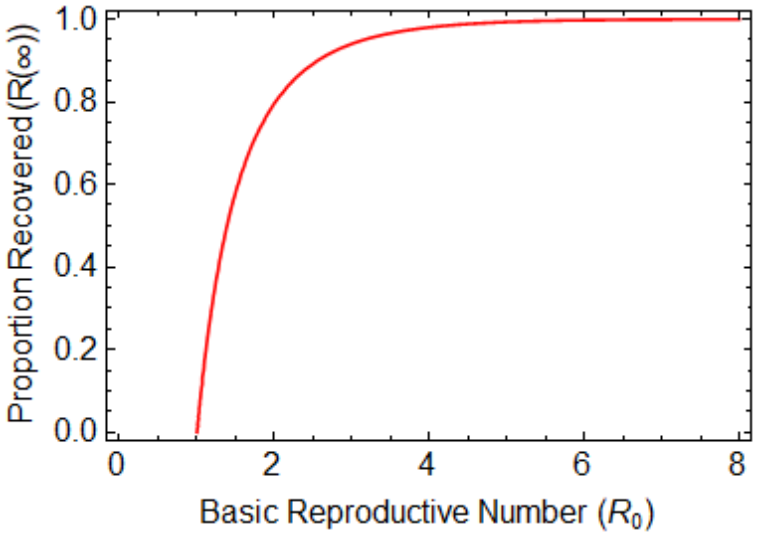

Figure 2.1: **Behavior of equation 2.14 for R_0 between 0 and 8.** Equation was evaluated using the software *Mathematica* [24].

Figure 2.1 confirms the result we encountered earlier: for values of $R_0 < 1$, the disease cannot invade and an epidemic does not occur. But for values of $R_0 \geq 1$, the disease can take off in the population. This makes intuitive sense because $R_0 = \frac{\beta}{\gamma}$ (where β is the infectivity rate and γ is the removal rate). If $R_0 < 1 \implies \beta < \gamma$, then individuals are removed from the population at a faster rate than they can infect others.

Figure 2.1 also yields a new and potentially counterintuitive insight: that the final proportion of the population that contracts the illness depends on R_0 and that for low values of R_0 (i.e. those between 1

and 3) a large proportion of the population will never get infected. This is because, for diseases with R_0 values in this range, disease spread is halted not due to a lack of susceptible individuals, but due to the number of infected individuals declining as a consequence of their removal rate being too large relative to their infectivity.

If $I \neq 0$, the end state is known as an endemic equilibrium. This state exists in formulations which include new individuals entering the susceptible pool, through birth or waning immunity. In these cases, we can find the perhaps counterintuitive result that disease can persist indefinitely at low frequency in the population, depending on the birth rate and the parameters γ and β. If we explore this case in our current formulation, we see from the equation for $\frac{dI}{dt}$ that, after dividing by I, $\beta S = \gamma \implies S = \frac{\gamma}{\beta}$. In our case, substituting this into the equation for $\frac{dS}{dt}$ gives $-\beta(\frac{\gamma}{\beta})I = 0 \implies -\gamma I = 0$. This implies that $\gamma = 0$, which means that $S = 0$, which is a contradiction. So, in our formulation the only equilibria are disease free. However, if there were a positive birth term in the equation for $\frac{dS}{dt}$, this would not be the case.

Chapter 3

Conscientious vaccination exemptions in kindergarten to eighth-grade children across Texas schools from 2012-2018: A regression analysis

3.1 Abstract

Background As conscientious vaccination exemption (CVE) percentages rise across the United States, so do the risk and occurrence of outbreaks of vaccine-preventable diseases such as measles. In the state of Texas, the median CVE percentage across school systems more than doubled between 2012 and 2018. During this period, the proportion of schools surpassing a CVE percentage of 3% rose from 2% to 6% for public schools, 20% to 26% for private schools, and 17% to 22% for charter schools. The aim of this study was to investigate this phenomenon at a fine scale.

Methods and Findings Here, we use beta regression models to study the socioeconomic and geographic drivers of CVE trends in Texas. Using annual counts of conscientious vaccination exemptions at the school system level from the 2012-2013 to the 2017-2018 school year, we identified county-level predictors of median CVE percentage among public, private, and charter schools, the proportion of schools below a high-risk threshold for vaccination coverage, and five-year trends in CVE's.Since the 2012-2013 school year, CVE percentages have increased in 41 out of 46 counties in the top ten metropolitan areas of Texas. We find that

77.6% of the variation in CVE percentages across metropolitan counties is explained by median income, the proportion of the population that holds a bachelor's degree, the proportion of the population that self-reports as ethnically white, the proportion of the population that is English speaking, and the proportion of the population that is under the age of five years old. Across the ten top metropolitan areas in Texas, counties vary considerably in the proportion of school systems reporting CVE percentages above 3%. Sixty-six percent of that variation is explained by the proportion of the population that holds a bachelor's degree and the proportion of the population affiliated with a religious congregation. Three of the largest metropolitan areas—Austin, Dallas-Fort Worth, and Houston—are potential vaccination exemption "hot spots" with over 13% of local school systems above this risk threshold. The major limitations of this study are inconsistent school-system-level CVE reporting during the study period and a lack of geographic and socioeconomic data for individual private schools.

Conclusions In this study, we have identified high-risk communities that are typically obscured in county-level risk assessments and found that public schools, like private schools, are exhibiting predictable increases in vaccination exemption percentages. As public health agencies confront the re-emerging threat of measles and other vaccine-preventable diseases, findings such as ours can guide targeted interventions and surveillance within schools, cities, counties, and sociodemographic subgroups.

3.2 Acknowledgments

The original work for this project was done under the guidance of Drs. Lauren Castro and Lauren Ancel Meyers, in The University of Texas at Austin's Department of Integrative Biology. It has been published in the journal *PLOS Medicine* [25].

I wrote the mathematical supplement independently for the purpose of this thesis, with reference to two papers on beta regression [26, 27].

3.3 Mathematical Supplement

In this project, we used regression to identify sociodemographic factors associated with conscientious vaccination exemptions. Over the course of the project's development we explored a range of methods, including logistic regression and the LASSO. In the end, we decided to fit our models to proportions and to use beta regression.

Beta regression was proposed in 2004 by Silvia Ferrari and Francisco Cribari-Neto as a method to predict a bounded dependent variable [26]. Prior methods recommended transforming the bounded dependent variable

so it took values on the real line (instead of within an interval) and then using linear regression to predict the transformed dependent variable. However, these methods have two main shortcomings [27]. First, predicting a transformed dependent variable $y^* = g(y)$ means that the regression parameters are interpreted in terms of the mean of y^*, not the mean of y. Second, bounded random variables are often heteroskedastic, with more variation near the center of the interval than near the boundaries, and asymmetric. Standard linear regression methods which assume that variables are normally distributed do not account for these qualities.

The Beta Regression Framework

Ferrari and Cribari-Neto developed a regression method that assumes that the dependent variable y comes from the beta distribution (equation 3.1, figure 3.1). This assumption has two key advantages: beta-distributed random variables are bounded between 0 and 1 (and the density function can be shifted and scaled to fit an arbitrary finite interval) and the beta distribution need not be symmetric (figure 3.1B).

$$f(y|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \text{ for } p,q > 0 \tag{3.1}$$

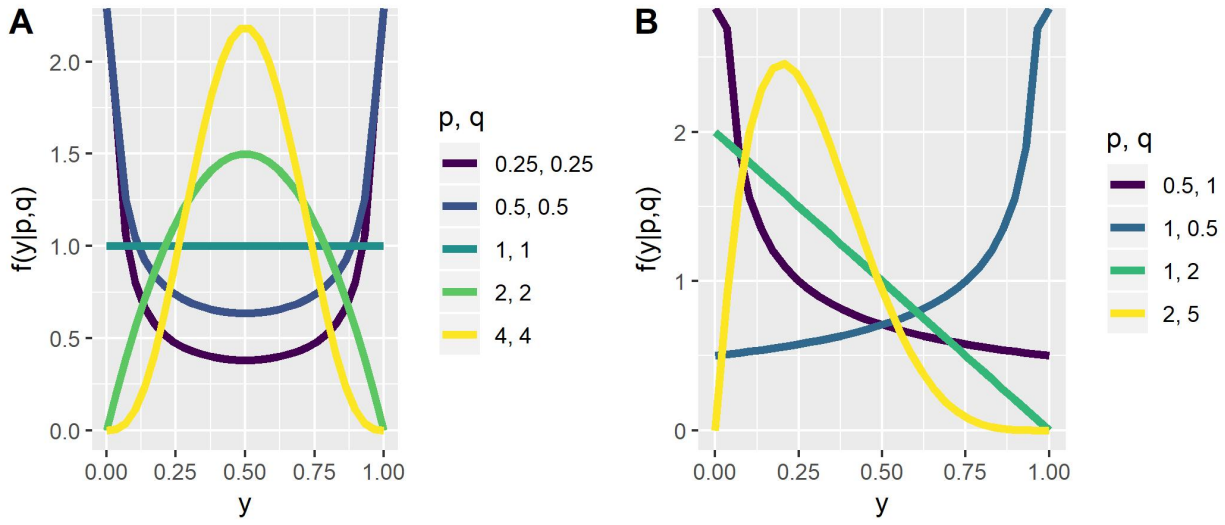

Figure 3.1: **The beta distribution (equation 3.1).** (A) The distribution is symmetric when $p = q$ and is uniform when $p = q = 1$. (B) The distribution is asymmetric when $p \neq q$.

While the beta distribution is parameterized in terms of p and q, regression methods typically employ densities whose parameters represent the mean and dispersal. Thus, the first step toward a regression method based on the beta distribution is to reparameterize the density function (equation 3.1) with parameters representing the mean and dispersal. In order to accomplish this, the authors define $\mu = E[y] = p/(p+q)$,

the mean of a beta-distributed random variable, and $\phi = p + q$. Because

$$Var[y] = \frac{pq}{(p+q)^2(p+q+1)} = \frac{\left(\frac{p(p+q\ \ p)}{(p+q)(p+q)}\right)}{1+(p+q)} = \frac{\left(\frac{p}{p+q}\right)\left(1-\frac{p}{p+q}\right)}{1+(p+q)} = \frac{\mu(1-\mu)}{1+\phi}$$

ϕ is inversely proportional to the variance of a beta-distributed random variable and consequently takes the role of a precision parameter.

If we note that $\mu\phi = p$ and $(1-\mu)\phi = q$, it is clear that we can rewrite equation 3.1 as

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi\ \ 1}(1-y)^{(1\ \ \mu)\phi\ \ 1} \text{ for } 0 < \mu < 1 \text{ and } \phi > 0 \tag{3.2}$$

Note that when $\mu = 1/2$, $p = q$ and the distribution is symmetric (figure 3.1A) regardless of ϕ.

In general, regression methods predict dependent variables $\{y_t|t = 1, ..., n\} = \{y_1, y_2, ..., y_n\}$ using predictors $\{x_t|t = 1, ..., n\} = \{(x_{t1}, x_{t2}, ..., x_{tk})|t = 1, ..., n\}$. Note that here, x_t is a vector of predictors associated with y_t. The matrix X whose t^{th} row is x_t is called a design matrix–this notation will return in the next section.

For example, in a linear regression model, we express each y_t as a linear combination of the predictors: $y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} + \epsilon = x_t^T \beta + \epsilon$, where ϵ is an error term and $\beta = (\beta_1, ..., \beta_k)$ is a vector of regression coefficients. The model's predicted value, $\bar{y}_t = \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk}$, is an estimate of the mean of y_t.

We express this relationship slightly differently for beta regression. Instead of predicting y_t directly, we predict $g(\mu_t)$, where $y_t \sim \mathcal{B}(\mu_t, \phi)$ (the beta distribution defined by equation 3.2) and g is a strictly increasing, twice-differentiable link function that maps the interval $(0, 1) \mapsto R$. A common g is the logit function: $g(\mu_t) = \log\left(\frac{\mu_t}{1\ \ \mu_t}\right)$. Similarly to linear regression, we estimate this quantity as a linear combination of predictors:

$$g(\mu_t) = \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} = x_t^T \beta \tag{3.3}$$

For the logit function, it follows that

$$\log\left(\frac{\mu_t}{1-\mu_t}\right) = x_t^T \beta \implies \frac{\mu_t}{1-\mu_t} = e^{x_t^T \beta} \implies \mu_t = \frac{e^{x_t^T \beta}}{1+e^{x_t^T \beta}}$$

The logit link function is popular because it makes the parameters, β, easily interpretable. Suppose we want to understand the effect of the parameter β_m. To explore this, we will increase one of the predictors, x_m, by some constant c, leaving all other predictors unchanged. That is, $x_m = x_m + c$. We denote the new predicted mean μ . We will now do some algebra with the goal of isolating β_m:

$$\begin{aligned}
\frac{\mu_t/(1-\mu_t)}{\mu_t/(1-\mu_t)} &= \frac{\exp\left\{\sum_{i \neq m} x_{ti}\beta_i + \beta_m x_m\right\}}{\exp\left\{\sum_{i=1}^{k} x_{ti}\beta_i\right\}} \\
&= \exp\left\{\left(\sum_{i \neq m} x_{ti}\beta_i + \beta_m(x_m + c)\right) - \left(\sum_{i \neq m} x_{ti}\beta_i + \beta_m x_m\right)\right\} \\
&= \exp(\beta_m c) \\
\implies \log\left(\frac{\mu_t/(1-\mu_t)}{\mu_t/(1-\mu_t)}\right) &= \beta_m c
\end{aligned}$$

Because $\mu_t, \mu_t \in (0,1)$, $\frac{\mu_t^*/(1 \quad \mu_t^*)}{\mu_t/(1 \quad \mu_t)}$ is an odds ratio. So, we can describe the parameter β_m in terms of the log odds of μ_t versus μ_t.

In our regression analysis of vaccination exemption rates, we employed a log-log link function $g(\mu_t) = -\log(-\log(\mu_t)) = x_t^T\beta$. Under this link function, the parameters have a less straightforward interpretation.

Maximum-Likelihood Parameter Estimation

The log-likelihood for β, ϕ given $y_1, ..., y_n$ and X (the $n \times k$ design matrix with rows $x_1, ..., x_n$) is:

$$\begin{aligned}
L(\beta, \phi | y_1, ..., y_n, X) &= \sum_{t=1}^{n} \log(f(y_t|\mu_t, \phi)) \text{ (where } \mu_t = g \ \ ^1(x_t^T\beta) \text{ by equation 3.3)} \\
&= \sum_{t=1}^{n} \log\Gamma(\phi) - \log\Gamma(\mu_t\phi) - \log\Gamma((1-\mu_t)\phi) + (\mu_t\phi - 1)\log(y_t) + ((1-\mu)\phi - 1)\log(1-y_t))
\end{aligned}$$

There is no closed-form way to find the maximum of this function, so we must employ numerical optimization methods. In the R package *betareg*, in which Cribari-Neto and Ferrari implement the beta regression method, they obtain the maximum likelihood parameter estimates by first computing the score function and then finding its maximum using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, which is a quasi-Newton method for unconstrained nonlinear optimization [27].

A score function is the gradient of a log-likelihood surface (here, the derivative of the log-likelihood with respect to the parameters β and ϕ). The score function is 0 at any parameter combination that locally maximizes the log-likelihood. Here, the score function is

$$(U_\beta(\beta, \phi)^T, U_\phi(\beta, \phi)^T)$$

where $U_\beta(\beta, \phi) = \phi X^T T(y \ \ -\mu \ )$ and $U_\phi(\beta, \phi) = \sum_{t=1}^{n}\{\mu_t(y_t - \mu_t) + \log(1-y_t) - \psi((1-\mu_t)\phi) + \psi(\phi)\}$ for

- $T = diag\{1/g^{\emptyset}(\mu_1), ..., 1/g^{\emptyset}(\mu_n)\}$

- X = the design matrix (the $n \times k$ matrix whose t^{th} row is x_t)
- $y_t = \log\{y_t/(1-y_t)\}$ and $\mu_t = \psi(\mu_t\phi) - \psi((1-\mu_t)\phi)$
- $\psi(x) = \frac{d}{dx}\log\{\Gamma(x)\}$ (ψ is called the digamma function)
- $y = (y_1, ..., y_n)^T$ and $\mu = (\mu_1, ..., \mu_n)$

The maximum-likelihood estimators of β and ϕ are obtained by using numerical methods to find the parameter combination where $(U_\beta(\beta,\phi)^T, U_\phi(\beta,\phi)^T) = (0,0)$. In the package *betareg*, this optimum is found through the R function *optim*'s implementation of the BFGS algorithm [19]. The BFGS algorithm begins at an initial estimate for β and ϕ, and iteratively approaches the peak using an approximation of the Hessian matrix which is updated at each step. This is far less computationally expensive than Netwon-type methods, which require the computation of the full Hessian matrix at each step.

To choose an initial value for β (the vector of regression parameters $\beta_1, ..., \beta_k$) the authors recommend using the ordinary least squares (OLS) estimate. We obtain this estimate by using maximum likelihood to fit an OLS regression model which uses X (the design matrix defined above) to predict the vector of transformed responses $z = (g(y_1), ..., g(y_n))^T$. In this framework, the maximum-likelihood parameter estimate is $\hat{\beta} = (X^TX)^{-1}X^Tz$. This calculation yields a vector of parameters that serve as a reasonable initial value.

Their estimation of an initial value for ϕ is slightly more involved. Recall that $var(y_t) = \mu_t(1-\mu_t)/(1+\phi)$. It follows that

$$\phi = \frac{\mu_t(1-\mu_t)}{var(y_t)} - 1 \tag{3.4}$$

To estimate μ_t, we use the OLS estimate of $g(\mu_t)$. That is, we estimate $\check{\mu}_t = g^{-1}(x_t^T\hat{\beta})$ where $\hat{\beta} = (X^TX)^{-1}X^Tz$ is the OLS maximum-likelihood parameter estimate, as defined above.

To estimate $var(y_t) = \sigma_t^2$, we first note that:

$$\begin{aligned} var(g(y_t)) &\approx var\{g(\mu_t) + (y_t - \mu_t)g^{\emptyset}(\mu_t)\} \quad \text{(by a first-order Taylor expansion)} \\ &= \left(g^{\emptyset}(\mu_t)\right)^2 \cdot var(y_t) \quad \text{(because } var(c+dz) = d^2var(z) \text{ for } c,d \in R \text{ and } z \text{ a random variable)} \\ \implies \check{\sigma}_t^2 &= \frac{var(g(y_t))}{(g^{\emptyset}(\mu_t))^2} \end{aligned}$$

$g^{\emptyset}(\mu_t)$ can be estimated using the OLS estimate, $\check{\mu}_t = g^{-1}(x_t^T\hat{\beta})$, as described above.

$var(g(y_t))$ is estimated as the OLS regression mean squared error. Let $\check{e} = z - X\hat{\beta}$, the vector of residuals from the OLS regression of $g(y)$. The i^{th} entry of $\check{e}$ is $\check{e}_i = g(y_i) - x_i\hat{\beta}$, where $g(y_i)$ is the observed value and $x_i\hat{\beta}$ is the value predicted by the OLS regression line. Thus the sum of squared residuals can be written as $\check{e}^T\check{e}$ and the OLS regression mean squared error can be written as: $var(g(y_t)) = \left(\check{e}^T\check{e}\right)/(n-k)$. So our

estimate for σ_t^2 is:

$$\check{\sigma}_t^2 = \frac{\check{e}^T \check{e}}{(n-k)} \cdot \frac{1}{(g^0(\check{\mu}_t))^2}$$

Putting all of these components back into the framework introduced in equation 3.4 and averaging across all values of t gives the following final estimate for ϕ:

$$\hat{\phi} = 1/n \cdot \sum_{t=1}^{n} \frac{\check{\mu}_t(1-\check{\mu}_t)}{\check{\sigma}_t^2} - 1$$

These estimates for β and ϕ are used to initialize the BFGS algorithm, which iteratively finds the maximum of the log-likelihood function. The optimal β and ϕ values found by this algorithm are then used to parameterize our final beta regression model.

Bibliography

[1] Greenbaum G, Getz WM, Rosenberg NA, Feldman MW, Hovers E, Kolodny O. Disease transmission and introgression can explain the long-lasting contact zone of modern humans and Neanderthals. Nature Communications. 2019;10(1). doi:10.1038/s41467-019-12862-7.

[2] Hippocrates. Ancient Medicine. Airs, Waters, Places. Epidemics 1 and 3. The Oath. Precepts. Nutriment.; 1923.

[3] Bernoulli D. Reflexions sur les avantages de l'inoculation. Mercure de France. 1760; p. 173–190.

[4] Ross R. An application of the theory of probabilities to the study of a priori pathometry.—Part I. Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character. 1916;92:204–230.

[5] Ross R, Hudson HP. An application of the theory of probabilities to the study of a priori pathometry.—Part II. Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character. 1917;93:212–225. doi:10.1098/rspa.1916.0007.

[6] Ross R, Hudson HP. An Application of the Theory of Probabilities to the Study of a priori Pathometry—Part III. Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character. 1917;93:225–240.

[7] Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character. 1927;115(772):700–721. doi:10.1098/rspa.1927.0118.

[8] Bailey NT. The Mathematical Theory of Epidemics; 1957.

[9] Hethcote HW. Asymptotic behavior in a deterministic epidemic model. Bulletin of Mathematical Biology. 1973;35(5-6):607–614. doi:10.1007/BF02458365.

[10] Hethcote HW. Qualitative analyses of communicable disease models. Mathematical Biosciences. 1976;28(3-4):335–356. doi:10.1016/0025-5564(76)90132-2.

[11] May RM, Anderson RM. Population biology of infectious diseases: Part II. Nature. 1979;280(5722):455–461. doi:10.1038/280455a0.

[12] Anderson RM, May RM. Population biology of infectious diseases: Part I. Nature. 1979;280(5721):361–367. doi:10.1038/280361a0.

[13] MacDonald G. The Analysis of Equilibrium in Malaria. Tropical Diseases Bulletin. 1952;49(9):813–829.

[14] Hethcote HW. The Mathematics of Infectious Diseases. SIAM. 2000;42(4):599–653. doi:10.1037/0003-066X.44.2.237.

[15] Metz JAJ. The Epidemic in a Closed Population with all Susceptibles Equally Vulnerable; Some Results for Large Susceptible Populations and Small Initial Infections. Acta Biotheoretica. 1978;27(1/2):75–123.

[16] Billard L, Lacayo H, Langberg NA. A new look at the simple epidemic process. Journal of Applied Probability. 1979;16(1):198–202. doi:10.2307/3213387.

[17] Wang X, Du Z, Pasco R, Pierce K, Petty M, Fox SJ, et al. COVID-19 Healthcare Demand Projections: Austin, Texas. Austin, TX: The University of Texas at Austin; 2020.

[18] Grimmett G, Stirzaker D. Probability and Random Processes. 3rd ed. New York: Oxford University Press; 2001.

[19] R Development Core Team 3 0 1 . A Language and Environment for Statistical Computing; 2013. Available from: http://www.r-project.org.

[20] Goulet V, Dutang C, Maechler M, Firth D, Shapira M, Stadelmann M. expm: Matrix Exponential, Log, etc; 2019. Available from: https://CRAN.R-project.org/package=expm.

[21] Johns Hopkins CSSE. 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository; 2020. Available from: https://github.com/CSSEGISandData/COVID-19.

[22] Texas Counties by Population; 2019. Available from: https://www.texas-demographics.com/counties_by_population.

[23] Keeling M, Rohani P. Modeling Infectious Diseases in Humans and Animals. Princeton, NJ: Princeton University Press; 2008.

[24] Wolfram Research I. Mathematica, Version 11.2; 2020. Available from: https://www.wolfram.com/mathematica.

[25] Morrison M, Castro LA, Ancel Meyers L. Conscientious vaccination exemptions in kindergarten to eighth-grade children across Texas schools from 2012 to 2018: A regression analysis. PLOS Medicine. 2020;doi:10.1371/journal.pmed.1003049.

[26] Ferrari SLP, Cribari-neto F. Beta Regression for Modelling Rates and Proportions Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics. 2004;31(7):799–815. doi:10.1080/0266476042000214501.

[27] Cribari-Neto F, Zeileis A. Beta regression in R. Journal of Statistical Software. 2010;34(2):1–24. doi:10.18637/jss.v034.i02.