Estimation and analysis of measures of disease for HIV infection in childbearing women using serial seroprevalence data

Ronel Sewpaul

June, 2011

Estimation and analysis of measures of disease for HIV infection in childbearing women using serial seroprevalence data

by

Ronel Sewpaul

Thesis submitted to the University of KwaZulu-Natal in fulfilment of the requirements for the Masters degree in Statistics.

Degree Assessment Board

Thesis supervisor	Prof. H. Mwambi
Thesis examiners	Examiners to be announced

UNIVERSITY OF KWAZULU-NATAL

PIETERMARITZBURG CAMPUS, SOUTH AFRICA

Disclaimer

This document describes work undertaken as a masters programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

Abstract

The prevalence and the incidence are two primary epidemiological parameters in infectious disease modelling. The incidence is also closely related to the force of infection or the hazard of infection in survival analysis terms. The two measures carry the same information about a disease because they measure the rate at which new infections occur. The disease prevalence gives the proportion of infected individuals in the population at a given time, while the incidence is the rate of new infections.

The thesis discusses methods for estimating HIV prevalence, incidence rates and the force of infection, against age and time, using cross-sectional seroprevalence data for pregnant women attending antenatal clinics. The data was collected on women aged 12 to 47 in rural KwaZulu-Natal for each of the years 2001 to 2006.

The generalized linear model for binomial response is used extensively. First the logistic regression model is used to estimate annual HIV prevalence by age. It was found that the estimated prevalence for each year increases with age, to peaks of between 36% and 57% in the mid to late twenties, before declining steadily toward the forties. Fitted prevalence for 2001 is lower than for the other years across all ages.

Several models for estimating the force of infection are discussed and applied. The fitted force of infection rises with age to a peak of 0.074 at age 15, and then decreases toward higher ages. The force of infection measures the potential risk of infection per individual per unit time. A proportional hazards model of the age to infection is applied to the data, and shows that additional variables such as partner's age and the number of previous pregnancies do have a significant effect on the infection hazard.

Studies for estimating incidence from multiple prevalence surveys are reviewed. The relative inclusion rate (RIR), accounting for the fact that the probability of inclusion in a prevalence sample depends on the individual's HIV status, and its role in incidence estimation is discussed as a possible future approach of extending the current work.

Keywords

prevalence, force of infection, incidence, HIV, pregnant women

Acknowledgements

This thesis was made possible by the assistance and support of several individuals and organizations to whom I would like to extend my sincere gratitude. Firstly I would like to thank Prof. Henry Mwambi, my supervisor, for his constant support and guidance throughout this research. Thank you for the advice, comments and motivation that has kept me working hard.

Thank you to SACEMA (the DST/NRF Centre for Epidemiological Modelling and Analysis) for the financing of my studies, as well as the academic support offered.

I would like to thank CAPRISA (Centre for the AIDS Program of Research In South Africa), for allowing me to use their rich dataset on HIV seroprevalence based on antenatal clinic visits. Thank you to Dr Ayesha Kharsany, Prof Quarraisha Abdool Karim, Dr Janet Frohlich, Ms Mukelisiwe Mlotshwa, Ms Nonhanhle Yende, Ms Ghetwana Mahlase, Mr Jayraj Ramotar and Ms Natasha Samsunder; and special thanks to the women who participated in the study. Thank you to the KwaZulu-Natal Department of Health, Umgungundlovu Health District for permission to conduct the study, which was ethically approved under UKZN BREC approval number E179/04.

Thank you to my colleagues and seniors at the Health Promotion Research and Development Unit, Medical Research Council, for their support and encouragement during this research.

A special thank you to my family and friends for their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 HIV in the World

The human immunodeficiency virus (HIV) is a retrovirus that infects cells of the immune system, destroying or impairing their function. As the infection progresses the immune system weakens, making the infected individual increasingly susceptible to opportunistic infections. The onset of acquired immunodeficiency syndrome (AIDS), the most severe stage of the HIV infection, occurs when the numbers of healthy cells of the immune system are very low and virus population grows without bound.

Transmission of HIV occurs through unprotected sexual intercourse, transfusion of contaminated blood and sharing of contaminated needles. The virus may also be transmitted from a mother to her infant during pregnancy, childbirth and breastfeeding. This mode of transmission is known as mother to child transmission (MCT). The majority of HIV infections among adults in Sub-Saharan Africa are transmitted through heterosexual intercourse. In other parts of the world, in particular North America and Europe, homosexual transmission among men is a significant transmission mode.

According to UNAIDS (2010) an estimated 33.3 million people worldwide are living with HIV. A staggering 68% of the global HIV infected population reside in Sub-Saharan Africa. Women account for a much higher proportion of all adults living with HIV worldwide.

The extent of the epidemic within Sub-Saharan Africa varies by region, with

Southern Africa carrying a particularly high proportion of the disease burden. UNAIDS (2010) estimates that in 2009 34% of all people living with HIV globally resided in the ten countries of Southern Africa, namely Angola, Botswana, Lesotho, Malawi, Mozambique, Namibia, South Africa, Swaziland, Zambia, and Zimbabwe. Swaziland has the highest adult HIV prevalence in the world, at an estimated 25.9%. Relatively lower prevalence rates are estimated in West and Central Africa, with Cameroon (5.3%) and Gabon (5.2%) having the highest prevalence rates in these regions. Lower prevalence is also estimated for East Africa, where the epidemics in many countries show signs of stabilising. Kenya and Uganda have had stable prevalence levels of around 5% and 7% respectively in the last few years.

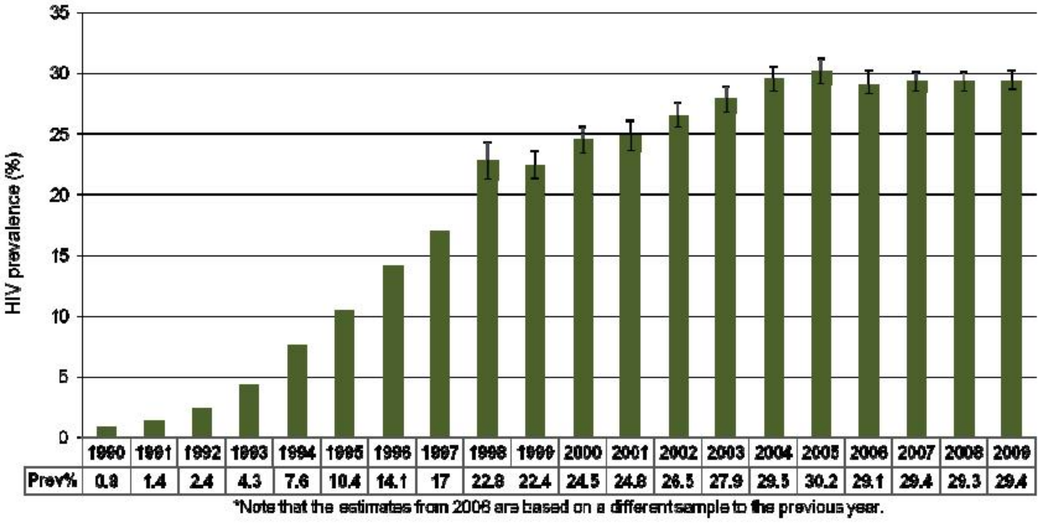
1.1.2 HIV in South Africa

An estimated 5.6 million people are living with HIV in South Africa, making it the country with the highest number of HIV positive people in the world. Approximately 3.3 million of these people are women aged 15 and older (UNAIDS, 2010). The HIV prevalence rate, which gives the proportion of HIV positive individuals out of all the individuals in the population, was 17.8% [17.2% - 18.3%] for 2009 as estimated by UNAIDS (2010).

There have been two major sources of data for monitoring the HIV epidemic in South Africa: seroprevalence surveys among women attending antenatal clinics and national population-based household surveys.

Data from antenatal surveillance

In many African countries, including South Africa, surveillance among pregnant women attending sentinel antenatal clinics (ANC) has been the primary source for estimating HIV prevalence in the general population and monitoring HIV trends over time. Pregnant women are seen to represent the general heterosexually active adult population. Furthermore, antenatal clinic surveys are relatively low cost and convenient, for developing countries in Africa, to conduct and the attendees serve as an easily accessible and stable population.



*Note that the estimates from 2006 are based on a different sample to the previous year.

Figure 1.1: HIV prevalence trends among antenatal women, South Africa, 1990 to 2009. Source: Department of Health (2010).

South Africa's National HIV and syphilis antenatal seroprevalence survey has been conducted annually since 1990. Participants in this cross-sectional survey are pregnant women who are attending public health antenatal clinic services for the first time during a pregnancy. ANC attendees undergo routine syphilis testing. Thus syphilis screening is used as an entry point for HIV testing using anonymous unlinked procedures. Since 2006, the survey's target sample size was expanded to increase geographical coverage down to district level. It now includes over 30 000 participants from 1457 clinics. This has allowed for estimates of prevalence to be calculated for each health district in each of the nine provinces.

The national HIV prevalence estimated from the annual ANC surveys rose sharply in the 1990's. Figure 1.1 shows that from 0.8% in 1990, the prevalence quickly grew to rates of over 20% in the following eight years. HIV prevalence reached a peak of 30.2% in 2005. For the four years thereafter, the national HIV prevalence has remained stable. The most recently published prevalence estimate is for the year 2009, and is estimated at 29.4% (Department of Health, 2010).

The HIV epidemic has progressed at a different pace in each province. The low-

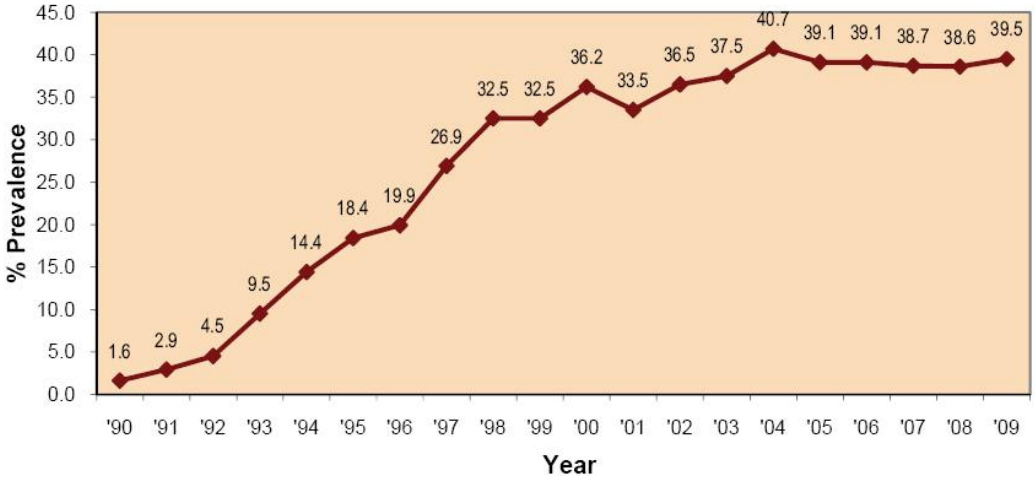


Figure 1.2: HIV prevalence trends among antenatal women, KwaZulu-Natal 1990 to 2009. Source: Department of Health (2010).

est HIV provincial prevalence rates in 2009 were observed in the Western Cape (16.9%) and the Northern Cape (17.2%). These two provinces have recorded consistently lower prevalence than the other provinces over time. KwaZulu-Natal has consistently recorded the highest provincial prevalence since 1990. In 2009 the estimated prevalence of 39.5% in KwaZulu-Natal far exceeded the national rate. Figure 1.2 shows the prevalence curve for KwaZulu-Natal. It can be seen that the prevalence increased steadily to a peak of 40.7% in 2004 before leveling off in the past few years. Results by district showed that in 2009 five of the eleven districts in KwaZulu-Natal, namely Ugu, Uthukela, eThekwini, ILembe and Umgungundlovu, produced very high HIV prevalence estimates of above 40% (Department of Health, 2010).

Table 1.1: HIV prevalence (%) among antenatal women by age group, South Africa 2001 to 2009.

Age Group	2001	2002	2003	2004	2005	2006	2007	2008	2009
<20	15.4	14.8	15.8	16.1	15.9	13.7	13.1	14.1	13.7
20-24	28.4	29.1	30.3	30.8	30.6	28.0	28.0	26.9	26.6
25-29	31.4	34.5	35.4	38.5	39.5	38.7	37.5	37.9	37.1
30-34	25.6	29.5	30.9	34.4	36.4	37.0	39.6	40.4	41.5
35-39	19.3	19.8	23.4	24.5	28.0	29.3	33.0	32.4	35.4

From 2001 to 2009 the prevalence among women aged 30 and older showed a general increasing trend. For the past seven years women in the age groups 25-

29 and 30-34 showed consistently higher HIV prevalence than those in other age groups. For the past three years the prevalence was highest among women in the 30-34 year age group, rising from 39.6% in 2007 to 40.4% in 2008 and 41.5% in 2009. Prior to this the 25-29 year age group had the highest prevalence rates with a peak rate of 39.5% in 2005 for this age group. In recent years the HIV prevalence among women under 30 years has gradually declined while the prevalence for the over 30's has continued to increase. The prevalence among women aged 35-39 increased by 6% over a period of four years, from 29.3% in 2006 to 35.4% in 2009. However, much of the increases in prevalence in recent years could be attributed to the increases in survival of those on antiretroviral treatment (ART) (Department of Health, 2010). Table 1.1 summarises the estimated annual prevalence rates by age group from 2001 to 2009. Data used is from the Department of Health National HIV and syphilis antenatal seroprevalence survey reports for the years 2004, 2007 and 2010. It is evident from the above information that HIV prevalence and incidence vary from place to place, by gender and over different time periods.

Data from population-based surveys

Many countries have recently conducted national population-based surveys for surveillance of the HIV epidemic in the general population. The South African National HIV Prevalence, Incidence, Behaviour and Communication Survey, 2008 was the third in a series of population-based household surveys designed for estimating HIV prevalence and incidence as well as HIV-related behaviour in South Africa (Shisana et al, 2009). The previous two surveys were conducted in 2002 and 2005. The survey sample included individuals aged 2 years and up. The multi-stage stratified sampling approach was designed to produce a nationally representative sample.

The overall estimated HIV prevalence rates for the South African population changed little across survey years - 11.4% in 2002, 10.8% in 2005 and 10.9% in 2008. The highest provincial prevalence in 2008 was found in KwaZulu-Natal (15.8%) and the lowest in Western Cape (3.8%). Females recorded disproportionately higher rates than males, with larger disproportions in the younger age groups. In the 15-19, 20-24 and 25-29 year age groups prevalence rates among females were more than 2 times higher than that of males. In 2008 prevalence among females was highest in the 25-29 year age category with a rate of 32.7%, while prevalence among males was highest in the 30-34 year age category with a rate of 25.8%. A decline in prevalence was found among youth aged 15-24, from 10.3% in 2005 to 8.6% in 2008. This can be attributed to several factors, among them the impact of education, which has a direct outcome of reducing new infections.

It should be noted that national population based surveys provide HIV prevalence data that is representative of the national population. These surveys are however very costly to conduct on an annual basis, and response rates can be low, particularly among specific demographic or socioeconomic population groups. Antenatal clinic surveillance, on the other hand, reflects the HIV prevalence of pregnant women attending public health antenatal clinics. It does not include males, non-pregnant women, or pregnant women who are not consulting public health antenatal clinics. However antenatal clinic surveys are relatively low cost to conduct, and antenatal clinic attendees serve as an easily accessible population. For these reasons antenatal surveys are able to be conducted on an annual basis in many low and middle income countries, providing prevalence data on large numbers of individuals.

1.2 Thesis Objectives

The objectives of the thesis are to:

- Model cross-sectional prevalence data

- Estimate HIV prevalence

- Estimate the force of infection for HIV

- Model the effect of covariates

The theoretical methods for implementing these objectives are explained and then applied using data recorded on women attending antenatal clinics in Vulindlela, a rural area of KwaZulu-Natal.

1.3 Thesis Overview

The thesis discusses the estimation of key measures of disease with reference to the HIV infection. The data used in this thesis is described in detail in Chapter 2. Exploratory analyses, examining the observed HIV prevalence by age group, as well as by other covariate groups, such as partner age group and the number of previous pregnancies, are also presented. Chapter 3 explains the generalized linear model and its use in modelling binary data. Logistic regression is then applied to the Vulindlela data, to model age-specific prevalence by year. Chapter 4 reviews methods for estimating the incidence of a disease from prevalence data. The usefulness of these methods in the context of HIV in South Africa is of importance, since the HIV epidemic may be reaching maturity in the country. A mature epidemic is one where the proportion of HIV infected people in the population levels off. The force of infection gives the rate at which individuals become infected with a disease, and the estimation of this important measure is discussed in Chapter 5. Some well known functions for the force of infection are illustrated using the Vulindlela antenatal data. The proportional hazards model is used in Chapter 6 to investigate the effect of explanatory variables, such as the number of previous pregnancies and the age of the participants' male partners, on the hazard of HIV infection, for women attending antenatal clinics in Vulindlela. SAS Version 9.2 was used for all the analyses in the current thesis.

Chapter 2

Exploratory Data Analysis

2.1 Data

The data was obtained from CAPRISA (Centre for AIDs Programme of Research in South Africa). It is in the form of serial cross-sectional prevalence data, obtained from anonymous HIV prevalence testing of pregnant women attending eight public health antenatal clinics in Vulindlela, an area of rural KwaZulu-Natal. The data was collected annually for each of the years 2001 to 2006. There are 2245 observations in total. The minimum and maximum age of pregnant women included in the study is 12 and 47 respectively.

The variables in the dataset are listed below.

- ID (each participant was allocated a unique identification number)
- Year (ranging from 2001 to 2006)
- Age (age of the participant at the time of testing)
- HIV status (either positive or negative)
- Partner Age (the age of each participant's male partner)
- Clinic (the name of the antenatal clinic attended by each participant)
- Previous Pregnancies (the number of previous pregnancies of each participant and the year in which each previous pregnancy occurred)

Note that the additional variables Partner Age, Clinic, and Previous Pregnancies were not recorded in the datasets for all of the survey years. Partner Age was recorded for participants in survey years 2003-2006. The clinic attended was recorded in the years 2002-2005 and Previous Pregnancies was recorded in years 2004-2006.

Vulindlela is a rural community in the KwaZulu-Natal midlands, situated about 150 km west of Durban. It has a population of approximately 400 000. According to sources at CAPRISA, a large proportion of this population is unemployed.

Peripheral blood specimens collected from women attending the antenatal clinics are routinely tested for syphilis, rhesus factor and ABO blood grouping, to detect and prevent haemolytic disease in the newborn babies. After removing personal identifiers, these blood specimens were then couriered to the CAPRISA research laboratories in Durban. HIV testing was performed using the standard enzyme linked immunosorbent antibody assay (Enzygnost, Dade Behring), (sensitivity: 100, specificity: 99.3). Reactive samples were confirmed using the Abbott Determine TM HIV -1/2 (Abbott Laboratories), (sensitivity: 100, specificity: 99.91).

The study comprised all pregnant women consulting any of the eight public health antenatal clinics in Vulindlela. However, only women attending these clinics for their first antenatal visit of their current pregnancy were included in the study, to prevent the same woman being included more than once in a specific year's dataset. The HIV status data from the above study leads to what is referred to as current status data. The data reports the age as well as the values of other covariates recorded for an individual at a given calender time, together with the disease status, here the HIV status.

2.2 Description of the sample

The total number of individuals observed in each year varied from 234 in 2003 to 552 in 2004. From Table 2.1 we see that the majority of individuals in each year (between 55% and 65%) were under 25 years of age. The proportion of 25-34 year olds ranged from 26.9% to 31.7%. Less than 12% of the individuals in each year were older than 35, with a very low proportion aged 40-47.

Table 2.1: Description of the sample for each of the years 2001-2006

	2001	**2002**	**2003**	**2004**	**2005**	**2006**	
	% (n)	% (n)	% (n)	% (n)	% (n)	% (n)	p-value
Age Group							
12-19	24.6% (41)	38.9% (159)	31.2% (73)	36.7% (202)	30.2% (109)	31.6% (96)	0.111
20-24	32.3% (54)	26.2% (107)	28.2% (66)	28.9% (159)	31.0% (112)	32.2% (98)	
25-29	19.2% (32)	20.5% (84)	21.4% (50)	15.1% (83)	16.9% (61)	15.5% (47)	
30-34	12.6% (21)	8.8% (36)	9.4% (22)	11.8% (65)	12.2% (44)	11.5% (35)	
35-39	8.4% (14)	3.9% (16)	7.3% (17)	5.6% (31)	6.4% (23)	6.9% (21)	
40-47	3.0% (5)	1.7% (7)	2.6% (6)	1.8% (10)	3.3% (12)	2.3% (7)	
Partner's age							
⩽ 19			9.6% (22)	12.4% (66)	10.6% (38)	10.3% (31)	0.348
20-24			31.3% (72)	34.7% (185)	32.5% (117)	31.6% (95)	
25-29			27.0% (62)	22.7% (121)	21.9% (79)	25.9% (78)	
30-34			15.7% (36)	13.7% (73)	15.6% (56)	15.9% (48)	
35-39			10.4% (24)	7.3% (39)	13.1% (47)	9.0% (27)	
⩾ 40			6.1% (14)	9.2% (49)	6.4% (23)	7.3%(22)	
Partner Age Difference							
Younger partner			5.2% (12)	4.7% (25)	7.2% (26)	5.3% (16)	0.848
0-3 years older			51.7% (119)	51.6% (275)	48.9% (176)	48.5% (146)	
4-7 years older			32.6% (75)	30.6% (163)	30.6% (110)	34.2% (103)	
8-11 years older			7.8% (18)	8.4% (45)	9.2% (33)	9.0% (27)	
⩾ 12 years older			2.6% (6)	4.7% (25)	4.2% (15)	3.0% (9)	
Clinic Attended							
Mafakhathini		11.1% (46)	11.1% (26)	5.1% (28)	16.3% (59)		0.000*
Mpumuza		17.4% (72)	9.8% (23)	12.1% (67)	26.3% (95)		
Mpophomeni		18.2% (75)	28.6% (67)	11.4% (63)	22.2% (80)		
Taylors		10.7% (44)	12.4% (29)	21.4% (118)	10.8% (39)		
Songonzima		14.0% (58)	15.0% (35)	20.5% (113)	12.2% (44)		
Elandskop		26.2% (108)	18.8% (44)	10.5% (58)	10.0% (36)		
Sondelani		-	4.3% (10)	19.0% (105)	2.2% (8)		
Ntembeni		2.4% (10)	-	-	-		
Number of previous pregnancies							
None				59.8% (312)	48.6% (171)	52.0% (168)	0.014*
One				25.1% (131)	33.0% (116)	30.0% (97)	
Two				10.3% (54)	10.5% (37)	9.9% (32)	
Three				3.1% (16)	6.0% (21)	4.0% (13)	
Four				1.7% (9)	2.0% (7)	4.0% (13)	
Total	100% (349)	100% (413)	100% (234)	100% (552)	100% (362)	100% (335)	

The variable Partner's age was recorded in the survey years from 2003 onwards. In each year, more than half of the individuals had partners aged between 20 and 29, while around a quarter had partners aged 30-39. The age difference between a woman and her male partner was calculated by subtracting the woman's age from that of her partner. A relatively low proportion of the individuals had partners who were younger than them. The most commonly observed age difference was 0-3 years, with around half of the individuals having partners up to three years older than them. Between 10% and 14% reported having a partner eight or more years their senior.

The name of the clinic attended by each individual was recorded only in years 2002-2005. The distribution of individuals by clinic seems to vary with each year, with the lowest proportions of individuals selected from the Sondelani and Ntembeni clinics. Cells containing a “ - ” indicate that there were no observations (i.e. $n = 0$) in those cells. For example the “ - ” corresponding to Sondelani in year 2002 means that there were no individuals who were tested at the Sondelani clinic in year 2002.

The variable Previous Pregnancies was recorded in the dataset from 2004 onwards. Note that this variable refers to the number of pregnancies experienced by an individual prior to their current pregnancy. The majority of the individuals in each year had never been pregnant before. Between a quarter and a third had experienced one previous pregnancy, and a lower proportion (15-18%) reported having had 2-4 previous pregnancies.

Chi-square tests were performed to test for significant associations between the year of observation and each of the explanatory variables. A “ $*$ ” next to each p-value indicates a significant association, that is, where $p < 0.05$. A significant association was found between clinic and year ($p = 0.000$), meaning that the distribution of individuals attending each clinic varied significantly by year. Similarly there was a significant association between the number of previous pregnancies and year ($p = 0.014$).

2.2.1 Description of the sample by age category and the number of previous pregnancies

The characteristics of the combined 2001-2006 sample by age category are presented in Table 2.2. Age was dichotomized into three groups, namely younger than 22 years, 22-31 years, and older than 31 years. These age groups were chosen in consultation with experts in the field of HIV prevalence data. There were significant associations between the women's age group and each of the variables partner's age, partner age difference and the number of previous pregnancies ($p < 0.05$ for all).

The age of the women's male partners generally increased with the women's ages. However, large age differences between women and their partners were more common among those older than 31 years. The prevalence of having a partner eight or more years their senior was higher among those aged 31 and older (23.2%) than those aged 22-31 years (13.0%) and younger than 22 years (8.9%). In addition, more individuals aged 31 and older had partners younger than themselves compared to those aged 22-31 and less than 22 years.

Older women tended to have more previous pregnancies. Two thirds of the individuals aged 32 and older had experienced two or more previous pregnancies, while the majority (89.6%) of the individuals younger than 22 years had never been pregnant before.

Table 2.2: Description of the sample by age category

	$<$ 22 years	**22 - 31 years**	**$>$ 31 years**	
	% (n)	% (n)	% (n)	p-value
Partner's age				
19 & younger	22.6% (156)	0.2% (1)	-	0.000*
20-24	57.3% (395)	14.1% (74)	-	
25-29	17.7% (122)	40.6% (213)	2.4% (5)	
30-34	1.9% (13)	31.1% (163)	17.5% (37)	
35-39	0.3% (2)	11.3% (59)	36.0% (76)	
40 or older	0.1% (1)	2.7% (14)	44.1% (93)	
Partner-age difference				
Younger partner	1.6% (11)	7.4% (39)	13.7% (29)	0.000*
0-3 years older	53.4% (368)	48.5% (254)	44.5% (94)	
4-7 years older	36.1% (249)	31.1% (163)	18.5% (39)	
8-11 years older	7.0% (48)	8.8% (46)	13.7% (29)	
$\geqslant$ 12 years older	1.9% (13)	4.2% (22)	9.5% (20)	
Number of previous pregnancies				
0	89.6% (510)	28.6% (122)	4.4% (8)	0.000*
1	10.0% (57)	54.0% (230)	28.7% (52)	
2 or more	0.4% (2)	17.4% (74)	66.9% (121)	

The partner's age and partner-age difference characteristics of the women according to the number of previous pregnancies experienced are presented in

Table 2.3, with significant associations found in each case ($p < 0.05$ for all). The data shows that the women who had multiple pregnancies tended to have older partners. The majority of those who had no previous pregnancies had partners aged below 30 years, the majority of those with one previous pregnancy had partners aged 20-34, and of those with two or more previous pregnancies the majority had partners aged 30 and above. Note however that this reflects the effect of the woman's age, because as shown in Table 2.2, partner's age tended to increase with the woman's age, and the older the woman the greater the number of previous pregnancies.

With regard to partner-age difference, a large partner-age difference was more common among those with multiple previous pregnancies. The proportion of individuals who had a partner eight or more years older than them was higher for those with one previous pregnancy and two or more previous pregnancies than for women who had never been pregnant before.

Note that in Table 2.3 as well as Table 2.2 above, blank ("-") values indicate that there were no observed individuals (i.e. $n = 0$) in those cells.

Table 2.3: Partner-age difference and partner age characteristics of the sample by the number of previous pregnancies

Number of previous pregnancies	**None**	**One**	**Two or more**	
	% (n)	% (n)	% (n)	p-value
Partner's age				
19 & younger	20.5% (130)	0.6% (2)	-	0.000^*
20-24	50.0% (317)	20.7% (70)	0.5% (1)	
25-29	21.9% (139)	31.7% (107)	11.3% (22)	
30-34	5.2% (33)	26.9% (91)	26.8% (52)	
35-39	1.4% (9)	14.8% (50)	26.3% (51)	
40 or older	0.9% (6)	5.3% (18)	35.1% (68)	
Partner-age difference				
Partner <8 years older than participant	90.9% (576)	87.0% (294)	76.3% (148)	0.000^*
Partner $\geqslant 8$ years older than participant	9.1% (58)	13.0% (44)	23.7% (46)	
Total	100.0% (634)	100.0% (338)	100.0% (194)	

2.3 Observed HIV prevalence by year

The observed HIV prevalence in each year was 27.5% in 2001, 34.2% in 2002, 40.9% in 2003, 42.6% in 2004, 37.3% in 2005 and 37.4% in 2006. Tables 2.4 to 2.8 present the observed HIV prevalence percentage and corresponding sample size by year, for each category of the explanatory variables. The estimated prevalence for a given year and a given level of an explanatory variable, is expressed as a percentage, and is calculated by $\hat{p}_{ti} = z_{ti}/n_{ti}$ x 100, where z_{ti} is the observed number of individuals testing HIV positive in year t and variable group i and n_{ti} is the total number of individuals observed in year t and variable group i. Many cell sample sizes are particularly small, for example, there are only between 5 and 12 individuals who are aged 40-47 (Table 2.4). One should exercise caution when making inferences based on results where the sample size (n) is too small, that is, when $n < 35$.

Chi-square tests were used to test for significant associations between HIV prevalence and the explanatory variables in each year. Significant associations between the the women's age group and the observed HIV prevalence were found in each of the years 2001-2006 (Table 2.4). Prevalence rates by age group showed a similar trend in each year. The observed HIV prevalence rose with age, then peaked in one of the age groups between ages 20 and 35, before declining towards the 40-47 year age group. In 2001 and 2002 HIV prevalence was highest in the 20-24 year age group, reaching rates of 44.4% and 45.8%; while in 2003 and 2004 the prevalence peaked in the 25-29 year age group, with rates of around 66% being observed. Note that the sample sizes in the two oldest age groups, 35-39 and 40-47, are fairly small. Therefore the calculated prevalence in these groups are less reliable than the prevalence obtained for the younger age groups, which contain large sample sizes.

In each year a significant association between HIV prevalence and partner's age was found (Table 2.5). The highest HIV prevalence was observed in individuals with partners in the 25-29 or 30-34 year age category, with the exception of year 2005, in which the highest prevalence was observed in the 35-39 year partner's age category. Relatively lower prevalence rates were observed for individuals whose partners were aged 19 and younger.

Observed HIV prevalence across the partner age difference categories are displayed in Table 2.6. In survey years 2004-2006, the prevalence appears to increase with age difference, from age differences of 0-3 years, 4-7 years, to 8-11 years. Although the prevalence rates for individuals with younger partners and those with partners 12 or more years older are very high, these estimates may not be reliable due to very small sample sizes for these groups.

Table 2.7 shows no clear pattern of prevalence rates by clinic. As a result no significant association was found between the clinic attended and the observed HIV prevalence ($p > 0.05$) in each year. The "-" in cells of Table 2.7 indicate that there were no observations (i.e. $n = 0$) in each of those cells, from which to calculate the prevalence. Table 2.8 shows that in each of the years 2004, 2005 and 2006, the HIV prevalence varies significantly by the number of previous pregnancies ($p < 0.05$). In 2004 and 2005 the prevalence appears to be highest for those individuals who have had one previous pregnancy, while in 2006 the prevalence is highest for those with three previous pregnancies.

Table 2.4: Observed HIV Prevalence (%) by age group for each year, 2001-2006

	2001	2002	2003	2004	2005	2006
Age Group	% Prev (n)	% Prev (n)	% Prev (n)	% Prev (n)	% Prev (n)	% Prev (n)
12-19	14.6% (41)	25.8% (159)	19.4% (67)	26.7% (202)	22.0% (109)	16.7% (96)
20-24	44.4% (54)	45.8% (107)	44.6% (65)	54.7% (159)	37.8% (111)	48.5% (97)
25-29	31.3% (32)	42.9% (84)	66.0% (50)	66.3% (83)	50.8% (61)	51.1% (47)
30-34	14.3% (21)	22.2% (36)	42.9% (21)	53.8% (65)	56.8% (44)	51.4% (35)
35-39	14.3% (14)	31.3% (16)	37.5% (16)	12.9% (31)	39.1% (23)	25.0% (20)
40-47	20.0% (5)	14.3% (7)	33.3% (6)	0.0% (10)	25.0% (12)	57.1% (7)
p-value	0.012*	0.003*	0.000*	0.000*	0.000*	0.000*

Table 2.5: Observed HIV Prevalence (%) by partner age group for each year, 2003-2006

	2003	2004	2005	2006
Partner Age Group	% Prev (n)	% Prev (n)	% Prev (n)	% Prev (n)
$\leqslant$19	5.3% (19)	16.7% (66)	10.5% (38)	9.7% (31)
20-24	40.0% (70)	34.1% (185)	31.9% (116)	27.4% (95)
25-29	46.7% (60)	65.3% (121)	38.0% (79)	51.9% (77)
30-34	63.9% (36)	63.0% (73)	46.4% (56)	54.2% (48)
35-39	22.7% (22)	41.0% (39)	59.6% (47)	34.6% (26)
$\geqslant$40	35.7% (14)	26.5% (49)	39.1% (23)	36.4% (22)
p-value	0.001*	0.000*	0.000*	0.000*

Table 2.6: Observed HIV Prevalence (%) by partner-participant age difference for each year, 2003-2006

Partner Age Difference Category	2003 % Prev (n)	2004 % Prev (n)	2005 % Prev (n)	2006 % Prev (n)
Younger partner	75.0% (12)	52.0% (25)	50.0% (26)	56.3% (16)
0-3 years older	41.6% (113)	40.0% (275)	27.8% (176)	29.9% (144)
4-7 years older	41.1% (73)	42.9% (163)	43.1% (109)	38.8% (103)
8-11 years older	11.8% (17)	53.3% (45)	51.5% (33)	55.6% (27)
$\geqslant$ 12 years older	33.3% (6)	44.0% (25)	53.3% (15)	55.6% (9)
p-value	0.018^*	0.434	0.006^*	0.026^*

Table 2.7: Observed HIV Prevalence (%) by Clinic for each year, 2002-2005

Clinic	2002 % Prev (n)	2003 % Prev (n)	2004 % Prev (n)	2005 % Prev (n)
Mafakhathini	34.8% (46)	30.8% (26)	35.7% (28)	39.0% (59)
Mpumuza	40.3% (72)	47.6% (21)	34.3% (67)	45.3% (95)
Mpophomeni	38.7% (75)	50.7% (67)	34.9% (63)	35.4% (79)
Taylors	38.6% (44)	37.9% (29)	43.2% (118)	30.8% (39)
Songonzima	31.0% (58)	39.3% (28)	53.1% (113)	29.5% (44)
Elandskop	28.7% (108)	31.8% (44)	43.1% (58)	36.1% (36)
Sondelani	-	40.0% (10)	41.9% (105)	25.0% (8)
Ntembeni	10.0% (10)	-	-	-
p-value	0.364	0.443	0.169	0.525

Table 2.8: Observed HIV Prevalence (%) by the number of previous pregnancies for each year, 2004-2006

Number of previous pregnancies	2004 % Prev (n)	2005 % Prev (n)	2006 % Prev (n)
0	36.2% (312)	26.5% (170)	29.2% (168)
1	61.8% (131)	49.1% (116)	45.8% (96)
2	46.3% (54)	43.2% (37)	45.2% (31)
3	12.5% (16)	47.6% (21)	53.8% (13)
4	0.0% (9)	28.6% (7)	46.2% (13)
p-value	0.000^*	0.002^*	0.033^*

2.4 Observed HIV prevalence by age category

The observed HIV prevalence by the number of previous pregnancies, partner's age and partner-age difference, for each of the age groups <22 years, 22-31 years and >31 years are presented in Table 2.9.

Observed HIV prevalence in the combined sample was highest (52.7%) among 22-31 years olds. The results suggest that for very young individuals (<22 years) and much older individuals (>31 years) the observed HIV prevalence

Table 2.9: Observed HIV Prevalence (%) by the number of previous pregnancies, partner's age and partner-age difference, for each of the age groups < 22 years, $22 - 31$ years and > 31 years

	< 22 years	**22-31 years**	**> 31 years**
	% Prev (n)	% Prev (n)	% Prev (n)
Partner's Age			
19 & younger	12.4% (153)	0.0% (1)	-
20-24	28.8% (392)	55.4% (74)	-
25-29	43.8% (121)	56.9% (211)	80.0% (5)
30-34	46.2% (13)	58.3% (163)	54.1% (37)
35-39	50.0% (2)	55.9% (59)	32.9% (73)
40 or older	0.0% (1)	64.3% (14)	28.0% (93)
p-value	0.000*	0.870	0.006*
Partner Age Difference			
Younger partner	9.1% (11)	66.7% (39)	58.6% (29)
0-3 years older	23.9% (364)	53.2% (252)	30.4% (92)
4-7 years older	31.6% (247)	58.3% (163)	36.8% (38)
8-11 years older	44.7% (47)	58.7% (46)	34.5% (29)
$\geqslant$ 12 years older	38.5% (13)	72.7% (22)	25.0% (20)
p-value	0.008*	0.251	0.066
Number of previous pregnancies			
None	26.5% (509)	56.6% (122)	12.5% (8)
One	33.3% (57)	59.4% (229)	48.1% (52)
Two or more	100.0% (2)	50.0% (74)	33.3% (120)
p-value	0.039*	0.364	0.064
Total	27.9% (970)	52.7% (751)	32.2% (292)

for those who have had at least one previous pregnancy is higher than for those who have not been pregnant before. However this association is only statistically significant for those in the <22 year age category ($p = 0.039$). For individuals aged 22-31 years, the HIV prevalence does not differ significantly across the number of previous pregnancies. The HIV prevalence in this age group was highest for those who have had one previous pregnancy (59.4%).

The observed HIV prevalence varied significantly across the partner age groups for women aged either below 22 years or older than 31 years, while HIV prevalence varied significantly across partner age difference categories only for women aged below 22 years. Among the individuals aged below 22 years, high HIV prevalence rates (between 43.8% and 50.0%) were observed for those with partners aged 25-39. For those aged 22-31 years however, prevalence remained high (above 50%) across all the partner age categories from 20-24 years and up.

Among individuals aged 22 and younger and 22-31 years, prevalence appears to be higher for those with partners either 8-11 years or 12 or more years older than themselves.

Chapter 3

The Generalized Linear Model (GLM)

Generalized linear models, first introduced by Nelder and Wedderburn (1972), are a class of statistical models that is a natural generalization of classical linear models. They allow for the modelling of response variables from a variety of distributions, when the relationship between the response and the explanatory variables is not of the simple linear form (Dobson, 1990). Generalized linear models include linear regression and ANOVA models, logit and probit models, log-linear and multinomial response models for counts and models used in survival analysis, which were previously studied as individual special topics. The above models share a number of properties which enable us to study them as a single class and to apply a common method for parameter estimation (McCullagh and Nelder, 1989).

Unlike using linear combinations such as $\mathbf{X}\beta$ to relate the mean of the distribution to the explanatory variables, as with linear regression, generalized linear models instead make use of general functions of linear combinations $h(\mathbf{X}\beta)$. In so doing, they can be applied to models where the relationship between the mean of the response variable and the explanatory variables is non-linear.

One of the key components in the development of generalized linear models is their link to the Exponential family of distributions. They exploit the fact that many of the properties of the Normal distribution were shared by this wider class of distributions. This allowed for a unified theory of extending the regression model to all distributions in the exponential family other than the normal distribution.

3.1 Exponential family of distributions

All response variables Y to which generalized linear models can be applied belong to the exponential family of distributions, which have the general form

$$f_Y(y;\theta,\phi) = exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\} \tag{3.1}$$

for some specific functions $a(.)$, $b(.)$ and $c(.)$.

The function $a(\phi)$ has the form $a(\phi) = \phi/w$ where ϕ is the dispersion or scale parameter which is constant over all observations and w is a known prior weight which varies with each observation. The parameter θ is known as the canonical parameter.

It can be shown that if a response Y has a distribution in the exponential family then it has mean and variance

$$E(Y) = \mu = b'(\theta) \tag{3.2}$$

$$Var(Y) = a(\phi)\ b''(\theta) \tag{3.3}$$

where $b'(\theta)$ and $b''(\theta)$ denote the first and second derivatives of $b(\theta)$.
The function $b''(\theta)$ is a function of the mean and hence, of θ. It is called the variance function, denoted by $V(\mu)$. Thus from (3.3), we have $Var(Y) = a(\phi)V(\mu)$.

Members of the exponential family include the Normal, binomial, Poisson, Gamma, exponential and inverse Gaussian distributions.

3.2 The GLM Model

Let $Y_i, i = 1, ..., n$ denote n independent observations of a random variable, not necessarily normally distributed. From McCullagh and Nelder (1989), a generalized linear model (GLM) has the following three components:

The random component
It is assumed that the responses Y_i ($i = 1, ..., n$) are independent random variables sharing the same distribution from the exponential family, with $E(Y_i) = \mu_i$ and a constant scale parameter. Thus Y_i satisfies equation (3.1) so that

$$f(y_i;\theta_i,\phi) = exp\{(y_i\ \theta_i - b(\theta_i))\ /\ a(\phi) + c(y_i,\phi)\}$$

The systematic component

A set of $p+1$, (usually) unknown parameters, β_i $(i = 0, 1, 2, ..., p)$, and the design matrix of known explanatory variables $\mathbf{X}_{n \,\mathrm{x}\, (p+1)}$, define a linear predictor η given by

$$\eta = \mathbf{X}\beta$$

where the $i'th$ row of $\mathbf{X}$ is given by $x_i = (1, x_{i1}, ..., x_{ip})'$ with $x_{ij}, \quad i = 1, ..., n$; equal to the value of the $j'th$ predictor or explanatory variable x_j, $j = 1, \ldots, p$ and $\beta' = (\beta_0, \beta_1, ..., \beta_p)$ is a vector of regression coefficients including the constant β_0 corresponding to $X_0 = 1$.

The link function

The link function, $g(.)$, gives the relationship between the mean of the $i'th$ observation and its linear predictor, so that

$$\begin{aligned} \eta_i &= g(\mu_i) \\ &= x_i' \, \beta \end{aligned}$$

The link function must be monotonic and differentiable. The *canonical link function* is that function which makes the linear predictor η_i the same as the canonical parameter θ_i from the exponential family member. With the canonical link function, all unknown parameters in β have sufficient statistics if the response distribution is a member of the exponential family with known scale parameter (Lindsey, 1997).

A normally distributed random variable therefore has a generalized linear model of the form $\eta_i = \mu_i = x_i'\beta$, where the canonical link is the identity. Thus the GLM is a unified approach which brings linear models under the same structure.

3.3 Parameter estimation

Maximum Likelihood estimation forms the theoretical basis for parameter estimation in generalized linear models. From (3.1) the log likelihood for a single observation in canonical form is given by

$$\ell_i = (y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi).$$

The maximum likelihood estimate $\hat{\beta}_j$ is the solution to the score equation $\frac{\partial \ell_i}{\partial \beta_j} = 0$. Note that since $Y_1, Y_2, ..., Y_n$ are independent the combined log likelihood is

$$\ell = \sum_{i=1}^{n} \ell_i.$$

By the chain rule,

$$\frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell}{\partial \theta} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

From $\eta_i = \sum \beta_j x_{ij}$ we have $\partial \eta_i / \partial \beta_{ij} = x_{ij}$, and from $b'(\theta_i) = \mu_i$ and $b''(\theta_i) = V_i$, we have $d\mu_i / d\theta_i = V_i$, where V_i is the variance function.

Therefore

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)} \frac{1}{V_i} \frac{d\mu_i}{d\eta_i} x_{ij} \\ &= \sum_{i=1}^{n} \frac{1}{a(\phi)} W_i \, (y_i - \mu_i) \, \frac{d\eta_i}{d\mu_i} \, x_{ij} \qquad (3.4) \end{aligned}$$

where

$$W_i = \left(\frac{d\mu_i}{d\eta_i} \right)^2 V_i^{-1}. \qquad (3.5)$$

Hence the parameter estimates $\hat{\beta}_j$ are given by the solutions to the maximum likelihood equations

$$\sum_{i=1}^{n} W_i \, (y_i - \mu_i) \, \frac{d\eta_i}{d\mu_i} \, x_{ij} = 0 \qquad (3.6)$$

where $\sum$ denotes summation over all units and x_{ij} is the $j'th$ covariate.

The vector $\partial \ell / \partial \beta$ is called the score vector, denoted by $\mathbf{U}$. It has expected value $E(\mathbf{U}) = \mathbf{0}$ and variance-covariance matrix

$$\jmath = E\bigg(- \frac{\partial^2 \ell}{\partial \beta^2} \bigg) = \frac{1}{a(\phi)} \mathbf{X}'\mathbf{W}\mathbf{X}, \qquad (3.7)$$

where $\mathbf{W}$ is the weight matrix with diagonal elements given by (3.5). $\jmath$ is referred to as the Information matrix. Its elements are the expected values of minus the second derivatives of the log likelihood. The information matrix will be referred to in the next subsections when we discuss the estimation algorithms.

The likelihood equations in (3.6) are non-linear functions of β. Solving them for β therefore requires iterative methods, whereby an initial approximation of the parameters is used and iterations proceed until the algorithm converges, that is, until the difference between successive approximations is sufficiently small. The three commonly used estimation approaches are outlined below.

3.3.1 The Newton-Raphson Method

The Hessian is the matrix of second derivatives of ℓ and is given by

$$\mathbf{H} = \frac{\partial^2 \ell}{\partial \beta_j \beta_k}.$$

Let $\beta^{(m)}$ denote the approximation for β at the $m'th$ iteration. Then the $(m+1)'st$ approximation for $\hat{\beta}$ is

$$\beta^{(m+1)} = \beta^{(m)} - (\mathbf{H}^{(m)})^{-1}\mathbf{U}^{(m)} \tag{3.8}$$

where $\mathbf{H}^{(m)}$ and $\mathbf{U}^{(m)}$ are the Hessian, $\mathbf{H}$ and the score vector, $\mathbf{U}$, evaluated at $\beta^{(m)}$. The term $\beta^{(0)}$ denotes the initial estimate of $\hat{\beta}$. At each iteration $\beta^{(m)}$ is used to obtain $\mathbf{H}^{(m)}$ and $\mathbf{U}^{(m)}$, which are then used in (3.8) to estimate $\beta^{(m+1)}$. In the next iteration $\beta^{(m+1)}$ is used to obtain $\beta^{(m+2)}$, and this process continues until convergence.

3.3.2 Fisher Scoring

An alternative procedure first suggested by Fisher is to replace minus the Hessian by its expected value, the information matrix. Therefore, the Fisher scoring formula for the $(m+1)'st$ estimate of $\hat{\beta}$ is

$$\beta^{(m+1)} = \beta^{(m)} + (\mathcal{J}^{(m)})^{-1}\mathbf{U}^{(m)} \tag{3.9}$$

where $\mathcal{J}^{(m)}$ and $\mathbf{U}^{(m)}$ are the information matrix $\mathcal{J}$ and the score vector $\mathbf{U}$ evaluated at $\beta^{(m)}$.

Multiplying both sides of equation (3.9) by $\mathcal{J}^{(m)}$ we obtain

$$\mathcal{J}^{(m)}\beta^{(m+1)} = \mathcal{J}^{(m)}\beta^{(m)} + \mathbf{U}^{(m)} \tag{3.10}$$

It can be shown that the right hand side of (3.10) is

$$\mathbf{X}'\mathbf{W}^{(m)}\mathbf{z}^{(m)}$$

where $\mathbf{W}^{(m)}$ is the weight matrix $\mathbf{W}$ with diagonal elements given in (3.5) evaluated at $\beta^{(m)}$ and $\mathbf{z}^{(m)}$ has elements

$$\begin{aligned} z_i^{(m)} &= \sum_j x_{ij}\beta_j^{(m)} + (y_i - \mu_i^{(m)})(\partial\eta_i^{(m)}/\partial\mu_i^{(m)}) \\ &= \eta_i^{(m)} + (y_i - \mu_i^{(m)})(\partial\eta_i^{(m)}/\partial\mu_i^{(m)}) \end{aligned} \tag{3.11}$$

Therefore the $(m+1)'st$ approximation of $\hat{\beta}$ is

$$\beta^{(m+1)} = (\mathbf{X}'\mathbf{W}^{(m)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(m)}\beta^{(m)}. \tag{3.12}$$

3.3.3 Iterative reweighted least squares

The representation in equation (3.12) shows that each iteration of Fisher scoring for numerical evaluation of the maximum likelihood estimate is a weighted least squares regression of the "working" or "adjusted" response variable z on the model matrix $\mathbf{X}$, with working weight matrix $\mathbf{W}$. This is a process known as *iterative reweighted least squares*, and, as shown in the steps leading up to (3.12), it is equivalent to Fisher scoring.

The vector z is a linearized form of the link function at μ, evaluated at y,

$$g(y) \approx g(\mu) + (y-\mu)g'(\mu) = \eta + (y-\mu)\frac{d\eta}{d\mu} = z.$$

The process is iterative because both z and $\mathbf{W}$ depend on the fitted values, for which only current estimates are available (McCullagh and Nelder, 1989).

In the normal linear model z is equal to y and $\mathbf{W}$ is the identity matrix, so that no iterations are required.

The parameter estimates $\hat{\beta}$ are asymptotically normally distributed with mean β and variance-covariance matrix, $\jmath^{-1} = a(\phi)(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$, which is the inverse of the information matrix.

3.3.4 Simplifications for the canonical links

Use of the canonical link results in simplification of the likelihood equations. With the canonical link the linear predictor is equal to the canonical parameter θ_i,

$$\eta_i = \theta_i = \sum_j \beta_j x_{ij} \ . \tag{3.13}$$

For this model,

$$\frac{d\mu_i}{d\eta_i} = \frac{d\mu_i}{d\theta_i} = \frac{d\, b'(\theta_i)}{d\theta_i} = b''(\theta_i).$$

Recall that $b''(\theta_i)$ is the variance function, denoted by V_i. Hence (3.4) simplifies to

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)} \frac{1}{V_i} V_i\, x_{ij} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)} x_{ij}.$$

The second derivatives of the log likelihood then take the form

$$\frac{\partial^2 \ell}{\partial \beta_j\, \partial \beta_h} = -\frac{x_{ij}}{a(\phi)} \left(\frac{\partial \mu_i}{\partial \beta_h}\right).$$

These do not depend on the observations $y_i,\ i = 1, ..., n$, so

$$\frac{\partial^2 \ell}{\partial \beta^2} = E\left(\frac{\partial^2 \ell}{\partial \beta^2}\right).$$

Therefore, under the canonical link, $\mathbf{H} = -\jmath$ and the Newton-Raphson and Fisher Scoring algorithms are identical. Since $a(\phi)$ is constant for all observations, the likelihood estimating equations are

$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \mu_i x_{ij}.$$

3.4 Inference

3.4.1 Measures of goodness of fit

The *deviance* was introduced by Nelder and Wedderburn (1972) as a measure of discrepancy or goodness of fit. It takes the form

$$D = 2[\ell(\mathbf{y}, \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}, \mathbf{y})] ,$$

where $\ell(\mathbf{y}, \mathbf{y})$ is the maximum log likelihood achievable in the *saturated* model allowing one parameter per observation, and $\ell(\hat{\boldsymbol{\mu}}, \mathbf{y})$ is the log likelihood evaluated in the model under consideration with $p + 1$ parameters ($p < n$). The deviance is a measure of the distance between the saturated model and the reduced model under investigation.

The scaled deviance D^* is the deviance expressed as a multiple of the dispersion parameter, ϕ, (assuming ϕ is known) so that

$$D^* = \frac{D}{\phi} = 2[(\ell(\mathbf{y}, \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}, \mathbf{y})]/\phi.$$

The deviance for the Normal-theory linear model is the residual sum of squares, $\sum(y - \hat{\mu})^2$.

The deviance ($\phi = 1$) or scaled deviance ($\phi \neq 1$, but known) measures the closeness of the fit of a model to the data. If a model describes the data well, then its log likelihood will be very close to $\ell(\mathbf{y}, \mathbf{y})$, resulting in a small deviance. Similarly, a large deviance indicates a poor fit to the data. The deviance has an approximate χ^2 distribution with $n - p - 1$ degrees of freedom, where p is the number of explanatory variables in the linear predictor. Generally, if $D \leqslant \chi^2_{\alpha,n-p-1}$ the fitted model is considered adequate. One may also divide the deviance by its degrees of freedom $n-p-1$. If the ratio $D/(n-p-1)$ is close to 1 we may conclude that the fitted model is adequate. A large value of this ratio could mean an incorrectly specified model.

The deviance may also be used for comparing nested models. This method uses deviances in a similar manner to the sums of squares in analysis of variance. It examines the change in deviance or scaled deviance (ϕ is known) between two nested models, therefore providing a useful guide for model selection. The procedure is detailed in McCullagh and Nelder (1989) and can also generally be viewed as the method of conditional deviance. Note that if ϕ is replaced by its unbiased estimate $\hat{\phi}$ when calculating D^*, the resulting scaled deviance is no longer useful as a means of checking the goodness of fit of a given model. However, the change in deviance or scaled deviance for model comparison works well both when ϕ is known as well as when it is replaced by its unbiased estimate, $\hat{\phi}$.

An alternative measure of discrepancy is the *generalized Pearson chi-square statistic*

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

where $V(\hat{\mu}_i)$ is the estimated variance function for the distribution concerned. For the Normal-theory linear model χ^2 is again the residual sum of squares, since $V(\hat{\mu}_i)$ is aprior generally equal to one.

Both the deviance and the generalized Pearson χ^2 - statistic have exact χ^2 distributions with $n - p - 1$ degrees of freedom for Normal-theory linear models. For the other distributions, however, the deviance and the generalized Pearson χ^2 have asymptotic χ^2_{n-p-1} distributions.

It is important to note that asymptotic results may not be relevant to statistics calculated from limited amounts of data, and in these cases either D or χ^2 may be a superior measure of discrepancy. Although χ^2 is sometimes preferred for its more direct interpretation, a general advantage of the deviance as a measure of discrepancy is that it is additive for nested sets of models if maximum likelihood estimates are used (McCullagh and Nelder, 1989).

3.4.2 Hypothesis testing

Likelihood ratio test

The likelihood ratio test is used to compare the fit of two models when one model is nested within the other, that is, when one model has parameters that are a subset of the other model (Cox and Hinkley, 1974). This test makes use of the difference or change in the deviances of the two models.

To see this consider two competing models with corresponding linear predictors given by

$$\begin{array}{ll} \eta_1 = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 & \text{Model 1} \\ \eta_2 = \mathbf{X}_1\beta_1 & \text{Model 2} \end{array}$$

Clearly Model 2 is nested within Model 1. Both models must have the same distribution and the same link function, and should only differ in their number of parameters. Because Model 2 has a reduced number of parameters it will almost always fit the data less well (have a lower log likelihood) than Model 1. However, it is necessary to test whether this difference in model fit is statistically significant, to avoid fitting a complex model unnecessarily. Therefore we shall test $H_0 : \beta_2 = \mathbf{0}$ versus $H_a : \beta_2 \neq \mathbf{0}$.

The deviance for Model 1 is $D_1 = 2[\ell(\mathbf{y},\mathbf{y}) - \ell(\hat{\mu}_1,\mathbf{y})]$ and for Model 2 it is $D_2 = 2[\ell(\mathbf{y},\mathbf{y}) - \ell(\hat{\mu}_2,\mathbf{y})]$.

We use

$$\begin{aligned} D_2 - D_1 &= 2[\ell(\mathbf{y},\mathbf{y}) - \ell(\hat{\mu}_2,\mathbf{y})] - 2[\ell(\mathbf{y},\mathbf{y}) - \ell(\hat{\mu}_1,\mathbf{y})] \\ &= 2[\ell(\hat{\mu}_1,\mathbf{y}) - \ell(\hat{\mu}_2,\mathbf{y})] \\ &= 2\, ln \left[\frac{L(\hat{\mu}_1)}{L(\hat{\mu}_2)}\right]. \end{aligned} \tag{3.14}$$

The statistic in (3.14) is the *likelihood ratio statistic*, which is also the difference in observed deviances. It has an asymptotic χ^2 distribution with degrees of freedom equal to $p_1 - p_2$, the difference in the number of explanatory variables between the two models. The distribution of χ^2 is exact in the case of the Normal-theory linear model.

If $2\, ln \left[\frac{L(\hat{\mu}_1)}{L(\hat{\mu}_2)}\right]$ is greater than the corresponding value from the chi-square table with appropriate degrees of freedom, then one would reject H_0 and conclude that the full model has a significantly smaller deviance than the reduced model, Model 2. We would then choose to fit Model 1 with the additional β_2 parameter. Otherwise we would choose the reduced Model 2.

If $\phi \neq 1$ then the change in deviance no longer matches the likelihood ratio statistic, but one need only to divide it by ϕ for it to be valid. Using the definition of the scaled deviance to compare Models 1 and 2 we would use the statistic

$$T = \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}}_2) - D(\mathbf{y}, \hat{\boldsymbol{\mu}}_1)}{\phi},$$

which is asymptotically χ^2 with q degrees of freedom, where q is the dimension of the sub-vector $\boldsymbol{\beta}_2$. If ϕ is unknown it is common practice to use an estimated value.

Wald test

The Wald test, like the likelihood ratio test, is used to assess the significance of each regression coefficient, β_j, in the model. The test statistic is

$$z_w = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}.$$

As mentioned in subsection 3.3 the variance-covariance matrix of the parameter vector $\hat{\beta}$ is the inverse of the information matrix given by (3.7). The square roots of the diagonal elements of the inverse information matrix are the standard errors of the regression coefficients.

The z_w statistic follows an approximate standard Normal distribution $N(0, 1)$. Many computer packages, including SAS, square the z_w statistic and compare it to a chi-square distribution with one degree of freedom. If $z_w^2 > \chi^2_{\alpha,1}$ one would reject the null hypothesis $H_0 : \beta_j = 0$ and conclude that the predictor variable associated with β_j is significant to the model.

Wald inference may be used to construct confidence intervals for individual regression coefficients. An approximate 100(1 - α)% confidence interval for the $j'th$ regression coefficient is

$$\hat{\beta}_j - z_{\alpha/2}\, s.e.(\hat{\beta}_j) \ < \ \beta_j \ < \ \hat{\beta}_j + z_{\alpha/2}\, s.e.(\hat{\beta}_j).$$

Score test

An alternative to the Wald test is the score test. The score statistic corresponding to the $j'th$ regression coefficient is the derivative of the log likelihood, ℓ, with respect to β_j. That is

$$U_j = \frac{\partial \ell}{\partial \beta_j}.$$

The test statistic takes the form

$$z_s = \frac{\hat{U}_j}{s.e.(\hat{U}_j)},$$

where the square root of the diagonal elements of the information matrix are the standard errors of the score statistics. The reference distribution for z_s is the standard Normal distribution.

3.4.3 Diagnostics

Goodness-of-fit statistics provide an overall measure of the adequacy of a model. Specific aspects of model adequacy, however, are examined by a number of specialized techniques which are collectively referred to as *diagnostics*. These specific aspects include the choice of variance function and link function and terms in the linear predictor, as well as the identification of outlying values requiring further investigation. The analysis of residuals is central to diagnostics.

3.4.3.1. The hat matrix

The estimate of the expected value of the response in a generalized linear model is

$$\begin{aligned} \widehat{E(Y_i)} &= \hat{\mu}_i \\ &= \hat{y}_i \end{aligned}$$

This is commonly known as the *fitted value*. It can be shown that

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

where $\mathbf{H}$ is called the *hat matrix*. This matrix is symmetric and idempotent. Its elements cannot exceed 1 and its trace is equal to p, the number of explanatory variables in the fitted model. It is explicitly given by

$$\mathbf{H} = \mathbf{W}^{1/2} X (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}^{1/2}.$$

In the above expression $\mathbf{W}$ is the weight matrix with elements given by (3.5) and $\mathbf{X}$ is the n by $p+1$ design matrix of explanatory variables. The diagonal elements of $\mathbf{H}$, denoted by h_{ii}, are called the *leverage measures*. They are used to detect influential observations.

3.4.3.2. Types of residuals

Raw residual

The $i'th$ raw residual, $y_i - \hat{\mu}_i$, is the crude difference between the $i'th$ observation and its fitted value. It therefore indicates how well the model fits each observation. Raw residuals are not easily comparable however, because each of the observed values, y_i, has a different standard error.

Pearson residual

Raw residuals are made more comparable by dividing them by the estimated standard deviation of y_i, so that we have

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \ .$$

This is called the *Pearson residual*, since the sum of the squares of these residuals form the Pearson χ^2-statistic.

Pearson residuals do not possess approximate unit variance, and so they must be standardized further. A better standardization is achieved by dividing the raw residual by its standard error, $s.e.(y_i - \hat{\mu}_i) = \sqrt{V(\hat{\mu}_i)(1 - h_{ii})}$, where h_{ii} are the diagonal elements of the hat matrix.

Hence the *standardized Pearson residual* is

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1 - h_{ii})}} \ .$$

Deviance residual

Each observed value of the response contributes a component deviance d_i to the (scaled) deviance, so that $D = \sum d_i$. For each observation y_i we define

$$d_i = 2 \int_{\hat{\mu}_i}^{y_i} \frac{y_i - s}{V(s)}\, ds$$

Then the *standardized deviance residual* is defined as

$$r_{D_i} = \frac{\text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}}{\sqrt{1 - h_{ii}}} \ .$$

The deviance is a measure of the overall goodness-of-fit of a model, and the deviance residuals are a means of identifying which individual observations contribute most to the lack of fit.

Deviance residuals approximate the normal distribution better than Pearson residuals, and are therefore better suited to check agreement to distributional assumptions in model checking statistics.

Other types of residuals also exist and two of them are briefly described below.

Anscombe residual

In order to make residuals of non-Normal distributions have similar properties to those of the Normal-theory residuals, Anscombe (1953) suggested defining a residual using a function $A(y_i)$ instead of y_i, where $A(.)$ is chosen so as to make the distribution of $A(Y)$ approximately Normal (McCullagh and Nelder, 1989). An *Anscombe residual* takes the form

$$r_{A_i} = \frac{A(y_i) - A(\hat{\mu}_i)}{s.e.[A(y_i) - A(\hat{\mu}_i)]}.$$

According to Barndorff-Nielsen (1978), the function $A(.)$ for a generalized linear model is given by

$$A(.) = \int \frac{d\mu}{V^{1/3}(\mu)}.$$

The standard error of $A(y_i)$ can be expressed as $A'(\mu_i)\sqrt{V(\mu_i)}$. McCullagh and Nelder (1989) provide further detail on Anscombe residuals for various distributions.

Likelihood residual

Comparing the deviance for a model fitted for the complete set of observations with the deviance when each observation, in turn, is omitted, gives rise to the *likelihood residual*,

$$r_{L_i} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{h_{ii}\, r_{P_i}^2 + (1 - h_{ii})r_{D_i}^2}\ ,$$

a weighted combination of the deviance and Pearson residuals (Collett, 1991).

Studies by Williams (1984) and Pierce and Schafer (1986) suggest that the values taken by the standardized deviance residuals, the likelihood residuals and the Anscombe residuals are usually very similar. Based on these studies Collett (1991) suggests that, since Anscombe residuals require more difficult computation, there is no great advantage in using them. Pearson residuals often have a skewed distribution and so may fail to have similar properties to those of Normal-theory residuals (McCullagh and Nelder, 1989). In addition they may not rank extreme observations appropriately. Collett (1991) therefore advises using either standardized deviance residuals or likelihood residuals for routine model checking and diagnostics.

3.4.3.3. Residual plots and model adequacy

Residuals can be plotted against a variety of statistics and other indices, each providing information on specific aspects of model adequacy. Lindsey (1997) suggests that the analysis of residuals is useful for relatively small sample sizes of at most 100 observations, and when the model fitted is far from being saturated.

There are a number of residual plot types but the most commonly used type is the normal probability plot defined below. To use the normal probability plot an assumption is made that the residuals follow the normal distribution.

A Normal probability plot shows the standardized residuals arranged in ascending order against an approximation to their expected values, under the assumption that they are normally distributed. These expected values are given by $\Phi^{-1}[(i-\frac{3}{8})/(n+\frac{1}{4})]$, where $i=1,...,n$. If the model fits well the plot should show a scatter of points around a straight line at 45^o. Any curvature or systematic deviations from the straight line indicate that the residuals do not have an approximate Normal distribution.

There are other methods of assessing model adequacy, as explained below.

Checking the form of the linear predictor may be done by plotting the standardized residuals against the fitted values $\hat{\mu}_i$. If the model fits well there should be no pattern in the plot. The occurrence of a systematic pattern indicates a poor choice of explanatory variables or incorrect functions of the explanatory variables in the model, including missing interaction terms. This can be further investigated by plotting the residuals against the explanatory variables. A plot of residuals against a particular explanatory variable will determine if that variable needs to be transformed. For example, $\mathbf{X}^2$ might contribute to a better model fit than **X**. Residuals may also be plotted against potential explanatory variables not included in the model. A trend in the plot shows that the explanatory variable in question does influence the response and should therefore be included in the model.

Outliers are an important reason for checking model fit or lack of fit. Observations which are surprisingly distant from the remaining observations in the data are known as *outliers* (Collett, 1991). They can either be the result of measurement error or they can just be naturally occurring rare values. Outliers can be detected using an *index plot*, which shows the residuals against their corresponding observation number or index. When outliers are detected their effect on the results of the analysis are determined by re-fitting the model after excluding them. If the results do not differ substantially for models with and without the outliers, then one need not be too concerned about them. However, if the outliers do affect inferences drawn from the data, an in-depth investigation is required into the cause of these extreme observations. Based on this one can decide on whether to include them, to revise the model or to omit them.

The analysis of influential observations is an important aspect of assessing or finding the best model. An observation is said to be influential if, when omitted or changed by a small amount, will substantially change the parameter estimates and, hence, the model fit. An influential observation may not necessarily be an outlier because it may be close to the main body of the data and thus have a small residual. When dealing with outliers it is important to pay attention to outliers that are influential than those which are not.

There are several measures of influence. Two of these are given below.

Leverage is a measure of how influential an observation is. The diagonal elements of the hat matrix, h_{ii}, are the leverage measures for the observation in a model. A plot of h_{ii} against index values indicates influential observations which require further investigation.

Cook's Distance (Cook, 1977) is a statistic which indicates how each observation affects the complete set of parameter estimates. It measures the squared distance between $\hat{\boldsymbol{\beta}}$, the estimated full parameter vector, and $\hat{\boldsymbol{\beta}}_i$, the estimated parameter vector with the $i'th$ observation removed. Cook's Distance is approximated by

$$C_i = \frac{1}{p}\, r_{D_i} \frac{h_{ii}}{1 - h_{ii}}.$$

Plots of Cook's distance against index number are also useful in locating observations with high influence.

3.5 Modelling binomial data

This thesis is based on the analysis of binary and binomial type data. Due to the heavy reliance on the binomial model in the current thesis we briefly review the GLM for a binomial response. The type of data that is considered in the thesis is current status data, giving the HIV disease status of an individual at a given age a_i at the time of testing; implying that if a test is positive then the actual age at infection is $a_i^* \leqslant a_i$. The aim is to use the data to model the prevalence of HIV and other disease rates.

3.5.1 Distinguishing between binary and binomial data

Consider the case where the response, Y, on an experimental unit can take on only one of two possible values. For example, the outcome can either be a 'success', denoted by 1 or a 'failure' denoted by 0. Let π_i and $1-\pi_i$ be the probabilities of 'success' and 'failure' respectively, on the $i'th$ $(i = 1, ..., N)$ experimental or observational unit. We may write this as

$$\Pr(Y_i = 1) = \pi_i \quad \text{and} \quad \Pr(Y_i = 0) = 1 - \pi_i \ .$$

The response on a single individual or experimental unit is a *Bernoulli outcome*. Ungrouped data which lists observations by individual experimental unit are known as *binary* data.

Associated with a response variable is a set of explanatory variables $x_1, ..., x_p$. Each possible combination of values of the explanatory variables is called a covariate class. When binary data are grouped by covariate class, the responses $0 \leqslant y_i \leqslant m_i$ are then the number of 'successes' among the m_i subjects in the $i'th$ covariate class. Then y_i/m_i denotes the observed response probability of 'success' in each covariate class. The number of subjects per covariate class sum up to the total number of subjects in the study, so that $m_1 + m_2 + ... + m_n = N$.

This form of grouped data is referred to as *binomial* data. It is then merely the sum of independent and homogeneous Bernoulli trials grouped by covariate class. The responses y_i follow a binomial distribution with probability function

$$P(Y_i = y_i) = \binom{m_i}{y_i} \ \pi_i^{y_i} \ (1 - \pi_i)^{m_i - y_i} \qquad y_i = 0, 1, ..., m_i \tag{3.15}$$

where π_i is the true probability of success in the $i'th$ covariate class.

It is important to distinguish between grouped (binomial) data and ungrouped (binary) data when conducting analysis. Methods of analysis involving the Normal approximation can be applied to binomial data, but not to binary data. In addition, for models of binomial data, asymptotic approximations are based on either of the asymptotes, $m \to \infty$ or $N \to \infty$, whereas for binary data only the latter asymptote applies (McCullagh and Nelder, 1989).

3.5.2 The model

Models for binary or binomial variables are used to describe the effect of a set of explanatory variables $(x_1, ..., x_p)$ on the response probability π.

They are generalized linear models of the form

$$g(\pi) = \eta = \sum_{j=1}^{p} \mathbf{x}_j \beta_j \qquad\qquad i = 1, ..., N \tag{3.16}$$

The three possible choices of link function $g(\pi)$ are:

- Logit or logistic function, $g(\pi) = log(\frac{\pi}{1-\pi})$.
- Probit function, $g(\pi) = \Phi^{-1}(\pi)$
- Complementary log-log function, $g(\pi) = log[-log(1-\pi)]$

3.5.3 Fitting the linear logistic model to binomial data

Suppose the responses y_i, $i = 1, 2, ..., n$, are the observed values of independent random variables $Y_1, Y_2, ..., Y_n$, where Y_i has the binomial distribution with index m_i and parameter π_i.

We want to fit a generalized linear model of the form given in (3.16) with the logit link $g(\pi_i) = log(\pi_i \ / \ (1-\pi_i))$.

The binomial distribution $B(m_i, \pi_i)$ belongs to the exponential family of Nelder and Wedderburn (1972) because the binomial probability distribution function (pdf) in (3.15) is equivalent to

$$f_Y(y_i; \pi_i) = exp\left(y_i \ log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \ log(1-\pi_i) + log\binom{m_i}{y_i} \right) \tag{3.17}$$

which has the general exponential form given in (3.1).

From the coefficient of y_i we see that the *canonical parameter* is the logit of π_i, namely

$$\theta_i = log\left(\frac{\pi_i}{1-\pi_i}\right) .$$

Solving for π_i in terms of θ_i leads to the solution

$$\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} \quad \text{and} \quad 1-\pi_i = \frac{1}{1+e^{\theta_i}} \ .$$

The *cumulant function* $b(\theta_i)$ is $-m_i \ log(1-\pi_i)$ or equivalently $m_i \ log(1+e^{\theta_i})$. The remaining term is $c(y_i, \phi) = log\binom{m_i}{y_i}$. Note that $a(\phi)=1$, so that $\phi = 1$.

It follows that the mean and variance relationship is given by

$$\mu_i = b'(\theta_i) = m_i \ \frac{e^{\theta_i}}{1+e^{\theta_i}} = m_i \pi_i$$

$$Var(y_i) = a(\phi) \ b''(\theta_i) = m_i \ \frac{e^{\theta_i}}{(1+e^{\theta_i})^2} = m_i\pi_i(1-\pi_i) \ ;$$

which is in agreement with the mean and variance of a binomial distribution $B(m_i, \pi_i)$.

3.5.4 Parameter estimation

We will follow the method for parameter estimation in a generalized linear model, as outlined in Section 3.3.

From (3.17) we conclude that the log likelihood for n independent binomial observations is given by

$$\ell(\pi; \mathbf{y}) = \sum_{i=1}^{n} \left[y_i \ log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \ log(1-\pi_i) \right] .$$

The constant function $\sum log\binom{m_i}{y_i}$ does not include π_i and is omitted from the expression as it plays no role in the estimation of model parameters.

Then the derivative of the log likelihood with respect to β_j is

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{\partial \ell}{\partial \pi_i} \frac{d\pi_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} \right) .$$

Now

$$\frac{\partial \ell}{\partial \pi_i} = \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)}$$

and

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \ ,$$

so that

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - m_i \pi_i}{\pi_i(1-\pi_i)} \ \frac{d\pi_i}{d\eta_i} \, x_{ij} \ . \qquad (3.18)$$

The Fisher information for β is

$$\begin{aligned} -E\left[\frac{\partial^2 \ell}{\partial \beta_j \, \partial \beta_k}\right] &= \sum_{i=1}^{n} \ \frac{m_i}{\pi_i(1-\pi_i)} \ \frac{\partial \pi_i}{\partial \beta_j} \ \frac{\partial \pi_i}{\partial \beta_k} \\ &= \sum_{i=1}^{n} \ \frac{m_i}{\pi_i(1-\pi_i)} \left(\frac{d\pi_i}{d\eta_i}\right)^2 x_{ij} \, x_{ik} \\ &= \sum_{i=1}^{n} \ w_i \, x_{ij} \, x_{ik} \end{aligned}$$

or in matrix notation, the information matrix is $\mathbf{X}'\mathbf{W}\mathbf{X}$ where

$$\mathbf{W} = \text{diag}\left(m_i \left(\frac{d\pi_i}{d\eta_i}\right)^2 \frac{1}{\pi_i(1-\pi_i)} \right) . \qquad (3.19)$$

The logit link is the canonical link for the binomial distribution. When this link is used we have $\theta_i = \eta_i = log(\pi_i/1-\pi_i)$, and therefore

$$\frac{\partial \eta_i}{\partial \pi_i} = \frac{1}{\pi_i(1-\pi_i)} \ .$$

Thus in the case of the linear logistic model (3.18) reduces to

$$\frac{\partial \ell}{\partial \beta} = \mathbf{X}'(\mathbf{Y} - \boldsymbol{\mu})$$

in matrix notation, and the weight matrix in the Fisher information reduces to

$$\mathbf{W} = \text{diag}\ (m_i \pi_i (1-\pi_i)).$$

Maximum likelihood estimates of the parameters β can then be calculated through an iterative procedure, using

$$\beta^{(k+1)} = (\mathbf{X}'\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(k)}\mathbf{z}^{(k)} \ ;$$

where $\mathbf{W}^{(k)}$ is $\mathbf{W}$ evaluated at the $k'th$ iteration and the adjusted variate $\mathbf{z}^{(k)}$ has elements

$$z_i^{(k)} = log\left(\frac{\pi_i^{(k)}}{1-\pi_i^{(k)}}\right) + \frac{y_i - m_i\pi_i^{(k)}}{m_i\pi_i^{(k)}(1-\pi_i^{(k)})}.$$

3.6 Application to Vulindlela antenatal clinic data

Generalized linear models for binomially distributed variables were fitted to the data collected from pregnant women attending antenatal clinics in Vulindlela, KwaZulu-Natal. The response variable for each subject was the HIV status, which can take on only one of two values, either 1 if the subject tested HIV positive or 0 if HIV negative. The response probability is therefore the probability of a subject being ever HIV positive by that age and time. The subjects' *age*, the *year* in which they were tested, the *antenatal clinic* they attended, their *partners' age* and their *number of previous pregnancies* were the explanatory variables in the data. A detailed description of the data is given in Chapter 2.

When the subjects are grouped by covariate class the responses, y_i, are then the number of HIV positive subjects out of the m_i subjects in the $i'th$ covariate class. The response probability y_i/m_i is the estimated proportion of HIV positive subjects in the $i'th$ covariate class. In other words it is the estimated *prevalence rate* for this covariate class. Note that y_i/m_i estimates the true prevalence rate π_i. In other words $\hat{\pi}_i = y_i/m_i$, $i = 1, \ldots, k$, where k is the number of covariate classes.

3.6.1 Grouped data

The subjects were grouped by individual age-year combinations. A total of 177 age-year covariate classes were formed. Figure 3.1 shows the observed prevalence rates, given by y_i/m_i, plotted against age.

A linear logistic model of the form

$$g(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\text{age} + \beta_2\text{year} \qquad (3.20)$$

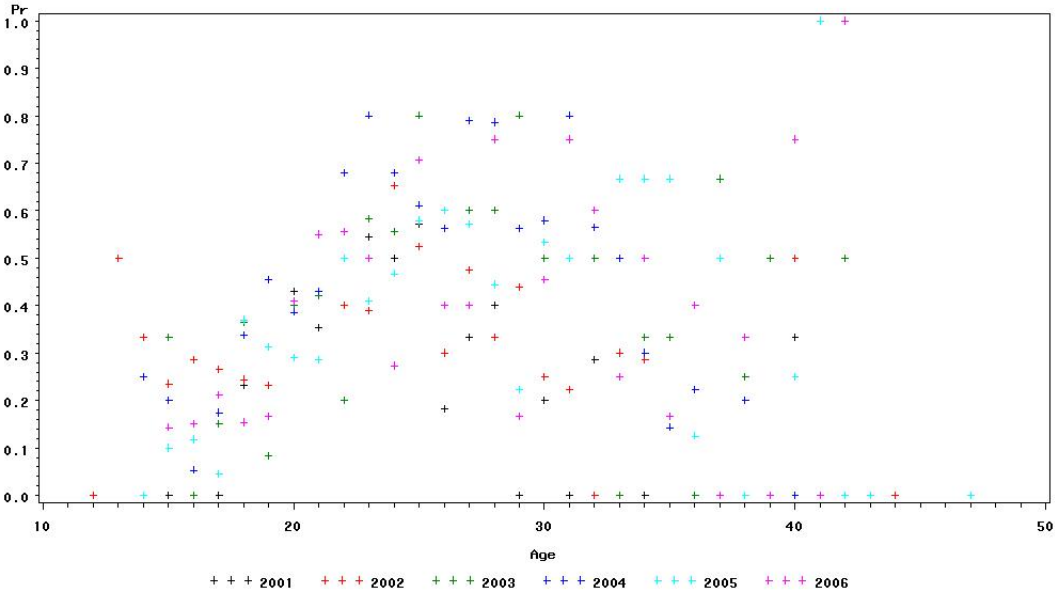


Figure 3.1: Scatter plot of the observed prevalence rates by age for all years

was fitted. Both age and year were assumed to be continuous variables. Results showed that the intercept as well as the model coefficients for age and year were all significant at $\alpha = 0.05$. The fitted prevalence rates against age for each year were graphed. The resulting logistic curves were straight lines which increased with age. The slopes for each year were the same, with each year's fitted values being proportionally higher than the previous year's.

The deviance and the Pearson chi-square statistic for the model in equation (3.20) is 356.6 and 310.08 respectively. The Deviance/df value of 2.0494 is far greater than 1, indicating that the model is not a very good fit to the data. Models with the probit and complementary log-log link functions yielded similar linear slopes. From Figure 3.1 it is evident that a more curvilinear model would fit the data better.

We noted that the best fitting linear predictor required that we take log and square root transformations of some covariates (age and time) rather than leaving them in their original scale. This is probably due to the complex dependance of the probability of being HIV positive on age and time.

Several models using various transformations of the age and year covariates were fitted. Likelihood ratio and Wald tests were used to determine the statistical significance of each covariate in a model, and the deviance, Pearson Chi-square statistic and the analysis of residuals was used to compare the goodness-of-fit between models.

The model found to be the best fit to the data is given by

$$g(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

where age and year are assumed to be continuous and

$X_1 =$ year
$X_2 =$ log(year)
$X_3 =$ log(age)
$X_4 = \sqrt{\text{age}}$
$X_5 =$ age x year (the interaction between age and year).

From the output shown in Table 3.1, we see that the deviance is 191.0438 and the deviance/df = 1.1172, which is very close to 1. All the model coefficients β_i are statistically significant at the 5% significance level. Their estimates, standard errors and confidence limits are displayed in Table 3.2.

Table 3.1: Criterion for assessing goodness of fit for the prevalence model with logit link and transformations of the age and year covariates

Criterion	DF	Value	Value/DF
Deviance	171	191.0438	1.1172
Scaled Deviance	171	191.0438	1.1172
Pearson Chi-Square	171	190.4962	1.1140
Scaled Pearson X2	171	190.4962	1.1140
Log Likelihood		-1243.5653	

The fitted prevalence rates by age for each of the years 2001 to 2006 are graphed in Figure 3.2. These parabolic slopes show that the prevalence in each year increases with age to peaks at $36\% - 57\%$ in the mid to late twenties and then steadily declines toward the early forties. The prevalence rates in 2001 are notably lower than those of the other years across all ages. For ages up to about

Table 3.2: Parameter estimates for the prevalence model with logit link and transformations of the age and year covariates

Parameter	DF	Estimate	Std Error	Wald 95% Upper	Wald 95% Lower	Chi-Square	Pr>ChiSq
Intercept	1	-42.4957	3.7824	-49.9091	-35.0824	126.23	<.0001
Year	1	-0.8110	0.1876	-1.1787	-0.4433	18.69	<.0001
log year	1	1.3919	0.3876	0.6322	2.1516	12.90	0.0003
log age	1	36.3169	3.3020	29.8451	42.7887	120.96	<.0001
Sqrt(age)	1	-14.9170	1.4106	-17.6818	-12.1522	111.82	<.0001
Year x Age	1	0.0177	0.0056	0.0068	0.0286	10.19	0.0014
Scale	0	1.0000	0.0000	1.0000	1.0000		

30 the prevalence slopes are very similar in all the years, and after age 30 the prevalence in the latter years is higher than those in the earlier ones. There also seems to be a shift in the peak prevalence over age. The 2001 peak is sooner in age than the subsequent years.

The parabolic prevalence curves in Figure 3.2 are similar in shape to those of Gouws (2006), who modelled HIV prevalence among women attending antenatal clinics using data from the national ANC surveys between 1995 and 2004, as well ANC data from Hlabisa, a rural district of Northern KwaZulu-Natal, for the years 1997, 1998, 1999 and 2001. The estimated prevalence from national antenatal clinics between 2001 and 2004, rose with age to highs of between 31.4% and 37.1% in the mid-twenties, and then declined to values of below 18% in the age group 40-49. The age at which the peak prevalence occurred had shifted over the study period, from around 23 years in 1995, to 25.4 years in 2001 and then to 26.7 years in 2004. Estimated prevalence for the Vulindlela antenatal data also shows a shift in the peak prevalence over age. In 2001 the prevalence estimates peak at age 24, whereas in 2005-2006 prevalence peaks at around the ages 27-28 (as shown in Figure 3.2). A possible explanation for this shift is due to easy access to antiretroviral treatment (ARV's) for HIV infected individuals as from 2004, when the South African government implemented a massive ARV roll out policy.

The prevalence curves for the Hlabisa district showed similar trends with age. Prevalence estimates for each year peaked at around the ages 22-26, with a shift in peak prevalence over age for each subsequent year. In 2001, the estimated prevalence for Hlabisa rose to a high of 51% at around age 25, before declining to around 16% in the late forties (Gouws, 2006). The estimated prevalence rates for Hlabisa, as well as the estimates obtained in this thesis for the

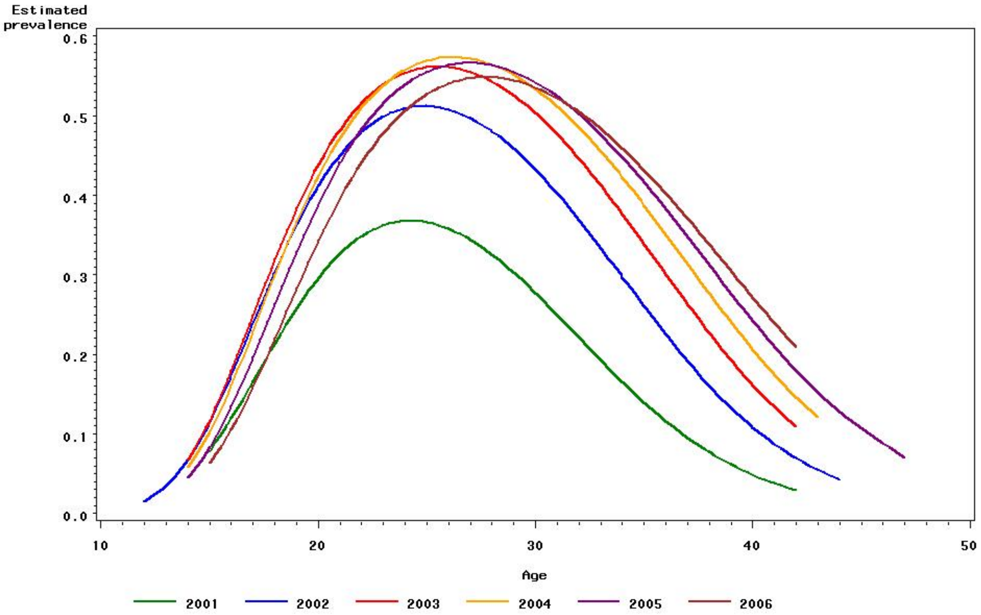


Figure 3.2: Fitted prevalence rates by age for each of the years 2001 to 2006, using a model with logit link and transformations of the age and year covariates.

Vulindlela area, are both higher than the prevalence estimates from the national antenatal surveys, for all age groups prior to age 40. This reflects the high HIV prevalence in the KwaZulu-Natal province.

Alternative models with year being treated as a categorical variable were also fitted. Age and transformations of the age variable were still treated as continuous. We then have five β_i coefficients, one for each of the years 2001 to 2005 with 2006 being the reference category. However, results showed that the coefficients corresponding to each year were not statistically significant ($p > 0.05\%$), with the exception of year 2001. The intercept, age and log(age) were however statistically significant.

Chapter 4

Methods for estimating incidence and prevalence

The prevalence of a disease is the proportion of all disease cases within a population at a given time. The incidence is the rate at which new cases of the disease occur. It is given by

$$\text{Incidence rate} = \frac{\text{No. of new cases of the disease during a given time period}}{\text{No. of uninfected individuals in the population x Time period of observation}}. \tag{4.1}$$

Prevalence includes both new and existing cases of the disease. There is therefore a strong interrelation between the prevalence and incidence, since any prevalent case means that a new infection has occurred before.

Both incidence and the force of infection are parameters used to describe the rate of new infections of a disease. The force of infection is defined as the rate (per unit time) at which susceptible individuals become infected with the disease.

This chapter will review some of the well-known methods for estimating the incidence of a disease from cross-sectional prevalence data. Note that all the methods discussed assume the disease to be irreversible.

4.1 Relation between incidence and prevalence

Freeman and Hutchison (1980) showed that when a population is in the *steady-state*, the prevalence, incidence and the duration of the disease are interrelated in such a way that any two of these measures may be used to calculate the third.

A steady-state population is defined as one in which: 1) Incidence and the distribution of durations of the disease remain constant over time; and 2) the population of affected and unaffected individuals remains constant over time (Freeman and Hutchison, 1980). The second steady-state assumption means that new individuals are added to the susceptible population at the same rate at which they are removed from the susceptible population, and that new cases are added to the diseased population at the same rate at which they are removed from the diseased population.

If D denotes the duration of the disease i.e. the length of time from infection with the disease to death (in the case of irreversible diseases), then the total number of infected individuals in the population is given by

$$P_N = P.\, N = I.\, E(D) \ , \tag{4.2}$$

where P is the prevalence or proportion of infected individuals, N is the total population size, I is the number of incident cases per unit time and $E(D)$ is the expected value of disease duration from a distribution of durations.

The interrelation of the *prevalence*, P, and the *incidence rate*, i, is given by

$$P = \frac{i.\, E(D)}{1 + i.\, E(D)} \qquad \Leftrightarrow \qquad \frac{P}{1-P} = i.\, E(D) \ \ . \tag{4.3}$$

The term $P/(1-P)$ is the *prevalence odds*, the odds of having the disease. It gives the probability of being disease prevalent relative to being disease free. From equation (4.3) we see that the prevalence odds can be expressed as the product of the incidence rate and the expected duration of disease. We can then obtain the incidence rate as

$$i = \frac{P}{(1-P).\, E(D)} \ . \tag{4.4}$$

Hence, under the assumption of a steady-state population, incidence can be obtained from prevalence data, when the expected duration of the disease is known. Equation (4.2) is exact when steady-state assumption (1) is met, that is, when the incidence and the distribution of disease durations remain constant over time. Furthermore, the constancy of both these quantities must have held for at least as long as the longest disease duration included in the distribution. This is because the number of infected individuals in a time period could be made up of incident cases only. Equations (4.3) and (4.4) are exact when both the steady-state assumptions are met.

The interrelations above were used by Brookmeyer and Quinn (1995) to estimate HIV incidence from a cross-sectional prevalence survey in India. Over 1900 individuals were tested for HIV, and the seronegatives were then tested for the p24 antigen. This antigen is detected in individuals in the preantibody or window period (the time from HIV infection to seroconversion). The preantibody period is relatively short and thus individuals in this period are likely to have been infected recently. The expected value of the duration of the p24 antigen period before seroconversion is then estimated from the data. This expected duration, along with the prevalence of the p24 antigenemia, are used in the above equations to estimate HIV incidence.

The models of Freeman and Hutchison (1980) do not model incidence as a function of age. However, the equations relating incidence, prevalence and duration may be applied to different age intervals. An extension of these models may be found in Alho (1992), for the case of exponentially growing populations.

4.2 Age specific incidence and the effect of disease specific mortality

4.2.1 Equal mortality for infected and uninfected individuals

Under the assumption of equal mortality for infected and uninfected individuals, Leske et al (1981) proposed a deterministic method to estimate incidence from age-specific prevalence data. An additional assumption of the model is that the disease incidence and the population composition (regarding disease risk factors) remain constant over time.

Consider the total number of disease prevalent individuals in a population at the beginning of age interval $k + 1$. This number includes (i) the individuals who were already infected with the disease at the beginning of age interval k and who survive to the end of the interval, plus (ii) those who become infected with the disease during interval k (incident cases) and survive to the end of the interval. This is expressed as

$$\begin{aligned} N_{k+1}\, P_{k+1} &= N_k P_k(1-q_k) \;+\; I_k(N_k - N_k P_k)(1-q_k) \\ &= N_k P_k(1-q_k) + I_k N_k(1-P_k)(1-q_k) \\ &= N_k[P_k(1-q_k) + I_k(1-P_k)(1-q_k)] \end{aligned} \tag{4.5}$$

where

$N_k =$ population size at the beginning of age interval k

$P_k =$ prevalence (proportion infected) at beginning of age interval k, $0 \leqslant P_k \leqslant 1$

$q_k =$ probability of dying during age interval k, $0 \leqslant q_k \leqslant 1$

$I_k =$ probability of becoming infected with the disease during age interval k, $0 \leqslant I_k \leqslant 1$

The term $N_k - N_k P_k$ represents the number of susceptible or uninfected individuals at the beginning of age interval k while $N_k P_k$ represents the number of disease prevalent individuals at the beginning of age interval k. The proportion who survive to the end of the $k'th$ age interval is represented by $(1-q_k)$.

Then by rearranging (4.5) and solving for I_k we have

$$I_k = \frac{N_{k+1}P_{k+1} \;-\; N_k P_k(1-q_k)}{(N_k - N_k P_k)(1-q_k)}\ .$$

If mortality risk is assumed equal for infected and uninfected individuals, then $N_{k+1} = N_k(1-q_k)$, and hence

$$I_k = \frac{P_{k+1} - P_k}{1 - P_k}\ . \tag{4.6}$$

For diseases with very low prevalence the denominator in (4.6) will be close to 1, so that in these cases I_k can be approximated by simply subtracting the prevalence proportions of successive age intervals.

The probability of being infected with the disease during age interval k, I_k, is also sometimes referred to as the *cumulative incidence* (Langohr, 1999). It is a probability and not a rate and hence differs from the *incidence rate*, i_k (Leske et al, 1981). Cumulative incidence may be viewed as the crude probability of developing the disease, as it does not account for the probability of dying. The incidence rate, on the other hand, is the net risk of developing the disease in the presence of competing mortality (Ederer, 1964), and it is therefore the parameter of greater epidemiological interest. The incidence rate, i_k, can however be estimated from I_k.

By definition, i_k is the number of new cases of disease in age interval k divided by the average population at risk (Elandt-Johnson, 1975). Then the incidence rate, i_k, is given by

$$i_k = \frac{I_k(N_k - N_k P_k)(1 - q_k) \; + \; I_k(N_k - N_k P_k) \; q_k \; f_k}{N_k - N_k P_k - [(N_k - N_k P_k)\frac{1}{2} q_k]} \; , \qquad (4.7)$$

where f_k is the proportion of new cases in age interval k who do not survive the interval.

Thus the numerator in (4.7) includes new cases of the disease in age interval k who survive to the end of the interval *and* the new cases in age interval k who die during the interval. It is in this way that i_k accounts for competing mortality, and differs from the cumulative incidence, I_k, in that I_k includes only those new cases of the disease who are alive at the end of the age interval (i.e. those surviving the interval). The denominator in (4.7) implies that periods from the beginning of the age interval until death follow a uniform distribution.

Equation (4.7) can be simplified so that

$$i_k = \frac{I_k[1 - q_k(1 - f_k)]}{1 - \frac{1}{2} q_k} \; . \qquad (4.8)$$

When estimating i_k two alternative assumptions can be made, regarding the distributions of the times until death and until acquiring infection:

1. If the times until death and until infection with the disease are assumed equal, then $f_k = \frac{1}{2}$; and the incidence rate equals the cumulative incidence, $i_k = I_k$. In this case, the incidence rate can then be estimated using

(4.6). However, as explained earlier, there are conceptual differences between I_k and i_k.

2. Alternatively, if the times until death and until infection are assumed to follow independent exponential distributions, then i_k takes the form

$$i_k = \frac{ln(1 - I_k)}{ln(1 - q_k) + ln(1 - I_k)} \left[\frac{1 - (1 - I_k)(1 - q_k)}{1 - \frac{1}{2} q_k} \right] . \qquad (4.9)$$

Details on the derivation of (4.9) can be found in Leske et al (1981). The delta method, also explained in Leske et al (1981), is used to obtain variance estimates of $\hat{i}_k$.

Thus, Leske et al (1981) showed that depending on the assumptions made about the survival times and disease-free times, the incidence rate, i_k, can be estimated using age-specific prevalence only, or using both age-specific prevalence and mortality.

Leske et al (1981) and Podgor et al (1983) applied these methods to age-specific prevalence data for various eye diseases, such as open angle glaucoma, in areas of the United Kingdom and the United States respectively.

4.2.2 Differential mortality for infected and uninfected individuals

Consider the case of differential mortality for infected and uninfected individuals (q'_k and q_k respectively). Then, from equation (4.5), the number of infected individuals in the population at the beginning of age interval $k + 1$ is given by

$$N_{k+1}\, P_{k+1} = N_k[P_k(1 - q'_k) + I_k(1 - P_k)(1 - q'_k)] \qquad (4.10)$$

Similarly, the total number of uninfected individuals at the beginning of age interval $k + 1$ is

$$\begin{aligned} N_{k+1} - N_{k+1}P_{k+1} &= (N_k - N_k P_k)(1 - I_k)(1 - q_k) \qquad (4.11) \\ &= N_k(1 - P_k)(1 - I_k)(1 - q_k) \end{aligned}$$

Now consider three types of time periods; from the beginning of age interval k to:

– death in the absence of disease;

– infection with the disease; and

– death in the presence of disease.

Podgor and Leske (1986) assumed that these three time periods followed independent exponential distributions with parameters λ_1, λ_2 and λ_3 respectively. The parameter λ_2 is equivalent to the disease incidence rate, i_k. Similarly, λ_3 may be interpreted as the disease-specific mortality rate and λ_1 as the mortality rate for uninfected individuals i.e. the background mortality rate.

Transforming equations (4.10) and (4.11) gives

$$\frac{1-P_k}{1-P_{k+1}} P_{k+1}\, e^{-(\lambda_1+\lambda_2)} = P_k\, e^{-\lambda_3} + (1-P_k)[e^{-\lambda_3} - e^{-(\lambda_1+\lambda_2)}]\, \frac{\lambda_2}{\lambda_1+\lambda_2-\lambda_3} \quad (4.12)$$

It then follows that, for age interval k, if λ_1 and λ_3 are known, then the estimates for P_k and P_{k+1} may be used to estimate λ_2. Equation (4.12) is non-linear in λ_2, and can therefore be solved by standard methods such as the Newton-Raphson algorithm. The estimated variances of λ_2 are obtained by the delta method (Podgor and Leske, 1986).

If $\lambda_1 = \lambda_3$ and substituting $I_k = 1 - e^{-\lambda_2}$, the model reduces to (4.6):

$$I_k = \frac{P_{k+1} - P_k}{1 - P_k}\ ,$$

which is the cumulative incidence under the model which assumes that disease does not affect mortality risk. The cumulative incidence is seen as the crude probability of acquiring infection as it does not account for mortality risk. Under the above parametrization, the incidence rate is given by

$$\lambda_2 = i_k = -ln\left[\frac{1-P_{k+1}}{1-P_k}\right]\ .$$

Podgor and Leske (1986) developed and applied this approach to estimate the incidence of various eye diseases in the United Kingdom. In terms of modelling disease progression $\lambda_2 = i_k$ is also known as the *hazard of infection* or the *force of infection*. Methods of estimating the force of infection from prevalence data are presented and applied in Chapter 5.

4.3 Probability framework for studying age-specific incidence and prevalence

This section describes the basic probability framework for studying the relation between age-specific incidence and prevalence in a cross-sectional sample. The framework, which is based on a simple three-state process, was discussed by Keiding (1991) in his much cited paper "Age-Specific Incidence and Prevalence: A Statistical Perspective" in the Journal of the Royal Statistical Society. This probability framework forms the basis of many of the methods developed for estimating incidence from prevalence data.

Let H denote the healthy state, or equivalently the susceptible state, when an individual is alive and disease-free, I the disease state and D dead (see Figure 4.1). Individuals are born into the healthy state. Then, during their lifetimes individuals can transit out of the healthy state either into state I, and then to state D, or directly to state D. The rates or intensities associated with the H → I and H → D transitions are $\theta(a,t)$ and $\mu(a,t)$ respectively, which are dependent on age and calendar time, while the rate $v(a,t,d)$ associated with the I → D transition is additionally dependent on the duration in state I. The transition intensity $\mu(a,t)$ can also be viewed as the natural mortality rate. The transition intensity $v(a,t,d)$ combines natural mortality and excess mortality due to disease.

Consider the following transition periods:
A_1 = time to death in the absence of disease (time of transition from state H directly to state D)
A_2 = time to infection (time of transition from state H to state I)
A_3 = time from infection to death (time of transition to state D, after entry to I)

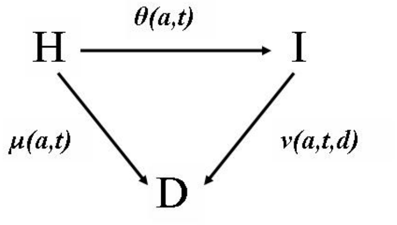


Figure 4.1: The three-state model for studying age-specific incidence and prevalence, where H=healthy, I=disease and D=death, with $\mu(a,t)$, $\theta(a,t)$ and $v(a,t,d)$, the rates or intensities corresponding to each transition time.

Then the rates or intensities associated with the first two transition periods are $\mu(a,t)$ and $\theta(a,t)$ respectively, which are dependent on age and calendar time, while the rate $v(a,t,d)$ associated with the third transition period is additionally dependent on disease duration d. The rate $\theta(a,t)$ is the incidence rate, or the rate of new infections.

Under this model, there are two potential times to death. If entry to the disease state, I, occurs, the time to death is $A_2 + A_3$. Otherwise the time to death is A_1. It is assumed that A_1 is independent of A_2 and A_3, so that we have a model of *independent competing risks*.

Consider the simple case of an exponential distribution: $A \sim \exp(\lambda)$, where

$$f(a) = \lambda e^{-\lambda a} \qquad a \geqslant 0$$

If $A =$ age to the occurrence of an event, then

$$\begin{aligned} P(A \leqslant a) &= F(a) = 1 - e^{-\lambda a} = P(\text{experiencing the event before age } a) \\ P(A > a) &= 1 - F(a) = S(a) = e^{-\lambda a} = P(\text{escaping or avoiding the event to beyond age } a) \end{aligned}$$

Now if λ is a function of a, then

$$\begin{aligned} 1 - e^{-\int_0^a \lambda(u)\,du} &= P(\text{experiencing the event before age } a) \\ e^{-\int_0^a \lambda(u)\,du} &= P(\text{escaping or avoiding the event to beyond age } a) \end{aligned}$$

Let us apply this result in the context of cross-sectional current status seroprevalence survey data, where an individual aged a is tested for disease at calendar time t. Note that the calendar time of birth for this individual is then $t-a$.

Hence the corresponding 'survival' times for A_1, A_2 and A_3 are:

$$\begin{aligned} P(A_2 > a) &= P(\text{escaping infection to beyond age } a) \\ &= exp\Big(- \int_0^a \theta(u, t-a+u)\, du \Big) \qquad (4.13) \end{aligned}$$

$$\begin{aligned} P(A_1 > a) &= P(\text{surviving to beyond age } a \text{ in the healthy state}) \\ &= exp\Big(- \int_0^a \mu(u, t-a+u)\, du \Big) \qquad (4.14) \end{aligned}$$

and

$$\begin{aligned} P(A_3 > d | A_2 = a-d) &= P(\text{surviving to beyond age } a \text{ in the disease state}|\text{infection at } a-d) \\ &= exp\Big(- \int_0^a v(u, t-a+u, u-a+d)\, du \Big)\ , \qquad (4.15) \end{aligned}$$

for an individual infected at age $a-d$.

In the current work we refer to the function $\theta(a,t)$ in equation (4.13) as the hazard or force of infection. This is because (4.13) is a type of survival function, where θ is its corresponding hazard rate.

Note that Keiding (1991) assumed that entry into state H, the alive-death transition and the onset of infection were all Poisson point processes, each with their own intensity function.

4.3.1 The prevalence odds

We shall now derive the prevalence odds, the ratio of the probability to be disease prevalent and the probability to be disease free.

The probability for an individual to be alive and in state I at age a and time

t is

$$k_1(a,t,d) = \int_0^a [P(\text{escaping infection until } a-d, t-d) \text{ x } P(\text{infection at } a-d, t-d)$$
$$\text{x } P(\text{survival in state I to } a, t)] \, dd$$

$$= \int_0^a \left\{ exp\left[-\int_0^{a-d} \{\mu(u, t-a+u) + \theta(u, t-a+u)\} du \right] \theta(a-d, t-d) \right.$$
$$\left. \text{x } exp\left[-\int_{a-d}^{a} v(u, t-a+u, u-a+d) \, du \right] \right\} dd \qquad (4.16)$$

This is therefore the probability of an individual aged a tested at time t, who was infected at some age $a-d$, and then survived in the disease state for the next d years until point of testing (a,t).

Note that $0 \leqslant d < a$, assuming that vertical transmission or infection at birth is not possible, otherwise $0 \leqslant d \leqslant a$. The integral is taken over all possible age-times $(a-d, t-d)$ at which infection may have occurred.

The probability for an individual to be alive and in the healthy state, H, at age a and time t is

$$\begin{aligned} k_2(a,t) &= P(\text{escaping infection to beyond } (a,t) \text{ and being tested at } (a,t)) \\ &= exp\left[-\int_0^a \{\mu(u, t-a+u) + \theta(u, t-a+u)\} du \right] \qquad (4.17) \end{aligned}$$

This probability corresponds to an individual aged a and tested at time t, having not experienced infection with the disease.

The prevalence odds are given by

$$\pi(a,t) = \frac{P_{at}}{1-P_{at}} \, ,$$

where P_{at} is the age specific prevalence at age a and time t, and can be expressed as

$$P_{at} = \frac{\text{P(an individual aged } a \text{ is alive and diseased at time } t)}{\text{P(an individual aged } a \text{ is alive at time } t)} \, .$$

Then $\pi(a,t)$ has the form

$$\pi(a,t) = \frac{k_1(a,t,d)}{k_2(a,t)} \qquad (4.18)$$

where $k_1(a,t,d)$ and $k_2(a,t)$ are defined in equations (4.16) and (4.17).

The prevalence odds, rather than the actual prevalence function, is often used in maximum likelihood procedures for estimating the parameters of interest.

4.4 Estimating the age and time incidence of HIV from antenatal seroprevalence data

We shall discuss the methods of Ades and Medley (1994) and Sakarovitch et al. (2007) for the estimation of age and time incidence of HIV based on seroprevalence data from women attending antenatal clinics. The methods used will be explained by referring to the data obtained from the Vulindlela antenatal clinic study.

The seroprevalence data contains observations from women aged 12 and above, attending antenatal clinics, in the rural area of Vulindlela in KwaZulu - Natal. The data were collected from 2001 to 2006.

The prevalence data consists of N_{at}, the number of individuals aged a ($a = 12, 13, \ldots, 47$) included in the survey at calendar time t ($t = 2001, 2002, \ldots, 2006$), and to Z_{at}, the number of HIV prevalent individuals at age a and calendar time t. The variable Z_{at} is assumed to be binomially distributed with size N_{at} and probability P_{at}. P_{at} denotes the probability that an individual aged a is HIV positive at time t. Hence P_{at} is the prevalence at age-time point (a,t).

Our aim is to estimate the age- and time-specific HIV incidence, denoted by $\theta(a,t)$. The seroprevalence odds, $\pi(a,t)$ are the odds of being HIV prevalent at (a,t). They are calculated as the probability of testing HIV-positive divided by the probability of testing HIV-negative, at a particular age and time. i.e. $\pi(a,t) = \frac{P_{at}}{1-P_{at}}$.

According to Sakarovitch et al. (2007), the seroprevalence odds are given by

$$\pi(a,t) = \frac{\int_0^a \{[\exp(-\int_0^{a-d} \theta(u,t-a+u)du)]\ \theta(a-d,t-d)\ \phi(a,t,d)\}dd}{\exp(-\int_0^a \theta(u,t-a+u)du)} \tag{4.19}$$

The numerator equates to P_{at}. The term $\phi(a,t,d)$ is the Relative Inclusion Rate (RIR), the probability of being included in the study for an individual aged a at calendar time t and infected d years before, relative to an uninfected individual aged a at time t. The details of this RIR parameter are described in Section 4.4.1. The term $[\exp(-\int_0^{a-d} \theta(u,t-a+u)du)]$ is the probability of escaping infection by the age-time point $(a-d,t-d)$. $\theta(a-d,t-d)$ is the HIV incidence rate, at age at infection $a-d$ and time at infection $t-d$. The denominator $\exp(-\int_0^a \theta(u,t-a+u)du)$ is an expansion of $(1-P_{at})$. It is the probability of escaping infection or not being infected by age-time point (a,t).

Note that the seroprevalence odds in (4.19) is similar in form to the expression for the prevalence odds used by Keiding (1991), given in equation (4.18). It differs from (4.18) in that it does not include the survival probabilities corresponding to the natural mortality rate $\mu(u,t-a+u)$ and the disease specific mortality rate $v(u,t-a+u,u-a+d)$. Instead the seroprevalence odds in (4.19) accounts for background mortality and disease specific mortality through $\phi(a,t,d)$, its Relative Inclusion Rate parameter. This parameter will be explained further in the text.

In the current work we refer to $\theta(a,t)$ as the hazard or force of infection.

The discrete version of equation (4.19) is

$$\pi_k = \sum_m \frac{(\exp(-\sum_{q<m} \theta_q X_{qk}))[1-\exp(-\theta_m X_{mk})]\phi(a_k,t_k,D_{qk})}{\exp(-\sum_q \theta_q X_{qk})} \tag{4.20}$$

where X_{qk} is the time spent in age-time interval q by persons in group k. We assume that the annual incidence rate θ_q is constant on each age-time interval $q=(a_i,t_i)$. If q refers to individual age-year periods, then only one year is spent in each interval q, and thus X_{qk} is equal to 1 for all q and k. D_{qk} is the time between survey date t_k and the time of HIV infection.

We can see that $(\exp(-\sum_{q<m} \theta_q X_{qk}))$ is the probability of escaping infection by

age-time interval q for all $q < m$ (Refer to the simpler example of the exponential distribution in Section 4.3). The term $[1 - \exp(-\theta_m X_{qk})]$ is the probability of being infected by age-time point m. $\sum_m$ implies that these infection probabilities are summed over all the possible age-time points at which infection may have occurred. This is because an individual aged a and tested at time t could have been infected at any age-time point (a_i, t_i) back up until age 12 or year 1984 i.e.($12 \leqslant a_i \leqslant a, 1984 \leqslant t_i \leqslant t$).

For example, the seroprevalence odds for an individual aged 14 and tested in year 2001 are

$$\pi_{14,01} = \frac{(1-e^{-\theta_{14,01}})e^{-(\theta_{13,00}+\theta_{12,99})}\phi(A \geqslant 14) + (1-e^{-\theta_{13,00}})e^{-\theta_{12,99}}\phi(A \geqslant 13) + (1-e^{-\theta_{12,99}})\phi(A \geqslant 12)}{e^{-(\theta_{14,01}+\theta_{13,00}+\theta_{12,99})}}.$$

The form of the numerator can be explained as follows:

The individual could have been infected at age 14 in 2001 (i.e. $\theta_{14,01}$) and not at age 13 in 2000 and at age 12 in 1999, and then survived AIDs related mortality to be tested and hence included in the study at age-time point (14, 2001);

or the individual could have been infected at age 13 in 2000 (i.e. $\theta_{13,00}$) and not at age 12 in 1999, and then survived AIDs related mortality to be tested at age 14 in 2001;

or the individual could have been infected at age 12 in 1999 (i.e. $\theta_{12,99}$) and then survived AIDs related mortality to be tested at age 14 in 2001.

4.4.1 The Relative Inclusion Rate

The Relative Inclusion Rate (RIR) takes into account the fact that the probability of being included in the sample depends on the HIV status. It is denoted by $\phi(a,t,d)$, which is the probability of being included in the sample at time t for an individual aged a and infected d years before, relative to the probability of being included in the sample for an uninfected individual aged a at time t.

According to Sakarovitch et al. (2007), the RIR is a product of three functions:

$$\phi(a,t,d) = f(t).g(a).h(d,a-d)$$

We shall discuss each of these three functions in detail.

f(t)

The function $f(t)$ is the change in the Relative Inclusion Rate over calendar time. The goal of this function is to take into account changes in the population, or changes in testing procedures or approaches to combating the disease over calendar time. This includes government policy on disease control.

We let f be a piecewise linear function, of the form

$$f(t) = (1+\beta_1(t-t_0)/100)I_{t\leqslant t_0} + (1+\beta_2(t-t_0)/100)I_{t>t_0}$$

This function has two slopes, β_1 and β_2, for before and after time t_0, and is equal to 1 at t_0. In the context of South Africa t_0 can be chosen to represent mid-2004. This is because in the year 2004 mass roll-out of antiretroviral drugs began in South Africa, particularly in the rural areas. This was a major change in the approach to combating the AIDs epidemic in the country. Therefore, in the Vulindlela study, β_1 is the slope from the beginning of the observation period to mid-2004 and β_2 is the slope from mid-2004 to 2006.

If no external information is available to help decide what values to set β_1 and β_2 to, then these two parameters must be estimated from the model. In the study by Sakarovitch et al. (2007) the two slopes were bounded between -10 and 10 per cent.

g(a)

The function $g(a)$ is the change in the RIR according to women's age. It takes into account the large number of HIV infected women among young pregnant women relative to young non-pregnant women. This is due to the fact that young pregnant women, say less than 20 years old, are likely to have entered sexual life earlier than non-pregnant women of the same age, and therefore to have been more exposed to HIV sexual transmission.

g is modelled as a piecewise linear function, which decreases toward one before age 20 and is equal to one thereafter, that is

$$\begin{aligned} g(a) &= 1 - \alpha(a-20) && \text{if } a \leqslant 20 \\ &= 1 && \text{if } a > 20 \end{aligned}$$

According to Carpenter (1997) the 'fertility risk ratio' for infected women compared to uninfected women is 1.35 for women between 15 and 19 years old. Therefore the parameter α may be fixed so that $g(17) = 1.35$ and $g(20) = 1$.

h(d, a-d)

The function h is the change in the RIR according to time since infection d and age at infection $a - d$. It accounts for the under-representation of HIV-infected women among pregnant women. The chances of an infected woman falling pregnant decline as the woman's and her partner's HIV infection progresses. This is due to a number of reasons. One major reason for this is that fertility declines as a woman's HIV infection progresses (Carpenter, 1997; Lewis et al, 2004). Infected women are also more likely to have infected partners, who will have a reduction in spermatozoid production (Krieger et al, 1991; Martin et al, 1992). Infected women may have sexual intercourse less often, due either to partner illness or that their partners have died from AIDs, and may therefore have lower chances of falling pregnant. Even a suspicion of HIV infection could make it difficult for an infected women to marry (or re-marry) and then have children (Ntozi, 1997).

We can therefore say that the probability of an infected woman becoming pregnant decreases as the number of years since infection (d) increases. The probability of becoming pregnant equals zero when the woman becomes infertile. This 'point of infertility' may occur with the onset of AIDs or a certain number of years before the onset of AIDs. Thus in order to model the time to infertility we need to first model the time to AIDs. This is done using a two parameter Weibull distribution with median μ and shape γ.

The parameters μ and γ were expressed as functions of age at infection $a - d$, so that $\mu = \mu_0 + (a-d)\mu_1$ and $\gamma = \gamma_0 + (a-d)\gamma_1$. The parameters μ_0, μ_1, γ_0 and γ_1 can be estimated from the model.

We will look at some properties of the Weibull distribution in order to further understand the function $h(d, a-d)$.

The Weibull distribution

The Weibull distribution is popular for modelling survival times or time-to-event data. In this case we will be modelling the time to AIDs i.e. the time from HIV infection to the onset of AIDs. This is also known as the AIDs incubation period.

The two parameter pdf of a Weibull distribution is given by

$$f(t) = \frac{\gamma}{\eta}\left(\frac{t}{\eta}\right)^{\gamma-1} exp\left[-\left(\frac{t}{\eta}\right)^{\gamma}\right] \tag{4.21}$$

where $\gamma =$ the shape (slope) parameter, $\eta =$ the scale parameter and $T =$ the time to AIDs.

The cumulative distribution function (cdf) of the Weibull distribution is

$$F(t) = P(T \leqslant t) = 1 - exp\left[-\left(\frac{t}{\eta}\right)^{\gamma}\right]. \tag{4.22}$$

The cdf gives the probability that the time to AIDs, T, is less than or equal to a specified time t.

The reliability function gives the probability that the time to AIDs T is greater than a specified time, t. That is:

$$R(t) = 1 - F(t) = exp\left[-\left(\frac{t}{\eta}\right)^{\gamma}\right]. \tag{4.23}$$

Sakarovitch et al. (2007) define the function $h(d, a-d)$ by

$$h(d, a-d) = exp\Bigg[-\left(\frac{d}{\mu_0 + \mu_1(a-d)}\right)^{\gamma_0+\gamma_1(a-d)}\Bigg]. \qquad (4.24)$$

This function gives the probability of developing AIDs either now or some time in the future, for a person infected d years before and at the age $a-d$. We can see that $h(d, a-d)$ is actually an adaptation of the Reliability function $R(t)$. Note that the random variable T being the time to AIDs is now D, the number of years since HIV infection, and d is a particular realization or value of D. Thus in this sense the function $h(d, a-d)$ does suffice in modelling the time to AIDs.

In order to determine whether the time to infertility decline occurs either at the onset of AIDs or 1, 2, 3 or more years before the onset of AIDs we express the median parameter $\mu = \mu_0 + \mu_1(a-d)$ as a linear function of the lag $a-d$.

4.5 Summary

This chapter discussed methods of estimating age specific incidence assuming equal and differential mortality for infected and uninfected individuals. The relationship between the incidence rate and the hazard of infection or the force of infection were discussed using approaches by Keiding (1991) and Sakarovitch et al (2007). Part of the aim of the current work is to estimate the rate of new infections using seroprevalence data. We use the force of infection or the hazard of infection as a measure of new infections. This task is accomplished in Chapter 5.

Chapter 5

Estimation of the force of infection

The force of infection, or hazard of infection, describing the rate at which susceptible individuals acquire infection, is one of the primary epidemiological parameters in infectious disease modelling. For many infectious diseases the force of infection is assumed to be age-dependent. Although the methods for estimating the force of infection can be applied to case notification data (reported cases of the disease in a specific time period), it is more common to have cross-sectional seroprevalence data from which the prevalence and the force of infection are estimated. The force of infection has the same interpretation as the incidence rate i_k in equation (4.7).

Let $q(a,t)$ be the proportion of susceptible individuals at age a and time t. Then the partial differential equation describing the change in the susceptible proportion at age a and time t is given by

$$\frac{\partial}{\partial a}q(a,t) + \frac{\partial}{\partial t}q(a,t) = -\ell(a,t)\; q(a,t) \tag{5.1}$$

where $\ell(a,t)$ is the force of infection. Note that in (5.1) it is assumed that the disease is irreversible, which we know to be the case for HIV infection, and that the mortality caused by the infection is negligible (Shkedy et al, 2006). In the case of HIV, we can assume negligible mortality at early stages of the disease which has a long incubation period.

In the steady state, under the time homogeneous assumption, we have $(\partial/\partial t)\;(q(a,t)) = 0$ and (5.1) reduces to

$$\frac{d}{da}q(a) = -\ell(a)\; q(a). \tag{5.2}$$

The equation above describes the change in the susceptible proportion with age, and $\ell(a)$ denotes the age-dependent force of infection. The prevalence is given by $\pi(a) = 1 - q(a)$.

Muench (1959) first proposed the concept of estimating the force of infection, using a catalytic model on summation data. In this model, the time spent in the susceptible class is exponentially distributed with rate β. The susceptible proportion was given by $q(a) = e^{-\int_0^a \beta \, ds} = e^{-\beta a}$ and $(d/da)q(a)$ was given by $-\beta e^{-\beta a}$. Hence Muench's model assumed a constant force of infection, $\ell(a) = \beta$, which is independent of age. This clearly was quite a strong assumption which many researchers later sought to relax. Griffiths (1974) suggested a model for measles in which the force of infection increases linearly with age over the age range $0 - 10$ years. The model was extended further by Grenfell and Anderson (1985) who used polynomial functions to model changes in the force of infection with age. The model assumed that $q(a) = e^{-\sum \beta_i a^i}$ resulting in a force of infection of the form $\ell(a) = \sum \beta_i i a^{i-1}$. The general solution for (5.2) is $q(a) = e^{-\gamma(a)}$, where $\gamma(a) = \int_0^a \ell(s) ds$ is the cumulative force of infection or hazard of infection.

When higher order polynomials are fitted estimates of the force of infection can be negative for some age values. Models resulting in the force of infection turning negative are those in which the estimated prevalence is a non-monotone function. Farrington (1990) addressed this problem by constraining the force of infection to be non-negative i.e. $\ell(a) \geqslant 0$. He developed a non-linear model for $\pi(a)$ which he applied to data on measles, mumps and rubella.

The models proposed by Muench (1959), Griffiths (1974), and Grenfell and Anderson (1985) may be fit as generalized linear models (GLMs) with binomial error and log link. Since then other models fitted within the framework of GLMs with binomial error have been considered, using different link functions. Becker (1989) and Diamond and McDonald (1992) parameterized the prevalence and force of infection as a Weibull model, using the complementary log-log link. Grummer-Strawn (1993) proposed a log-logistic model with logit link function as well as a Weibull proportional hazards model with complementary log-log link.

Keiding (1991), who was the first to explicitly use a non-parametric technique to estimate the force of infection, used isotonic regression models to estimate the prevalence and kernel smoothers to estimate the force of infection. More recently Shkedy et al (2003) proposed using local polynomials to model the prevalence and the force of infection.

Semi-parametric models were proposed by Shiboski (1998), Hastie and Tibshirani (1990) as well as Rossini and Tsiatis (1996), in which the age-specific prevalence is modelled non-parametrically and possible covariate effects are included in the parametric component of the model.

In order to better understand the models used to estimate the force of infection, we need to gain an understanding of survival analysis methods. We give a brief background to survival analysis, then discuss how these methods can be applied to current status data to estimate the prevalence and the force of infection. We then present commonly used models of the force of infection, and apply them to the Vulindlela antenatal clinic data used in the current thesis.

5.1 Background to survival analysis

Survival analysis is the term used to describe the analysis of *time to event data*, where the variable of interest is the time from a well defined time origin until the occurrence of a particular event or endpoint. In the context of infectious disease data the survival time is the time that elapses before an individual is infected with the disease. The time origin may be birth or the date of entry into the study. A measurement scale for the passage of survival times should be well defined, and in most studies calendar time serves as a common and meaningful measure. When the time origin is birth the survival time is equal to the age of the individual, as is the case in the current application.

Standard statistical procedures are not appropriate in the analysis of survival data, as data of this type are generally not symmetrically distributed and tend to be positively skewed. It is therefore unreasonable to assume that the data have a normal distribution (Collett, 2003). Examples of distributions that are better suited to survival data are the Weibull, exponential and gamma distributions. The use of the exponential distribution, however, may be unrealistic in many applications, as it results in a constant hazard rate.

A main feature of survival data is *censoring*. A censored survival time occurs when the event or endpoint of interest is unknown or has not been observed for an individual. This could be because the individual has not yet experienced the event at the point of observation or because he or she has been lost to follow-up. *Right censoring* occurs when the actual survival time is after (or to the right of) the last known or observed survival time while *left censoring* occurs when the actual survival time is less than (or to the left of) the last known survival time.

In current status data all the observations are censored (Namata et al, 2007). The Vulindlela antenatal clinic data is cross sectional data where the current HIV status of each individual is determined at a particular age-time point, either 1 for infected or 0 for uninfected. At the time of HIV testing each individual's time to infection is unknown. The only available information is the age and time at which they are tested, which becomes their censored survival time.

Suppose an HIV test for individual i is done at age a_i. If the test is positive then the actual age at infection is $a_i^* \leqslant a_i$, hence the age at infection is left-censored. On the other hand if the test is negative then we know that the age at infection is some future age $a_i^* > a_i$ and in this case the age at infection is right censored.

Another type of censoring is *interval censoring*, where the actual age at infection a_i^* is known to lie between two age points a_i^1 and a_i^2, that is $a_i^1 < a_i^* < a_i^2$. The current data however does not present the problem of interval censored data because the individuals in the study were not tested at two time points.

5.2 The survival function and the hazard function

Suppose $t_1, t_2, ..., t_n$ are observed survival times of the continuous random variable T, which has probability density function $f(t)$. The cumulative distribution function of T, $F(t)$, is given by

$$\begin{aligned} F(t) &= P(T \leqslant t) \\ &= \int_0^t f(u)\,du, \end{aligned}$$

which represents the probability of survival time being less than some value t.

The *survival function* gives the probability of an individual surviving to beyond time t, i.e. the individual experiences the event or endpoint after time t, and is given by

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - F(t). \end{aligned} \tag{5.3}$$

The *hazard function* is the instantaneous rate of occurrence of the event of interest at time t, given survival to time t. It is defined by

$$h(t) = lim_{\delta t \to 0} \frac{P(t \leqslant T \leqslant t + \delta t \mid T \geqslant t)}{\delta t}$$

By the laws of conditional probability it can be shown that

$$\begin{aligned} h(t) &= lim\ _{\delta t \to 0} \frac{P(t \leqslant T \leqslant t + \delta t)}{\delta t\ \ P(T \geqslant t)} \\ &= lim\ _{\delta t \to 0} \left(\frac{F(t + \delta t) - F(t)}{\delta t} \right) \frac{1}{P(T \geqslant t)} \\ &= \frac{F'(t)}{1 - F(t)} && (5.4) \\ &= \frac{f(t)}{S(t)}. && (5.5) \end{aligned}$$

The derivations imply that

$$S(t) = e^{-\int_0^t h(u)\ du} = e^{-H(t)} \tag{5.6}$$

where $H(t)$ is called the cumulative or integrated hazard function.

5.3 Current status data - The force of infection and the prevalence

In the context of current status infectious disease data, the event/endpoint is infection with the disease and survival time is the time to infection.

If the time origin is at birth, then the observed survival time is the age of an individual at the time of testing for infection. Then the survival function $S(a)$ gives the probability of an individual escaping infection to beyond age a, i.e. the individual is infected with the disease after age a. Note that from equation (5.6), $S(a)$ is a function of $h(a)$, the hazard of infection. At a particular age a the survival function gives the probability to be uninfected, or *susceptible* to the disease.

Hence

$$S(a) = q(a),$$

where $q(a)$ is the proportion of susceptible individuals at age a.

The proportion of infected individuals, or the *prevalence*, at age a is given by

$$\pi(a) = 1 - q(a) = 1 - S(a).$$

From (5.3) we see that the prevalence is equivalent to the cumulative distribution function $F(a)$ of the age to infection variable.

Using infectious disease data we equate the hazard function to the *force of infection*, the rate at which susceptible individuals become infected. It can be seen from (5.2) that the force of infection is represented as

$$\ell(a) = \frac{\pi'(a)}{1 - \pi(a)} \tag{5.7}$$

where $\pi(a)$ is the age-specific prevalence.

5.4 Models for the force of infection using different link functions

Consider a cross-sectional prevalence sample of size N and let a_i be the age of the $i'th$ subject. We observe the binary variable Y_i such that

$$Y_i = \begin{cases} 1 \text{ if subject } i \text{ experienced infection before or at age } a_i \\ 0 \text{ otherwise} \end{cases}$$

Let $\pi(a_i) = P(Y_i = 1)$ and suppose $\pi(a_i)$ is a function of a vector valued parameter β.

The log likelihood for β is then given by

$$\ell(\beta) = \sum_{i=1}^{N} Y_i \; log\{\pi(a_i)\} + (1 - Y_i) \; log\{1 - \pi(a_i)\} \tag{5.8}$$

with $\pi(a_i) = 1 - q(a_i)$ being the probability to be infected before or at age a_i.

A generalized linear model for a binomial response takes the form

$$g(\pi(a)) = \eta(a)$$

where $\eta(a)$ is the linear predictor and g is the link function. Then the value of $\pi(a)$ is found by $\pi(a) = g^{-1}(\eta(a))$.

The models in which the force of infection was constant (discussed by Muench (1959)), linear (Griffiths, 1974) and flexible (Grenfell and Anderson, 1985) all assume that g is the log link function for $(1 - \pi)$ and that $\eta(a) = \sum_{i=0}^{k} \beta_i a^i$, where k is equal to 1 (constant), 2 (linear) and K (flexible). In these models $\pi(a) = 1 - e^{-\eta(a)}$, and using the definition for the force of infection in (5.7), we have

$$\ell(a) = \frac{\pi'(a)}{1 - \pi(a)} = \frac{\eta'(a) \; e^{-\eta(a)}}{e^{-\eta(a)}} = \eta'(a). \tag{5.9}$$

Therefore when using a model with the log link the force of infection is simply the first derivative of the linear predictor. This means that the linear predictor $\eta(a)$ is the cumulative hazard function.

When a link function other than the log link is used, the force of infection may still be derived from (5.7). It can be shown that for the binomial distribution, the force of infection is expressed as

$$\ell(a) = \eta'(a) \; \delta(\eta(a)) \tag{5.10}$$

The form of $\delta(.)$ is dependent on the link function g. We see that when g is the log link the value of $\delta(\eta(a))$ is equal to 1. Table 5.1 shows the structural forms of the force of infection, $\pi(a)$ and $\delta(\eta(a))$ associated with the three commonly used link functions in models with binomial error.

Table 5.1: Structural forms of the force of infection for commonly used link functions

Link function	$\pi(a)$	$\ell(a)$	$\delta(\eta(a))$
log	$1 - e^{-\eta(a)}$	$\eta'(a)$	1
complementary log-log	$1 - e^{-e^{\eta(a)}}$	$\eta'(a)\ e^{\eta(a)}$	$e^{\eta(a)}$
logit	$\frac{e^{\eta(a)}}{1 + e^{\eta(a)}}$	$\eta'(a)\ \frac{e^{\eta(a)}}{1 + e^{\eta(a)}}$	$\frac{e^{\eta(a)}}{1 + e^{\eta(a)}}$

Using the Vulindlela antenatal clinic data, we will illustrate three well known examples of generalized linear models for the force of infection, which assume that the time to infection follows either an exponential, Weibull or log-logistic distribution. The variables used are participant's age a_i and post-test HIV status (infected or uninfected), Y_i. Model fitting was conducted using SAS Proc Genmod.

5.4.1 Exponential distribution

Survival analysis

If the age to infection a is exponentially distributed, we have probability density function

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

The survival function or probability to be susceptible is given by

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

and the hazard function is given by

$$\frac{f(a)}{S(a)} = \frac{\lambda e^{-\lambda a}}{e^{-\lambda a}} = \lambda.$$

Current status data

The proportion of susceptible individuals $q(a)$ is the survival function, so that $q(a) = S(a) = e^{-\lambda a}$.

The prevalence is then given by

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

The force of infection is the hazard rate, $l(a) = \lambda$. We can show this directly using equation (5.7) by

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

Hence the force of infection is a constant value which does not depend on age.

Model fitting and parameter estimation

We fit a generalized linear model with the complementary log-log link function, of the form

$$g(\pi(a)) = log(-log(1-\pi(a))),$$

using the $log(age)$ as an offset variable.

Substituting for $\pi(a)$ we have

$$g(\pi(a)) = log(\lambda) + log(a). \tag{5.11}$$

Table 5.2 below shows the deviance and Pearson Chi-square statistics for the fitted model, as well as the minimised value of the log likelihood. The parameter estimates for the fitted model are displayed in Table 5.3.

Table 5.2: Fit statistics from the exponential model for the force of infection

Criterion	DF	Value	Value/DF
Deviance	32	219.7307	6.8666
Scaled Deviance	32	219.7307	6.8666
Pearson Chi-Square	32	234.3620	7.3238
Scaled Pearson X2	32	234.3620	7.3238
Log Likelihood		-1350.6600	

Table 5.3: Parameter estimates from the exponential model for the force of infection

Parameter	DF	Estimate	Standard Error	Wald 95% Lower	Wald 95% Upper	Chi-Square	Pr > ChiSq
Intercept	1	-3.2165	0.0367	-3.2885	-3.1445	7672.57	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

The intercept $\mu = -3.2165$ in Table 5.3 gives the value of $log(\lambda)$ in (5.11) above, so that the true parameter λ can be found by exponentiating the intercept. That is,

$$g(\pi(a)) = \mu + log(a) \;\; \Rightarrow \; \lambda = e^{\mu}.$$

The force of infection rate is therefore $\lambda = e^{-3.2165} = 0.0401$ or 4.01%.

As per the invariance property the lower and upper 95% confidence limits for λ are given by

$$\text{LCL} = e^{\hat{\mu}-1.96 \text{ x } s.e.(\hat{\mu})} = e^{-3.2165-1.96 \text{ x } 0.0367} = 0.0373$$

$$\text{UCL} = e^{\hat{\mu}+1.96 \text{ x } s.e.(\hat{\mu})} = e^{-3.2165+1.96 \text{ x } 0.0367} = 0.0431$$

Figure 5.1 and Figure 5.2 respectively show the plotted prevalence and force of infection rates against age. Prevalence of HIV infection rises with age toward a maximum of 0.73 or 73% in the mid-forties.

Recall that the prevalence $\pi(a)$ corresponds to $F(a)$, the cumulative distribution function of a survival time, in this case the age to infection. This means that $\pi(a)$ is a cumulative prevalence. It represents the probability to be infected before age a. For example, $\pi(20)$ gives the probability to be infected before age 20, i.e. the probability of an individual having been infected at age 19 or 18 or 17 or..., in other words at any age less than 20. The cumulative prevalence at the highest age is an estimate of the prevalence rate in the total sample. This is due to the fact that the probability to be infected before this age is the overall probability to be infected in the sample of individuals ranging in age from 12 to 47.

The force of infection in Figure 5.2 is constant at 0.0401 for all ages and implies a force of infection that is independent of age. While this may be appropriate in the case of some diseases such as Rubella (Hens et al, 2010), a constant force of infection for HIV may not be realistic.

The interpretation of the force or hazard of infection $\lambda = 0.0401$ is that given a participant had not tested positive before age a the probability that the individual will test positive at age a is 0.0401. The disadvantage with the exponential model is that it assumes that the rate of new infections is constant over all ages.

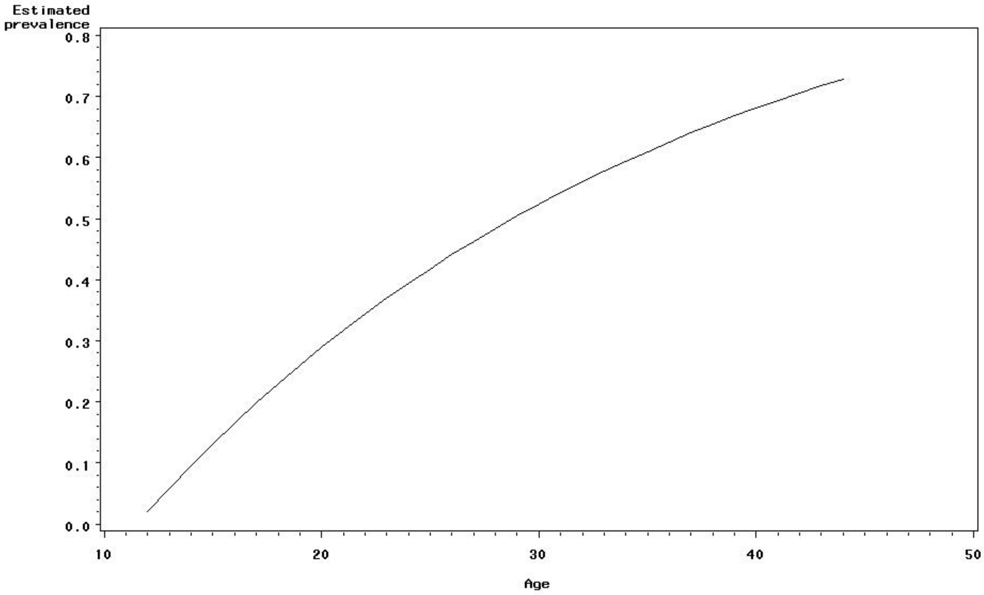


Figure 5.1: The fitted prevalence function for the exponential model

In this case a rate of approximately 4 per hundred at every age, which is quite unrealistic as shown in Figure 5.2.

5.4.2 Weibull distribution

Survival analysis

If we assume that the age A spent in the susceptible class before infection (HIV positive test) follows a Weibull distribution, then A has probability density function

$$f(a) = \alpha\lambda a^{\alpha-1}\, e^{-\lambda a^{\alpha}}.$$

The probability of infection beyond age a, given by the survival function, is

$$S(a) = e^{-\lambda a^{\alpha}},$$

and hence the hazard (or force of infection) takes the form

$$\frac{f(a)}{S(a)} = \frac{\alpha\lambda a^{\alpha-1}\, e^{-\lambda a^{\alpha}}}{e^{-\lambda a^{\alpha}}} = \alpha\lambda a^{\alpha-1}.$$

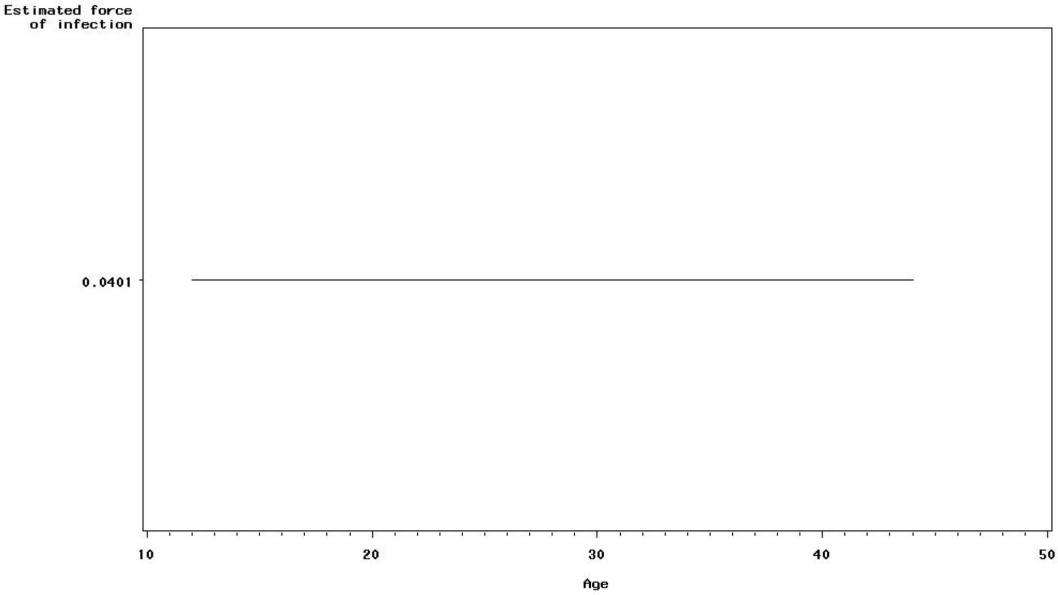


Figure 5.2: The fitted force of infection function for the exponential model

Current status data

The proportion of susceptible individuals is given by the survival function,

$$q(a) = e^{-\lambda a^{\alpha}}.$$

The prevalence is therefore

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

and the force of infection is the hazard rate

$$\ell(a) = \alpha\lambda a^{\alpha-1},$$

which may be verified using the formula in (5.7). The Weibull force of infection is therefore age-dependent and monotone (either increasing or decreasing), depending on the sign of α. Note that when $\alpha = 1$ the Weibull model is an exponential model.

Model fitting and parameter estimation

A generalized linear model with the complementary log-log link function, of the form

$$g(\pi(a)) = log(-log(1 - \pi(a)))$$

is fit to the data. $Log(age)$ instead of age is used as the independent variable. By substituting $\pi(a) = 1 - e^{-\lambda a^{\alpha}}$ into the equation above, the model then equates to

$$g(\pi(a)) = log(\lambda) + \alpha\ log(a). \tag{5.12}$$

The Weibull model was fit to the antenatal clinic data from Vulindlela. The resulting prevalence function rises monotonically with age, with the slope or rate of increase becoming flatter towards higher ages. The Weibull force of infection decreases monotonically with age with its slope levelling off towards higher age values. The fitted prevalence and force of infection functions for the Weibull model were very similar in form to that of the log- logistic model, which is discussed next. The goodness of fit however proved to be slightly better for the log-logistic model, and therefore we shall present the results of the log logistic model.

5.4.3 Log-logistic distribution

Survival analysis

Let us assume that the age to infection A has a log-logistic distribution, with probability density function

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

Then, the survival function is

$$S(a) = \frac{1}{1+\lambda a^{\alpha}}$$

and the hazard function takes the form

$$\frac{f(a)}{S(a)} = \frac{\lambda \alpha a^{\alpha-1}}{1+\lambda a^{\alpha}}.$$

Current status data

The proportion of susceptible individuals is equivalent to the survival function, so that we have

$$q(a) = \frac{1}{1+\lambda a^{\alpha}}$$

The prevalence is given by

$$\begin{aligned} \pi(a) &= 1 - q(a) \\ &= 1 - \frac{1}{1+\lambda a^{\alpha}} \\ &= \frac{\lambda a^{\alpha}}{1+\lambda a^{\alpha}}. \end{aligned}$$

The force of infection is the hazard rate,

$$\ell(a) = \frac{\lambda \alpha a^{\alpha-1}}{1+\lambda a^{\alpha}}.$$

To verify this using the formula in (5.7), we have

$$\begin{aligned} \ell(a) &= \frac{\pi'(a)}{1-\pi(a)} \\ &= \frac{\frac{d}{da}\left(\frac{\lambda a^{\alpha}}{1+\lambda a^{\alpha}}\right)}{\left(1 - \frac{\lambda a^{\alpha}}{1+\lambda a^{\alpha}}\right)} \end{aligned}$$

Using the quotient rule for differentiation,

$$\begin{aligned} \ell(a) &= \frac{[\lambda \alpha a^{\alpha-1}(1+\lambda a^{\alpha}) - \lambda \alpha a^{\alpha-1}(\lambda a^{\alpha})]/(1+\lambda a^{\alpha})^2}{1/(1+\lambda a^{\alpha})} \\ &= \frac{\lambda \alpha a^{\alpha-1}(1+\lambda a^{\alpha} - \lambda a^{\alpha})}{1+\lambda a^{\alpha}} \\ &= \frac{\lambda \alpha a^{\alpha-1}}{1+\lambda a^{\alpha}}. \end{aligned}$$

Model fitting and parameter estimation

A generalized linear model is fit to the data using a logit link function and with $log(age)$ as the independent variable. The model takes the form

$$g(\pi(a)) = logit(\pi(a)) = log\left(\frac{\pi(a)}{1-\pi(a)}\right).$$

The right hand side of $g(\pi(a))$ equates to

$$\begin{aligned} g(\pi(a)) &= log(\lambda a^{\alpha}) \\ &= log(\lambda) + \alpha\ log(a). \end{aligned} \tag{5.13}$$

The goodness of fit statistics in Table 5.4 show that the deviance, the Pearson Chi-square and the log likelihood are smaller than those of the exponential model, and marginally smaller than those of the Weibull model.

It can be seen from Table 5.5 that the estimate of the intercept is -1.8878. Due to functional form of the intercept in (5.13), the estimate of λ is derived by exponentiating the intercept, so that $\lambda = e^{\mu} = e^{-1.8878} = 0.1514$. The coefficient of $log(age)$ is α, and is estimated to be 0.5859.

Figure 5.3 and 5.4 shows the fitted prevalence and force of infection functions against age. The prevalence increases monotonically with age, from 0.09 to 0.54 over the age range, and with a slope that becomes flatter with increasing age. As discussed before, the prevalence at the highest age is an estimate of the overall prevalence in the sample.

The force of infection declines with age, from 0.108 at age 12 to 0.01 in the mid to late forties. The slope becomes flatter with increasing age.

Table 5.4: Fit statistics from the log-logistic model for the force of infection

Criterion	DF	Value	Value/DF
Deviance	31	142.9536	4.6114
Scaled Deviance	31	142.9536	4.6114
Pearson Chi-Square	31	137.3814	4.4317
Scaled Pearson X2	31	137.3814	4.4317
Log Likelihood		-1312.2715	

Table 5.5: Parameter estimates from the log-logistic model for the force of infection

Parameter	DF	Estimate	Standard Error	Wald 95% Lower	Wald 95% Upper	Chi-Square	Pr > ChiSq
Intercept	1	-1.8878	0.2129	-2.3050	-1.4705	78.62	<.0001
lage	1	0.5859	0.0869	0.4156	0.7562	45.46	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

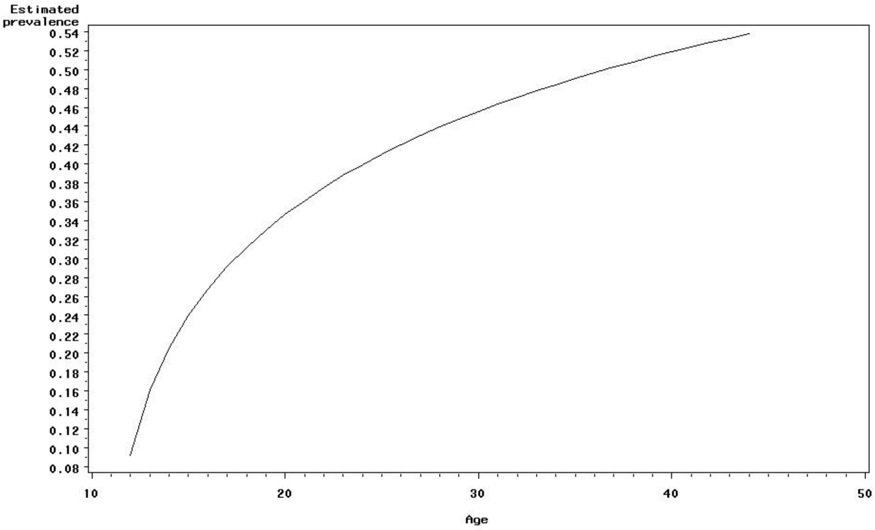


Figure 5.3: The fitted prevalence function for the log-logistic model

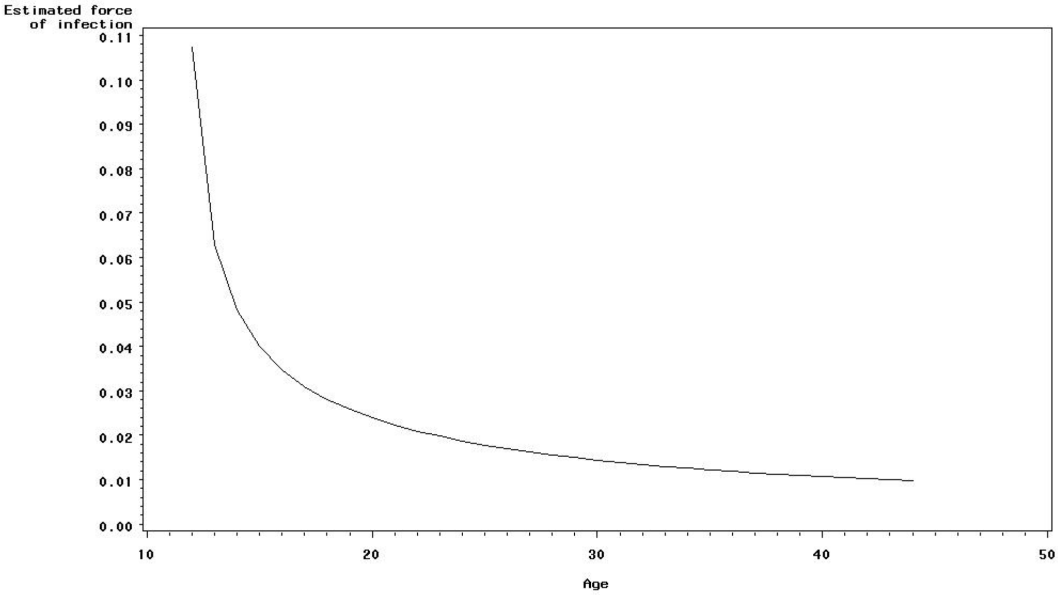


Figure 5.4: The fitted force of infection function for the log-logistic model

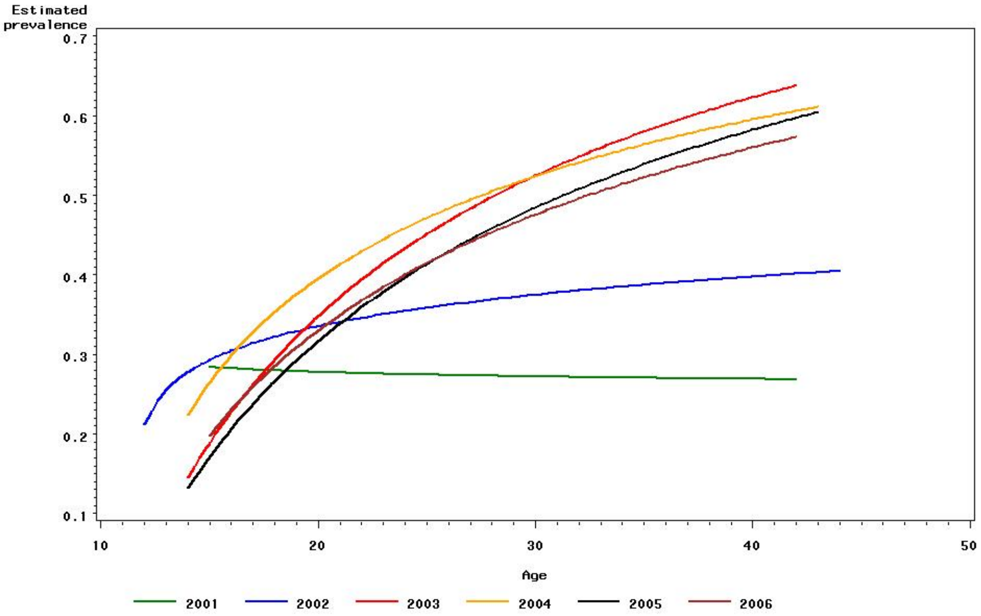


Figure 5.5: The fitted prevalence functions for each year from the log-logistic model

Fitting the model separately by year (Figure 5.5) shows that the prevalence is highest in years 2003 and 2004 for all ages 20 and older. The prevalence increases with age in all the years from 2002 - 2006, while in 2001 the prevalence shows a slight decline with age. The slopes for the latter years, 2003-2006, are notably steeper than in 2002, indicative of a higher rate of increase in the prevalence with age in these latter years.

Figure 5.6 shows the corresponding force of infection functions for each year using the log logistic model. In 2001 the force of infection remains almost constant, showing a very slight increase over age. The force of infection is lowest in 2001 across all age values. The force of infection functions for the other years decrease with age. In 2002 the force of infection has a steep decline from 0.094 at age 12 to 0.02 at age 15. For ages 15 and older the force of infection for years 2003, 2004, 2005 and 2006 all have similar slopes, and are notably higher than the rates for 2002. The disadvantage with the log-logistic model is that it assumes a declining rate of new infections with age, which is not realistic for a disease such as HIV. Thus more realistic models are preferred.

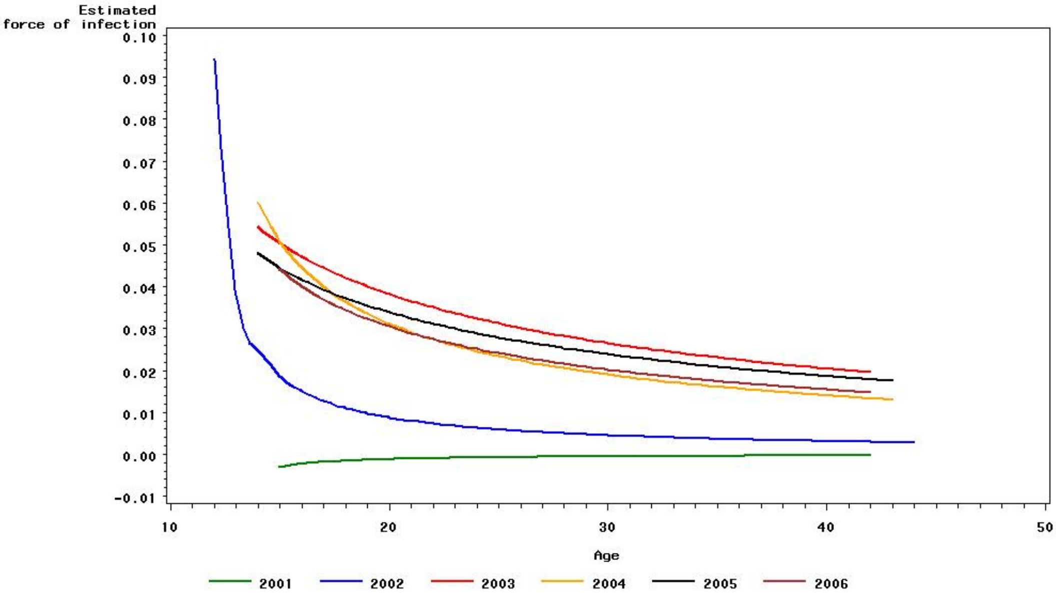


Figure 5.6: The fitted force of infection functions for each year from the log-logistic model

5.5 Non-linear models

The following models for the force of infection do not assume that the age to infection a follows a specific distribution. However an assumption is made about the form of the force of infection $\ell(a)$. They are non-linear models. Hence we use SAS Proc NLMIXED to fit these models to the data.

5.5.1 Linear force of infection

A model in which the force of infection increases linearly with age was first introduced by Griffiths (1974). This model specifies the prevalence so that the linear predictor is a quadratic function of age.

The linear force of infection has the form

$$\ell(a) = \beta_1 + 2\beta_2 a$$

and the prevalence is given by

$$\pi(a) = 1 - exp(\beta_0 \ + \ \beta_1 a \ + \ \beta_2 a^2). \tag{5.14}$$

Note that this model can be seen in the context of survival analysis where $\pi(a)$ is the cumulative distribution function of the time to infection or alternatively, $1-$ survival time, and where $\ell(a)$ is the corresponding hazard rate. The function $\lambda(a) = \int_0^a \ell(a)\ da$ gives the cumulative hazard function.

Model fitting and parameter estimation

A model for the linear force of infection is fit using SAS Proc NLMIXED. The functional form of the prevalence in (5.14) must be specified in NLMIXED and initial values are required for β_0, β_1 and β_2.

Note that within the framework of generalized linear models for binary data, the model for the linear force of infection can be fitted using a log link. The model takes the form

$$\begin{aligned} g(\pi(a)) &= log(1 - \pi(a)) \\ &= \beta_0 + \beta_1 a + \beta_2 a^2, \end{aligned}$$

where the number of susceptible individuals is the response. This model leads to a linear force of infection, increasing or declining with age, and therefore unrealistic in relation to the HIV infection process.

5.5.2 Farrington's model

Farrington (1990) proposed a non-linear model for the force of infection, defined by

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

To ensure that the force of infection satisfies $\ell(a_i)\ \geqslant\ 0,\ i = 1, 2, ..., n$ the parameter space was constrained to be non-negative ($\alpha_j\ \geqslant\ 0,\ j = 1, 2, 3$).

The model assumes that the force of infection is zero at birth, then increases linearly to a peak before decreasing exponentially. The age at which the force of infection reaches a peak corresponds to the maximum contact rate of susceptibles with infectious individuals.

The parameter α_3 is referred to as the long term residual value of the force of infection. When $\alpha_3 = 0$ the force of infection decreases to 0 as age tends to infinity. This results in the $2-$ instead of $3-$parameter model.

Farrington (1990) defined the prevalence by

$$\pi(a) = 1 - exp\bigg(\frac{\alpha_1}{\alpha_2}ae^{-\alpha_2 a} + \frac{1}{\alpha_2}\bigg(\frac{\alpha_1}{\alpha_2} - \alpha_3\bigg)(e^{-\alpha_2 a} - 1) - \alpha_3 a\bigg). \qquad (5.15)$$

This type of function is more realistic for a disease such as HIV, and hence it is discussed and applied below.

Model fitting and parameter estimation

I. Farrington's 3-parameter model

The model is fit using SAS Proc NLMIXED, where initial estimates of α_1, α_2 and α_3 are required. The expression for prevalence in (5.15) must be specified and the binomial distribution must be specified in the model statement.

There are several optimization techniques in Proc NLMIXED, and since the default technique, the Dual Quasi-Newton method, could not achieve convergence after a large number of iterations, the other non-linear optimization techniques were tried. The Newton-Raphson Ridge Optimization (NRRIDG) method was used, as it achieved convergence of the algorithm while yielding the lowest values of the fit statistics, when compared to the other optimization techniques.

The fitted model resulted in a monotonically decreasing force of infection. However, the Farrington's 2-parameter model, discussed next, produced a better fit to the data.

II. The 2-parameter model ($\alpha_3 = 0$)

When $\alpha_3 = 0$, Farrington's model for the force of infection reduces to a 2-parameter model, where

$$\ell(a) = \alpha_1 a e^{-\alpha_2 a}$$

and

$$\pi(a) = 1 - exp\left(\frac{\alpha_1}{\alpha_2} a e^{-\alpha_2 a} + \frac{1}{\alpha_2}\left(\frac{\alpha_1}{\alpha_2}\right)(e^{-\alpha_2 a} - 1)\right)$$

SAS Proc NLMIXED is again used to fit the model above and the resulting fit statistics as well as parameter estimates are displayed in Tables 5.6 and 5.7 below.

Table 5.6: Fit statistics from Farrington's 2-parameter model for the force of infection

Criterion	Value
-2 Log Likelihood	223.2
AIC	227.2
AICC	227.6
BIC	230.2

Table 5.7: Parameter estimates from Farrington's 2-parameter model for the force of infection

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
alpha1	0.06432	0.008626	33	7.46	<.0001	0.05	0.04677	0.08187	-0.00004
alpha2	0.3180	0.02675	33	11.89	<.0001	0.05	0.2636	0.3725	0.00002

Fitted prevalence (Figure 5.7) rises monotonically with age, with a slope that becomes flatter toward higher ages. The prevalence increases from almost zero at age 12 to 0.47 in the mid forties. The force of infection (Figure 5.8) on the other hand increases from age 12 to a peak of 0.074 at age 15 then decreases toward zero beyond age 35.

Analysis by year (Figure 5.9) shows that the estimated prevalence functions all rise steeply toward the early twenties and then level off, with the slopes for years 2003-2006 being much steeper than those for 2001 and 2002. For ages 22 and above, the prevalence is highest in 2003 and 2004.

The force of infection functions for each year are plotted in Figure 5.10. The functions rise to a peak at around age 16 and decline thereafter, tending towards zero in the mid to late forties. The slope for 2001 does not rise toward a peak, as it simply decreases monotonically from 0.055 at age 15 toward zero in the higher age groups. The slope for 2002 increases toward age 15 at a much

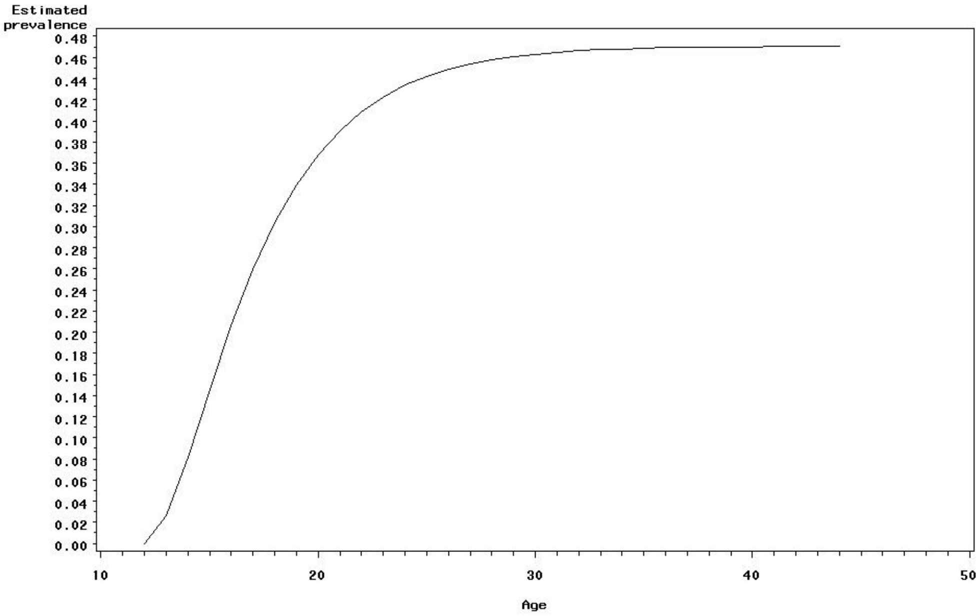


Figure 5.7: The fitted prevalence function for Farrington's 2-parameter model i.e. when $\alpha_3 = 0$.

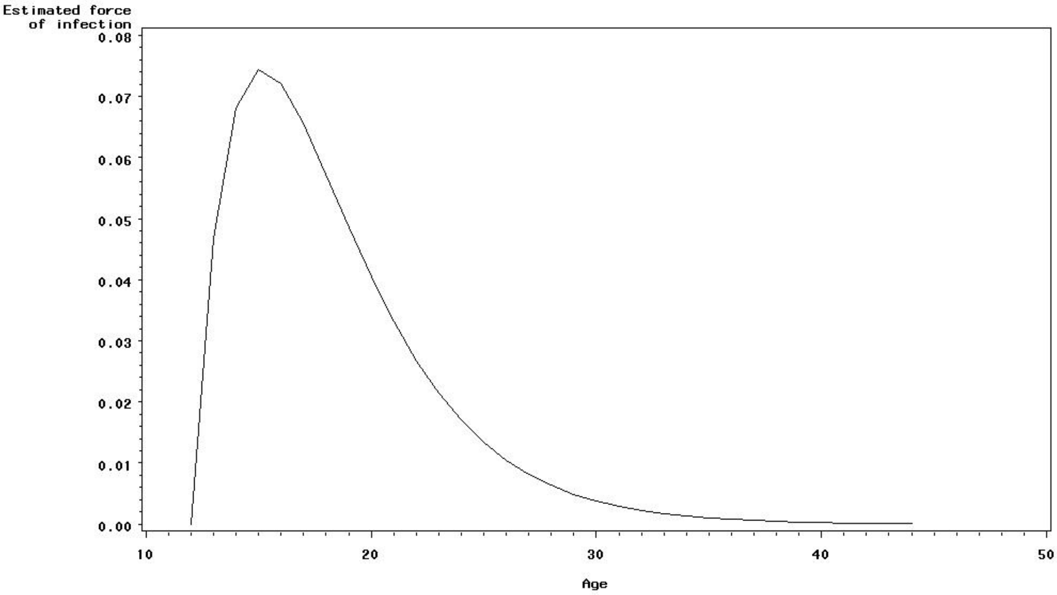


Figure 5.8: The fitted force of infection function for Farrington's 2-parameter model i.e. when $\alpha_3 = 0$.

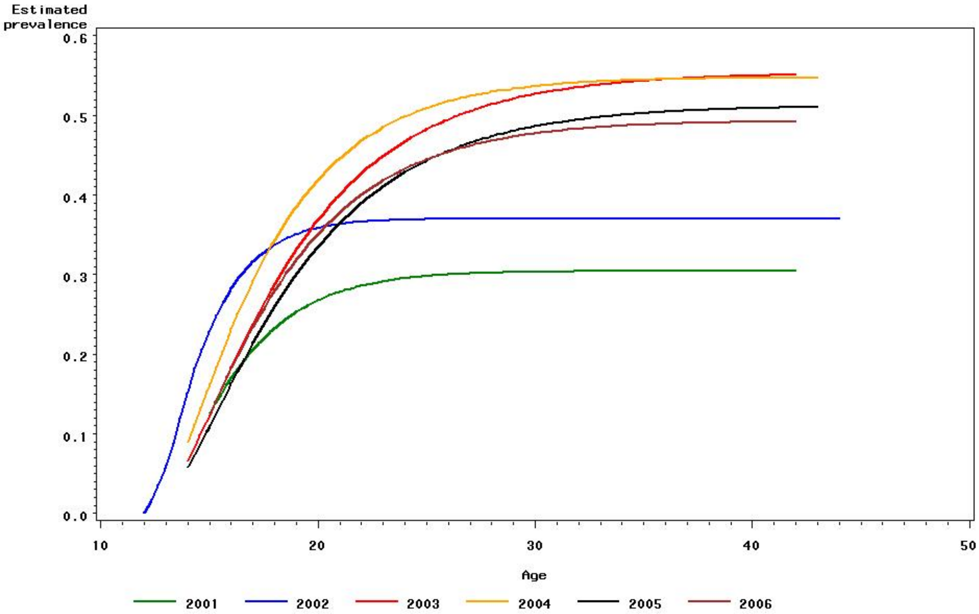


Figure 5.9: The fitted prevalence functions for each year using the Farrington's 2-parameter model.

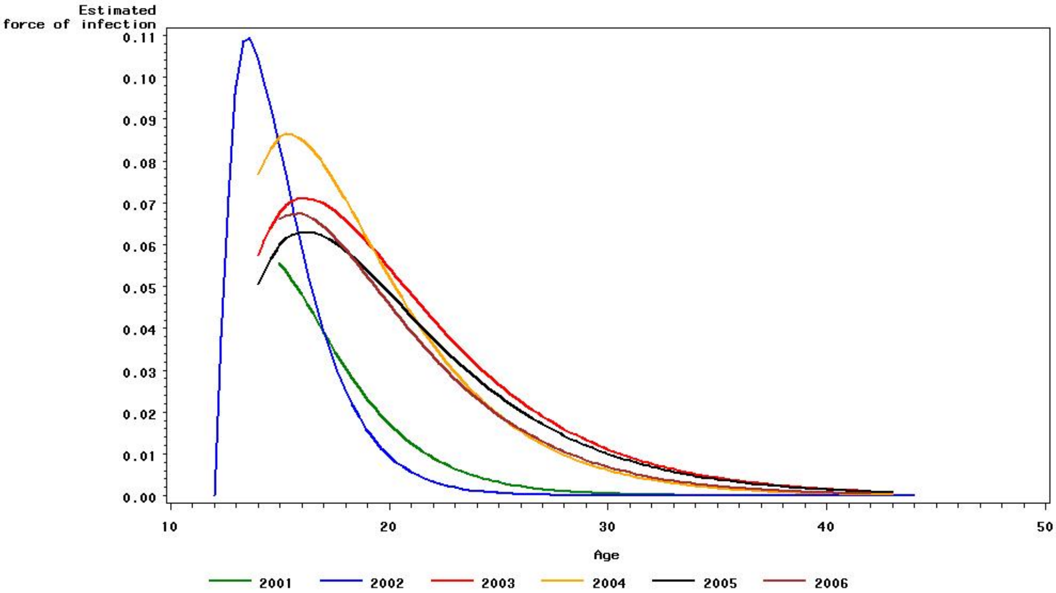


Figure 5.10: The fitted force of infection functions for each year using the Farrington's 2-parameter model.

higher rate than the other years and has a notably steeper decline than the other years.

The results show that both the log-logistic model and the Farrington's 2-parameter model estimate the HIV prevalence and the force of infection relatively well, when applied to the Vulindlela antenatal clinic data. The Farrington's 2-parameter model does seem to produce more realistic prevalence and force of infection curves. Using this model, the force of infection is particularly high at lower age values, being above 0.04 between ages 14 and 20. A peak of 0.074 at age 15 is observed. The force of infection then declines with age, becoming very low from age 30 onwards. Thus in the context of HIV among pregnant women, these trends of the force of infection with age result in a more epidemiologically plausible curve, making the Farrington's 2 parameter model a more preferable fit to the Vulindlela antenatal data.

Recall that in cross sectional prevalence data, all the observations are censored. For an HIV infected individual the true age at infection would have occurred some time before the age at testing, in which case the age at testing is left censored. In the case of right censored individuals, infection may occur at some age beyond the age at testing. Interval censoring occurs when the true age at infection is known to lie between two observed testing ages. It is thus important to develop models which account for the different types of censoring.

Chapter 6

Covariate dependence on the hazard of a positive HIV test

6.1 Introduction

The proportional hazards regression model, introduced by Cox (1972), has become by far the most widely used procedure for modelling the relationship of many explanatory variables on survival times. The model assumes a parametric form for the effects of the explanatory variables but no specified form of probability distribution is assumed for the survival times. Cox's model is therefore a *semiparametric model*.

The model is however based on the assumption of proportional hazards, which refers to the fact that the ratio of hazard functions for any two individuals will be constant over the survival time, but only depend on measured subject-specific explanatory variables.

6.2 The general proportional hazards model

Suppose we wish to examine the relationship of a set of explanatory variables on the hazard of a particular event of interest. Let $x_1, x_2, ..., x_p$ be the values of p explanatory variables $X_1, X_2, ..., X_p$ observed for each of n individuals. These observed values form the covariate vector x $= (x_1, x_2, ..., x_p)'$. Let $h_0(t)$ be the hazard function for an individual for whom the values of all the explanatory variables in x are zero. The function $h_0(t)$ is an unspecified nonnegative function and is known as the *baseline hazard*.

The hazard for the $i'th$ individual is then given by

$$\begin{aligned} h_i(t) &= e^{\beta' \mathbf{x}_i}\, h_0(t) \\ &= exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi})\, h_0(t), \end{aligned} \tag{6.1}$$

where β is a p x 1 vector of coefficients corresponding to the explanatory variables.

The term $e^{\beta' \mathbf{x}_i}$ is called the *hazard ratio* or *relative hazard*. It gives the hazard at time t for an individual with vector of observed explanatory variables $\mathbf{x}_i$ relative to the hazard for an individual with $\mathbf{x} = \mathbf{0}$.
The hazard ratio cannot be negative and the exponential in (6.1) plays an important role in ensuring that this is indeed the case.

To better understand the concept of proportional hazards, consider the ratio of hazards for two individuals with fixed vectors of explanatory variables $\mathbf{x}_i$ and $\mathbf{x}_j$,

$$\frac{h_i(t)}{h_j(t)} = \frac{e^{\beta' \mathbf{x}_i} h_0(t)}{e^{\beta' \mathbf{x}_j} h_0(t)} = \frac{e^{\beta' \mathbf{x}_i}}{e^{\beta' \mathbf{x}_j}}\ . \tag{6.2}$$

Taking the log of this hazard ratio gives $ln[h_i(t)/h_j(t)] = \beta'(\mathbf{x}_i - \mathbf{x}_j)$, which is constant over time. In other words, the hazard for one individual is a fixed proportion of the hazard for the other individual. A graph of the log hazards for the two individuals would therefore show parallel curves over time.

It is important to note that $h_0(t)$ cancels out of the numerator and denominator in (6.2). Furthermore, no functional form for $h_0(t)$ was specified. We will see in Section 6.3 that the β-coefficients of a proportional hazards model can be estimated without specifying the functional form of $h_0(t)$.

6.2.1 The survivor function

The survivor function for the $i'th$ individual is given by

$$S_i(t) = [S_0(t)]^{exp(\beta' \mathbf{x}_i)}, \tag{6.3}$$

where $\mathbf{x}_i$ is the vector of observed explanatory variables for the $i'th$ individual. $S_0(t)$ is the *baseline survivor function* and can be expressed as

$$S_0(t) = exp(-H_0(t)) = exp\Big(- \int_0^t h_0(t)\ dt \Big) \tag{6.4}$$

where $H_0(t)$ denotes the *cumulative baseline hazard function*.

6.3 Fitting the proportional hazards model

The unknown parameters $\beta_1, \beta_2, \ldots, \beta_p$ of the proportional hazards model are estimated using the *method of maximum likelihood*. Cox (1972) proposed the *partial likelihood function*, which depends only on the parameters of interest. This partial likelihood allows for estimation of the model parameters without having to specify the baseline hazard function $h_0(t)$.

The resulting parameter estimates have similar distributional properties to full maximum likelihood estimates. In particular, they are asymptotically normal and approximately unbiased (Allison, 1995). While there is some loss of information about β in using the partial likelihood instead of the full likelihood function, so that the resulting estimates are not fully efficient, Efron (1977) shows that the loss of efficiency is very small.

Another interesting property of the partial likelihood estimates is that they depend only on the ranks of the event times, rather than their numerical values. Thus any monotonic transformation of the event times will not alter the parameter estimates (Allison, 1995).

6.3.1 The partial likelihood function

Suppose that in a sample there are n observed individuals, $n - r$ right-censored survival times and r event times. We assume for the moment that only one event occurs at any one time. The r event times are ordered so that $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$ and $t_{(j)}$ is the $j'th$ ordered event time. The set of individuals who are at risk of experiencing the event at a time just prior to $t_{(j)}$ is called the *risk set*, denoted by $R(t_{(j)})$.

Then the partial likelihood function proposed by Cox (1972) is

$$L(\beta) = \prod_{j=1}^{r} \frac{exp(\beta' \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} exp(\beta' \mathbf{x}_l)} \;, \tag{6.5}$$

where $\mathbf{x}_{(j)}$ is the vector of explanatory variables for the individual who experiences the event at time $t_{(j)}$.

The denominator in (6.5) is the sum of values of $exp(\beta' \mathbf{x})$ over all individuals who are at risk at time $t_{(j)}$. It thus includes both right-censored individuals and those who experience the event at $t_{(j)}$. In other words censoring occurs after observed event times. Note that the product in the partial likelihood is taken only over the individuals who experience the event, whereas a full likelihood function takes the product over all the individuals in the sample. Therefore, a right-censored individual (i.e. one who does not experience the event at $t_{(j)}$ or any time prior to $t_{(j)}$) will not feature in the numerator, but will be included in the summation over the risk sets in the denominator.

Furthermore, we see that the partial likelihood depends only on the ranking of event times, as this is what determines the risk set at each event time $t_{(j)}$. Cox (1972) argued that the intervals between successive event times carry no information about β. Hence, since $h_0(t)$ has no specified functional form, it is possible that $h_0(t)$ and thus $h(t)$ is zero in these intervals.

Let us look briefly at how Cox's partial likelihood function in equation (6.5) is derived.

Deriving the partial likelihood function

Consider the probability that an individual experiences an event at time $t_{(j)}$, given that only one event may occur at any one time;

$$P(\text{individual with explanatory variables } \mathbf{x}_{(j)} \text{ has an event at } t_{(j)} | \text{one event at } t_{(j)}). \tag{6.6}$$

By the law of conditional probability $P(A|B) = P(A \text{ and } B)/P(B)$, so that (6.6) becomes

$$\frac{P(\text{individual with explanatory variables } \mathbf{x}_{(j)} \text{ has an event at } t_{(j)})}{P(\text{one event at } t_{(j)})}.$$

Event times are assumed to be independent of each other. Hence the denominator in the above expression is the sum of the probabilities of experiencing an event at time $t_{(j)}$ over all individuals who are at risk for the event at $t_{(j)}$. The expression above then equates to

$$\frac{P(\text{individual with explanatory variables } \mathbf{x}_{(j)} \text{ has an event at } t_{(j)})}{\sum_{l \in R(t_{(j)})} P(\text{individual } l \text{ has an event at } t_{(j)})} .$$

The single time point $t_{(j)}$ is replaced with the interval $(t_{(j)}, t_{(j)} + \delta t)$, and by dividing the numerator and denominator by δt we have

$$\frac{P[\text{individual with explanatory variables } \mathbf{x}_{(j)} \text{ has an event in } (t_{(j)}, t_{(j)} + \delta t)]/\delta t}{\sum_{l \in R(t_{(j)})} P[\text{individual } l \text{ has an event in } (t_{(j)}, t_{(j)} + \delta t)]/\delta t} .$$

By taking limits as $\delta t \rightarrow 0$ and by the definition of a hazard function, this expression becomes

$$\frac{h_i(t)}{\sum_{l \in R(t_{(j)})} h_l(t_{(j)})} ,$$

for the $i'th$ individual. Substituting (6.1) into this expression gives

$$\frac{h_0(t)\ exp(\beta' \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} h_0(t)\ exp(\beta' \mathbf{x}_l)} .$$

The baseline hazard cancels from the numerator and the denominator, so that we have

$$\frac{exp(\beta' \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} exp(\beta' \mathbf{x}_l)} .$$

Taking the product of these probabilities over all event times $j = 1, ..., r$ we obtain the partial likelihood function as given in equation (6.5).

6.4 Estimation of the β coefficients

Following the method of maximum likelihood, the first derivatives of the log likelihood with respect to $\beta_k, k = 1, ..., p$ are equated to zero, and solving these equations yields the parameter estimates $\hat{\beta}_k$. The vector of parameters is denoted as $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$. From equation (6.5) the log partial likelihood is given by

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{r} \left(\beta' \mathbf{x}_{(j)} - ln \Big[\sum_{l \in R(t_{(j)})} exp(\beta' \mathbf{x}_l) \Big] \right) \tag{6.7}$$

The first derivative of the log partial likelihood with respect to β_k is then

$$\begin{aligned} \frac{\partial \ell(\beta)}{\partial \beta_k} &= \sum_{j=1}^{r} \left(x_{(jk)} - \frac{\sum_{l \in R(t_{(j)})} x_{lk}\; exp(\beta' \mathbf{x}_l)}{\sum_{l \in R(t_{(j)})} exp(\beta' \mathbf{x}_l)} \right) \\ &= \sum_{j=1}^{r} \left(x_{(jk)} - \sum_{l \in R(t_{(j)})} w_{jl}(\beta)\; \mathbf{x}_l \right) \end{aligned} \tag{6.8}$$

where

$$w_{jl}(\beta) = \frac{exp(\beta' \mathbf{x}_l)}{\sum_{s \in R(t_{(j)})} exp(\beta' \mathbf{x}_s)} \; .$$

The term $x_{(jk)}$ in (6.8) denotes the value of the explanatory variable X_k for the individual who has an event at time $t_{(j)}$.

We need to find the set of first derivatives for all p explanatory variables, that is $(\partial\ell(\beta)/\partial\beta_1), (\partial\ell(\beta)/\partial\beta_2), \ldots, (\partial\ell(\beta)/\partial\beta_p)$. Setting $(\partial\ell(\beta)/\partial\beta_k) = 0$ for each β_k, $k = 1, ..., p$ gives rise to the score equations, which when simultaneously solved, give the vector of parameter estimates $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)'$.

To estimate the variance of $\hat{\boldsymbol{\beta}}$ consider the second derivative of the log partial likelihood in equation (6.7) with respect to β_k, which is given by

$$\begin{aligned} \frac{\partial^2 \ell(\beta)}{\partial \beta_k^2} &= -\sum_{j=1}^{r} \left(\frac{\big(\sum_{l \in R(t_{(j)})} exp(\beta' \mathbf{x}_l)\big)\big(\sum_{l \in R(t_{(j)})} x_{lk}^2 exp(\beta' \mathbf{x}_l)\big) - \big(\sum_{l \in R(t_{(j)})} x_{lk} exp(\beta' \mathbf{x}_l)\big)^2}{\big(\sum_{l \in R(t_{(j)})} exp(\beta' \mathbf{x}_l)\big)^2} \right) \\ &= -\sum_{j=1}^{r} \left(\frac{\sum_{l \in R(t_{(j)})} x_{lk}^2\; exp(\beta' \mathbf{x}_l)}{\sum_{l \in R(t_{(j)})} exp(\beta' \mathbf{x}_l)} - \left(\frac{\sum_{l \in R(t_{(j)})} x_{lk}\; exp(\beta' \mathbf{x}_l)}{\sum_{l \in R(t_{(j)})} exp(\beta' \mathbf{x}_l)} \right)^2 \right). \end{aligned} \tag{6.9}$$

The p x p information matrix takes the form

$$\jmath(\beta) = -\frac{\partial^2 \ell(\beta)}{\partial \beta^2} \ ,$$

thereby including all second-order partial derivatives of $\ell(\beta)$.

The variance-covariance matrix of $\hat{\beta}$ is the inverse of the observed information matrix evaluated at $\hat{\beta}$,

$$\widehat{Var(\hat{\beta})} = \jmath^{-1}(\hat{\beta}).$$

The vector of parameter estimates $\hat{\beta}$ is consistent and asymptotically normally distributed with mean β, the true parameter vector, and variance $[E(\jmath(\hat{\beta}))]^{-1}$, the inverse of the expected information matrix.

Many statistical packages, including SAS, use the Newton-Raphson algorithm to fit a proportional hazards model. This algorithm is remarkably robust for the Cox partial likelihood (Therneau and Grambsch, 2000). Details of the Newton-Raphson estimation procedure are discussed in Chapter 3 Section 3.3.1.

6.5 Handling ties

Up until now we have assumed that only one event can occur at any one time. Tied event times are the result of more than one event taking place at a time $t_{(j)}$.

The proportional hazards model assumes that the hazard function is continuous, and therefore tied event times are not possible (Collett, 2003). However, event times are often recorded to the nearest day, month or year, and ties are therefore likely to be present as a result of these imprecise measurement units.

If many events can occur at a time $t_{(j)}$, then it is also possible for there to be more than one censored observation at that time $t_{(j)}$. When there are both events and censored observations at any time point, the censoring is assumed to occur after the events. This eliminates any uncertainty about which individuals are included in the risk set at that time point.

In order to incorporate tied event times into a proportional hazards model, the likelihood function in (6.5) must be modified. The exact expression for the partial likelihood for tied data is given by

$$L(\beta) = \prod_{j=1}^{r} \left(\int_0^{\infty} \prod_{k \in D(t_{(j)})} \left[1 - exp\left(- \frac{exp(\beta' \mathbf{s}_j)}{\sum_{l \in R(t_{(j)})} exp(\beta' \mathbf{s}_l)} \, t \right) \right] exp(-t) \, dt \right) , \quad \textbf{(6.10)}$$

where s_j is the vector of sums of each of the p explanatory variables for all individuals who experience an event at $t_{(j)}$, $j = 1, ..., r$. $D(t_{(j)})$ denotes the set of all individuals who experience an event at $t_{(j)}$.

The basis for constructing the exact partial likelihood is to assume that the d_j tied events at time $t_{(j)}$ are due to imprecise measurement of time, and that there is a true ordering of the d_j events. Thus the tied events could have occurred in any one of the d_j! possible arrangements of their values. The denominator in the exact partial likelihood is modified so as to include each of these arrangements.

Computation of the exact partial likelihood can be very time consuming, particularly when there are a large number of ties at one or more event times (Collett, 2003). Approximations to the likelihood function that are less computationally intensive and that still account for the presence of tied event times were proposed by Breslow (1974) and Efron (1977).

The simpler approximation of the two was given by Breslow (1974), who proposed the approximate partial likelihood

$$L(\beta) = \prod_{j=1}^{r} \frac{exp(\beta' \mathbf{s}_j)}{\left[\sum_{l \in R(t_{(j)})} exp(\beta' \mathbf{x}_l) \right]^{d_j}} \qquad \textbf{(6.11)}$$

where d_j is the number of events occurring at time $t_{(j)}$ and s_j is the vector of sums of each of the explanatory variables for all individuals who experience the event at time $t_{(j)}$. The d_j events are assumed to be distinct and to occur sequentially. The denominator in (6.11) is the summation over all possible sequences of the events. The Breslow approximation is relatively straightforward to compute and is an adequate approximation when the number of tied observations at any one time is not too large.

The approximation proposed by Efron (1977) uses as the partial likelihood

$$L(\beta) = \prod_{j=1}^{r} \frac{exp(\beta' \mathbf{s}_j)}{\prod_{k=1}^{d_j} \left[\sum_{l \in R(t_{(j)})} exp(\beta' \mathbf{x}_l) - (k-1) d_j^{-1} \sum_{l \in D(t_{(j)})} exp(\beta' \mathbf{x}_l) \right]} \ . \quad (6.12)$$

This yields a closer approximation to the exact partial likelihood than that proposed by Breslow. However, in practise the two approximations often give similar results (Collett, 2003).

The exact, Efron and Breslow methods for handling tied observations are all available in the SAS software package. The Breslow method is the default method for handling ties in many statistical software packages, including SAS.

Cox (1972) introduced an alternative approximation for the model where the time-scale is viewed as being discrete, so that under this model tied observations are permissable. The approximation is given by

$$L(\beta) = \prod_{j=1}^{r} \frac{exp(\beta' \mathbf{s}_j)}{\sum_{l \in R(t_{(j)}; d_j)} \ exp(\beta' \mathbf{s}_l)} \ , \quad (6.13)$$

where $R(t_{(j)}; d_j)$ is a set of d_j individuals drawn from $R(t_{(j)})$, the risk set at time $t_{(j)}$. The proportional hazards model with discrete time-scale takes the form

$$\frac{h_i(t)}{1 - h_i(t)} = exp(\beta' \mathbf{x}_i) \ \frac{h_0(t)}{1 - h_0(t)} \ ,$$

with corresponding partial likelihood function given by equation (6.13). Under this model, the hazard function, $h_i(t)$, for an individual with explanatory variables $\mathbf{x}_i$, is the probability of experiencing the event in the interval $(t, t+1)$ given survival to time t. When the width of the time intervals tends to zero the model tends to the proportional hazards model in equation (6.1).

When there are no tied observations (i.e. $d_j = 1$ at each time $t_{(j)}$) then equations (6.10), (6.11), (6.12) and (6.13) all reduce to the partial likelihood function in equation (6.5).

6.6 Estimating the survivor and hazard functions

The survivor function for the $i'th$ individual in a proportional hazards model, found in (6.3) and repeated here for convenience, is

$$S_i(t) = [S_0(t)]^{exp(\beta' \mathbf{x}_i)}. \tag{6.14}$$

This equation indicates that once we have estimates of the regression coefficients, all we need is an estimate of the baseline survivor function $S_0(t)$, in order to produce estimates of the survival probability at each time $t_{(j)}$ and for different values of the explanatory variables. Similarly, the hazard function in (6.1) for the $i'th$ individual can only be estimated once an estimate of $h_0(t)$ has been found.

It is assumed that the hazard is constant between adjacent event times. The estimated baseline hazard function at time $t_{(j)}$ has the form

$$\hat{h}_0(t_{(j)}) = 1 - \hat{\alpha}_j \ , \tag{6.15}$$

where $\hat{\alpha}_j$ is the solution of the equation

$$\sum_{l \in D(t_{(j)})} \frac{exp(\hat{\beta}' \mathbf{x}_l)}{1 - \hat{\alpha}_j^{exp(\hat{\beta}' \mathbf{x}_l)}} = \sum_{l \in R(t_{(j)})} exp(\hat{\beta}' \mathbf{x}_l) \tag{6.16}$$

for $j = 1, ..., r$. $D(t_{(j)})$ in (6.16) denotes the set of all individuals who experience an event at time $t_{(j)}$ and $R(t_{(j)})$ is the set of all individuals (both censored and uncensored) who are at risk at $t_{(j)}$.

When there are no tied event times $D(t_{(j)})$ contains only one individual, and the solution to (6.16) is

$$\hat{\alpha}_j = \left(1 - \frac{exp(\hat{\beta}' \mathbf{x}_{(j)})}{\sum_{l \in R(t_{(j)})} exp(\hat{\beta}' \mathbf{x}_l)}\right)^{exp(-\hat{\beta}' \mathbf{x}_{(j)})}$$

When there are tied event times, iterative methods are required for obtaining a solution to equation (6.16).

The estimator of the baseline survivor function is then given by

$$\hat{S}_0(t) = \prod_{j=1}^{k} \hat{\alpha}_j \tag{6.17}$$

for $t_{(k)} \leqslant t < t_{(k+1)}$, $k = 1, ..., r-1$, and where $\hat{\alpha}_j$ is the solution to (6.16). This is the estimator for α_j used in several statistical packages including SAS. An alternative estimator for α_j due to Breslow (1974) has the form

$$\hat{\alpha}_j = exp\left(\frac{-d_j}{\sum_{l \in R(t_{(j)})} exp(\hat{\beta}'\mathbf{x}_l)} \right),$$

where d_j is the number of individuals who have an event at $t_{(j)}$. Using this estimator, $\hat{S}_0(t)$ is again given by (6.17).

The estimate of the survivor function in (6.14) is then obtained by substituting the estimated baseline survivor function as well as the maximum partial likelihood parameters $\hat{\beta}_k$ and the observed values of the explanatory variables, so that

$$\hat{S}_i(t) = [\hat{S}_0(t)]^{exp(\hat{\beta}'\mathbf{x}_i)}.$$

It is then also possible to estimate the baseline hazard at time $t_{(j)}$ using equation (6.15). However, Hosmer and Lemeshow (1999) show that individual pointwise estimates of the baseline hazard function are typically too unstable or "noisy". When these estimates are graphed against time it is difficult to identify the general shape of the underlying baseline hazard.

The baseline cumulative hazard function, $H_0(t)$, on the other hand, gives more stable estimates. Using the relationship $\hat{S}_0(t) = exp(-\hat{H}_0(t))$, the estimated baseline cumulative hazard is

$$\hat{H}_0(t) = -ln[\hat{S}_0(t)].$$

The cumulative hazard for the $i'th$ individual with explanatory variables $\mathbf{x}_i$ is then

$$\hat{H}(t) = exp(\hat{\beta}'\mathbf{x}_i)\ \hat{H}_0(t),$$

which when plotted against time may provide a useful graphical descriptor of the "risk experience" (Hosmer and Lemeshow, 1999). Note that if the estimated survivor function for individual i, $\hat{S}_i(t)$, has been obtained, then one can also use $\hat{H}_i(t) = -ln[\hat{S}_i(t)]$ to obtain the cumulative hazard for this individual.

6.7 Application to Vulindlela antenatal clinic data

This section discusses the fitting of a proportional hazards model to the Vulindlela antenatal clinic data, to examine the effect of explanatory variables such as partner's age and the number of previous pregnancies on the hazard of an HIV positive test.

Besides an individual's age and HIV status, additional variables recorded in the dataset are *partner's age*, *number of previous pregnancies* and *antenatal clinic attended*. These additional variables are discussed in Chapter 2. However not all three of these variables were recorded in each year's data. The variables captured in each year are:

2001 - age   status

2002 - age   status   clinic

2003 - age   status   clinic   partner's age

2004 - age   status   clinic   partner's age   previous pregnancies

2005 - age   status   clinic   partner's age   previous pregnancies

2006 - age   status   partner's age   previous pregnancies

Hence, a different proportional hazards model was fitted for each year, due to the different combinations of variables recorded in each year's data.

The individual's age is the *time to event* or *survival time*, and infection with HIV is the event of interest. The *censored* event times are the ages of those individuals who are found to be uninfected (i.e. those whose HIV status = 0 at the time of testing). This is because these individuals had not yet experienced the event of interest at the time of testing. Note that we really need not know the exact age at infection but rather we are using the age at the time of HIV testing as a proxy for the actual time to event.

Recall that because we are using cross-sectional data; for uncensored individuals the age at testing is not necessarily the age at which the individual was infected. An individual who tests HIV positive at age a could have been infected at any age prior to a.

The proportional hazards model will examine the effect of explanatory variables, such as partner's age and the number of previous pregnancies, on the hazard of HIV infection, with the age at HIV testing as the time to event.

6.7.1 Model fitting

SAS Proc Phreg was used to fit a series of proportional hazards regression models to the Vulindlela antenatal clinic data. The Phreg procedure readily incorporates explanatory variables that are measured on a continuous scale, such as partner's age. Categorical explanatory variables (those with three or more finite response categories) must be specified using the *class* statement in Proc Phreg. Each response category or level is then compared to a reference category. SAS uses the last category as its reference category by default, unless otherwise specified.

The number of previous pregnancies experienced by an individual took on one of five values ($0, 1, 2, 3$ or 4). These five categories were collapsed to create a variable defined as follows:

$$\text{Previous pregnancies} = \begin{cases} 0 \text{ if the individual has never been pregnant before} \\ 1 \text{ if the individual has had 1 previous pregnancy} \\ 2 \text{ if the individual has had} \geqslant 2 \text{ previous pregnancies} \end{cases}$$

This variable was specified as categorical. Note that previous pregnancies refers to the number of pregnancies experienced by the individual prior to their current pregnancy. The first category (no previous pregnancies) was used as the baseline or reference category for the analysis. Therefore, the hazard of infection corresponding to individuals having either 1 or $\geqslant$2 previous pregnancies would each be interpreted relative to those who had never been pregnant before.

The variable *clinic* is also categorical. It has eight categories, representing any one of the antenatal clinics an individual attended.

The age of each individual's male partner was recorded under the variable *Partner age*, which was treated as a continuous variable when fitting the model. To investigate the effect of a large age difference between an individual and her male partner, the binary variable *Agediff_eight* was created, so that

$$\text{Agediff\_eight} = \begin{cases} 1 \text{ if the individual's partner is } \geqslant 8 \text{ years older than the individual} \\ 0 \text{ if age difference between the individual and her partner is } < 8 \text{ years} \end{cases}$$

The Efron method was used for handling ties. Analyses using the exact, Efron and Breslow methods showed that estimates produced using the Efron method were closer to those of the exact method, than the Breslow method.

6.7.2 Results

A model was fitted to the combined data for years 2004-2006, using the explanatory variables Partner age, Previous pregnancies and Agediff_eight. Observations for years 2001-2003 could not be used as they do not have values for all of these explanatory variables. A total of 1180 observations were used of which 469 were events (HIV positive individuals) and 711 were censored (HIV negative individuals). The percentage of censored individuals is 60.25%. Table 6.1 shows the model fit statistics from the SAS output. The lower the values of AIC, SBC and -2 log likelihood the better the fit of the model to the data. The output shows that inclusion of the above-mentioned explanatory variables results in lower values of these fit statistics.

Table 6.1: Model fit statistics for the proportional hazards model of combined 2004-2006 data

Criterion	Without covariates	With covariates
-2 log L	5665.641	4978.527
AIC	5665.641	4986.527
SBC	5665.641	5003.129

Table 6.2: Testing Global Null Hypothesis: BETA=0 for the proportional hazards model of combined 2004-2006 data

Test	Chi-Square	DF	Pr > Chi-Sq
Likelihood ratio	687.1138	4	<.0001
Score	527.1107	4	<.0001
Wald	485.0505	4	<.0001

Table 6.3: Parameter estimates for the proportional hazards model of combined 2004-2006 data

Parameter		DF	Estimate	Std Err	Chi-Square	Pr > ChiSq	HR
Partner age		1	-0.25978	0.01341	375.1616	<.0001	0.771
Prev preg -	One	1	-0.42583	0.11278	14.2552	0.0002	0.653
	Two or more	1	-1.25147	0.16817	55.3789	<.0001	0.286
Agediff_eight		1	2.46719	0.16976	211.2174	<.0001	11.789

The Global Null Hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is used to test if the overall effect of all the explanatory variables is significant. Table 6.2 shows that using the likelihood ratio, score or Wald tests, we can reject H_0 and conclude that the overall model is significant.

Of greater interest, however, is the effect of each individual explanatory variable, given by the table of parameter estimates (Table 6.3). Values under the column labeled 'Estimate' give the estimates of β_k for the $k'th$ explanatory variable and 'Std Err' gives the standard error of each estimate, $s.e.(\hat{\beta}_k)$. It can be seen that all the explanatory variables do each have a significant effect on the hazard of infection ($p < 0.05$).

The hazard ratio (labeled HR) is the value of $exp(\hat{\beta}_k)$. For example, the parameter estimate associated with having one previous pregnancy is $\hat{\beta}_k = -0.42583$, with hazard ratio $exp(-0.42583) = 0.653$. Therefore an individual who has had one previous pregnancy has a hazard rate that is about two thirds of that of an individual who has had no previous pregnancies (i.e. who has never been pregnant before). Hazard ratios lower than 1 can be better interpreted by their inverse, in this case $1/0.653 = 1.531$. A better interpretation is that the hazard of infection for an individual who has had no previous pregnancies is 1.53 times higher than for an individual who has had one previous pregnancy. The hazard ratio associated with having two or more previous pregnancies is 0.286, which is lower than the hazard ratio for having had one previous pregnancy. In other words, the hazard of infection for an individual who has had no previous pregnancies is $1/0.286 = 3.5$ times higher than for an individual who has had two or more previous pregnancies. In summary, having one previous pregnancy is associated with a lower hazard of infection than having had no previous pregnancies, but having had two or more previous pregnancies is associated with an even lower infection hazard. The results therefore suggest that the greater the number of previous pregnancies the lower the hazard rate. One would then expect individuals who have never been pregnant before to have the highest infection hazard, and those who have experienced two or more previous pregnancies to have the lowest.

The hazard ratio corresponding to the variable Partner age is less than 1, meaning that an increase in partner's age by one year results in a decreased hazard of infection. An individual whose partner is aged $a + 1$ has a hazard of infection that is 0.771 of that of an individual whose partner is aged a. Alternatively, in terms of its inverse, the hazard of infection for an individual with partner aged a is 1.3 times higher than for an individual with partner aged $a + 1$. A high hazard ratio was observed for the binary variable Agediff_eight. Hence the hazard of infection for individuals whose partners were older than them by eight or more years was 11.8 times higher than for those who had a partner either less than eight years older than them or younger than them.

In summary, the results from fitting this proportional hazards model show that the age of an individual's partner, the number of previous pregnancies and whether a partner was older than an individual by eight or more years, all do have a significant effect on the hazard of HIV infection. When interpreting these results one should bear in mind, however, that strictly speaking we are referring to the hazard of a positive HIV test. The results showed that the older the partner, regardless of the age of the individual, the lower the hazard rate. However, when a partner was eight or more years older than an individual, this resulted in a much higher hazard relative to cases when the age difference between a partner and individual was less than eight years.

The hazard was lower for individuals who had one previous pregnancy than for those who had never been pregnant before, and even lower for those who had two or more previous pregnancies. A possible reason for this is that women who have many children are likely to be more responsible, more focused in taking care of their children and more careful to avoid engaging in behaviours that would place themselves at risk for HIV infection.

Proc Phreg produces estimates of the survivor function $S(t)$ either for specified values of the explanatory variables, or for their mean values. Figure 6.1 shows the estimated survival curves for each of the years 2004, 2005 and 2006, for

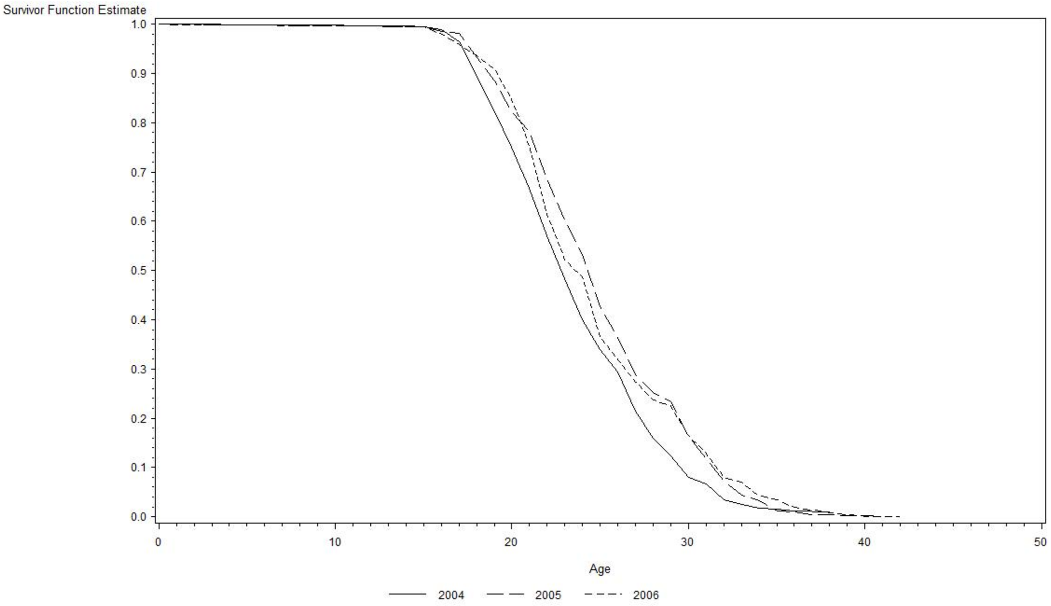


Figure 6.1: The estimated survivor functions for each of the years 2004, 2005 and 2006 for individuals with no previous pregnancies and whose partners' ages do not exceed theirs by eight or more years.

individuals who have experienced no previous pregnancies and whose partners' ages do not exceed theirs by eight or more years. This was done by fitting a proportional hazards model for each year, using only the explanatory variables Previous pregnancies and Agediff_eight. The baseline survivor functions in Figure 6.1 were then obtained by setting the values of both of these variables to be equal to 0. It can be seen that the slopes of the estimated survivor functions do not differ substantially by year.

In order to examine the effect of sets of explanatory variables on the hazard rate for each year, a proportional hazards model was fitted for each of the years 2003, 2004, 2005 and 2006. The analysis of parameter estimates for each year are presented in Table 6.4.

The model for year 2003 shows that Agediff_eight and Partners age do have a significant effect on the hazard of infection. The effect of the variable Clinic was not found to be significant in the models for each year. The hazard rate did not differ significantly across the antenatal clinics attended by participants

Table 6.4: Parameter estimates for the proportional hazards models for each of the years 2003-2006

2003							
Parameter		DF	Estimate	Std Err	Chi-Sq	Pr > ChiSq	HR
Partner age		1	-0.33576	0.03515	91.2672	<.0001	0.715
Agediff_eight		1	1.61693	0.59313	7.4317	0.0064	5.038
2004							
Parameter		DF	Estimate	Std Err	Chi-Sq	Pr > ChiSq	HR
Partner age		1	-0.30334	0.02172	195.1163	<.0001	0.738
Prev preg -	One	1	-0.39679	0.16506	5.7788	0.0162	0.672
	Two or more	1	-1.18753	0.25795	21.1940	<.0001	0.305
Agediff_eight		1	2.63795	0.25957	103.2801	<.0001	13.984
2005							
Parameter		DF	Estimate	Std Err	Chi-Sq	Pr > ChiSq	HR
Partner age		1	-0.23447	0.02658	77.8132	<.0001	0.791
Prev preg -	One	1	-0.41037	0.22689	3.2713	0.0705	0.663
	Two or more	1	-1.11460	0.30632	13.2401	0.0003	0.328
Agediff_eight		1	2.47434	0.32884	56.6181	<.0001	11.874
2006							
Parameter		DF	Estimate	Std Err	Chi-Sq	Pr > ChiSq	HR
Partner age		1	-0.22694	0.02534	80.2263	<.0001	0.797
Prev preg -	One	1	-0.50343	0.23162	4.7240	0.0297	0.604
	Two or more	1	-1.28087	0.35657	12.9039	0.0003	0.278
Agediff_eight		1	2.16740	0.33191	42.6407	<.0001	8.736

($p > 0.05$ for all). The *Clinic* variable was therefore excluded from the models for years 2003, 2004 and 2005, the years in which this variable was recorded.

The analysis by individual year showed similar results to that of the overall model for the combined 2004-2006 data. Partners age and Agediff_eight have a significant effect on the hazard of infection in each year. An individual with a partner of age $a + 1$ has a hazard of infection that is roughly 0.7 of that of an individual with partner aged a. Individuals whose partners were older than them by eight or more years are expected to have a much higher hazard of infection than those with a smaller partner-individual age difference. The effect of having experienced previous pregnancies was significant in each of the years, except that in 2005 the hazard associated with having had one previous pregnancy did not significantly differ from that associated with having no previous pregnancies ($p = 0.0705$). The hazard ratios associated with having had one previous pregnancy or two or more previous pregnancies were similar to that observed in the combined model (see Table 6.3). In each year the hazard rate was lower for those who had one previous pregnancy relative to those who had no previous pregnancies, and even lower for those who had two or more previous pregnancies relative to those who had no previous pregnancies.

In understanding why women with previous pregnancies have a lower hazard rate than those who have never been pregnant before, an analysis of the demographic and behavioural characteristics of women who have multiple pregnancies is required. Tables 2.2 and 2.3 show the age distribution, partner age distribution and the partner-age difference among individuals having one, two or more, or no previous pregnancies. It can be seen that individuals with previous pregnancies tended to be older women. This may be indicative of stable relationships, which are associated with higher age. Those with two or more previous pregnancies were also more likely to have older partners. Hence, women having multiple pregnancies are likely to be older more mature women, in stable relationships with one partner. These women could be more responsible, and less prone to engaging in behaviours that would place themselves at risk for HIV infection.

Table 2.9 suggests that for very young women (<22 years old) and much older women (>31 years old) the observed HIV prevalence is higher for those with either one or two or more previous pregnancies than for those who have had no previous pregnancies. This could be in disagreement with the results of the proportional hazards models in this chapter. These results show that women with one or at least two previous pregnancies have a lower hazard rate than those who have had no previous pregnancies. It is important to note that the assumption of age at testing as the time to event should be interpreted with caution. This is because the actual age at infection is left censored, namely lower than the age at which the HIV test was performed. Therefore the true time to event is not correctly captured and this may introduce a bias in either direction. In fact in the current analysis the data can be both left and right censored. In survival analysis we assume right censoring. A more robust analysis should therefore account for both left and right censored data. However the problem of interval censored data is not present in the current thesis because HIV testing per individual was not carried out between two different age points. Future extensions of this model in the case of antenatal clinic data would thus need to account for the error arising from the fact that the actual age at infection is not accurately known.

Chapter 7

Conclusion

The prevalence, incidence and the force of infection are well known and very useful measures of disease. They play an important role in quantifying the burden of disease within a population at a given time or over a specified time period. The prevalence gives the probability to be diseased at a given time. It is the proportion of existing cases of the disease in the population, and thus includes both new and pre-existing cases of the disease, that is, those individuals who have acquired infection during the given time interval, as well as those who were infected prior to this time interval. Incidence, on the other hand, refers to the rate of new infections. Prevalence and incidence rates are therefore strongly related, since any prevalent case means that a new infection has occurred before. The force of infection, also referred to as the hazard of infection, is closely related to the incidence, in that both measures quantify the rate of new infections. Information on the rate and number of new infections is especially important, as it aids in evaluating interventions and disease prevention strategies, and in determining the rate at which treatment is to be supplied.

Data from pregnant women attending antenatal clinics has been the primary source for assessing HIV trends in South Africa. Antenatal clinic attendees serve as a large and easily accessible population. This thesis discusses methods for modelling the HIV prevalence, incidence and the force of infection, by age and time, for women of child bearing age using cross-sectional seroprevalence data. The data used is from pregnant women attending antenatal clinics in Vulindlela, an area of rural KwaZulu-Natal, in each of the years 2001-2006.

Logistic regression was used to model the HIV prevalence across time and age. The fitted prevalence slopes for each year showed a similar trend, with estimated prevalence rates rising with age to peaks of between 36% and 57% in the mid to late twenties and then steadily declining toward the early forties. There seems to be a shift in the peak prevalence over age, as the 2001 peak occurs at an earlier age than the other years. Furthermore, prevalence estimates in 2001 were notably lower than those of the subsequent years across all ages.

The simple interrelation of prevalence and incidence $P = I.E(D)$, as given by Freeman and Hutchison (1980), can be applied to many diseases including HIV. However information on the distribution of disease durations is needed. The early method of Leske et al (1983), which assumes equal mortality for infected and uninfected individuals, may not model HIV incidence sufficiently well, due to the high mortality rates among HIV infected individuals relative to uninfected individuals. Keiding (1991) provides the basis for studying prevalence and incidence within a probability framework. Keiding's expression for the prevalence odds was later used and adapted in subsequent studies of incidence estimation, including Sakarovitch et al. (2007).

Pregnant women attending public health antenatal clinics are generally not representative of all adult women. The antenatal samples do not include non-pregnant women, or pregnant women who are not consulting a public health antenatal clinic. The risk profiles and prevalence rates for these individuals may differ from those who are included in the antenatal samples. Hence, the Relative Inclusion Rate (RIR) is a valuable parameter of interest to take into account when modelling HIV incidence, as it adjusts for the probability of an infected woman being included in the sample relative to an uninfected woman.

Application of various well known models for the force of infection to the Vulindlela antenatal data showed that the log-logistic and the Farrington's 2-parameter model both fit the data reasonably well. The Farrington's 2-parameter model was chosen as the preferred model, since it seemed to give a more realistic reflection of the rate of new infections over age. The estimated force of infection functions for each year rise to peaks of between 0.055 and 0.109 at around age 16 and decline thereafter, tending toward zero in the mid to late forties. Beyond age 30 the force of infection is estimated at less than 0.01 for each year.

The proportional hazards regression model was used to determine the effect of additional explanatory variables in the dataset on the hazard of HIV infection. Note that this actually refers to the hazard of a positive HIV test, since the age at testing occurs either at or some time after the age at infection. The age of a woman's male partner, a large partner-participant age difference, and the number of previous pregnancies experienced all had a significant effect on the hazard rate. An individual with a partner of age $a+1$ had a hazard rate that was roughly 0.7 of that of an individual with partner aged a. Thus an increase in partner's age was associated with a decrease in the hazard rate. Women whose partners were eight or more years older than themselves had a much higher hazard rate than those whose partners were either younger than them or older by less than eight years. The hazard was lower for women who had one previous pregnancy relative to those who had never been pregnant before, and even lower for those who had two or more previous pregnancies than for those who had never been pregnant before. The models used the age at testing as a proxy for the age at infection, the latter being the true time to event. Future extensions of this model in the case of antenatal clinic data could explore incorporating both left and right censored observations. This future work would need to account for possible bias arising from the fact that the actual age at infection is not accurately known.

Bibliography

Ades AE, Medley GB. (1994). Estimates of disease incidence in women based on antenatal or neonatal seroprevalence data: HIV in New York City. *Statistics in Medicine*, **13**:1881-1894.

Alho J. (1992). On prevalence, incidence, and duration in general stable populations. *Biometrics*, **48**:587-592.

Allison P. (1995). *Survival Analysis Using the SAS System: A Practical Guide*. SAS Institute Inc. Cary, North Carolina.

Anscombe FJ. (1953). Contribution to the discussion of a paper by H. Hotelling. *Journal of the Royal Statistical Society*, **B15**:229-230.

Barndorff-Nielsen O. (1978). *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, Inc. New York.

Becker NG. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall. New York.

Breslow NE. (1974). Covariance analysis of censored survival data. *Biometrics*, **30**:89-100.

Brookmeyer R, Quinn T. (1995). Estimation of current Human Immunodeficiency Virus incidence rates from cross-sectional survey using early diagnostic tests. *American Journal of Epidemiology*, **141**(2):166-172.

Carpenter L. (1997). Estimates of the impact of HIV infection on fertility in a rural Ugandan population cohort. *Health Transition Review*, **7**(Suppl. 2):113-126.

Collett D. (1991). *Modelling Binary Data*. Chapman & Hall. London, U.K.

Collett D. (2003). *Modelling Survival Data in Medical Research*. 2nd edition. Chapman & Hall/CRC. Boca Raton, Florida.

Cook RD. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**:15-18.

Cox DR. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*: Series B, **34**:187-220.

Cox DR, Hinkley DV. (1974). *Theoretical Statistics*. Chapman & Hall. London, U.K.

Department of Health. (2004). *National HIV and syphilis antenatal seroprevalence survey in South Africa: 2003*.

Department of Health. (2007). *National HIV and syphilis antenatal seroprevalence survey in South Africa: 2006*.

Department of Health. (2010). *National Antenatal Sentinel HIV and Syphilis Prevalence Survey in South Africa, 2009*.

Diamond ID, McDonald JM. Analysis of current-status data. In *Demographic Application of Event History Analysis*, Trussel J, Hankinson R, Tiltan J (eds), Chapter 12. Oxford University Press. Oxford, 1992.

Dobson AJ. (1990). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC. London, U.K.

Ederer F. (1964). The effect of adjusting for competing mortality risks. *American Journal of Public Health*, **54**(7):1129-1133.

Efron B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72**:557-565.

Elandt-Johnson RC. (1975). Definition of Rates: Some remarks on their use and misuse. *American Journal of Epidemiology*, **102**:267-271.

Farrington CP. (1990). Modeling forces of infection for measles, mumps and rubella. *Statistics in Medicine*, **9**:953-967.

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

Gouws E. (2006). *Incidence of HIV infection in rural KwaZulu-Natal*. PhD thesis. Nelson R Mandela School of Medicine, University of KwaZulu-Natal. Durban, South Africa.

Grenfell BT, Anderson RM. (1985). The estimation of age-related rates of infection from case notifications and serological data. *Journal of Hygiene*, **95**(2):419-436.

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

Grummer-Strawn LM. (1993). Regression analysis of current status data: an application to breast feeding. *Journal of the American Statistical Association*, **88**:758-765.

Hastie TJ, Tibshirani RJ. (1990). *Generalized Additive Models*. Chapman & Hall. New York.

Hens N, Aerts M, Faes C, et al. (2010). Seventy-five years of estimating the force of infection from current status data. *Epidemiol Infect*, **138**:802-812.

Hosmer DW, Lemeshow S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. John Wiley & Sons, Inc. New York.

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

Krieger JN, Coombs RW, Collier AC, et al. (1991). Fertility parameters in men infected with human immunodeficiency virus. *Journal of Infectious Diseases*, **164**(3):464-469.

Langohr K. (1999). *Estimation of the Incidence of Disease with the Use of Prevalence Data*. Technical Report 12/1999. Fachbereich Statistik, Universitat Dortmund.

Leske M, Ederer F, Podgor M. (1981). Estimating incidence from age-specific prevalence in Glaucoma. *American Journal of Epidemiology*, **113**(5):606-613.

Lewis JJ, Ronsmans C, Ezeh A, et al. (2004). The population impact of HIV on fertility in sub-Saharan Africa. *Aids*, **18**(Suppl. 2):S35-S43.

Lindsey JK. (1997). *Applying Generalized Linear Models*. Springer. New York.

Martin PM, Gresenguet G, Herve VM, et al. (1992). Decreased number of spermatozoa in HIV-1-infected individuals. *Aids*, **6**(1):130.

McCullagh P, Nelder JA. (1989). *Generalized Linear Models*. 2nd edition. Chapman & Hall/CRC. London, U.K.

Muench H. (1959). *Catalytic Models in Epidemiology*. Harvard University Press. Boston.

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

Nelder JA, Wedderburn RWM. (1972). Generalised linear models. *Journal of the Royal Statistical Society: Series A*, **135**:370-384.

Ntozi JP. (1997). Widowhood, remarriage and migration during the HIV/AIDS epidemic in Uganda. *Health Transition Review*, **7**(Suppl.):125-144.

Pierce DA, Schafer DW. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, **81**:977-986.

Podgor M, Leske M, Ederer F. (1983). Incidence estimates for lens changes, macular changes, open-angle Glaucoma and diabetic retinopathy. *American Journal of Epidemiology*, **118**(2):206-212.

Podgor M, Leske M. (1986). Estimating incidence from age-specific prevalence for irreversible diseases with differential mortality. *Statistics in Medicine*, **5**:573-578.

Rossini AJ, Tsiatis AA. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*, **91**(423):713-721.

Sakarovitch C, Alioum A, Ekouevi DK, et al. (2007). Estimating incidence of HIV infection in childbearing age African women using serial prevalence data from antenatal clinics. *Statistics in Medicine*, **26**:320-335.

SAS Institute Inc. 2008. *SAS/STAT 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Shiboski SC. (1998). Generalized additive models for current status data. *Lifetime Data Analysis*, **4**:29-50.

Shisana O, Rehle T, Simbayi LC, et al. (2009). *South African National HIV Prevalence, HIV Incidence, Behavioural and Communications Survey, 2008*. Human Sciences Research Council Publishers. Cape Town.

Shkedy Z, Aerts M, Molenberghs G, et al. (2006). Modelling age-dependent force of infection from prevalence data using fractional polynomials. *Statistics in Medicine*, **25**:1577-1591.

Shkedy Z, Aerts M, Molenberghs G, et al. (2003). Modelling forces of infection by using monotone local polynomials. *Applied Statistics*, **52**(4):469-486.

Therneau TM, Grambsch PM. (2000). *Modelling Survival Data: Extending the Cox Model*. Springer. New York.

UNAIDS (2010). *UNAIDS Report on the Global AIDS Epidemic, 2010*. World Health Organisation. Geneva, Switzerland.

Williams DA. (1984). Residuals in generalized linear models. *Proceedings of the 12th International Biometrics Conference*. Tokyo, 59-68.