A Multiplicative Model of the Transmission Rate
and its Statistical Inferences

by

Cong Cong Michelle Fan

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Public Health Sciences
University of Toronto

© Copyright by Cong Cong Michelle Fan 2001
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-63595-3
A Multiplicative Model of the Transmission Rate

and its Statistical Inferences

by

Cong Cong Michelle Fan

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Public Health Sciences
University of Toronto

Abstract

The transmission rate -- a key parameter in epidemic modelling -- is formulated as the product of a sequence of conditional probabilities based on the pathogenesis of an infection. These conditional probabilities represent the effectiveness of the host defence mechanisms (susceptibility), the amount of pathogen released by an infective (infectiousness), and the contact pattern of a susceptible and infectives. With this formulation, an incidence process is derived as the product of the transmission rate, the number of susceptibles, and a power function of the number of infectives under a general assumption of the infectiosity of the pathogen. A martingale is constructed as the difference of the counting process that counts the number of infections and the integral of the incidence process (the compensator). This martingale is used to
estimate the cumulative transmission rates at the observed infection times. The cubic spline method and the kernel method are used to further obtain an estimate of the transmission rate as a continuous function of time. To study population heterogeneity, two approaches are employed: (1) the martingale method is used to make estimation within each stratified population and (2) the Cox model is used to estimate the effects of various factors on the transmission rate.
Acknowledgement

My deepest appreciation goes to my supervisor, Prof. John J. Hsieh. This work would not be possible without his wisdom and insight. He taught me the subject; he taught me the research; and he provided me with an ideal learning environment. His support was generous, as a father's, and his interest in my success went beyond duty. For the years to come I shall remain a disciple of his immense wisdom.

My sincere gratitude also goes to my research committee members, Prof. Pauline van den Driessche and Prof. Paul N. Corey, for their involvement from the beginning of my research, for reading versions of my dissertation, and for making valuable corrections and instructive remarks. I also wish to thank Dr. Raymond Tellier from the Hospital for Sick Children who also contributed to this work.

I would like to thank the internal examiner, Prof. Michael D. Escobar, and the external examiner, Prof. David E. Matthews, from the University of Waterloo, for their valuable comments and suggestions.

I am very fortunate to have the guidance, support and encouragement from Prof. John J. Hsieh, Prof. Paul N. Corey, Prof. Michael D. Escobar, during my graduate studies. The financial support provided for my research by the University of Toronto, the Ontario Graduate Scholarship program, and MITACS is greatly appreciated.

My sincere thanks also go to the Department of Clinical and Medical Affair, Aventis Pasteur. This industry working experience has been invaluable for my research and it is the starting point of my career.

I am grateful to my friends and colleagues, Ruxandra Spijevic-Pinto, Qilong Yi, Aleksandra Kolenc-Saban, who supported me. I am particularly grateful to my friend and my former colleague from Aventis Pasteur, Gosia Jaskot who offered me her genuine friendship. And most importantly, she spent hours and hours after work to answer my questions in Microbiology and provided me very helpful textbooks that are referenced frequently in this thesis.

My husband has been most supportive of my quest and believed in me throughout my years of schooling. And finally my sincerest thought goes to my mother. Her hard work and determination has made the biggest impact in my life. To my mom, I dedicate this thesis.
# Table of Contents

1. **Introduction and Literature Review** ......................................................... 1
   1.1 **General Introduction** ................................................................. 1
   1.2 **Mathematical Epidemic Modelling** ............................................... 6
      1.2.1 **The Deterministic Approach** ............................................... 8
      1.2.2 **The Stochastic Approach** .................................................. 9
      1.2.3 **Modifications to the SIR Model** ........................................ 9
   1.3 **Statistical Inferences** ................................................................. 11
   1.4 **Applications of the Transmission Rate** ....................................... 12

2. **A Multiplicative Model of the Transmission Rate and Derivation of Incidence Processes** .................................................. 16
   2.1 **Transmission Rate** ................................................................. 16
   2.2 **Pathogenesis of an Infection** ................................................... 19
   2.3 **Issues in Modelling and Data Availability** .................................. 21
   2.4 **Multiplicative Models in Statistics** .......................................... 22
   2.5 **Contact Initiation Process** ....................................................... 23
   2.6 **Incidence Process under Basic Assumptions** ............................ 25
   2.7 **Low and High Infectiosities** .................................................... 29
   2.8 **Population Heterogeneity** ....................................................... 32
   2.9 **Time Homogeneity** ................................................................. 33
   2.10 **SIR and SEIR Models** ............................................................. 34

3. **Non-Parametric Inferences** ................................................................. 36
   3.1 **Martingale and Counting Process Review** .................................. 37
   3.2 **Martingale Inferences for the Transmission Rate** ....................... 39
      3.2.1 **SIR Models** ................................................................. 40
      3.2.2 **SEIR Models** ............................................................... 42
   3.3 **Estimating the Transmission Rate as a Continuous Function** ....... 43
      3.3.1 **The Cubic Spline Method** .............................................. 43
      3.3.2 **The Kernel Smoothing Method** ....................................... 44
   3.4 **Estimating the Number of Adequate Contacts** ........................... 45
   3.5 **Applications** ............................................................................ 47
      3.5.1 **Smallpox Outbreak** ....................................................... 47
3.5.2 Respiratory Disease Epidemic .................................................................51
3.6 Population Heterogeneity – Stratified Analysis ........................................54
  3.6.1 Stratified Analysis – Respiratory Disease Epidemic ..............................55
  3.6.2 Relative Risk Calculation – Respiratory Disease Epidemic .................57
4 Infection Tables .................................................................................................................59
  4.1 Waiting Time to Infection ..................................................................................59
  4.2 Other Functions of the Waiting Time to Infection .........................................61
  4.3 Predicting Epidemic .........................................................................................63
  4.4 Goodness-of-fit .................................................................................................65
    4.4.1 Smallpox Outbreak ..................................................................................66
    4.4.2 Respiratory Disease Epidemic .................................................................72
4.5 Construction of Infection Tables ...........................................................................75
4.6 Applications of Infection Tables ...........................................................................76
5 Multiplicative Model in Studying Observed Heterogeneity ......................79
  5.1 A Survival Analysis Review .............................................................................79
  5.2 The Cox Model Applied to Epidemic Data ..................................................81
  5.3 Generalized Baseline Function for Cox Model ............................................83
  5.4 Semi-Parametric Estimate of the Transmission Rate .................................86
  5.5 Application - Respiratory Disease Data .......................................................87
    5.5.1 Without Time-Dependent Covariates ....................................................88
    5.5.2 With Time-Dependent Covariates ..........................................................90
6 Conclusions and Further Suggestions ...............................................................93
  6.1 Conclusions ......................................................................................................93
  6.2 Limitations and Further Suggestions .............................................................96
References .......................................................................................................................99
Appendices ..................................................................................................................107
List of Tables

Table 3-1 Parameter estimation using GLM – smallpox outbreak .............................................. 47
Table 3-2 Parameter estimation using GLM – respiratory disease epidemic .............................. 47
Table 3-3 Estimate and variance of $\tilde{B}(t)$ – smallpox outbreak ........................................ 50
Table 3-4 Estimate and variance of $\tilde{B}(t)$ – respiratory disease epidemic .......................... 53
Table 3-5 Estimate and variance of $\tilde{B}(t)$ by age group – respiratory disease epidemic .......... 56
Table 3-6 Estimated overall transmission rates $\beta$, RR and the test statistic for $RR = 1$ .......... 58
Table 4-1 Expected and observed total number of infections – smallpox outbreak ...................... 68
Table 4-2 Expected and observed total number of infections over the course of the epidemic – respiratory disease epidemic ($m = 1$) ........................................................................................................... 73
Table 4-3 Infection table – respiratory disease epidemic ($m = 1$) ............................................. 77
Table 5-1 Model estimations – without time-dependent covariates ........................................... 89
Table 5-2 Baseline transmission rate estimation ........................................................................ 89
Table 5-3 Estimated transmission rates by age ........................................................................... 90
Table 5-4 Model estimations – with time-dependent covariates .............................................. 92
List of Figures

Figure 2-1 Pathogen multiplication results in different outcomes (Mims et. al., 1993) .......... 17
Figure 3-1 Cubic spline estimate of the transmission rate - smallpox outbreak.......................... 51
Figure 3-2 Cubic spline estimate of the transmission rate - respiratory disease epidemic......... 54
Figure 3-3 Cubic spline estimate of transmission rate by age group - respiratory disease epidemic .................................................................................................................. 57
Figure 4-1 Expected and observed total number of infections - smallpox outbreak \( m = 1 \) .... 70
Figure 4-2 Expected and observed total number of infections - smallpox epidemic \( m = 0.8 \) ... 71
Figure 4-3 Expected and observed total number of infections - respiratory disease epidemic \( m = 1 \) ........................................................................................................................................ 74
Figure 4-4 Lorenz curve and \( STTT \) - respiratory disease epidemic \( m = 1 \) ....................... 78
List of Appendices

A. Population classification – respiratory disease epidemic data.................................107
B. Longitudinal data – respiratory disease epidemic data.............................................108
1 Introduction and Literature Review

Infectious disease can drastically change the morbidity and mortality of a population, and therefore, play an important role in human life and the course of human history. Although vaccines are available for many infectious diseases (for example, polio, hepatitis B and measles), some of these diseases still have direct impacts in some parts of the world. So, it is important to understand the process of propagation of infectious diseases in order to control their spread. At the population level, the infection process manifests itself as an epidemic, which has been studied by three main approaches: observational, experimental and mathematical. The mathematical modelling of an epidemic in this thesis is concerned with the infection processes mainly via close person-to-person contact within a population. Modelling of an infectious disease is a rather involved task because transmission interactions are very complex. The transmission rate $\beta(t)$, which is one of the key parameters in epidemic models, defines the rate of infection transmission at time $t$ through direct contact between infectives and susceptibles. The focus of this work is to study the disease propagation by suggesting a probabilistic model of the transmission rate and by applying statistical methods to make inferences. This chapter will provide an introduction to the probability models and methods developed in this dissertation, followed by a review of the epidemiology of infectious diseases and the necessary ingredients (hosts, agents and the environment) affecting the spread of an infectious disease. Section 1.1 gives an outline description of this dissertation. The literature review on epidemic modelling given in the rest of this chapter will focus on existing works on the relevant mathematical epidemic modelling, the applications of statistical methods on infectious disease data, and the applications of the transmission rate.

1.1 General Introduction

Most epidemic models discussed in the literature use the law of mass action and its generalizations, which fail to connect the effects of hosts, agents and the environment on the spread of an infectious disease. This work integrates knowledge from various disciplines including epidemiology, microbiology, probability theory, stochastic processes and statistics to
model the infection transmission, to estimate the transmission rate, and to study the important ingredients of an epidemic (hosts, agents and the environment).

Chapter 2 will introduce a probability formulation of the transmission rate $\beta(t)$ based on the pathogenesis of an infection. The transmission rate for a given type of pathogen will be broken down into three components: (1) host factors – susceptibility, (2) agent factors – infectivity, and (3) contacts initiated by infectives with susceptibles – contact pattern. Because pathogen is carried and transmitted by infectives for person-to-person transmitted diseases, the infectivity is decomposed further into two parts. One is the infection generating power of the pathogen, its infectiosity (Tellier, 2000); the other is the variable amount of pathogen carried by an infective individual, infectiousness. Environmental conditions, which affect the survival of the pathogens, are less important for diseases that are transmitted via close person-to-person contact. So, they are not considered in this formulation. This probability formulation of the transmission rate will serve as a foundation for model developments.

The logic for this foundation comes from the pathogenesis of an infection. Upon contact by an infective, a susceptible host may or may not become infective, and so may or may not further cause disease in other susceptibles. Infections can be conceptually viewed as directly related to the infectiosity of the pathogen and the amount of pathogen present, but inversely related to host defence mechanisms; this is known as the pathogenesis of an infection. For a given pathogen, the infection transmission is the intersection of the following three events: (1) one or more contacts initiated by infectives with a susceptible, (2) a certain amount of pathogen is emitted by infectives in these contacts, (3) a certain amount of pathogen starts to multiply within the susceptible host body. Consequently, the probability of an infection transmission is the product of the appropriate conditional probabilities derived from these three events. As this multiplicative model reflects the underlying disease spreading mechanism and the complex interplay of the host and the pathogen, it is more realistic than many existing models that assume the transmission rate to be independent of these factors.

An infection can only occur if there are contacts between a susceptible and one or more infectives. The population contact process is a point process, whether the infection transmission occurred at the contact time is treated as a mark, and together gives a marked point process. An infection process is a compound point process determined by the marked point process. The stochastic incidence process, which define the expected number of new infections that occur
instantaneously within a population given the strict past, are then derived under various assumptions in Chapter 2. A general incidence process $\dot{B}(t)S(t)I^\pi(t)dt$ is obtained.

Stochastic models are often formulated as systems of stochastic differential equations. Rather than solving a system of stochastic differential equations to obtain desired estimates (which can be extremely difficult), statistical methods provide an efficient way of making inferences based on an intricate probability formulation. Counting processes and the martingale theory have been applied in survival analysis by Fleming and Harrington (1991) and Andersen et. al. (1993), and have been adopted in epidemic modelling for a special case (the bilinear incidence rate for an SIR model) by Becker (1989, 1993). In Chapter 3, using the counting process and the incidence processes derived in Chapter 2, the martingale method will be applied to general situations that are extensions of Becker's work. The martingale method will also be applied to SEIR models, which are more realistic models than SIR models.

It is desirable to treat the transmission rate as a continuous function of time, which provides a direct and precise measurement of the disease spread over time. To estimate this transmission rate, Becker (1989) uses the Epanechnikov kernel function to smooth the increments in the cumulative transmission rate estimated at observed infection times from the martingale method. Since smoothing tends to reduce the goodness-of-fit, the cubic spline method will be used in Chapter 3 as an alternative to interpolate the cumulative transmission rate at the observed infection times. The derivative of this cubic spline gives an estimate of the transmission rate, which is a continuous quadratic spline that is also first order differentiable. Data from a respiratory disease epidemic and from a smallpox outbreak, which have been studied by many investigators, including Bailey (1975) and Becker (1989), will be used to illustrate the estimation of the transmission rate. In general, the exact time of infection is not observable, but it is usually approximated using appropriate assumptions of the latent and infectious periods. The latent period is the time interval during which the development of the pathogen takes place purely internally, without the emission of any kind of infectious material. The infectious period is the interval during which the pathogen may be emitted.

Host factors (for example, age, previous infection and genetic make-up) that are determinants of an epidemic may create heterogeneity in susceptibles. Household environmental factors (whether a household is infected or how long the household has been affected at a given time) may also create heterogeneity in susceptibles. Heterogeneity also exists among infectives
because infectives at different disease stages have different levels of microbes within their systems. For each infection, which infective is responsible for the infection is not observable, so that estimating the effect of infectives on the transmission rate would be impossible from incidence data. If, however, the infectiousness is replaced by an average infectiousness of the population (or by assuming that infectives are homogeneous), the host factors and the household environmental factors, which may affect the transmission rate, may be studied from incidence data using statistical models. A stratified analysis is useful when dealing with one or two factors. A heterogeneous susceptible population is stratified into a small number of strata according to these factors, and the martingale method will be applied within each stratum in which individuals are homogeneous. The respiratory disease epidemic data have information on each susceptible, and they will be used to illustrate the stratified analysis in Chapter 3.

The waiting time to infection $T$ of each susceptible during an epidemic is the random variable of interest for which the transmission rate $\beta(t)$ is defined at time $T = t$. The hazard function of $T$ or the infection rate, which defines the instantaneous rate of infection per unit time for each susceptible, is the product of the transmission rate and the expected number of infectives in the population. It is estimated by the product of the estimated transmission rate and the observed number of infectives. Other life table functions depicting the distribution of $T$, such as the probability of escaping infection, the density, the conditional infection probability, the expected number of infections within a time interval, the average waiting time to infection, the infection-free expectancy and the two derived functions, the Lorenz curve and the scaled-total-time-on-test, will be estimated from the estimated hazard function. These functions are of epidemiological significance and are useful in constructing test statistics.

Infection tables, which consist of the above estimated functions, are similar to life tables used in demography, (Chiang, 1968 and Hsieh, 1991b). Infection tables provide more information than it is offered by just the estimated transmission rate that is conventionally given in the literature. The construction of an infection table is a new application of the existing method in demography. Data from the respiratory disease epidemic are used to illustrate the construction of the infection table.

Predicting an epidemic will also be discussed in Chapter 4. Given a transmission rate over time and the number of infectives and susceptibles at the beginning of the epidemic, the disease spreading process will be predicted using the two functions of $T$ (the hazard function and
conditional infection probability). Assumptions regarding the latent and infectious periods are also required. Two situations will be discussed in Chapter 4: (1) constant latent and infectious periods, (2) a constant latent period and a random infectious period. Predicting the number of new infections over the course of an epidemic are useful for health officials to estimate the total incidence and the costs incurred during epidemic attacks. This prediction procedure will also be used to check the goodness-of-fit of a model by comparing the observed and predicted number of new infections over the epidemic period.

A stratified analysis is performed in Chapter 3 to study heterogeneity introduced by susceptibles. However, the stratification method becomes inefficient when too many factors are considered simultaneously. In Chapter 5, the well-known Cox model that handles the observable heterogeneity in survival analysis is extended to model heterogeneity in the context of disease spreading process. In survival analysis, factors that make a population heterogeneous are called covariates. The same definition of covariates will also be used in the context of epidemic modeling. The hazard function of the time to infection for a heterogeneous population is the product of the hazard function for a homogeneous population, also known as the baseline hazard, and an exponential function for the additional risk introduced by covariates.

The number of infectives constitutes part of the baseline function, and thus, the baseline hazard is different for different model assumptions. Under the Greenwood assumption, the infection rate is independent of the number of infectives. Under the Reed-Frost assumption, the infection rate depends linearly on the number of infectives. Under a general assumption of the infectiosity, the infection rate is a power function of the number of infectives. Therefore, the likelihood function is different for different model assumptions. However, the effects of covariates will be estimated using a partial likelihood function in which the baseline hazard will not be included. Consequently, the estimates of the covariate effects will be the same for all models. The data from the respiratory disease epidemic will be used to fit this statistical model and to estimate the effects of covariates. Furthermore, the transmission rate will be estimated adjusting for covariates in Chapter 5.

Chapter 6 summarizes the results from previous chapters, discusses limitations of this work, and suggests further research problems.
1.2 Mathematical Epidemic Modelling

During the scientific revolution of the 1600s, as mathematical relationships are used to study the physical universe, many scientists reasoned that similar relationships must exist in the biological world. They originated the "Law of mortality" which formed one of the bases of mathematical application in biology, that is the life table. John Graunt constructed the first known life table at that time (Lilienfeld and Stolley, 1994). Another significant contribution to the development of mathematical epidemiology is due to Daniel Bernoulli. In 1760, Bernoulli concluded that inoculation not only protects against smallpox, but also confers life-long immunity. He used a life table to estimate the life expectancy if smallpox were eliminated from the population, and determined that inoculation at birth would increase the life expectancy (Gani, 1978).

The earliest contribution to modern mathematical epidemiology came from En'ko (1889). He formulated the progression of an epidemic in a similar approach as done by Reed and Frost, that is, the incidence is proportional to the number of susceptibles as well as the number of infectives. Around 1928, Reed and Frost introduced a simple chain binomial model to study the spread of infectious diseases in their lectures. Their work is later reported in the literature by, for example, Wilson and Burke (1942) and Abbey (1952). Greenwood (1931) independently formulated a different chain binomial model in which the incidence is proportional to the number of susceptibles alone. In chain models, cases (infections) are classified by generations and so time is considered discrete. Gart (1972) and Becker (1981) also introduced a multi-parameter chain binomial epidemic model, which included the Reed-Frost and Greenwood models as special cases. Chain binomial models are applied to diseases such as measles and the common cold (Lidwell and Somerville, 1951, Brimblecombe, 1958, and Heasman and Reid, 1961). The formulation is straightforward, and the classic likelihood method is applied to obtain estimates.

In the 1920s, Ross, Kermack and McKendrick suggested using compartment models to study the spread of infectious diseases. Their approach formed the foundation of the mathematical epidemic modelling. During an epidemic the population at any time consists of different proportions of susceptibles, infectives and possibly others (latent individuals and recovered individuals). Each compartment specifies a class of individuals (a compartment is referred to as a state in the stochastic context). Kermack and McKendrick's deterministic threshold theorem provides a mathematical foundation for the basis of disease control programs.
Contacts between susceptibles and infectives may lead to infection, and infectives may recover at different times after they become infective. This dynamics is stochastic in nature, but, for a large population, the statistical fluctuations may be ignored and the change in the size of each compartment becomes deterministic. So there are two basic ways to formulate this dynamics: *deterministic* and *stochastic*. Changes in the number of individuals in each state (compartment) is formulated as a system of stochastic differential equations, and in most case deterministic models is obtained by replacing the expected value in stochastic models by the state variable.

Various compartment models have been suggested for studying person-to-person transmitted diseases, but most of these models are based on the *general epidemic model*, also known as the *SIR* model, with three compartments:

(a) Susceptibles (*S*) are individuals who are capable of contracting the disease and becoming infectives.

(b) Infectives (*I*) are individuals who had contracted the disease and are able to infect others.

(c) Recovered (*R*) are individuals who have recovered with life-long immunity to the diseases.

Exclusion of mortality is valid when the disease is non-fatal and the epidemic duration is short. In the sequel, *S(t), I(t)* and *R(t)* denote the size of the states *S, I* and *R* at time *t* respectively. Assume there are *n* initial susceptibles, *a* initial infectives and no recovered (immune) individuals at the beginning of the epidemic. The *SIR* model assumptions are (Hethcote, 1989): at each *t*,

(1) The population is closed: there is no birth or death, that is, *S(t) + I(t) + R(t) = n + a*.

(2) The population is homogeneously mixed: the susceptibles move from *S* to *I* at a rate proportional to the number of susceptibles and the number of infectives, namely, \( \beta S(t)I(t) \).

(3) The infectives move from *I* to *R* at a rate proportional to the number of infectives, namely, \( \gamma I(t) \).

(4) The latent period is of zero length: a susceptible host who contacted an infective becomes infective immediately.
The basic model parameters are $\beta$ and $\gamma$, where $\beta$ is the constant transmission rate (sometimes referred to as the contact rate in deterministic models), and $\gamma$ is the constant recovery rate (it becomes the removal rate if the last compartment consists of the removed individuals rather than recovered individuals). Assumption (2) implies that all susceptibles are identical and independent, while assumption (3) implies all infectives are identical and independent.

1.2.1 The Deterministic Approach

A deterministic model assumes that for given transmission and recovery rates, the number of new infectives and new recoveries at any time are certain. The system of differential equations for an SIR model is (Bailey, 1975):

$$\frac{dS(t)}{dt} = -\beta S(t)I(t)$$

$$\frac{dI(t)}{dt} = \beta S(t)I(t) - \gamma(t) \quad (1-1)$$

$$\frac{dR(t)}{dt} = \gamma(t)$$

As the change of $S(t)$ is linear in $S(t)$ and $I(t)$ (see the first differential equation), it is often referred to as a bilinear incidence function.

The solution to the above differential equation approximated by using the Taylor's expansion of the exponential term, was first obtained by Kermack and McKendrick (1927a, b). The solution is given in terms of the relative removal rate ($\rho = \gamma / \beta$) as

$$R(t) = \rho^2 \left( \frac{n}{\rho} - 1 + \alpha \tanh(\alpha \frac{t}{2} - \nu) \right) \quad (1-2)$$

where $\alpha = \sqrt{(n/\rho - 1)^2 + 2na/\rho^2}$ and $\nu = \tanh^{-1}((n/\rho - 1)/\alpha)$. Kermack and McKendrick (1927a, b) also provide approximate solutions for the case in which the transmission and recovery rates depend on the infectious period.
1.2.2 The Stochastic Approach

Sizes of the states at any time $t \{S(t), I(t), R(t)\}$ form a multi-dimensional continuous-time discrete-state stochastic process, which reduces to a two-dimensional process $\{S(t), I(t)\}$ under assumption (1). The Reed-Frost assumption determines the infinitesimal probabilities as follows (Becker, 1989).

\[
\begin{align*}
P(dS(t) = -1, dI(t) = 1 \mid F_t) &= S(t^-)I(t^-)\beta dt + o(dt) \\
P(dS(t) = 0, dI(t) = -1 \mid F_t) &= I(t^-)\gamma dt + o(dt) \\
P(dS(t) = 0, dI(t) = 0 \mid F_t) &= 1 - S(t^-)I(t^-)\beta dt - I(t^-)\gamma dt + o(dt) \\
P(dS(t) = -r, dI(t) = r \mid F_t) &= o(dt), \text{ for } r > 1
\end{align*}
\]

where $F_t$ is a filtration, which is an increasing family of right continuous sub-$\sigma$-algebras and contains information up to just before $t$. The system of equations (1-3) is useful for martingale inferences, in which the last condition implies no simultaneous changes of states (to satisfy the definition of a counting process).

Let $p_{i,j}(t)$ denote the probability that at time $t$ there are $i$ susceptibles and $j$ infectives, that is, $p_{i,j}(t) = P(S(t) = i, I(t) = j)$. The system of stochastic differential equations can also be formulated in terms of state probabilities (Bailey, 1975). The generating function and the Laplace transform are used to obtain the solution to this differential-difference equation. Unfortunately, no completely satisfactory way has been suggested for solving this equation to obtain explicit expressions of individual probabilities, moments, distribution of duration time, epidemic curves, and other parameters. Even approximations of the above quantities are clumsy in their expressions and difficult to use (Bailey, 1975).

1.2.3 Modifications to the SIR Model

The latent period is an important epidemiological feature of an infectious disease, which is omitted by the SIR models for computational simplicities. It would be more realistic to include this state in the model, as the latent state $E$, thus, resulting in an SEIR model. This compartment contains individuals who have the microbes in their system, but are not yet able to transmit them to others. During the latent period, the microbes multiply within the host system until they reach
a threshold level, at which the individual is capable of releasing a sufficient amount of such microbes to others. When the microbes reach the threshold level, the latent individual then moves from state $E$ to state $I$. With this additional state, assumption (2) and (4) are modified as follows

(2a) Susceptibles move from $S$ to $E$ at a rate proportional to state sizes of the susceptibles and the infectives, $\beta S(t)I(t)$.

(4a) The latent period is not zero; latent individuals move from $E$ to $I$ at a rate proportional to the state size, $\xi E(t)$.

An SEIR model has been discussed in a deterministic setting in the literature (Liu et al., 1987 and Hethcote and van den Driessche, 1991). However, its application in statistical inferences is rather limited due to its complicated nature and the limited availability of data.

The population under study is generally heterogeneous that needs to be divided into homogeneous sub-populations, which can be determined, not only by intrinsic factors, but also by extrinsic factors and agent-related factors, such as the mode of transmission, latent period, infectious period, genetic susceptibility and resistance. Rushton and Mautner (1955) discussed a deterministic simple epidemic model in a heterogeneous population. Haskey (1954) first extended the stochastic simple epidemic model to study a heterogeneous population.

The basic model parameters may be time dependent, in which case the constant parameters in assumption (2), (2a), (3) and (4a) are replaced by $\beta(t)$, $\gamma(t)$, and $\xi(t)$. Yang and Chiang (1972) provided some discussions on such complicated models. Becker (1989) applied statistical methods to estimate $\beta(t)$.

The bilinear incidence assumption is the result of the law of mass action, first discussed by Wilson and Worcester (1945a, b). Many researchers have modified the bilinear incidence to a non-linear incidence. The non-linear incidence in the stochastic approach is discussed by Severo (1969) as follows. The "probability" of a new infective in $[t, t+dt)$ has the form $\beta S(t)I(t)\gamma dt$, where $\beta$ is the transmission rate, $a$ is the infection power and $b$ is the safety-in-number power. The "probability" of any infective being removed in $[t, t+dt)$ is $\gamma(I(t))^{c} dt$, where $\gamma$ is the removal rate or recovery rate and $c$ is the removal power. Liu et al. (1986, 1987), Hethcote et al. (1989), Hethcote and van den Driessche (1991), Hochberg (1991) and many others, use various function forms as the incidence in the deterministic settings.
1.3 Statistical Inferences

Because the spread of a disease is a complicated process, assumptions must be made to simplify it, so that the process may be described using a few epidemiologically meaningful parameters. Statistical inference procedures are needed to perform estimations and to evaluate model assumptions. *Chain binomial* models classify infection cases within a group, such as a household, into generations, where infectives are assumed to recover at the end of each generation. The maximum likelihood estimates of the infection transmission probability from one generation to the next are obtained for different models. The Pearson Chi-square test is used to check the model fit in Becker (1981, 1989).

Longini and Koopman (1982) used the likelihood function of final size of household infection to obtain the maximum likelihood estimates of the probability of escaping infection within the community and the probability of escaping infection within a household. Logini-Koopman model does not require any assumptions concerning the length or the distribution of the latent and infection periods. Haber et. al. (1988) and Longini et. al. (1988) extended this model to include individual risk factors into the probability of escaping infection. Haber et. al. (1991) extended this model further by allowing stratification in susceptibility. This approach falls into the category of Binary regression.

Becker and Hopper (1983) presented a method to assess the heterogeneity of the spread of a disease in a single epidemic or series of epidemics through a community. The heterogeneity of the population is assessed by using a two-dimensional contingency table. The method employed is essentially the Cox-Mantel test (Cox and Oakes, 1984), which is equivalent to the log-rank test.

Grenfell and Anderson (1985) described a maximum likelihood method for estimation of the age-related rates of infection. Becker (1986, 1989) applied the *Generalized Linear Model* (GLM) to epidemic modelling. The Poisson regression model is used to analyse data from a single epidemic in a large community where the observed infection times are used. The transmission rate is estimated as either time-dependent or time-independent. These statistical methods provide estimates of the basic model parameters and their variances, so that hypothesis testing can also be performed.
Becker (1989, 1993) discussed the non-parametric method, the martingale method, to estimate the transmission rate as a continuous function of time. The main benefit of this method is that statistical inferences about certain parameters, such as the transmission rate, can be derived without making a complete specification of the epidemic model. The martingale method is applied to study SIR models. Times of infection that are generally not observable are approximated assuming the fixed infectious period assumption. The transmission rate as a continuous function of time is estimated using kernel smoothing of the martingale estimates of the cumulative transmission rate.

Rhodes et. al. (1996) assumed the contact process of an infective \( i \) and a susceptible \( j \) has intensity \( \lambda_{ij}(t) \). The infection process is treated as the thinned version of the contact process with intensity \( \lambda_{ij}(t)p(t; z_i, z_j, \theta) \), where \( z_i \) is the covariates associated with the infective, \( z_j \) is the covariates associated with the susceptible, \( \theta \) is a vector of unknown parameters, and \( p(t) \) is the probability that a contact between \( i \) and \( j \) at time \( t \) would result in an infection. This approach is modified in this dissertation.

Rhodes et. al. (1996) also discussed in detail the three categories of statistical models: Poisson regression, Cox regression, and binary regression models. Statistical analyses are described when different amount of information, regarding contacts, time of infection and infectious period, are available. The likelihood functions on different levels of availability of information are constructed using the properties of marked and thinned counting processes. So, the score function for each parameter and the information matrix are derived.

1.4 Applications of the Transmission Rate

The deterministic threshold theorem of Kermack and McKendrick (1927a, b) for the general deterministic epidemic model maintains that the relative removal rate \( \rho (\rho = \gamma / \beta) \) is a threshold density for the initial size of the susceptible population, and at the end of the epidemic the total number of infections or removals is \( \lim_{t \to \infty} R(t) = 2 \left( \rho - \frac{\rho}{n} \right) \). The population size is taken as the number of individuals in a unit area, which is in fact the population density. If \( n < \rho \), an epidemic will not take place. If a small number of infected individuals are introduced, an epidemic will start to build if, and only if, \( n > \rho \). If \( n \) just exceeds the threshold density
\( n = \rho + \varepsilon \), then the resulting total size of the epidemic is approximated by \( \lim_{t \to \infty} R(t) = 2\varepsilon \), and the susceptible population size reduces to \( \lim_{t \to \infty} S(t) = \rho - \varepsilon \). Notice that the susceptible population is as far beneath the threshold value as it was above it originally.

Whittle (1955) derived the stochastic threshold theorem corresponding to the deterministic threshold theorem. Instead of stating that the epidemic will or will not occur, the stochastic threshold theorem gives the probability that an epidemic of not more than a given intensity will occur. The intensity of an epidemic is the proportion of the total number of susceptibles that finally contracts the disease. If \( \pi_j \) denotes the probability that not more than a proportion \( j \) of the \( n \) susceptibles are eventually infected, the theorem states that:

1. If \( \rho < n(1-j) \), then \( \left( \frac{\rho}{n} \right)^a \leq \pi_j \leq \left( \frac{\rho}{n(1-j)} \right)^a \);
2. If \( n(1-j) \leq \rho < n \), then \( (\rho/n)^a \leq \pi_j \leq 1 \);
3. If \( \rho \geq n \), then \( \pi_j = 1 \),

where \( a \) is the initial number of infectives as defined earlier. This theorem says that for small intensity with \( \rho < n \), the probability of an epidemic is approximately \( 1 - (\rho/n)^a \), rather than 1, as in the deterministic case. The threshold theorem is an important result in mathematical epidemic modelling.

The basic reproduction rate \( R_0 \) is the average number of secondary infections produced when one infected individual is introduced into a population where everyone is susceptible (Anderson and May, 1991) during his/her infectious period. When a disease starts to spread among a population, the number of susceptibles decreases. Eventually equilibrium may be attained, with the rate of infection being balanced against a rate of the appearance of new susceptibles. At equilibrium, each infection will on average reproduce exactly one secondary infection. The basic reproduction rate is an important concept in biology because it connects with the fitness of evolutionary theory. A parasitic species must have \( R_0 > 1 \) if it is capable of establishing itself in a host population. Estimating the basic reproductive rate has been discussed by Anderson and May (1991).

In general, the reproductive rate is estimated indirectly. If the population is homogeneously mixed, then the number of secondary infections produced by an infective will be
linearly proportional to the probability that any random contact initiated by this infective is with a susceptible. Therefore, the effective reproductive rate $R$ must equal the basic reproductive rate discounted by the fraction of the susceptibles $x$; that is $R = R_0 x$. Using the fact that at equilibrium each infection will on average reproduce exactly one secondary infection, the following relation can be used to estimate $R_0$

$$R_0 x^* = 1 \quad (1-4)$$

where $x^*$ is the fraction of susceptibles at equilibrium. By setting the second equation of (1-1) equal to zero, $x^*$ can be approximated as

$$x^* = \frac{\gamma}{n\beta} = \frac{\rho}{n} \quad (1-5)$$

Therefore, the basic reproductive rate is given as

$$R_0 = \frac{n}{\rho} \quad (1-6)$$

The deterministic threshold theorem can also be stated as follows: if $R_0 < 1$, an epidemic will not take place; if $R_0 > 1$, an epidemic will start to build.

For the SIR model defined in (1-1) with the transmission rate and the recovery rate being time-dependent, the basic reproductive rate may be estimated directly as

$$R_0 = n \int_0^\infty e^{-\int_0^u \beta(t) dt} \beta(t) dt \quad (1-7)$$

where $e^{-\int_0^u \beta(t) dt}$ is the probability that an infective is still capable of infecting others at time $t$, $\beta(t)$ is the force of transmitting the disease to a susceptible, and $n$ is the number of susceptibles at the time this infective is introduced to the population. When the recovery rate and transmission rate are both constants, (1-7) simplifies to (1-6). If mortality is incorporated into the SIR model with the death rate $\mu(t)$, the infective has to be alive and within his infectious period to infect others, then (1-7) is modified as

$$R_0 = n \int_0^\infty e^{-\int_0^u \beta(t) + \mu(t) dt} \beta(t) dt \quad (1-8)$$
Here the transmission rate is used to obtain the basic reproductive rate.

If a proportion \( f \) of the population is vaccinated, then the susceptible population is approximated by

\[
x = 1 - (VE) f \quad (1.9)
\]

where \( VE \) denotes the vaccine efficacy. Estimating vaccine efficacy has been discussed in the literature by many researchers, for example, Smith et. al. (1984), Halloran et. al. (1992), Longini and Halloran (1996). Using the threshold density of susceptibles given by (1-5), the vaccine fraction that is necessary to prevent an epidemic from occurring can be approximated by setting \( x \) to be \( x^* \), which yields \( VE = (n\beta - \gamma)/(n\beta f) \). Thus, the transmission rate is also useful in estimating the vaccine efficacy.

The transmission rate, a key parameter in epidemic modelling, captures the complex interaction of susceptibles and infectives and is also useful in estimating other important key parameters such as the threshold value, the basic reproductive rate and the vaccine efficacy. Therefore, models of the transmission rate that are more realistic in their assumptions and are useful for estimation are worthy of investigation.
2 A Multiplicative Model of the Transmission Rate and Derivation of Incidence Processes

In this chapter, the pathogenesis of an infection will be used as a basis to form a multiplicative model of the transmission rate. The pathogenesis of an infection and the complex interplay of pathogens and hosts will be reviewed, and the proprieties of multiplicative models in statistics will also be discussed. The proposed multiplicative model will be shown to model the reality more closely than the models that are frequently used in the literature. More importantly, this model will help the derivation of the incidence process, which is useful for statistical inferences.

An infection can only occur if there are contacts between a susceptible and one or more infectives. The population contact process is treated as a point process and whether the infection transmission occurred at each contact time is treated as a mark, and together constitutes a marked point process. An infection process is a compound point process determined by the marked point process. In SIR models, for different assumptions regarding the infectiosity of the pathogen, different marked point processes will be used, and thus, will result in different incidence processes. The incidence process for heterogeneous susceptible population, for a time-homogeneous contact rate, and for SEIR models will also be formulated.

2.1 Transmission Rate

When a microorganism overcomes the host defenses and multiplies inside the host body, the host is infected. When this results in tissue damage, or causes an alteration in tissue function, the host is diseased. However, in epidemic modelling these two terms are interchangeable because epidemic modelling is concerned with the number of individuals in the community who are capable of infecting others. The term infectives includes both infected and diseased individuals.

When susceptibles are exposed to the same amount of pathogen with a given infectiosity, different outcomes among individuals are determined by their defence mechanisms, in other words, by the susceptibility of the host. Efficient defence mechanisms result in a rapid destruction of the invading microbes, so that the pathogen never multiplies sufficiently to reach a threshold level and cause infection in the host. This kind of infection is known as asymptomatic
Figure 2-1 Pathogen multiplication results in different outcomes (Mims et. al., 1993)
infection. This situation is illustrated by the upper part of Figure 2-1 (Mims et. al., 1993), while the lower part of this figure shows a situation where the pathogen multiplied sufficiently to cause disease in the individual. Individuals experiencing asymptomatic infection will not be captured in most infectious disease data, and they will not be considered any different from susceptibles.

Upon entering the host body, the pathogen multiplies until the host defence systems start to react effectively. As shown in Figure 2-1, the amount of pathogen that may be released by an infective changes with time through the infectious period. Therefore, infective individuals may have different infectiousness during the infectious period. During a contact, the amount of pathogen actually released by an infective and received by a susceptible depends mostly on the form of the contact.

For a person-to-person transmitted disease, infection may take place by one or more contacts with infectives at a time depending on the infectiosity of the pathogen and the amount of pathogens being released. If the infectiosity is high, one contact is adequate to produce a new infection; otherwise, more contacts may be required to transmit the disease. Nonetheless, for a pathogen with specific infectiosity, the following four conditions must be satisfied for an infection to occur:

(1) There are individuals in the population infected by the pathogen.
(2) A susceptible makes one or more contacts with infectives.
(3) During the contacts, a certain amount of the pathogen is emitted by infectives.
(4) The pathogen enters the host body and reaches a certain threshold because the susceptible host defence mechanisms are inefficient in responding to the invasion.

An infection is the result of (2)–(4) occurring consecutively. Therefore, for a given pathogen, the probability of an infection can be expressed as the product of a sequence of probabilities:

\[
P\{\text{an infection}\} = P\{\text{a certain amount of pathogen begins to multiply in the susceptible host body}\} \times P\{\text{a certain amount of pathogen is emitted by the infectives in one or more contacts}\} \times P\{\text{a certain amount of pathogen is emitted by the infectives}\} \times P\{\text{one or more contacts between a susceptible and infectives}\} \times P\{\text{one or more contacts between a susceptible and infectives}\}
\]
This is a multiplicative model of the transmission probability. For a given infectiosity, the first conditional probability specifies the host susceptibility. The second conditional probability reveals the infectiousness of the infectives who emitted the pathogen. The third probability reflects the contact pattern of individuals in the population.

The amount of pathogen released by the infective and the amount that enters the susceptible host system differ mainly due to the first line of host defence (a component of the host susceptibility). This difference may also be influenced by the effect of the environment on the survival of the pathogen. However, the environment effect is minimal for a close person-to-person transmitted disease, and is omitted in this model.

Consequently, the infection probability is decomposed into the measure of the population contact pattern, susceptibility, the infectiousness of the infective host, and infectiosity of the pathogen. Not only is this multiplicative model of infection transmission based on the pathogenesis of an infection, but it is also sensitive to determinants of an epidemic. Host factors are incorporated in the susceptibility term, and agent factors are partially reflected by the infectiousness term. However, some data, for example, the individual who is responsible for the transmission of the disease and the exact time of infection, are not available in general. So this model cannot be applied to specific diseases without further assumptions. The random mixing of a population and a constant latent period are normally used to simplify the reality in order to obtain some results. Issues regarding epidemic modelling will be further discussed in Section 2.3.

### 2.2 Pathogenesis of an Infection

On a macroscopically level, epidemics are determined by hosts, agents and the environment (Kelsey et al., 1996, and Lilienfeld and Stolley, 1994). The outcomes of the interactions between susceptible hosts and pathogens are the results of battles between the hosts defence mechanisms and the microbes. This interaction needs to be considered microscopically. A review of the pathogenesis of an infection will help in understanding the epidemiology of infectious diseases, and it will also help in justifying the necessity of the multiplicative model given in the earlier section.

Many terms have been used to categorize microorganisms with respect to their interactions with their hosts. The term pathogen refers to a microorganism that is able to cause
diseases, while nonpathogen refers to a microorganism that does not cause disease. Microbes may be pathogenic or nonpathogenic depending on the interaction between the host and the pathogen. Infectious diseases are caused by infectious pathogens. To be a successful infectious disease pathogen, a microorganism must be able to survive the passage from one host to another, attach and enter the host tissues, resist the host defenses, replicate rapidly, and cause tissue damage or tissue malfunction. The disease-producing power or potency of a pathogen is called its virulence, while the infection-producing power or potency of a pathogen is called its infectiosity (Tellier, 2000). The multiplication of the microorganism in a host body is determined by the microorganism's survival mechanisms and the host defence system at the time of invasion (Jesen, 1984, Darnell et. al., 1990, and Mims et. al., 1993).

Hosts are endowed with a variety of mechanisms that help protect them from microbes that they may encounter during their lives. External tissues that prevent microbes from getting into deeper tissues form the first line of defence known as the non-specific external defence. The second line or non-specific inner defence refers to the innate part of the body that reacts immediately against invading microbes, for example, phagocytosis and inflammation. The third line, the acquired immunity is a dual component system. One component is the humoral antibody response, the formation of antibody molecule specific to the invading microbe. The other component is the cell-mediated immune response, which involves T cell-type lymphocytes. For some antigens, the acquired immunity may remain in the body for years.

Hosts have developed highly efficient recognition systems to detect foreign invaders as well as effective inflammation and immune responses that prevent the growth and eventually eliminate the invading pathogen from their bodies. On the other hand, pathogens have evolved and developed a variety of characteristics, which enable them to bypass or overcome a host defenses. Consequently, every infection is a battle between the capacity of the microbe to replicate, spread and cause disease, and the ability of the host to control and terminate the invading pathogen. Such battles involve complex interplay of both parts; one can not simply conclude that microbe + host = disease (Mims et. al., 1993). However, diseases can be viewed as having the following relationship, (Jensen and Wright, 1984)

\[
\text{Infectious Disease} \propto N V / \text{HF} \quad (2-1)
\]

where V is the virulence of the microbes, N is the number of microorganisms, and HF are host factors or the level of effectiveness of the host defence mechanisms. The chances of being
infected by a pathogen increase with the amount of pathogen emitted by the infective host that successfully enters the susceptible host. The chances also increase as the pathogenicity of the microbe increases, but decrease as the level of effectiveness of host defence mechanisms increase. A similar relationship holds for infection, which describes infection in terms of infectiosity, amount of pathogen and host defences. Such a relationship is known as the pathogenesis of an infection. This relationship provides the basis for the multiplicative model of the transmission probability given in Section 2.1.

### 2.3 Issues in Modelling and Data Availability

Population contact patterns are extremely complicated and are almost impossible to model without making relatively strong assumptions. The most frequently made assumption in epidemic modelling is the random mixing assumption, which states that every individual has an independent and equal chance of contacting any other individual in the population. This assumption may over-simplify the reality. For example, the chance of contacting someone depends highly on an individual’s daily activities, which may vary from day to day and from individual to individual. However, with a large or stratified population, random mixing is generally close to the reality.

The amount of pathogen ready to be released by an infective depends on the amount of pathogen in his system that in turn depends on the duration of being in the infective state (infectious period). The amount of pathogen actually released during a contact depends on the type of contacts, for example, sneeze releases more pathogen than talking does. Consequently, infectiousness is a function of the infectious period and the form of contacts during this period. However, forms of contacts are not always observable, and thus, are generally not available in incidence data and are ignored in the first stage of modelling. The infective that is responsible for the transmission of the disease is not generally observable. To account for the variability introduced by infectives, the average of the infectious period or other individual characteristics of infectives may be incorporated in modelling. The infectiosity of many pathogens and how they may vary over time are not well understood. For historical data, such information may not even be available.

The host defences are directly explained in terms of non-specific defence mechanisms (inflammation and phagocytosis), and specific defence mechanisms (antibodies, B cells and T
Data on measurements of humoral immune responses are more difficult to obtain than data on age, gender and household environmental factors. So the multiplicative model will not incorporate the direct measurements of host defence mechanisms.

A general problem with epidemic data is the accuracy of data. There is a latent period for most infectious diseases, which is the time required for the pathogen to multiply and to reach a threshold from its initial entrance into the host body. Very often a disease is identified in an individual when certain symptoms have appeared, but by then, the infectious period has already started or even ended for many diseases. This period is difficult to measure and is not always available in the incidence data. Asymptomatic infections are common for many infectious diseases, but are not generally available in data, as in incidence data infection are often identified based on the appearance of symptoms. It may be questionable if the observed number of infectives represents the true number of infectives in a population.

Most of these issues exist in epidemic modelling in general, and therefore, data collection becomes an important aspect of mathematical epidemic modelling. Data on each individual’s disease progression need to be carefully collected, that is, when a susceptible is infected, when he becomes infective and further becomes immune. To use the transmission rate $\beta(t)$ or the overall transmission rate $\beta$ to study the spread of a disease without using the multiplicative model leads to inability to explain the details of an infection. Breaking this parameter down into more parameters according to the pathogenesis of an infection is more precise, and allows the variability introduced by each component to be studied separately. Consequently, the data collection becomes more difficult; for example, data on the pathogens also need to be collected.

### 2.4 Multiplicative Models in Statistics

Regression models play an important role in many applied settings for making predictions, testing hypotheses, and identifying interactions among variables. Although attractively simple, the traditional linear regression often fails in practical situations because the effects of many factors on the outcome are usually non-linear. For example, multiplicative models are used to study Poisson count data.

Multiplicative models are frequently used in cancer research in the study of rate or risk $\lambda$, and are shown to be superior to additive models, (Breslow and Day, 1987). The purpose of
statistical methods in cancer research is to reveal the association between exposures and the
disease. An additive model or the excess risk due to a particular exposure for the two
dimensional sets of rates is given as

\[ \lambda_{it} = \lambda_{it} + b \quad (2-2) \]

where \( i \) indicates the \( i \)-th stratum, 1 for exposed and 0 for unexposed, and \( b \) represents the
additive effect of the particular exposure. A multiplicative model of rate or the relative risk of the
exposure is given as

\[ \lambda_{it} = r \lambda_{it} \quad (2-3) \]

where \( r \) is the relative risk due to the exposure. The concept of independence leading to the
additive model is rather simplistic and breaks down when plausible mechanisms for the disease
process are considered (Koopman, 1977). The multiplicative model is superior to the additive
model because of its empirical behavior and several logical properties that it possesses (Breslow

Section 2.1 suggests a multiplicative model for studying the effect of the population
contact pattern, infectiousness and susceptibility on the spread of a disease. Contacts between
infectives and a susceptible lead to transmissions of pathogens, and further, the multiplication of
the pathogen within the host system leads to an infection in the susceptible. Susceptible,
infectives and pathogens do not act independently; the joint effect of each part naturally falls into
a product of a sequence of conditional probabilities, and thus, leads to a multiplicative model.

### 2.5 Contact Initiation Process

A stochastic process \{\( N(t) : t \geq 0 \)\} with an integral-valued state space that is associated with
counting (one by one) of some events, is called a counting process or a point process, provided
that it is right continuous and has jumps of size one only. The incidence process of the counting
process, as defined in this dissertation, is given as \( dN(t) = E(dN(t) \mid F_t) \), where
\( dN(t) = N(t) - N(t^-) \) and \( F_t \) is a filtration (the \( \sigma \)-algebra generated by \( N(t) \) up to time \( t^- \)). In
the deterministic setting, it is referred to as an incidence function.
In *SIR* models, numbers of susceptibles, infectives and recovered individuals at time \( t \), denoted respectively by \( S(t) \), \( I(t) \) and \( R(t) \), form a three-dimensional, continuous-time and discrete-state Markov process. Assume the following:

1. Susceptible population is homogeneous.
2. Infective population is homogeneous.
3. Population is homogeneously mixed and the contact rate is time inhomogeneous.
4. Per “contact” transmission probability is time-homogeneous.

In addition to assumptions (1)-(4), the following is also assumed,

5. The population is closed without births or deaths with fixed size \( n + a \), where \( n \) is the number of susceptibles and \( a \) is the number of infectives at the beginning of the epidemic.

Assumption (5) implies that \( S(t) + I(t) + R(t) = n + a \), for all \( t \). Under assumption (5), the three-dimensional process is reduced to a two-dimensional process \( \{ S(t), I(t) \} \) (or \( \{ I(t), R(t) \} \)).

Infection caused by contacts between infectives and susceptibles will decrease the number of susceptibles and add new infectives to existing infectives. A multi-dimensional Markov process models this dynamics. Rhodes et. al. (1996) introduced the contact process as a point process and the infection process as a marked point process. This idea will be extended here to derive incidence processes.

A contact involves two individuals, a *contacter* who *initiated* the contact, and a *contactee* who *received* it. Any individual \( i \) in the population may initiate contacts with other individuals over time. Suppose individual \( i \) initiated \( c_{ij}(t) \) number of contacts with \( j \) up to time \( t \) and each contact occurred at times \( t_{u1}, \ldots, t_{uk}, \ldots, t_{u,v(t)} \). The collection of points \( \{ t_{u,k} \}_{k=1}^{v(t)} \) is a *point process*. The *contact initiation process* of \( i \) to \( j \), which counts the number of contacts that \( i \) initiated with individual \( j \) up to time \( t \), is a *counting process* and is denoted by \( C_{ij}(t) \).

If \( i \) initiated a contact with \( j \) at \( t \), then \( dC_{ij}(t) = 1 \); otherwise, \( dC_{ij}(t) = 0 \). Under assumption (3), the incidence of the contact initiation process is time inhomogeneous, as

\[
\mathbb{E}\left[dC_{ij}(t) \mid F_r \right] = \alpha(t)dt \quad (2-4)
\]

\( \forall i, j \) and \( i \neq j \), where \( F_r = \sigma\{ C_{ij}(u), u < t \} \) is the filtration.

Assume that at each instant an individual can *initiate* at most one contact, then

\( P\{ [dC_{i1}(t) = 1] \cap [dC_{i2}(t) = 1] \mid F_r \} = o(dt) \), \( \forall i, j_1, j_2, i \neq j_1 \neq j_2 \). Note that an individual may
act as a contacter and a contactee at the same time. However, only if the contacter is infective, may this contact lead to an infection in the susceptible contactee, that is, a contact initiated by a susceptible with an infective would not result in any transmissions of pathogens.

It is possible that an individual receives multiple contacts simultaneously from different individuals at one instant. The contact accepted process is a compound point process because the number of contacts at each contact time point is a discrete random variable that takes integer values \{1, 2, \ldots, n + a - 1\}. Let \( \{t_{i, k}\}_{k=1}^{c_{i,j}(t)} \) be the contact points initiated by \( i_1 \) with \( j \) up to time \( t \), and \( \{t_{i_2,k}\}_{k=1}^{c_{i_2,j}(t)} \) be the contact points initiated by \( i_2 \) with \( j \) up to \( t \). Suppose \( i_1 \) and \( i_2 \) simultaneously initiated a total of \( c_{(i_1, i_2), j}(t) \) contacts with \( j \) up to time \( t \) at times \( t_{i_1,1}, \ldots, t_{i_1,k} \), \( t_{i_2,1}, \ldots, t_{i_2,k} \), \( t_{i_1,1}, \ldots, t_{i_1,k} \), \( t_{i_2,1}, \ldots, t_{i_2,k} \). This sequence of time points that is the intersection of the two sets \( \{t_{i_1,k}\}_{k=1}^{c_{i_1,j}(t)} \) and \( \{t_{i_2,k}\}_{k=1}^{c_{i_2,j}(t)} \) is also a point process. Consequently, the contact initiated process of \( (i_1, i_2) \) simultaneously with \( j \), which counts the number of contacts initiated by \( (i_1, i_2) \) simultaneously with \( j \) up to time \( t \), is a counting process denoted by \( C_{(i_1, i_2), j}(t) \). The intensity process of \( C_{(i_1, i_2), j}(t) \) is given as

\[
E[dC_{(i_1, i_2), j}(t) | F_{t-}] = \alpha(t)dt \quad (2-5)
\]

The contact initiated process of \( m \) individuals simultaneously with \( j \) may be defined similarly as \( C_{(i_1, \ldots, i_m), j}(t) \) with the intensity process given as

\[
E[dC_{(i_1, \ldots, i_m), j}(t) | F_{t-}] = \alpha(t)dt \quad (2-6)
\]

### 2.6 Incidence Process under Basic Assumptions

In this section, the incidence process, which gives the expected number of new infections in a short interval, will be derived using a probability approach based on the multiplicative model discussed in Section 2.1. The incidence is derived similarly for all models (SIR and SEIR). For the sake of demonstration the SIR model is considered. It is important to clarify assumptions used to derive the incidence process and to show how incidence may vary as the assumptions are modified.
Infectiosity, which represents the agent factor, is an important aspect for the incidence process formulation. Another assumption, in addition to those assumptions specified in Section 2.5, says that

(6) One infectious contact is required for an infection transmission to occur with a certain probability.

Assumption (6) implies that the type of pathogen causing the epidemic has the infectiosity such that the amount of pathogen released in one contact initiated by an infective is adequate to produce an infection in a susceptible. If two infectives contacted a susceptible simultaneously, then a double dosage of the pathogen was released to this susceptible. With assumption (6), one of the two infectious contacts will not result in any infection transmission because the threshold level of infectious pathogen is reached with just one infectious contact. Assumption (2) implies that every contact initiated by an infective is identical. Therefore, it is not necessary to make the distinction as which contacter is responsible for the infection transmission. Note that this information is generally not available in incidence data.

Let $X_{ik}$, where $k=1,2,\ldots,c_y(t)$, indicate whether the infection transmission occurred ($j$ has become infected) during the contact initiated by $i$ at $t_{ik}$, then $X_{ik}$ are marks for the contact initiation process $\{C_y(t)\}$. If two infectives initiated contacts simultaneously at $t_{i/k} = t_{j/k}$ for some $k$, only one of the infectious contacts has the mark equal to 1. So $\{C_y(t), X_{ik}\}$ forms a marked point process (Rhodes et. al., 1996)

The infection transmission may take place with a certain probability at the moment when an infective initiated a contact with a susceptible. Let

$$P\{\text{a certain amount of pathogen starts to multiply within the susceptible host body | a certain amount of pathogen is emitted by an infective in one contact}\} = \phi,$$

$$P\{\text{a certain amount of pathogen is emitted by an infective | one infectious contact }\} = \xi.$$

Both parameters are time-homogeneous under assumption (4). Using the multiplicative model given in Section 2.1, this probability, the per "contact" transmission probability, is denoted by $\eta$ and $\eta = \phi \xi$. Under assumption (4), marks $X_{ik}$ for all $k$ are independent and identical Bernoulli ($\eta$) random variables. If the contacter $i$ is infective and the contactee $j$ is susceptible, its conditional expectation is given as
where \( I_i(t^-) \) is the indicator function of individual \( i \) being infective just before time \( t \), and \( S_j(t^-) \) is the indicator function of individual \( j \) being susceptible just before time \( t \). Note that the filtration in this case will consist of the information on the contacter’s and contactee’s infectivity status (being infective or non-infective) and their contact history. \( F_{i^-} = \sigma\{I_i(u), S_j(u), 0 \leq u < t\} \).

A mark is only defined at a time point when there is an initiated contact. Obviously, marks are equal to 0 if the contacts are not initiated by infectives and received by susceptibles.

Let \( N_q(t), i \neq j, \) be the infection process of \( i \) to \( j \) that counts the number of times that \( j \) is infected by \( i \) up to time \( t \), then the infection process is a compound point process determined by the marked process \( \{C_q(t), X_{\cdot \cdot q}\} \) and \( N_q(t) = \sum_{k=1}^{C_q(t)} X_{\cdot \cdot q} \). Assume a susceptible can be infected only once during an epidemic (that is recovery with long term immunity after infection), then the state space for this compound point process is \( \{0, 1\} \). The infection process is 1 if \( j \) is infected by \( i \), and is 0 if \( j \) is not infected by \( i \) up to time \( t \). This process is right continuous, and the instantaneous change at time \( t \) is given as

\[
dN_q(t) = N_q(t) - N_q(t^-) = \sum_{k=1}^{C_q(t)} X_{\cdot \cdot q} - \sum_{k=1}^{C_q(t^-)} X_{\cdot \cdot q} = X_{\cdot \cdot q}(t^-) dC_q(t) \quad (2-8)
\]

The intensity process of the compound point process is generally evaluated by conditioning on the contact initiation process, \( E\{dN_q(t) \mid F_{i^-}\} = E\{E\{X_{\cdot \cdot q} \mid F_{i^-}\} \mid dC_q(t) \mid F_{i^-}\} \), where the filtration is defined as \( F_{i^-} = \sigma\{I_i(u), S_j(u), C_y(u), 0 \leq u < t\} \). The inner expectation is given by (2-7), and it can be taken out of the outer expectation because \( \phi\phi\phi I_i(t^-) S_j(t^-) \) is a predictable process with respect to \( F_{i^-} \). So it follows from (2-4) and (2-7) that the intensity process of the infection process defined between \( i \) and \( j \) is given as

\[
E\{dN_q(t) \mid F_{i^-}\} = \phi\phi\phi \alpha(t) I_i(t^-) S_j(t^-) dt \quad (2-9)
\]

If the contact initiation process and the infection process are completely observable, that is, for each infection the individual who is responsible for this infection is fixed, then the infection process for susceptible \( j \) \( dN_{ij}(t) \), where \( i^- \) indicates the responsible infective, has the
intensity process $E\{dN_{i,j}(t) \mid F_{j^-}\} = \alpha(t)\phi\xi(t^-)dt$. If $N(t)$ is the total number of infections up to time $t$ for the population, then $dN(t) = \sum_{j=1}^{a\alpha} dN_{i,j}(t)$. Under assumption (1), the incidence process of $N(t)$ is given as

$$E\{dN(t) \mid F_{j^-}\} = E\{ \sum_{j=1}^{a\alpha} dN_{i,j}(t) \mid F_{j^-}\} = \beta(t)S(t^-)dt \quad (2-10)$$

where $\beta(t) = \alpha(t)\phi\xi = \alpha(t)\eta$ is the transmission rate, which is time inhomogeneous, and $S(t^-) = \sum_{j=1}^{a\alpha} S_j(t^-)$, the number of susceptibles before time $t$. Because $t^-$ is not random for each infection, the incidence process does not involve the number of infectives $I(t^-)$, that is, the incidence process is linear with respect to the number of susceptibles. This is the Greenwood model.

If the contact initiation process and the infection process are not completely observable, which is generally the case in practice, then any infectives may initiate a contact with susceptible $j$ and further cause infection in this individual. Define $N_j(t)$ as the infection process of susceptible $j$. Because any infective can infect this susceptible individual, $dN_j(t) = \sum_{i=1}^{a\alpha} dN_{i,j}(t)$. And define $N(t)$ as the infection process of the population. Because any susceptible may become infective, $dN(t) = \sum_{j=1}^{a\alpha} dN_j(t)$. Thus, under assumptions (1) and (2), the incidence process of $N(t)$ is given as

$$E\{dN(t) \mid F_{j^-}\} = \beta(t)I(t^-)S(t^-)dt \quad (2-11)$$

Consequently, with assumptions (1) to (6), the incidence process is bilinear, which is commonly used and discussed in the literature (Wilson and Worcester, 1945a, b). This is the Reed-Frost model.

For incidence processes as given in (2-10) and (2-11), the martingale method can be used to make inferences on the transmission rate. If the contact pattern is also observable, the likelihood method can be used to estimate each parameter ($\alpha(t), \phi$ and $\xi$).
2.7 Low and High Infectiosities

If a pathogen has a low infectiosity, then the amount of pathogen emitted by an infective during one contact may not be sufficient to infect a susceptible. If a double dosage of such pathogen at the same epoch is required to produce an infection, then the assumption (6) needs to be modified to:

(6a) Two infectious contacts are required to produce an infection.

Under assumption 6(a), a susceptible will be infected if this individual received at least two infectious contacts at \( t \), i.e., \([dC_{i,j}(t) = 1]\) and \([dC_{i,j}(t) = 1]\), for a pair of individuals \((i_1, i_2)\) who are infective, and \( i_1 \neq i_2 \).

Let \( X_{(i_1, i_2), k} \), where \( k = 1, 2, \ldots, c_{(i_1, i_2)} \), indicate whether the infection transmission occurred during the two simultaneous contacts initiated by \((i_1, i_2)\) with \( j \) at \( t_{(i_1, i_2), k} \); then \( X_{(i_1, i_2), k} \) are marks for the contact initiation process \( \{C_{(i_1, i_2)}(t)\} \). If three infectious contacts occurred at time \( t \), by assuming (6a), one of the three infectious contacts is ignored. Thus, \( \{C_{(i_1, i_2)}(t), X_{(i_1, i_2), k}\} \) forms a marked point process.

The infection transmission may take place with a certain probability at the moment when at least two of those simultaneous contacts are initiated by infectives. This probability is 0 if at most one contact, among those simultaneous contacts, is initiated by an infective. Let

\[
P\{\text{a certain amount of pathogen is emitted by infectives | two infectious contacts}\} = \xi.
\]

Using the multiplicative model given in Section 2.1, the infection transmission probability is given as \( \phi \xi = \eta \). Under assumption (4), marks \( X_{(i_1, i_2), k} \) for all \( k \) are independent and identical Bernoulli (\( \eta \)) random variables. If \((i_1, i_2)\) are infectives and \( j \) is a susceptible, its conditional expectation is given as

\[
E[X_{(i_1, i_2), k} | F_j^-] = \eta I_{i_1}(t^-) I_{i_2}(t^-) S_j(t^-) \quad (2-12)
\]

The infection process \( \{N_{(i_1, i_2)}(t)\} \) is a compound point process determined from this marked point process \( \{C_{(i_1, i_2)}(t), X_{(i_1, i_2), k}\} \). \( N_{(i_1, i_2)}(t) = \sum_{k=1}^{c_{(i_1, i_2)}} X_{(i_1, i_2), k} \). Note that there is no distinction between contacts initiated by \((i_1, i_2)\) and those initiated by \((i_2, i_1)\), so \( \{N_{(i_1, i_2)}(t)\} \) and
\{N_{(i,j)}(t)\} are considered as one process. Similarly, the intensity process is evaluated by conditioning on the contact initiation process. Because $\phi(t, t^-)I_{i_j}(t^-)S_{i_j}(t^-)$ is predictable with respect to $F_r$, by using (2-5) and (2-12) the incidence process of \{N_{(i,j)}(t)\} is given as

$$E\{dN_{(i,j)}(t) | F_r\} = \eta \alpha(t) I_{i_j}(t^-) I_{i_j}(t^-) S_{i_j}(t^-)dt$$ \hspace{1cm} (2-13)

Let $N_j(t)$ be the infection process of susceptible $j$ and $N(t)$ be the infection process for the population at time $t$. If the infectives responsible for each infection are fixed, then, under assumption (1), the incidence process of $N(t)$ is given as

$$E\{dN(t) | F_r\} = \eta \alpha(t)S(t^-)dt$$ \hspace{1cm} (2-14)

Because the two infectives who were responsible for the infection are not random, the incidence process does not involve the number of infectives $I(t^-)$.

If the contact initiation process and the infection process are not completely observable, then any two infectives may make contacts with susceptible $j$ and further cause infection in this individual, that is, $dN_j(t) = \sum_{(i_j, \ell \in I(t^-), \ell \in I(t^-), \ell \notin \ell_j)} dN_{(i_j, \ell \in I(t^-), \ell \notin \ell_j)}(t)$, which sums over all possible combinations of pairs of infectives. Furthermore, $dN(t) = \sum_{j=1}^n dN_j(t)$, then the incidence process of $N(t)$ is given as

$$E\{dN(t) | F_r\} = \frac{\eta \alpha(t)}{2} (I(t^-) - I(t^-)I(t^-)S(t^-))dt$$ \hspace{1cm} (2-15)

Note that the 2 in the denominator arises because \{N_{(i_j, \ell \in I(t^-), \ell \notin \ell_j)}(t)\} and \{N_{(i_{j_1}, \ell)}(t)\} are considered as one process. If $I(t^-)$ is relatively large, which is generally true in most epidemics except at the beginning and at the end of an epidemic, then the incidence process may be approximated as

$$E\{dN(t) | F_r\} = \frac{\eta \alpha(t)}{2} I^2(t^-)S(t^-)dt$$ \hspace{1cm} (2-16)

Therefore, under assumptions (1)-(5) and (6a), the incidence process is approximately linear in the number of susceptibles and quadratic in the number of infectives.
In general, the number of contacts required may take any positive values \( m \geq 0 \). Note that \( m \) is introduced to reflect the infectiosity of the pathogen. If the infectiosity is high, then fewer contacts are required to produce an infection; but if the infectiosity is low, then more contacts are required. Then (6a) is generalized to

(6b) \( m \) infectious contacts are required to produce an infection.

Therefore, the incidence process is given as

\[
E\{dN(t) \mid F_r\} = \eta \alpha(t) \left( \frac{I(t^-)}{m} \right) S(t^-) dt \quad (2-17)
\]

and if \( I(t^-) \) is large relative to \( m \), the above may be approximated as

\[
E\{dN(t) \mid F_r\} = \frac{\eta \alpha(t)}{m!} I^n(t^-) S(t^-) dt = \beta(t) I^n(t^-) S(t^-) dt \quad (2-18)
\]

where \( \beta(t) = \phi \xi \alpha(t) / m! \) is the transmission rate. Therefore, under assumptions (1)-(5) and (6b), the incidence process is approximately linear in the number of susceptibles and a power function in the number of infectives.

The martingale method in conjunction with a smoothing method can be used to estimate the transmission rate for incidence processes given as in (2-17) and (2-18). However, as (2-18) is easy to work with, it will be used for statistical inferences. The estimating method will be discussed in Chapter 3.

Naively, when \( m \) is a positive integer, it can be interpreted as the number of contact required for an infection transmission. If the population is sufficiently large, then the sizes of each class (susceptible, infective and recovered) can be considered as continuous variables (see Hethcote 1989). Similarly, the number of contacts becomes a continuous variable, which is an extension of the idea of contacts. Thus, \( m \) may be a fraction or an integer which is greater than zero. If \( m \) is zero, the mode of transmission is via a common vehicle. If \( m \) is greater than zero, the mode of transmission is by a direct person-to-person contact. Furthermore, if \( m \) is a number less than 1, the pathogen is of high infectiosity; if \( m \) is a number greater than 1, the pathogen is of low infectiosity.
2.8 Population Heterogeneity

In reality, a population is not homogenous. Susceptibles may have different degrees of immune responses, while infectives may have different levels of infectiousness. Consequently, assumptions (1) and (2) are modified as:

(1a) Susceptible population is heterogeneous.

(2a) Infective population is heterogeneous.

Also, susceptibility and infectiousness may vary over time. Therefore, parameters for the susceptibility and the infectivity need to be defined for each individual. Let

\[ \phi_j(t) = P\{ \text{a certain amount of pathogen starts to multiply within } j's \text{ body at } t \mid \text{a certain amount of pathogen is released by an infective in a contact at } t \} , \]

\[ \xi_i(t) = P\{ \text{a certain amount of pathogen is released by } i \text{ at } t \mid i \text{ contacts a susceptible at } t \} . \]

In addition, some infectives may be more active than other infective individuals. So these individuals may contact and further infect more susceptibles in the population. Similarly, some susceptibles may have higher risk behaviours. Therefore, they may have a higher probability of contacting some infective individuals. Consequently, assumption (3) is modified as:

(3a) The population is not randomly mixed and the contact rate is time inhomogeneous.

The probability of \( i \) contacting \( j \) at the time \( t \) is given by \( P\{ C_q(t) = 1 \mid F_x \} = \alpha_q(t)dt + o(dt) \).

Here the filtration may contain information on each individual’s risk behaviour that may affect \( \alpha_q \). Let \( Z_j(t) \) denote the covariate process for contactee \( j \), \( V_i(t) \) denote the covariate process for contacter \( i \), then the filtration is \( F_r = \sigma\{ I_i(u), S_j(u), C_q(u), Z_j(u), V_i(u), 0 \leq u < t \} \). The incidence process of \( N(t) \) under assumptions (1a), (2a), (3a) and (4)-(6) is given as

\[ E\{ dN(t) \mid F_r \} = \sum_{j=1}^{n(t)} \sum_{l=1}^{n(t)} \phi_j(t)\xi_l(t)\alpha_q(t)dt \quad (2-19) \]

where \( \phi_j(t) \) is a function of the covariates \( Z_j(t) \), and \( \xi_l(t) \) is a function of \( V_l(t) \), and \( \alpha_q(t) \) is a function of both \( Z_j(t) \) and \( V_l(t) \). The incidence process of \( N(t) \) under assumptions (1a), (2a), (3a), (4), (5) and (6b) is given as
Incidence processes as given in (2-19) and (2-20) are not useful for inferences, because they are computationally unmanageable. However, if there are only a small number of factors that make susceptibles heterogeneous, the intensity process can be used to make inferences on the effect of these factors. Susceptibles are classified into \( b \) groups, \( S^1(t), \ldots, S^b(t) \), and within each group, susceptibility is assumed to be homogeneous. If \( N^k(t) \) denotes the infection process of the \( k \)-th susceptible population stratum, then its incidence process is given as

\[
E\{dN(t) \mid F_r \} = \sum_{j=1}^{b} \sum_{i=1}^{\infty} \phi_j(t) \prod_{i=1}^{m} \xi_i(t) \alpha_{ij}(t) dt \quad (2-20)
\]

for \( k = 1, \ldots, b \), where \( \eta^k(t) = \phi^k(t) \xi \) is the per "contact" transmission rate for stratum \( k \). The infection transmission rate is denoted by \( \beta^k(t) = \eta^k(t) \alpha(t) \). The intensity process (2-21) can be studied by stratified martingale method (non-parametric method), which will be discussed in Chapter 3. The relative rate will be used to evaluate covariate effects, and the transmission rate within each stratum will be estimated using martingale method. Furthermore, the Cox model (semi-parametric method) will be used to estimate the effect of these factors in Chapter 5, and a semi-parametric estimate of the transmission rate that accounts for other covariates will also be obtained.

\section*{2.9 Time Homogeneity}

In the literature, the transmission rate is often assumed to be time-homogeneous. Thus, the contact initiation process is simplified as a time-homogeneous point process. As susceptibility and infectiousness are time-homogeneous, assumption (3) is modified as

(3b) Population is homogeneously mixed and the contact rate is time-homogeneous.

Then the infection process has intensity \( E\{dN^k(t) \mid F_r \} = \phi^k \xi \alpha(t) S^k(t^-) dt \), and the incidence process of \( N(t) \) under assumptions (1), (2), (3b), (4)-(6) is given as

\[
E\{dN(t) \mid F_r \} = \phi \xi \alpha(t^-) S(t^-) dt = \beta(t^-) S(t^-) dt \quad (2-22)
\]
where $\beta = \phi \xi \alpha$, which is time-homogeneous, represents the overall transmission rate. This incidence process (2-22) is bilinear with time homogeneous transmission rate. The incidence process of $N(t)$ under assumption (1), (2), (3b), (4), (5) and (6b) is given as

$$E\{dN(t) \mid F_t\} = \phi \xi \alpha \begin{pmatrix} I(t^-) \\ m \end{pmatrix} S(t^-) dt \quad (2-23)$$

which may be approximated as

$$E\{dN(t) \mid F_t\} = \frac{\phi \xi \alpha}{m!} I^{-n} S(t^-) dt = \beta t^n S(t^-) dt \quad (2-24)$$

This incidence process is linear in the number of susceptibles and a power function in the number of infectives, which is more general than (2-22). The martingale method can be used to estimate the transmission rate for models with incidence processes as given in (2-22) and (2-24).

### 2.10 SIR and SEIR Models

Previous sections focus on deriving incidence processes using various assumptions. The incidence process of $N(t)$ that was derived under assumption (5), closed population without births or deaths, is also the intensity process for the process $-S(t)$ because $N(t) = n - S(t)$. To study an SIR model, the changes in $I(t)$ also needs to be considered. The infective process $I(t)$ can be expressed as $I(t) = n + a - S(t) - R(t)$ under assumption (5). Therefore, it is the difference of two counting processes $n + a - S(t)$ and $R(t)$. The intensity of this process is simply obtained by subtracting the infection incidence and the recovery incidence. Define the recovery rate $\gamma(t)$, a continuous function of $t$, as

$$\lim_{\Delta t \to 0} P\{\text{the individual is in state R at } t \mid \text{an individual is in I at time } t - \Delta t \}/\Delta t = \gamma(t)$$

Then the recovery intensity is $\gamma(t) I(t^-)$. The system of stochastic differential equations for an SIR model is given as

$$E\{dS(t) \mid F_t\} = -\phi(t) \xi(t) \alpha(t) S(t^-) I(t^-) dt \quad (2-25)$$

$$E\{dI(t) \mid F_t\} = \phi(t) \xi(t) \alpha(t) S(t^-) I(t^-) dt - \gamma(t) I(t^-) dt$$
This is more general than the system of stochastic differential equations used in the literature that is derived using (1-3).

When additional classes of individuals are considered, there will be more compartments in the system so that the dimension of the process is increased. An SEIR model forms a four-dimensional process \( \{S(t), E(t), I(t), R(t)\} \), where \( E(t) \) is the number of latent individuals. It reduces to a three-dimensional process \( \{S(t), E(t), I(t)\} \) for a closed population. Similarly, \( n - S(t) \) and \( R(t) \) are counting processes, but \( E(t) \) and \( I(t) \) are not counting processes.

The incidence process of infection depends on the infectiosity of the pathogen, the number of infectives, the number of susceptibles, the contact rate, infectiousness and susceptibility, but not on the number of individuals in other classes. Formulating the incidence process of infection for SEIR models in a closed population is similar to what has been discussed for SIR models, the derivation of which will not be discussed in detail. The stochastic differential equations are given as

\[
\begin{align*}
E\{dS(t) | F_{t^-}\} &= -\phi(t)\xi(t)\alpha(t)S(t^-)I^w(t^-)dt \\
E\{dE(t) | F_{t^-}\} &= \phi(t)\xi(t)\alpha(t)S(t^-)I^w(t^-)dt - \zeta(t)E(t^-)dt \\
E\{dI(t) | F_{t^-}\} &= \zeta(t)E(t^-)dt - \gamma(t)I(t^-)dt
\end{align*}
\]

(2-26)

where \( \zeta(t) \) is the transition rate from \( E \) to \( I \), defined as a continuous function of \( t \), by

\[
\lim_{\Delta t \to 0} P\{ \text{the individual is in state } I \text{ at } t | \text{the individual is in state } E \text{ at } t - \Delta t \}/\Delta t = \zeta(t).
\]

In this chapter, using the multiplicative model and properties of counting processes, various incidences are derived for different assumptions. In Chapter 3, these incidence processes will be used in conjunction with the martingale method and a smoothing method to estimate the transmission rate. Chapter 4 will study other functions that are transformations of the transmission rate. The observed heterogeneity will be studied using the stratified martingale method in Chapter 3 and using the Cox model in Chapter 5. A semi-parametric estimate of the transmission rate accounting for other covariates will also be introduced in Chapter 5.
3 Non-Parametric Inferences

The transmission rate is one of the key parameters in epidemic modelling. Estimating this parameter and its standard error is one of the main modelling interests. However, the process of spreading an infectious disease is seldom observed completely which often makes estimation difficult if not impossible. The martingale method can be profitably used to make inferences on the cumulative transmission rate using only the observable realization of the underlying process. Becker (1989) has used this method to estimate the cumulative transmission rate under the assumption of bilinear incidence for an SIR model. In this chapter, the martingale method will be applied to estimate the transmission rate under general assumptions for SEIR models and for a heterogeneous population. A review of counting processes, martingales, the Doob-Meyer decomposition theorem, and the martingale transform theorem will be given at the beginning of the chapter. The number of infections occurring during an epidemic can be treated as a counting process. The incidence processes derived in Chapter 2 are the hazard processes (under various assumptions) of this counting process. A martingale that involves both the observed counting process and the incidence process will be used to obtain an unbiased estimate of the cumulative transmission rate under general assumptions (with the infectiosity of a pathogen). Furthermore, variances of these estimators will also be estimated using the variation process. In this chapter, the cubic spline method will be used to obtain an estimation of the transmission rate as a continuous function of time. Also, the kernel smoothing method used by Becker (1989) for estimating the transmission rate will be briefly discussed. Two sets of data will be used to illustrate the inference procedure: a smallpox epidemic in a Nigerian village, and a respiratory disease epidemic on the island of Tristan da Cunha. Bailey (1975), Becker (1989) and other investigators have studied these two data sets. Host factors that create heterogeneity among susceptibles will be studied using a stratified analysis by applying the model given in Chapter 2. The respiratory disease data will be used to illustrate the stratified analysis so that certain conclusions on the effect of the host factors will be drawn for this epidemic.
3.1 Martingale and Counting Process Review

The martingale method plays an important role in non-parametric inferences, as demonstrated in the seminal paper by Aalen (1978). The martingale method has several advantages over conventional statistical methods, Hsieh (1997), and has been successfully used in survival analysis, Fleming and Harrington (1991), Hsieh (1992) and Andersen et. al. (1993). For details of martingale theory see Liptser and Shiryayev (1989). The martingale method can be similarly generalized to study epidemics. Relevant results in martingale theory will be reviewed in this section.

A martingale \( \{M(t): t \geq 0\} \) is a right continuous stochastic process with finite mean adapted to a filtration \( F_t \) if, and only if, for all \( t \geq s \),

\[
E[M(t) \mid F_s] = M(s) \quad (3-1)
\]

The Doob-Meyer decomposition theorem, (Meyer, 1966), provides a way of constructing a martingale from a counting process. This theorem states that a counting process \( N(t) \) is uniquely decomposed into its compensator \( \Lambda(t) \) and a martingale \( M(t) \),

\[
M(t) = N(t) - \Lambda(t) \quad (3-2)
\]

if, and only if, \( \Lambda(t) \) satisfies

\[
d\Lambda(t) = E[dN(t) \mid F_t] \quad (3-3)
\]

Recall that \( dN(t) = N(t) - N(t^-) \) and \( F_t \) is the filtration (on the \( \sigma \)- algebra generated by \( N(t) \) up to time \( t^- \)).

The variation process of a martingale is given by

\[
<M>(t) = \int_0^t \text{Var}[dM(u) \mid F_u] \quad (3-4)
\]

so that

\[
d<M>(t) = \text{Var}[dM(t) \mid F_t] \quad (3-5)
\]

Also, \( d<M>(t) \) can be rewritten as,
\[ d < M > (t) = E\{dM^2(t) \mid F_t \} - E\{dM(t) \mid F_t \}^2 = E\{dM^2(t) \mid F_t \} \quad (3-6) \]

where the second equality follows from (3-1).

The martingale central limit theorem for a sequence of martingales (Lipster and Shiryaev, 1989), states that for a sequence of zero mean martingales \( M_1, M_2, \ldots, M_n, \ldots \), if there exists a non-negative continuous non-decreasing deterministic function \( V(t) \) with \( V(0) = 0 \) such that:

(a) for each \( t \in (0, \tau) \), \( < M_n > (t) \) converges to \( V(t) \) in distribution;

(b) there exist constants \( c_n \to 0 \) as \( n \to \infty \), \( P\{\sup_t |dM_n(t)| \leq c_n\} \to 1 \),

then, for each \( t \in (0, \tau) \), \( M_n(t) \) converges to \( M(t) \) in distribution, where \( M(t) \) is normally distributed with mean zero and variance equal to \( < M_n > (t) = V(t) \).

The martingale transform theorem states that if \( G(t) \) is an \( F_t \)-predictable process and \( M(t) \) is a martingale, then

\[ M^*(t) = \int_0^t G(u)dM(u) \quad (3-7) \]

is also a martingale. This theorem provides a way of constructing a different martingale from a given martingale and an \( F_t \)-predictable process, which is more useful for estimation than the original martingale. The variation process of such martingale transform is

\[ \int_0^t d < M^* > (u) = \int_0^t G^2(u)d < M > (u) \quad (3-8) \]

This also follows from (3-1). The purpose of using a martingale is to estimate the parameter in the compensator and to construct test statistics using the martingale transform and the martingale central limit theorem by choosing appropriate \( F_t \)-predictable processes \( G(t) \).

The martingale method has been successfully applied in survival analysis. Suppose there are \( n \) independent and identical individuals. The counting process \( N(t) = \sum_{i=1}^n \mathbb{1}(T_i \leq t) \), where \( \mathbb{1}(A) \) is the indicator function of \( A \) and \( T_i \) is the time to death of individual \( i \), counts the number of deaths up to time \( t \). The compensator \( \Lambda(t) \) of this counting process is the integral of product
of the left continuous process $Y(t) = \sum_{i=1}^{\infty} 1(T_i \geq t)$ and a deterministic function $dH(t)$ called the hazard probability (see, for example, Fleming and Harrington 1991, Section 5.2). Thus, $d\Lambda(t) = Y(t)dH(t)$. By the Doob-Meyer decomposition theorem, a martingale is constructed as $M(t) = N(t) - \int_0^t Y(u)dH(u)$. Taking $G(t) = J(t)/Y(t)$, where $J(t) = I(Y(t) > 0)$, as the $\mathcal{F}_t$-predictable process, the transformed martingale is then given by

$$
\int_0^t J(u)/Y(u)dM(u) = \int_0^t J(u)/Y(u)dN(u) - \int_0^t J(u)dH(u).
$$

Because $t = \inf\{u > 0, Y(u) = 0\}$, $J(t) = 1$ for all $t \leq t$. Using the moment method, the martingale estimate of the cumulative hazard is $
\tilde{H}(t) = \int_0^t \frac{dN(u)}{Y(u)}.
$

Stochastic models are also used to study fly mating selection, Barndorff-Nielsen (1968), Bartlett et. al. (1971), Aalen (1975, 1978) and Andersen et. al. (1993). Male and female virgin flies were inserted into an observation arena called a 'pomoscope'. The flies were observed continuously, and the number of matings, time of initiation and time of termination were recorded. Mating involves the interactions of the male and female flies, and is similar to the mechanism of the spread of a disease as it involves contact between a susceptible and infectives (may be more than one). The mating process is simpler because it may only appear between a pair of flies (one male and one female). In the mating selection problem, $N(t)$ is defined as the number of matings initiated up to time $t$. Let $F(t)$ and $M(t)$ denote the number of female and male flies, respectively, not yet having initiated a mating just before $t$ and $\alpha(t)$ denote the instantaneous mating rate. The intensity process for $N(t)$ is $d\Lambda(t) = \alpha(t)F(t)M(t)$ (Aalen, 1978, Section 8). A martingale can be constructed using this intensity process, and similar inferences can be made as shown above. This type of generalization will be employed in the following sections of this chapter to formulate martingale inferences in epidemic modelling.

### 3.2 Martingale Inferences for the Transmission Rate

Becker (1989) has used the martingale method to estimate the cumulative transmission rate under the assumption of a bilinear incidence for an SIR model. This section will apply the
martingale method to the general incidence process as derived in Chapter 2. Indeed, the compensator of the counting process, which will be used in the construction of a martingale for estimating the cumulative transmission rate, is derived probabilistically by using the multiplicative model in Chapter 2.

The general incidence process is \( d\Lambda(t) = \beta(t)S(t^-)I^n(t^-)dt \), so the compensator, which is the integral of the incidence process, is given as

\[
\Lambda(t) = \int_0^t \beta(u)S(u^-)I^n(u^-)du \quad (3-9)
\]

By the Doob-Meyer theorem, a martingale is constructed as

\[
dM(t) = dN(t) - \beta(t)S(t^-)I^n(t^-)dt \quad (3-10)
\]

However, this martingale is not sufficient for use in estimation. A \( F_t \)-predictable process needs to be chosen to transform (3-10) to new martingales using the martingale transform theorem for simplifying the estimation.

### 3.2.1 SIR Models

The number of infectives or the number of susceptibles being zero indicates the end of an epidemic in an SIR model. So the end time point of an epidemic is denoted by \( \tau = \inf\{t:S(t)I(t) = 0\} \). However, it is seldom the case that all susceptibles are infected when an epidemic stops. So it is sufficient to define the end time point as \( \tau = \inf\{t: I(t) = 0\} \). If the \( F_t \)-predictable process is taken to be \( G(t) = J(t)/(S(t^-)I^n(t^-)) \), where \( J(t) = 1(I(t^-) > 0) \), then \( J(t) \) is always equal to 1 for \( t \leq \tau \). By the martingale transform theorem, multiplying (3-10) by \( G(t) \) yields a martingale transform

\[
M^*(t) = \int_0^t G(u)dM(u) = \int_0^t G(u)dN(u) - \int_0^t \beta(u)J(u)du \quad (3-11)
\]

Since the martingale given in (3-10) has zero mean, it follows from (3-2) that

\[
E\left[ \int_0^t G(u)dN(u) \right] - E\left[ \int_0^t \beta(u)J(u)du \right] = 0 \quad (3-12)
\]
Let \( B(t) = \int_0^t \beta(u) \, du \). As \( J(t) = 1 \), for all \( t \leq \tau \), by the moment method, it follows that

\[
\int_0^t G(x) \, dN(x) - \tilde{B}(t) = 0 \quad (3-13)
\]

where \( \tilde{B}(t) \) is the martingale estimate of \( B(t) \) given as

\[
\tilde{B}(t) = \int_0^t G(x) \, dN(x) = \int_0^t \frac{1}{S(t^-)I^n(t^-)} \, dN(x) \quad (3-14)
\]

If \( \{t_1, t_2, \ldots, t_d\} \) are observed infection times over the course of an epidemic, then \( dN(t) \) equals to one at these times and equals to zero elsewhere. To be a counting process, the number of infections must be one at each jump point. However, because of the coarser unit used in data collection, some of these \( t_j \)'s may have the same value. If the total number of infections up to \( t \) is given as \( N(t) \), then the martingale estimation of \( B(t) \) is simplified to

\[
\tilde{B}(t) = \sum_{k=1}^{N(t)} \frac{1}{(n-k+1)I^n(t^-_k)} \quad (3-15)
\]

where \( n-k+1 \) is the number of susceptibles at \( t^-_k = t_{k-1} \). The estimated variance is given by the predictable variation process,

\[
\tilde{Var}\{\tilde{B}(t)\} = \int_0^t G^2(x) \, dN(x) = \sum_{k=1}^{N(t)} \frac{1}{(n-k+1)I^n(t^-_k)^2} \quad (3-16)
\]

Becker (1989) provides the special case of (3-15) and (3-16) for \( m = 1 \).

If the epidemic has duration \( \tau \), then the overall transmission rate is estimated as

\[
\tilde{\beta} = \frac{\tilde{B}(\tau)}{\tau} = \sum_{k=1}^{N(\tau)} \frac{1}{\tau(n-k+1)I^n(t^-_k)} \quad (3-17)
\]

with the estimated variance as

\[
\tilde{Var}(\tilde{\beta}) = \sum_{k=1}^{N(\tau)} \frac{1}{\tau(n-k+1)I^n(t^-_k)^2} \quad (3-18)
\]
If the transmission rate is constant over time, denoted by $\beta_r$, then

$$B(\tau) = \int_0^\tau \beta_r \, du = \beta_r \tau \quad (3-19)$$

The martingale estimate for the overall transmission rate (3-17) and its estimated variance (3-18) can be used to estimate $\beta_r$ and $\text{var}(\beta_r)$.

### 3.2.2 SEIR Models

Let $E(t)$ denote the number of latent individuals at time $t$. In an SEIR model, the end of an epidemic is $\tau = \inf\{t : S(t)(I(t) + E(t)) = 0\}$, which may be similarly simplified as $\tau = \inf\{t : I(t) + E(t) = 0\}$. Furthermore, if the number of infectives in a data set is always greater than zero, then the same $F_r$-predictable process as used in the SIR model can be used for an SEIR model. Consequently, the same martingale estimate (3-15) and estimated variance (3-16) apply.

At times, it is possible in a SEIR model that $I(t) = 0$, yet $E(t) > 0$, then the $F_r$-predictable process needs to be reformulated as $G(t) = 1/[S(t^-)(I(t^-) + \varepsilon)^\gamma]$, where $\varepsilon$ is a small positive number to guarantee that the denominator never equals zero. Indeed, in an interval $a < t < b$, if $I(t^-) = 0$ then $dN(t)$ must be zero (if there are no infectives, then no infections may occur) and $G(t)dN(t) = 0$; if $I(t^-) \geq 1$, then $S(t^-)(I(t^-) + \varepsilon)^\gamma = S(t^-)I^\gamma(t^-)$. Thus, the same martingale estimate (3-15) and the estimated variance (3-16) are approximately obtained for an SEIR model.

Similarly, the same martingale estimate (3-17) for the overall transmission rate applies if the number of infectives is greater than zero during the course of the epidemic; otherwise (3-17) is obtained approximately.

When performing estimation using a SEIR model, it is not necessary to use the above $F_r$-predictable process discussed here. If the data are such that $I(t^-) \geq 1$ for all $t \leq \tau$, the $F_r$-predictable process as given in Section 3.2.1 may be used. Then the martingale estimate (3-15) and the estimated variance (3-16) are obtained without approximation.
3.3 Estimating the Transmission Rate as a Continuous Function

The martingale method provides the non-parametric estimates of the cumulative transmission rate \( B(t) = \int_0^t \beta(u) \, du \), but in practice, the transmission rate \( \beta(t) \) as a continuous function of time is more useful because it provides a direct and precise measurement of its variations over time. Difficulties associated with estimating the transmission rate are similar to those encountered in estimating density functions. In general, there are two ways to obtain a continuous function estimate of the transmission rate, interpolation and smoothing.

3.3.1 The Cubic Spline Method

In this section, the histospline interpolation method developed by Hsieh (1991a, 2001) for estimating hazard functions is extended to estimate the transmission rate. The cumulative transmission rate can be estimated as a differentiable function by interpolation. In general, a piecewise cubic spline provides a good estimate of the underlying function. As cubic polynomials have four coefficients, there is sufficient flexibility to ensure that the approximate function is first and second order continuously differentiable in any interval. The first derivative of the cubic polynomial gives the estimate of the transmission rate as a differentiable function of time.

The following definitions for various types of cubic splines are taken from Burden and Faires (1981, Section 3.4). For a function \( f \) defined on \( a = x_0 < x_1 < \cdots < x_n = b \), called nodes, a cubic spline interpolant \( g \), of \( f \) is a function that satisfies the following conditions:

1. \( g \) is defined piecewise by patching together cubic polynomials \( g_j \), defined on the subinterval \([x_j, x_{j+1}]\) for each \( j = 0, 1, \ldots, n - 1\);
2. \( g(x_j) = f(x_j) \) for each \( j = 0, 1, \ldots, n \);
3. \( g_{j+1}(x_{j+1}) = g_j(x_{j+1}) \) for each \( j = 0, 1, \ldots, n - 2 \);
4. \( g'_{j+1}(x_{j+1}) = g'_j(x_{j+1}) \) for each \( j = 0, 1, \ldots, n - 2 \);
(5) $g_{j+1}^{*}(x_{j+1}) = g_{j}^{*}(x_{j+1})$ for each $j = 0, 1, \ldots, n - 2$;

(6) one of the following boundary conditions

\begin{align*}
g^{*}(x_0) &= g^{*}(x_n) = 0 \quad \text{(a natural spline)}; \\
g^{'}(x_0) &= f^{'}(x_0) \quad \& \quad g^{'}(x_n) = f^{'}(x_n) \quad \text{(a complete spline)}.
\end{align*}

The cumulative transmission rate $B(t_j)$ is estimated at each observed infection time $x_j$, $j = 0, 1, \ldots, n$. A piecewise twice differentiable cubic spline function $g(x)$, which interpolates these points $(x_j, B(t_j))$, gives a continuous estimation of $B(t)$, then $\beta(t)$ can be estimated as $\beta(t) = g^{'}(t)$. Thus, the transmission rate is a piecewise continuous quadratic function. The natural spline is a reasonable choice to interpolate data points $(x_j, B(t_j))$ because the instantaneous changes in the transmission rate $(\beta^{'}(t))$ at the beginning and at the end of an epidemic are close to zero.

Because $B^{'}(t)$ is the transmission rate, it should be nonnegative at all times. If the increment of the function values between two consecutive nodes is too large relative to the length of the interval, then $B^{'}(t)$ may have negative values in this interval. If this occurs, one of these two nodes may be omitted.

### 3.3.2 The Kernel Smoothing Method

Becker (1989) used the kernel smoothing method to obtain the transmission rate. A kernel estimator for the infection transmission rate $\beta(t)$ from the estimated cumulative rate $\tilde{B}(t)$ is

$$\beta(t) = \frac{1}{b} \int_{0}^{1} K\left(\frac{t - x}{b}\right) d\tilde{B}(x) \quad (3-20)$$

where $K$ is a kernel function and the bandwidth $b$ is a positive constant. In this case, the integration is actually a summation over the infection times. The Epanechnikov kernel function

$$K(x) = 0.75(1-x)^2, \quad -1 < x < 1 \quad (3-21)$$

is most frequently used in the literature. There have been many discussions about the choice of the kernel function and the bandwidth in the context of density estimation (Bean and Tsokos, 1980). Hastie and Tibshirani (1990) point out that the choice of bandwidth is much more
important than the choice of kernel. Bandwidths can be calculated as the average of the distance from each data point to its 1st, 2nd, and 3rd nearest neighbours among the remaining data points (data points to its left and right), or to its 1st, 2nd, and 3rd nearest neighbours of all remaining data points (data points to its right only).

Each method (the cubic spline method and the kernel method) has its pros and cons, see Bean and Tsokos (1980) and Hsieh (2001). Comments on these issues will not be discussed further here, but for completeness, the kernel smoothing method is also used to estimate the transmission rate and various bandwidths are also examined.

3.4 Estimating the Number of Adequate Contacts

The martingale method cannot provide estimates on the number of adequate contacts \( m \) required for the successful transmission of pathogen. This \( m \) needs to be provided in order to apply the martingale method. Because \( m \) reflects the infectiosity of the pathogen, it should be assessable from lab experiments together with statistical models. However, in this section, only a statistical model is used to estimate \( m \) because the lab results on the infectiosity is not available in this historical data.

The generalized linear model (GLM) may be used to obtain an estimate of \( m \). Becker (1986) used the GLM to make inferences on the transmission rate. Let \( O_k \) denote the number of new infections at epoch \( k \). Taking a log transformation of (2-18), assuming a constant transmission rate, and removing the expectation by adding the error term, yields the log-linear model

\[
\log(O_k) = \log(\beta) + \log S_{k-1} + m \log I_{k-1} + \epsilon_k
\]

where \( S_{k-1} \) and \( I_{k-1} \) are the number of susceptibles and infectives respectively at epoch \( k-1 \). The error term may be taken as a Poisson random variable because the number of infections at an epoch conditioning on a filtration follows the Poisson distribution. Parameters of this model are \([\beta, m]\). An SAS procedure "Genmod" can provide the maximum likelihood estimate (MLE) for \( \log \beta \) and \( m \) by specifying \( \log S_{k-1} \) as "OFFSET".

In general, the transmission rate is not constant. If the transmission rate is time dependent with the following form \( \beta = \beta_0 e^{\alpha t} \), then the log-linear model is
\[
\log(O_k) = \log(\beta_0) + \alpha k + \log S_{k-1} + m \log I_{k-1} + \varepsilon_k \quad (3-23)
\]

with parameters \(\{\beta_0, \alpha, m\}\) in this model. Similarly, \(\log S_{k-1}\) will be used as "OFFSET", and the MLE of these parameters are obtained from "Genmod". More importantly, the significance of \(\alpha\), in other word, whether the transmission rate is time dependent, can be examined using Wald’s test.

Estimates from GLM for the smallpox outbreak and the respiratory disease epidemic are summarized in Table 3-1 and Table 3-2 where models (3-22) and (3-23) are used. The transmission rates for both diseases are time dependent as the time is a significant term with P-value less than 0.05. The predicted number of adequate contacts is 0.82 for the smallpox outbreak data and 0.65 for the respiratory disease epidemic. Note that in both examples, \(m\) significantly differs from 0, which indicates the mode of transmission is via person-to-person transmission. However, in the smallpox outbreak example, \(m\) estimated in the time dependent model does not significantly differ from 1; but in the respiratory disease epidemic, \(m\) significantly differs from 1 in both time-dependent and time-independent models.
Table 3-1 Parameter estimation using GLM – smallpox outbreak

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>P-value</th>
<th>Log Likelihood (Degree of Freedom)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Independent</td>
<td>log β</td>
<td>-5.9867</td>
<td>0.2978</td>
<td>0.0001</td>
<td>-57.80 (82)</td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>0.3905</td>
<td>0.2673</td>
<td>0.1441</td>
<td></td>
</tr>
<tr>
<td>Time Dependent</td>
<td>log β₀</td>
<td>-5.3615</td>
<td>0.3519</td>
<td>0.0001</td>
<td>-54.42 (81)</td>
</tr>
<tr>
<td></td>
<td>α</td>
<td>-0.0273</td>
<td>0.0115</td>
<td>0.0181</td>
<td></td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>0.8156</td>
<td>0.3229</td>
<td>0.0115</td>
<td></td>
</tr>
</tbody>
</table>

Table 3-2 Parameter estimation using GLM – respiratory disease epidemic

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>P-value</th>
<th>Log Likelihood (Degree of Freedom)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Independent</td>
<td>log β</td>
<td>-6.010</td>
<td>0.3088</td>
<td>0.0001</td>
<td>-26.22 (31)</td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>0.4969</td>
<td>0.1529</td>
<td>0.0012</td>
<td></td>
</tr>
<tr>
<td>Time Dependent</td>
<td>log β₀</td>
<td>-5.2833</td>
<td>0.3637</td>
<td>0.0001</td>
<td>-22.61 (30)</td>
</tr>
<tr>
<td></td>
<td>α</td>
<td>-0.0652</td>
<td>0.0273</td>
<td>0.0167</td>
<td></td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>0.6497</td>
<td>0.1757</td>
<td>0.0002</td>
<td></td>
</tr>
</tbody>
</table>

3.5 Applications

Data from a smallpox outbreak in Abakaliki and a respiratory disease epidemic on the island of Tristan da Cunha will be studied in this section.

3.5.1 Smallpox Outbreak

Data from a smallpox outbreak in Abakaliki, a community in south-eastern Nigeria, made available by Bailey and Thomas (1971), have been studied by Becker (1989) to estimate the transmission rate under the bilinear incidence assumption. A total of 30 cases appeared among 120 individuals who were initially susceptible.
The latent and infectious periods for smallpox given by Benenson (1970) are 13 and 7 days respectively. Note that these figures are only the mean for the latent and infectious period and variability is associated with the actual latent or infectious period of each infected individual. If, using these values, the primary case would have recovered by the end of day 7, then it would have been impossible for any susceptibles to receive any infectious contacts between day 8 and day 13. However, the fourth case received an infectious contact on day 9 and another three cases had theirs on day 12. Becker (1989) suggested using 14 days as the infectious period for the primary case and 7 days for the remaining cases. He argued that the effective infectious period is likely to be longer for the initial case than for those cases arising after the presence of this serious disease had been discovered, that is, the infectious period for the primary case had a larger variance than the remaining cases. Nonetheless, the latent and infectious periods for smallpox are rather accurate, that is, the variance associated with each period is small, so that the infection time for each infection is approximated rather accurately. With these assumptions, an SEIR model with a fixed latent period and a fixed infectious period is considered to study this outbreak. This data set is available in Becker (1989).

Smallpox is caused by an orthopoxvirus, which is normally of a high infectiosity. The GLM provides an estimate of $m$ as 0.8156 when the transmission rate is time dependent. For analytical purposes, one adequate contact ($m = 1$) is also studied. The martingale estimates of the cumulative transmission rates $B(t)$ using (3-15) and estimated variances using (3-16) for two different values of $m$ are summarized in Table 3-3. Becker also arrived at the estimates for $m = 1$ (Becker, 1989, Section 7.6). The estimates of the cumulative transmission rate are similar for different values of $m$, but the estimated cumulative transmission rate under $m = 0.8$ is higher than the estimate under $m = 1$ after the 26th day. This is because there were more than one infective in the population after that day and $\tilde{B}(x)$ is inversely related to $I^*(t)$, see (3-15).

The last infection occurred on day 64, so that the interpolation was done from day 0 to day 64 with the initial conditions as $B^*(0) = B^*(64) = 0$. Because an infection occurred on the first day, the increment of the cumulative transmission rates from day 0 to day 1 is very large, and the derivative of the cubic function has negative values. Consequently, the node at day 1 is omitted in both cases ($m = 0.8$ and $m = 1$). There is also a relatively large increment from day 38 to day 39. The derivative of the cubic function also has negative values between these two days. So the node at day 39 is also omitted in both cases ($m = 0.8$ and $m = 1$). These adjustments
are arrived at after evaluating other possible adjustments. So a cubic spline \( g(t) \), which interpolates the cumulative transmission rate, is obtained for each case.

The estimated transmission rates \( \tilde{g}(t) = g'(t) \) for \( m = 0.8 \) and \( m = 1 \), as shown in Figure 3-1, are piecewise quadratic functions with a continuous first derivative. These two functions have similar shapes, and are identical before day 20. The transmission rate for this outbreak peaks around day 13 before dropping and becoming relatively stable for the rest of the epidemic period.

For the smallpox outbreak data, Becker (1989) uses the Epanechnikov kernel function and a bandwidth of 9 to obtain the estimated transmission rate under \( m = 1 \). The plot of the estimated transmission rate is available in Section 7.9 in Becker (1989). Because a large bandwidth will sacrifice the goodness-of-fit, Becker's estimation does not have small oscillations as shown in the cubic spline estimation. The estimated transmission rate using kernel method with bandwidth 9 is different from the one estimated using the cubic spline method. So different bandwidths calculated from this particular data set are also examined. Bandwidths calculated as the average of the distance from each data point to its 1\(^{st}\), 2\(^{nd}\), and 3\(^{rd}\) nearest neighbours among the remaining data points are 3.0, 5.7 and 8.3, respectively. Using the average of the distance from each data point to its 1\(^{st}\), 2\(^{nd}\), and 3\(^{rd}\) nearest neighbours of all remaining data points, the bandwidths are 2.4, 4.7, and 7.2, respectively. Using a bandwidth between 3 and 5 in the kernel method produces similar estimates of the transmission rate as the cubic spline method in both cases (\( m = 0.8 \) and \( m = 1 \)).
Table 3-3 Estimate and variance of $\tilde{B}(t)$ – smallpox outbreak

<table>
<thead>
<tr>
<th>$k$</th>
<th>$t_k$</th>
<th>$S(t_k^-)$</th>
<th>$I(t_k^-)$</th>
<th>$O_k$</th>
<th>$m = 0.8$</th>
<th>$m = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\tilde{B}(t_k^-)$</td>
<td>s.e. $\tilde{B}(t_k^-)$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>119</td>
<td>1</td>
<td>1</td>
<td>0.00840</td>
<td>0.00840</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>118</td>
<td>1</td>
<td>1</td>
<td>0.01688</td>
<td>0.01193</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>117</td>
<td>1</td>
<td>1</td>
<td>0.02542</td>
<td>0.01468</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>116</td>
<td>1</td>
<td>3</td>
<td>0.05151</td>
<td>0.02103</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>113</td>
<td>1</td>
<td>1</td>
<td>0.06036</td>
<td>0.02282</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>112</td>
<td>1</td>
<td>1</td>
<td>0.06929</td>
<td>0.02450</td>
</tr>
<tr>
<td>7</td>
<td>23</td>
<td>111</td>
<td>1</td>
<td>1</td>
<td>0.07830</td>
<td>0.02611</td>
</tr>
<tr>
<td>8</td>
<td>26</td>
<td>110</td>
<td>2</td>
<td>1</td>
<td>0.08352</td>
<td>0.02662</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
<td>109</td>
<td>6</td>
<td>2</td>
<td>0.08792</td>
<td>0.02680</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>107</td>
<td>5</td>
<td>2</td>
<td>0.09310</td>
<td>0.02705</td>
</tr>
<tr>
<td>11</td>
<td>35</td>
<td>105</td>
<td>1</td>
<td>1</td>
<td>0.10262</td>
<td>0.02868</td>
</tr>
<tr>
<td>12</td>
<td>38</td>
<td>104</td>
<td>2</td>
<td>1</td>
<td>0.10815</td>
<td>0.02921</td>
</tr>
<tr>
<td>13</td>
<td>39</td>
<td>103</td>
<td>1</td>
<td>1</td>
<td>0.11786</td>
<td>0.03078</td>
</tr>
<tr>
<td>14</td>
<td>43</td>
<td>102</td>
<td>4</td>
<td>2</td>
<td>0.12436</td>
<td>0.03112</td>
</tr>
<tr>
<td>15</td>
<td>44</td>
<td>100</td>
<td>5</td>
<td>1</td>
<td>0.12712</td>
<td>0.03124</td>
</tr>
<tr>
<td>16</td>
<td>45</td>
<td>99</td>
<td>5</td>
<td>1</td>
<td>0.12990</td>
<td>0.03137</td>
</tr>
<tr>
<td>17</td>
<td>46</td>
<td>98</td>
<td>5</td>
<td>1</td>
<td>0.13272</td>
<td>0.03149</td>
</tr>
<tr>
<td>18</td>
<td>48</td>
<td>97</td>
<td>4</td>
<td>2</td>
<td>0.13956</td>
<td>0.03186</td>
</tr>
<tr>
<td>19</td>
<td>49</td>
<td>95</td>
<td>3</td>
<td>1</td>
<td>0.14393</td>
<td>0.03216</td>
</tr>
<tr>
<td>20</td>
<td>54</td>
<td>94</td>
<td>3</td>
<td>2</td>
<td>0.15281</td>
<td>0.03277</td>
</tr>
<tr>
<td>21</td>
<td>59</td>
<td>92</td>
<td>5</td>
<td>1</td>
<td>0.15581</td>
<td>0.03291</td>
</tr>
<tr>
<td>22</td>
<td>64</td>
<td>91</td>
<td>6</td>
<td>1</td>
<td>0.15843</td>
<td>0.03301</td>
</tr>
</tbody>
</table>

$S(t_k^-)$ - Number of susceptibles at $t_k^-$.  
$I(t_k^-)$ - Number of infectives at $t_k^-$.  
$O_k$ - Number of infections at epoch $k$ or in the time interval $(t_{k-1}, t_k]$.  
$\tilde{B}(t_k^-)$ - Estimated cumulative transmission rate.  
\(\text{s.e. } \tilde{B}(t_k^-)\) - Standard error of the estimated cumulative transmission rate.
3.5.2 Respiratory Disease Epidemic

Data from the October/November 1967 respiratory disease epidemic on the island Tristan da Cunha has been studied by Becker (1989) for other purposes. These data were collected by medical officers of the British Medical Research Council. A total of 39 cases were recorded among the community of 255 islanders. Becker and Hopper (1983) found evidences to support the assumption that all islanders were initially susceptible. The first case among islanders was taken as the beginning of the epidemic. The date on which symptoms first appeared, and days that symptoms lasted were recorded for each consequent case. Becker (1989) assumed each infection took place one day before the first onset of symptoms and the infectious period is taken to be the days at which the individual displayed symptoms. Thus, the incubation period is
assumed to be 1 day. Because the infectious period is assumed in this case as the period in which the symptom displayed, the latent period is the same as the incubation period. With these assumptions, it is possible to trace the number of susceptibles, latent, individuals, infectives and recoveries over the course of the epidemic. An \textit{SEIR} model with a fixed latent period is used to study this epidemic. The created data are given in Appendix A.

The GLM estimates \( m \) to be 0.6497 when the transmission rate is time dependent. The cumulative transmission rates (using (3-15)) and estimated variances (using (3-16)) for \( m = 0.6 \) and \( m = 1 \) are summarized in Table 3-4. Similarly, because there were many infectives in the population after the 10\textsuperscript{th} day, the estimated cumulative transmission rate under \( m = 0.6 \) is similar to the estimate under \( m = 1 \) before that day, but it is higher after that day.

The spline method is applied to estimate the transmission rate. Two points, day 9 and day 27, need to be omitted in order to have all positive estimates of the transmission rate. The continuous estimates of the transmission rate \( \beta(t) \) are shown in Figure 3-2 for \( m = 0.6 \) and \( m = 1 \). The cubic spline estimate of the transmission rate for \( m = 0.6 \) is also shown in Figure 3-2. The transmission rates for \( m = 0.6 \) and \( m = 1 \) follow closely.

Kernel estimates are also performed with various bandwidths: 0.73, 1.05, 1.95, 2.05, and 2.73. The kernel method using a bandwidth between 0.73 and 1.2 produces similar estimates as in the cubic spline method.
Table 3-4 Estimate and variance of $\tilde{B}(t)$ – respiratory disease epidemic

<table>
<thead>
<tr>
<th>$k$</th>
<th>$t_k$</th>
<th>$S(t_k)$</th>
<th>$I(t_k)$</th>
<th>$O_k$</th>
<th>$m = 0.6$</th>
<th>$m = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\tilde{B}(t_k)$</td>
<td>s.e. $\tilde{B}(t_k)$</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>254</td>
<td>1</td>
<td>1</td>
<td>0.00394</td>
<td>0.00394</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>253</td>
<td>1</td>
<td>2</td>
<td>0.01186</td>
<td>0.00685</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>251</td>
<td>1</td>
<td>4</td>
<td>0.02789</td>
<td>0.01054</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>247</td>
<td>3</td>
<td>4</td>
<td>0.03632</td>
<td>0.01083</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>243</td>
<td>7</td>
<td>6</td>
<td>0.04408</td>
<td>0.02300</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>237</td>
<td>12</td>
<td>3</td>
<td>0.04694</td>
<td>0.02306</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>234</td>
<td>11</td>
<td>5</td>
<td>0.05206</td>
<td>0.02317</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>229</td>
<td>12</td>
<td>1</td>
<td>0.05304</td>
<td>0.02319</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>228</td>
<td>15</td>
<td>2</td>
<td>0.05477</td>
<td>0.02323</td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>226</td>
<td>14</td>
<td>3</td>
<td>0.05751</td>
<td>0.02328</td>
</tr>
<tr>
<td>11</td>
<td>18</td>
<td>223</td>
<td>16</td>
<td>2</td>
<td>0.05921</td>
<td>0.02331</td>
</tr>
<tr>
<td>12</td>
<td>19</td>
<td>221</td>
<td>16</td>
<td>1</td>
<td>0.06007</td>
<td>0.02333</td>
</tr>
<tr>
<td>13</td>
<td>20</td>
<td>220</td>
<td>17</td>
<td>2</td>
<td>0.06173</td>
<td>0.02336</td>
</tr>
<tr>
<td>14</td>
<td>27</td>
<td>218</td>
<td>1</td>
<td>1</td>
<td>0.06632</td>
<td>0.02380</td>
</tr>
<tr>
<td>15</td>
<td>28</td>
<td>217</td>
<td>1</td>
<td>1</td>
<td>0.07093</td>
<td>0.02424</td>
</tr>
</tbody>
</table>

$S(t_k)$ - Number of susceptibles at $t_k$.

$I(t_k)$ - Number of infectives at $t_k$.

$O_k$ - Number of infections during day $t_k$.

$\tilde{B}(t_k)$ - Estimated cumulative transmission rate.

s.e. $\tilde{B}(t_k)$ - Standard error of the estimated cumulative transmission rate.
3.6 Population Heterogeneity – Stratified Analysis

The assumption of a homogeneous population may not hold true in reality. Population heterogeneity may be introduced by infectives and/or by susceptibles. The amount of pathogen within an infective individual’s system determines his infectiousness, and this amount varies during his infectious period. However, at the time of infection, which infective is responsible for the transmission of the pathogen is unobservable. So the infectiousness is often considered to be the same for each infective, that is, infectives are homogeneous.

The host defence mechanisms are best explained by nonspecific and specific defence mechanisms, such as antibodies, B cells and T cells. However, other variables, such as age, gender, behavior, and genetic make-up that may cause heterogeneity among susceptibles, are
often available. Because these factors are considered as the host factors in epidemiology, it is certainly of interest to estimate and to test the effect of these factors on the transmission rate.

The stratified non-parametric approach is useful when dealing with one or two such factors. Susceptibles are classified into a small number of strata specified by these factors and within each stratum susceptibles are homogenous. The incidence process for a heterogeneous population is derived in Section 2.8, and the martingale method discussed in Section 3.2 can be applied within each stratum. Similarly, estimates of $\beta^k(t)$ can be obtained using the cubic spline method or the kernel method, which may provide certain understanding on the factor effects.

The relative risk defined as the ratio of the stratum-specific incidences in epidemiology (see Breslow and Day, 1980, Section 2.4), can be used to study the factor effects. In Section 3.2, the overall transmission rate is directly estimated using the martingale method. If $\beta^k$ is the overall transmission rate for the $k$-th stratum, then the relative risk of one group to another provides an estimate of the effect of this factor. For example, the age factor classifies the susceptibles into three groups ($k = 3$): infants, schoolchildren and adults, then the relative risk of infants as compared to adults is defined as

$$RR = \beta^1 / \beta^3 = \phi^1 / \phi^3 \quad (3-24)$$

where the second equality follows from the assumption (2), that is, infectives are homogenous. A hypothesis of $RR$ equal to 1 can be tested to see if the age effect is significant, that is, whether infants have higher susceptibility than adults. This type of hypothesis testing has been discussed frequently in epidemiology (Lilienfeld ans Stolley, 1994. Breslow and Day, 1980, and Andersen et. al., 1993).

### 3.6.1 Stratified Analysis – Respiratory Disease Epidemic

During the October/November 1967 respiratory disease epidemic on the Island of Tristan da Cunha, all the natives were initially susceptible. So heterogeneity may exist in the susceptible population because of age. The community consisted of 27 infants (0-4 years), 40 schoolchildren (5-16 years), and 188 adults (>16 years). Table 3-5 shows the cumulative transmission rate estimated using the martingale method for each age group using $m = 1$. The cubic spline estimate of the transmission rate for each groups is shown in Figure 3-3. Infants have a higher transmission rate than schoolchildren and adults. The differences between schoolchildren and
adults are not obvious, except at the beginning of the epidemic, and the transmission rate for adults is rather constant over time. Figure 3-3 also shows the transmission rate without stratification, which lies in between the two curves for infants and schoolchildren.

Table 3-5 Estimate and variance of $\tilde{B}(t)$ by age group – respiratory disease epidemic

<table>
<thead>
<tr>
<th>$k$</th>
<th>$t_k$</th>
<th>$\tilde{B}^1(t_k)$</th>
<th>s.e. $\tilde{B}^1(t_k)$</th>
<th>$\tilde{B}^2(t_k)$</th>
<th>s.e. $\tilde{B}^2(t_k)$</th>
<th>$\tilde{B}^3(t_k)$</th>
<th>s.e. $\tilde{B}^3(t_k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00535</td>
<td>0.00003</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.02500</td>
<td>0.00083</td>
<td>0.01072</td>
<td>0.00006</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>0.11550</td>
<td>0.00445</td>
<td>0.05064</td>
<td>0.00128</td>
<td>0.01072</td>
<td>0.00006</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>0.14388</td>
<td>0.00485</td>
<td>0.05941</td>
<td>0.00136</td>
<td>0.01253</td>
<td>0.00006</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>0.16432</td>
<td>0.00499</td>
<td>0.05941</td>
<td>0.00136</td>
<td>0.01487</td>
<td>0.00006</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>0.17333</td>
<td>0.00503</td>
<td>0.06167</td>
<td>0.00136</td>
<td>0.01487</td>
<td>0.00006</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>0.17333</td>
<td>0.00503</td>
<td>0.06419</td>
<td>0.00137</td>
<td>0.01689</td>
<td>0.00006</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>0.17333</td>
<td>0.00503</td>
<td>0.06419</td>
<td>0.00137</td>
<td>0.01736</td>
<td>0.00006</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>0.17333</td>
<td>0.00503</td>
<td>0.06610</td>
<td>0.00137</td>
<td>0.01774</td>
<td>0.00006</td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>0.17333</td>
<td>0.00503</td>
<td>0.06610</td>
<td>0.00137</td>
<td>0.01897</td>
<td>0.00006</td>
</tr>
<tr>
<td>11</td>
<td>18</td>
<td>0.17333</td>
<td>0.00503</td>
<td>0.06610</td>
<td>0.00137</td>
<td>0.01970</td>
<td>0.00006</td>
</tr>
<tr>
<td>12</td>
<td>19</td>
<td>0.17333</td>
<td>0.00503</td>
<td>0.06610</td>
<td>0.00137</td>
<td>0.02007</td>
<td>0.00006</td>
</tr>
<tr>
<td>13</td>
<td>20</td>
<td>0.17333</td>
<td>0.00503</td>
<td>0.06610</td>
<td>0.00137</td>
<td>0.02077</td>
<td>0.00007</td>
</tr>
<tr>
<td>14</td>
<td>27</td>
<td>0.17333</td>
<td>0.00503</td>
<td>0.06610</td>
<td>0.00137</td>
<td>0.02676</td>
<td>0.00010</td>
</tr>
<tr>
<td>15</td>
<td>28</td>
<td>0.17333</td>
<td>0.00503</td>
<td>0.06610</td>
<td>0.00137</td>
<td>0.03278</td>
<td>0.00014</td>
</tr>
</tbody>
</table>

$\tilde{B}^1(t_k)$, $\tilde{B}^2(t_k)$ and $\tilde{B}^3(t_k)$ - Estimated cumulative transmission rates for infants, schoolchildren and adults, respectively.

s.e. - Standard error.
3.6.2 Relative Risk Calculation – Respiratory Disease Epidemic

Table 3-6 summarizes the overall transmission rate and its standard error for each age group. Infants are five times more likely to be infected than adults ($RR$ is 5.29). Schoolchildren are twice as likely to be infected as compared to adults ($RR$ is 2.02). Using the martingale central limit theorem, the Chi-square tests show that the difference of the infection risks between infants and adults are statistically significant and the difference between schoolchildren and adults are not statistically significant. Furthermore, the relative risk of infants as compared to schoolchildren is 2.62 and the Chi-square test statistic is 1.93, which is not significant. (Notice that this result is not available in Table 3-6).
Table 3-6 Estimated overall transmission rates $\beta$, $RR$ and the test statistic for $RR = 1$

<table>
<thead>
<tr>
<th></th>
<th>Infants</th>
<th>Schoolchildren</th>
<th>Adults</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>0.005098</td>
<td>0.001944</td>
<td>0.000964</td>
</tr>
<tr>
<td>s.e.($\beta$)</td>
<td>0.002087</td>
<td>0.00109</td>
<td>0.000345</td>
</tr>
<tr>
<td>$RR$</td>
<td>5.29</td>
<td>2.02</td>
<td>-</td>
</tr>
<tr>
<td>$\chi$</td>
<td>9.40</td>
<td>1.11</td>
<td>-</td>
</tr>
</tbody>
</table>

$\beta$ – Overall transmission rate.

s.e.($\beta$) – Standard error of the estimated transmission rate.

$RR$ – Relative risk.
4 Infection Tables

For epidemic models without vital dynamics, the waiting time to infection \( T \) is the sojourn time in the susceptible state \( S \). The hazard function of \( T \) is the infection rate, which is defined as the product of the transmission rate and the expected number of infectives. In this Chapter, functions depicting the distribution of \( T \), which are not available in the epidemic modelling literature, will be derived from the hazard function. These functions, similar to life table functions in demography, include the probability of escaping infection, the number of infections over a time interval, the conditional infection probability, and the infection-free expectancy. These "life table" functions are of epidemiological significance; the interpretation of each function will be discussed. Two other derived functions, the Lorenz curve (Goldie, 1977) and the scaled-total-time-on-test (Barlow and Campo, 1975) will also be discussed. Some of these functions are used for predicting epidemics in a population by further specifying the initial number of infectives, the initial number of susceptibles, the latent period, the infectious period (or the recovery rate), the infectiosity of the pathogen (the number of adequate contacts \( m \)), and the transmission rate. Infection tables, similar to life tables in demography, see Chiang (1968) and Hsieh (1991b, 2001), provide a better picture of the spread of a disease during an epidemic than the transmission rate itself. Infection tables can be constructed from data collected after the epidemic has run its course. The infection table for the respiratory disease data will be constructed based on the estimated transmission rate estimated in Section 3.5. Infection tables may be created before epidemic attacks, so that they can be used to estimate the medical costs incurred. This is useful in helping health officials make decisions on disease control strategies. Using the estimated transmission rate and the conditional probability of infection, the number of infections will be predicted for each day during the epidemic for the smallpox outbreak data and the respiratory disease data. The goodness-of-fit will be evaluated by comparing the predicted and the observed infections.

4.1 Waiting Time to Infection

In a community, the start of an epidemic is indicated by the occurrence of the first case of a disease, and the end of an epidemic is known when the number of infectives and latent
individuals both reach zero. After an epidemic starts, each infective may contact any susceptibles in the community and may transmit the disease. As a result, each susceptible is associated with a random time, the \textit{waiting time of infection} in state $S$, also known as the \textit{sojourn time}, denoted by $T$. The \textit{length of an epidemic} $\tau$ is a random variable with sample space $(0,\infty)$. For epidemics, $\tau$ is finite; for endemics, $\tau$ is infinite. If the epidemic stops before all susceptibles become infective, the waiting time to infection $T$ is a mixed random variable with the sample space $(0,\tau)\cup\{\infty\}$. For those who eventually escaped infection, $T$ is equal to $\infty$; while for those infected during the epidemic, $T \in (0,\tau)$. If the epidemic stops when all susceptibles become infective, then the sample space of $T$ is $(0,\tau)$.

The hazard function of $T$ is defined as

$$h_T(t) = \lim_{\Delta t \to 0} \frac{P[T \leq t + \Delta t | T > t]}{\Delta t}$$

(4-1)

Because for models without vital dynamics, $T$ is the sojourn time in state $S$, the event $T > t$ implies that the individual is in $S$ at time $t$, while the event $T \leq t + \Delta t$ implies that the individual is no longer in $S$ at time $t + \Delta t$. Consequently, the hazard function is also the instantaneous rate of leaving state $S$ and moving to the next state. Regardless of the model type applied (\textit{SIR} or \textit{SEIR}), the infection rate is the instantaneous rate of a susceptible leaving state $S$. Therefore, the infection rate is the hazard function of the sojourn time in state $S$.

Under the general assumption of the infectiosity (as seen in Section 3.7), the hazard function of $T$ for a homogenous population is

$$h_T(t) = \beta(t)i^n(t) = \phi(t)\xi(t)\alpha(t)i^n(t)$$

(4-2)

where $i(t)$ is the expected number of infectives (a continuous function of $t$). When there are no infectives in the population, the hazard function of $T$ is a zero function.

Hazard functions are important when studying the distribution of random variables, but other functions that are one-to-one transformations of hazard functions are often used in statistics and demography.
4.2 Other Functions of the Waiting Time to Infection

Other functions of the waiting time to infection $T$ can be derived from the hazard function. These functions have been discussed in demography, Chiang (1968) and Hsieh (1991b, 2001). In demography, the random variable is the *time to death*, which is defined from zero to some upper limit, for example, 100 for human. Every individual in a birth cohort will eventually die before that limit. Therefore, the hazard function of the time to death tends to infinity as time approaches this upper limit (100). On the other hand, during an epidemic, the hazard function of $T$ depends on the number of infectives in the population, and thus, it is a zero function after the epidemic stopped (no infectives exist in the population). The detailed derivation of each function will not be shown because they are straightforward generalization of what are available in Chiang (1968) and Hsieh (1991b, 2001).

The distribution function of $T$

$$ F_T(t) = P[T \leq t] = 1 - \exp\left\{-\int_0^t \beta(u) i^n(u) du\right\} \quad (4-3) $$

gives the probability of a susceptible being infected up to time $t$ during the course of the epidemic. Note that $t < \tau$ and $\tau$ may be finite or infinite. The survival function of $T$

$$ S_T(t) = P[T > t] = 1 - F_T(t) = \exp\left\{-\int_0^t \beta(u) i^n(u) du\right\} \quad (4-4) $$

gives the probability of escaping infection at any time during the course of the epidemic. The density of $T$

$$ f_T(t) = \beta(t) i^n(t) \exp\left\{-\int_0^t \beta(u) i^n(u) du\right\} \quad (4-5) $$

gives the probability per unit time of being infected during the course of the epidemic. The functions mentioned above have been discussed in the epidemic modelling literature, see Becker (1989).

The probability of infection during $(t, t + x]$ is given as

$$ d_T(x; t) = P[t < T \leq t + x] = S_T(t) - S_T(t + x) \quad (4-6) $$
This function, multiplied by the number of susceptibles at the beginning of the epidemic, gives an estimate of the number of infections that may occur in the interval \((t, t + x]\).

The conditional infection probability

\[
q_T(x; t) = P[T \leq t + x | T > t] = \frac{d_T(x; t)}{S_T(t)} = 1 - \exp\left[-\int_t^{t+x} \beta(u) S_T(u) du\right]
\]  

(4-7)
gives the probability of being infected during the interval \((t, t + x]\) given the individual has not yet been infected at time \(t\). This conditional probability provides a different estimate than the probability of infection \(d_T(x; t)\), because it accounts for the fact that the individual is able to escape infection up to time \(t\). The product of \(q_T(x; t)\) and the expected number of susceptibles at the beginning of an interval \(s(t)\) gives an estimate of the number of infections that expected to occur in the interval \((t, t + x]\).

The two functions, \(d_T(x; t)\) and \(q_T(x; t)\), are of epidemiological importance because they can be used to predict the number of new cases based on current observations. Health officials can then use this information to distribute health care resources. If contracting the disease requires hospitalization, this estimation can provide a direct approximation of how much medical resources, such as hospital beds, medical personnels, and medical supplies, would be required at different periods during the course of an epidemic.

Because \(T\) is a mixed random variable, \(w = P\{T = \infty\}\) denotes the probability of a susceptible eventually escaping infection by the end of the epidemic. In an epidemic caused by moderate infectious diseases, many susceptibles will not be infected at the end of the epidemic, and thus, \(w\) is often a positive number less than 1. The expectation of \(T\) equals to infinity because of the non-zero weight at \(T = \infty\). However, the average waiting time to infection \(E\{T \wedge \tau\}\), if the observed epidemic length \(\tau\) is finite, is calculated as:

\[
e_T(0) = E(T \wedge \tau) = \int_0^\tau S_T(t) dt
\]

(4-8)

where \(T \wedge \tau\) is the minimum of \(T\) and \(\tau\). Similarly, for individuals who have escaped infection up to a time \(x\), the average time of being infection-free during the remaining course of the epidemic is defined as
This is called the *infection-free expectancy*.

The scaled total time on test $STTT(x)$ (Barlow and Campo, 1975), gives the proportion of the total time of infection after the epidemic starts $x$ time units:

$$STTT(x) = \frac{E[T \wedge x]}{E[T \wedge \tau]} = 1 - \frac{\int_{0}^{r} S_{\tau}(t) dt}{\int_{0}^{r} S_{\tau}(t) dt} \quad (4-10)$$

The Lorenz curve $L(x)$ (Goldie, 1977), gives the fraction of the total mean time of being infection-free contributed by those individuals who have been infected by time $x$

$$L(x) = \frac{E[T \mid T \leq x]}{E[T \wedge \tau]} = 1 - \frac{\int_{0}^{r} S_{\tau}(t) dt + xS_{\tau}(x)}{\int_{0}^{r} S_{\tau}(t) dt} \quad (4-11)$$

where $\mathbb{1}(A)$ is the indicator function of the event $A$.

From the plot of $STTT(x)$ vs $F(x)$, the hazard function can be seen to be increasing, decreasing, or constant (that is, if $T$ is close to a Exponential distribution). From the plot of $L(x)$ vs $F(x)$, it can be seen whether or not $T$ can be degenerated to a constant.

### 4.3 Predicting Epidemic

Using an *SEIR* model, the disease spreading process may be predicted by specifying the initial number of susceptibles $s_0$ (also denoted by $n$), the initial number of infectives $i_0$ (also denoted by $a$), the transmission rate $\beta(t)$, the infectiosity of the pathogen (or the adequate number of contacts $m$), the latent period $U$ and the infectious period $V$ of the disease.

Given a transmission rate, the expected number of infections that may occur in a time interval $(t, t + x]$ is
\[ o(x,t) = s(t)q_r(x,t) = s(t) \left[ 1 - \exp \left( - \int_t^{\infty} \beta(u)i^m(u)du \right) \right] \quad (4-12) \]

where \( s(t) \) is the expected number of susceptibles at time \( t \).

Suppose the number of infections is predicted for each day, and the epidemic lasted for \( W \) days. Let \( O_k \) denote the number of new infections that occurred during day \( k \), \( k = 1,\ldots,W \). Let \( i_k \) and \( s_k \) be the number of infectives and susceptibles at the end of day \( k \), where \( i_0 \) and \( s_0 \) are known. Note that in continuous time, the number of individuals at the end of the day equals to the number of individuals at the beginning of the next day. For simplicity, the transmission rate is assumed to be constant within each day and denoted by \( \tilde{\beta} \), where \( k = 1,\ldots,W \). For each individual, the probability of infection during day \( k \) given that the individual is susceptible at the beginning of day \( k \) is denoted by \( q_k \), and is estimated as

\[ \tilde{q}_k = 1 - \exp(-\tilde{\beta}\tilde{i}_{k-1}^{m}) \quad (4-13) \]

where \( \tilde{i}_{k-1} \) is the predicted number of infectives at the end of day \( k - 1 \). Under random mixing, homogeneous susceptibles, and homogeneous infectives assumptions, \( O_k \) is estimated as

\[ \tilde{\sigma}_k = \tilde{s}_{k-1} \left( 1 - \exp(\tilde{\beta}\tilde{i}_{k-1}^{m}) \right) \quad (4-14) \]

where \( \tilde{s}_{k-1} \) is the predicted number of susceptibles at the end of day \( k - 1 \). The terms \( \tilde{s}_{k-1} \) and \( \tilde{i}_{k-1} \), for \( k \geq 1 \), also need to be predicted because the number of susceptibles and infectives are known only at the beginning of the epidemic (that is, only \( s_0 \) and \( i_0 \) are known).

The number of susceptibles at the end of each day \( k \geq 1 \) is equal to the number of susceptibles at the beginning of the day, minus the number of infections that occurred during that day

\[ \tilde{s}_k = \tilde{s}_{k-1} - \tilde{\sigma}_k \quad (4-15) \]

The number of infectives at the end of each day, using (2-26) given in Section 2.10, equals to the number of infectives at the beginning of the day, minus the number of recoveries, plus the number of latent individuals becoming infective during that day. The later two components need to be predicted based on the assumption of the latent and infectious periods.
If the latent period \( U = \mu \) and infectious period \( V = \alpha \) are both constants, then \( \tilde{\sigma}_{k-\mu} \) estimates the number of recoveries and \( \bar{\sigma}_{k-\mu} \) estimates the number of those who become infective on day \( k \) (or infected \( \mu \) days ago). The number of infectives is then given as

\[
\tilde{I}_k = \tilde{I}_{k-1} - \bar{\sigma}_{k-\mu} + \tilde{\sigma}_{k-\mu} \tag{4-16}
\]

for \( k \geq 1 \). If \( k \) is less than \( \alpha + \mu \) or \( \mu \), then \( \tilde{\sigma}_{k-\alpha-\mu} \) and \( \bar{\sigma}_{k-\mu} \) are taken to be zero.

If the latent period is constant \( U = \mu \), but the infectious period \( V \) is random with a certain distribution on the interval, then the above formula can be used by substituting the mean of the infectious period as \( \alpha \). However, if \( V \) has a large variance, this approximation will be inaccurate. Instead, the recovery rate \( \gamma(t) \) needs to be specified so that the number of recoveries at the end of each day may be obtained. Let the probability of recovery during day \( k \) denote by \( \nu_k \); then

\[
\nu_k = 1 - \exp\left(-\int_{k-1}^{k} \gamma(u) du\right) \tag{4-17}
\]

Similarly, assuming constant daily recovery rate denoted by \( \bar{\gamma}_k \), the estimated daily recovery probability is \( \bar{\nu}_k = 1 - \exp(-\bar{\gamma}_k) \). So the expected number of recoveries on day \( k \geq 1 \) is given as

\[
\bar{r}_k = \bar{I}_{k-1} \bar{\nu}_k \tag{4-18}
\]

and then the number of infectives is given as

\[
\tilde{I}_k = \tilde{I}_{k-1} - \bar{I}_{k-1} \bar{\nu}_k + \bar{\sigma}_{k-\mu} \tag{4-19}
\]

If \( k \) is less than \( \mu \), then \( \bar{\sigma}_{k-\mu} \) is taken to be zero.

### 4.4 Goodness-of-fit

Since an epidemic can be predicted, the goodness-of-fit can be evaluated by comparing the predicted and observed numbers of infections over the course of the epidemic. The number of infections are predicted using (4-14) with an estimated transmission rate.
4.4.1 Smallpox Outbreak

In the smallpox data, the latent and infectious periods are assumed constant; namely $\mu = 13$ and $\alpha = 7$, except for the adjustment of the infectious period for the first infective as 14 days. By using equations (4-14), (4-15) and (4-16), the number of infections that may occur during each day can be predicted. The transmission rate was estimated using the cubic spline method, which was shown in Figure 3-1 in Section 3.5.1 for the adequate number of contacts being 1 and 0.8. The number of infections up to a certain day is the sum of the daily infections up to that day. The predicted number of infections up to each day for the adequate contact being 1 ($m = 1$) is summarized in Table 4-1, which has fewer clusters of cases than the observation does. For example, the three cases that occurred on day 13 in the observation are separated into different days in the prediction. This is because the estimated transmission rate is a continuous function of time. Because the spline method fixes the transmission rate as 0 at the last occurrence of infection, the predicted epidemic length is the same as the observed. The total number of infections is predicted to be 32.58.

The number of infections over the course of the epidemic is similarly predicted using the kernel estimation of the transmission rate with a bandwidth 9 (suggested in Becker (1989)). The results for the adequate contacts being 1 ($m = 1$) are also summarized in Table 4-1. With a bandwidth of 9 days, the transmission rate is non-zero for a period after the last infection. Consequently, the predicted epidemic length is 7 days longer than the observed. The total number of infections by the end of the epidemic is 30.45. As shown in Figure 4-1, the predicted number of infections given by the cubic spline method is relatively close to the observed number of infections before day 40, while the prediction given by the kernel method is close to the observed number of infections after day 40.

The cumulative transmission rate was also estimated using $m = 0.8$ in Section 3.5.1. The predicted number of infections is estimated similarly using the cubic spline method and the kernel method, which are 31.04 and 28.73 respectively. The predicted number of the total infections using both methods is shown in Figure 4-2. The predicted number of infections using the kernel method follows the observed number of infections very closely.

The cubic spline method and the kernel method both give good fit to the data for $m = 1$ and $m = 0.8$. However, by comparing Figure 4-1 and Figure 4-2, $m = 0.8$ might be a better
assumption for this epidemic since the prediction given by $m = 0.8$ improves consistently in both methods than the prediction given by $m = 1$. 
Table 4-1 Expected and observed total number of infections – smallpox outbreak

<table>
<thead>
<tr>
<th>$t$</th>
<th>Observed</th>
<th>$m = 0.8$</th>
<th>$m = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cubic Spline</td>
<td>Kernel</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.08</td>
<td>0.12</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.24</td>
<td>0.26</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.47</td>
<td>0.43</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.74</td>
<td>0.62</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1.06</td>
<td>0.87</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1.41</td>
<td>1.20</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1.79</td>
<td>1.57</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2.18</td>
<td>1.99</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>2.65</td>
<td>2.43</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>3.36</td>
<td>2.90</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>4.31</td>
<td>3.41</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>5.39</td>
<td>3.92</td>
</tr>
<tr>
<td>13</td>
<td>6</td>
<td>6.48</td>
<td>4.44</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
<td>7.31</td>
<td>4.95</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>7.37</td>
<td>5.04</td>
</tr>
<tr>
<td>16</td>
<td>7</td>
<td>7.42</td>
<td>5.20</td>
</tr>
<tr>
<td>17</td>
<td>7</td>
<td>7.44</td>
<td>5.42</td>
</tr>
<tr>
<td>18</td>
<td>8</td>
<td>7.49</td>
<td>5.70</td>
</tr>
<tr>
<td>19</td>
<td>8</td>
<td>7.65</td>
<td>6.03</td>
</tr>
<tr>
<td>20</td>
<td>8</td>
<td>7.93</td>
<td>6.41</td>
</tr>
<tr>
<td>21</td>
<td>8</td>
<td>8.31</td>
<td>6.82</td>
</tr>
<tr>
<td>22</td>
<td>8</td>
<td>8.75</td>
<td>7.22</td>
</tr>
<tr>
<td>23</td>
<td>9</td>
<td>9.18</td>
<td>7.65</td>
</tr>
<tr>
<td>25</td>
<td>9</td>
<td>10.12</td>
<td>8.65</td>
</tr>
<tr>
<td>26</td>
<td>10</td>
<td>10.77</td>
<td>9.17</td>
</tr>
<tr>
<td>27</td>
<td>10</td>
<td>11.63</td>
<td>9.72</td>
</tr>
<tr>
<td>28</td>
<td>12</td>
<td>12.71</td>
<td>10.30</td>
</tr>
<tr>
<td>29</td>
<td>12</td>
<td>13.76</td>
<td>10.84</td>
</tr>
<tr>
<td>30</td>
<td>14</td>
<td>14.57</td>
<td>11.34</td>
</tr>
</tbody>
</table>
Table 4-1 Expected and observed total number of infections – smallpox outbreak

(continued)

<table>
<thead>
<tr>
<th>31</th>
<th>14</th>
<th>15.09</th>
<th>11.83</th>
<th>13.83</th>
<th>12.07</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>14</td>
<td>15.47</td>
<td>12.30</td>
<td>14.54</td>
<td>12.52</td>
</tr>
<tr>
<td>33</td>
<td>14</td>
<td>15.80</td>
<td>12.74</td>
<td>15.04</td>
<td>12.96</td>
</tr>
<tr>
<td>34</td>
<td>14</td>
<td>16.11</td>
<td>13.17</td>
<td>15.34</td>
<td>13.37</td>
</tr>
<tr>
<td>35</td>
<td>15</td>
<td>16.44</td>
<td>13.59</td>
<td>15.46</td>
<td>13.78</td>
</tr>
<tr>
<td>36</td>
<td>15</td>
<td>16.97</td>
<td>14.09</td>
<td>15.63</td>
<td>14.30</td>
</tr>
<tr>
<td>37</td>
<td>15</td>
<td>17.65</td>
<td>14.64</td>
<td>15.85</td>
<td>14.91</td>
</tr>
<tr>
<td>38</td>
<td>16</td>
<td>18.37</td>
<td>15.27</td>
<td>16.19</td>
<td>15.62</td>
</tr>
<tr>
<td>39</td>
<td>17</td>
<td>18.95</td>
<td>15.93</td>
<td>16.76</td>
<td>16.41</td>
</tr>
<tr>
<td>40</td>
<td>17</td>
<td>19.34</td>
<td>16.67</td>
<td>17.64</td>
<td>17.27</td>
</tr>
<tr>
<td>41</td>
<td>17</td>
<td>19.67</td>
<td>17.46</td>
<td>18.77</td>
<td>18.21</td>
</tr>
<tr>
<td>42</td>
<td>17</td>
<td>20.13</td>
<td>18.28</td>
<td>20.04</td>
<td>19.17</td>
</tr>
<tr>
<td>44</td>
<td>20</td>
<td>22.06</td>
<td>19.88</td>
<td>22.34</td>
<td>20.99</td>
</tr>
<tr>
<td>45</td>
<td>21</td>
<td>23.10</td>
<td>20.62</td>
<td>23.31</td>
<td>21.81</td>
</tr>
<tr>
<td>46</td>
<td>22</td>
<td>24.16</td>
<td>21.33</td>
<td>24.35</td>
<td>22.57</td>
</tr>
<tr>
<td>47</td>
<td>22</td>
<td>25.30</td>
<td>21.99</td>
<td>25.40</td>
<td>23.25</td>
</tr>
<tr>
<td>48</td>
<td>24</td>
<td>26.60</td>
<td>22.60</td>
<td>26.56</td>
<td>23.85</td>
</tr>
<tr>
<td>49</td>
<td>25</td>
<td>27.68</td>
<td>23.17</td>
<td>27.87</td>
<td>24.40</td>
</tr>
<tr>
<td>50</td>
<td>25</td>
<td>28.34</td>
<td>23.71</td>
<td>28.92</td>
<td>24.91</td>
</tr>
<tr>
<td>51</td>
<td>25</td>
<td>28.75</td>
<td>24.21</td>
<td>29.52</td>
<td>25.41</td>
</tr>
<tr>
<td>52</td>
<td>25</td>
<td>29.02</td>
<td>24.68</td>
<td>29.84</td>
<td>25.89</td>
</tr>
<tr>
<td>53</td>
<td>25</td>
<td>29.20</td>
<td>25.13</td>
<td>30.02</td>
<td>26.36</td>
</tr>
<tr>
<td>54</td>
<td>27</td>
<td>29.35</td>
<td>25.55</td>
<td>30.16</td>
<td>26.84</td>
</tr>
<tr>
<td>55</td>
<td>27</td>
<td>29.49</td>
<td>25.96</td>
<td>30.29</td>
<td>27.31</td>
</tr>
<tr>
<td>56</td>
<td>27</td>
<td>29.63</td>
<td>26.35</td>
<td>30.46</td>
<td>27.78</td>
</tr>
<tr>
<td>57</td>
<td>27</td>
<td>29.79</td>
<td>26.71</td>
<td>30.67</td>
<td>28.21</td>
</tr>
<tr>
<td>58</td>
<td>27</td>
<td>29.98</td>
<td>27.04</td>
<td>30.92</td>
<td>28.62</td>
</tr>
<tr>
<td>59</td>
<td>28</td>
<td>30.19</td>
<td>27.36</td>
<td>31.20</td>
<td>29.01</td>
</tr>
<tr>
<td>60</td>
<td>28</td>
<td>30.44</td>
<td>27.65</td>
<td>31.52</td>
<td>29.36</td>
</tr>
<tr>
<td>61</td>
<td>28</td>
<td>30.68</td>
<td>27.90</td>
<td>31.87</td>
<td>29.65</td>
</tr>
<tr>
<td>62</td>
<td>28</td>
<td>30.90</td>
<td>28.10</td>
<td>32.18</td>
<td>29.86</td>
</tr>
<tr>
<td>63</td>
<td>28</td>
<td>31.04</td>
<td>28.23</td>
<td>32.44</td>
<td>30.00</td>
</tr>
<tr>
<td>64</td>
<td>29</td>
<td>31.04</td>
<td>28.35</td>
<td>32.58</td>
<td>30.12</td>
</tr>
<tr>
<td>65</td>
<td>29</td>
<td>31.04</td>
<td>28.46</td>
<td>32.58</td>
<td>30.22</td>
</tr>
<tr>
<td>66</td>
<td>29</td>
<td>31.04</td>
<td>28.54</td>
<td>32.58</td>
<td>30.30</td>
</tr>
<tr>
<td>67</td>
<td>29</td>
<td>31.04</td>
<td>28.61</td>
<td>32.58</td>
<td>30.36</td>
</tr>
<tr>
<td>68</td>
<td>29</td>
<td>31.04</td>
<td>28.65</td>
<td>32.58</td>
<td>30.39</td>
</tr>
<tr>
<td>69</td>
<td>29</td>
<td>31.04</td>
<td>28.69</td>
<td>32.58</td>
<td>30.43</td>
</tr>
<tr>
<td>70</td>
<td>29</td>
<td>31.04</td>
<td>28.72</td>
<td>32.58</td>
<td>30.45</td>
</tr>
<tr>
<td>71</td>
<td>29</td>
<td>31.04</td>
<td>28.73</td>
<td>32.58</td>
<td>30.45</td>
</tr>
</tbody>
</table>
Figure 4-1 Expected and observed total number of infections - smallpox outbreak ($m = 1$)
Figure 4-2 Expected and observed total number of infections - smallpox epidemic ($m = 0.8$)
4.4.2 Respiratory Disease Epidemic

During the respiratory disease epidemic, the latent period is treated as a constant equal to 1 day, while the infectious period is random. Because the infectious period ranges from 2 to 11 days, the variance of the infectious period is large. The recovery rate has to be estimated in order to predict the number of recoveries.

The recovery rate is estimated by dividing the total number of recoveries within an interval by the total time that all infectives spend in the infective state within that interval. The recovery rate is assumed to be constant throughout this interval. In calculating the total time spend in the infective state, it ignores those individuals who recovered on a specific day because they are not considered to be in the infective state for a full day (for that specific day). By experimenting with various fractional values, the most suitable unit to use is 0.5 day, except on days 13, 21 and 22, where there were many recoveries. On these days, 0.75 day is used. The length of each interval is chosen by experimenting with different combinations, so that the number of recoveries within an interval is not too small and the interval itself is not too long. The recovery rate is evaluated by comparing the predicted number of recoveries with the observed number of recoveries. By comparing the observed with the expected number of recoveries, the recovery rate that gives the best fit to the data is selected and used for prediction. The estimated recovery rate is shown in Table 4-2.

The expected number of infections are estimated using (4-14), (4-15) and (4-19). The expected total number of infections up to each day over the course of the epidemic is shown in Table 4-2. As shown in Figure 4-3, the prediction by using the cubic spline method is closer to the observed; this can be explained by the following. Firstly, there are too many infections clustered during the period from day 9 to day 14, so that the estimated transmission rate is not as stable as the estimated transmission rate in the smallpox data. Secondly, the latent period for common cold has large variability among individuals. Thirdly, the infectious period is taken as from 1 day after the infection contact to the disappearance of symptoms, which may not be accurate. The infectious period for common cold often ends before the disappearance of the symptoms. Thus, the infective population $I(t)$ at time $t$ may be overestimated, which leads to the transmission rate being underestimated. Also, the estimated recovery rate may introduce errors into the prediction.
Table 4-2 Expected and observed total number of infections – respiratory disease epidemic
\( (m = 1) \)

<table>
<thead>
<tr>
<th>( t )</th>
<th>Recovery Rate</th>
<th>Observed</th>
<th>Expected</th>
<th>Cubic Spline</th>
<th>Kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1818</td>
<td>0</td>
<td>0.03</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.1818</td>
<td>0</td>
<td>0.07</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.1818</td>
<td>0</td>
<td>0.15</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.1818</td>
<td>0</td>
<td>0.25</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.1818</td>
<td>0</td>
<td>0.39</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.1818</td>
<td>0</td>
<td>0.56</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.1818</td>
<td>1</td>
<td>0.79</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.2469</td>
<td>1</td>
<td>1.40</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.2469</td>
<td>3</td>
<td>3.17</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.2469</td>
<td>7</td>
<td>6.86</td>
<td>1.88</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.2469</td>
<td>11</td>
<td>11.55</td>
<td>3.79</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.2469</td>
<td>17</td>
<td>13.27</td>
<td>6.87</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.2469</td>
<td>17</td>
<td>13.48</td>
<td>10.40</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.1505</td>
<td>20</td>
<td>16.86</td>
<td>13.15</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.1505</td>
<td>25</td>
<td>19.29</td>
<td>15.51</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.1505</td>
<td>26</td>
<td>19.67</td>
<td>17.68</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.1505</td>
<td>28</td>
<td>22.04</td>
<td>20.06</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.0909</td>
<td>31</td>
<td>23.94</td>
<td>22.14</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.0909</td>
<td>33</td>
<td>24.85</td>
<td>23.87</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.0909</td>
<td>34</td>
<td>25.81</td>
<td>25.65</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>0.7059</td>
<td>36</td>
<td>27.91</td>
<td>27.08</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.7407</td>
<td>36</td>
<td>29.75</td>
<td>28.01</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>0.1818</td>
<td>36</td>
<td>31.49</td>
<td>28.32</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.1818</td>
<td>36</td>
<td>33.69</td>
<td>28.40</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.5455</td>
<td>36</td>
<td>36.06</td>
<td>28.40</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>0.5455</td>
<td>36</td>
<td>37.93</td>
<td>28.40</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>0.2609</td>
<td>36</td>
<td>39.30</td>
<td>28.68</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>0.2609</td>
<td>37</td>
<td>40.13</td>
<td>29.17</td>
<td></td>
</tr>
</tbody>
</table>
Table 4-6 Expected and observed total number of infections – respiratory disease epidemic ($m = 1$) (continued)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>0.2609</td>
<td>38</td>
<td>40.13</td>
<td>29.84</td>
</tr>
<tr>
<td>30</td>
<td>0.2609</td>
<td>38</td>
<td>40.13</td>
<td>30.62</td>
</tr>
<tr>
<td>31</td>
<td>0.2609</td>
<td>38</td>
<td>40.13</td>
<td>31.29</td>
</tr>
<tr>
<td>32</td>
<td>0.2609</td>
<td>38</td>
<td>40.13</td>
<td>31.57</td>
</tr>
<tr>
<td>33</td>
<td>0.2609</td>
<td>38</td>
<td>40.13</td>
<td>31.57</td>
</tr>
</tbody>
</table>

Figure 4-3 Expected and observed total number of infections - respiratory disease epidemic ($m = 1$)
4.5 Construction of Infection Tables

Infection tables, similar to life tables, consist of a number of “life table” functions, many of which were discussed in Sections 4.2 and 4.3. This section describes the structure and construction of infection tables.

Infection tables contain the estimates of a number of basic life table functions:

1. time (in days), \( t \)
2. transmission rate, \( \beta(t) \)
3. number of infectives, \( i(t) \)
4. hazard function, \( h_r(t) \)
5. probability of escaping the infection, \( S_r(t) \)
6. infection probability over a time interval, \( d_r(t; x) \)
7. conditional infection probability, \( q_r(t; x) \)
8. infection-free expectancy, \( e_r(t) \)
9. number of infections, \( o(t) \)

Other functions that have been discussed in Section 4.2 may be included in an infection table. An infection table can be constructed after the epidemic has run its course. Constructing an infection table is done in the following four steps,

(a) obtain the martingale estimation of the cumulative transmission rate,
(b) estimate the transmission rate as a differentiable function using either the cubic spline method, or using the kernel smoothing method,
(c) estimate the hazard function by multiplying the observed number of infectives,
(d) estimate other infection table functions based on the estimated hazard function.

If a population is heterogeneous, a stratified analysis is required. After classifying the population into a small number of strata, an infection table can be constructed for each stratum by performing steps (a) to (d).

During an epidemic, from data collected up to a certain time \( t \), an infection table can be also created. The cubic spline method can be applied to estimate the cumulative transmission rate. The quadratic spline can be extended, or taken to be a certain function based on the trend.
shown up to time \( t \) for an extra period \( (t, t + d) \). The infection table can then be extended for the period \( (t, t + d) \), and the behaviour of the disease propagation in the population for the next period of time \( d \) can be predicted. These predictions are useful because they allow health officials to make preparations for the remaining period of the epidemic. The infection table may be updated as more data are collected over time, so decisions and control strategies may be adjusted accordingly.

Similarly, infection tables can be created for various epidemics by specifying the conditions discussed in Section 4.3. Health officials can use them to prepare the best solution, in terms of the health resource arrangements, and to anticipate the occurrences of specific epidemics.

### 4.6 Applications of Infection Tables

The infection table for the respiratory disease epidemic is shown in Table 4-3. The hazard function from day 8 to day 23 is relatively high as more susceptibles are likely to be infected during this period than on other days (see column 6). The infection-free expectancy decreases monotonically from 31.18 days at the start of the epidemic to 1 day at day 33 (see column 8). At the end of the epidemic, 83.9% of the population were not infected (see column 5).

If a random variable can be degenerated to a constant, then the Lorenz curve \( L(x) \) is a straight line (Goldie, 1977). However, \( L(x) \) for \( T \) (waiting time to infection in this epidemic) as shown in Figure 4-4 is not close to the straight line, and thus, \( T \) cannot be degenerated to a constant.

If a random variable has an Exponential distribution, then the \( STT(T) \) is a straight line (Barlow and Campo, 1975). However, \( STT(T) \) for \( T \) as shown in Figure 4-4 is not close to the straight line. So \( T \) is not close to the Exponential distribution; in other words, the hazard function of \( T \) is not constant over time. Because the number of infectives varies over time and the estimated transmission rate (see Figure 3-2) varies rapidly over time as well, the hazard function, which is the product of the two functions, will not be constant over time.
Table 4-3 Infection table – respiratory disease epidemic \((m = 1)\)

<table>
<thead>
<tr>
<th>(t_k)</th>
<th>(\beta(t_k))</th>
<th>(i(t_k))</th>
<th>(h_r(t_k))</th>
<th>(S_r(t_k))</th>
<th>(d_r(t_k; i_{k+1}))</th>
<th>(q_r(t_k; i_{k+1}))</th>
<th>(e_r(t_k))</th>
<th>(o(t_k))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000000</td>
<td>-</td>
<td>0.000000</td>
<td>1.000000</td>
<td>0.000000</td>
<td>0.000103</td>
<td>31.18</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>0.000208</td>
<td>1</td>
<td>0.000208</td>
<td>0.999897</td>
<td>0.000208</td>
<td>0.000315</td>
<td>30.19</td>
<td>0.03</td>
</tr>
<tr>
<td>2</td>
<td>0.000423</td>
<td>1</td>
<td>0.000423</td>
<td>0.999582</td>
<td>0.000423</td>
<td>0.000533</td>
<td>29.19</td>
<td>0.04</td>
</tr>
<tr>
<td>3</td>
<td>0.000645</td>
<td>1</td>
<td>0.000645</td>
<td>0.999049</td>
<td>0.000645</td>
<td>0.000759</td>
<td>28.21</td>
<td>0.08</td>
</tr>
<tr>
<td>4</td>
<td>0.000875</td>
<td>1</td>
<td>0.000875</td>
<td>0.998291</td>
<td>0.000873</td>
<td>0.000992</td>
<td>27.23</td>
<td>0.11</td>
</tr>
<tr>
<td>5</td>
<td>0.001112</td>
<td>1</td>
<td>0.001112</td>
<td>0.997300</td>
<td>0.001109</td>
<td>0.001233</td>
<td>26.26</td>
<td>0.14</td>
</tr>
<tr>
<td>6</td>
<td>0.001356</td>
<td>1</td>
<td>0.001356</td>
<td>0.996071</td>
<td>0.001351</td>
<td>0.002067</td>
<td>25.29</td>
<td>0.17</td>
</tr>
<tr>
<td>7</td>
<td>0.003371</td>
<td>0</td>
<td>0.000000</td>
<td>0.994012</td>
<td>0.000867</td>
<td>0.011140</td>
<td>23.34</td>
<td>0.22</td>
</tr>
<tr>
<td>8</td>
<td>0.005821</td>
<td>1</td>
<td>0.005821</td>
<td>0.984012</td>
<td>0.000867</td>
<td>0.011140</td>
<td>23.34</td>
<td>0.61</td>
</tr>
<tr>
<td>9</td>
<td>0.012108</td>
<td>1</td>
<td>0.012108</td>
<td>0.982938</td>
<td>0.011902</td>
<td>0.010208</td>
<td>22.59</td>
<td>1.76</td>
</tr>
<tr>
<td>10</td>
<td>0.007037</td>
<td>3</td>
<td>0.021111</td>
<td>0.972904</td>
<td>0.020539</td>
<td>0.010636</td>
<td>21.81</td>
<td>3.70</td>
</tr>
<tr>
<td>11</td>
<td>0.001219</td>
<td>7</td>
<td>0.008534</td>
<td>0.962557</td>
<td>0.008214</td>
<td>0.003062</td>
<td>21.04</td>
<td>4.69</td>
</tr>
<tr>
<td>12</td>
<td>0.000093</td>
<td>11</td>
<td>0.001027</td>
<td>0.959609</td>
<td>0.000985</td>
<td>0.006811</td>
<td>20.10</td>
<td>1.72</td>
</tr>
<tr>
<td>13</td>
<td>0.001586</td>
<td>12</td>
<td>0.019026</td>
<td>0.953074</td>
<td>0.018133</td>
<td>0.023238</td>
<td>19.23</td>
<td>0.21</td>
</tr>
<tr>
<td>14</td>
<td>0.001306</td>
<td>11</td>
<td>0.014370</td>
<td>0.930927</td>
<td>0.013377</td>
<td>0.003995</td>
<td>18.66</td>
<td>3.38</td>
</tr>
<tr>
<td>15</td>
<td>0.000159</td>
<td>12</td>
<td>0.001905</td>
<td>0.927208</td>
<td>0.001767</td>
<td>0.007008</td>
<td>17.73</td>
<td>2.43</td>
</tr>
<tr>
<td>16</td>
<td>0.000908</td>
<td>15</td>
<td>0.013527</td>
<td>0.920710</td>
<td>0.012546</td>
<td>0.014184</td>
<td>16.85</td>
<td>0.38</td>
</tr>
<tr>
<td>17</td>
<td>0.000823</td>
<td>14</td>
<td>0.011519</td>
<td>0.907650</td>
<td>0.010456</td>
<td>0.007834</td>
<td>16.08</td>
<td>2.37</td>
</tr>
<tr>
<td>18</td>
<td>0.000343</td>
<td>16</td>
<td>0.005486</td>
<td>0.900539</td>
<td>0.004940</td>
<td>0.004515</td>
<td>15.20</td>
<td>1.90</td>
</tr>
<tr>
<td>19</td>
<td>0.000340</td>
<td>16</td>
<td>0.005433</td>
<td>0.896474</td>
<td>0.004870</td>
<td>0.008539</td>
<td>14.26</td>
<td>0.91</td>
</tr>
<tr>
<td>20</td>
<td>0.000755</td>
<td>17</td>
<td>0.012840</td>
<td>0.888819</td>
<td>0.011413</td>
<td>0.016412</td>
<td>13.38</td>
<td>0.96</td>
</tr>
<tr>
<td>21</td>
<td>0.001167</td>
<td>9</td>
<td>0.010506</td>
<td>0.874232</td>
<td>0.009185</td>
<td>0.011748</td>
<td>12.58</td>
<td>2.10</td>
</tr>
<tr>
<td>22</td>
<td>0.001435</td>
<td>6</td>
<td>0.008608</td>
<td>0.863961</td>
<td>0.007437</td>
<td>0.009008</td>
<td>11.72</td>
<td>1.84</td>
</tr>
<tr>
<td>23</td>
<td>0.001557</td>
<td>6</td>
<td>0.009344</td>
<td>0.856178</td>
<td>0.008000</td>
<td>0.009307</td>
<td>10.82</td>
<td>1.74</td>
</tr>
<tr>
<td>24</td>
<td>0.001535</td>
<td>4</td>
<td>0.006141</td>
<td>0.848210</td>
<td>0.005209</td>
<td>0.005839</td>
<td>9.91</td>
<td>2.20</td>
</tr>
<tr>
<td>25</td>
<td>0.001369</td>
<td>3</td>
<td>0.004106</td>
<td>0.843257</td>
<td>0.003462</td>
<td>0.003688</td>
<td>8.96</td>
<td>2.37</td>
</tr>
<tr>
<td>26</td>
<td>0.001057</td>
<td>1</td>
<td>0.001057</td>
<td>0.840164</td>
<td>0.000888</td>
<td>0.000841</td>
<td>7.99</td>
<td>1.87</td>
</tr>
<tr>
<td>27</td>
<td>0.000601</td>
<td>1</td>
<td>0.000601</td>
<td>0.839458</td>
<td>0.000504</td>
<td>0.000312</td>
<td>7.00</td>
<td>1.37</td>
</tr>
<tr>
<td>28</td>
<td>0.000000</td>
<td>1</td>
<td>0.000000</td>
<td>0.839196</td>
<td>0.000000</td>
<td>0.000000</td>
<td>6.00</td>
<td>0.84</td>
</tr>
<tr>
<td>29</td>
<td>0.000000</td>
<td>2</td>
<td>0.000000</td>
<td>0.839196</td>
<td>0.000000</td>
<td>0.000000</td>
<td>5.00</td>
<td>0.00</td>
</tr>
<tr>
<td>30</td>
<td>0.000000</td>
<td>2</td>
<td>0.000000</td>
<td>0.839196</td>
<td>0.000000</td>
<td>0.000000</td>
<td>4.00</td>
<td>0.00</td>
</tr>
<tr>
<td>31</td>
<td>0.000000</td>
<td>2</td>
<td>0.000000</td>
<td>0.839196</td>
<td>0.000000</td>
<td>0.000000</td>
<td>3.00</td>
<td>0.00</td>
</tr>
<tr>
<td>32</td>
<td>0.000000</td>
<td>2</td>
<td>0.000000</td>
<td>0.839196</td>
<td>0.000000</td>
<td>0.000000</td>
<td>2.00</td>
<td>0.00</td>
</tr>
<tr>
<td>33</td>
<td>0.000000</td>
<td>1</td>
<td>0.000000</td>
<td>0.839196</td>
<td>0.000000</td>
<td>-</td>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Figure 4-4 Lorenz curve and STTT – respiratory disease epidemic ($m = 1$)
5 Multiplicative Model in Studying Observed Heterogeneity

Heterogeneity often exists among infectives and among susceptibles. Heterogeneity introduced by infectives cannot be estimated using statistical methods, because at the time of an infection the infectives responsible for the infection are seldom observable. By using the multiplicative model, it is now possible to study the observable heterogeneity introduced by susceptibles using statistical methods. In Section 3.6, stratified non-parametric analyses are used to study heterogeneity where it is shown that the transmission rate for infants and adults are significantly different. However, many factors can cause heterogeneity among susceptibles, such as the host (intrinsic) factors and household environmental factors that are specific to the host. Stratified analyses become inefficient when dealing with a large number of factors simultaneously. In this chapter, the Cox model (Cox, 1972), which handles the observed heterogeneity in survival analysis, will be extended to study the observed heterogeneity in infectious disease data. The hazard function of waiting time to infection $T$ for a heterogeneous population is a product of the baseline hazard function (the hazard function of time to infection $T$ for a homogeneous population) and the risks introduced by other factors. The maximum likelihood estimate of the baseline hazard is a generalization of the baseline hazard estimates given by Breslow (1974). The partial likelihood function of covariates for infectious disease data, which is independent of the baseline hazard, is the same as the one used for non-infectious disease data. Rhodes et. al. (1996) also arrives at a similar partial likelihood and score functions using a stochastic approach. The transmission rate that accounts for covariates is obtained by multiplying the estimated baseline hazard function and the estimated risk associated with covariates. The statistical method discussed in this chapter will be illustrated using the respiratory disease epidemic data that has been studied earlier.

5.1 A Survival Analysis Review

Survival analysis is a class of statistical methods for studying the occurrence and timing of events. These methods are often applied to study events such as deaths, onsets of diseases, equipment failures and social events, such as marriages and job terminations. An event is
considered to be a qualitative change that occurs in time. A qualitative change is a transition from one state to another, for example from a healthy state $H$ to a death state $D$. Note that the number of individuals in the death state does not have any impact on the rate of leaving the healthy state. Thus, the death process can be treated as a Markov process.

Covariates are factors that make a population heterogeneous, such as age, health status and treatment received. Before entering the death state, if the duration within an intermediate transient state may change the rate of entering the disease state, such as an illness state, the duration in the illness state is a time-dependent covariate. Thus, the death process becomes a semi-Markov process.

The heterogeneity that exists among individuals may cause different event times. Survival analysis is designed for longitudinal data on the occurrence of an event that aims at estimating causal or predictive models in which the risk of the event occurring depends on covariates. Several approaches have been developed to study the heterogeneity, such as non-parametric, semi-parametric and parametric, see Kalbfleisch (1980), Cox and Oakes (1984), Fleming and Harrington (1991), and Andersen et. al. (1993). As the Cox model can handle many different types of covariates simultaneously, it has received a lot of attention and has been applied in different research areas, such as social science, physics and engineering.

A cohort of individuals who enrolled onto a study at various times are observed over a period of time. Some of them may experience a certain event. Times observed for these individuals are event times. Some of them may not experience the event by the end of the study, or some may stop participating before the study ends. Times observed for such individuals are not event times, but the time of last contact. These are called censored observations. Censoring is an important feature of survival data; it takes many forms and occurs for different reasons, and have been dealt with using different methods in the literature. However, censoring seldom occur in infectious disease data because the epidemic length is relatively short and most individuals are observed over the entire epidemic.

If the event time, denoted by $U$, is the sojourn time in state $H$, then the Cox model states that the hazard function of $U$ has the following relationship with the covariates (Cox, 1972, Cox and Oakes, 1984, Chapter 7)

$$h_u(u; \mathbf{z}_u(u)) = h_0(u)\exp(b^T \mathbf{z}_u(u)) = h_0(u)\exp(\sum^p_{i=1} \beta_i z_{i,u}(u)) \quad (5-1)$$
where $b' = (b_1, \ldots, b_p)$ is a $p$-dimensional vector and $b'$ denotes transpose of vector $b$, $z_i(u) = (z_{i1}(u), \ldots, z_{ip}(u))'$ is a $p$-dimensional covariate vector at time $u$ for individual $i$ and $h_0(u)$ is the baseline hazard, which is common for all individuals. The baseline hazard is the hazard function for a homogenous population in which all covariates equal to zero. The additional risk related to a unit change in the $k$-th covariate is given by $\exp(b_k)$. This model is also referred to as a multiplicative risk model. Advantages of this model are discussed in Breslow and Day (1987). If all covariates are not time dependent, then the Cox model is often referred to as the proportional hazards model. Cox models, with or without time-dependent covariates, use the partial likelihood function rather than the conventionally used likelihood function for statistical analysis.

### 5.2 The Cox Model Applied to Epidemic Data

The event of most interest in epidemic modelling is the infection. Because the number of infectives in state $I$ has a direct impact on the infection rate, the infection process differs from the death process. If the infection process is captured by an SIR model, the time to event is the waiting time to infection $T$, as defined in Section 4.1. The hazard function of $T$ or the infection rate for a homogeneous population is given by (4-1). Host (intrinsic) factors and household environmental factors may cause heterogeneity among susceptibles. For simplicity, factors that cause heterogeneity among susceptibles are also referred to as covariates. The hazard function for a heterogeneous population is defined as

$$h_T(t | z(t), i(t)) = \lim_{\Delta t \to 0} \frac{P[T \leq t + \Delta t | T > t, z(t), i(t)]}{\Delta t} \quad (5-2)$$

where $z(t)' = (z_i(t), z_2(t), \ldots, z_p(t))$ is a covariate vector at time $t$. If all covariates are time independent, then $z(t)$ simplifies to $\bar{z}$.

The hazard function of the time to infection of susceptible $j$ or the infection rate of a susceptible $j$ is given as

$$h_{\xi_j}(t) = \phi_j(t) \xi_j(t) \alpha(t) \xi^n(t) \quad (5-3)$$
where infectiousness is assumed to be homogeneous among all infectives and the contact rate is also assumed to be homogenous among population. Suppose susceptibles with the same covariates specified by a particular value of $z_j$ have the same susceptibility, similar to the Cox model as given by (5-1), the term $\phi_j(t)$ is written as a function of covariates.

$$
\phi_j(t \mid z) = \phi_0(t) \exp(b'z_j(t)) = \phi_0(t) \sum_{i=1}^{p} \exp(b_i z_i(t)) \quad (5-4)
$$

where $b'$ and $z_j(t)$ are defined similarly as in (5-1). $\phi_0(t)$ is the baseline susceptibility for a homogeneous population with $z(t) = 0$, and $\exp(b'z_j(t))$ is the additional risk related to the covariates at time $t$ for susceptible $j$.

Using (5-4), the hazard function of $T$ for a heterogeneous susceptible population is given as

$$
h_T(t \mid z(t), i(t)) = \phi_0(t) \xi(t) \alpha(t) i^m(t) \exp(b'z(t)) \quad (5-5)
$$

If $h_0(i(t), t)$ denotes the baseline hazard, which is the infection rate for an individual with $z = 0$ at time $t$, then (5-5) is simplified as

$$
h_T(t \mid z(t), i(t)) = h_0(i(t), t) \exp(b'z(t)) \quad (5-6)
$$

where

$$
h_0(i(t), t) = \phi_0(t) \xi(t) \alpha(t) i^m(t) = g_0(t) i^m(t) \quad (5-7)
$$

The parameter $m$ in the hazard function accommodates two types of diseases. For $m = 0$, the hazard function represents the occurrence rate for non-infectious diseases (the Cox model in survival analysis) and infectious diseases under the Greenwood assumption. For $m > 0$, it represents the infection rate for person-to-person transmitted infectious diseases. Thus, the Cox model’s baseline function is generalized. The next section aims at introducing an estimate of the baseline hazard function, which encompasses both situations. This generalization provides a semi-parametric method to study the effects of the covariates on the infection rate or furthermore, on the transmission rate.

Some concepts regarding this model are worth clarifying. The end of the epidemic is often indicated when the number of infective and infected both become zero. For susceptibles
who escaped infection, the time to infection $T$ is infinite. These observations are not censored as often seen in survival analysis.

A closed population was assumed in all models discussed in earlier chapters. However, in this model, this assumption is not required because the partial likelihood function is a product of the risk ratios at each infection time. The risk ratio is defined as the risk of the individual becoming infected over the total risk of all susceptibles becoming infected. If there are deaths or births at time $t$, the total risk will be adjusted by adding the risk of births or subtracting the risk of deaths from that point onward without changing it before $t$.

5.3 Generalized Baseline Function for Cox Model

A Cox model application relies on the estimation of the baseline hazard function. In this section, a general baseline hazard estimate is introduced to generalize the Cox model applications to study a larger collection of diseases.

If the number of adequate contacts is $m$, the hazard function of $T$ for a heterogeneous population is given by (5-5). The likelihood function of $T$ can be constructed and used for statistical inferences. In theory, because time is continuous, it is reasonable to assume that two infections cannot occur at the same time. However, in practice, observations are often grouped into coarser units, such as days, and multiple infections in one day are very common.

Infection times observed during an epidemic are ranked into distinctive infection times as \( \{t_{(1)}, t_{(2)}, \ldots, t_{(d)}\} \), where $d$ is the total number of distinct infection times at the end of the epidemic. Let $S_{(j)}$ denote the class of susceptibles just before each infection transmission time $t_{(j)}$ and $S_{(d+1)}$ denote the class of susceptibles who eventually escaped infection over the course of the epidemic. Let $r_{j}$ denote the total number of individuals infected at time $t_{(j)}$ and these individuals have covariates $z_{(j)}, z_{(j+1)}, \ldots, z_{(r_{j})}$. Let $E_{(j)} = \sum_{t \leq t_{(j)}} e_{(t)}$ and $\{e_{(1)}, e_{(2)}, \ldots, e_{(d)}\}$ denote the sum of covariates at time $\{t_{(1)}, t_{(2)} \ldots, t_{(d)}\}$. The full likelihood in terms of both $g_{a}(t)$ and $b$ is given as
The full likelihood function involving $g_0(t)$ and $b$ is difficult to evaluate, so the partial likelihood of $b$ is used. If $g_0(t)$ is assumed to be constant between two consecutive infection times, then it becomes a step function (Breslow, 1974)

$$g_0(t) = g, \quad (5-9)$$

for $t_{(j-1)} < t \leq t_{(j)}$, where $t_{(d+1)} = \tau$ and $j = 1, 2, \ldots, d, d+1$. Then the baseline infection rate becomes $h_0(i(t), t) = g_j i^m(t)$. Under this assumption, the maximum likelihood estimate of $g$, within each consecutive interval is derived from (5-8) as

$$\hat{g}_j = \frac{r_j}{\left( \int_{t_{(j-1)}}^{t_{(j)}} i^m(u) du \sum_{\in S_{(j)}} \exp(b'z_i) \right)} \quad (5-10)$$

for $j = 1, 2, \ldots, d$. and for $t_{(d)} < t \leq \tau$, $g_0(t)$ is estimated as

$$\hat{g}_{d+1} = \frac{1}{\left( \int_{t_{(d)}}^{\tau} i^m(u) du \sum_{\in S_{(d+1)}} \exp(b'z_i) \right)} \quad (5-11)$$

Note that (5-11) is not the maximum likelihood estimate. Using (5-8), it is impossible to obtain the maximum likelihood estimate of $g_{d+1}$.

The integral factor in the denominator of (5-10) generalizes the baseline estimates given in Breslow (1974) for the Cox model as follows.

(1) For $m = 0$, equation (5-10) is the same as the Breslow's baseline estimate.

(2) For $m > 0$, (5-10) gives a weighted adjustment of Breslow's baseline estimate. Indeed, it incorporates the weights as an integral of $i^m(t)$ in that interval.
In theory, the expected number of infectives \( i(t) \) is a continuous function of time. For estimation, the observed number of infectives is used, which is a step function.

Substituting the estimates (5-10) and (5-11) into the full likelihood function (5-8) yields

\[
L(g_0(t), b \mid z, i(t)) = \prod_{j=1}^d \left( \frac{r_j i^m(t_{(j)})}{\int_{t_{j-1}}^{t_j} i^m(u)du} \right)^{r_j} e^{-r_j} \frac{\exp(b^t z_{(j)})}{\sum_{i \in S_{(j)}} \exp(b^t z_i)}
\]

(5-12)

Note that the first part of the product does not include covariates, it will be part of the score functions (the derivative of (5-12) with respect to each \( b \)). Thus, the partial likelihood of \( b \) is given as

\[
PL(b) = \prod_{j=1}^d \frac{\exp(b^t z_{(j)})}{\sum_{i \in S_{(j)}} \exp(b^t z_i)}
\]

(5-13)

which is a product of the risk ratios at infection times. The numerator of the ratio is the risk of those individuals who are infected at time \( t_{(j)} \), while the denominator is the total risk of all the susceptibles in \( S_{(j)} \). The score functions and the information are similar to those described in Cox and Oakes (1984, Section 7.4), which will not be repeated here.

If some covariates of interest are time-dependent, the covariates used in (5-8) is replaced by \( z(t) \). Similarly, assuming a constant baseline transmission rate, the maximum likelihood estimate baseline hazard in \( t_{(j-1)} \leq t \leq t_{(j)} \) is given as

\[
g_j = \frac{r_j}{\left( \int_{t_{j-1}}^{t_j} i^m(u) \sum_{i \in S_{(j)}} \exp(b^t z_i(u))du \right)} = \frac{r_j}{\left( \int_{t_{j-1}}^{t_j} \sum_{i \in S_{(j)}} \exp(b^t z_i(u) + m \log i(u))du \right)}
\]

(5-14)

In the last interval, the baseline transmission rate is assumed to be

\[
\hat{g}_{d+1} = \frac{1}{\left( \int_{t_{d+1}}^{T} i^m(u) \sum_{i \in S_{d+1}} \exp(b^t z_i(u))du \right)}
\]

(5-15)
Note that (5-15) is not the maximum likelihood estimate of \( g_{d+1} \). Again, additional information regarding the infectiosity of the pathogen and the number of infectives is incorporated into Breslow’s baseline estimate as an extra dimension. With these baseline estimates, the partial likelihood function is given as

\[
PL(\hat{b}) = \prod_{j=1}^{d} \frac{\exp(\hat{b}' \varepsilon_{i(j)},(t_{j}))}{\left[ \sum_{i \in S_{t_{j}}} \exp(\hat{b}' \varepsilon_{i}(t_{j})) \right]^r} \tag{5-16}
\]

The score function of each covariate and the information matrix are similar to those given in Section 8.3 in Cox and Oaks (1984).

Because the infectiosity of the pathogen, the infectiousness of the population, and the number of infectives in the population are baseline characteristics, and the partial likelihood function of covariates does not involve any baseline characteristics, a similar partial likelihood is derived for various model assumptions. Various statistical software packages can perform the Cox model analysis. A SAS procedure, "PHREG", can perform statistical inferences for the Cox models with time-independent and time-dependent covariates. With the estimated baseline given in (5-10) and (5-14), covariate effects are easily estimated for infectious disease data by using this procedure. This will be shown in Section 5.5. However, the baseline transmission rate cannot be estimated directly using this procedure.

### 5.4 Semi-Parametric Estimate of the Transmission Rate

The baseline hazard function \( g_o(t) = \phi_o(t) \xi(t) \alpha(t) \) is the baseline transmission rate, and the maximum likelihood estimates of the piecewise baseline hazards are given in (5-10) and (5-14).

Substituting the estimated covariate effects \( \hat{b} \) and the observed number of infectives \( i(t) \) yields the estimated transmission rate in each interval adjusted by covariates \( \xi^* \)

\[
\hat{\beta}_j = \frac{r_j \exp(\hat{b}' \xi^*)}{\left[ \int_{t_{j-1}}^{t_j} \alpha(u) du \right] \left[ \sum_{i \in S_{t_j}} \exp(\hat{b}' \xi_i) \right]} \tag{5-17}
\]

This is a semi-parametric estimate of the transmission rate adjusted for covariates, which is a step function.
In Section 3.6, the transmission rate for each age group is estimated using the martingale method. Let \( z_1 \) denote the indicator variable for infants, \( z_2 \) denote the indicator variable for schoolchildren, and \( \hat{b}_1 \) and \( \hat{b}_2 \) be the estimated covariate effects respectively. The transmission rate for adults is given as \( \hat{\beta}_j = \frac{r_j}{\left[ \int_{t_{j,n}}^{t_{j,n+1}} \exp(\hat{b}_1) \right] \sum S_i, \exp(\hat{b}_1)} \), which is the baseline transmission rate. The transmission rate for infants is given as \( \hat{g}_j \exp(\hat{b}_1) \), and the transmission rate for schoolchildren is given as \( \hat{g}_j \exp(\hat{b}_2) \).

If there are time-dependent covariates, the transmission rate adjusted for covariates \( z^* \) is given as

\[
\hat{\beta}_j = \frac{r_j \exp(\hat{b}_2 z^*)}{\left[ \int_{t_{j,n}}^{t_{j,n+1}} \sum S_i, \exp(\hat{b}_2 z^*) \right] du} \quad (5-18)
\]

Note that in this case, the estimated baseline transmission rate is more complicated and additional programming is required to make such estimation.

### 5.5 Application - Respiratory Disease Data

The same respiratory disease epidemic data that has been studied in Chapters 3 and 4 is studied here using the Cox model. From the data presented in Becker (1989), covariates such as age and the household infection information are identified. The rearranged data have 254 observations (individuals) not including the introductory case, see Appendix B. The following variables are created:

1. household id number,
2. time of infection in days,
3. infected: 1 if individual was infected during this epidemic, 0 otherwise,
4. infants: 1 for infants, 0 otherwise,
5. schoolchildren: 1 for schoolchildren, 0 otherwise,
6. time of recovery in days,
If both variables (4) and (5) equal to 0, the individual is an adult. The age factor in an epidemic with short duration is treated as a time-independent covariate.

5.5.1 Without Time-Dependent Covariates

The Cox model provides good fit to data if the hazard functions for different values of a covariate considered in the model are proportional. For example, if age is a time-independent covariate in the model, then the plots of the logarithm of the cumulative hazard functions against the logarithm of time for infants, schoolchildren and adults should be parallel. The factor age satisfies the proportional hazards assumption from the plot. So, the proportional hazards model may be used to study these covariates, infants and schoolchildren.

The fitted models for age is summarized in Table 5-1. The relative risk is also calculated in Section 3.6.2. The relative risk of infants to adults is 5.2 in the non-parametric method and it is predicted as 4.1 (Model 1) using this semi-parametric model. The relative risk of schoolchildren to adults is 2.0 using the non-parametric method and it is predicted as 1.4 (Model 1) using this semi-parametric model. In both methods, the risks of infection for schoolchildren and adults are not significantly different.

Using the method introduced in Section 5.4, the baseline transmission rates are estimated and summarized in Table 5-2. Columns (1) to (3) are taken from the data given in Appendix A. Note that the risk set for different age categories before each infection time are given by columns (4), (5) and (6), which are used to estimate the total risk, that is, the denominator of the formula given in column (7). Column (8) is the total time in the infective state by all infective individuals during the interval, which is calculated using data given in Appendix A. Column (9) is the estimated baseline transmission rate, which is also the estimated transmission rate for adults. Table 5-3 shows the transmission rate for each age category.
### Table 5-1 Model estimations – without time-dependent covariates

<table>
<thead>
<tr>
<th>Models</th>
<th>Covariates</th>
<th>Estimated $\hat{b}$</th>
<th>s.e $\hat{b}$</th>
<th>P-Value</th>
<th>Relative Risk ($=e^{\hat{b}}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Infant</td>
<td>1.4173</td>
<td>0.3871</td>
<td>0.003</td>
<td>4.126</td>
</tr>
<tr>
<td></td>
<td>School children</td>
<td>0.3000</td>
<td>0.4629</td>
<td>0.5170</td>
<td>1.350</td>
</tr>
</tbody>
</table>

### Table 5-2 Baseline transmission rate estimation

<table>
<thead>
<tr>
<th>$t_{(j-1)}$</th>
<th>$t_{(j)}$</th>
<th>$r_j$</th>
<th>Infants</th>
<th>School-children</th>
<th>Adults</th>
<th>$\sum_{i=1}^{n} \exp(\hat{b} X_i)$</th>
<th>$\int_{t_{(j-1)}}^{t_{(j)}} i^m(u)du$</th>
<th>$\hat{g}_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
<td>(8)</td>
<td>(9)</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>1</td>
<td>27</td>
<td>40</td>
<td>187</td>
<td>0.002838</td>
<td>6</td>
<td>0.000473</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>2</td>
<td>27</td>
<td>40</td>
<td>186</td>
<td>0.005692</td>
<td>1</td>
<td>0.005692</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>4</td>
<td>27</td>
<td>39</td>
<td>185</td>
<td>0.011460</td>
<td>1</td>
<td>0.011460</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>4</td>
<td>24</td>
<td>38</td>
<td>185</td>
<td>0.011929</td>
<td>1</td>
<td>0.011929</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>6</td>
<td>22</td>
<td>37</td>
<td>184</td>
<td>0.018478</td>
<td>3</td>
<td>0.006159</td>
</tr>
<tr>
<td>11</td>
<td>13</td>
<td>3</td>
<td>19</td>
<td>37</td>
<td>181</td>
<td>0.009698</td>
<td>18</td>
<td>0.000539</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>5</td>
<td>17</td>
<td>36</td>
<td>181</td>
<td>0.016681</td>
<td>12</td>
<td>0.001390</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>1</td>
<td>17</td>
<td>35</td>
<td>177</td>
<td>0.003397</td>
<td>11</td>
<td>0.000309</td>
</tr>
<tr>
<td>15</td>
<td>16</td>
<td>2</td>
<td>17</td>
<td>35</td>
<td>176</td>
<td>0.006817</td>
<td>12</td>
<td>0.000568</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>3</td>
<td>17</td>
<td>34</td>
<td>175</td>
<td>0.010308</td>
<td>15</td>
<td>0.000687</td>
</tr>
<tr>
<td>17</td>
<td>18</td>
<td>2</td>
<td>17</td>
<td>34</td>
<td>172</td>
<td>0.006944</td>
<td>14</td>
<td>0.000496</td>
</tr>
<tr>
<td>18</td>
<td>19</td>
<td>1</td>
<td>17</td>
<td>34</td>
<td>170</td>
<td>0.003496</td>
<td>16</td>
<td>0.000219</td>
</tr>
<tr>
<td>19</td>
<td>20</td>
<td>2</td>
<td>17</td>
<td>34</td>
<td>169</td>
<td>0.007017</td>
<td>16</td>
<td>0.000439</td>
</tr>
<tr>
<td>20</td>
<td>27</td>
<td>1</td>
<td>17</td>
<td>34</td>
<td>167</td>
<td>0.003533</td>
<td>46</td>
<td>0.000077</td>
</tr>
<tr>
<td>27</td>
<td>28</td>
<td>1</td>
<td>17</td>
<td>34</td>
<td>166</td>
<td>0.003546</td>
<td>1</td>
<td>0.003546</td>
</tr>
<tr>
<td>28</td>
<td>33</td>
<td>0</td>
<td>17</td>
<td>34</td>
<td>165</td>
<td>0.003558</td>
<td>10</td>
<td>0.000356</td>
</tr>
</tbody>
</table>
Table 5-3 Estimated transmission rates by age group

<table>
<thead>
<tr>
<th>$t_{(j-1)}$</th>
<th>$t_{(j)}$</th>
<th>Adults</th>
<th>School-children</th>
<th>Infants</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>0.000473</td>
<td>0.001951</td>
<td>0.000638</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>0.005692</td>
<td>0.023483</td>
<td>0.007683</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0.011460</td>
<td>0.047283</td>
<td>0.015469</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>0.011929</td>
<td>0.049218</td>
<td>0.016102</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>0.006159</td>
<td>0.025413</td>
<td>0.008314</td>
</tr>
<tr>
<td>11</td>
<td>13</td>
<td>0.000539</td>
<td>0.002223</td>
<td>0.000727</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>0.001390</td>
<td>0.005736</td>
<td>0.001876</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>0.000309</td>
<td>0.001274</td>
<td>0.000417</td>
</tr>
<tr>
<td>15</td>
<td>16</td>
<td>0.000568</td>
<td>0.002344</td>
<td>0.000767</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>0.000687</td>
<td>0.002835</td>
<td>0.000928</td>
</tr>
<tr>
<td>17</td>
<td>18</td>
<td>0.000496</td>
<td>0.002046</td>
<td>0.000669</td>
</tr>
<tr>
<td>18</td>
<td>19</td>
<td>0.000219</td>
<td>0.000902</td>
<td>0.000295</td>
</tr>
<tr>
<td>19</td>
<td>20</td>
<td>0.000439</td>
<td>0.001809</td>
<td>0.000592</td>
</tr>
<tr>
<td>20</td>
<td>27</td>
<td>0.000077</td>
<td>0.000317</td>
<td>0.000104</td>
</tr>
<tr>
<td>27</td>
<td>28</td>
<td>0.003546</td>
<td>0.014629</td>
<td>0.004786</td>
</tr>
<tr>
<td>28</td>
<td>33</td>
<td>0.000356</td>
<td>0.001468</td>
<td>0.000480</td>
</tr>
</tbody>
</table>

5.5.2 With Time-Dependent Covariates

From the plots shown in Section 3.5.2 and Section 3.6.1, the transmission rate varies over time; this observation may suggest the existence of time-dependent covariates. Household environmental factors vary over time, and should be treated as time-dependent covariates. For each household, the time of the appearance of the first infective and the recovery of the last infective are created from the data given in Becker (1989). Because new infections may occur within an infected household, the number of infectives varies at different time points. These time points and the corresponding number of infectives within the households (number of infective members) are also created, which are given in Appendix B. The household infected (1 if the household currently has more than one infectives at a given time, 0 otherwise) and the household infected
duration (the duration from the appearance of the first infection to a given time) are additional time-dependent covariates. These two variables are created within the SAS "PHREG" procedure.

Various models by fitting different covariates are summarized in Table 5-4. The effect of schoolchildren is insignificant (see Model 1 in Table 5-1 and Model 4 in Table 5-4), so schoolchildren and adults are grouped into one category. Model 5 gives the best fit by the stepwise selection – a build in option of the "PHREG". From these fitted models, the following conclusions may be drawn.

1. Infants have a higher chance of getting infected than do schoolchildren and adults (the relative risk of infants to adults is greater than 4 in all models).
2. Schoolchildren and adults have similar risk in getting infection (the relative risk of schoolchildren to adults is about 1 in Model 1 and 4).
3. Susceptibles have a higher chance of being infected if the household has been infected for a longer period (the relative risk of every additional day is greater than 1 in Model 2, 4, 5 and 7).
4. The factor household infected reduces the infection rate when adjusting for the household infection duration (see Model 5). However, this effect is reversed when the household infection duration is not included in the model (Model 3 and 6).
5. Susceptibles have a higher chance of being infected if there are more infective members in the household (see Model 8) if the model does not include other household infection factors. However, this effect becomes insignificant when adjusting for other household environment factors.
6. The infant factor, the household infected duration and the household infected have significant effects on the transmission rate using the stepwise selection (Model 5).

In Model 6, the relative risks for the household infected is 7.5, but it has a negative effect, 0.32, when adjusting for the household infected duration in Model 5. This is because the three household factors carry similar information. For the same reason, when the household infected or the household infected duration is included in the model, the number of infective members is insignificant. In this epidemic, household environmental factors are important aspects to study the infection rate and the infection process (see Model 5). However, these factors are time-dependent, which can not be estimated using the stratified analysis, and so, the Cox model needs to be used to understand the effect of household factors on the infection transmission.
<table>
<thead>
<tr>
<th>Models</th>
<th>Covariates</th>
<th>Estimated $\hat{b}$</th>
<th>s.e $\hat{b}$</th>
<th>P-Value</th>
<th>Relative Risk ($=e^{\hat{b}}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Infant</td>
<td>1.5768</td>
<td>0.3888</td>
<td>0.0001</td>
<td>4.840</td>
</tr>
<tr>
<td></td>
<td>Household infection duration</td>
<td>0.2734</td>
<td>0.1144</td>
<td>0.0169</td>
<td>1.314</td>
</tr>
<tr>
<td></td>
<td>Number of infective members</td>
<td>0.3405</td>
<td>0.2681</td>
<td>0.2041</td>
<td>1.406</td>
</tr>
<tr>
<td>3</td>
<td>Infant</td>
<td>1.5050</td>
<td>0.3811</td>
<td>0.0001</td>
<td>4.504</td>
</tr>
<tr>
<td></td>
<td>Household infected</td>
<td>1.2833</td>
<td>0.5396</td>
<td>0.0174</td>
<td>3.609</td>
</tr>
<tr>
<td></td>
<td>Number of infective members</td>
<td>0.5256</td>
<td>0.2602</td>
<td>0.0434</td>
<td>1.691</td>
</tr>
<tr>
<td>4</td>
<td>Infant</td>
<td>1.4961</td>
<td>0.4080</td>
<td>0.0002</td>
<td>4.464</td>
</tr>
<tr>
<td></td>
<td>Schoolchildren</td>
<td>0.4818</td>
<td>0.4754</td>
<td>0.3108</td>
<td>1.619</td>
</tr>
<tr>
<td></td>
<td>Household infected</td>
<td>-1.2844</td>
<td>0.5743</td>
<td>0.0253</td>
<td>0.277</td>
</tr>
<tr>
<td></td>
<td>Household infected duration</td>
<td>0.3629</td>
<td>0.1341</td>
<td>0.0068</td>
<td>1.438</td>
</tr>
<tr>
<td></td>
<td>Number of infective members</td>
<td>0.5145</td>
<td>0.2955</td>
<td>0.09816</td>
<td>1.673</td>
</tr>
<tr>
<td>5</td>
<td>Infant</td>
<td>1.4147</td>
<td>0.3918</td>
<td>0.0003</td>
<td>4.115</td>
</tr>
<tr>
<td></td>
<td>Household infected duration</td>
<td>0.5071</td>
<td>0.0950</td>
<td>0.0001</td>
<td>1.660</td>
</tr>
<tr>
<td></td>
<td>Household infected</td>
<td>-1.1225</td>
<td>0.5628</td>
<td>0.0461</td>
<td>0.325</td>
</tr>
<tr>
<td>6</td>
<td>Infant</td>
<td>1.4617</td>
<td>0.3785</td>
<td>0.0001</td>
<td>4.313</td>
</tr>
<tr>
<td></td>
<td>Household infected</td>
<td>2.0165</td>
<td>0.3511</td>
<td>U.WUI</td>
<td>7.512</td>
</tr>
<tr>
<td>7</td>
<td>Infant</td>
<td>1.5284</td>
<td>0.3813</td>
<td>0.0001</td>
<td>4.611</td>
</tr>
<tr>
<td></td>
<td>Household infected duration</td>
<td>0.3810</td>
<td>0.0648</td>
<td>0.0001</td>
<td>1.464</td>
</tr>
<tr>
<td>8</td>
<td>Infant</td>
<td>1.5133</td>
<td>0.3799</td>
<td>0.0001</td>
<td>4.542</td>
</tr>
<tr>
<td></td>
<td>Number of infective members</td>
<td>0.9511</td>
<td>0.1483</td>
<td>0.0001</td>
<td>2.589</td>
</tr>
</tbody>
</table>
6 Conclusions and Further Suggestions

6.1 Conclusions

The contribution this work makes to mathematical epidemic modelling is the multiplicative model of the transmission rate. By using this model, some existing statistical methods that have not been discussed in the mathematical epidemic modelling are brought in, and the existing statistical methods that have been used in other contexts are extended to some general situations. Historical data from two person-to-person transmitted diseases (by airborne transmission) are analyzed using the proposed model. This analysis provides additional information about the two epidemics. However, the multiplicative model and these statistical methods are not exactly intended for these two particular diseases, but may be used directly or modified to study diseases transmitted by close person-to-person contact in different susceptible populations.

Agent factors, host (intrinsic) factors, and environmental (extrinsic) conditions are determinants of an epidemic in general, but environmental factors, which affect the survival of the pathogen, are less critical in epidemics caused by close person-to-person transmitted diseases. Estimating the effect of each determinant is important in understanding the complicated mechanism that underlies the spread of a disease. The multiplicative model of the transmission rate decomposes the effect of hosts into susceptibility and contact pattern of population, and the effect of the pathogen as infectiousness and infectiosity. As it is based on the pathogenesis of an infection, the multiplicative model models the infection process more closely and forms a basis for studying an epidemic and its essential factors.

Rhodes et. al. (1996) uses the thinned version of the contact process to study the infection process. This approach is generalized by assuming $m$ adequate contacts. The per “contact” transmission probability is modelled as the product of two conditional probabilities, which represent the susceptibility and the infectiousness respectively. Instead of studying the maximum likelihood estimates, the incidence processes are derived for statistical inferences.

Using a probability approach and making various model assumptions, different incidence processes are derived using this multiplicative model of the transmission rate. The incidence process and the number of observed infections occurring during an epidemic are used to estimate
the transmission rate and the covariate effects on the transmission rate. Thus, it is important to derive an incidence processes that are biologically meaningful and computationally manageable for a particular epidemic. Various functional forms of the incidence have been assumed in the literature, and the incidence processes derived using this model have verified or generalized some of the existing ones.

Infectious diseases have complicated modes of transmissions that are affected by the interaction mechanisms between hosts and agents. Because it is unethical to conduct experiments on people, data are only available from naturally occurring epidemics. Simple assumptions have to be made so that estimates may be obtained and certain conclusions may be drawn, but such estimates could be misleading and have less practical meaning when assumptions are too far away from reality. Therefore, it is important to test the precision of the estimates and the realism of the model assumptions used in deriving such estimates. Consequently, statistical methods become important in achieving this because they work directly from the model formulation, provide meaningful estimates, and allow hypothesis testing. The martingale method, the life table method, and the Cox model were used to perform inferences, in which model checking were also carried out.

Becker (1989) used the martingale method to estimate the transmission rate for the bilinear incidence in SIR models. This dissertation extended the martingale application from three aspects. Firstly, the martingale method is generalised to make inference for SEIR models by choosing appropriate \( \tau \) (the end of an epidemic) and a \( F_t \)-predictable process. If the number of infectives never reaches zero before the end of an epidemic, the same \( F_t \)-predictable process used in the SIR model was used in the SEIR model to construct the martingale transform (as seen in the smallpox epidemic data). If the number of infectives reaches zero before the end of an epidemic, a different \( F_t \)-predictable process was used (as seen in the respiratory disease data).

Secondly, the martingale method estimates the transmission rate using a general incidence process \( \beta(t)S(t)I^n(t) \) derived from the multiplicative model. The estimated transmission rate, \( m \) and other model assumptions were assessed graphically by comparing the observed and predicted numbers of infections. In the smallpox epidemic data, the transmission rates estimated under \( m = 0.8 \) and \( m = 1 \) (the bilinear incidence) both provided adequate fit to the data. The number of infections predicted under the assumption of \( m = 0.8 \) is closer to the
observed number of infections than the predictions under $m = 1$. Keeping the rest of the model assumptions, the transmission rate is also estimated for $m = 1.2, 1.5, 1.8, 2, 2.5$. When $m$ is greater than 1.5, the predicted number of infections becomes unrealistically large, which indicates a poorer fit of the data. These results demonstrate the importance of considering different degrees of infectiosity in studying the spread of a disease. Therefore, the infectiosity is indeed an important parameter in modelling epidemic data.

Thirdly, the martingale method is used to estimate and test heterogeneity within susceptibles. In the respiratory disease data, the overall transmission rate within a stratum specified by a host factor (age), was estimated using the martingale method. Testing whether the relative risk of one group to another equal to 1 assesses the significance of the effect of age on the transmission rate. Age was shown to be an important determinant of the spread of the disease. Population homogeneity that is generally assumed in the literature failed in this particular case. This indicates the necessity of examining basic model assumptions.

This multiplicative model includes many factors (covariates) that make a population heterogeneous. By assuming a particular functional form of the association of covariates on the transmission rate, this multiplicative model falls into the framework of the Cox model. The effects of covariates were estimated simultaneously using the software procedure developed for the Cox model. The significance of each covariate effect was tested using the Wald’s statistic. As seen in the stratified martingale method, age has a significant effect on the transmission rate along with some other time-dependent covariates. The estimates of baseline the transmission rate is a generalization of Breslow (1974) and the transmission rate adjusting for other covariates is also estimated.

The life table method from demography was brought to epidemic modelling, and offered many meaningful functions for studying an epidemic. The Lorenz curve and the STTT provide certain information on the distribution of the waiting time to infection. An infection table provides a detailed description of how a disease progresses in a population and it is useful in assisting health officials to decide the right control strategies. Another important use of the life table functions is to predict the number of infections over time, so that the goodness-of-fit can be examined graphically.
6.2 Limitations and Further Suggestions

The availability and the accuracy of data limit the application of the multiplicative model and statistical methods - the martingale method, the life table method and the Cox model - used in this work. The most serious limitation is that the infectious disease data generally consist of times of first appearance of symptoms. However, the data required for statistical modelling are the times of infection. It is reasonable to assume that the incubation period is the sum of the latent and infectious periods for some diseases, such as smallpox. This assumption would not hold true for most infectious diseases, as it is reflected in the results of goodness-of-fit in the two data sets studied in this dissertation. The model does not yield good fit to the respiratory disease epidemic data, which indicates certain problems with the model assumptions, particularly the assumption regarding the relation between incubation and infectious periods.

The multiplicative model takes into account the variability of infectiousness. However, the infective that is responsible for the infection is not observable. Homogeneity of infectives is often presumed to obtain results regarding the effects of susceptible host factors on the transmission rate. It was used to obtain estimates of covariate effects in the Cox model, and it was also used to conclude that the difference in the overall transmission rate is due to the difference in susceptibility in the stratified martingale analysis. As this assumption may not hold true in general, biases may be introduced to the conclusions. One improvement is to replace the infectiousness term in the multiplicative model by an average infectiousness of the infectives. The average infectiousness can be treated as a function of the infectious period, as the amount of pathogen within an infective during the infectious period is different. For many diseases, the infectiousness of an infective increases at the beginning of the infectious period, and then decreases. Thus, a unimodal distribution, such as a Gamma distribution, may be appropriate for describing the distribution of the infectiousness. If the infective individual who is responsible for the infection is observable, then the individual infectious period should be calculated for use.

The multiplicative model was applied to two airborne transmitted diseases, smallpox and the respiratory disease, where the assumption of randomly mixed populations seems reasonable because contacts have general meanings. If this model were used to study sexually transmitted diseases, then the assumption of an equal chance of contacting any infectives in the population would have to be altered which may result in a complicated incidence process. However, because
the mode of transmission and the infective responsible for the infection in sexually transmitted
diseases are easier to identify than in airborne transmission, the multiplicative model may be
useful in identifying covariates associated with a high transmission rate due to infectives. Note
that this may not hold true for AIDS, acquired immune-deficiency-syndrome, because the long
incubation period.

Contact form determines the amount of pathogen that is released to a susceptible. But, because of lack of available data, it was not incorporated into the multiplicative model. It is possible to introduce another conditional probability to signify forms of contacts in the multiplicative model. However, this would complicate the choice of the number of adequate contacts for a given pathogen, that is, the parameter $m$ would be treated as a function of contact forms. As a result, the incidence processes would become too complicated and thus require inference procedures other than the martingale method and the Cox model.

The limitation with the martingale method is that it cannot be used to estimate the adequate number of contacts. In this work, the generalized linear model was used to estimate $m$, and this estimate was then used to obtain the martingale estimate of the transmission rate $\beta(t)$. It is desirable to search for a statistical method or model that can estimate the number of adequate contacts and the transmission rate simultaneously.

When using the martingale method, the model checking is done graphically. The test statistic for comparing the expected and observed number of infections follows a non-central Chi-square. However, the non-centrality (the degree of freedom) of this Chi-square is difficult to derive in this case, and thus, it is lacking for rigorous tests.

A number of suggestions will be described for further studies. The transmission rate is estimated for a fixed number of adequate contacts (which reflect the infectiosity of a pathogen) in the martingale method. It may be reasonable to question if the infectiosity of a pathogen varies over time $m(t)$, precisely, over the course of the epidemic. The martingale method can easily incorporate a random infectiosity by including the infectiosity of the pathogen or the number of adequate contacts into the filtration ($F_t$). For example, suppose the adequate contacts is a step function of time: for $0 = a_0 \leq t < a_1$, $m(t) = m_1$, ..., for $a_{i-2} \leq t < a_{i-1}$, $m(t) = m_{i-1}$, and, for $a_{i-1} \leq t \leq a_i = \tau$, $m(t) = m_i$, where $l$ is the number of intervals that the entire epidemic is divided into. The filtration $F_t$ is given as $F_t = \sigma\{ I_i(u), S_j(u), C_k(u), M(u), 0 \leq u < t\}$, where $M(t)$ is the
sample path for the infectiosity of the pathogen. The incidence process is $\beta(t)S(t)I^m(t)$ for $a_{k-1} \leq t < a_k$, where $k = 1, \ldots, l$. GLM can be used to estimate $m_k$ within each interval, and furthermore, the transmission rate can then be estimated using the counting process and the corresponding incidence process. This suggestion provides one possible solution to the limitation with the martingale method discussed earlier.

The two data sets that are considered in this work do not contain information on the vaccination status of the subjects, so vaccine efficacy is not studied. Vaccination status can be incorporated as a covariate in the Cox model and its effect may be estimated. The risk ratio associated with this covariate provides an estimate of the vaccine efficacy. Vaccine efficacy can also be estimated using the stratified martingale method because it is signified by the relative risk of the vaccinated and the unvaccinated groups.

Many serious infectious diseases appear to be under control in developed countries. However, this is not the case for diseases such as tuberculosis and AIDS. Infectious diseases are by no means eliminated or well under control (with the exception of some childhood diseases such as smallpox). International travel brings more susceptibles into areas where diseases are endemic, and conversely; thus facilitating further spread of infectious diseases. The ageing of the population in developed countries raises the size of susceptible population that sustains epidemics or endemic diseases. The next important step is to employ the work done in this dissertation to develop control strategies for infectious diseases.
References


Wilson, E. B. and Worcester, J (1945 b): The law of mass action in epidemiology II. Proc. N.A.S.


Appendices

A. Population classification - respiratory disease epidemic data

<table>
<thead>
<tr>
<th>$t$</th>
<th>Susceptibles $S(t)$</th>
<th>Latent Individuals $E(t)$</th>
<th>Infectives $I(t)$</th>
<th>Recoveries $R(t)$</th>
<th>Infection Transmissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>254</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>254</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>254</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>254</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>254</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>254</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>254</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>253</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>253</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>251</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>247</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>243</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>237</td>
<td>6</td>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>237</td>
<td>0</td>
<td>12</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>234</td>
<td>3</td>
<td>11</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>229</td>
<td>5</td>
<td>12</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>228</td>
<td>1</td>
<td>15</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>226</td>
<td>2</td>
<td>14</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>223</td>
<td>3</td>
<td>16</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>221</td>
<td>2</td>
<td>16</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>220</td>
<td>1</td>
<td>17</td>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>218</td>
<td>2</td>
<td>9</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>218</td>
<td>0</td>
<td>6</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>218</td>
<td>0</td>
<td>6</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>218</td>
<td>0</td>
<td>4</td>
<td>33</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>218</td>
<td>0</td>
<td>3</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>218</td>
<td>0</td>
<td>1</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>218</td>
<td>0</td>
<td>1</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>217</td>
<td>1</td>
<td>1</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>216</td>
<td>1</td>
<td>2</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>216</td>
<td>0</td>
<td>2</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>216</td>
<td>0</td>
<td>2</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>32</td>
<td>216</td>
<td>0</td>
<td>2</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>216</td>
<td>0</td>
<td>1</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>34</td>
<td>216</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>0</td>
</tr>
</tbody>
</table>
B. Longitudinal data - respiratory disease epidemic data

Variables:
(1) Household id number
(2) Time of first infection within a household
(3) Time of last recovery within a household
(4) Time of infection of an individual in days
(5) Infected (an individual is infected)
(6) Time of recovery of an individual in days
(7) Infants
(8) Schoolchildren
(9) – (15) Times when the number of infectives within a household changes
(16) –(22) Numbers of infectives within a household at each corresponding time point.

The community consists of 255 individuals of which 104 individuals, not including the primary case, are from the 24 infected households. Data for these individuals are shown in detail. The remaining 150 individuals from the unaffected households have the same observations for each covariate, and thus, only a few of them are shown. Note that it is not necessary to separate the 150 individuals into households.
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 1 | 0 | 6 | 34 | 0 | . | 0 | 1 | 0 | 6 | . | . | . | . | . | . | 1 | 0 | . | . | . | . | . | . |
| 2 | 6 | 16 | 6 | 1 | 12 | 0 | 0 | 0 | . | . | . | . | . | . | . | 0 | . | . | . | . | . | . | . |
| 3 | 8 | 24 | 8 | 1 | 20 | 0 | 1 | 0 | . | . | . | . | . | . | . | 0 | . | . | . | . | . | . | . |
| 4 | 8 | 12 | 8 | 1 | 12 | 0 | 0 | 0 | . | . | . | . | . | . | . | 0 | . | . | . | . | . | . | . |
| 5 | 9 | 12 | 9 | 1 | 12 | 0 | 1 | 0 | . | . | . | . | . | . | . | 0 | . | . | . | . | . | . | . |
| 6 | 9 | 21 | 9 | 1 | 20 | 1 | 0 | 0 | . | . | . | . | . | . | . | 0 | . | . | . | . | . | . | . |
| 7 | 9 | 12 | 9 | 1 | 12 | 1 | 0 | 0 | . | . | . | . | . | . | . | 0 | . | . | . | . | . | . | . |
| 8 | 10 | 19 | 10 | 1 | 19 | 0 | 1 | 0 | . | . | . | . | . | . | . | 0 | . | . | . | . | . | . | . |
| 9 | 10 | 16 | 10 | 1 | 15 | 0 | 0 | 0 | . | . | . | . | . | . | . | 0 | . | . | . | . | . | . | . |
| 10 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | . | . | . | . | . | . | 0 | . | . | . | . | . | . | . |
| 11 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | 13 | 16 | . | . | . | 0 | 1 | 2 | 0 | . | . | . | . | . | . | . |
| 12 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | 13 | 16 | 10 | 19 | . | . | 0 | 1 | 0 | . | . | . | . | . | . | . |
| 13 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | 13 | 16 | 10 | 19 | . | . | 0 | 1 | 0 | . | . | . | . | . | . | . |
| 14 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | 13 | 16 | 10 | 19 | . | . | 0 | 1 | 0 | . | . | . | . | . | . | . |
| 15 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | 13 | 16 | 10 | 19 | . | . | 0 | 1 | 0 | . | . | . | . | . | . | . |
| 16 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | 13 | 16 | 10 | 19 | . | . | 0 | 1 | 0 | . | . | . | . | . | . | . |
| 17 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | 13 | 16 | 10 | 19 | . | . | 0 | 1 | 0 | . | . | . | . | . | . | . |
| 18 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | 13 | 16 | 10 | 19 | . | . | 0 | 1 | 0 | . | . | . | . | . | . | . |
| 19 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | 13 | 16 | 10 | 19 | . | . | 0 | 1 | 0 | . | . | . | . | . | . | . |
| 20 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | 13 | 16 | 10 | 19 | . | . | 0 | 1 | 0 | . | . | . | . | . | . | . |
| 21 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | 13 | 16 | 10 | 19 | . | . | 0 | 1 | 0 | . | . | . | . | . | . | . |
| 22 | 10 | 16 | 10 | 1 | 16 | 1 | 0 | 0 | 10 | 13 | 16 | 10 | 19 | . | . | 0 | 1 | 0 | . | . | . | . | . | . | . |
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 15| 23| 34| 0 |   | 0 | 0 | 0 | 0 | 16| 23|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 16| 25| 16| 1 | 25| 0 | 1 | 0 |   |   |   |   |   |   |   |   | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 16| 25| 34| 0 |   | 0 | 0 | 0 | 0 | 16| 25|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 16| 25| 34| 0 |   | 0 | 0 | 0 | 0 | 16| 25|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 16| 25| 34| 0 |   | 0 | 0 | 0 | 0 | 16| 25|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 17| 25| 17| 1 | 20| 0 | 0 | 0 |   |   |   |   |   |   |   |   | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 17| 25| 17| 1 | 21| 0 | 0 | 0 |   |   |   |   |   |   |   |   | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 17| 25| 20| 1 | 25| 0 | 0 | 0 | 0 | 17| 21|   |   |   | 0 | 2 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 18| 23| 18| 1 | 23| 0 | 0 | 0 |   |   |   |   |   |   |   |   | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 18| 23| 34| 0 |   | 0 | 0 | 0 | 0 | 18| 23|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 18| 23| 34| 0 |   | 0 | 0 | 0 | 0 | 18| 23|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 20| 29| 20| 1 | 29| 0 | 0 | 0 |   |   |   |   |   |   |   |   | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 20| 29| 34| 0 |   | 0 | 0 | 0 | 0 | 20| 29|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 20| 29| 34| 0 |   | 0 | 0 | 0 | 0 | 20| 29|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 27| 32| 27| 1 | 32| 0 | 0 | 0 |   |   |   |   |   |   |   |   | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 27| 32| 34| 0 |   | 0 | 1 | 0 | 27| 32|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 27| 32| 34| 0 |   | 0 | 1 | 0 | 27| 32|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 27| 32| 34| 0 |   | 0 | 0 | 0 | 27| 32|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 27| 32| 34| 0 |   | 0 | 0 | 0 | 27| 32|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 28| 33| 28| 1 | 33| 0 | 0 | 0 |   |   |   |   |   |   |   |   | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 28| 33| 34| 0 |   | 0 | 0 | 0 | 28| 33|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 28| 33| 34| 0 |   | 0 | 0 | 0 | 28| 33|   |   |   | 0 | 1 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

******************************************************************************************** Not Infected Households **********************************************************************************************
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   | 34| 0 |   | 1 | 0 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   | 34| 0 |   | 1 | 0 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   | 34| 0 |   | 1 | 0 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   | 34| 0 |   | 1 | 0 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|   | 34| 0 |   | 1 | 0 | 0 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |