Estimating the Force of Infection from Prevalence Data: Infectious Disease Modelling

by

YUSENTHA BALAKRISHNA

A thesis presented to

THE UNIVERSITY OF KWAZULU-NATAL

in fulfilment of the requirement for the degree of

MASTER OF SCIENCE IN STATISTICS

UNIVERSITY OF KWAZULU-NATAL

PIETERMARITZBURG CAMPUS, SOUTH AFRICA

2013

Disclaimer

This document describes work undertaken as a Masters programme of study at the University of KwaZulu-Natal (UKZN) with the financial assistance of the National Research Foundation (NRF). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of UKZN or the NRF.

Declaration

The work described in this dissertation was carried out from in the School of Mathematics, Statistics and Computer Science at the University of KwaZulu-Natal, under the supervision of Professor Henry G. Mwambi.

The thesis represents original work by the author and has not otherwise been submitted in any form for any degree or diploma to any other tertiary institution. Where use has been made of the work of others, it is duly acknowledged in the text.

Yusentha Balakrishna

Student number 209503118 Signature

Professor Henry G. Mwambi

Supervisor Signature

Abstract

By knowing the incidence of an infectious disease, we can ascertain the high risk factors of the disease as well as the effectiveness of awareness programmes and treatment strategies. Since the work of Hugo Muench in 1934, many methods of estimating the force of infection have been developed, each with their own advantages and disadvantages.

The objective of this thesis is to explore the different compartmental models of infectious diseases and establish and interpret the parameters associated with them. Seven models formulated to estimate the force of infection were discussed and applied to data obtained from CAPRISA. The data was age-specific HIV prevalence data based on antenatal clinic attendees from the Vulindlela district in KwaZulu-Natal.

The link between the survivor function, the prevalence and the force of infection was demonstrated and generalized linear model methodology was used

to estimate the force of infection. Parametric and nonparametric force of infection models were used to fit the models to data from 2009 to 2010. The best fitting model was determined and thereafter applied to data from 2002 to 2010. The occurring trends of HIV incidence and prevalence were then evaluated. It should be noted that the sample size for the year 2002 was considerably smaller than that of the following years. This resulted in slightly inaccurate estimates for the year 2002.

Despite the general increase in HIV prevalence (from 54.07% in 2003 to 61.33% in 2010), the rate of new HIV infections was found to be decreasing. The results also showed that the age at which the force of infection peaked for each year increased from 16.5 years in 2003 to 18 years in 2010.

Farrington's two parameter model for estimating the force of HIV infection was shown to be the most useful. The results obtained emphasised the importance of HIV awareness campaigns being targeted at the 15 to 19 year old age group. The results also suggest that using only prevalence as a measure of disease can be misleading and should rather be used in conjunction with incidence estimates to determine the success of intervention and control strategies.

Acknowledgements

Foremost, I would like to thank Professor Henry Mwambi for his supervision. His dedication to my work, guidance and reassuring demeanour contributed immensely to the completion of this thesis. I have been honoured to have had you as my supervisor and will eternally be grateful to you. Sincere gratitude to Dr. Thomas Achia who never failed to give me his undivided attention whenever consulted.

Thank you to SACEMA (South African Centre for Epidemiological Modelling and Analysis) for all your academic support and to the NRF (National Research Foundation) for your financial assistance. Many thanks to CAPRISA (Centre for the AIDS Programme of Research in South Africa) and a most sincere and humble thank you especially to Dr. Ayesha Kharsany who willingly gave me kind permission for use of their data in this thesis. Your valued support has made this thesis possible.

My deep gratitude and thanks to senior tutor Mr Yougan Aungamuthu who gave me insight and clarity during the last few hurdles which I encountered. Finally, a very special thank you to my family and friends. Your encouragement and relentless faith in my capabilities inspired me during the most daunting of times.

I dedicate this thesis to my mother Pravina Naidoo. You are the greatest blessing in my life and my deep desire to always make you proud is the paramount reason for all that I have accomplished.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Infectious diseases frequently dominate news headlines because of their ability to spread rapidly amongst the population and debilitate a country's healthcare facilities and policies. Thus by having prior knowledge of the transmission patterns of an infectious disease, we are able to discover whether an epidemic is a likely outcome or not, amongst other vital conclusions. It is possible to mathematically model the progress of most infectious diseases. Modelling infectious diseases allows us to gain insight into mechanisms influencing the spread of the disease. Such models force a clear statement of assumptions and hypotheses and help to derive new insights and hypotheses. Modelling infectious diseases establishes relative importance of different

processes and parameters so as to focus research or management efforts, and explores management options. Modelling and analysis of infectious diseases further helps to guide the design of studies that lead to the collection of relevant data, to inform the design of public health policies and in the design of control strategies.

Models can support, add to, and sometimes even overturn prevailing wisdom. Models support research flow for the development of new treatment and vaccination plans. They can aid public health policy decisions concerning a country's, such as South Africa's, most dangerous infectious diseases such as HIV, TB, malaria and many more. Essentially, infectious disease modelling plays a key role in policy making, health-economic aspects, emergency planning and risk assessment and control-programme evaluation.

The force of infection is a crucial parameter in epidemiological models and characterizes the instantaneous rate at which susceptible individuals acquire an infectious disease. The force of infection consists of the contact rate between susceptible and infected individuals, the probability of disease transmission given contact and the probability that the randomly chosen partner

is infectious (Akpa et al. 2010).

Relying solely on prevalence to determine the level of disease within a population can lead to inaccurate conclusions. Prevalence is defined as the proportion of cases in a population at a given time. It indicates how widespread a disease is whereas the force of infection indicates the risk of contracting the disease. Prevalence is sensitive to the age structure of the population. Younger individuals have less time to become infected and will thus record lower prevalence estimates than those of older individuals. Further, if the disease mortality is high (infected individuals dying sooner than the uninfected), raw prevalence may underestimate the impact of the disease since living individuals are less likely to be infected. Force of infection models are helpful in this regard and can analyse prevalence data when age can be determined at the time of the sample and when the disease is endemic in a population (Conn et al. 2012).

1.1 Literature review and preliminary concepts

In 1760, a publication by Daniel Bernoulli was the first to introduce mathematical modelling of infectious diseases but it was Hugo Muench who first proposed the idea of estimating the force of infection. In the year 1934, Hugo Muench stated,

"The thing to do, then, is to find out what curve describes the growth of the summation data and to find the derivative, which will be the rate at which the curve is rising at different ages."

This formed the basis of Muench's catalytic model and the in depth research of the estimation of the force of infection that followed. However, it was only in 1959 that his work became widely known, through a publication entitled *Catalytic Models in Epidemiology*. Muench's model assumed a constant force of infection that was applicable to the entire population at any point in time i.e. at any age. It also catered for the fraction of the population that could not be infected at all. Muench (1934) modelled prevalence as

$$\pi(a) = 1 - e^{-\lambda a}$$

where λ was assumed to be a constant effective exposure rate. He determined the rate of change of the prevalence to be:

$$\Delta\pi(a) = \lambda e^{-\lambda a}$$

Thus the rate of change per susceptible was modelled by Muench as

$$l(a) = \frac{\Delta\pi(a)}{1-\pi(a)} = \frac{\lambda e^{-\lambda a}}{e^{-\lambda a}} = \lambda$$

Muench's work triggered further research into the force of infection and both parametric and nonparametric methods were developed to estimate the force of infection.

Griffiths (1974) proposed an age-dependent (linear) force of infection. Letting $l(a)$ denote the force of infection (or equivalently the hazard of infection as in survival analysis) as a function of age, then the model by Griffiths can be written as

$$l(a) = \begin{cases} \beta_1 + 2\beta_2 a & , \quad a > \tau \\ 0 & , \quad a \leq \tau \end{cases}$$

τ=end of the maternal antibody period

Since τ has been specified as a parameter in the model, the model can be viewed as a changepoint model. Griffiths deduced that the prevalence for his model would be

$$\pi(a) = 1 - e^{-(\beta_0+\beta_1 a+\beta_2 a^2)}$$

and showed the linear trend by plotting

$$l(a) = \frac{\Delta\pi(a)}{1-\pi(a)}$$

against age.

Eleven years later, Grenfell and Anderson (1985) expanded on Griffiths (1974) model and proposed a polynomial function to model the force of infection. The proposed approach has the advantage of flexible curve shapes because of the higher order of the polynomials used. The force of infection under Grenfell and Anderson's model did not constrain the force of infection to be constant or linear but instead allowed the data to lead the results. The model assumed can be written as

$$\pi(a) = 1 - e^{-\sum \beta_i a^i},$$

implying

$$l(a) = \sum \beta_i i a^{i-1}.$$

However, the Grenfell and Anderson model had the complication of yielding negative estimates for the force of infection. Farrington (1990) sought to correct this by considering a nonlinear model:

$$l(a) = (\alpha_1 a - \alpha_3)e^{-\alpha_2 a} + \alpha_3.$$

To ensure that his model produced positive estimates, Farrington constrained the parameter space to be non-negative.

Finally, Keiding (1991) proposed a nonparametric method of estimating the force of infection from serological data, based on the Kaplan-Meier estimator of $1 - \pi(a)$ which finds its origin in survival analysis. Keiding addressed the issues of time homogeneity, monotonicity, and censoring. He used a kernel smoother in his method and in 1996 he proposed to replace the kernel smoother with a smoothing spline. Semi-parametric models were later explored, in which the age-specific prevalence was modelled nonparametrically and possible covariate effects such as gender were included in the parametric component of the model. In more recent times, Shkedy et al (2003 and 2006) proposed local and fractional polynomials to model the force of infection.

The five milestone papers of Muench (1934), Griffiths (1974), Grenfell and Anderson (1985), Farrington (1990), and Keiding (1991), form the framework of the estimation of the force of infection that is still used today. In more recent times, researchers have built upon this work using more modern statistical models and advanced estimation procedures that are subsequently more computer intensive.

1.2 HIV prevalence among antenatal clinic attendees in South Africa

Until 1998, South Africa had one of the fastest growing HIV epidemics in the world, but since 2006, HIV prevalence among pregnant women has remained relatively stable.

The South African Department of Health carried out the *National Antenatal Sentinel HIV and Syphilis Prevalence Survey in South Africa* in 2010, published in 2011. The annual study looks at data from antenatal clinics and uses it to estimate HIV prevalence amongst pregnant women. According to the study, KwaZulu-Natal's estimated HIV prevalence among the antenatal clinic attendees was 39.5% for both 2009 and 2010. KwaZulu-Natal had the

highest HIV prevalence for both years. Table 1.1 shows the estimated HIV prevalence among antenatal clinic attendees in KwaZulu-Natal by age.

Table 1.1: The estimated HIV prevalence (%) among antenatal clinic attendees in KwaZulu-Natal, by age.

Age Group	2009	2010
10-14	7.9	9.1
15-19	13.7	14.0
20-24	26.6	26.7
25-29	37.1	37.3
30-34	41.5	42.6
35-39	35.4	38.4
40-44	25.6	30.9
45-49	23.9	28.2

In such a large and diverse country as is South Africa, the true figures can not be known exactly. What is essential however, is that the limitations of each study are acknowledged whenever their results are interpreted. For this reason, the advantages and disadvantages of each force of infection estimation method is considered when comparing the results obtained in this thesis. We

also discuss the suitability of using antenatal clinic data.

1.3 Statement of problem

The project starts by first exploring the different infectious disease compartmental models and the parameters associated with them. In particular, we discuss the SIR (Susceptible-Infected-Recovered), MSLIR (Maternal Protected-Susceptible-Latent-Infected-Recovered) and SIS (Susceptible-Infected-Susceptible) models. Note that the transmission of HIV is described by an SI (Susceptible-Infected) model. These compartmental models are the basic models for transmission of a disease and are established by classifying the affected population by disease status, namely susceptible, infected and recovered. The rates at which individuals move from one state to the next are key to the model and in understanding the risk level of the disease. We shall be studying the interpretation of these rates which are key parameters of the models and transmission dynamics. The specific rates in question are the incidence rates, force of infection and recovery rates. The research will also seek to estimate the key disease parameters using different statistical assumptions and methods.

1.4 Specific objectives

The specific objectives for this project are:

1. To explore the different compartmental models of infectious diseases and establish the parameters associated with them including their interpretation.

2. Demonstrate the link between the survivor function as used in time to event analysis, the prevalence and the hazard or force of infection.

3. Use the generalized linear model formulation and methodology to estimate the force of infection.

4. Show the application to HIV antenatal clinic sero-prevalence data.

1.5 Data

At the heart of any statistical research project lies the data. The project makes use of current status data obtained from CAPRISA (Centre for AIDS Programme Research In South Africa). The data was collected from the Vulindlela Clinical Research Site, one of CAPRISA's five clinical research

sites.

Vulindlela is a sub-district situated 90 minutes west of Durban. About 230 000, predominantly Zulu-speaking, people reside in this rural community. Access to seven primary health care clinics is available. Residents are able to receive extensive primary care as well as acquire advice on family planning, sexually transmitted infection (STI) treatment, antenatal care, treatment of opportunistic infections and minor ailments and are able to undergo voluntary HIV counseling and testing. Grey's Hospital and Edendale Hospital, the regional referral hospitals, are within a 30 minute drive away from the Vulindlela district. Many of the 60 community-based organizations in the district provide residents with HIV prevention and home-based care services. These organisations are also closely linked with CAPRISA.

The data we used was obtained from pregnant women visiting the antenatal clinics in the Vulindlela district. They were asked a variety of questions and voluntary anonymous testing for HIV was carried out at the clinic. From the many variables available, we extracted the age of each patient and their HIV status. The reason is because age is a key determinant of the onset of most

infectious diseases and HIV infection, which is the focus of the current study, is not an exception.

Antenatal clinic data is used since pregnant women attending antenatal clinics are thought more likely to best represent the general population of adult women as they constitute an easily identifiable, accessible and stable population. This is because they consist of the proportion of the population that are sexually active and are not using contraceptives. Thus, this makes them susceptible to HIV infection. They are thus used to estimate HIV prevalence. However, there are many biases that may arise because of using only this select population. Only pregnant women are tested suggesting that only fertile women are sampled. Also, HIV-infected women may be less likely to become pregnant and studies have shown that HIV reduces fertility (Gray et al, 1997). In addition, not all pregnant women may be attending antenatal clinics.

Chapter 2

General Disease Transmission Models

Individuals in a population can be classified into different states with respect to their disease status. This method of dynamical rules is applicable to only transmissable diseases such as AIDS, SARS, measles and other infectious diseases. This is because the number of new cases of infection or incidence of the disease is dependent on the number of existing cases. Compared with non-transmissable diseases such as cardio-vascular diseases, the number of new cases is independent of the number of existing cases.

2.1 The SIR model

The SIR model is the most basic infectious disease model. The model is applicable mostly to viral and bacterial diseases that can confer immunity. For simple infectious diseases, the population can be modelled by (Shkedy et al. 2009):

$$N(a,t) = X(a,t) + Y(a,t) + Z(a,t) \tag{2.1}$$

N(a,t) - total population at age a and time t

X(a,t) - number of susceptible individuals at age a and time t

Y(a,t) - number of infected individuals at age a and time t

Z(a,t) - number of recovered/immune individuals at age a and time t

The SIR model operates on the following assumptions:

i) Newborns enter directly into the susceptible phase i.e the period during which newborns are protected from infection maternal antibodies is ignored.
ii) Infection, the latent period, the infectious period and disease occur simultaneously.

iii) Once an individual has recovered from the disease, they cannot be re-infected.

We also assume the birth rate to be equal to the natural death rate so that the total population size is constant.

The compartmental nature of the SIR model leads to interest in estimating the rate at which individuals move from state to state. Figure 2.1 depicts the basic SIR model.

There are five rates that need to be considered when modelling the flow of individuals within the population, with respect to age and time:

- Birth rate $\mu(a,t)$: The rate at which individuals enter the susceptible class.
- Force of infection λ(a,t): The rate at which individuals leave the susceptible class and enter the infected class.
- Recovery rate $\nu(a,t)$: The rate at which individuals leave the infected

Figure 2.1: The basic SIR model.

class and enter the recovered class.

- Disease death rate $\alpha(a,t)$: The excess mortality rate at which individuals leave the infected class as a consequence of death due to disease.

- Natural death rate $\mu(a,t)$: The rate at which individuals leave the susceptible, infected and recovered classes as a consequence of natural death.

Change in number of susceptible individuals

Change in susceptibles = number of newborns - number leaving susceptible

class

$\Rightarrow \frac{\partial X(a,t)}{\partial a} + \frac{\partial X(a,t)}{\partial t} = N(a,t)\mu(a,t) - [\lambda(a,t) + \mu(a,t)]X(a,t)$

Change in number of infected individuals

Change in infected = number of newly infected - number of recovering and dying infected

$\Rightarrow \frac{\partial Y(a,t)}{\partial a} + \frac{\partial Y(a,t)}{\partial t} = \lambda X(a,t) - [\nu(a,t) + \alpha(a,t) + \mu(a,t)]Y(a,t)$

Change in number of immune individuals

Change in immunes = number of newly recovered - number of dying immune

$\Rightarrow \frac{\partial Z(a,t)}{\partial a} + \frac{\partial Z(a,t)}{\partial t} = \nu(a,t)Y(a,t) - \mu(a,t)Z(a,t)$

The above three equations are together known as the governing disease transmission equations.

2.1.1 The time homogeneous model

The time homogeneous model is a steady state SIR model. Where the SIR model employs two timescales of host age and time, the time homogeneous model employs only one timescale of host age. This means that there is

no time dependence in the variables in the governing disease transmission equations. We also assume that the disease death rate $\alpha = 0$. it will be shown that by assuming an equal birth and death rate of μ and a disease death rate of $\alpha = 0$, we will arrive at a system of two coupled ordinary differential equations which becomes easier to analyse. The transmission models under the time homogeneous model become (Shkedy et al. 2009):

$$N(a) = X(a) + Y(a) + Z(a) \tag{2.2}$$

$$\frac{dX(a)}{da} = N(a)\mu(a) - [\lambda(a) + \mu(a)]X(a), \tag{2.3}$$

$$\frac{dY(a)}{da} = \lambda(a)X(a) - (\nu(a) + \mu(a))Y(a), \tag{2.4}$$

$$\frac{dZ(a)}{da} = \nu(a)Y(a) - \mu(a)Z(a). \tag{2.5}$$

The total number of individuals at age a denoted N(a) can be calculated as:

$$\text{N(a)} = \text{N(0)P(survives to age } a) = \text{N(0)S(a)}$$

The survival function $S(a)$ can assume two forms which can be labelled as Type I and Type II survivor function.

Type I survivor function

Under Type I survivor function, we consider only two cases: a person dying before age L or a person dying after age L. The survival function $S(a)$ then becomes:

$$S(a) = \begin{cases} 1 & , \quad a \leq L \\ 0 & , \quad a > L \end{cases} \tag{2.6}$$

Type II survivor function

Type II survivor function assumes the death rate to be a constant rate of μ over time. The survival function $S(a)$ then becomes:

$$S(a) = e^{-\mu a} \tag{2.7}$$

Thus N(a) can be determined assuming either Type I or Type II survival. Given Type I survival:

$$N(a) = \begin{cases} N(0) & , \quad a \leq L \\ 0 & , \quad a > L. \end{cases} \tag{2.8}$$

Given Type II survival:

$$\begin{aligned} N(a) &= N(0)S(a) \\ &= N(0)e^{-\mu a}. \end{aligned} \tag{2.9}$$

In equation (2.2), X(a) is the number of individuals present in the susceptible class. It is possible to calculate X(a) by using our knowledge of Markov chains. The introduction of a stochastic process accommodates for the unequal time intervals between observations (Mwambi et al. 2011). Also, real processes are stochastic and adding stochasticity to a model gives it greater flexibility ensuring a better tool of estimation (Haran. 2009). Let T_X be the amount of time spent in the susceptible class (and the outcome of an exponential distribution). Ignoring the natural death rate (since an epidemic outbreak moves faster than the vital rates), individuals leave the susceptible class at a rate of λ. Thus:

$$T_X \sim Exp(\lambda)$$

Therefore, the probability that an individual becomes infected before age a (moves from class X to class Y before age a) is the cumulative distribution of T_X.

$$\begin{aligned} P(T_X \leq a) &= \int_0^a \lambda e^{-\lambda x} dx \\ &= 1 - e^{-\lambda a}. \end{aligned} \tag{2.10}$$

Hence, the probability of staying in the susceptible class at age a is

$$P(T_X > a) = e^{-\lambda a}.$$

X(a) can be calculated as:

$$X(a) = N(a)P(T_X > a) = N(a)e^{-\lambda a}. \tag{2.11}$$

Under Type I mortality:

$$\begin{aligned} X(a) &= N(a)e^{-\lambda a} \qquad (2.12)\\ &= \begin{cases} N(a)e^{-\lambda a} & , \quad a \leq L \\ 0 & , \quad a > L \end{cases} \end{aligned}$$

$$\frac{dX(a)}{da} = -\lambda N(a)e^{-\lambda a} = -\lambda X(a) \tag{2.13}$$

Under Type II mortality:

$$\begin{aligned} X(a) &= N(a)e^{-\lambda a} \qquad (2.14)\\ &= N(0)e^{-\mu a}e^{-\lambda a} \\ &= N(0)e^{-(\lambda+\mu)a} \end{aligned}$$

$$\frac{dX(a)}{da} = -(\lambda+\mu)N(0)e^{-(\lambda+\mu)a} = -(\lambda+\mu)X(a) \tag{2.15}$$

Let $x(a)$ be the proportion of the population that is susceptible at age a.

$$x(a) = \frac{X(a)}{N(a)} = \frac{N(0)e^{-\lambda a}e^{-\mu a}}{N(0)e^{-\mu a}} = e^{-\lambda a} \tag{2.16}$$

By using the proportion of susceptible individuals at a given age or time, we eliminate the natural death rate μ and thus any change in the proportion of susceptible hosts would be due solely to individuals moving into the infected class.

Similarly, we can eliminate the natural death rate for the infected class and the recovered class by calculating the proportion of the population that each of these classes represent. For the infected class, we integrate the differential equation over all ages and calculate the proportion y(a).

$$\frac{dY(a)}{da} = \lambda X(a) - (\nu + \mu)Y(a). \tag{2.17}$$

$$\Rightarrow Y(a) = \frac{\lambda}{\lambda - \nu}N(a)[e^{-\nu a} - e^{-\lambda a}],$$

$$y(a) = \frac{Y(a)}{N(a)} = \frac{\lambda}{\lambda - \nu}[e^{-\nu a} - e^{-\lambda a}]. \tag{2.18}$$

Likewise, for the recovered class we have:

$$\frac{dZ(a)}{da} = \nu Y(a) - \mu Z(a) \tag{2.19}$$

But N(a) = X(a) + Y(a) + Z(a),

$$\Rightarrow Z(a) = N(a) - X(a) - Y(a)$$

Therefore,

$$z(a) = \frac{Z(a)}{N(a)} = 1 - x(a) - y(a) \tag{2.20}$$

2.2 The MSLIR model

The MSLIR model is an expansion of the SIR model. It accounts for two more stages in the transmission of an infectious disease (Shkedy et al. 2009).

- The period during which an newborn is temporarily protected from infection by maternal antibodies.
- The period during which an individual is infected, but not yet infectious. This is known as the *latent* period.

Figure 2.2 depicts where these two additional stages fit into the SIR model.

Figure 2.2: The MSLIR model.

2.3 The SIS model

The SIS model has only two compartments: susceptible and infected. This is because it models diseases where re-infection after recovery is possible. Some forms of childhood respiratory diseases and some STIs such as gonorrhoea may fall under such a model. Figure 2.3 depicts this model. Some STI's such as gonorrhoea follow this model (Shkedy et al. 2009).

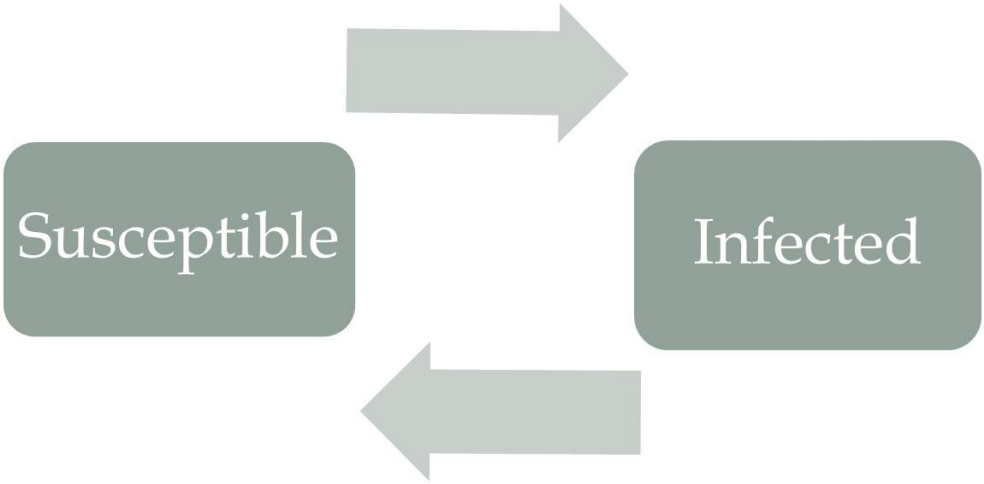

Figure 2.3: The SIS model.

Chapter 3

Exploratory Data Analysis

The data obtained from CAPRISA was collected for the years of 2002 through to 2010. Each observation consisted of the age of the patient (pregnant woman attending the antenatal clinic in Vulindlela), and their status of HIV infection. Disease status was ascertained via the ELISA test. Patient ages ranged from a minimum of 12 years to a maximum of 45 years.

3.1 Prevalence estimation by year

The yearly prevalence estimate can be calculated using the following formula

$$p = \frac{x}{n} \tag{3.1}$$

p=prevalence estimate

x=number of patients who tested positive for HIV infection

n=total number of individuals in the sample

A cursory glance at the prevalence (estimated using equation 3.1) over the years (Figure 3.1) reveals a general increasing trend. We find 2004 to have the highest prevalence estimate and 2007 to have the lowest. Focusing on more recent times, the prevalence estimate for 2009 is shown to be 0.3898 (95% CI 0.3416-0.4402) or 38.98% and that of 2010 to be 0.4088 (95% CI 0.3578-0.4618) or 40.88%. This is realistic as it echoes KZN's HIV prevalence rate of 39.5%. Further it should be noted that Vulindlela is one of the high risk HIV areas within KwaZulu-Natal so it should not be a surprise that its prevalence estimates tend to be higher than those at the province level.

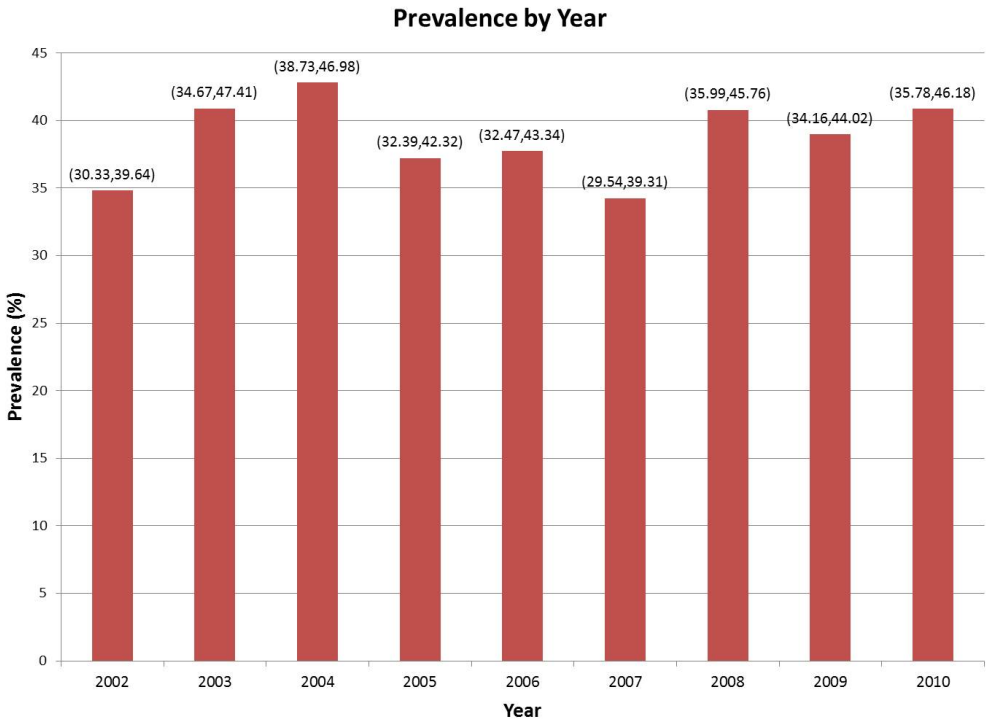

Figure 3.1: The estimated prevalence of HIV amongst pregnant women in Vulindlela by year. The prevalence bars also include confidence intervals at the top.

3.2 Prevalence estimation by age group

Making use once again of equation 3.1, we can calculate the estimated prevalence of HIV amongst the women for each age group. The data was grouped into three age groups: 24 years and younger, 25-34 years, and 35 years and above. Figure 3.2 shows estimated prevalence for each of these age groups.

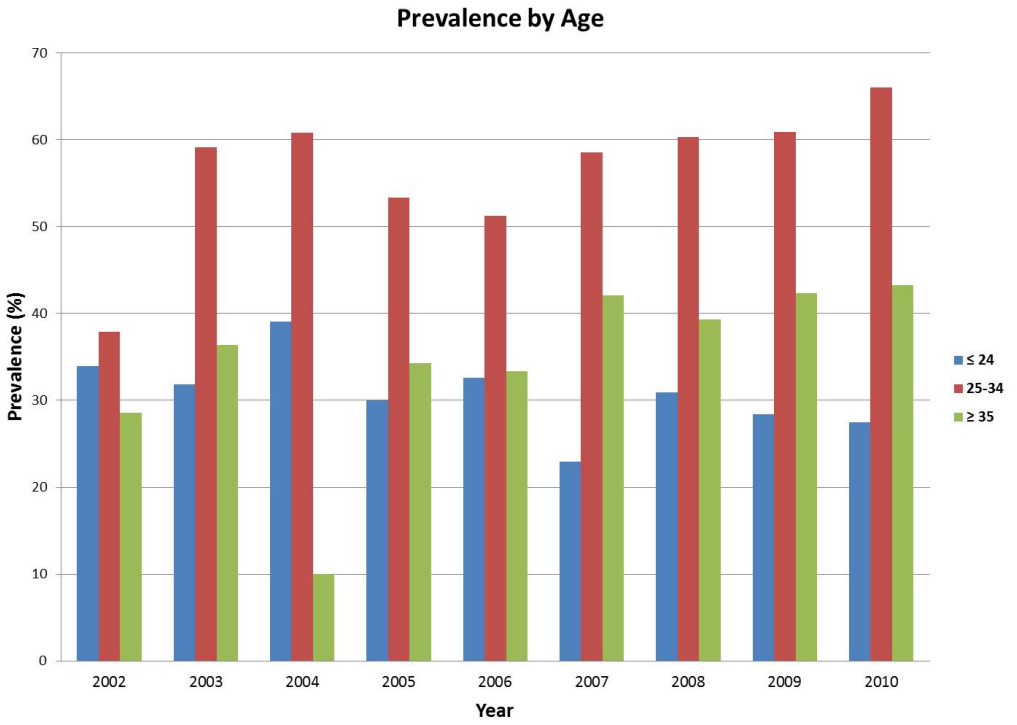

Figure 3.2: The estimated prevalence of HIV amongst pregnant women in Vulindlela by age group.

Figure 3.2 shows the prevalence by age. The most striking feature we find is the age group of 25-34 years having the highest prevalence for each year.

In general, focusing on recent years, we notice that 2010 has a higher overall observed prevalence than 2009. The models we fit to estimate the force of infection should reflect this. The higher prevalence in 2010 may tend to

support the argument that while we might think we are winning the war against HIV, it might not be the case in some high risk areas. This also touches on the question of spatial temporal heterogeneous dynamics of the disease.

Chapter 4

Estimation of the Force of Infection for HIV

4.1 A note about the generalized linear model

The Generalized Linear Model (GLM) was first introduced by Nelder and Wedderburn (1972). The model provides a unified theory of regression modelling that encompasses the most important models for continuous and discrete variables. There are three main characteristics present in GLMs as described in Section 4.1.2. There are two important issues to consider:

- The distribution of the response

- The model (*link function*) that relates the mean response to the explanatory variables

GLMs are restricted to the exponential family of distributions for the response Y because the algorithm applies to the entire family, for any choice of the link function. Note that Y here is a random variable that can be continuous or discrete, depending on the nature of the outcome of the event of interest.

4.1.1 The exponential family of distributions

The canonical form of distributions that are members of the exponential family is

$$f(y, \theta, \phi) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)} \tag{4.1}$$

where $a(.)$, $b(.)$ and $c(.)$ are specific functions, θ is the natural location parameter and ϕ is the scale parameter.

Two important properties of the exponential family of distributions are:

$$\mu = E(y) = b'(\theta) \tag{4.2}$$

$$var(y) = b''(\theta)a(\phi) \tag{4.3}$$

The normal, binomial, negative binomial, exponential and gamma distributions are just some of the distributions that belong to this family.

4.1.2 The GLM structure

A GLM has three important components:

- *Random Component*: the response variables $Y_1, Y_2, \ldots, Y_n$ are independently and identically distributed from the exponential family having the canonical form

- *Systematic Component*: a linear predictor as a function of explanatory variables,

$$\eta_i = \mathbf{x}_i\beta$$

- *Link Function*: a relationship between the linear predictor η_i to the expected value $\mu_i = E(Y_i)$,

$$\eta_i = g(\mu_i)$$

There are many link functions available, depending on the response variable of interest, but the most commonly used, and the ones that we make use of when modelling current status data in the form of binary outcomes are listed below. Assuming Y_i $Bernoulli(1, \mu_i)$:

- The log link function

$$\eta_i = ln(\mu_i).$$

- The logit link function

$$\eta_i = ln\{\frac{\mu_i}{1-\mu_i}\}.$$

- The complementary log-log link function

$$\eta_i = ln[-ln(1-\mu_i)].$$

It is important to understand that the link function is a transformation on the population mean, and not the data. This is the main idea as introduced by Nelder and Wedderburn (1972). Note that because of specific modelling needs, the use of a canonical link function is not a necessity as will be demonstrated in the current application.

4.1.3 Parameter estimation

When carrying out parameter estimation under the GLM models we can make use of three possible iterative methods to estimate the parameters of a the GLM model.

- Newton-Raphson method
- Fisher Scoring algorithm
- Iterative Reweighted Least Squares method

The Newton-Raphson method is a commonly used method for finding zero approximations of a real-valued function. It can also be used to find a minimum or maximum of a function by applying the method to the derivatives (Newton-Raphson method in optimization). The Fisher scoring algorithm is a form of the Newton-Raphson method and is used to solve maximum likelihood equations numerically. Nelder and Wedderburn (1972) used it to estimate $\hat{\beta}$ in GLMs. Fisher scoring is a special case of the iterative reweighted least squares method.

Newton-Raphson method

Given a function $f(x)$ and its derivative $f'(x)$, we can use the Newton-

Raphson updating equation

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, n = 0, 1, \ldots \tag{4.4}$$

Fisher Scoring algorithm

Specifically, given an initial estimate β, the algorithm update equation for Fisher Scoring is

$$\beta^{new} = \beta + \left[E\left(-\frac{\partial^2 I}{\partial\beta\partial\beta^T} \right) \right]^{-1} \frac{\partial I}{\partial\beta}. \tag{4.5}$$

It can be shown that equation (4.5) can be rewritten as

$$\beta^{new} = \beta + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}, \tag{4.6}$$

where $\mathbf{z}$ is the n-vector with ith component

$$z_i = (Y_i - \mu_i)g'(\mu_i),$$

and $\mathbf{W}$ is the $n \times n$ diagonal matrix with

$$W_{ii} = \left[g'(\mu_i)^2 b''(\theta_i) \right]^{-1}.$$

Iterative Reweighted Least Squares method

In GLM, if we let $\mathbf{z}^*$ be an n-vector with the i^{th} component given by

$$z_i^* = (Y_i - \mu_i)g'(\mu_i) + \mathbf{x}_i^T\beta,$$

then the updating equation of GLM becomes

$$\beta^{new} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}^*, \tag{4.7}$$

where

$$\mathbf{W} = diag[(g'(\mu_1)^2 b''(\theta_1))^{-1}, \ldots, (g'(\mu_n)^2 b''(\theta_n))^{-1}].$$

4.2 Modelling current status data

Before we can begin modelling current status, there are some concepts that need to be established first.

(i) In general, the change in the probability of being susceptible, with respect to age and time, can be expressed as

$$\frac{\partial q(a,t)}{\partial a} + \frac{\partial q(a,t)}{\partial t} = -l(a)q(a,t).$$

$q(a,t)$ = P(susceptible at age a, time t)

$l(a,t)$ = force of infection/hazard

The above expression follows from ignoring the vital dynamics in $\frac{\partial X(a,t)}{\partial a} + \frac{\partial X(a,t)}{\partial t} = N(a,t)\mu(a,t) - (\lambda(a,t) + \mu(a,t))X(a,t)$. Since a single epidemic outbreak moves faster than the normal birth and death rates in a given population, normal birth and death rates are excluded. A possible limitation is assuming the disease death rate to be zero. Future research could involve accommodating this parameter for increased accuracy.

However, we are interested in the change of the number of susceptible individuals under the assumption of time homogeneity ($\frac{dq(a,t)}{\partial t} = 0$). Hence

$$\frac{dq(a)}{da} = -l(a)q(a).$$

(ii) Current status data can be interval-censored, left- or right-censored. With *interval-censored* data, each individual is tested more than once with a specified period of time between each test. If an individual tests negative for the disease at age a_1 but tests positive for the disease at age a_2, this implies that although the exact age of infection is unknown, infection occurred within the age interval (a_1, a_2).

With left- and right-censored data (the type of data that we have available),

individuals are tested only once, at certain point in time, and two observations are recorded: the age of the individual and their disease status. Let age a^* be the age at which the individuals are tested. If an individual tests negative, it implies that the individual may be infected *after* age a^*. This is known as *right-censored* data (assumed to occur with probability $q(a)$).

If an individual tests positive, it implies that the individual was infected *before* age a^*. This is known as *left-censored* data (assumed to occur with probability $1 - q(a)$).

We can then observe the binary random variable Y_i where for sample size N and age of the i^{th} individual a_i

$$Y_i = \begin{cases} 1 \quad if \quad left - censored \\ \\ 0 \quad if \quad right - censored \end{cases} \tag{4.8}$$

In survival analysis, the probability of an individual surviving beyond time x (event occurs after time x) is given by the *survival* function S(x). The *hazard rate* is the rate at which the event occurs and is given by

$$\frac{f(x)}{S(x)} = \frac{F'(x)}{1 - F(x)}, \tag{4.9}$$

where $f(x) = F'(x) = [1 - S(x)]' = -S'(x)$.

Adapting (4.9) for our purposes, the probability of an individual becoming infected after age a (surviving or escaping infection beyond age a) is

$$S(a) = 1 - prevalence = 1 - \pi(a)$$

Thus $F(x)$ in (4.9) corresponds to $\pi(a)$. Since the hazard rate is congruent to the force of infection:

$$l(a) = \frac{\pi'(a)}{1 - \pi(a)}. \tag{4.10}$$

Letting the probability that an individual is susceptible at age a i.e. infected after age a, be represented by q(a), the prevalence $\pi(a)$ is given by

$$\pi(a) = 1 - q(a). \tag{4.11}$$

It is important to note that $\pi(a)$ represents a cumulative prevalence. It denotes the probability of being infected before or at age a. Thus when interpreting our results, the value of $\pi(a)$ at the maximum age in the sample estimates the prevalence rate of the disease in the total sample. This is because the cumulative prevalence at age a_{max} includes the cumulative prevalences of all ages less than a_{max} in the sample.

Note

Consider an age-specific cross-sectional prevalence sample of size N where a_i is the age of the i^{th} subject. Instead of observing the age at infection, we observe a binary variable Y_i such as in equation (4.8).

Remembering that $\pi(a_i) = \text{P(infected before age } a_i) = 1 - q(a_i)$, our pdf becomes

$$f[Y_i, \pi(a_i)] = \pi(a_i)^{Y_i}(1 - \pi(a_i))^{1-Y_i}.$$

Thus our log-likelihood is

$$L = \sum_{i=1}^{N} \{Y_i ln[\pi(a_i)] + (1 - Y_i) ln[1 - \pi(a_i)]\}. \tag{4.12}$$

Incorporating a link function g and a linear predictor η, we arrive at

$$g[\pi(a)] = \eta(a) \Rightarrow \pi(a) = g^{-1}[\eta(a)] \tag{4.13}$$

g is often taken to be the logit link function $\left\{ln(\frac{\pi}{1-\pi})\right\}$, but other link functions can be used as well. For example, the complementary log-log link $\{ln[-ln(1 - \pi)]\}$ and the log link $\{-ln(1 - \pi)\}$.

Using a model with a log link function leads to a simple interpretation of the first derivative of the linear predictor. $\eta(a)$ is the cumulative hazard therefore the force of infection is equivalent to the first derivative of $\eta(a)$. In the general case, when the link function is not restricted, the force of infection

Table 4.1: Table of link functions and their corresponding δ structures

Link	$\pi(a)$	$l(a)$	$\delta[\eta(a)]$
log	$1 - e^{-\eta(a)}$	$\eta'(a)$	1
clog-log	$1 - e^{-e^{\eta(a)}}$	$\eta'(a)e^{\eta(a)}$	$e^{\eta(a)}$
logit	$\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$	$\eta'(a)\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$	$\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$

can still be derived using the definition of the hazard rate.

It is easy to see that for the binomial distribution, the force of infection can be expressed as a product of two functions:

$$l(a) = \eta'(a)\delta[\eta(a)]. \tag{4.14}$$

where δ is determined by the link function. Table 4.1 shows three possible link functions with their corresponding δ structure for the force of infection.

4.2.1 Constant force of infection

The catalytic model assumes a constant force of infection applicable to the entire population at any point in time, that is, at any age. Assuming a constant force of infection for HIV, however, implies like risk factors for both younger and older individuals. Thus, age becomes our defining risk factor

and the force of infection should be modelled as such.

If we assume that the time spent in the susceptible class follows an exponential distribution with parameter λ, we obtain a constant force of infection. The pdf of the exponential distribution is

$$f(x) = \lambda e^{-\lambda x}.$$

From the pdf above, we can derive the probability of being susceptible at age a, the prevalence and the force of infection as

$$q(a) = P(X > a) = e^{-\lambda a},$$

$$\pi(a) = 1 - q(a) = 1 - e^{-\lambda a}. \tag{4.15}$$

$$l(a) = \frac{\pi'(a)}{1 - \pi(a)} = \frac{\lambda e^{-\lambda a}}{e^{-\lambda a}} = \lambda. \tag{4.16}$$

We can fit a general linear model with a complementary log-log link function in SAS to estimate our constant force of infection, λ. That is,

$$\begin{aligned} g[\pi(a)] &= ln[-ln[1 - \pi(a)]], \\ &= ln[-ln[e^{-\lambda a}]], \end{aligned} \tag{4.17}$$

$$\begin{aligned} &= ln(\lambda a), \\ &= ln(\lambda) + ln(a), \\ &= \mu + ln(a). \end{aligned}$$

Thus we can fit the above model in SAS using Proc GENMOD and obtain the intercept μ. Thus

$$\mu = ln(\lambda) \Rightarrow \lambda = e^{\mu}.$$

4.2.2 Linear force of infection

The age-dependent(linear) force of infection model accounts for the maternal antibody period though it has the disadvantage of constraining the force of infection to be linear.

Expanding and generalizing on the assumption of the constant force of infection estimation, let us assume that $\lambda(a) = \beta_0 + \beta_1 a + \beta_2 a^2$. Thus our prevalence $\pi(a)$ is given by $\pi(a) = 1 - e^{-\lambda(a)}$. Then

$$\pi(a) = 1 - e^{-(\beta_0 + \beta_1 a + \beta_2 a^2)}. \tag{4.18}$$

Thus

$$l(a) = \frac{\pi'(a)}{1 - \pi(a)}, \tag{4.19}$$

$$= \frac{(\beta_1 + 2\beta_2 a)e^{-(\beta_0+\beta_1 a+\beta_2 a^2)}}{e^{-(\beta_0+\beta_1 a+\beta_2 a^2)}},$$

$$= \beta_1 + 2\beta_2 a.$$

Thus to estimate the linear functional form of the force of infection a non-linear optimization algorithm can be used first to fit a prevalence model specifying initial values for β_0, β_1 and β_2. In our case Proc NLMIXED in SAS was used and since no information on the values of the β's is available we assume a minimal value for each. If estimates from previous research are known, they can be used as initial values. However, this thesis determines the initial values through trial-and-error and convergence of the employed algorithm.

Proc NLMIXED then iteratively provides us with estimates of β_0,β_1 and β_2 and thus the linear force of infection can be estimated, for a given age.

4.2.3 Weibull force of infection

If we assume that the time spent in the susceptible class follows a Weibull distribution with parameters k and λ, we obtain a monotone force of infection. The pdf of the Weibull distribution is

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(\frac{x}{\lambda})^k}.$$

Thus

$$q(a) = P(X > a) = e^{-(\frac{a}{\lambda})^k},$$

$$\pi(a) = 1 - q(a) = 1 - e^{-(\frac{a}{\lambda})^k}. \tag{4.20}$$

$$\begin{aligned} l(a) &= \frac{\pi'(a)}{1-\pi(a)}, \\ &= \frac{k\lambda^{-k}a^{k-1}e^{-(\frac{a}{\lambda})^k}}{e^{-(\frac{a}{\lambda})^k}}, \\ &= k\lambda^{-k}a^{k-1}. \end{aligned} \tag{4.21}$$

Letting $\alpha = \lambda^{-k}$ and $\beta = k$, we get

$$l(a) = \alpha\beta a^{\beta-1}. \tag{4.22}$$

We can fit a general linear model with a complementary log-log link function in SAS to estimate the force of infection.

$$\begin{aligned} g[\pi(a)] &= ln[-ln[1-\pi(a)]], \\ &= ln[-ln[e^{-\alpha a^\beta}]], \\ &= ln(\alpha a^\beta), \\ &= ln(\alpha) + \beta ln(a), \\ &= \mu + \beta ln(a). \end{aligned} \tag{4.23}$$

We can fit the above model in SAS using Proc GENMOD with $ln(a)$ as the covariate or predictor variable and obtain the estimates for μ and β. Note

$$\mu = ln(\alpha) \Rightarrow \alpha = e^{\mu}$$

and β is the regression coefficient for $ln(a)$. We can then substitute the values of α and β into $l(a)$ to estimate the force of infection for a given age.

4.2.4 Log-logistic force of infection

If we assume that the time of spent in the susceptible class follows a log-logistic distribution with parameters α and β, we obtain a single-peak force of infection. We know the pdf of the log-logistic distribution is

$$f(x) = \frac{(\frac{\beta}{\alpha})(\frac{x}{\alpha})^{\beta-1}}{[1+(\frac{x}{\alpha})^{\beta}]^2}.$$

Thus

$$\begin{aligned} q(a) = P(X > a) &= 1 - \frac{1}{1+(\frac{a}{\alpha})^{-\beta}}, \\ \pi(a) = 1 - q(a) &= \frac{1}{1+(\frac{a}{\alpha})^{-\beta}}. \end{aligned} \tag{4.24}$$

Hence, to derive our force of infection $l(a)$,

$$\begin{aligned} \pi'(a) &= \frac{d}{da}[1+a^{-\beta}\alpha^{\beta}]^{-1}, \\ &= -[1+a^{-\beta}\alpha^{\beta}]^{-2}.-\beta a^{-\beta-1}\alpha^{\beta}, \end{aligned} \tag{4.25}$$

$$= \frac{\beta a^{-\beta-1}\alpha^{\beta}}{(1+a^{-\beta}\alpha^{\beta})^2}.$$

$$\begin{aligned} 1-\pi(a) &= 1-\frac{1}{1+a^{-\beta}\alpha^{\beta}}, \\ &= \frac{1+a^{-\beta}\alpha^{\beta}-1}{1+a^{-\beta}\alpha^{\beta}}, \\ &= \frac{a^{-\beta}\alpha^{\beta}}{1+a^{-\beta}\alpha^{\beta}}. \end{aligned} \tag{4.26}$$

$$\begin{aligned} l(a) &= \frac{\pi'(a)}{1-\pi(a)}, \\ &= \frac{\beta a^{-\beta-1}\alpha^{\beta}}{(1+a^{-\beta}\alpha^{\beta})^2} \div \frac{a^{-\beta}\alpha^{\beta}}{1+a^{-\beta}\alpha^{\beta}}, \\ &= \frac{\beta a^{-\beta-1}\alpha^{\beta}}{(1+a^{-\beta}\alpha^{\beta})^2} \times \frac{1+a^{-\beta}\alpha^{\beta}}{a^{-\beta}\alpha^{\beta}}, \\ &= \frac{\beta a^{-1}}{1+a^{-\beta}\alpha^{\beta}}, \\ &= \frac{\beta a^{-1}}{a^{-\beta}(a^{\beta}+\alpha^{\beta})}, \\ &= \frac{\beta a^{\beta-1}}{a^{\beta}+\alpha^{\beta}}, \\ &= \frac{\beta a^{\beta-1}\alpha^{-\beta}}{\alpha^{-\beta}a^{\beta}+1}, \\ &= \frac{\beta a^{\beta-1}\lambda}{1+\lambda a^{\beta}}. \end{aligned} \tag{4.27}$$

where $\lambda = \alpha^{-\beta}$. We can fit a general linear model with a logit link function in SAS to estimate the force of infection.

$$g[\pi(a)] = logit[\pi(a)], \tag{4.28}$$

$$\begin{aligned} &= ln\left[\frac{(1+a^{-\beta}\alpha^{\beta})^{-1}}{(a^{-\beta}\alpha^{\beta})(1+a^{-\beta}\alpha^{\beta})^{-1}}\right], \\ &= -ln(a^{-\beta}\alpha^{\beta}), \\ &= \beta ln(a) - \beta ln(\alpha), \\ &= \beta ln(a) + ln(\alpha^{-\beta}). \end{aligned}$$

Letting $\lambda = \alpha^{-\beta}$;

$$\begin{aligned} g[\pi(a)] &= ln(\lambda) + \beta ln(a), \\ &= \mu + \beta ln(a). \end{aligned} \tag{4.29}$$

We can fit the log-logistic model in SAS using Proc GENMOD and obtain the estimates for μ and β. Note

$$\mu = ln(\lambda) \Rightarrow \lambda = e^{\mu}.$$

We can then substitute the values of β and λ into $l(a)$ to estimate the force of infection for a given age.

4.2.5 Farrington's force of infection

In 1985, Grenfell and Anderson proposed polynomial functions to model the force of infection. The method had the advantages of

- flexible curve shapes due to the higher order of polynomials

- an unconstrained force of infection since the model allowed the data to lead the results

and the disadvantages of

- flexibility limited the type of polynomial used
- unbounded and uncontrolled behaviour at the extremes of the age scale
- non-monotonicity
- negative force of infection estimates

In 1990, Farrington sought to correct the negative estimates of Grenfell and Anderson and proposed non-linear models for the force of infection. Under the three parameter Farrington (1990) model, the prevalence is modelled as

$$\begin{aligned} \pi(a) &= 1 - e^{\frac{\alpha_1}{\alpha_2} a e^{-\alpha_2 a} + \frac{1}{\alpha_2}(\frac{\alpha_1}{\alpha_2} - \alpha_3)(e^{-\alpha_2 a} - 1) - \alpha_3 a}, \\ &= 1 - e^{r}. \end{aligned} \tag{4.30}$$

where $r = \frac{\alpha_1}{\alpha_2} a e^{-\alpha_2 a} + \frac{1}{\alpha_2}(\frac{\alpha_1}{\alpha_2} - \alpha_3)(e^{-\alpha_2 a} - 1) - \alpha_3 a$.

Note that $\pi(a) \in (0, 1)$ implies that $r \in (-\infty, 0)$.

Hence, to derive our force of infection $l(a)$,

$$\begin{aligned} \frac{dr}{da} &= \frac{\alpha_1}{\alpha_2}e^{-\alpha_2 a} - \alpha_1 a e^{-\alpha_2 a} - \frac{\alpha_1}{\alpha_2}e^{-\alpha_2 a} + \alpha_3 e^{-\alpha_2 a} - \alpha_3, \qquad (4.31) \\ &= -\alpha_1 a e^{-\alpha_2 a} + \alpha_3 e^{-\alpha_2 a} - \alpha_3, \\ &= (-\alpha_1 a + \alpha_3)e^{-\alpha_2 a} - \alpha_3. \end{aligned}$$

$$\begin{aligned} \pi'(a) &= -e^r \times \frac{dr}{da} \qquad (4.32) \\ &= -e^r \times \left[(-\alpha_1 a + \alpha_3)e^{-\alpha_2 a} - \alpha_3\right]. \end{aligned}$$

$$\begin{aligned} 1 - \pi(a) &= 1 - (1 - e^r), \qquad (4.33) \\ &= e^r. \end{aligned}$$

$$\begin{aligned} l(a) &= \frac{\pi'(a)}{1 - \pi(a)}, \qquad (4.34) \\ &= \frac{-e^r \times (-\alpha_1 a + \alpha_3)e^{-\alpha_2 a} - \alpha_3}{e^r}, \\ &= -[(-\alpha_1 a + \alpha_3)e^{-\alpha_2 a} - \alpha_3], \\ &= (\alpha_1 a - \alpha_3)e^{-\alpha_2 a} + \alpha_3. \end{aligned}$$

To ensure that the force of infection is always positive across all ages, Farrington constrained the parameter space to be non-negative. Thus,

$$\alpha_j \geq 0, j = 1, 2, 3$$

$$l(a_i) \geq 0, i = 1, 2, \ldots, n$$

Note that $\lim_{a\to\infty}[(\alpha_1 a - \alpha_3)e^{-\alpha_2 a} + \alpha_3] = \alpha_3$. Therefore, $l(a) \geq 0$ because $\alpha_3 \geq 0$ is constrained.

Farrington's model assumes that the force of infection at birth is zero, and then rises linearly to a peak, thereafter decreasing exponentially. When the contact rate between susceptible individuals and infectious individuals reaches a maximum, the peak in the model is obtained at the corresponding age.

The long-term residual value of the force of infection is represented by the parameter α_3. Under Farrington's two parameter model, $\alpha_3 = 0$. This results in the force of infection decreasing to zero as $a\to \infty$.

We can fit Farrington's prevalence model in SAS using the flexible nonlinear optimization procedure Proc NLMIXED which allows the user to specify their

own likelihood model. We fit Farrington's model with two parameters and three parameters and use model fit statistics to choose the best fitting model. Since no information on the parameters is given, we assume a minimal initial value for each. The initial values are determined through trial-and-error. Proc NLMIXED then provides us with parameter estimates which can be substituted into $l(a)$ to estimate the force of infection for any given age.

4.2.6 Fractional polynomials

Standard high order polynomials offer a variety of curves but have a weakness in that they fit the data badly at the extremes of the age scale. It also fits data poorly whenever asymptotic behaviour of the infection process is expected. Fractional polynomials allow for flexible changes in the force of infection since they extend standard polynomials to multiple non-integer powers.

In general, a fractional polynomial of degree m for the linear predictor is defined as

$$\eta_m(a, \underline{\beta}, p_1, p_2, \cdots, p_m) = \sum_{i=o}^{m} \beta_i H_i(a). \tag{4.35}$$

where $m \in Z$, $p_1 \leq p_2 \leq \ldots \leq p_m$ is a sequence of powers and $H_i(a)$ is a

transformation function given by

$$H_i(a) = \begin{cases} a^{p_i} & , \quad p_i \neq p_{i-1} \\ H_{i-1}(a)ln(a) & , \quad p_i = p_{i-1} \end{cases} \tag{4.36}$$

with initial conditions $p_0 = 1$ and $H_0 = 1$.

Royston and Altman (1994) argued that, in practice, fractional polynomials of order higher than 2 are rarely needed and suggested to choose the value of the powers from the set:

$$\{\text{-2,-1,-0.5,0,0.5,1,2,max(3,m)}\}.$$

First order fractional polynomials

$$\begin{aligned} \eta_1(a, \underline{\beta}, p) &= \sum_{i=0}^{1} \beta_i H_i(a), \\ &= \beta_0 H_0(a) + \beta_1 H_1(a), \\ &= \beta_0 + \beta_1 H_1(a). \end{aligned} \tag{4.37}$$

$$H_1(a) = \begin{cases} a^{p_1} & , \quad p_1 \neq p_0 \\ H_0(a)ln(a) & , \quad p_1 = p_0 \end{cases} \tag{4.38}$$

$$= \begin{cases} a_1^p & , \quad p_1 \neq 1 \\ ln(a) & , \quad p_1 = 1 \end{cases}$$

Therefore

$$\eta_1(a) = \begin{cases} \beta_0 + \beta_1 a^{p_1} & , \quad p_1 \neq 1 \\ \beta_0 + \beta_1 ln(a) & , \quad p_1 = 1 \end{cases} \tag{4.39}$$

Second order fractional polynomials

$$\begin{aligned} \eta_2(a, \underline{\beta}, p_1, p_2) &= \sum_{i=0}^{2} \beta_i H_i(a), \\ &= \beta_0 H_0(a) + \beta_1 H_1(a) + \beta_2 H_2(a), \\ &= \beta_0 + \beta_1 H_1(a) + \beta_2 H_2(a). \end{aligned} \tag{4.40}$$

$$H_2(a) = \begin{cases} a^{p_2} & , \quad p_2 \neq p_1 \\ H_1(a) ln(a) & , \quad p_2 = p_1 \end{cases} \tag{4.41}$$

Therefore

$$\eta_2(a) = \begin{cases} \beta_0 + \beta_1 a^{p_1} + \beta_2 a^{p_2} & , \quad p_2 \neq p_1 \\ \beta_0 + \beta_1 a^{p_1} + \beta_2 a^{p_1} ln(a) & , \quad p_2 = p_1 \end{cases} \tag{4.42}$$

Note

The constant, linear, Weibull and log-logistic force of infections can all be shown to be fractional polynomials.

Constant force of infection

$$g[\pi(a)] = \mu + ln(a)$$

is a first order fractional polynomial with $p = 0$,$\beta_0 = \mu$ and $\beta_1 = 1$.

Linear force of infection

The model with a linear force of infection can be parameterized as a first-order fractional polynomial with a complementary log-log link for which $p = 0$ and $\beta_1 = 2$. This implies that

$$\eta_1(a) = \beta_0 + \beta_1 ln(a) = \beta_0 + 2ln(a)$$

Thus

$$\eta_1'(a) = \frac{2}{a}$$

and

$$e^{\eta_1(a)} = e^{\beta_0 + 2ln(a)} = e^{\beta_0} a^2 = \kappa a^2$$

where $\kappa = e^{\beta_0}$. Therefore

$$l(a) = \eta_1'(a) e^{\eta_1(a)} = \frac{2}{a} \kappa a^2 = 2\kappa a$$

The form of $l(a)$ implies that the force of infection is zero at birth and thereafter, increases linearly.

Weibull force of infection

$$g[\pi(a)] = \mu + \beta ln(a)$$

is a first order fractional polynomial with $p = 0$, $\beta_0 = \mu$ and $\beta_1 \neq 1$.

Log-logistic force of infection

$$g[\pi(a)] = \mu + \beta ln(a)$$

is a first order fractional polynomial with $p = 0$, $\beta_0 = \mu$ and $\beta_1 \neq 1$.

Model selection

In general, we make use of the quantity $G(m, \underline{p})$ to measure the fit of different models. More specifically,

$$G(m, \underline{p}) = D(1, 1) - D(m, \underline{p}) \tag{4.43}$$

where $D(1, 1)$ is the deviance of the model with fractional polynomial of order 1 and power 1, and $D(m, \underline{p})$ is the deviance of the model with fractional

polynomial of order m and sequence of powers $\underline{p} = (p_1, p_2, \ldots, p_m)$.

Note that $D(1,1)$ is taken to be the reference or baseline deviance and we measure the improvement of other models upon this. The larger the value of G, the better the fit of the model.

To determine the most adequate order of the model, we begin by selecting the best first order fractional polynomial and the best second order fractional polynomial - the best model being determined by that which has the highest likelihood or, equivalently, the smallest deviance. The criterion

$$D(1, \tilde{\underline{p}}) - D(2, \tilde{\underline{p}}) > \chi^2_{2,0.9}$$

where $\tilde{\underline{p}}$ is the power sequence for the model with the best goodness of fit, is recommended by Royston and Altman as a decision tool to select between the first and second order fractional polynomial models. The first order fractional polynomial model is rejected if the above criterion is met.

Constrained fractional polynomials

Regardless of the variety of curve slopes that fractional polynomials offer, there is still no guarantee that $\pi(a)$ will be a monotone function of age. This can result in a negative estimate for the force of infection. From Table 4.1 we can see that the force of infection is negative whenever $\eta'(a)$ is negative

(since $\delta[\eta(a)]$ is strictly positive). Thus, our models should be fitted subject to the constraint

$$\eta'_m(a, \hat{\beta}, \underline{p}) \geq 0$$

for all ages a in the predefined range. In practice, we can fit a large number of fractional polynomials over a grid of powers, and check if the above constraint is met for all ages a. If a given sequence of powers leads to a negative derivative of the linear predictor, the model is considered inappropriate. This implies that the model with the best goodness-of-fit amongst all fractional polynomials for which $\eta'_m(a, \hat{\beta}, \underline{p}) \geq 0$ is selected.

4.2.7 Monotone local polynomials

Parametric models in literature require assumptions about the parametric structure. This restrains the linear predictor thus detracting from the true shape of the estimated force of infection. Nonparametric models assume only smoothness for the prevalence or the force of infection. Local polynomials:

- estimate the prevalence and force of infection simultaneously resulting in a smooth estimated probability curve.
- have the advantageous property of automatic boundary correction.

- are fully unconstrained and highly data driven resulting in a revelation of data aspects that was previously ignored or hidden by parametric models.

Monotone local polynomials is a nonparametric method that can be used to estimate the force of infection. We know that within the fractional polynomial framework

$$\eta(a) = g[\pi(a)] \Rightarrow \pi(a) = g^{-1}[\eta(a)]$$

where $\eta(a)$ is the linear predictor having a flexible parametric structure. Within the local polynomial framework, we assume the same structure with the exception of not specifying a parametric structure for $\eta(a)$. This method accommodates for the simultaneous nonparametric estimation of the force of infection and prevalence. At certain ages, however, negative estimates of the force of infection may arise as a result of the unconstrained model. Thus we make use of a local smoother, constrained to be monotone, to ensure maximum flexibility in addition to overcoming the negative force of infection estimate problem.

As shown before, when observing the age of infection of an age-specific cross-sectional prevalence sample, the log-likelihood is given by

$$L = \sum_{i=1}^{N} Q_i \left\{ Y_i, g^{-1}[\eta(a_i)] \right\} \tag{4.44}$$

where Q_i is the contribution of the i^{th} subject to the Bernoulli log-likelihood with success probability $\pi(a_i) = g^{-1}[\eta(a_i)]$. As shown before the force of infection can be expressed as

$$l(a) = \eta'(a)\delta[\eta(a)] \tag{4.45}$$

where δ is determined by the link function g. The choice of link function when using a local polynomial likelihood method, however, is less important. The local polynomial likelihood method provides consistent estimates for $\eta(a)$ and $\eta'(a)$, without any parametric restriction on the functional form. They only have to satisfy some smoothness condition. Thus, for a given link function, the local force of infection can be found using equation (4.45).

The local likelihood estimation is based on the maximization of

$$\sum_{i=1}^{n} Q_i \left\{ Y_i, g^{-1}[\eta(a_i - a)] \right\} K[\frac{(a_i - a)}{h}],$$

where K is a kernel, assigning higher weights to data points in the neighbourhood of some fixed a, and h is a bandwidth parameter. The linear predictor

is locally approximated by a polynomial of order p.

$$\begin{aligned}\eta(a_i - a) &\approx \eta(a) + \eta'(a)(a_i - a) + \ldots + \eta^{(p)}(a_i - a)[\frac{(a_i - a)^p}{p!}], \quad (4.46)\\ &= \beta_0(a) + \beta_1(a)(a_i - a) + \ldots + \beta_p(a_i - a)^r,\\ &= \sum_{r=0}^{p} \beta_r(a)(a_i - a)^r.\end{aligned}$$

Let us consider the first order Taylor expansion for the linear predictor. Thus

$$\begin{aligned}\eta(a_i - a) &\approx \eta(a) + \eta'(a)(a_i - a), \quad (4.47)\\ &= \beta_0(a) + \beta_1(a)(a_i - a).\end{aligned}$$

Hence, the linear predictor and the first derivative can be estimated as follows:

$$\eta(a) = \hat{\beta}_0(a), \quad (4.48)$$

$$\eta'(a) = \hat{\beta}_1(a). \quad (4.49)$$

Table 4.2 demonstrates how we can estimate the force of infection using local polynomials, for a given link function.

For each value of a, the estimation of $\beta_r(a)$ must be repeated. The choice of kernel is less important but conventional choices are the symmetrical beta family given by

Table 4.2: Table of local estimates for $l(a)$ for a given link function

Link	$\pi(a)$	$\delta[\eta(a)]$	Local estimate for $l(a)$
log	$1-e^{-\eta(a)}$	1	$\hat{\beta_1(a)}$
clog-log	$1-e^{-e^{\eta(a)}}$	$\mathrm{e}^{\eta(a)}$	$\hat{\beta_1(a)e^{\beta_0(a)}}$
logit	$\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$	$\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$	$\hat{\beta_1(a)}\frac{e^{\hat{\beta_0}(a)}}{1+e^{\hat{\beta_0}(a)}}$

$$K(u) = \frac{(1-u^2)^\gamma}{\beta(0.5, \gamma+1)}, \tag{4.50}$$

for $|u| \leq 1$, $\gamma = 0, 1, 2, \ldots$ and the Gaussian kernel given by

$$K(u) = \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}}. \tag{4.51}$$

However, the choice of the smoothing parameter h is crucial. The optimal local bandwidth h is the minimizer of the average mean square error (AMSE) for the force of infection.

Chapter 5

Application and Results

As mentioned before, our dataset consists of left- and right-censored observations. These are dealt with by using the cumulative and survivor functions which in our data, correspond to a prevalent and susceptible event respectively. Ages of the women attending the antenatal clinics in Vulindlela ranged from 14 to 45 years. To ensure that the models fitted as best as possible, the ages were scaled such that the minimum age of 14 years would be represented by the value 0.5. Hence, the maximum age of 45 years is represented by the value 31.5. The minimum age is represented by 0.5 instead of 0 so as to accommodate for the function $ln(a)$ that occurs in the Weibull and log-logistic models ($ln(a)$ is undefined for $a = 0$). The models were first fitted to the

2009 data and then to the 2010 data. These are the most recent data sets obtained from CAPRISA on HIV sero-prevalence in their Vulindlela research sites.

SAS was used to obtain parameter estimates for each model then MATLAB was used to plot the prevalence and force of infection functions.

5.1 Constant force of infection

The constant force of infection model was fitted using the GENMOD procedure with a complementary log-log link function. This procedure revealed estimates such that

$$\pi(a)_{2009} = 1 - e^{-0.0523a}, \tag{5.1}$$

$$\pi(a)_{2010} = 1 - e^{-0.0536a}, \tag{5.2}$$

and

$$l(a)_{2009} = 0.0523, \tag{5.3}$$

$$l(a)_{2010} = 0.0536. \tag{5.4}$$

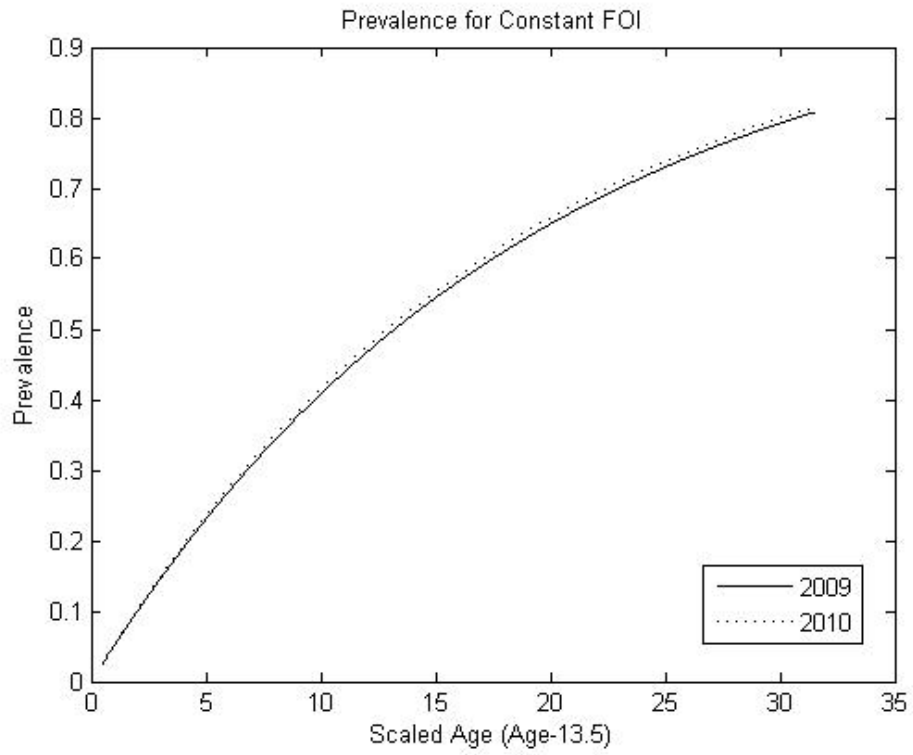

Figure 5.1: The fitted prevalence function for the constant model.

Figure 5.1 shows us that the prevalence of HIV among the pregnant women in Vulindlela peaked at 81% in 2009 and 82% in 2010. These figures do not seem realistic. This may indicate that a constant force of infection is very unrealistic. Prevalence rises as age increases since $\pi(a)$ is a cumulative probability. The force of infection as can be seen in figure 5.2 for 2009 is

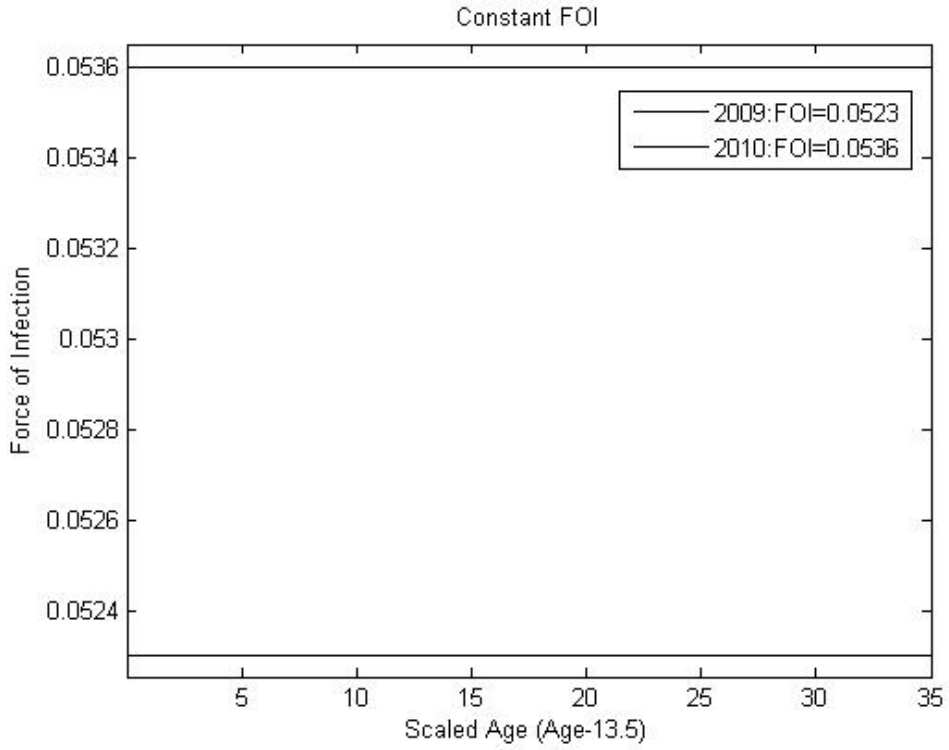

Figure 5.2: The fitted force of infection for the constant model.

a constant 0.0523 across all ages, while for 2010 there was a slight increase to 0.0536. Under the constant force of infection these rates imply that only approximately 5 per every 100 individuals will be newly infected at each age level. For a widely spread disease, this figure seems to be underestimated. Also, it would be disadvantageous to assume a constant rate of new infections regardless of age.

5.2 Linear force of infection

The NLMIXED procedure was used to fit the linear force of infection. Assuming initial values $\beta_0 = \beta_1 = \beta_2 = 0.001$ (note that the β's can each assume different initial values as per trial-and-error), estimates produced were such that

$$\pi(a)_{2009} = 1 - e^{-(-0.1067+0.0973a-0.0024a^2)}, \tag{5.5}$$

$$\pi(a)_{2010} = 1 - e^{-(-0.1414+0.1159a-0.0033a^2)}, \tag{5.6}$$

and

$$l(a)_{2009} = 0.0973 - 0.0048a, \tag{5.7}$$

$$l(a)_{2010} = 0.1159 - 0.0066a. \tag{5.8}$$

From Figure 5.3, we can see that the prevalence only becomes significant after the age of 15 for both 2009 and 2010. The prevalence increases monotonically and reaches its peak of 58.5% during the early thirties for the year 2009. For 2010, prevalence peaks to 58.4% during the early thirties as well. The estimated prevalence for the total sample is 43.8% for 2009 and 20.9%

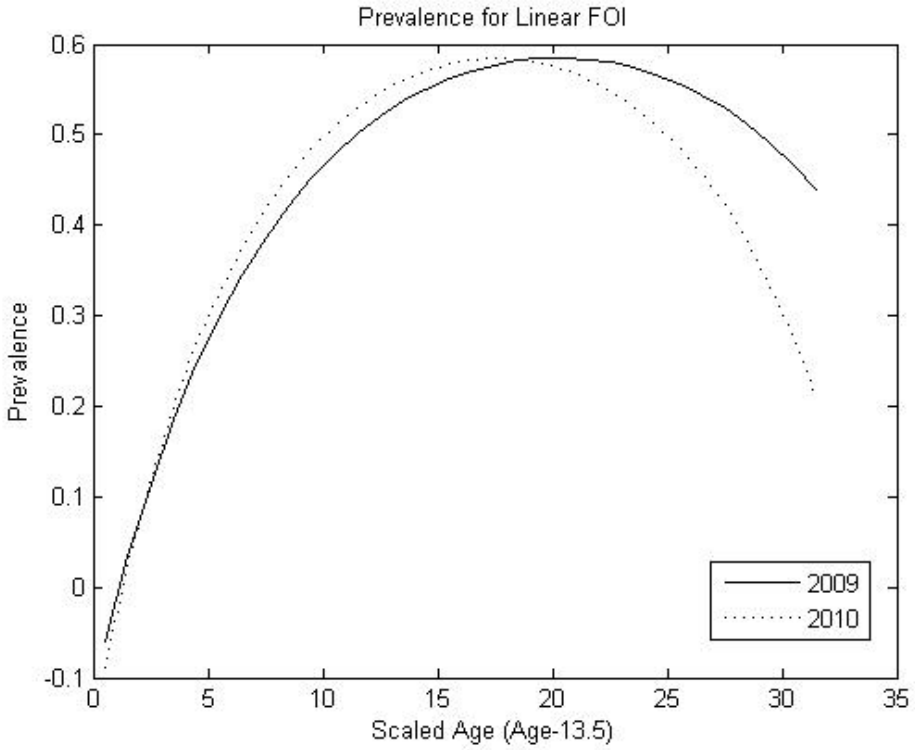

Figure 5.3: The fitted prevalence function for the linear model.

for 2010. Compared to the estimates produced under the constant force of infection model, these estimates make more sense.

Figure 5.4 shows the plot of the force of infection under the linear model. Immediately, we notice the force of infection to be a declining function. However, after the ages of 33 and 31 for years 2009 and 2010 respectively, the force of infection becomes negative. This suggests that the rate of new infections beyond these ages is too insignificant to contribute to prevalence.

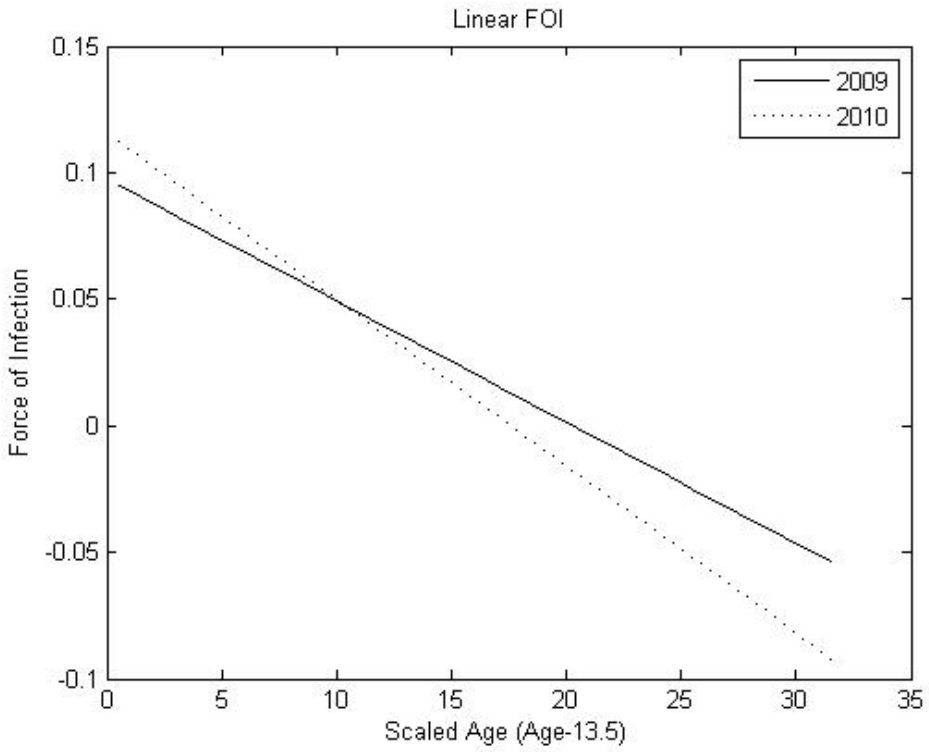

Figure 5.4: The fitted force of infection function for the linear model.

5.3 Weibull force of infection

We fit the Weibull model in SAS using the GENMOD procedure with a complementary log-log link function. The procedure reveals estimates of the functional forms of age dependent prevalence and force of infection as

$$\pi(a)_{2009} = 1 - e^{-0.0865a^{0.7916}}, \tag{5.9}$$

$$\pi(a)_{2010} = 1 - e^{-0.0874a^{0.8023}}, \tag{5.10}$$

and

$$l(a)_{2009} = 0.0685a^{-0.2084}, \tag{5.11}$$

$$l(a)_{2010} = 0.0701a^{-0.1977}. \tag{5.12}$$

The plotted prevalence in Figure 5.5 reveals an increasing prevalence as age increases, which is expected. The prevalence reaches a maximum of 73% in 2009 and 75% in 2010. These prevalence figures are less than that of the figures produced under the constant model but larger than the figures produced under the linear model. Figure 5.6 reveals the plotted force of infection. The force of infection has the pattern of decreasing as age increases. For 2009, the force of infection decreases from 0.079 at age 14 to 0.033 at age 45. For 2010, the force of infection decreases from 0.08 at age 14 to 0.035 at age 45. As age increases, the slope of the Weibull force of infection becomes flatter.

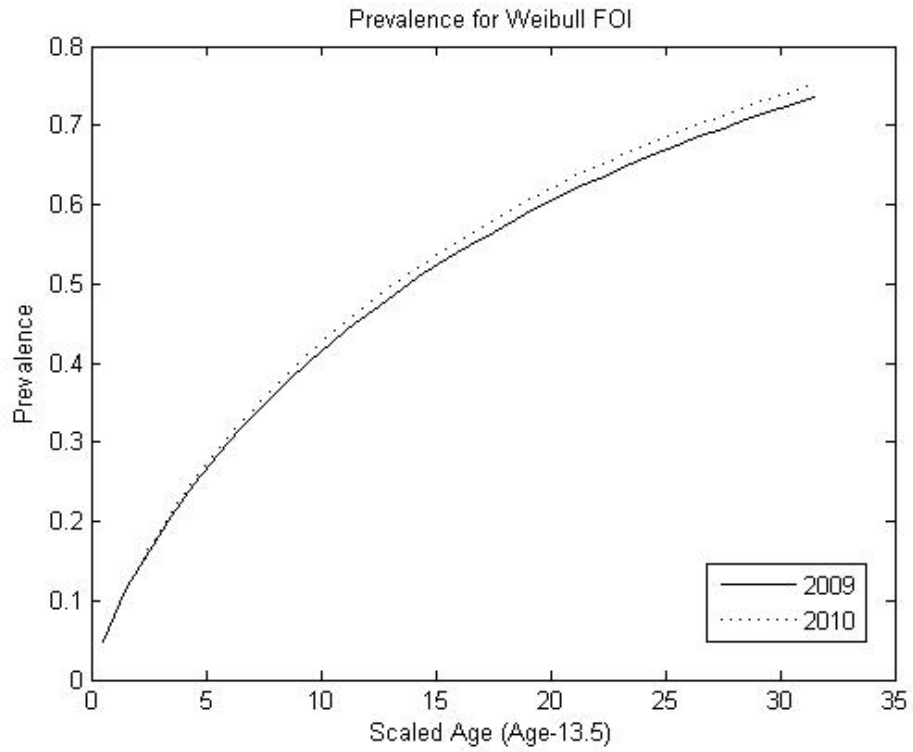

Figure 5.5: The fitted prevalence for the Weibull model.

5.4 Log-logistic force of infection

The GENMOD procedure with a logit link function was used to fit the log-logistic model. Estimates produced were such that

$$\pi(a)_{2009} = \frac{0.0621a^{1.0777}}{1 + 0.0621a^{1.0777}}, \tag{5.13}$$

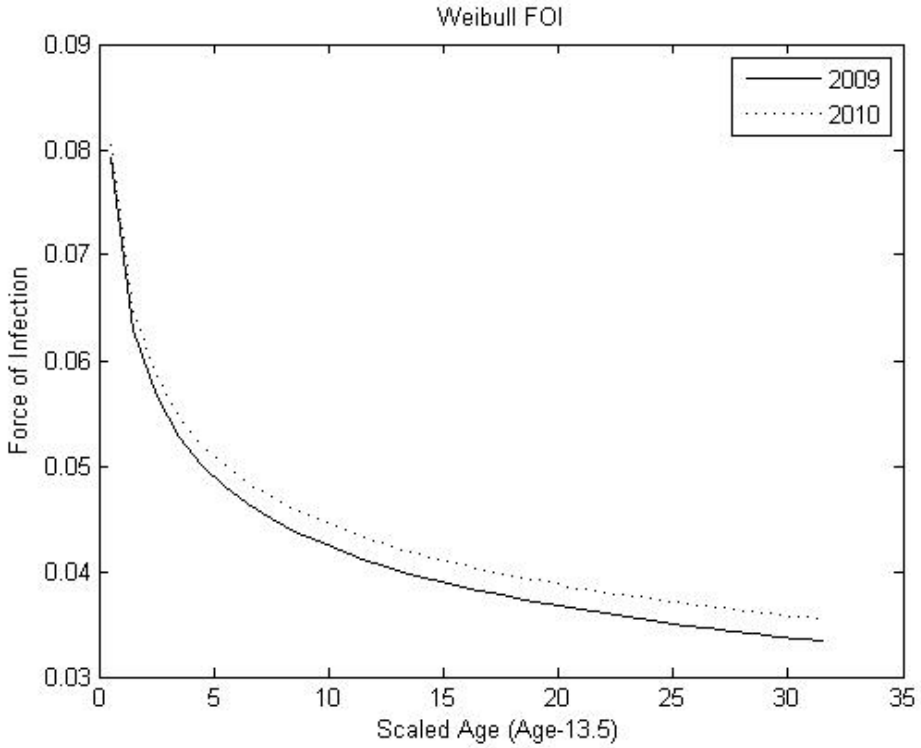

Figure 5.6: The fitted force of infection for the Weibull model.

$$\pi(a)_{2010} = \frac{0.0597a^{1.1188}}{1 + 0.0597a^{1.1188}}, \tag{5.14}$$

and

$$l(a)_{2009} = \frac{0.0669a^{0.0777}}{1 + 0.0621a^{1.0777}}, \tag{5.15}$$

$$l(a)_{2010} = \frac{0.0668a^{0.1188}}{1 + 0.0597a^{1.1188}}. \tag{5.16}$$

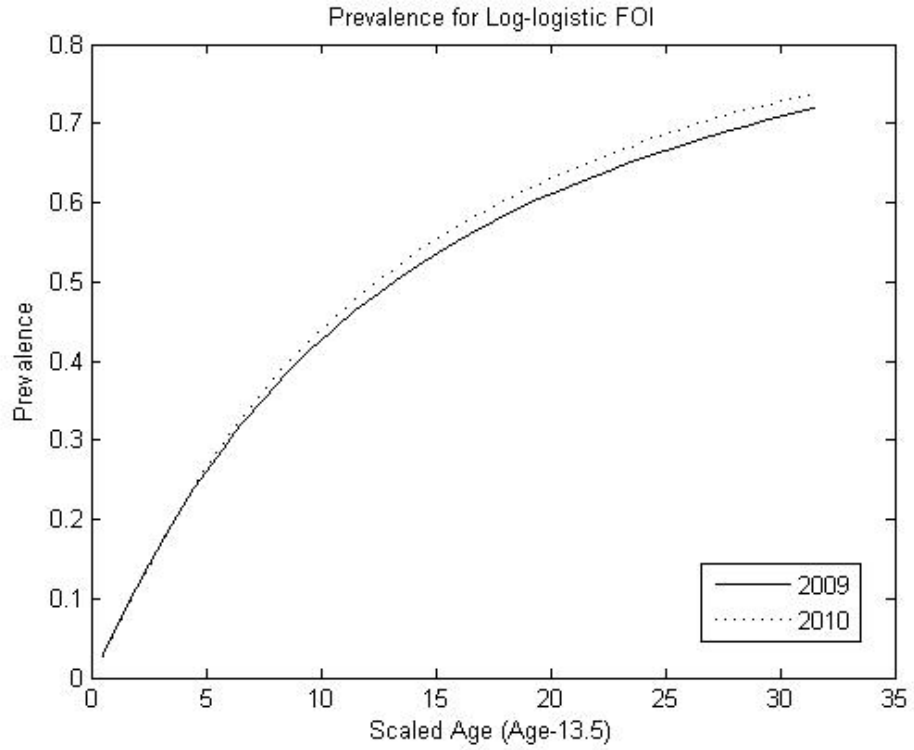

Figure 5.7: The fitted prevalence for the log-logistic model.

The estimated prevalence function under the log-logistic model, as plotted in Figure 5.7, closely resembles the estimated prevalence function under the Weibull model (Figure 5.5). The prevalence rate, as per Figure 5.7, of HIV is estimated to be 72% in 2009 and 74% in 2010. The log-logistic force of infection as plotted in Figure 5.8, rises to a peak and then proceeds to decline

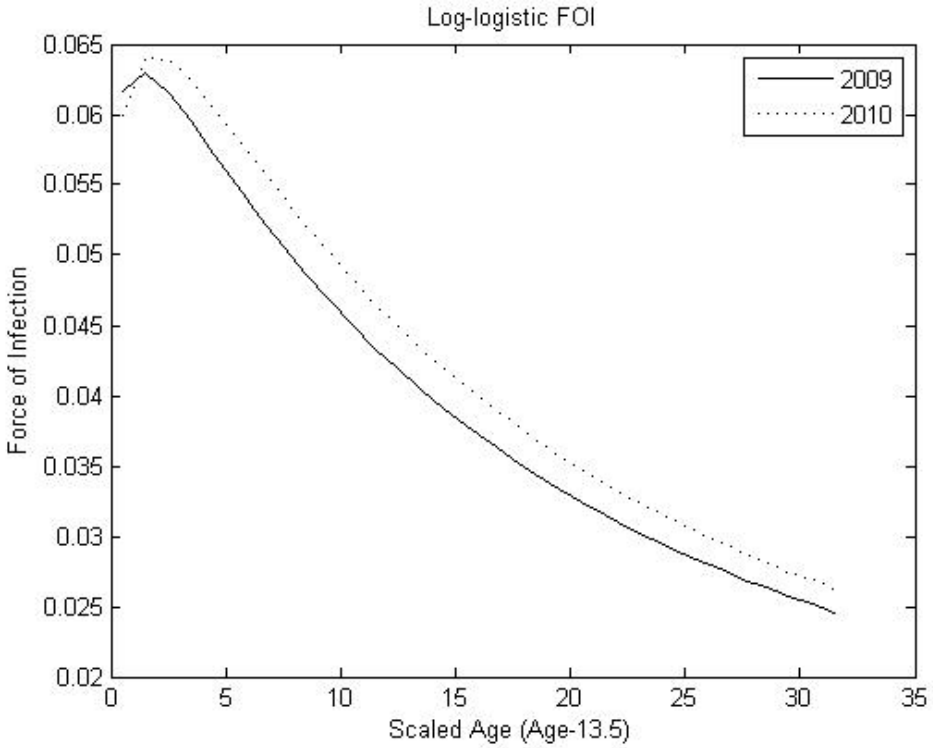

Figure 5.8: The fitted force of infection for the log-logistic model.

with increasing age. In 2009, the graph shows that the force of infection reached a maximum of 0.063 at the age of 15. In 2010, the force of infection reached its peak of 0.064 at the age of 15.5. As Figure 5.8 shows, the log-logistic model assumes a decreasing rate of new infections as age increases.

5.5 Farrington's force of infection

Farrington's two parameter model and Farrington's three parameter model were fitted using the NLMIXED procedure, with the default Dual-Quasi Newton optimization technique assuming

$$\alpha_1 = \alpha_2 = \alpha_3 = 0.1,$$

for the three parameter model, and

$$\alpha_1 = \alpha_2 = 0.1,$$

$$\alpha_3 = 0,$$

for the two parameter model. Note that the α's can each assume different initial values.

5.5.1 Farrington's two parameter model

The estimates for the two parameter model were iteratively obtained such that

$$\pi(a)_{2009} = 1 - e^{0.2139ae^{-0.2324a} + 0.9202(e^{-0.2324a} - 1)}, \tag{5.17}$$

$$\pi(a)_{2010} = 1 - e^{0.2285ae^{-0.2394a} + 0.9544(e^{-0.2394a} - 1)}, \tag{5.18}$$

and

$$l(a)_{2009} = 0.0497ae^{-0.2324a}, \tag{5.19}$$

$$l(a)_{2010} = 0.0547ae^{-0.2394a}. \tag{5.20}$$

Figure 5.9 shows a fitted prevalence function that increases monotonically as age increases and then proceeds to flatten out at higher ages. The prevalence of HIV infection among the women attending antenatal clinics in Vulindlela was 60% in 2009 and increased to 61% in 2010, the maximum values of prevalence in Figure 5.9. Figure 5.10 shows the estimated force of infection for Farrington's two parameter model. The general function rises to a peak and thereafter declines until it flattens out at higher ages. In 2009, the estimated force of infection reached a peak of 0.079 around the age of 18 and decreased to almost zero in the mid-forty age group. In 2010, the estimated force of infection reached a peak of 0.084 around the age of 18 as well and thereafter declined to close to zero in the mid-forty age group.

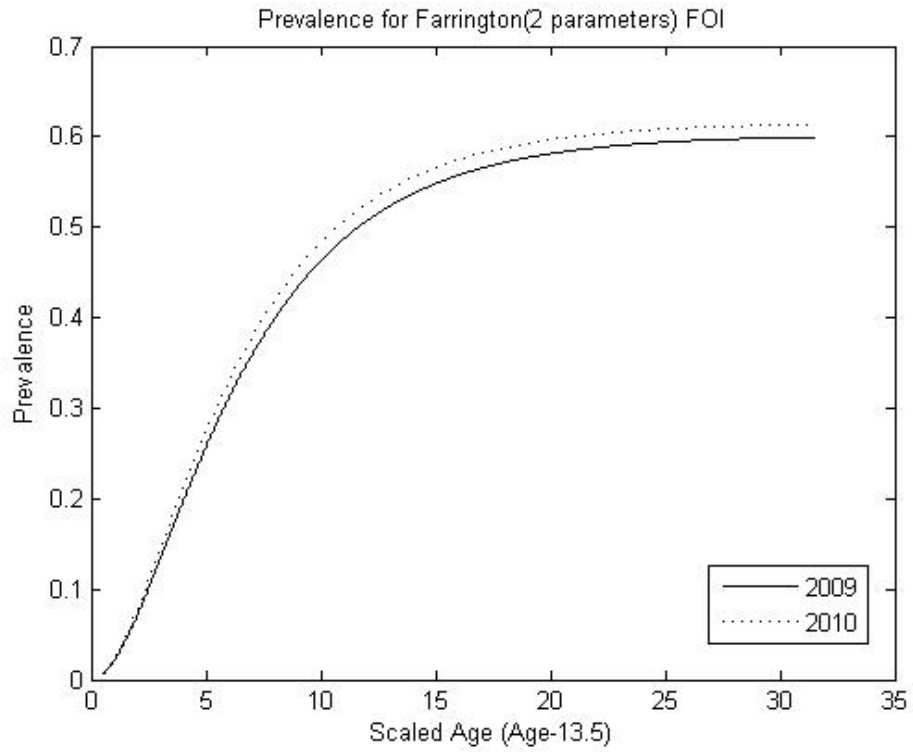

Figure 5.9: The fitted prevalence for Farrington's two parameter model.

5.5.2 Farrington's three parameter model

The estimates for the three parameter model were iteratively obtained such that

$$\pi(a)_{2009} = 1 - e^{0.3342ae^{-0.1457a} + 2.727(e^{-0.1475a} - 1) + 0.068a}, \tag{5.21}$$

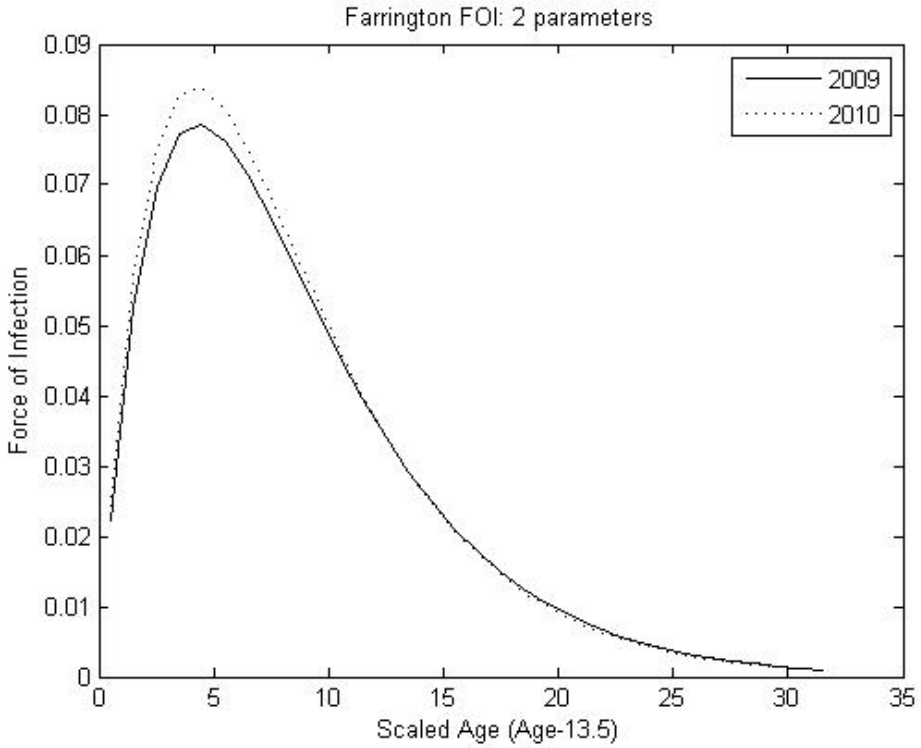

Figure 5.10: The fitted force of infection for Farrington's two parameter model.

$$\pi(a)_{2010} = 1 - e^{0.4876ae^{-0.1167a} + 5.5079(e^{-0.1167a} - 1) + 0.1552a}, \tag{5.22}$$

and

$$l(a)_{2009} = (0.0493a + 0.068)e^{-0.1475a} - 0.068, \tag{5.23}$$

$$l(a)_{2010} = (0.0569a + 0.1552)e^{-0.1167a} - 0.1552. \tag{5.24}$$

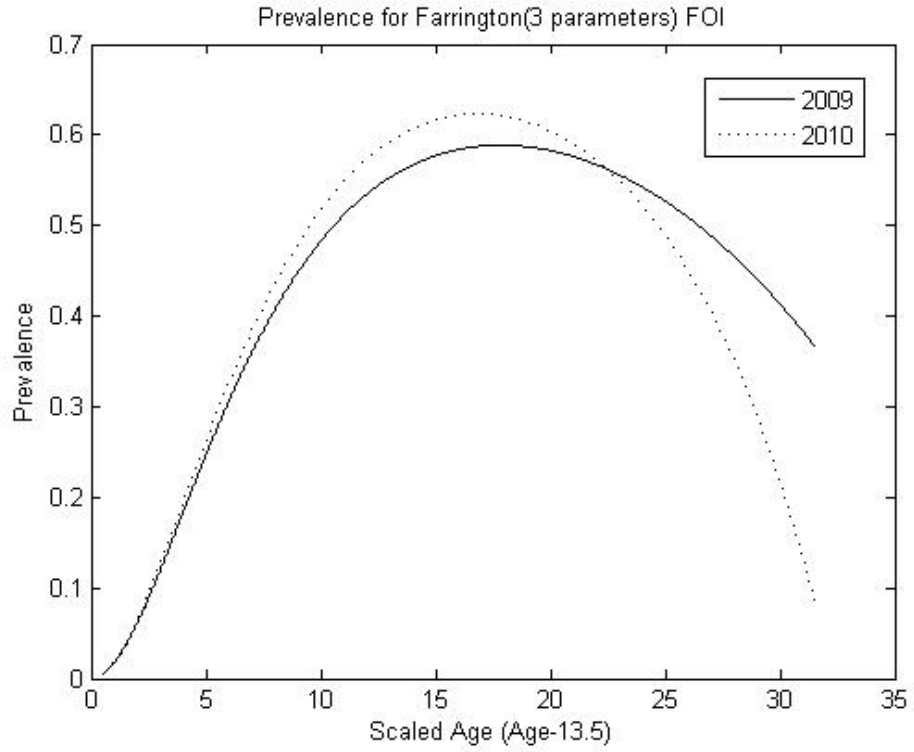

Figure 5.11: The fitted prevalence for Farrington's three parameter model.

The plotted prevalence under Farrington's three parameter model is shown in Figure 5.11. It shows the prevalence increasing from almost zero to a peak and thereafter decreasing, as age increases. For 2009, the peak of 58.87% is reached around the age of 31, while for 2010, the peak of 62.17% is reached around age 30. The estimated prevalence for the total sample was 36.73%

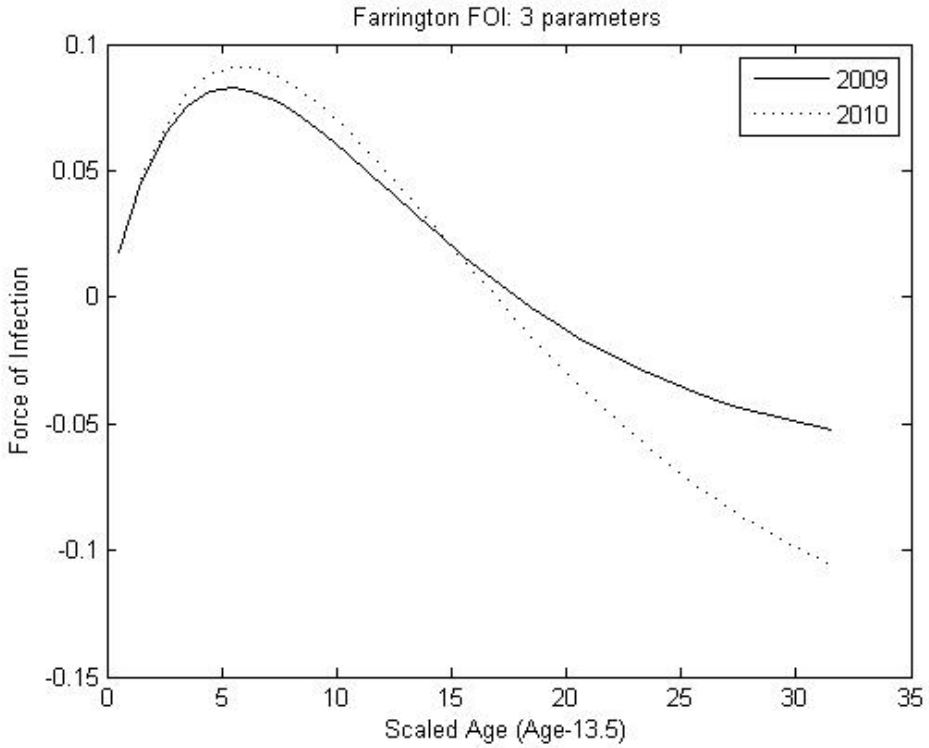

Figure 5.12: The fitted force of infection for Farrington's three parameter model.

for the year 2009. However, for 2010, we find that the prevalence is grossly underestimated to be 8.67%.

We find, in Figure 5.12 Farrington's three parameter force of infection rising from almost zero to a peak and thereafter decreasing as age increases. The force of infection reaches a peak of 0.083 at age 19 for 2009. For 2010, the peak of 0.091 is reached around the age of 19 as well. We find that beyond age

30, the force of infection becomes negative. Thus, the rate of new infections beyond the age of 30 become insignificant. This is a worrying prediction as it is a productive age. It further suggests that Farrington's three parameter model may not realistically capture this particular force of infection.

5.6 Fractional polynomial force of infection

A variety of first order fractional polynomials and a variety of second order fractional polynomials was fitted using a SAS MACRO containing the GENMOD procedure. The power sequence and the link used was varied. The best fitting polynomial for each order was then noted.

Table 5.1: Table of best fitting fractional polynomials

Year	First and Second Order Fractional Polynomial	Link	Deviance
2009	$\eta_1(a) = 1.747 - 6.1826a^{-0.5}$	logit	56.0457
	$\eta_2(a) = -16.8353 + 11.4565a^{-0.2} + 6.3036a^{-0.2}ln(a)$	clog-log	54.7315
2010	$\eta_1(a) = 0.2336 - 6.5838a^{-1}$	clog-log	26.1467
	$\eta_2(a) = -2.0175 - 5.1644a^{-0.5} + 4.5238a^{-0.5}ln(a)$	clog-log	24.6872

Remembering that the first order fractional polynomial is rejected in favour

of the second order if the criterion

$$D(1,\underline{\tilde{p}}) - D(2,\underline{\tilde{p}}) > \chi^2_{2,0.9}$$

is met, we find that the second order fractional polynomial is the most adequate for both 2009 and 2010. By using Tables 4.2 and 5.1, we can find the prevalence and force of infection. We find the prevalence to be

$$\pi(a)_{2009} = 1 - e^{-u}, \tag{5.25}$$

where $u = e^{-16.8353+11.4565a^{-0.2}+6.3036a^{-0.2}ln(a)}$;

$$\pi(a)_{2010} = 1 - e^{-v}, \tag{5.26}$$

where $v = e^{-2.0175-5.1644a^{-0.5}+4.5238a^{-0.5}ln(a)}$;

and the force of infection to be

$$l(a)_{2009} = (4.0123a^{-1.2} - 1.2607a^{-1.2}ln(a))e^{-16.8353+11.4565a^{-0.2}+6.3036a^{-0.2}ln(a)}, \tag{5.27}$$

$$l(a)_{2010} = (7.106a^{-1.5} - 2.2619a^{-1.5}ln(a))e^{-2.0175-5.1644a^{-0.5}+4.5238a^{-0.5}ln(a)}. \tag{5.28}$$

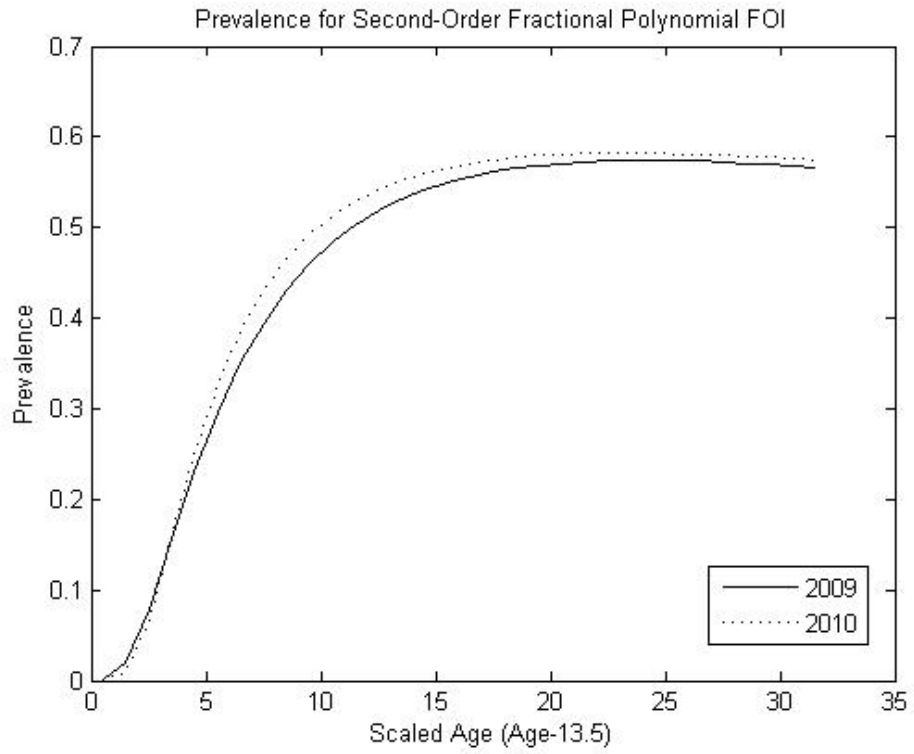

Figure 5.13: The fitted prevalence for the second-order fractional polynomial model.

From Figure 5.13, we can immediately see that the fractional polynomial prevalence function closely resembles Farrington's two parameter prevalence function (Figure 5.9). Both increase monotonically as age increases and

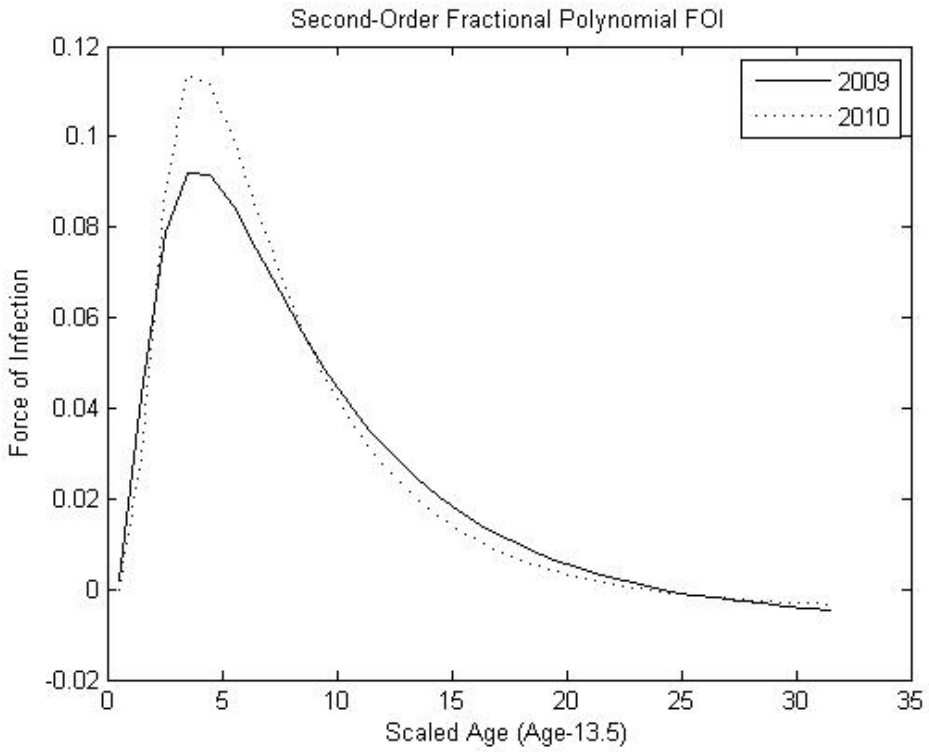

Figure 5.14: The fitted force of infection for the second-order fractional polynomial model.

flattens out at higher ages. Figure 5.13 shows the prevalence of HIV in this community was 57% in 2009 and increased slightly to 58% in 2010. These estimates are lower than that produced using Farrington's two parameter model.

Figure 5.14 shows the estimated force of infection for the fractional polynomial model. Again we see the resemblance when compared to Farrington's

two parameter force of infection function (Figure 5.10). The function of Figure 5.14 rises to a peak and thereafter declines until it flattens out at higher ages. The estimated force of HIV infection reached a peak of 0.0929 in 2009 and increased to 0.1148 in 2010 - both around the age of 17. After the age of 38, we find that the force of infection becomes negative suggesting that after this age, the rate of new infections is too insignificant to contribute to prevalence.

5.7 Fit statistics

Tables 5.2 and 5.3 show the fit statistics of each force of infection model. For 2009, the AIC value of 117.4 (smaller is better) determines that both Farrington's two and three parameter models are the best fitting models. Farrington's three parameter model for the force of infection (Figure 5.12), however, does not fall within our desired range (we desire both the prevalence and force of infection to be positive). Hence, we will rule out Farrington's three parameter model and select Farrington's two parameter model as our best fitting model. Farrington's two parameter model realistically captures the prevalence and force of infection.

For 2010, the AIC value of 95.3 determines Farrington's three parameter model as the best fitting model. However, we once again find the estimated force of infection under this model to be negative at higher age groups (Figure 5.12). Thus, we will disregard Farrington's three parameter model and accept our second best fitting model (as determined by the AIC value of 98.0), Farrington's two parameter model, as our best fitting model.

To emphasise, the force of infection is by definition a non-negative quantity. Therefore, the linear model together with Farrington's three parameter model, which lead negative estimates of the force of infection, are strictly not suitable.

5.8 Evolution of HIV infection in the Vulindlela district

As determined by our previous analysis of data from the years 2009 and 2010, we find that Farrington's two parameter model best fitted both sets of data. Using this model, we can now analyse a larger dataset to discover trends

Table 5.2: Table of fit statistics for force of infection models fitted to the 2009 data

Model	Deviance(df)	Pearson Chi-Square	Log Likelihood	AIC
Constant	63.2071 (29)	54.6925 (29)	-229.8957	124.4049
Linear	-	-	-	118.7
Weibull	60.7889 (28)	49.9468 (28)	-228.6866	123.9866
Log-logistic	58.7540 (28)	48.1866 (28)	-227.6692	121.9517
Farrington(2)	-	-	-	117.4
Farrington(3)	-	-	-	117.4
Frac. Polynomial	54.7315 (27)	44.6818 (27)	-225.6579	119.9292

occurring in the prevalence and force of HIV infection. More specifically, we look at data ranging from the year 2002 through to 2010. Figures 5.15 and 5.16 depict the cumulative prevalence and force of infection estimates that were obtained respectively.

After applying Farrington's two parameter model to the increased dataset, we observe three key statistics: the prevalence, the peak of the estimated force of infection and the age at which the force of infection peaks. Note that Table 5.4 shows the true age at which the force of infection peaks and not the scaled age.

Table 5.3: Table of fit statistics for force of infection models fitted to the 2010 data

Model	Deviance(df)	Pearson Chi-Square	Log Likelihood	AIC
Constant	37.3420 (28)	38.9755 (28)	-208.5575	109.0099
Linear	-	-	-	98.2
Weibull	35.1002 (27)	33.8385 (27)	-207.4366	108.7680
Log-logistic	31.9906 (27)	31.2173 (27)	-205.8818	105.6585
Farrington(2)	-	-	-	98.0
Farrington(3)	-	-	-	95.3
Frac. Polynomial	24.6872 (26)	26.8188 (26)	-202.2301	100.3551

Using Table 5.4, we can make use of line graphs to enable us to clearly see trends in these statistics over the years.

Immediately from Table 5.4, we can see that the estimates for the year 2002, are considerably lower than the estimates produced for the following years. We can attribute this inaccuracy to the small sample size available for that year. Figure 5.17 shows the prevalence under Farrington's two parameter model over the years. Prevalence increased in general from 54% in 2003 to 61% in 2010. It can also be seen that the prevalence decreased between 2004

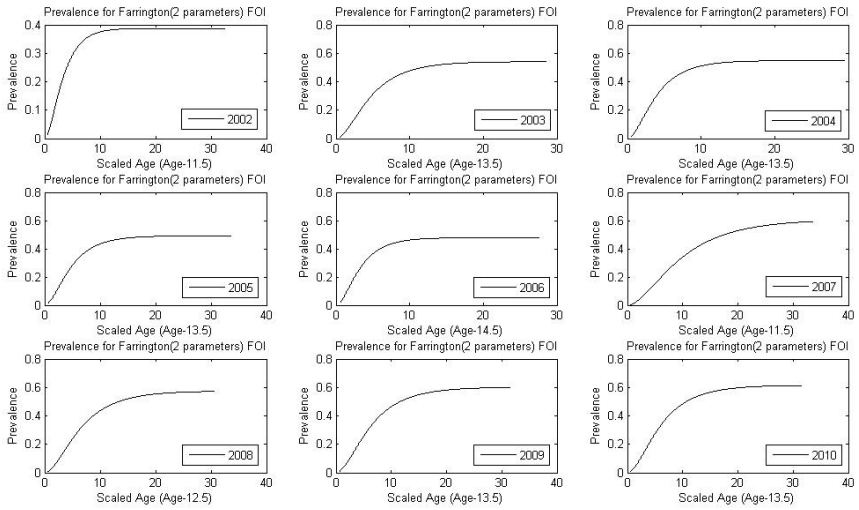

Figure 5.15: The cumulative prevalence estimates under Farrington's two parameter model for the years 2002 to 2010.

and 2006.

Figure 5.18 shows the estimated force of infection peaks over the years. The graph has a decreasing trend (0.0893 in 2002 to 0.0838 in 2010). Despite the occurring increase in prevalence, the rate of new HIV infections is decreasing. This may imply successful HIV prevention strategies and programmes.

Finally, Figure 5.19 depicts the age at which the force of infection peaked for each year. The general trend is increasing (16.5 years in 2003 to 18 in 2010). The increasing trend is desirable, in a sense, as it suggests that the risk of HIV infection is highest at an older age group rather than at a previously

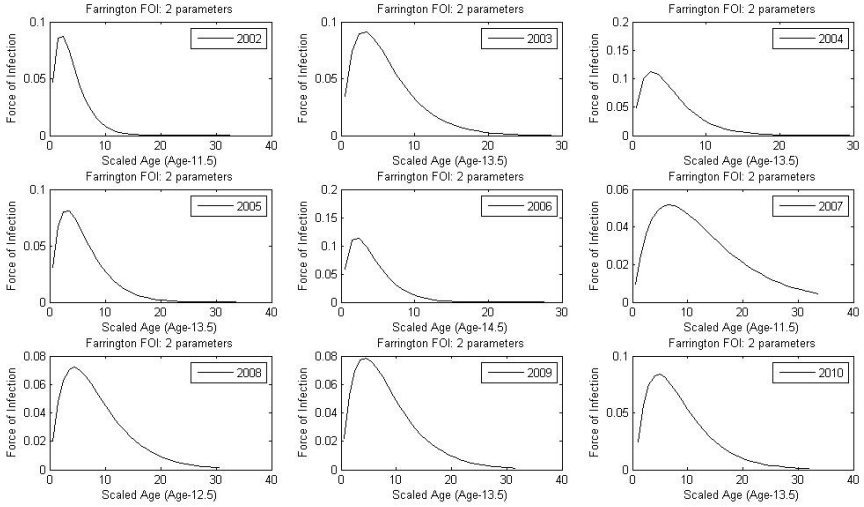

Figure 5.16: The force of infection estimates under Farrington's two parameter model for the years 2002 to 2010.

younger group.

Determining the best strategy for measuring incidence remains a challenge. The prospective follow-up of a cohort of HIV-negative individuals provides a direct estimate of HIV incidence; however, such studies are expensive, challenging, resource-consuming and raise the ethical dilemma of collecting data from cohorts without implementing interventions. Furthermore, the enrolment of individuals into a cohort study often leads to behavioural changes that result in a lower observed HIV incidence than in the wider population

Table 5.4: Table of observed key statistics gathered from the plotted estimated prevalences and forces of infection under Farrington's model.

Year	Prevalence (%)	Peak of FOI	Age of Peak
2002	38.75	0.0893	13.5
2003	54.07	0.0918	16.5
2004	54.86	0.1127	16
2005	49.12	0.0818	16.5
2006	47.81	0.1152	16.5
2007	59.32	0.0517	18
2008	57.2	0.072	17
2009	59.95	0.0786	18
2010	61.33	0.0838	18

of interest. Cross-sectional, age-specific prevalence data provides valuable age-specific incidence estimates as an alternative to measuring the incidence of HIV infection directly from cohort studies.

The most rudimentary of the statistical models that have been developed to estimate incidence from cross-sectional age-prevalence surveys assume that HIV incidence rates in the population are stable over time and that the preva-

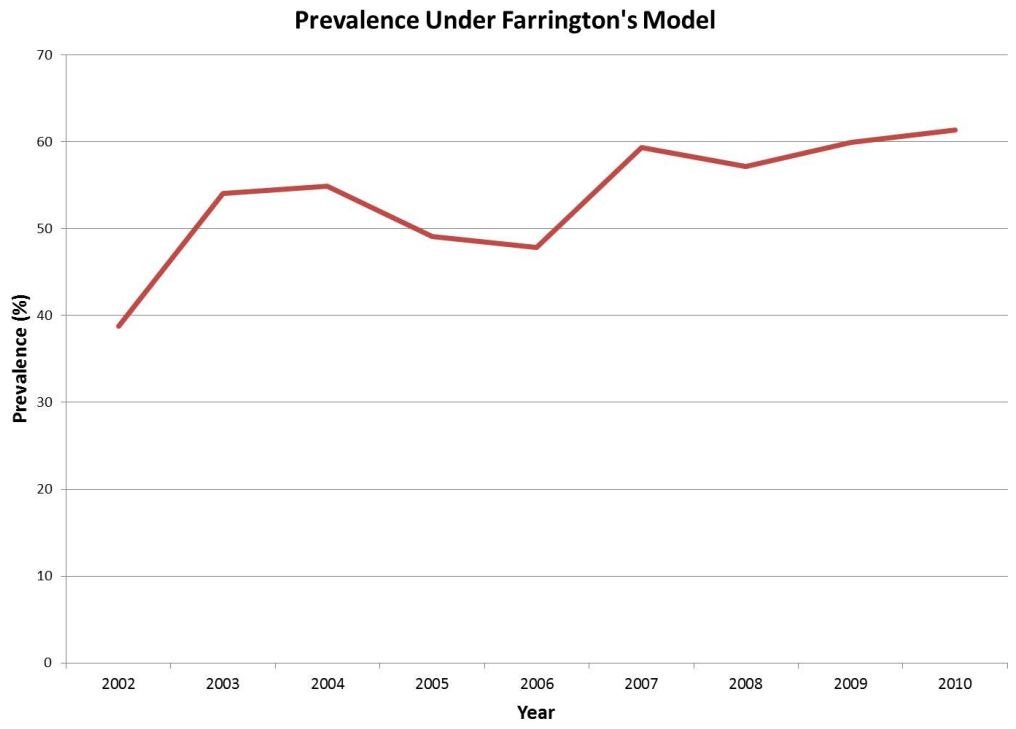

Figure 5.17: The estimated Farrington's prevalence of HIV amongst pregnant women in Vulindlela by year.

lence in young people increases linearly with age. Based on this assumption, we can apply Farrington's model to the area HIV because of the distinctive property of the Farrington force of infection rising to a peak and declining until flattening out at higher ages. However, Farrington applied his model to measles, mumps and rubella - all being reversible infections. Podgor and Leske (1986) developed a method for estimating the incidence of irreversible

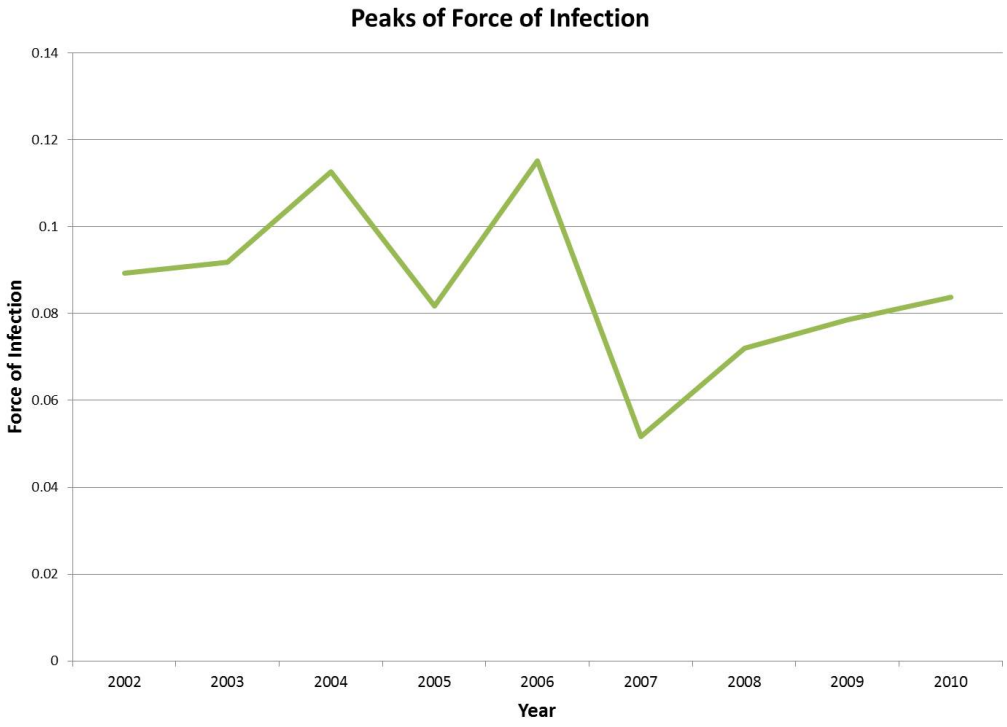

Figure 5.18: The estimated Farrington's force of HIV infection amongst pregnant women in Vulindlela by year.

diseases (eg. HIV) from age-specific prevalence data. Their method adjusts for differential mortality, assumes that the force of infection is constant and allows incidence to be estimated over age bands. We have seen, however, that a constant force of infection is unlikely for HIV. Williams et al (2001) formulated an extended dynamic model which allowed for a changing force of infection, age-dependence of the risk of infection and differential mortality.

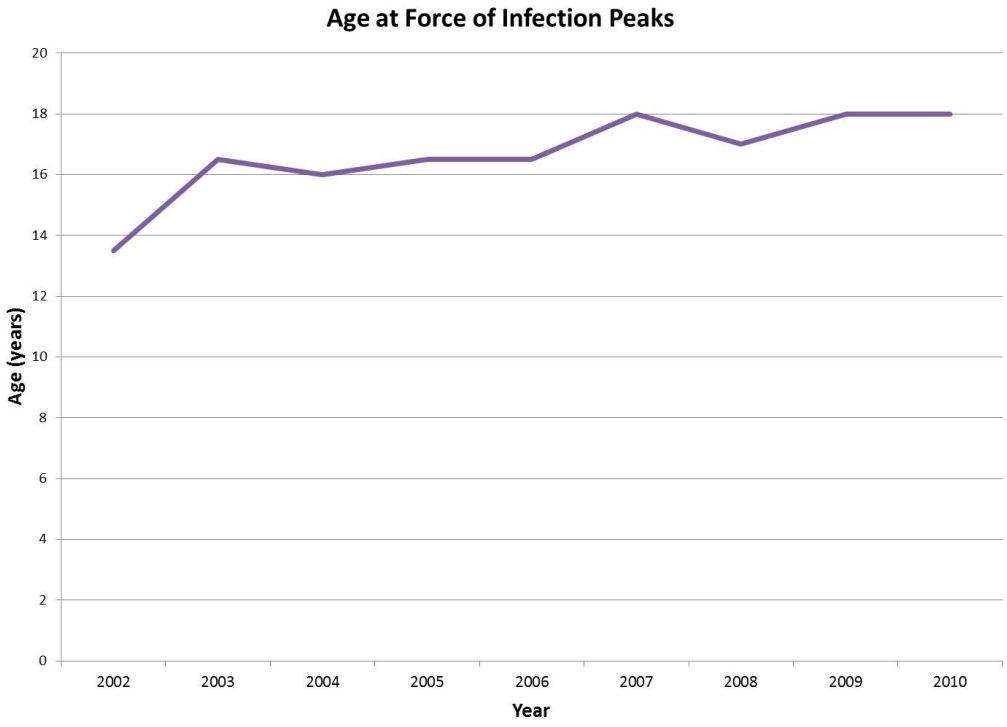

Figure 5.19: The age at which Farrington's force of infection peaks for each year.

Chapter 6

Conclusion

Prevalence and incidence are two vital measures of disease when considering infectious disease modelling. The prevalence gives us an idea about the proportion of the population that is living with the disease. However, since this proportion includes both old and new infections, it tells us little about the present moment status of the disease. Thus, in order for us to discern between old and new infections, we make use of incidence or equivalently the force of infection. Incidence refers to the rate of appearance of new infections or the force of infection. Hence, incidence affects the prevalence. By knowing the force of infection, we can ascertain the high risk factors of the disease as well as the effectiveness of awareness programmes and treatment strategies.

The project started by first exploring compartmental modelling of infectious diseases and the key parameters associated with it. The key parameters associated with the SIR, MSLIR and SIS models were explored as well and an interpretation of them was made. We then explored six different methods of estimating the force of HIV infection amongst pregnant women attending antenatal clinics in the Vulindlela district, for the years of 2009 and 2010. We also concluded that Farrington's two parameter model best described the prevalence and force of infection. The results showed that HIV prevalence increased from 60% in 2009 to 61% in 2010. It was shown that as age increases, the probability of being infected increased with age. We also found that the force of infection was at its highest in the 15-19 year old age group for both years, with the highest force of infection being 0.084 in 2010 - an increase from 2009's 0.079 force of infection. This suggests that HIV awareness campaigns should be aimed at the 15-19 year old age group in an effort to bring the rate of new infections down. Applying Farrington's two parameter model to a larger dataset (data for the years 2002-2008), we were able to identify any trends that may have been present. This application yielded three key findings - an increasing trend in prevalence, a decreasing trend of the force

of infection or equivalently the incidence rate and an increasing trend of the age at which the force of infection peaks. The ages, at which the force of infection was found to peak, ranged from 13-18 years. This key finding further emphasises the importance of HIV awareness campaigns targeting the 15-19 year old age group. Further research could focus on the 15-19 year old age group and determine other factors that could be contributing to this increased force of HIV infection.

The results also suggest that using only prevalence as a measure of disease intensity can be misleading. The force of infection is the best measure to use in evaluating the success of intervention and control strategies for a disease such as the use of ARVs in the case of HIV/AIDS. Prevalence should therefore be interpreted in conjunction with the force of infection.

Bibliography

[1] Akpa O.M., Oyejolu B.A. (2010). Modeling the transmission dynamics of HIV/AIDS epidemics: An introduction and a review. *J. Infect. Dev. Ctries.*, 4(10):597-608

[2] Becker N.G. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall. New York

[3] Brookmeyer R., Konikoff J. (2011). Statistical Considerations in Determining HIV Incidence from Changes in HIV Prevalence. *Statistical Communications in Infectious Diseases*, 3(1):1-14

[4] Conn P.B., Cooch E.G., Caley P. (2012). Accounting for detection probability when estimating force-of-infection from animal encounter data. *J. Ornithol.*, 152(2):5511-5520

[5] Dietz K., Heesterbeek J.A.P. (2002). Daniel Bernoulli's epidemiological model revisited. *Mathematical Biosciences 180 (2002)*, 1-21

[6] Farrington C.P. (1990). Modeling forces of infection for measles, mumps and rubella. *Statistics in Medicine*, 9(8):953-967

[7] Freeman J., Hutchison G. (1980). Prevalence, incidence and duration. *American Journal of Epidemiology*, 112(5):707-723

[8] Gray J.A., Dore G.J., Li Y., Supawitkul S., Effler P., Kaldor J.M. (1997). HIV-1 infection among female commercial sex workers in rural Thailand. *AIDS*, 11(1):89-94

[9] Grenfell B.T., Anderson R.M. (1985). The estimation of age-related rates of infection from case notifications and serological data. *Journal of Hygiene*, 95(2):419-436

[10] Griffiths D. (1974). A catalytic model of infection for measles. *Applied Statistics*, 23(3):330-339

[11] Hallett T.B., Zaba B., Todd J., Lopman B., Mwita W., Biraro S., Gregson S., Boerma J. (2008). Estimating incidence from prevalence

in generalized HIV epidemics: Methods and Validation. *PLoS Medicine*, 5(4):611-622

[12] Haran M. (2009). *An introduction to models for disease dynamics*. Penn State University

[13] Hardin J.W., Hilbe J.M. (2007). *Generalized Linear Models and Extensions*. 2nd edition. StataCorp LP. United States of America

[14] Heisy D., Joly D., Messier F.. (2006). The fitting of general force-of-infection models to wildlife disease prevalence data. *Ecology*, 87(9):2356-2365

[15] Hens N., Aerts M., Faes C. (2010). Seventy-five years of estimating the force of infection from current status data. *Epidemiology and Infection*, 138(6):802-812

[16] Hens N., Faes C., Aerts M., Shkedy Z., Mintiens K., Laevens H., Boelaert F. (2007). Handling missingness when modelling the force of infection from clustered seroprevalence data. *Journal of Agricultural, Biological and Environmental Statistics*, 12(4):498-513

[17] Hens N., Shkedy Z., Aerts M., Faes C., Van Damme P., Beutels P. (2012). *Modelling Infectious Disease Parameters Based on Serological and Social Contact Data*. Springer

[18] Indrayan A. (2013). Prevalence and Incidence. *Ganga Ram Journal*, 3(1):38-41

[19] Johnson L.F., Mossong J., Dorrington R.E., et al. (2013). Life Expectancies of South African Adults Starting Antiretroviral Treatment: Collaborative Analysis of Cohort Studies. *PLoS Medicine*, 10(4):e1001418

[20] Keiding N. (1991). Age-specific incidence and prevalence: a statistical perspective (with discussion). *Journal of the Royal Statistical Society, Series A*, 154(3):371-412

[21] Keiding N., Begtrup K., Scheike T.H., Hasibeder G. (1996). Estimation from current status data in continuous time. *Lifetime Data Analysis*, 2(2):119-129

[22] Mahiane G., Brand H. (2012). Update on HIV incidence estimation from prevalence data. *SACEMA* quarterly update

[23] McCullagh P., Nelder J.A. (1989). *Generalized Linear Models*. 2nd edition. Chapman & Hall/CRC. London, UK

[24] Muench H. (1934). Derivation of rates from summation data by the catalytic curve. *Journal of the American Statistical Association*, 29(185):25-38

[25] Muench H. (1959). *Catalytic Models in Epidemiology*. Harvard University Press, Boston

[26] Mwambi H., Ramroop S., White L.J., Okiro E.A., Nokes D.J., Shkedy Z., Molenberghs G. (2011). A frequentist approach to estimating the force of infection for a respiratory disease using repeated measurement data from a birth cohort. *Stat. Methods Med. Res.*, 20(5):551-570

[27] Namata H., Shkedy Z., Faes C. (2007). Estimation of the force of infection from current status data using generalized linear mixed models. *Journal of Applied Statistics*, 34(8):923-939

[28] National Department of Health. (2011). *National HIV and Syphilis Antenatal Seroprevalence Survey in South Africa,2010*, accessed 30 November 2012 (Available at http://www.hst.org.za/publications/2010-national-antenatal-sentinel-hiv-and-syphilis-prevalence-survey-south-africa)

[29] National Department of Health. (2012). *National HIV and Syphilis Antenatal Seroprevalence Survey in South Africa,2011*, accessed 3 October 2013 (Available at http://www.health.gov.za/docs/reports/2013/Antenatal_survey_report_2012_web_optimized.pdf)

[30] Nelder J.A., Wedderburn R.W.M. (1972). Generalised linear models. *Journal of the Royal Statistical Society: Series A*, 135(3):370-384

[31] Pettifor A.E., Measham D.M., Rees H.V., Padian N.S. (2004). Sexual Power and HIV Risk, South Africa. *Emerging Infectious Diseases online*, 10(11):1-9

[32] Rehle T., Shisana O., Pillay V., Zuma K., Puren A., Parker W. (2007). National HIV incidence measures - new insights into the South African epidemic. *South African Medical Journal*, 97(3):194-199

[33] Rodriguez G. (2010). *Parametric Survival Models*. Princeton University

[34] Sakarovitch C., Alioum A., Ekouevi D.K., Msellati P., Leroy V., Dabis F. (2007). Estimating incidence of HIV infection in childbearing age African women using serial prevalence data from antenatal clinics. *Statistics in Medicine*, 26(2):320-335

[35] Saphonn V., Hor L.B., Ly S.P. (2002). How well do antenatal clinic attendees represent the general population?. *International Journal of Epidemiology*, 31(2):449-455

[36] SAS Institute Inc. (2008). *SAS/STAT 9.2 User's Guide*. Cary, NC: SAS Institute Inc

[37] Sewpaul R. (2011). *Estimation and Analysis of Measures of Disease for HIV Infection in Childbearing Women Using Serial Seroprevalence Data*. Masters thesis. University of KwaZulu-Natal. Pietermaritzburg, South Africa

[38] Shkedy Z., Aerts M., Molenberghs G. (2003). Modelling forces of infection by using monotone local polynomials. *Journal of the Royal Statistical Society, Series C*, 52(4):469-485

[39] Shkedy Z., Aerts M., Molenberghs G., Beutels P., Van Damme P. (2006). Modelling age-dependent force of infection fro prevalence data using fractional polynomials. *Statistics in Medicine*, 25(9):1577-1591

[40] Shkedy Z., Mwambi H. (2009). *Modeling Infectious Diseases: An Introduction, Practical Sessions*. SUSAN Conference. Kenya

[41] Shkedy Z., Van Effelterre T., Namata H. (2009). *Modeling Infectious Diseases: An Introduction*. SUSAN Conference. Kenya

[42] UNAIDS (2010). *UNAIDS Report on The Global Aids Epidemic, 2010*. World Health Organisation. Geneva, Switzerland, accessed 2 September 2013 (Available at http://www.unaids.org/documents/20101123_GlobalReport_em.pdf)

[43] Whitaker H.J., Farrington C.P. (2004). Estimation of infectious disease parameters from serological survey data: the impact of regular epidemics. *Statistics in Medicine*, 23(15):2429-2443

[44] Williams B., Campbell C. (1998). Understanding the epidemic of HIV in South Africa. *South African Medical Journal*, 88(3):247-251

[45] Williams B., Gouws E., Wilkinson D., Abdool Karim S. (2001). Estimating HIV incidence rates from age prevalence data in epidemic situations. *Statistics in Medicine*, 20(13):2003-2016

[46] World Health Organization. (2004). *National AIDS programmes: a guide to indicators for monitoring and evaluating national HIV/AIDS prevention programmes for young people*, accessed 3 August 2013 (Available at http://www.who.int/hiv/pub/epidemiology/napyoungpeople.pdf)