CURRENT TRENDS AND RECENT ADVANCES IN APPLIED BIOMETRY

Selected Topics in Statistical Modeling of Infectious Diseases Data

International Livestock Research Institute (ILRI)
Nairobi, Kenya

Khangelani Zuma and Ziv Shkedy
Human Sciences Research Council, Pretoria, South Africa
Center for Statistics, Limburgs Universitair Centrum, Diepenbeek, Belgium

August 30 - September 3, 2004
Part I

Introduction
Chapter 1

Introduction: Infectious Disease Data

- Examples: SARS, AIDS, Measles, Hepatitis B.
- Mathematical modeling:
  - Time homogenous model.
  - Mixing patterns.
  - The force of infection.
- The impact of migration on the spread of HIV/STI.
1.1 Outbreak of SARS

- Severe acute respiratory syndrome (SARS) is a viral respiratory illness caused by a coronavirus, called SARS-associated coronavirus (SARS-CoV).
- SARS was first reported in Asia in February 2003.
- Over a period of few months, the disease spread to more than two dozen countries in: North America, South America, Europe, and Asia.
- The SARS global outbreak of 2003 was contained; however, it is possible that the disease could re-emerge.
1.1.1 Cumulative Number of Cases

- Figure 1.1 shows the cumulative number of reported SARS cases. In March, 17, 2003 the number of reported cases was 167 and it increased to 7919 in May, 20, 2003.

Figure 1.1: Reported cumulative number of SARS cases (www.cdc.gov/ncidod/sars/ and www.who.int/csr/sars).
1.1.2 Spread of SARS Over the World

- The rapid spread of SARS can also be seen in Figure 1.2 which shows the number of countries with at least 1 reported case of SARS. In March 17, 2003, the number of countries is 7 and it increased to 31 by May, 20, 2003.

![Figure 1.2: Number of countries with at least 1 SARS case reported in May, 20, 2003.](www.cdc.gov/ncidod/sars/ and www.who.int/csr/sars).
1.2 Outbreak of AIDS in US

- Up to December 1992, the centers for disease control in USA (CDC) has reported on 249,199 cases of AIDS in USA (Hay and Wolak, 1994) and 169,623 AIDS related deaths.
- Figure 2.16 (solid line) shows the quarterly number of reported cases AIDS in USA. The decline in number of cases in the last few quarters is due to underreporting.

Figure 1.3: Quarterly number of AIDS cases in US.
1.3 Measles in England and Wales

- The national immunization program against measles in England and Wales began in 1968.
- Figure 1.4 shows the weekly number of measles cases from 1950 to 1979.

Figure 1.4: Weekly number of measles cases in England and Wales (from Fine and Clarkson, 1982).
1.4 Incidence of Measles

- Incidence of measles over the world.
- Africa versus Europe.

Figure 1.5: Measles over the world.
1.5 Hepatitis B St Petersburg, Russia (1994-1999)

- Figure 1.6 shows the monthly number of cases by months and reveals, as expected, differences in incidence evolution over time between age groups.

Figure 1.6: Hepatitis B St Petersburg, Russia (1994-1999).
1.6 Age-time Dependent Incidence of Hepatitis B in Bulgaria (1985-2000)

- Another example of incidence data of hepatitis B is shown in Figure 1.7 which shows the reported number of hepatitis B in Bulgaria from 1985 to 2000 by age.

- Similar to the pattern observed in St Petersburg patterns can be observed with respect to time and age.

- Since hepatitis B is mainly a sexual transmitted disease the highest incidence is observed for age group which are the most sexual active.
1.6.1 Incidence of Symptomatic Cases

Figure 1.7: Age-time dependent incidence. From top-left to right-bottom: (a) Observed incidence rate per 100000 from 1983 to 2000, (2) Loess model for the incidence, (3) Loess model for the incidence (oposite perspective relative to (2)) and (4) contour plot for the observed incidence rate.
1.7 Prevalence of hepatitis B

- Prevalence of Hepatitis B (1997) over the world.

![Geographic Pattern of Hepatitis B Prevalence, 1997](image)

Data as of 06/04/98

Figure 1.8: Hepatitis B over the world.
1.8 Sero Prevalence of Hepatitis A in Bulgaria and Belgium

- An example of current status data.
- For each individual:
  \[ Y_i = \begin{cases} 
  1, & \text{Infected,} \\
  0, & \text{Not infected.} 
\end{cases} \]

- Figure 1.9 shows two examples of serological data of hepatitis B in Bulgaria and Belgium (left panels).
Chapter 2

Mathematical models for Infectious Diseases

• Compartmental models.
• Basic models for transmission.
• Disease states: susceptible, infected, immune.
• The force of infection.
2.1 The Basic SIR Model

It is assumed (Anderson 1982, Anderson and May 1991, Bailey 1975) that for simple infectious diseases the individuals in the population can be classified, according to their infection status, into three states:

- Susceptible to infection: individuals who have not been exposed yet, this is the population at risk.
- Infected and infectious to others.
- Immune to re-infection.

Such a model is called the SIR model.
2.1.1 The SIR Model

- Disease states.

Figure 2.1: The Basic SIR model.
2.1.2 Notation

- The number of hosts at risk at time $t$ and age $a$ is denoted by $X(a, t)$.
- $Y(a, t)$ is the number of infected hosts at time $t$ and age $a$.
- $Z(a, t)$ is the number of immune hosts at age $a$ and time $t$.
- The total population is given by
  \[ N(a, t) = X(a, t) + Y(a, t) + Z(a, t). \]

Two time scales: age (a) and time (t).
2.1.3 Age and Time

- Incidence of hepatitis B in Bulgaria.

Figure 2.2: Hepatitis B, Bulgaria.
2.1.4 Assumptions

- It assumes that all newborns are entered directly into the susceptible class and therefore, the model ignores the maternal antibodies period.

- Another assumption is that the infection, the infectious period and the disease occur simultaneously, i.e. that the $SIR$ model ignores the latent period (the period in which the individual is infected but not infectious to others).
2.1.5 Flow Diagram

Figure 2.3: The Basic SIR model.
2.1.6 The Force of Infection

- The rate in which susceptible become infected.

**Figure 2.4:** The Basic SIR model.
2.1.7 Duration

• The duration of stay within each class varies from class to class.
  – For measles (in the developed countries) the mean age of exposure is approximately 4–5 years.
  – The infectious period is around 7 days.
  – The immunity is life long.

• These lengths of time hold for most of the childhood infectious diseases, the susceptible period lasts years, the infectious period days and the immunity is assumed to be life long.
2.1.8 Duration: A Graphical Display

Duration in the infection states

Birth

susceptible (years)

infectious (days)

immune (life long)

AGE

Figure 2.5: Duration in the infection states.
2.1.9 Duration: The Statistical Aspects

- If we take a sample of $N$ individuals from the population we expect that:

  \[
  \text{proportion of susceptibles} \approx 1 - \text{proportion of immune}
  \]

- This means that we model the change from the susceptible class to the immune class, ignoring the other compartments in the model.

![Diagram showing the transition from susceptible to immune](image)
2.2 Transmission

- A central characteristic of the population dynamics of infection diseases is the transmission of the infection from the infected class (assumed to be infectious in the SIR model) to the susceptible class.
- The mass-action principle states (Bailey 1975):
  \[
  \text{new cases of infection (the incidence)} = \beta Y X.
  \]
2.2.1 The Mass-action Principle

- Assumption: Mixing at random.
- Does it make sense?

\[
\text{Number of new cases} = P(\text{transmission}) \times \# \text{ of infectious} \times \# \text{ of susceptible}
\]

Figure 2.7: The mass-action principle.
2.2.2 The Mass-action Principle

- Number of susceptible.

\[
\text{Number of new cases} = P(\text{transmission}) \times \# \text{ of infectious} \times \# \text{ of susceptible}
\]

Figure 2.8: The mass-action principle.
2.2.3 Homogeneous Mixing and the Force of Infection

- The underlying assumption of the mass-action principle is that the infectious and the susceptible individuals are mixing in a random manner (homogeneous mixing), i.e., that $\beta$ is age and time independent.

- The force of infection $\lambda$, is the rate at which the host moves from the susceptible to the infected class.

- It is assumed that $\lambda$ is a linear function of the total number of infected hosts $Y$, i.e.,

$$\lambda = \beta Y$$
2.3 Mathematical Models for Transmission Dynamics

- Bailey (1975, 1982), Anderson (1982) and Anderson and May (1991) proposed a set of 3 partial differential equations to describe the flow of the individuals within the population with respect to time and host age:

\[
\begin{align*}
\frac{dX}{da} + \frac{dX}{dt} &= N\mu - (\lambda(a, t) + \mu)X(a, t), \\
\frac{dY}{da} + \frac{dY}{dt} &= \lambda X - (v + \alpha + \mu)Y(a, t), \\
\frac{dZ}{da} + \frac{dZ}{dt} &= vY - \mu Z(a, t).
\end{align*}
\]
2.3.1 Parameters in the Model

- $\mu$ is the natural rate of death (the life expectancy is equal to $1/\mu$). This is also the birth rate.
- $\lambda(a, t)$ is the force of infection for age $a$ at time $t$.
- $\nu$ is the recovery rate.
- $\alpha$ is the rate of death caused by the disease.
2.3.2 The Flow of Individuals (I)

- The change in the number of susceptibles, $X(a, t)$, is equal to the difference between the rate at which the individuals leave the susceptible class $(\lambda(a, t) + \mu) \times X(a, t)$ and the number of newborns who are entered into the susceptible class $(N \mu)$.

- The change in the number of infected individuals is the difference between the number of the newly infected $(\lambda(a, t)X(a, t))$ and the number of infected individuals developing immunity or dying $(\nu + \mu + \alpha)Y(a, t)$.

- The number of “new” immunes is equal to $vY(a, t)$ while the number of immune hosts that die and leave the population is $\mu Z(a, t)$. 
2.3.3 The Flow of Individuals (II)

Figure 2.9: The flow of individuals.
2.4 The Basic Reproductive Number

- The *basic reproductive rate*, $R_0$, is the expected number of new infections produced when one infected individual is entering into a completely susceptible population (Halloran 1998).

$$R_0 = \text{Number of contacts} \times \text{Transmission probability per contact} \times \text{Duration of infection}.$$  

- If $R_0 < 1$, it implies that the rate of generation of new cases is smaller than the rate of loss of existing cases, therefore the parasite cannot invade the population. **In this case the disease cannot maintain itself and will be eliminated.**
2.5 The Effective Reproductive Number

• The effective reproductive number, $R$, is assumed to be linearly dependent on the basic reproductive number:

$$R = R_0 x,$$

where $x$ is the proportion of susceptible hosts,

$$x = \frac{\text{susceptible hosts}}{\text{total number of hosts}}.$$

• $R$ is the expected number of new infections in practice, taking into account that only a proportion $x$ of the individuals are susceptible.
2.6 Equilibrium

- At equilibrium, the rate at which susceptible hosts are infected is the same as the rate at which new susceptible hosts appear, so $R = 1$.

- Since $R = R_0x = 1$, we derive

$$R_0 = \frac{1}{x^*} = \left(\frac{\bar{N}}{X}\right).$$
2.7 Natural Rate of Mortality

- The natural rate of mortality is the rate of death that does not depend on the disease. There are two types of natural mortality that usually used in the model.

- Type I mortality: The assumption is that everyone survives to age $L$ and then die. The survival function $\ell(a)$ is given by

$$\ell(a) = \begin{cases} 
1 & a \leq L, \\
0 & a > L.
\end{cases}$$

- Type II mortality: Here, the assumption is that the rate of death $\mu$ is constant over time and that the survival function is given by

$$\ell(a) = e^{-\mu a}.$$
2.7.1 Life Expectancy

- The life expectancy is \( L \) for type II mortality and
  \[
  \frac{1}{\mu}
  \]

- For type I mortality. Usually, we define
  \[
  \frac{1}{\mu} = L
  \]

- At each class the hosts can die (a natural death) at rate \( \mu \). Since hosts in the infected classes can die from the disease too, the survival function at the \( Y \) class is given by
  \[
  \ell(a) = e^{-(\mu + \alpha)a},
  \]
  where \( \alpha \) is called the disease induced death rate.
2.8 Mixing Pattern

• The above assumption about the random mixing of the population usually does not hold.

• Most populations do not mix in a random fashion but have subgroups that mix more with their own members.

• For example, different age groups within a school, contacts within households and sexual contacts within the population.
2.8.1 Contact Matrix (I)

• Halloran (1998) considered a population with two groups, A and B. The contact pattern between these two groups can be described by the following mixing matrix:

\[
C = \begin{pmatrix}
\beta_{aa} & \beta_{ab} \\
\beta_{ba} & \beta_{bb}
\end{pmatrix}
\]  

(2.1)

• The contact rate for individuals of group A with those of group B is denoted as \(\beta_{ab}\). Individuals can also contact with individuals from the same group: \(\beta_{aa}\) and \(\beta_{bb}\) denote the within-group contact rate for groups A and B, respectively.
2.8.2 Contact Matrix (II)

- Figure 5.1 illustrates the mixing pattern in this simple example.

\[
\begin{align*}
\beta_{aa} & \quad A \quad \beta_{ab} \\
& \quad \text{within group contact rate} \\
& \quad \beta_{ba} \quad B \quad \beta_{bb} \\
& \quad \beta_{ab} \quad \beta_{ba} \quad \text{between groups contact rate}
\end{align*}
\]

Figure 2.10: Mixing pattern between two hypothetical sub-populations. $\beta_{aa}$ is the contact rate within group A, $\beta_{bb}$ is the contact rate within group B. $\beta_{ab} = \beta_{ba}$ is the contact rate between the two groups.
2.8.3 Migration

![Diagram of Rural-Urban areas]

Figure 2.11: Rural-Urban areas.
2.8.4 Migration

Figure 2.12: Rural-Urban areas.
2.9 Example I: Hepatitis B

- Infection with Hepatitis B (HBV) occurs in many parts of the world. It is estimated that between 1 billion (Edmunds et al. 1996) to 2 billion (Van Damme et al. 1997, 1998) individuals who are alive have been infected with HBV.

- There are around 300–350 million chronic carriers of the virus all over the world. Van Damme et al. (1997, 1998) described the three major outcomes of HBV:
  - After infection the individuals may present symptomatic acute infection and develop lifelong immunity.
  - The infected individuals may become a chronic carrier.
  - The infected individuals may die within a few days from infection.
• Edmunds et al. (1996) discussed a mathematical model for the dynamics of HBV in developing countries. The model is an extension of the SIR model, in which the latent class (H) and the carriers class (C) are added into the model.

• A set of 5 partial differential equations was used by Edmunds et al. (1996) in order to describe the process mentioned above:

\[
\begin{align*}
\frac{dX}{da} + \frac{dX}{dt} &= -\left(\lambda(a, t) + \mu\right)X(x, t), \\
\frac{dH}{da} + \frac{dH}{dt} &= \lambda(a, t)X(a, t) - (\sigma + \mu)H(a, t), \\
\frac{dY}{da} + \frac{dY}{dt} &= \sigma H(a, t) - (\mu + \gamma_1)Y(a, t), \\
\frac{dC}{da} + \frac{dC}{dt} &= p\gamma_1 Y(a, t) - (\mu + \gamma_2 + w)C(a, t) \\
\frac{dZ}{da} + \frac{dZ}{dt} &= (1 - p)\gamma_1 Y(a, t) + \gamma_2 C(a, t) - \mu Z(a, t). 
\end{align*}
\]
2.9.1 Flow Diagram For Hepatitis B

![Flow Diagram for Hepatitis B](image)

Figure 2.13: The flow of individuals within the population for the HBV model described in Edmunds et al. (1996). At each state the individual can die at rate $\mu$. At the carrier state the rate of death is $\mu + w$. $1/\sigma$, $1/\gamma_1$ and $1/\gamma_2$ are the average durations in the latent, infectious and carriers states, respectively.
2.10 Example II: AIDS in an open homosexual community

• The first case of AIDS in UK was reported in 1981, 3 cases were reported in 1982 and during 1986 308 cases of AIDS were detected (Healy and Tillett, 1988).

• Baily (1988) presented results from few representative blood products samples of the homosexual community in San Francisco. In 1978, 4.5% of the blood products sample were infected by HIV, 12.6% in 1979, 24.1% in 1980 and 67.4% were HIV positive in the 1984 blood products sample.
The disease AIDS was identified in 1981 and the virus HIV in 1983-1984. For an individual infected with HIV, antibodies are detectable few weeks after the infection. The incubation period for AIDS (the time from infection to the disease symptoms) is much longer and it is measured in years rather than days.
2.10.1 Mathematical Model For AIDS/HIV

- Let $X(t)$ be the number of susceptible in an open homosexual population. The force of infection, $\lambda$, is the rate in which susceptible individuals are infected.

- One part of the population $p$ will be develop AIDS, i.e. that this part will become infected ($Y_1$) and then transfers to the AIDS class (A) in rate $v_1$.

- The other part of the susceptible class will be infected ($Y_2$) but will not develop the disease, this part is assume to move into a non-infectious class (Z) in rate $v_2$.

- Since there is no available cure for AIDS the immune class is not in the model.
2.10.2 Flow Diagram for AIDS

- Figure 2.14 shows the individuals flow mentioned above.

\[
\begin{align*}
X & \quad \overset{p\lambda}{\rightarrow} Y_2 \quad \overset{v_2}{\rightarrow} Z \\
\quad & \quad \overset{(1-p)\lambda}{\rightarrow} Y_1 \quad \overset{v_1}{\rightarrow} \ A \quad \text{Clinical AIDS}
\end{align*}
\]

Figure 2.14: *The flow of individuals for the HIV transmission.*
2.10.3 Differential Equations

- The set of differential equations that describe the process above is given by:

\[
\begin{align*}
\frac{dX}{dt} &= B - (\lambda + \mu)X, \\
\frac{dY_1}{dt} &= (1 - p)\lambda X - (\mu + v_1)Y_1, \\
\frac{dY_2}{dt} &= p\lambda X - (\mu + v_2)Y_2, \\
\frac{dA}{dt} &= v_1 Y - (\mu + \alpha)A, \\
\frac{dZ}{dt} &= v_2 Y_2 - \mu Z.
\end{align*}
\]
2.10.4 Solution for the Model

Figure 2.15: Numerical solution for the AIDS model for hypothetical population with $N_0 = 10000$, $Y_0 = 5$ and $\frac{1}{\mu} = 75$. Panel (a): $\frac{1}{v_1} = \frac{1}{v_2} = 8$, $p = 0.2$. Panel (b): same as (a) but with $p = 0.5$. Panel (c): same as (a) except $\frac{1}{v_2} = 20$. Panel (d): same as (c) except $p = 0.05$. Continuous line - Seropositives ($Y_1 + Y_2 + A + Z$, Dotted line - Cases $\times 10$, Long dashed line - Cumulative number of deaths ($\sum_{i=1}^{t} A(i) \times \alpha$).
Outbreak of AIDS in US: The Data

- Observed data.

Figure 2.16: Quarterly number of AIDA cases in US.
Chapter 3

The Static Model

• Only one time scale: age.
• Following one cohort of individuals from birth to death.
• Main assumption: there is no change over time.
3.1 The Static Model

- The static model is the steady state of the basic model, i.e., the state in which the variables $I$, $X$, $H$, $Y$ and $Z$ do not depend on the time but only on the host age.

- We assume that force of infection, $\lambda$, is constant and independent of time or age and that $\alpha = 0$.

- Differential equations:

\[
\frac{dX}{da} = N\mu - (\lambda(a) + \mu)X(a), \quad \text{susceptible},
\]

\[
\frac{dY}{da} = \lambda X - (v + \alpha + \mu) Y(a), \quad \text{infected},
\]

\[
\frac{dZ}{da} = vY - \mu Z(a), \quad \text{Immune}.
\]

- $\lambda(a)$: age-dependent force of infection.
3.1.1 The Number of Host at Age $a$

- The number of hosts at age $a$ is given by
  \[
  N(a) = N(0)\ell(a) = N(0)P(\text{survive at age } a).
  \]

  For type I mortality we get that
  \[
  N(a) = \begin{cases} 
  N(0) & a \leq L, \\
  0 & a > L.
  \end{cases}
  \]

- For type II mortality the number of hosts at age $a$ is given by
  \[
  N(a) = N(0)e^{-\mu a}.
  \]
3.1.2 Susceptible Class

- Let $T_X$ be the lifetime at the susceptible class.

- We assume that $T_X$ is an outcome of exponential distribution, $T_X \sim \exp(\lambda)$, and the probability to move from the susceptible to the infected class is given by

$$P(X \rightarrow Y \text{ before age } a) = F(T_X) = 1 - e^{-\lambda a},$$

- It follows that the probability to stay in $X$ is $e^{-\lambda a}$.

- The expected number of host in the susceptible class at age $a$, $X(a)$, is given by

$$X(a) = N(0)\ell(a)e^{-\lambda a}.$$
Number of Susceptible

• For type I mortality we get

\[ X(a) = \begin{cases} 
N(0)e^{-\lambda a} & a \leq L, \\
0 & a > L.
\end{cases} \]

• The change in the susceptible class:

\[ \frac{dX(a)}{da} = -\lambda N(0)e^{-(\lambda+\mu)a} = -\lambda X(a). \]

• For type II

\[ X(a) = N(0)e^{-(\lambda+\mu)a}. \]

The derivative with respect to the age is

\[ \frac{dX(a)}{da} = -(\lambda + \mu)N(0)e^{-(\lambda+\mu)a} = -(\lambda + \mu)X(a), \]
Proportion of Susceptibles

- the differential equation above describes the change in the susceptible class as the age change: $\lambda X$ hosts are transferred to the infected class and $\mu X$ die.

- Instead of the total number of susceptibles we can use the proportion of susceptible hosts at age $a$:

  $$x(a) = \frac{X(a)}{N(a)} = \frac{N(0)\ell(a)e^{-\lambda a}}{N(0)\ell(a)} = e^{-\lambda a}.$$ 

  Note that we eliminate the natural rate of death, $\mu$, when we use the proportion of susceptible hosts.

- The change in the susceptible class:

  $$\frac{dx(a)}{da} = -\lambda x(a).$$
3.1.3 Infected Class

- The change in the infected class can be described in the following way:
  - New hosts are entered from the susceptible class (at rate $\lambda X(a)$) and at the same time hosts are transferred to the immune class or die at rate $(v - \mu)Y(a)$. The corresponding differential equation is given by
    \[
    \frac{dY(a)}{da} = \lambda X(a) - (v + \mu)Y(a).
    \]
  - Integrating overall ages we get that the total number of hosts in the infected class is given by
    \[
    Y(a) = \frac{\lambda}{\lambda - v} N(0) \ell(a) \left[ e^{-va} - e^{-\lambda a} \right].
    \]
  - The proportion of infected host by
    \[
    y(a) = \frac{\lambda}{\lambda - v} \left[ e^{-va} - e^{-\lambda a} \right].
    \]
3.1.4 Immune Class

- Hosts from the infected class enter into the immune class in the rate $vY(a)$ and die at rate $\mu Z(a)$. The differential equation for the immune class is given by

$$\frac{dZ(a)}{da} = vY(a) - \mu Z(a).$$

- The total number of host in the immune class can be calculated by

$$Z(a) = N(a) - X(a) - Y(a).$$

- The proportion of immune

$$z(a) = 1 - x(a) - y(a).$$
3.2 Example for the SIR model

- In this example, and for all the following examples, we present the results for the steady-state model. This means that the parameters in the model do not change over the time.

![Graphs of the SIR model](image)

*Figure 3.1: The SIR model. Left panels: $x(a)$ and $z(a)$. Right panels: $y(a)$. Upper panels: the model with average duration of 10 years in the susceptible class and 10 days in the infected class. Lower panels: the model with average duration of 10 year and 10 days for the susceptible and the infected classes respectively.*
3.3 Model with Maternal Antibodies Period

Figure 3.2: The SIR model with maternal antibody period.
3.3.1 The Maternal Antibodies Class

- The proportion of hosts protected by maternal antibodies at age $a$ is given by
  
  $$i(a) = e^{-da},$$

  so the average duration of the maternal antibodies period is
  
  $$\frac{1}{d}$$

  (which is typically between 3 to 9 months).
3.4 Model with Maternal Antibodies and Latent Periods

Figure 3.3: The SIR model with maternal antibody and latent periods.
3.4.1 Duration

- We assume that susceptibles transfer to the latent period (H) at rate $\lambda$ and become infected at rate $\sigma$. The table below presents the different classes with their average duration (the values of the parameters correspond to the Rubella in developing countries, these values were taken from Table 3.1 in Anderson and May).

<table>
<thead>
<tr>
<th>Class</th>
<th>Parameter</th>
<th>Expected duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$d=4$</td>
<td>0.25 years</td>
</tr>
<tr>
<td>X</td>
<td>$\lambda=0.4$</td>
<td>2.5 years</td>
</tr>
<tr>
<td>H</td>
<td>$\sigma=34.76$</td>
<td>11.5 days</td>
</tr>
<tr>
<td>Y</td>
<td>$v=31.74$</td>
<td>10.5 days</td>
</tr>
</tbody>
</table>
### 3.4.2 Solution for the Model

- Note that for this example the average duration at the latent period (10.5 days) and at the infected class (11.5 days).

- the proportion of immune hosts is approximately

\[
z(a) \approx 1 - i(a) - x(a) .
\]

![Figure 3.4: Patterns for the five classes in the model. The average duration is 3 months in the maternal antibodies class, 2.5 years in the susceptible class, 11.5 days in the latent class and 10.5 days in the infected class. Note that h(a) and y(a) are relatively small, therefore i(a) + z(a) ≈ i(a) + (1 - i(a) - x(a)) = 1 - x(a). This can be seen from figure (b) and (f).](image)
3.4.3 $\lambda = 0.1$

- The table below presents the parameters for the model (the only change is in $\lambda$, the average duration in the susceptible class in 10 years).

<table>
<thead>
<tr>
<th>Class</th>
<th>Parameter</th>
<th>Expected duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$d=4$</td>
<td>0.25 years</td>
</tr>
<tr>
<td>X</td>
<td>$\lambda=0.1$</td>
<td>10 years</td>
</tr>
<tr>
<td>H</td>
<td>$\sigma=34.76$</td>
<td>11.5 days</td>
</tr>
<tr>
<td>Y</td>
<td>$v=31.74$</td>
<td>10.5 days</td>
</tr>
</tbody>
</table>
3.4.4 Solution for the Model

- Since the duration in the susceptible class increases (as $\lambda$ decreases) the proportion of immune hosts $i(a)$ converge in slow rate to 1.

![Graphs showing the proportion of hosts with antibodies, susceptibles, host in the latent class, infected, and immune over age]

Figure 3.5: The model with average duration of 10 years in the susceptible class.
3.4.5 Proportion of Susceptible as Function of $\lambda$.

Figure 3.6: Proportion of susceptible hosts $x(a)$ for different values of $\lambda$, average duration in the maternal antibodies class - 3 months, 10.5 days in the latent class and 11.5 days in the infected class.
3.4.6 Proportion of Immune as Function of $\lambda$

Figure 3.7: Proportion of immune hosts $z(a)$ for different values of $\lambda$, average duration in the maternal antibodies class - 3 months, 10.5 days in the latent class and 11.5 days in the infected class.
Chapter 4

Immunization in The Static Model
4.1 Immunization Against Measles

- One picture is better than 1000 words.

Figure 4.2: Immunization program against measles and its impact on the global incidence
4.2 Immunization in the SIR Model

• The direct effect is the following:
  suppose that one part of the host population, \( p \) say, vaccinates at birth and the other part, \( (1 - p) \), does not. This means that \( (1 - p) \) from the hosts are entered to the susceptible class while the part that was vaccinated, \( p \), is entered directly into the immune class.

• In other words, out of \( N(0) \) hosts at age 0, 
  \( (1 - p)N(0) \) are susceptible and \( pN(0) \) are immune.
4.2.1 The Time Window For immunization

Figure 4.3: Immunization window
4.2.2 Indirect Effect of Immunization

- The indirect effect is the following: immunization programs reduce the number of infected hosts and therefore cause a reduction in the force of infection $\lambda$.

- Recall that the basic reproductive rate at equilibrium is given by

$$R_0 = \frac{1}{x^*} = \left( \frac{\bar{N}}{\bar{X}} \right)_e = \frac{\lambda + \mu}{\mu}.$$  

The last equality follows from the definition of $\bar{X}$ and $\bar{N}$. Now, if $p$ of the hosts are immunized then at equilibrium $x^*$ is at most $1 - p$. 

4.2.3 Critical Proportion of Vaccination

- As we mentioned, for $R = R_0 x < 1$ the infection cannot spread, so the critical proportion of host that has to be vaccinated in order to eliminate the disease is the one with $R_0 (1 - p_c) < 1$ or

$$p_c = 1 - \frac{1}{R_0}.$$
4.2.4 The New Equilibrium

• For such an immunization program ($p$ vaccinated at birth) the basic reproductive rate is

\[ R_0 = \frac{\lambda + \mu}{(1 - p)\mu}. \]

• The new force of infection is

\[ \lambda' = \mu(1 - R_0 - p) = \mu(p_c - p). \]

This means that the force of infection linearly depends on the proportion of vaccinated host.
4.2.5 $\lambda' \to 0$

- As $p \to p_c$ then $\lambda' \to 0$ and the disease will be eliminated. Note that for $\lambda' \to 0$ the proportion of susceptible hosts

$$x(a) \to (1 - p)e^{-(\lambda'a)} \to (1 - p).$$

- These host are not vaccinated but will not transfer to the infected class since the force of infection is converging to 0 (the other part of the host population will be transferred directly to the immune class).
4.2.6 Equilibrium Before the Vaccination

- The table below shows the parameters for the SIR model before the immunization.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Expected duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class X</td>
<td>$\lambda = \frac{1}{9}$</td>
</tr>
<tr>
<td>Class Y</td>
<td>$\nu = 31.74$</td>
</tr>
<tr>
<td>rate of death</td>
<td>$\mu = \frac{1}{75}$</td>
</tr>
</tbody>
</table>

- The expected duration the susceptible class (before immunization) is 9 years. Note that in this case the basic reproductive rate is

$$R_0 = \frac{\lambda + \mu}{\mu} = 9.333,$$
4.2.7 Proportion os Susceptible

• and the critical proportion is

\[ p_c = 1 - \frac{1}{R_0} = 0.89. \]

Figure 8 shows the proration of susceptible hosts as a function of \( p \) and the host age.
• When $p = 0$, the probability to be in the susceptible class decreases exponentially with rate

$$\lambda = \frac{1}{9}.$$ 

• For this value of $p$ the curve looks as the curve in Figure 1. As $p$ increases (and $\lambda$ decreases) the probability profiles become more and more flat, indicating that it is “more and more difficult” to transfer from the susceptible to the infected class, the probability to transfer from $X$ to $Y$,

$$1 - e^{-(\lambda'a)}.$$
4.3 The Average Age at Infection

- As mentioned before, for a constant force of infection and type II mortality,

\[ R_0 = \frac{L}{A}, \]

- For a given value of \( L \), as the average age of infection increases \( R_0 \) decreases. As \( A \to L \), \( R_0 \to 1 \) and for \( A > L \) \( R_0 \) is below 1.

- Anderson and May (1991) showed that the new average age at infection is

\[ A' = \frac{A}{1 - p}. \]

- Figure 4.5 shows the average age at infection, for several values of \( R_0 \), as a function of \( p \). Note that as \( p \to p_c \) the average age at infection converge to the life expectancy and therefore the disease is eliminated.
Figure 4.5: Immunization program. The change of the average age at infection for the cases with $R_0 = 5$ ($p_c = 0.8$), $10$ ($p_c = 0.9$) and $15$ ($p_c = 1 - 1/15$). In all cases $L=75$. 
Chapter 5

HIV/AIDS and other STIs

5.1 Overview

- Unprecedented epidemic of HIV and other STIs in the world.
- Disproportionately high in sub-Saharan Africa (Gouws and Williams, 2000).
- In developing countries, HIV and STIs among leading causes of substantial morbidity (Gerbase, et al 1998).
- Serious threat to public health.
5.2 Adults and Children Estimated To Be Living With HIV/AIDS, 2003

Total: 34 – 46 million
5.3 Estimated Number of Adults and Children Newly Infected With HIV, 2003

Total: 4.2 – 5.8 million
5.4 Estimated Adult and Child Deaths from HIV/AIDS, 2003

Total: 2.5 – 3.5 million
5.5 Spread of HIV/STIs


- Predominant factor is rural-urban and cross-border labour migration (Jochelson, et al 1991).
5.5.1 Contact Matrix

- Figure 5.1 illustrates the mixing pattern in this simple example.

\[ \beta_{aa} \quad \beta_{ab} \quad \beta_{bb} \]

\( \beta_{aa} \) is the contact rate within group A,\( \beta_{bb} \) is the contact rate within group B. \( \beta_{ab} = \beta_{ba} \) is the contact rate between the two groups.

Figure 5.1: Mixing pattern between two hypothetical sub-populations. \( \beta_{aa} \) is the contact rate within group A, \( \beta_{bb} \) is the contact rate within group B. \( \beta_{ab} = \beta_{ba} \) is the contact rate between the two groups.
5.6 Epidemic of STIs

- Africa and other developing countries bear heavy burden of STIs.
- In 1995, 19.7% of these cases in sub-Saharan Africa.
5.7 Impact of STIs

- STIs cause acute symptoms such as genital ulcers and discharge.
- Most common cause of infertility.
5.8 Evolution of HIV

- AIDS first recognized as a new disease early 80s.
- Existed at least since late 70s.
- First recognized among homo/bisexual men in US.
- Two years later, HIV virus identified as the cause of AIDS.
5.9 Burden of HIV

- HIV/AIDS affect all countries of the globe, making it the emerging infection of the late 20th century (Nicoll and Gill, 1999).
- First 2 cases of AIDS reported in S.A. in 1983.
- HIV spread relentlessly from urban to rural areas in the mid 90s.
- Macro-simulation models predicted prevalence of 30% between 2000 to 2005 (Schall, 1990).
5.10 Relationship Between HIV and STIs

- Share same risk factors.
- Presence of STI could be a marker of high risk of HIV (Mertens, et al 1990).
- STIs promote transmission of HIV, high shedding of HIV virus into the genital fluids in the presence of genital ulcers.
5.11 Human Migration

- Extraordinary high rates of migration between and within countries (Crush, 1995).
- Roots of migration since 1880s, discovery of gold.
- Circular labour migration is the predominant type.
- Apartheid system of governance.
5.12 Migration and Spread of HIV/STIs

- HIV follow the movement of people (Quinn, 1994; Decosas, et al 1995; Decosas and Adrien, 1997; Mabey and Mayaud, 1997).

5.13 Studies of Migration and HIV/STIs

- The role of migration in the spread of HIV has been described primarily as a result of men becoming infected while away and infecting their partners when they return. Migrant men have contact with e.g. sex workers during migration.

- In extreme cases migrant men establish parallel families in urban areas and rural areas, forming sexual networks.

- Women also play a role (see case studies).
5.13.1 Sexual Networks

- Integral part in understanding the epidemiology of HIV/STIs within human populations (intervention programs).
- Sexual networks derived from partner tracing.
- Studies in infectious diseases and sexual networks demonstrate higher likelihood of an infection within core group sexual networks.
- Partner-reporting fail to provide sufficient information necessary to estimate the degree of heterogeneity between sexual networks.
5.14 South African Migration Project

5.14.1 Designed to

- Investigate the link between circular migration and spread of HIV.
- Test hypothesis that migrant men and their rural partners are at high risk of HIV infection.
- Measure the extent to which rural epidemic of HIV is fuelled by circulation within rural areas as opposed to introduction from outside home community of returning migrants.
5.14.2 Migration Study: Design

- Long and short distance migrant men who come from Hlabisa (a rural area in KwaZulu-Natal province of South Africa).
- Rural female partners of both long and short distance migrant men based in Hlabisa.
- Non-migrant men and their partners based in Hlabisa.
5.14.3 **Migration Study: Sampling Method**

- Migrant men in Carletonville.
- Migrant men in Richard’s Bay.
- Non-migrant men in Hlabisa.
- Formation of three cohorts.
  - *Long distance migrants* $\leftrightarrow$ *rural female partners*.
  - *Short distance migrants* $\leftrightarrow$ *rural female partners*.
  - *Non-migrants* $\leftrightarrow$ *rural female partners*.
- Combine short & long distance migrants.
CHAPTER 5. HIV/AIDS AND OTHER STIS

5.14.4 Study Map
5.14.5 Migration Study

- Collected information regarding
  - socio-demographic characteristics.
  - sexual behavioural factors.
  - Biomedical factors.
- Blood and urine for tests of HIV and other STIs.
- Four-monthly visits and same procedure.
- Total of seven visits.
5.14.6 Available Data

- Migration data consists of 631 men and women aged between 18 and 69 years who were interviewed during the first clinical visit.
- 49% of women were partners of migrant men and 51% were partners of non-migrant men.
- 0 to 4 female partners interviewed for each man.
- Composition of sexual network partnerships consisted of 187 dyads, 40 triads, 4 quadriads and 1 pentad.
5.14.7 Remark

- Provided valuable information related to the effects of migration on the spread of HIV
- Enabled deeper understanding of the consequences of migration to the spread of HIV/STIs
- Discussed in the next lecture
- Similar study among migrant women has been proposed.
Chapter 6

Summary

- Time to infection.
- Follow up studies and cross-sectional studies.
- Time to event and current status data.
Childhood infections: age is time scale, and $t_0=$ birth

STI; $t_0=$ the age of first sexual activity
- **Follow up studies:**
  - Follow a cohort of individuals for a given time intervals \((t_0, t_n)\).
  - Observations: \(t_1\) or \(t_0\) if the individual is not infected before the end of the study.
  - Analysis: Survival analysis.

- **Current status data:**
  - Serological test for a group of individuals.
  - Observations: infected- YES or NO.
  - Analysis: Models for binary data.

- **Hazard rate:** the rate in which individuals become infected.

- **The force of infection:** the rate in which individuals become infected.

The parameter of primary interest is the same!
Infected between $t_1$ and $t_2$. 
Right censored
CURRENT STATUS DATA

séro-positive

séro-negative

\( a_1 \quad a_2 \)

Age at test

\( (a_1, \text{Positive}) \)

\( (a_2, \text{Negative}) \)

We do not know the age at infection.
Part II

Estimation for Serological Data
Chapter 7

Estimation from Current-Status Data

- Survival analysis and current status data.
- Parametric models in survival analysis and their corresponding models for current status data.
- Parametric models for current status data and their fractional polynomial formulation.
7.1 Time Homogeneous Model for Infectious Diseases

- The change in the susceptible class can be described by:

\[
\frac{\partial}{\partial a} q(a, t) + \frac{\partial}{\partial t} q(a, t) = -\ell(a, t) [q(a, t)],
\]

where \( q(a, t) \) is the probability to be susceptible at age \( a \) and time \( t \).

- if we assume time homogeneity, that is \( \frac{\partial}{\partial t} q(a, t) = 0 \) then

\[
\frac{d}{da} q(a) = -\ell(a) [q(a)].
\]

- \( \ell(a) \) is the rate in which susceptible become infected and called the hazard or the force of infection.
Example

- Assume that the time spent at the susceptible class is exponential with rate $\beta$
- $\Rightarrow \frac{d}{da}q(a) = -\beta e^{-\beta a} = -\beta q(a)$.
- The force of infection in this case, $\beta$, is constant.
- This is the so-called catalytic model.
First Connection to Statistics

- Susceptible class and immune class.

Figure 7.1: A follow up study.
7.2 Follow-up Studies

- In a follow-up study we follow individuals of one cohort from birth to infection.
- In this case, the epidemiological parameters can be estimated directly, using survival analysis methods.

Figure 7.2: A follow up study.
7.3 Example: Exponential Distribution

- \( Y_i \sim \exp(2) \).
- A random sample of 1000 observations.

Figure 7.3: 1000 observations.
7.4 Current-Status Data

7.4.1 Cross-sectional Sample

- In practice, data are not available from follow-up studies but from cross-sectional samples.

Figure 7.4: Cross-sectional sample.
7.4.2 The Data

• The information about each individual in the sample:
  – Age at test.
  – Infection status (yes or no).

Figure 7.5: Current-status data.
7.4.3 Right censoring

- Right censored data: Age at infection is higher than the age at test.

Figure 7.6: Right censoring.
7.4.4 Left censoring

- Left censored data: Age at infection is smaller than the age at test.

Figure 7.7: Left censoring.
7.4.5 Serological Datasets

- 5 serological datasets.

Figure 7.8: *Five cross-sectional seroprevalence datasets.*
7.4.6 Modeling Current-Status Data

- Consider a prevalence sample of size $N$ and let $a_i$ be the age of the $i$th subject.
- Instead of observing the age of infection we observe a binary variable $Y_i$ such that
\[
Y_i = \begin{cases} 
1 & \text{subject } i \text{ experienced infection before age } a_i \text{ (left-censored)} \\
0 & \text{elsewhere (right-censored)}
\end{cases}
\]
- The prevalence: $\pi(a_i) = 1 - q(a_i)$.
- The likelihood:
\[
L(\beta) = \sum_{i=1}^{N} \{Y_i \log(\pi(a_i)) + (1 - Y_i) \log(1 - \pi(a_i))\}.
\]
7.5 Parametric Models in Survival Analysis

7.5.1 The Survival Function

- In survival analysis the survival function $S(x)$ is the probability of an individual surviving beyond time $x$, i.e. the individual experience an event after time $x$.

- In context of infection disease data, the event is the infection, and the survival function is the probability that an individual was infected beyond age $a$. 
7.5.2 The Hazard Function

- An informal definition of the hazard rate in survival analysis: "the rate in which individuals experience the event".

\[
\text{hazard} = \frac{f(x)}{S(x)} = \frac{F'(x)}{1 - F(x)}.
\]

where \( f(x) = F(x)' = (1 - S(x))' \).

- In context of infection disease data, the hazard function is the force of infection, the rate in which individuals become infected.
7.6 Parametric Models for Current Status Data

7.6.1 The Prevalence

- The probability to be susceptible at age $a$, $q(a)$ is the survival function.
- The prevalence, $\pi(a) = 1 - q(a)$.

The Force of infection

$$\ell(a) = \frac{\pi'(a)}{1 - \pi(a)}$$
7.6.2 Force of Infection

- For models with log link function, 
  \[- \log[1 - \pi(a)] = \eta(a)\] we have

  \[\ell(a) = \frac{\pi'(a)}{1 - \pi(a)} = \frac{\eta'(a)e^{-\eta(a)}}{e^{-\eta(a)}} = \eta'(a)\]

- In the general case for binary response the force of infection has the form of

  \[\ell(a) = \eta'(a)\delta[\eta(a)]\]

  where \(\delta[\eta(a)]\) is determined by the link function.

- E.g., with logit link function we have

  \[\frac{\pi'(a)}{1 - \pi(a)} = \eta'(a, \beta)\frac{e^{\eta(a)}}{1 + e^{\eta(a)}} = \eta'(a)\pi(a)\]

  \[\Rightarrow \delta[\eta(a)] = \frac{e^{\eta(a)}}{1 + e^{\eta(a)}} = \pi(a)\]
7.7 Three Examples

7.7.1 Exponential Distribution

Survival Analysis

- Let us assume that the time spent in the susceptible class is exponential. Then, the survival function is

\[ S(a) = e^{-\lambda a}. \]

- The hazard (or the force of infection):

\[ \frac{f(a)}{S(a)} = \frac{\lambda e^{\lambda a}}{e^{\lambda a}} = \lambda. \]
Current-Status Data

- The likelihood:
  \[ L(\beta) = \sum_{i=1}^{N} \{Y_i \log(\pi(a_i)) + (1 - Y_i) \log(1 - \pi(a_i))\} . \]

- The prevalence
  \[ \pi(a) = 1 - e^{-\lambda a} . \]

- The force of infection: \( \ell(a) = \lambda \). A constant force of infection which does not depend on age.

- GLM with log link
  \[ \log(1 - \pi(a)) = -\lambda a . \]

- GLM with complementary log-log link function
  \[ g(\pi(a)) = \log(-\log(1 - \pi(a))) = \log(\lambda) + \log(a) . \]
7.7.2 Weibull Distribution

Survival Analysis

- Let us assume that the time spent in the susceptible class is a Weibull random variable. Then, the survival function is

\[ S(a) = e^{-\lambda a^\alpha}. \]

- The hazard (or the force of infection)

\[ \frac{f(a)}{S(a)} = \frac{\alpha \lambda a^{\alpha-1} e^{-\lambda a^\alpha}}{e^{-\lambda a^\alpha}} = \lambda \alpha a^{\alpha-1}. \]
Current-Status Data

- The prevalence
  \[ \pi(a) = 1 - e^{-\lambda a^\alpha}. \]

- GLM with complementary log-log link function
  \[ g(\pi(a)) = \log(-\log(1 - \pi(a))) = \log(\lambda) + \alpha \log(a). \]

- Age dependent force of infection
  \[ \ell(a) = \lambda \alpha a^{\alpha - 1}, \]
  which is monotone (increasing or decreasing) force of infection, depends on the sign of \( \alpha \).

- Note that for \( \alpha = 1 \) the Weibull model is an exponential model.
7.7.3 Log-logistics Distribution

Survival Analysis

- Let us assume that the time spent in the susceptible class follows a log-logistic distribution. Then, the survival function is
  \[ S(a) = \frac{1}{1 + \lambda a^\alpha}. \]

- The hazard (or the force of infection)
  \[ f(a) = S(a) = \frac{\alpha a^{\alpha-1} \lambda}{1 + \lambda a^\alpha}. \]
Current-Status Data

- The prevalence
  \[ \pi(a) = \frac{1}{1 + \lambda a^\alpha}. \]

  \[ 1 - \pi(a) = 1 - \frac{1}{1 + \lambda a^\alpha} = \frac{\lambda a^\alpha}{1 + \lambda a^\alpha}. \]

- GLM with logit link function
  \[ \logit \left( \frac{\pi(a)}{1 - \pi(a)} \right) = \log(\lambda a^\alpha) = \log(\lambda) + \alpha \log(a). \]

- Age dependent force of infection
  \[ \ell(a) = \frac{\alpha a^{\alpha-1} \lambda}{1 + \lambda a^\alpha}. \]

  which is a single peak force of infection.
Chapter 8

Modeling the Force of Infection Using Fractional Polynomials

8.1 Fractional Polynomials

- Grenfell and Anderson (1985) assumed a $m$ order polynomial for the linear predictor.
- Prevalence: $\pi(a) = 1 - e^{-\eta(a)}$,
  $$\eta(a) = \beta_0 + \beta_1 a + \beta_2 a^2 + \ldots \beta_p a^m.$$  
- Force of infection: $\eta'(a)$. 
Application to the Data

- Models with logit and c-log-log link functions.

Figure 8.1: Hepatitis A in Belgium. Left panel: data and estimated models for the prevalence. Right panel: estimated forces of infection. Solid line: model with logit link function. Dashed line: model with complementary log-log link function.
8.1.1 First Order Fractional Polynomials

- Linear predictor: \( g(\pi(a)) = \eta_1(a) \). With
  \[
  \eta_1(a, \beta, p) = \beta_0 + \beta_1 H(a).
  \]

- \( H(a) \) is a transformation given by
  \[
  H(a) = \begin{cases} 
  a^{pi} & \text{if } p \neq 0 \\
  H_{i-1}(a) \times \log(a) & \text{if } p = 0
  \end{cases}
  \]

- \( \eta(a) \):
  \[
  \eta(a) = \begin{cases} 
  \beta_0 + \beta_1 a^p & \text{if } p \neq 0 \\
  \beta_0 + \beta_1 \log(a) & \text{if } p = 0
  \end{cases}
  \]
8.1.2 Second Order Fractional Polynomials

- **Linear predictor:**
  \[
  \eta_2(a, \beta, p_1, p_2) = \sum_{i=0}^{2} \beta_i H_i(a),
  \]
  where \( p_1 \leq p_2 \) is a sequence of powers and \( H_i(a) \), \( i = 0, 1, 2 \) is a transformation function given by
  \[
  H_i(a) = \begin{cases} 
  a^{p_i} & \text{if } p_i \neq p_{i-1} \\
  H_{i-1}(a) \times \log(a) & \text{if } p_i = p_{i-1}
  \end{cases}
  \]
  Here, \( H_0(a) = 1 \).

- **Linear predictor:**
  \[
  \eta(a) = \begin{cases} 
  \beta_0 + \beta_1 a^{p_1} + \beta_2 a^{p_2} & \text{if } p_2 \neq p_1 \\
  \beta_0 + \beta_1 a^{p_1} + a^{p_1} \log(a) & \text{if } p_2 = p_1
  \end{cases}
  \]
8.1.3 The Powers

- Royston and Altman (1994) argued that, in practice, fractional polynomials of order higher than 2 are rarely needed.

- WE choose the value of the powers from the set \{-2, -1, -0.5, 0, 0.5, 1, 2, \max(3, m)\}.

- The models considered by Grenfell and Anderson (1985) have the general form of \( \eta_m(a, \beta, p_1, p_2, \ldots, p_m) \) with \( p_i = i \) for \( i = 1, 2, \ldots, m \).
8.2 Model selection

• Choose the best fractional polynomial of the first order and the best fractional polynomial of the second order.

• Reject the first order model if

\[ D(1, \hat{p}) - D(2, \hat{p}) > \chi^2_{2,0.9} \]

where \( \hat{p} \) is the power sequence for the model that has the best goodness-of-fit.
8.3 Link function (1): logit

- We consider models of the general form

\[
\pi(a) = \frac{\exp(\eta_m)}{1 + \exp(\eta_m)}
\]

where \( \eta_m \) is the fractional polynomial. For a first order fractional polynomials we specify

\[
\pi(a) = \frac{\exp(\beta_0 + \beta_1 H(a))}{1 + \exp(\beta_0 + \beta_1 H(a))}
\]

- Second Order fractional polynomial

\[
\pi(a) = \frac{\exp(\beta_0 + \beta_1 H_1(a) + \beta_2 H_2(a))}{1 + \exp(\beta_0 + \beta_1 H_1(a) + \beta_2 H_2(a))}
\]
8.4 Link function (2) : c-log-log

- We consider models of the general form
  \[ \pi(a) = 1 - \exp(-\gamma(a)). \]

- More precisely, for the first order fractional polynomials we specify
  \[ \pi(a) = \begin{cases} 
  1 - \exp(-\beta_0 e^{\beta_1 H_1(a)}) & p \neq 0, \\
  1 - \exp(-\beta_0 a^{\beta_1}) & p = 0. 
\end{cases} \]

- For the second order fractional polynomials, we consider the following specification
  \[ \pi(a) = 1 - \exp(-\beta_0 e^{\beta_1 H_1(a) + \beta_2 H_2(a)}). \]
8.5 Constrained Fractional Polynomials

- Although fractional polynomials provide a wide range of curve shapes, there is no guarantee that $\pi(a)$ will be a monotone function of age and therefore fractional polynomials can still result in a negative estimate for the force of infection.

- The force of infection is negative whenever $\eta'_m(a, \hat{\beta}, p) < 0$ (since $\delta(\eta_m(a, \hat{\beta}, p))$ is strictly positive).

- We can fit a large number of fractional polynomials, over a grid of powers, and check for each fitted model if $\eta'_m(a, \hat{\beta}, p) \geq 0$, for all ages $a$.

- In case that a given sequence of powers leads to a negative derivative of the linear predictor, the model is not considered an appropriate model.
8.6 First Order Fractional Polynomials

- Recall that a first order fractional polynomials has the form

\[ \eta(a) = \beta_0 + \beta_1 a^p. \]

8.6.1 Exponential Model

- The GLM with c-log-log link

\[ g(\pi(a)) = \log(\lambda) + \log(a) \]

is a first degree fractional polynomial model with

\( p = 0, \beta_0 = \log(\lambda) \) and \( \beta_1 = 1. \)
8.6.2 Weibull Model

- The GLM with c-log-log link

\[ g(\pi(a)) = \log(\lambda) + \alpha \log(a) \]

is a first degree fractional polynomial model with \( p = 0 \), \( \beta_0 = \log(\lambda) \) and \( \beta_1 = \alpha \neq 1 \).

8.6.3 Log-logistic Model

- The GLM with logit link function

\[ g(\pi(a)) = \log(\lambda) + \alpha \log(a) \]

is a first order fractional polynomial with \( p = 0 \), \( \beta_0 = \log(\lambda) \) and \( \beta_1 = \alpha \).
8.6.4 Other Examples

- Munch, Griffiths, Grenfell and Anderson and many others.....all used fractional polynomials...but they did not know that !!

<table>
<thead>
<tr>
<th>Publication</th>
<th>Force of infection</th>
<th>Fractional polynomial</th>
<th>Link function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Munch (1959), Farrington (2001), Jewell and Van der laan (1995)</td>
<td>constant</td>
<td>$\eta(m = 1, p = 0, \beta = 1)$</td>
<td>cloglog</td>
</tr>
<tr>
<td>Griffiths (1974)</td>
<td>linear</td>
<td>$\eta(m = 1, p = 0, \beta = 2)$</td>
<td>cloglog</td>
</tr>
<tr>
<td>Grenfell and Anderson (1985)</td>
<td>polynomial</td>
<td>$\eta(m = k, p_i = i)$</td>
<td>log</td>
</tr>
<tr>
<td>Keiding (1996), Becker (1989), Diamond and McDonald (1992), Grummer-Strawn (1993)</td>
<td>monotote</td>
<td>$\eta(m = 1, p = 0, \beta)$</td>
<td>cloglog</td>
</tr>
<tr>
<td>Grummer-Strawn (1993)</td>
<td>flexible</td>
<td>$\eta(m = 1, p = 0, \beta)$</td>
<td>logit</td>
</tr>
</tbody>
</table>
8.7 Application to the data (logit link)

- First and second order fractional polynomials.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>df</th>
<th>Deviance</th>
<th>( p )</th>
<th>( G(1, p) )</th>
<th>df</th>
<th>Deviance</th>
<th>( p_1, p_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hepatitis A (Be)</td>
<td>83</td>
<td>115.34</td>
<td>0.32</td>
<td>34.21</td>
<td>81</td>
<td>97.61</td>
<td>1.0, 1.3</td>
</tr>
<tr>
<td>Hepatitis A (Bul)</td>
<td>80</td>
<td>79.51</td>
<td>1</td>
<td>0</td>
<td>78</td>
<td>77.77</td>
<td>1.9, 1.9</td>
</tr>
<tr>
<td>Varicella</td>
<td>41</td>
<td>50.94</td>
<td>0.07</td>
<td>69.59</td>
<td>39</td>
<td>43.90</td>
<td>-0.7, -0.6</td>
</tr>
<tr>
<td>Rubella</td>
<td>41</td>
<td>56.28</td>
<td>0.03</td>
<td>165.13</td>
<td>39</td>
<td>42.34</td>
<td>-0.9, -0.4</td>
</tr>
<tr>
<td>Mumps</td>
<td>41</td>
<td>82.31</td>
<td>-0.2</td>
<td>516.88</td>
<td>39</td>
<td>47.94</td>
<td>-1.2, -0.9</td>
</tr>
</tbody>
</table>
8.8 Hepatitis A

- Fractional polynomials for Hepatitis A (first and second order with logit link function).
8.9 Rubella, Mumps and Varicella

- Second order fractional polynomials for rubella, mumps and varicella (logit link).
8.10 Link function

- Second order model with logit and c-log-log link functions.
8.11 Hepatitis A in Bulgaria

- The best model is the first order logit model.
8.12 Adding covariates to the model

- A model with simple linear effect for a categorical covariate
  \[ \eta(a) = \phi(a) + Z \beta \]
  where \( \pi(a) = g^{-1}(\eta(a)) \) and \( Z \) is categorical covariate (such as gender).

- We model \( \phi(a) \) with fractional polynomial.
Part III

Interval Censored Data: Application to AIDS/HIV
Chapter 9

Interval Censored Data

9.1 Introduction

- Interval-censored data arise where the exact event time is not observed directly.
- Often due to the periodic assessment.
- For example, HIV seroconversion time is not known exactly, only that it occurred sometime within a specific time interval (Jewell, et al 1994).
- Time of last HIV- test and first HIV+ test.
9.2 HIV Detection

- HIV-status is detected by performing laboratory analysis of blood samples.

- HIV diagnosis is delayed compared to HIV transmission.

- Infection can happen before the first exam time or might not have happened at last examination (left, right censoring/truncation).
9.3 Methods for Interval Censored Data

- Methods of analysing interval-censored data stem from the Cox proportional hazards model (Cox, 1972).
- Finkelstein (1986) generalised the Cox proportional hazards model.
- Parametric methods readily available (Lindsey and Ryan, 1998).
- Suitable for independent data.
9.4 Drawbacks of These Methods

- Do not reduce to standard survival setting when data are not interval-censored.
- Rigorous theoretical account (Huang and Wellner, 1997).
9.5 Correlated Interval Censored data

- interval-censored data already complicated.
- Dependency complicates even further.
- Arise from sampling method used (multistage sampling).
- For example, migration data.
9.6 Migration Data Networks
9.7 Models for Correlated Data

- Dependency modelled as random effects or frailties.
- Frailties are considered unobserved mutually independent random variables.
- r.v. specified by some parametric distribution (See current workshop on shared frailty models).
- Frailties in likelihoods of (Finkelstein, 1986) or (Huang and Wellner, 1997) results in complex & intractable likelihood functions.
9.8 Fix up Approaches

- Event occurred at the beginning or end of each examination time.


- Ignore correlation in the data and utilize standard univariate interval censored data techniques (Finkelstein, 1986; Huang and Wellner, 1997).

- Resultant naive standard errors lead to invalid inference (Guo and Lin, 1994; Wei, et al 1989).
9.9 Proposed Approach

- Consider frailties as missing/unobserved.
- Treat infection time as unobserved.
- Use known visiting time.
- Separate data into observed and unobserved parts.
9.10 Estimation Methods

- Facilitate the EM algorithm (Dempster, Laird and Rubin, 1977).

- Unobserved data is augmented conditional on the observed data.

- Reduces to techniques for standard correlated survival data (Guo and Rodriguez, 1992; Klein, 1992).

- Bayesian paradigm, very appropriate (Zuma and Lurie, 2005).
9.11 Recall

- Migration data with periodic HIV assessment.
- HIV data clustered within sexual networks.
- Sexual networks are distinct sub-groups connected by sexual relationships.
9.12 The Data

- Distribution of sexual partnerships and HIV infection.

<table>
<thead>
<tr>
<th>Sexual partnership size</th>
<th>Number of sexual partnerships</th>
<th>Percentage</th>
<th>HIV infection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>122</td>
<td>36.0</td>
<td>48</td>
</tr>
<tr>
<td>2</td>
<td>175</td>
<td>51.6</td>
<td>88</td>
</tr>
<tr>
<td>3</td>
<td>37</td>
<td>10.9</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1.2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.3</td>
<td>0</td>
</tr>
</tbody>
</table>

- Analysis restricted to 339 sexual partnership networks.
- Mean sexual partnership size is 1.78 individuals.
- Sexual partnership size ranges from 1 to 5 with only one man in each sexual partnership network.
9.13 **Descriptive Statistics**

- Descriptive statistics of variables used in HIV infection.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Percent</th>
<th>Variable</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Migration status</strong></td>
<td></td>
<td><strong>Recent sexual partners</strong>¹</td>
<td></td>
</tr>
<tr>
<td>Migrant men</td>
<td>31.6</td>
<td>More than one</td>
<td>20.2</td>
</tr>
<tr>
<td>Partners of migrant men</td>
<td>24.8</td>
<td><strong>Lifetime partners</strong></td>
<td></td>
</tr>
<tr>
<td>Non-migrant men</td>
<td>18.6</td>
<td>More than one</td>
<td>67.8</td>
</tr>
<tr>
<td>Partners of non-migrant men</td>
<td>25.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Age in years</strong></td>
<td></td>
<td><strong>Active syphilis</strong></td>
<td></td>
</tr>
<tr>
<td>18 to 24</td>
<td>5.5</td>
<td>Positive</td>
<td>15.7</td>
</tr>
<tr>
<td>25 to 34</td>
<td>29.9</td>
<td><strong>Status of other STIs</strong></td>
<td></td>
</tr>
<tr>
<td>35 or above</td>
<td>64.6</td>
<td>Positive</td>
<td>28.3</td>
</tr>
</tbody>
</table>
9.13.1 Data Notation

- $X_{ij}$ is a vector of covariates associated with
  - $j$th member ($j = 1, \cdots, J_i$) of
  - $i$th sexual network ($i = 1, \cdots, I$)
- $\beta$ vector of coefficients.
- Ordered clinical exam times
  \[ \{0 < v_{ij,1} < v_{ij,2} < \cdots < v_{ij,n_{ij}} < \infty\} \] and HIV indicators
  \[ \{\delta_{ij,1}, \delta_{ij,2}, \cdots, \delta_{ij,n_{ij}}\} \].
- $t_{ij} < v_{ij,1}$ or $t_{ij} \in \{v_{ij,k}; v_{ij,k+1}\}$ or $t_{ij} > v_{ij,n_{ij}}$. 
9.14 Irreversibility of HIV Status

- HIV is irreversible.

- Interval $v_{ij} = \{v_{ij,1}; v_{ij,2}\}$ is sufficient, where
  - $v_{ij,1} = \max\{v_{ij,k} | \delta_{ij,k} = 0\}$ and
  - $v_{ij,2} = \min\{v_{ij,k} | \delta_{ij,k} = 1\}$. 
9.15  Note

• $v_{ij}$ possibly varies.

• Techniques for grouped multivariate survival data (Guo and Lin, 1994) inappropriate.

• $y_{ij} = t_{ij} \in (v_{ij,1}; v_{ij,2}]$ if infection occurred and $y_{ij} = v_{ij,1}$ if right-censored.

• $\delta_{ij} = 1$ if infected with HIV and 0 otherwise.

• $b_i$ denotes the $ith$ sexual network specific frailty.
9.16 Model Formulation

- Cox proportional hazards model (Cox, 1972).
- Hazard function $\lambda(y_{ij}|X_{ij})$ related to $\lambda_0(y_{ij})$ through
  $$\lambda(y_{ij}|X_{ij}) = \lambda_0(y_{ij})e^{\beta'X_{ij}}.$$  
- Cumulative hazards $\Lambda_0(y_{ij}) = \int_0^{y_{ij}} \lambda_0(y_{ij})dy_{ij}$.
- Integrated fixed effects hazard
  $$\Lambda(y_{ij}|X_{ij}) = \int_0^{y_{ij}} \lambda_0(y_{ij})e^{\beta'X_{ij}}dy_{ij}.$$
9.17 Conditional Survival Model

- Satisfy a multiplicative frailty model
  \[ h(y_{ij}|b_i, X_{ij}) = b_i \lambda_0(y_{ij}) e^{\beta'X_{ij}}. \]

- Integrated hazards
  \[ H(y_{ij}|b_i, X_{ij}) = b_i \Lambda(y_{ij}|X_{ij}). \]

- Conditional survival function
  \[ S(y_{ij}|b_i, X_{ij}) = \exp \left\{ -H(y_{ij}|b_i, X_{ij}) \right\}. \]

- HIV+ person contributes to the likelihood
  \[ f(y_{ij}|b_i, X_{ij}) = S(y_{ij}|b_i, X_{ij}) \times h(y_{ij}|b_i, X_{ij}). \]
9.18 Baseline Hazard

- A parametric form for the baseline hazards is assumed.
- Baseline hazards are assumed constant, $\lambda_0(y_{ij}) = \lambda_0$.
- Other forms of baseline hazards can be assumed.
- $\lambda_0(y_{ij}) = \lambda_0 \eta y_{ij}^{\eta-1}$. 
9.19 Sexual Network Frailty

- Parametric or nonparametric frailty distribution can be assumed (Guo and Rodriguez, 1992).

- A gamma distribution with shape and scale parameters $\alpha$ and $\alpha^{-1}$ respectively.

- Several authors in both the frequentist (Klein 1992; Sastry, 1997) and the Bayesian literature (Clayton, 1991; Manda, 1998) have used gamma frailty.
9.19.1 EM Implementation

- Requires calculation of
  \[ Q(\theta; \theta^{(r)}) = E_{\theta^{(r)}}[\log f_c((x; \theta)|y; \theta^{(r)})]. \]

- Marginal conditional distribution of \( b_i \) given the observed data.

- Marginal conditional distribution of \( t_{ij} \) given the observed data.
9.20 Parameter Estimation

- Conditional independence between $y_i$ given $b_i$.

- Complete-data likelihood

$$L_i(b_i, v_i, t_i|\theta) = f(b_i|\alpha) \prod_{j=1}^{J_i} \left[ S(t_{ij}|b_i, X_{ij}) \times h(t_{ij}|b_i, X_{ij}) \right]^{\delta_{ij}}$$

$$\times \left[ S(v_{ij,k}|b_i, X_{ij}) \right]^{1-\delta_{ij}}$$

where $\theta = \{\alpha, \lambda_0, \beta\}$. 
9.21 Log Likelihood

- Complete-data log-likelihood of sexual partnership $i$

\[
l_i(\theta) = \alpha \log \alpha - \log \Gamma(\alpha) + (\alpha - 1) \log b_i - \alpha b_i
\]

\[
+ \sum_{j=1}^{J_i} \delta_{ij}[-b_i \lambda_0 t_{ij} \exp(\beta' X_{ij}) + \log(b_i) + \log(\lambda_0) + \beta' X_{ij}]
\]

\[
- (1 - \delta_{ij})b_i \lambda_0 v_{ij,k} \exp(\beta' X_{ij}).
\]

- Consists (of functions) of unobserved data.
9.21.1 Conditional Distribution of Frailties

- Joint distribution of $b_i$ and $v_i$

$$
\begin{align*}
\frac{f_i(b_i, v_i|\theta)}{f_i(b_i, v_i|\theta)} &= \int_{v_i \delta_i + \cdot \cdot \cdot}^{v_i \delta_i + k + 1} \int_{v_i \delta_i + k}^{v_i \delta_i + k + 1} L_i(b_i, v_i, t_i|\theta) \, dt_i \, dt_i \, \cdots \, dt_i \, \delta_i \\
&= \frac{\alpha^\alpha}{\Gamma(\alpha)} b_i^{\alpha - 1} e^{-b_i(\alpha + \sum_{j=1}^{J_i} (1 - \delta_{ij}) \Lambda(v_{ij,k}|X_{ij})} \\
&\quad \prod_{j=1}^{\delta_i + k} (e^{-b_i \Lambda(v_{ij,k}|X_{ij})} - e^{-b_i \Lambda(v_{ij,k+1}|X_{ij})}).
\end{align*}
$$

- $g_i(b_i|v_i; \theta)$ computed from $f_i(b_i, v_i; \theta)$ as

$$
g_i(b_i|v_i; \theta) = \frac{f_i(b_i, v_i; \theta)}{L_i(\theta; v_i)}.
$$

- Use to calculate conditional expectations.
9.21.2 Conditional Distribution of Infection Time

- Joint conditional distribution
  \[ f_{ij'}(b_i, t_{ij'}|v_i; \theta) = f_i(b_i|v_i; \theta) \cdot f_{ij'}(t_{ij'}|b_i, v_i; \theta). \]

- \( t_{ij'} \) conditional distribution
  \[
  f_{ij'}(t_{ij'}|b_i, v_i, \lambda_0, \beta) = \frac{e^{-b_i \Lambda(t_{ij'}|X_{ij'})} \cdot b_i \lambda_0 e^{\beta'X_{ij'}}}{S(v_{ij',k}|b_i, \lambda_0, \beta) - S(v_{ij',k+1}|b_i, \lambda_0, \beta)}.
  \]

- \( f_i(b_i|v_i; \theta) \) already calculated.
9.21.3 E-step

- $E[b_it_{ij}|v_{ij,k} < t_{ij} \leq v_{ij,k+1}, \theta]$ from $f_{ij}(b_i, t_{ij}|\theta)$.
- $b_i$ and $\log(b_i)$ from $f_i(b_i|v_i; \theta)$
- Require numerical approximations.
- Denote $E[b_i|v_i; \theta]$, $E[\log b_i|v_i; \theta]$ and $E[b_it_{ij}|v_{ij,k} < t_{ij} \leq v_{ij,k+1}; \theta]$ by $\overline{b_i}$, $\overline{\log b_i}$ and $\overline{b_t t_{ij}}$ respectively.
- $\overline{b_i}$, $\overline{\log b_i}$ and $\overline{b_t t_{ij}}$ evaluated at current estimate of $\theta$. 
9.21.4 M-step

- M-step involves maximizing $Q(\theta; \theta^{(r)})$ after inserting $\overline{b_i}$, $\log \overline{b_i}$ and $\overline{b_i t_{ij}}$.

- Not always closed form expression.

- Maximize via Newton-Raphson algorithm which requires first and second derivatives of $Q(\theta; \theta^{(r)})$. 
9.21.5 Computation

- Computation of \( \{\hat{\lambda}_0, \hat{\beta}\} \) using profile likelihood.
- Maximize \( Q(\theta; \theta^{(r)}) \) over \( \lambda_0 \) for all values of \( \beta \) to obtain \( \hat{\lambda}_0(., \beta) \), thereafter maximize \( Q(\alpha, \beta, \hat{\lambda}_0(., \beta)) \) over \( \beta \) to find \( \hat{\beta} \).
- \( \hat{\theta} \) is then used to compute \( b_i, \log b_{ij} \) and \( b_{ij}t_{ij} \).
9.21.6 Computation Steps

- Let $\theta^{(0)} = \{\alpha^{(0)}, \lambda_0^{(0)}, \beta^{(0)}\}$ be the initial parameter estimates and set $r = 1$.
  - (i) Compute $b_i, \log b_i$ and $b_it_{ij}$ using $\theta^{(r-1)}$.
  - (ii) Maximize $l_1(\alpha)$ w.r.t. $\alpha$ to get $\alpha^{(r)}$.
  - (iii) Maximize $l_2(\lambda_0, \beta^{(r-1)})$ w.r.t. $\lambda_0$ to get $\lambda_0^{(r)}$.
  - (iv) Maximize $l_2(\lambda_0^{(r)}, \beta)$ w.r.t. $\beta$ to get $\beta^{(r)}$.
  - (v) Let $\theta^{(r)} = \{\alpha^{(r)}, \lambda_0^{(r)}, \beta^{(r)}\}$. Set $r=r+1$ and repeat steps (i) to (v) until convergence.
9.21.7 Standard Errors

- EM does not provide standard errors.
- Methods of calculating the observed information matrix in the EM context (Meng and Rubin, 1991; McLachlan and Krishnan, 1997; Jamshidian and Jennrich, 2000).
9.22 Fitted Models

- Model without frailty term.
- Model with frailty term.
9.23 Model Fit

- Frailty model contains effects covariates not included in the model.

- AIC for frailty model and standard model is 1445.11 and 1377.61 respectively.

- Frailty model fits the data better.
9.24 Results

- The sexual partnership frailty parameter represents the sexual partnership effect.

- Men whose partners were not included in the study are affected by their own frail.

- Frailty model controls for unobserved effects.
9.24.1 Determinants of HIV Infection

- Circular migrant networks at high risk of HIV infection.
- Migrant men at lesser risk of HIV infection than their partners.
- Young (aged 18 to 24 years) associated with high risk.
- Large number of sexual partners.
- Presence of syphilis or other STIs is a high risk.
### 9.24.2 Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model without frailty</th>
<th>Model with frailty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Standard Errors</td>
</tr>
<tr>
<td><strong>Baseline hazards</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.016</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>Migration status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Migrant men</td>
<td>0.157</td>
<td>0.220</td>
</tr>
<tr>
<td>Partners of migrant men</td>
<td>0.156</td>
<td>0.204</td>
</tr>
<tr>
<td>Non-migrant men</td>
<td>-0.205</td>
<td>0.258</td>
</tr>
<tr>
<td><strong>Age in years</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 to 24</td>
<td>1.031</td>
<td>0.292</td>
</tr>
<tr>
<td>25 to 34</td>
<td>0.524</td>
<td>0.160</td>
</tr>
<tr>
<td><strong>Recent sexual contact partners</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>More than one</td>
<td>0.611</td>
<td>0.192</td>
</tr>
<tr>
<td><strong>Number of lifetime partners</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>More than one</td>
<td>0.349</td>
<td>0.171</td>
</tr>
<tr>
<td><strong>Syphilis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>0.437</td>
<td>0.157</td>
</tr>
<tr>
<td><strong>Status of other STIs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>0.484</td>
<td>0.180</td>
</tr>
<tr>
<td><strong>Frailty variance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexual partnership</td>
<td>0.462</td>
<td>0.054</td>
</tr>
</tbody>
</table>
9.24.3 Frailty Variance Estimate

- Variance $= 0.462$ with $z$-test $= 8.56$.
- Moderate degree of intra-cluster correlation within sexual partnerships, 0.188.
9.24.4 Variance Estimate

- Standard errors are slightly magnified in the frailty model.

- Indicates that fixed effect parameters are now estimated more realistically, but with lower precision.

- Treating each response as independent gives false impression that there is more information in the data than there really is.
9.24.5 Comparing Models

- $\beta$ for recent sexual contact partners and syphilis status were reduced.
- Risk associated with recent sexual reduced by about 23%.
- Intuitively, individuals from large sexual partnerships are at higher risk of HIV infection.
- Inclusion of frailties account for unobserved proneness.
- Results indicate proneness to syphilis by sexual partnerships.
9.24.6 Conclusion

- EM algorithm most suited in this setting.
- Treat infection time and frailties as unobserved.
- Simplifies to standard frailty model.
- No need to augment right-censored data.
9.24.7 Remarks

- Formulated the model and estimation via EM algorithm.
- Frequentist approach.
- Likelihood intractable, numerical methods.
9.24.8 Bayesian Approach

- Bayesian avoid evaluating complex integrals.
- Treat unknown parameters as random variables.
- Assigns prior distribution to unknown parameters.
- Samples generated from joint posterior.
9.25 Bayesian Model

- Framework already hierarchical.
- Fully specified from the frequentist perspective.
- Parameters estimated via EM algorithm.
- Bayesian perspective, specify priors for the
  - fixed effects vector $\beta$,
  - baseline hazard $\lambda_0$ and
  - hyperparameter $\alpha$. 
9.25.1 Some Notation

- $\theta$ unknown parameter vector.
- $g(\theta)$ prior distribution.
- $f(y|\theta)$ likelihood function.
- $g(\theta|y)$ posterior distribution.
9.25.2 Bayes Theorem

- Bayes’ theorem: posterior distribution of $\theta$

$$g(\theta|y) = \frac{f(y|\theta)g(\theta)}{\int f(y|\theta)g(\theta)\,d\theta}.$$  

- Proportional form

$$g(\theta|y) \propto f(y|\theta)g(\theta).$$
9.26 Prior Specification

- Fixed effects $\beta$ prior is MVN ($d_0 = 0, \Sigma_0 = \nu_0 I$)
  
  $- \nu_0$ a suitably chosen large number.

- Baseline hazards, Gamma($\xi_0 = 1, \zeta_0 = 20$).

- Gamma ($\nu_0 = 1, \kappa_0 = 1$) prior for $\alpha$. 
CHAPTER 9. INTERVAL CENSORED DATA

9.27 Directed Acyclic Graph

Diagram:

- **\( (d_0, \sum_0) \)**
- **\( (\xi_0, \zeta_0) \)**
- **\( (\nu_0, \kappa_0) \)**

Nodes:
- **\( \beta \)**
- **\( \lambda_0 \)**
- **\( b_i \)**
- **\( t_{ij} \)**
- **\( y_{ij} \)**
- **\( v_{ij} \)**

Arrows indicate the direction of influence.
9.28 Model Specifications

\[ h(y_{ij} | \beta, b_i) = b_i \lambda(y_{ij}) e^{\beta'X_{ij}} \quad \text{where} \quad \lambda(y_{ij}) = \lambda_0 \]

\[ \beta \sim \text{MVN}(d_0, \Sigma_0) \]

\[ \lambda_0 \sim \text{Ga}(\xi_0, \zeta_0) \]

\[ b_i \sim \text{Ga}(\alpha, \alpha) \]

\[ \alpha \sim \text{Ga}(\nu_0, \kappa_0) \]

\[ t_{ij} \sim \text{Exp}(b_i \lambda_0 \exp(\beta'X_{ij})) \]
9.29 Joint Posterior Distribution

- Bayesian inference requires the joint posterior density of all parameters and hyperparameters given the data.

\[
f(\text{data}, \beta, \lambda_0, t_{ij}, b_i, \alpha) = f(\beta)f(\lambda_0)f(\alpha) \times \left\{ \prod_{i=1}^I f(b_i|\alpha) \prod_{j=1}^{J_i} L_i(y_{ij}|\beta, \lambda_0, b_i) \right\}.
\]

- Posterior density cannot be obtained analytically.

- Use Gibbs sampler.
9.30 Gibbs Conditionals

• Required Gibbs conditional distributions are for
  – sexual network random effects
    \( f(b_i|\text{data, } \beta, \lambda_0, t_{ij}, \alpha) \),
  – infection time
    \( f(t_{ij}|v_{ijk} < t_{ij} \leq v_{ijk+1}, \text{data, } \beta, \lambda_0, b_i, \alpha) \),
  – sexual network random effect inverse variance
    \( f(\alpha|\text{data, } \beta, \lambda_0, t_{ij}, b_i) \),
  – baseline hazards \( f(\lambda_0|\text{data, } \beta, t_{ij}, b_i, \alpha) \)
  – fixed effects \( f(\beta|\text{data, } \lambda_0, t_{ij}, b_i, \alpha) \).
9.30.1 Sexual Network Random Effects

\[
f(b_i|\text{data}, \beta, \lambda_0, t_{ij}, \alpha) \\
\propto b_i^{\sum_{j=1}^{J_{ij}} \delta_{ij} + \alpha - 1} e^{-b_i \left[ \alpha + \sum_{j=1}^{J_{ij}} \delta_{ij} \Lambda(t_{ij}|X_{ij}) + \{1-\delta_{ij}\} \Lambda(v_{ij,k}|X_{ij}) \right]} \\
\]

- **Kernel** of a gamma distribution
  - shape \( \alpha + \sum_{j=1}^{J_{ij}} \delta_{ij} \) and
  - inverse scale
    \[
    \alpha + \sum_{j=1}^{J_{ij}} \left[ \delta_{ij} \Lambda(t_{ij}|X_{ij}) + (1-\delta_{ij}) \Lambda(v_{ij,k}|X_{ij}) \right].
    \]
9.30.2 Infection Time

\[ f(t_{ij} | v_{ij,k} < t_{ij} \leq v_{ij,k+1}, \text{data}, \beta, \lambda_0, b_i, \alpha) \propto \exp(-t_{ij} b_i \lambda_0 e^{\beta'X_{ij}}) \]

- **Kernel** of a gamma distribution
  - shape 1 and
  - inverse scale \( b_i \lambda_0 e^{\beta'X_{ij}} \).

- Sampled directly on condition that \( t_{ij}^* \in (v_{ij,k}, v_{ij,k+1}] \).
CHAPTER 9. INTERVAL CENSORED DATA

9.30.3 Random Effects Inverse Variance

\[
f(\alpha|\text{data}, \beta, \lambda_0, b_i) \\
\propto \alpha^{\nu_0 - 1} \times \left( \frac{\alpha^\alpha}{\Gamma(\alpha)} \right)^I \left( \prod_{i=1}^I b_i \right)^{\alpha-1} e^{-\alpha \left[ \kappa_0 + \sum_{i=1}^I b_i \right]}.
\]

- Does not simplify.
- Methods for sampling from an arbitrary conditional distribution.
- \textit{log-concave} in \( \alpha \) use \textit{adaptive-rejection sampling} scheme (Gilks and Wild, 1992).
9.30.4 Baseline Hazard

\[ f(\lambda_0 | \text{data}, \beta, t_{ij}, b_i, \alpha) \]

- *Kernel* of a gamma distribution
  - shape parameter \( \xi_0 + \sum_{i=1}^{I} \sum_{j=1}^{J_i} \delta_{ij} \) and
  - scale parameter
    \[ \zeta_0 + \sum_{i=1}^{I} \sum_{j=1}^{J_i} \left[ \delta_{ij} b_i t_{ij} e^{\beta'X_{ij}} + (1-\delta_{ij}) b_i v_{ij,k} e^{\beta'X_{ij}} \right]. \]
9.30.5 Fixed Effects

\[ f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha) \approx N(\tilde{\beta}, [I(\tilde{\beta})^{-1}] \]

- Flat prior.
- Multivariate normal candidate.
- Metropolis step (Metropolis et al, 1953; Hastings, 1970; Tierney, 1984; Chib and Greenberg, 1995).
9.30.6 M-H step

- Step 0: Chose starting point $\theta_0$ and set $i = 1$.
- Step 1: Generate candidate point $\theta^*$ from $q(\theta_i, .)$ and $u$ from $U(0, 1)$.
- Step 2: Set $\theta_{i+1} = \theta^*$ if $u \leq a(\theta_i, \theta^*)$ and $\theta_{i+1} = \theta_i$ otherwise, where acceptance probability is given by

$$a(\theta, \vartheta) = \min\left\{\frac{\pi(\vartheta|y)q(\vartheta, \theta)}{\pi(\theta|y)q(\theta, \vartheta)}; 1\right\}.$$
9.31 Applications

- Model parameters: 526 parameters
  - 9 fixed effects,
  - 1 baseline hazard,
  - 339 sexual network specific random effects,
  - 176 infection times and
  - inverse scale for sexual network frailty distribution.
9.31.1 Chains

- Five parallel chains were run from independent starting points.
- Equal sampling rate led to correlated successive values for $b_i, t_{ij}, \alpha$ and $\lambda_0$.
- Iteration modified to iterate through $b_i, t_{ij}, \alpha$ and $\lambda_0$ five times for each draw of $\beta$.
- Improved efficiency.
9.31.2 Sampling Scheme

- Involved EM for MLE and Fisher information.
- MLE for M-H steps.
- Computationally intensive.
- Lead to inefficient sampling and contributed to the decision to modify the sampling scheme.
9.31.3 Diagnostics

- GR reduction factor satisfactory after modifications (Gelman and Rubin, 1992).
- Geweke convergence diagnostics (Geweke, 1992).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Chain 1</th>
<th>Chain 2</th>
<th>Chain 3</th>
<th>Chain 4</th>
<th>Chain 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline hazard</td>
<td>-1.280</td>
<td>-0.333</td>
<td>1.490</td>
<td>-0.054</td>
<td>-1.760</td>
</tr>
<tr>
<td>Migrant men</td>
<td>1.020</td>
<td>-0.542</td>
<td>0.659</td>
<td>-0.971</td>
<td>0.317</td>
</tr>
<tr>
<td>Part. of migrant men</td>
<td>1.260</td>
<td>-0.614</td>
<td>0.196</td>
<td>0.472</td>
<td>1.300</td>
</tr>
<tr>
<td>Non-migrant men</td>
<td>1.560</td>
<td>-1.404</td>
<td>1.530</td>
<td>1.250</td>
<td>-0.180</td>
</tr>
<tr>
<td>Age:18 to 24</td>
<td>-0.128</td>
<td>1.130</td>
<td>-1.280</td>
<td>-0.806</td>
<td>1.130</td>
</tr>
<tr>
<td>Age:25 to 34</td>
<td>0.875</td>
<td>0.510</td>
<td>-1.050</td>
<td>-0.924</td>
<td>-1.130</td>
</tr>
<tr>
<td>Current partners</td>
<td>-0.978</td>
<td>0.560</td>
<td>-1.170</td>
<td>0.556</td>
<td>2.600</td>
</tr>
<tr>
<td>Lifetime partners</td>
<td>0.914</td>
<td>-0.163</td>
<td>-2.140</td>
<td>0.670</td>
<td>0.217</td>
</tr>
<tr>
<td>Syphilis</td>
<td>-0.840</td>
<td>-2.860</td>
<td>-0.877</td>
<td>-1.950</td>
<td>1.950</td>
</tr>
<tr>
<td>Other STIs</td>
<td>-0.967</td>
<td>-0.192</td>
<td>2.290</td>
<td>2.280</td>
<td>0.853</td>
</tr>
<tr>
<td>Frailty variance</td>
<td>1.140</td>
<td>-0.228</td>
<td>-0.011</td>
<td>0.267</td>
<td>-0.861</td>
</tr>
</tbody>
</table>
9.31.4 Parameter Estimates

- Histograms of the fixed effects parameters.
### 9.31.5 EM and Gibbs estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>EM algorithm</th>
<th>Gibbs sampler</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
</tr>
<tr>
<td><strong>Baseline hazard</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.007</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>Migration status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Migrant men</td>
<td>0.460</td>
<td>0.216</td>
</tr>
<tr>
<td>Partners of migrant men</td>
<td>0.299</td>
<td>0.210</td>
</tr>
<tr>
<td>Non-migrant men</td>
<td>-0.219</td>
<td>0.259</td>
</tr>
<tr>
<td><strong>Age in years</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 to 24</td>
<td>2.455</td>
<td>0.296</td>
</tr>
<tr>
<td>25 to 34</td>
<td>1.072</td>
<td>0.163</td>
</tr>
<tr>
<td><strong>Recent sexual contact partners</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>More than one</td>
<td>0.558</td>
<td>0.189</td>
</tr>
<tr>
<td><strong>Number of lifetime partners</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>More than one</td>
<td>0.328</td>
<td>0.172</td>
</tr>
<tr>
<td><strong>Syphilis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0=Negative, 1=Positive</td>
<td>0.284</td>
<td>0.158</td>
</tr>
<tr>
<td><strong>Status of other STIs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0=Negative, 1=Positive</td>
<td>0.503</td>
<td>0.181</td>
</tr>
<tr>
<td><strong>Frailty variance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sexual network</td>
<td>0.459</td>
<td>0.069</td>
</tr>
</tbody>
</table>
9.31.6 Parameter Estimates

- Estimates similar to the respective modes obtained from the EM algorithm.
- Gibbs estimates more variable.
- Gibbs sexual network frailty variance is large compared to EM
- Posterior median and mean is 0.788 and 0.812.
- EM mode was estimated 0.462.
9.31.7 Remark

- Agreement expected, prior nearly flat.

- ML variance biased downwards: ML unable to correct for variability of unobserved frailties and infection time.

- THIS IN COMPLETELY UNDESIRABLE.
9.31.8 Bayesian Appraisal

- Capable of not only incorporating information about frailties and infection time, but also uncertainties about available information.

- Uncertainty about variance components is incorporated through the choice of a plausible prior.
9.31.9 Conclusion

- Gibbs conditionals simplify to two iterative steps.
  - *imputation step* which draws $b_i^{(r)}$ and $t_{ij}^{(r)}$ from $f(b_i|\text{data}, \beta, \lambda_0, t_{ij}^{(r-1)}, \alpha)$ and $f(t_{ij}|v_{ijk} < t_{ij} \leq v_{ijk+1}, \text{data}, \beta, \lambda_0, b_i^{(r)}, \alpha)$ and
  - *posterior step* which draws $\theta^{(r)} = \{\beta^{(r)}, \lambda_0^{(r)}, \alpha^{(r)}\}$ from $f(\theta^{(r)}|\text{data}, b_i^{(r)}, t_{ij}^{(r)}).

- Stochastic counterparts to the E-step and M-step of the EM algorithm.

- Broad regularity conditions, $(b_i^{(r)}, t_{ij}^{(r)}, r = 1, 2, \cdots)$ converges to $f(b_i^{(r)}, t_{ij}^{(r)}|\text{data})$ (Gilks, et al 1996).
9.31.10 Software

• The model was implemented in C++.
• Exploring WinBUGS.
Part IV

Case Studies
Chapter 10

Modeling the Outbreak of AIDS in US (I)

Modeling the outbreak of AIDS in US using

- Generalized linear models.
- Nonlinear models.
- Hierarchical Bayesian models using Winbugs.
10.1 Introduction

- Number of AIDS cases in US from 1982 to 1990.
- Under reporting problem.

Figure 10.1: Number of reported cases of AIDS in US.
10.2 Generalized Linear Models

- The exponential family:
  \[ f(y_i, \theta_i) = \exp[y_i \theta_i - b(\theta_i) + c(y_i)] \]

- Example Poisson distribution:
  \[ f(y_i, \theta_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} = \exp[y_i \log(\mu_i) + \mu_i - \log(y_i!)] \]
  - \( \theta_i = \log(\mu_i) \).
  - \( b(\theta_i) = \exp(\theta_i) \).
  - \( c(y_i) = -\log(y_i!) \).
10.2.1 Scale Parameter

- The exponential family with scale parameter

\[
f(y_i, \theta_i, \phi) = \exp \left[ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right]
\]

- Normal distribution:

\[
f(y_i, \theta_i, \phi) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left[ \frac{-(y_i - \mu_i)^2}{2\sigma^2} \right]
\]

\[
= \exp \left[ \frac{y_i \mu_i}{\sigma^2} - \frac{\mu_i^2}{2\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2} \log(2\pi \sigma^2) \right]
\]

- \( \theta_i = \mu_i \).
- \( b(\theta_i) = \frac{\theta_i^2}{2} \).
- \( \phi = \sigma^2 \).
- \( c(y_i, \phi) = \frac{y_i^2}{2\sigma^2} - \frac{1}{2} \log(2\pi \sigma^2) \).
10.2.2 The Components of GLM

- The Distribution of the response.
- Linear Predictor.
- Link function.

Example: Linear Regression

- Distribution: $y_i = \mu(x_i) + \varepsilon_i \Rightarrow y_i \sim N(\mu(x_i), \sigma^2)$.
- Linear predictor: $\mu(x_i) = X_i \beta$.
- Link function: $\theta_i = \mu(x_i) \Rightarrow$ identity link function.

Example: Poisson Regression

- Distribution: $y_i \sim \text{Poisson}(\mu(x_i))$.
- Linear predictor: $\eta(x_i) = X_i \beta$ and $\mu(x_i) = \exp(\eta_i)$.
- Link function: $\theta_i = \log(\mu(x_i)) \Rightarrow \log$ link function.
10.3 Nonlinear Models

- The basic model has two main components: the nonlinear function for the mean response and distribution of the response.
- For example,
  \[ y_i = \mu(x_i, \beta) + \varepsilon_i. \]
  Here \( \mu(x_i, \beta) \) is a nonlinear function.
- If \( \varepsilon_i \sim N(0, \sigma^2) \) then \( E(y_i) = \mu(x_i, \beta) \).

Examples

- Biexponential model:
  \[ \mu(x, \beta) = \beta_1 \exp(-\beta_2 x) + \beta_3 \exp(-\beta_4 x). \]
- Gomperz growth model:
  \[ \mu(x, \beta) = \beta_1 \left(1 - e^{-\beta_2 e^{\beta_3 x}}\right). \]
10.4 A Simple Model for an Epidemic

- Let us assume that an epidemic starts at time $t_0$.
- At $t_0$, the number of infected individuals is $Y(0)$.
- It is assumed that the number of seropositives (the number of infected individuals) initially rises exponentially, that is,

$$Y(t) = Y(0) \exp(\Lambda t).$$

- Note that from time $t_1$, $Y(0)$ is constant.
- If we assume that $Y(0) = 1$ then,

$$Y(t) = \exp(\Lambda t).$$
10.5 Growth Curves

- Growth curves aim to describe the growth of the response with respect to a predictor, often the predictor is time.
- In general, the growth profile is nonlinear function of time and often reaching an asymptote.
- Let us assume that the response is a Poisson random variable,
\[ Y(t) \sim \text{Poisson}(\mu(t)), \]
where \( \mu(t) \) is the mean which depends on time.
10.6 Exponential Growth Curves

- The most simple model assumes that the mean is a nonlinear exponential function of the time,

\[ \mu(t) = \alpha e^{\beta t}. \]

- Note that for \( Y(0) = \alpha \) and \( \Lambda = \beta \), the exponential growth model corresponds to the simple model for the epidemic discussed above.

- For a log link function we have

\[ \log(\mu(t)) = \log(\alpha) + \beta t. \]

- Hence, the exponential growth model can be fitted as a GLM with Poisson distribution and log link function.
10.6.1  SAS Code for the GLM

- For the AIDS data, we use data from 1982 to the end of 1987 (to overcome the problem of under reporting).

- The response is cases and the time is ti.

- Procedure GENMOD:

  ```
  proc genmod data=hiv;
  model cases=ti/dist=poission;
  run;
  ```
10.6.2 Output

The GENMOD Procedure

Model Information

Data Set: WORK.HIV
Distribution: Poisson
Link Function: Log
Dependent Variable: cases
Observations Used: 24

Criteria For Assessing Goodness Of Fit

<table>
<thead>
<tr>
<th>Criterion</th>
<th>DF</th>
<th>Value</th>
<th>Value/DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>22</td>
<td>1654.5843</td>
<td>75.2084</td>
</tr>
<tr>
<td>Scaled Deviance</td>
<td>22</td>
<td>1654.5843</td>
<td>75.2084</td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>22</td>
<td>1640.1287</td>
<td>74.5513</td>
</tr>
<tr>
<td>Scaled Pearson X2</td>
<td>22</td>
<td>1640.1287</td>
<td>74.5513</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td></td>
<td>486477.9610</td>
<td></td>
</tr>
</tbody>
</table>

Algorithm converged.

Analysis Of Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald 95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>5.6634</td>
<td>0.0138</td>
<td>5.6363 - 5.6905</td>
<td>167627</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>ti</td>
<td>1</td>
<td>0.1445</td>
<td>0.0007</td>
<td>0.1431 - 0.1459</td>
<td>39824.1</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Scale</td>
<td>0</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000 - 1.0000</td>
<td>39824.1</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
10.6.3 SAS Code for the nonlinear model

- The same model can be fitted using procedure NLMIXED.
- The response is cases and the time is ti.
- Procedure NLMIXED:

```sas
proc nlmixed data=hiv;
parms K=11, alpha=0.9;
mu = K*exp(alpha*ti);
model cases ~ poisson(mu);
run;
```
### 10.6.4 Output

| Parameter | Estimate | Standard Error | DF  | t Value | Pr > |t| | Alpha | Lower | Upper | Gradient |
|-----------|----------|----------------|-----|---------|------|---|-----|-------|-------|---------|
| K         | 288.12   | 3.9855         | 24  | 72.29   | < .0001 | 0.05 | 279.89 | 296.35 | -1.49E-7 |
| alpha     | 0.1445   | 0.000724       | 24  | 199.56  | < .0001 | 0.05 | 0.1430 | 0.1460 | -0.0005 |

- Note that 288.12 = \( \exp(5.66) \) = the intercept of the GLM.
Figure 10.2: Exponential model.
10.7 Logistic Growth Curve

- The growth in the exponential model increases with time.
- For logistic growth model the mean is given by
  \[ \mu(t) = K \frac{ae^{\beta t}}{1 + ae^{\beta t}}. \]
- \( K \) is the asymptote for \( t \rightarrow \infty \).
10.7.1 Fitting Logistic Growth Curve as a GLM

- \( Y(t) \sim \text{Poisson}(\mu(t)) \).

- Let us assume that \( K \) is known. Using logit link function, i.e.,
  \[
  \log \left( \frac{\mu(t)}{K} \right) = \log(\alpha) + \beta t,
  \]
  since
  \[
  \frac{\mu(t)}{K} = \frac{\alpha e^{\beta t}}{1 + \alpha e^{\beta t}}.
  \]

- For known \( K \) the logistic model is a GLM with Poisson distribution and logit link function.

- \( K \) can be found in a grid search.
10.7.2 SAS Macro

%macro glmgc2(data,k,linki);
   data a;
      set &data;
   y1=cases/(&k);
run;
proc genmod data=a;
model y1=ti/link=&linki dist=POISSON;
title &k,&linki;
run;
%mend;
Output GENMOD

K=10235,Link=logit  10:48 Wednesday, August 11, 2004 170
The GENMOD Procedure
Model Information
Data Set WORK.A
Distribution Poisson
Link Function Logit
Dependent Variable y1
Observations Used 25

Criteria For Assessing Goodness Of Fit

<table>
<thead>
<tr>
<th>Criterion</th>
<th>DF</th>
<th>Value</th>
<th>Value/DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>23</td>
<td>0.0422</td>
<td>0.0018</td>
</tr>
<tr>
<td>Scaled Deviance</td>
<td>23</td>
<td>0.0422</td>
<td>0.0018</td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>23</td>
<td>0.0426</td>
<td>0.0019</td>
</tr>
<tr>
<td>Scaled Pearson X2</td>
<td>23</td>
<td>0.0426</td>
<td>0.0019</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td></td>
<td>-13.3876</td>
<td></td>
</tr>
</tbody>
</table>

Analysis Of Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald 95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-4.2367</td>
<td>1.9843</td>
<td>-8.1259</td>
<td>-0.3476</td>
<td>0.0328</td>
</tr>
<tr>
<td>ti</td>
<td>1</td>
<td>0.2209</td>
<td>0.1263</td>
<td>-0.0265</td>
<td>0.4684</td>
<td>0.0802</td>
</tr>
<tr>
<td>Scale</td>
<td>0</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

NOTE: The scale parameter was held fixed.
10.7.3 Fitting Logistic Growth Curve as a Nonlinear Model

- Using SAS procedure NLMIXED

```
proc nlmixed data=hiv;
parms K=7800, alpha=5, beta=4;
mu = K*(alpha*exp(beta*ti)/(1+alpha*exp(beta*ti)));
model cases ~ poisson(mu);
run;
```

**Output NLMIXED**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
<th>Gradient</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>10763</td>
<td>0</td>
<td>24</td>
<td>Infty</td>
<td>&lt;.0001</td>
<td></td>
<td>0.05</td>
<td>-Infty</td>
<td>Infty</td>
<td>-1.04E-7</td>
<td></td>
</tr>
<tr>
<td>alpha</td>
<td>0.01413</td>
<td>0.000277</td>
<td>24</td>
<td>51.05</td>
<td>&lt;.0001</td>
<td></td>
<td>0.05</td>
<td>0.01356</td>
<td>0.01470</td>
<td>-0.06017</td>
<td></td>
</tr>
<tr>
<td>beta</td>
<td>0.2171</td>
<td>0.001246</td>
<td>24</td>
<td>174.28</td>
<td>&lt;.0001</td>
<td></td>
<td>0.05</td>
<td>0.2146</td>
<td>0.2197</td>
<td>-0.00804</td>
<td></td>
</tr>
</tbody>
</table>
Figure 10.4: Logistic model.
10.8 Gomperz Growth Curve

- For Gomperz growth model the mean is given by
  \[ \mu(t) = K \left( 1 - e^{-\alpha e^{\beta t}} \right). \]
- \( K \) is the asymptote when \( t \rightarrow \infty \).
- Note that for a known \( K \) we have:
  \[ \frac{\mu(t)}{K} = \left( 1 - e^{-\alpha e^{\beta t}} \right). \]
10.8.1 Fitting Gomperz Growth Curve as a GLM

- Using complementary-log-log link function, we have,
  \[
  \log \left[ - \log \left( 1 - \frac{\mu(t)}{K} \right) \right] = \log(\alpha) + \beta t.
  \]
  \[\Rightarrow\] A GLM with Poisson distribution and log link function.
- \( K \) can be found in a grid search.
Output SAS procedure GENMOD

Output of SAS procedure GENMOD:

```
K=7936.79, link=cloglog  10:48 Wednesday, August 11, 2004 171
The GENMOD Procedure
  Model Information
  Data Set          WORK.A
  Distribution      Poisson
  Link Function     CLL
  Dependent Variable y1
  Observations Used 25

Criteria For Assessing Goodness Of Fit

<table>
<thead>
<tr>
<th>Criterion</th>
<th>DF</th>
<th>Value</th>
<th>Value/DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>23</td>
<td>0.0437</td>
<td>0.0019</td>
</tr>
<tr>
<td>Scaled Deviance</td>
<td>23</td>
<td>0.0437</td>
<td>0.0019</td>
</tr>
<tr>
<td>Pearson Chi-Square</td>
<td>23</td>
<td>0.0444</td>
<td>0.0019</td>
</tr>
<tr>
<td>Scaled Pearson X2</td>
<td>23</td>
<td>0.0444</td>
<td>0.0019</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td></td>
<td>-14.8814</td>
<td></td>
</tr>
</tbody>
</table>

Analysis Of Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald 95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-3.9147</td>
<td>1.7522</td>
<td>-7.3489</td>
<td>-0.4804</td>
<td>4.99</td>
</tr>
<tr>
<td>ti</td>
<td>1</td>
<td>0.2078</td>
<td>0.1140</td>
<td>-0.0157</td>
<td>0.4312</td>
<td>3.32</td>
</tr>
<tr>
<td>Scale</td>
<td>0</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: The scale parameter was held fixed.
```
10.8.2 Fitting a Gomperz Growth Curve as a Nonlinear Model

- Procedure NLMIXED.

```plaintext
proc nlmixed data=hiv;
parms K=4000,beta=0.05,alpha=1;
mu = K*(1-exp(-alpha*exp(beta*ti)));
model cases ~ poisson(mu);
run;
```

Output

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
<th>Gradient</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>7976.24</td>
<td>101.92</td>
<td>24</td>
<td>78.26</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>7765.89</td>
<td>8186.58</td>
<td>-5.98E-9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta</td>
<td>0.2074</td>
<td>0.001799</td>
<td>24</td>
<td>115.28</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>0.2036</td>
<td>0.2111</td>
<td>-0.00052</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alpha</td>
<td>0.01992</td>
<td>0.000393</td>
<td>24</td>
<td>50.69</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>0.01911</td>
<td>0.02073</td>
<td>-0.00256</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 10.6: Gomperz model.
Chapter 11

Modeling the Outbreak of AIDS in US (II)
11.1 Hierarchical Bayesian Models

11.1.1 Very Short Introduction

Bayes’ Theorem, the Likelihood and the Prior Distribution

• Suppose we have a set of hypotheses

\[ H_1, \ldots, H_k, \]

with prior probabilities

\[ P(H_1), \ldots, P(H_k). \]

• For example, let \( H_1 \) represent the event “HIV-positive” and \( H_2 \) the event “HIV-negative” and let \( P(H_1) \) be the proportion of HIV-positives in the population and \( P(H_2) \) the proportion of HIV-negatives in the population.
• Suppose that we collect data, $A$, then the likelihood of the data given the hypotheses $H_i$ is, for $i = 1, \ldots, k$,

$$P(A|H_i).$$

• For example, suppose that an individual was tested for HIV, then $A$ is the result of the test. Note that in this example $A$ could be a negative or positive test.

• Bayes states that the posterior probabilities for the hypothesis $H_i$, given the data $A$, can be expressed as

$$P(H_i|A) = \frac{P(A|H_i) \times P(H_i)}{\sum_{i=1}^{k} P(A|H_i) \times P(H_i)}.$$ 

• The general expression for posterior probabilities is

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\sum_{i=1}^{k} \text{Likelihood} \times \text{Prior}}.$$
Components of Bayesian Inference

- Let $y$ be the observed data and $\theta$ the unknown parameter
  - The **likelihood**: $P(y|\theta)$.
  - The **prior distribution** of the parameter: $P(\theta)$.
  - The **posterior distribution** of the parameter given the data: $P(\theta|y)$.

- Bayes theorem tells us how to calculate the posterior distribution
  \[
  P(H_i|A) = \frac{P(A|H_i) \times P(H_i)}{\sum_{i=1}^{k} P(A|H_i) \times P(H_i)}. 
  \]
• In the general case

\[ P(\theta|y) = \frac{P(y|\theta) \times P(\theta)}{\int P(y|\theta)P(\theta)d\theta} \propto P(y|\theta) \times P(\theta). \]

or, using other notation as in the previous chapter,

\[ \pi(\theta|x) \propto \pi(\theta)f(x|\theta). \]

• Basic form can be formulated as:

\textbf{POSTERIOR} \propto \textbf{LIKELIHOOD} \times \textbf{PRIOR}
How to Perform Bayesian Inference?

• Within the Bayesian framework, inference and estimation is based on the posterior distribution. This is in contrast with the frequentist approach in which inference and estimation is based on (asymptotic) sampling theory.

• The basic steps
  – Specify prior for $\theta$.
  – Apply Bayes’ theorem: $P(\theta|y) \propto P(y|\theta) \times P(\theta)$.
  – Examine the posterior distribution $P(\theta|y)$. 
Markov Chain Monte Carlo (MCMC) Simulation

- The basic idea is to generate a sample for the unobserved parameter $\theta^{(1)}, \ldots, \theta^{(n)}$ from the full conditional distribution of $\theta$,

$$P(\theta | \text{all other parameters in the model and the data}).$$

See Section

- Once a sample for $\theta^{(i)}$ is generated, we can approximate the posterior distribution of $\theta$ and estimate the posterior mean.

**How to Generate the Sample for $\theta$?**

In Section ??- ??, we use two methods:

- the accept-rejection method.
- the Gibbs sampler.
11.1.2 How to Calculate the Posterior Mean? Monte Carlo Integration

- The posterior mean of any function \( t(\theta) \) is define by

\[
E(t(\theta)|y) = \int t(\theta)P(\theta|y)d\theta,
\]

for example for \( t(\theta) = \theta \) we have

\[
E(\theta|y) = \int \theta P(\theta|y)d\theta,
\]

which is the mean of the posterior distribution.

- If we use MCMC methods to approximate the posterior distribution we can use the so called Monte Carlo integration method to estimate the posterior mean:

\[
\bar{\theta} = \frac{1}{N} \sum_{i=1}^{N} \theta^{(i)}
\]

Note that \( \bar{\theta} \) is the sample mean of the sequence of \( \theta^{(i)} \).
11.1.3 Hierarchical Logistic Growth Model

The Likelihood

• The first level of the model is the likelihood:

First stage:

\[
\begin{aligned}
    y(t) & \sim \text{Poisson}(\mu(t)), \\
    \mu(t) & = K \frac{\alpha e^{\beta t}}{1 + \alpha e^{\beta t}}.
\end{aligned}
\]

• Within the Bayesian framework, \( \alpha, \beta \) and \( K \) are random variables for which a probabilistic model is needed to be specified.
The Prior Distributions

- We assume independent normal prior for the parameters:

\[
\begin{align*}
\alpha & \sim \mathcal{N}(\mu_\alpha, \tau_\alpha^2), \\
\beta & \sim \mathcal{N}(\mu_\beta, \tau_\beta^2), \\
K & \sim \mathcal{N}(\mu_K, \tau_K).
\end{align*}
\]

- \(\mu_\alpha\) and \(\tau_\alpha^2\) are called hyperparameters. Note that

\[
\tau_\ell^2 = \frac{1}{\sigma_\ell^2}.
\]

Hyperprior Distribution

- To complete the specification of a full hierarchical model we need to specify prior distributions for the hyperparameters.
• For our analysis we choose flat (or noninformative priors)

Third stage:

\[
\begin{aligned}
&\mu_\alpha \sim \text{N}(0, 100000), \\
&\mu_\beta \sim \text{N}(0, 100000), \\
&\mu_K \sim \text{N}(0, 100000), \\
&\tau^2_\alpha \sim \text{gamma}(0.0001, 0.0001), \\
&\tau^2_\beta \sim \text{gamma}(0.0001, 0.0001), \\
&\tau^2_K \sim \text{gamma}(0.0001, 0.0001).
\end{aligned}
\]
**Graphical Model**

Figure 11.1: Graphical model for the breakout of AIDS.
Full Conditional Distributions

- Recall that $y(t) \sim \text{Poisson}(\mu(t))$ and $\alpha \sim \text{N}(\mu_\alpha, \sigma^2_\alpha)$, $\sigma^2_\alpha = 1/\tau^2_\alpha$.

- The full conditional distribution of $\alpha$, given all other parameters and the observed data,

$$P(\alpha|\beta, K) = \frac{1}{\sqrt{2\pi}\sigma^2_\alpha e^{\frac{1}{2\sigma^2_\alpha}(\alpha-\mu_\alpha)^2}} \times \prod_{t=1}^{n} \frac{1}{y_t!} e^{(-\mu_t)} \mu_t^{y_t}$$
Winbugs Code

model
{
for( i in 1 : T ) {
cases[i] ~ dpois(mu[i])
mu[i]<-K*(alpha*exp(beta*timei[i])/(1+alpha*exp(beta*timei[i])))
}
alpha ~ dnorm(mu.alpha,tau.alpha)
beta ~ dnorm(mu.beta,tau.beta)
K ~ dnorm(mu.K,tau.K)
uu.mu.alpha ~ dnorm(0.0,1.0E-6)
uu.mu.beta ~ dnorm(0.0,1.0E-6)
uu.mu.K ~ dnorm(0.0,1.0E-6)
uu.tau.alpha~dgamma(0.001,0.001)
uu.tau.beta~dgamma(0.001,0.001)
uu.tau.K~dgamma(0.001,0.001)
sig.alpha<-1/tau.alpha
sig.beta<-1/tau.beta
}
Winbugs Code

- The likelihood:

```c
for(i in 1:T) {
  cases[i] ~ dpois(mu[i])
  mu[i] <- K*(alpha*exp(beta*time[i])/(1+alpha*exp(beta*time[i])))
}
```

- Priors and hyperprior:

```c
alpha ~ dnorm(mu.alpha,tau.alpha)
beta ~ dnorm(mu.beta,tau.beta)
K ~ dnorm(mu.K,tau.K)
mu.alpha ~ dnorm(0.0,1.0E-6)
mu.beta ~ dnorm(0.0,1.0E-6)
mu.K ~ dnorm(0.0,1.0E-6)
tau.alpha~dgamma(0.001,0.001)
tau.beta~dgamma(0.001,0.001)
tau.K~dgamma(0.001,0.001)
```

- In Winbugs, for normal distribution, $\sigma^2 = 1/\tau^2$.

```c
sig.alpha<-1/tau.alpha
sig.beta<-1/tau.beta
```
### Posterior Means

- MCMC simulation with 150000 iterations. We monitor the chain every 50 iterations.
- Sample size: 3000.

<table>
<thead>
<tr>
<th>node</th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>2.5%</th>
<th>median</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>10760.0</td>
<td>217.4</td>
<td>7.655</td>
<td>10350.0</td>
<td>10760.0</td>
<td>11210.0</td>
</tr>
<tr>
<td>alpha</td>
<td>0.01413</td>
<td>2.86E-4</td>
<td>6.204E-6</td>
<td>0.01358</td>
<td>0.01413</td>
<td>0.01469</td>
</tr>
<tr>
<td>beta</td>
<td>0.2172</td>
<td>0.002349</td>
<td>8.649E-5</td>
<td>0.2126</td>
<td>0.2172</td>
<td>0.2219</td>
</tr>
</tbody>
</table>
Density Estimate for the Posterior Distributions

Figure 11.2: Logistic model. Kernel density estimate for the posterior distribution of $\alpha, \beta$ and $K$. 
CHAPTER 11. MODELING THE OUTBREAK OF AIDS IN US (II)  

Trace Plots and Autocorrelation functions

Figure 11.3: Logistic model. Trace plots and autocorrelation for the chains of $\alpha$, $\beta$ and $K$. 
The Burn-in Period

![Trace plots for β and K. The first 6500 iterations.](image)

Figure 11.4: Logistic model. Trace plots for \( \beta \) and \( K \). The first 6500 iterations.
Figure 11.5: Logistic model. Trace plots for $\beta$ and $K$. Iteration 7500-11000.
Autocorrelation Function

Figure 11.6: Logistic model. Autocorrelation function.
11.1.4 Prediction

Posterior Predictive Distribution

- Let $y = (y_1, \ldots, y_n)$ be the observed data.
- After the data $y$ have been observed, we can predict an unknown observation $\tilde{y}$ from the same process.
- The distribution of $\tilde{y}$ is called posterior predictive distribution.

$$
P(\tilde{y}|y) = \int P(\tilde{y}, \theta|y)d\theta = \int P(\tilde{y}|\theta, y)P(\theta|y)d\theta = \int P(\tilde{y}|\theta)P(\theta|y)d\theta$$

- Note that $P(\tilde{y}|\theta, y) = P(\tilde{y}|\theta)$ since $\tilde{y}$ and $y$ are conditionally independent given $\theta$. 
Prediction the number of AIDS cases from 1998-1992

Figure 11.7: Logistic and Gomperz models. Posterior means for the number of cases.
Chapter 12

Who Infects Who In the Spiraling Epidemic of HIV

12.1 Couples and HIV Infection

- Most studies focus on couple counselling.
- Coping with dissimilar HIV status in a couple.
- Important to identify the direction of the epidemic among couples.
12.2 HIV in Couples

- Have same or dissimilar HIV status.
- Concordance or discordance status.
- Concordance if partners have same HIV (+ or - status) status.
- Discordance of partners have different HIV status.
12.3 Migration and Man

- Considered to be the ones bringing an infection to their female partners on their return.
- If so, one would expect in discordant couples that a man is HIV+ and woman HIV-, waiting to be infected.
- Case study of migrant couples and non-migrant couples designed to study this aspect of discordancy and the effect of migration.
12.4 Migrant and Non-migrant Couples

- Subset of migration project.
- 168 couples recruited into the project.
- 98 (58%) involved a migrant man.
- 70 (42%) involved a non-migrant man.
- All women were non-migrants.
- HIV prevalence 19.9%.
- Men 24.4% and women 15.5%, p=0.04, OR=1.8.
12.5  **Patterns of Infection**

<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
<th>Overall n=168(%)</th>
<th>Migrant couples n=98(%)</th>
<th>Non-migrant couples n=70(%)</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV-</td>
<td>HIV+</td>
<td>25(14.9)</td>
<td>10(6.0)</td>
<td>16(9.5)</td>
<td>0.48</td>
</tr>
<tr>
<td>HIV-</td>
<td>HIV-</td>
<td>117(69.6)</td>
<td>64(65.3)</td>
<td>53(75.7)</td>
<td>0.15</td>
</tr>
<tr>
<td>HIV+</td>
<td>HIV-</td>
<td>25(14.9)</td>
<td>19(19.4)</td>
<td>6(8.6)</td>
<td>0.05</td>
</tr>
<tr>
<td>HIV-</td>
<td>HIV+</td>
<td>10(6.0)</td>
<td>7(7.1)</td>
<td>3(4.3)</td>
<td>0.66</td>
</tr>
<tr>
<td>HIV+</td>
<td>HIV+</td>
<td>16(9.5)</td>
<td>8(8.2)</td>
<td>8(11.4)</td>
<td>0.48</td>
</tr>
</tbody>
</table>

- Migrant couples as likely as non-migrant couples to have neither partner infected.
- 21% couples discordant.
- Migrant couples 2.5 times likely to be discordant.
- Of discordant couples, man HIV+(71%) and woman HIV+(29%).
12.6 Construction of Variables

- Primary endpoint: presence of at least one HIV+ individual in a couple.
- Determine demographic and biomedical factors related to HIV.
- Age and age at first sexual intercourse refer to the average characteristics within the couple.
- Lifetime partners and current regular partners refer to combined couple total.
- STI is last 4 months refer to status of at least one member.
12.7 Statistical Methods and Mathematical Model

- Logistic regression model used to investigate the relationship between HIV and possible risk factors.
- To investigate the RR of infection for migrant and non-migrant men and women from their spouses and from partners outside the relationship, a mathematical model was constructed.
12.8 Mathematical Model

- Directions of infection in the couples.
12.8.1 Probabilities of Infection

\[ P_{nn} = (1 - \alpha)(1 - \beta) \]
\[ P_{pn} = \alpha(1 - \beta)(1 - \delta) \]
\[ P_{np} = \beta(1 - \alpha)(1 - \gamma) \]
\[ P_{pp} = \alpha \beta + \alpha \gamma + \beta \gamma - \alpha \beta (\gamma + \delta) \]

- Assume various transmissibility rates of 2:1, 3:1, 1:1.
- Model using a Binomial distribution.
- Contact matrix in mixing pattern.
### 12.8.2 Estimated Infection Probabilities

<table>
<thead>
<tr>
<th>Gender</th>
<th>Overall Outside</th>
<th>Migrant couples Spouse</th>
<th>Non-migrant couples</th>
<th>RR</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migrants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Man</td>
<td>26.8± 3.7</td>
<td>1.0± 0.6</td>
<td></td>
<td>26.3</td>
<td>0.000</td>
</tr>
<tr>
<td>Woman</td>
<td>10.8±3.4</td>
<td>5.1±2.5</td>
<td></td>
<td>2.1</td>
<td>0.169</td>
</tr>
<tr>
<td>Non-migrants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Man</td>
<td>18.6±3.9</td>
<td>1.8±1.1</td>
<td></td>
<td>10.5</td>
<td>0.000</td>
</tr>
<tr>
<td>Woman</td>
<td>7.0±3.8</td>
<td>9.3±3.6</td>
<td></td>
<td>0.8</td>
<td>0.657</td>
</tr>
</tbody>
</table>

- Infection from outside is more likely among migrant couples.
12.8.3 RR for Man and Woman

- The ff is the risk ratio for infection comparing migrants and nonmigrants for men and women to be infected by their spouses or from outside the relationship.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Outside</th>
<th>Spouses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man</td>
<td>1.44</td>
<td>0.58</td>
</tr>
<tr>
<td>Woman</td>
<td>1.53</td>
<td>0.54</td>
</tr>
</tbody>
</table>

- Both men and women are likely to be infected from outside and less likely to be infected by their spouses.
- Changing the transmissibility rate from either direction changes the RR estimate by 1.5% in all cases.
12.9 Statistical Model

- Identify important risk factors.
- Model the data at couple level.
- Logistic regression model

```plaintext
proc logistic data=couple descending;
class migration agecat(ref=last) marital agefis;
model hiv=migration agecat marital agefis /risklimits;
run;
```
12.9.1 Results

- Logistic regression models of risk factors for HIV in a couple.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Univariate</th>
<th>Model I</th>
<th>Model II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR</td>
<td>[95%CI]</td>
<td>OR</td>
</tr>
<tr>
<td>Migration status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1=Migrant, 0=Non-migrant)</td>
<td>1.66</td>
<td>[0.93-3.29]</td>
<td>1.75</td>
</tr>
<tr>
<td>Age in years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 to 24</td>
<td>6.69</td>
<td>[1.23-36.31]</td>
<td>3.94</td>
</tr>
<tr>
<td>25 to 34</td>
<td>1.30</td>
<td>[0.62-2.72]</td>
<td>0.89</td>
</tr>
<tr>
<td>35+</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1=Unmarried, 0=Married</td>
<td>1.98</td>
<td>[0.97-4.03]</td>
<td>1.53</td>
</tr>
<tr>
<td>Age at first sexual debut</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1=16 years or less, 0=over 16</td>
<td>2.79</td>
<td>[1.30-5.98]</td>
<td>2.45</td>
</tr>
<tr>
<td>Lifetime partners</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1=More than one, 0=only one</td>
<td>1.78</td>
<td>[0.63-5.08]</td>
<td></td>
</tr>
<tr>
<td>Current sexual partners</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1=More than one, 0=only one</td>
<td>1.14</td>
<td>[0.58-2.27]</td>
<td></td>
</tr>
<tr>
<td>STI symptoms last 4 months</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1=Yes, 0=No</td>
<td>2.23</td>
<td>[1.14-4.36]</td>
<td>1.19</td>
</tr>
</tbody>
</table>

- Migrant couples were more likely than non-migrant
12.9.2 Conclusion

- Migration and HIV: a double edged sword.
- New dimension to the effects of migration.
- Not all partners were infected, even after a long time.
- Raises some interest into microbiological research and genetics.
Chapter 13

New Dimension of Migration

13.1 Migrant Women

- In recent years, female circular migration has increased in South Africa.
- The impact that migration has on the health of female migrants has not been investigated as extensively as it has been for men.
• Most studies have concentrated on the migration of men and the risk that this entails for them and their non-migrant female partners.

• In this case study, we investigate the risk factors for HIV infection in self-identified migrant and non-migrant women in Carletonville, South Africa.
13.2 Study Design

- Cross-sectional community based household survey.
- 13 to 60 years men and women.
- Two-stage random sampling used to select an index household.
- From an index household, a systematic sample of households was taken.
- Focus on women who identified their migration status.
13.3 Descriptive Statistics

- 709 women stated their migration status.
- 150 (21.2%) were migrants.
- 559 (78.8%) were nonmigrants.
- 33.1% of women were born in other urban areas of South Africa, Carletonville district (32%), rural areas of South Africa (26.8%) and other countries (8.1%).
- 40% of migrant women were born in other rural areas of South Africa.
13.4  Note

- The data of migrant and non-migrant women in urban areas has been downloaded onto our computers.
- The data is in SAS format.
- A SAS program that reads-in the data is included.
- The program also includes models that were fitted on these data.
13.5 Task

- Analyze these data.
- Identify important determinants of HIV infection.
- Condom use appears to increase the risk of HIV (strange result!). How do you explain this?
- What conclusions do you make from this study?
- What new knowledge do these results add?
- What policy interventions should be put in place as a result of your findings?
Chapter 14

National population surveys

14.1 Issues Related to Data

- Knowledge of HIV based on antenatal data.
- Data based on some special population groups, e.g. sex workers.
- Lack generalizability.
- Need for more generalizable studies.
14.2 HIV Prevalence Surveys

- HSRC conducted the first ever national HIV prevalence survey in South Africa, 2002.
- Study conducted biennially, due in the field in October 2004.
- The survey is based on a multistage random sampling of individual participants.
14.3 National Enumerator Areas

• Country divided into 85,000 EAs.
• EAs consists of about 200 HH/VPs.
• HSRC randomly sampled a representative sample of 1,000 EAs, stratified by province, geo-type and race.
• EAs and VPs are geo-located and a database of these EAs and their corresponding VPs was developed.
• MASTER SAMPLE, our updated sampling frame.
14.4 Master Sample EAs
14.5 Sampling Techniques

- Nelson Mandela/HSRC national HIV prevalence survey.
- VPs are randomly selected.
- At each VP, individuals are grouped (strata) into 3 different age groups, 2-14; 15-24; 25+.
- Kish grid used to identify respondents.
- Questionnaire administration and blood/saliva samples taken.
14.6 Visiting Points Identification
14.7 National Surveys

- Give better picture of the extent of HIV prevalence.
- Provide more generalizable results.
- Provide statistical methodological challenges.
  - Small Area Estimation techniques,
  - Spatial mapping of the disease;
  - Estimating HIV incidence from panel studies.
- Approach is being extended to other African countries.
Chapter 15

Modeling Forces of Infection Using Beta-Binomial Model

- Non parametric estimation of the force of infection using hierarchical Bayesian model with binomial likelihood and beta prior.
15.1 Bayesian Approach: Formulation of the Model

- Binomial likelihood:
  \[ y_i \sim \text{Bin}(n_i, \pi_i), \quad \text{for} \quad i = 1, 2, \ldots, n, \]

- Product-beta prior:
  \[
P(\pi | \alpha, \beta) \propto \prod_{i=1}^{n} \pi_i^{\alpha_i - 1} (1 - \pi_i)^{\beta_i - 1}, \quad (\alpha_i > 0, \beta_i > 0),
\]

where \( \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n), \beta = (\beta_1, \beta_2, \ldots, \beta_n) \).

- For the unconstrained case
  \[
P(\pi | y, \alpha, \beta) \propto \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \prod_{i=1}^{n} \pi_i^{\alpha_i - 1} (1 - \pi_i)^{\beta_i - 1}
  = \prod_{i=1}^{n} \pi_i^{y_i + \alpha_i - 1} (1 - \pi_i)^{n_i - y_i + \beta_i - 1},
\]
15.2 Estimating Under Order Restrictions

- Recall that the problem is to estimate $\pi$ under the order restrictions, $\pi_1 \leq \pi_2 \leq \ldots \leq \pi_n$.

- The $n$ dimensional parameter vector is constrained to lie in a subset $S^n$ of $\mathbb{R}^n$. The constrained set $S^n$ is determined by the order among the components of $\pi$. 
15.3 Constrained Posterior Distribution

- The posterior distribution of $\pi$ given the constraints is the unconstrained posterior distribution normalized such that

$$P(\pi|y) \propto \frac{P(y|\pi)P(\pi|\alpha, \beta)}{\int_{S^n} P(y|\pi)P(\pi|\alpha, \beta)d\pi}, \quad \pi \in S^n.$$ 

- Let $S^n_j(\pi_j, j \neq i)$ be a cross section of $S^n$ defined by the constraints for the component $\pi_i$ at a specified set of $\pi_j, j \neq i$.

- In our setting, $S^n_j(\pi_j, j \neq i)$ is the interval $[\pi_{i-1}, \pi_{i+1}]$. 
15.4 Constrained Posterior Distribution

- Constrained posterior distribution:
\[
P(\pi_i | y_i, \alpha, \beta, \pi_j, j \neq i) \propto P(y_i | \pi) P(\pi | \alpha, \beta) \quad \pi_i \in S^m_j(\pi_j, j \neq i),
\]
\[
0, \quad \pi_i \notin S^m_j(\pi_j, j \neq i).
\]
- When the likelihood and the prior distribution are combined, the posterior conditional distribution of \(\pi_i | y_i, \alpha, \beta, \pi_j (j \neq i)\) is the standard posterior distribution restricted to \(S^m_j(\pi_j, j \neq i)\).
- The Constrained posterior distribution:
\[
\text{Beta}(y_i + \alpha_i, n_i - y_i + \beta_i) \text{ restricted to the interval } [\pi_{i-1}, \pi_{i+1}]
\]
15.5 Constrained Posterior Distribution

- For the special cases with where \( \alpha = (1, 1, \ldots, 1) \) and \( \beta = (1, 1, \ldots, 1) \) we have
  \[
  y_i \sim \text{Bin}(n_i, \pi_i), \quad \text{for } i = 1, 2, \ldots, n,
  \]
  with the prior
  \[
  \pi_i|y_i, \alpha, \beta, \pi_j, j \neq i \sim \text{U}(\pi_{i-1}, \pi_{i+1}) \quad i = 2, n - 1
  \]
  For \( i = 1 \):
  \[
  \pi_1|y_1, \alpha, \beta, \pi_2 \sim \text{U}(0, \pi_2)
  \]
  For \( i = n \):
  \[
  \pi_2|y_n, \alpha, \beta, \pi_{n-1} \sim \text{U}(\pi_{n-1}, 1)
  \]
  For \( i = n \) and \( f \neq 0 \)
  \[
  \pi_2|y_n, \alpha, \beta, \pi_{n-1} \sim \text{U}(\pi_{n-1}, 1 - f)
  \]
15.6 Constrained Posterior Distribution

- If $\alpha \neq (1, 1, \ldots, 1)$ and $\beta \neq (1, 1, \ldots, 1)$ we can replace the uniform distribution with truncated beta distribution.

- This means that during the MCMC simulation the sampling from the full conditional distribution can be reduced to interval restricted sampling from the standard posterior distribution.
15.7 Posterior Mean for $\ell(x)$

- Once the prevalence values are obtained, the problem of estimating the force of infection becomes straightforward.
- Let $\pi^{(k)}$ be the constrained value of $\theta$, obtained in the $k$’th iteration of the MCMC simulation.
- The force of infection $\ell^{(k)}(x)$ can be estimated by

$$\hat{\ell}^{(k)}(x) = \frac{\hat{\theta}^{(k)}(x)}{(1 - \hat{\theta}^{(k)}(x))}.$$ 

- Since we assume that the force of infection is a smooth function, we smooth $\ell^{(k)}(x)$ with a third order moving average

$$\ell_A^{(k)}(x) = A\ell^{(k)}(x)$$
15.8 Results for Mumps

- Parametric versus nonparametric: the level of the force of infection at the peak.

Figure 15.1: Mumps.
15.9 Results for Rubella

- Parametric versus nonparametric: the age at which the force of infection reach the maximum.

Figure 15.2: *Rubella.*
15.10 Force of Infection at the Maximum (mumps)

- Posterior distribution for the force of infection at age 3.5-6.5.
- Variability.
- Level of the force of infection at the peak.

Figure 15.3: Force of infection at several age groups.
15.11 Influence of $f$

- The fraction of uninfected individuals and the estimated force of infection.

Figure 15.4: Prevalence and force of infection for hepatitis A in Belgium.
15.12 Average Age at Infection

- Posterior means for the average age at infection increase with $f$.

<table>
<thead>
<tr>
<th></th>
<th>$f$</th>
<th>$\bar{f}$</th>
<th>$\bar{A}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rubella</td>
<td>0.00</td>
<td>0.007</td>
<td>9.27</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.016</td>
<td>9.58</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>0.025</td>
<td>9.85</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.033</td>
<td>10.24</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.043</td>
<td>10.63</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.053</td>
<td>11.08</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.063</td>
<td>11.55</td>
</tr>
<tr>
<td>mumps</td>
<td>0.00</td>
<td>0.004</td>
<td>5.29</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.012</td>
<td>5.58</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>0.021</td>
<td>5.99</td>
</tr>
</tbody>
</table>