ANALYSIS OF TIME-TO-EVENT DATA INCLUDING FRAILTY MODELING

By
Belinda Phipson

SUBMITTED IN FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
AT
UNIVERSITY OF KWAZULU-NATAL
PIETERMARITZBURG, SOUTH AFRICA
OCTOBER 2006

Abstract

There are several methods of analysing time-to-event data. These include nonparametric approaches such as Kaplan-Meier estimation and parametric approaches such as regression modeling. Parametric regression modeling involves specifying the distribution of the survival time of the individuals, which are commonly chosen to be either exponential, Weibull, log-normal, log-logistic or gamma distributed. Another well known model that does not require assumptions about the hazard function to be made is the Cox proportional hazards model.

However, there may be deviations from proportional hazards which may be explained by unaccounted random heterogeneity. In the early 1980s, a series of studies showed concern with the possible bias in the estimated treatment effect when important covariates are omitted. Other problems may be encountered with the traditional proportional hazards model when there is a possibility of correlated data, for instance when there is clustering.

A method of handling these types of problems is by making use of frailty modeling. Frailty modeling is a method whereby a random effect is incorporated in the Cox proportional hazards model. While this concept is fairly simple to understand, the method of estimation of the fixed and random effects becomes complicated. Various methods have been explored by several authors, including the Expectation-Maximisation (EM) algorithm, penalized partial likelihood approach, Markov Chain Monte Carlo (MCMC) methods, Monte Carlo EM approach and different methods using Laplace approximation.

The lack of available software is problematic for fitting frailty models. These models are usually computationally extensive and may have long processing times. However, frailty modeling is an important aspect to consider, particularly if the Cox proportional hazards model does not adequately describe the distribution of survival time.

UNIVERSITY OF KWAZULU-NATAL
SCHOOL OF
STATISTICS AND ACTUARIAL SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Science for acceptance a thesis entitled "**Analysis of Time-to-Event Data Including Frailty Modeling**" by **Belinda Phipson** in fulfillment of the requirements for the degree of **Master of Science**.

Dated: October 2006

Supervisor: _____
Dr Henry Mwambi

UNIVERSITY OF KWAZULU-NATAL

Date: **October 2006**

Author: **Belinda Phipson**

Title: **Analysis of Time-to-Event Data Including Frailty Modeling**

Department: **Statistics and Actuarial Science**

Degree: **M.Sc.** Convocation: **April** Year: **2007**

Signature of Author

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Dr Henry Mwambi, my supervisor, for his guidance and constant support during this research. Thank you for motivating me and keeping me working hard.

I would also like to thank Tianxi Cai from Harvard School of Public Health, Carl Lombard and Biddy Buchanan-Lee for supplying me with the data sets.

To everyone who has offered input into this project, Anneke Grobler, Samuel Manda, Kangelani Zuma, Noël Veraverbeke, Paul Janssen, Cathy Connolly, Lize van der Merwe, Immo Kleinschmidt, Linda Haines; thank you for all your suggestions and for answering any queries and questions that I had. To Rebecca Shanmugam, this thesis would not look like this if it wasn't for your help!

Thank you to CAPRISA (Centre for the AIDS Program of Research In South Africa) for the opportunity to do a Masters Fellowship with them for the first year of my Masters, and for financing my studies for that year.

Thank you to the Medical Research Council of South Africa for allowing me a generous amount of time to complete this thesis.

And a special thanks to all my friends and family for your support.

Chapter 1

Introduction to Survival Analysis

1.1 Basic Concepts

In many statistical applications and analyses the variable of interest is often the time that elapses before some event occurs (known as the end-point). The analysis of this type of data is known as survival analysis, and is applicable in many areas such as animal production in agriculture, and medical research, for example, in clinical trials.

A very important and necessary requirement in survival analysis is the precise and unambiguous definition of the time origin and endpoint of a study. The time origin could for example be the diagnosis of the disease, the recruitment of an individual into a study, the commencement of treatment or the onset of AIDS symptoms. The time of entry into a study is not always the same for every patient, which results in what is referred to as staggered entry. The end-point of a study should be defined before the onset, and it includes events such as death, recurrence of symptoms or relief of pain. For each individual the time-to-event is measured from the date of entry to the time of the event. Thus staggered entries do not pose a problem when it comes to the analysis, as all that is required is the total duration of time spent in the study. The time that a patient spends in the study is known as the patient time.

The main reasons why standard statistical procedures are not useful in the analysis

of survival data are because firstly, the survival data are generally not symmetrically distributed, and usually tend to be positively skewed. Thus a symmetrical distribution such as the normal distribution cannot be assumed; more realistic distributions would be, for example, the Weibull, gamma or exponential distribution. Another reason why the normal distribution cannot be assumed is because survival data must be positive. Secondly, the survival times are often censored. This occurs when the end-point of the study has not been observed for some individuals. There are a few reasons as to why an individual has been censored. The study could have reached an end, and an individual may not yet have experienced the event of interest in the study. (For example the patient could still be alive at the end of the study when the event is death.) Another reason that an individual could be censored is that he or she could have been lost due to follow up (this could happen if, say, he or she relocated to another city or area). The only information that the investigator has about the individual is the last time the individual was known to be alive, usually the last time he or she attended the clinic. Another form of censored data is if the individual died from causes unrelated to the treatment. However this may be quite difficult to establish.

The above forms of censoring are known as right censored data, and these patient times are less than the actual survival time. Left censoring occurs when an individual experienced the event before the study commenced. For example, say the interest in a particular study is time to remission, and the first visit is 3 months after the individual is known to be cancer free. If there is evidence of a tumour at this first visit, then the event occurred before the study started, and thus this individual is left censored. Another type of censoring is interval censoring, when the event occurs within an interval of time. A good example of interval censoring is when the event is defined as the first HIV+ test. If an individual is HIV- on the fourth visit, say, and is HIV+ on the fifth visit; then the exact date of seroconversion is only known to be between the two visits, and the individual is interval censored.

Investigators are mainly concerned with right censoring, which can be expressed as

follows. If an individual enters the study at time t_0, and dies at time $t_0 + t$, then t is the uncensored survival time. However if the individual is last known to be alive at time $t_0 + c$, then c is known as the right censored survival time. (The individual may have been lost due to follow up, or may not have experienced the event by the end of the study.) Alternatively, if we define t to be the time to event, and c the time at which censoring occurs, then an individual is right censored if $t > c$ and uncensored if $t \leq c$.

1.2 The Survivor Function and the Hazard Function

Suppose we have a group of patients with survival times $t_1, t_2, \ldots, t_N$, some of which may be censored. These values can be regarded as realizations of the continuous random variable T, which has probability density function $f(t)$, and cumulative distribution function $F(t)$, where $F(t)$ is given by

$$\begin{aligned} F(t) &= P(T < t) \\ &= \int_0^t f(u)\mathrm{d}u. \end{aligned}$$

This represents the probability that the survival time is less than some value t (Collet, 1994). The survival function, which represents the probability that an individual will survive beyond time t, is given by

$$\begin{aligned} S(t) &= P(T \geq t) \\ &= 1 - F(t). \end{aligned} \tag{1.1}$$

Because survival distributions are usually skewed and there may be censored observations, the mean and variance are not used to summarise the distribution of T, but rather the median and quantiles are used. These can be estimated from the survival function. For example the median survival time is that value m of T satisfying $S(m) = 0.5$.

The hazard function is defined as the probability that an individual dies at time t, given that he or she has survived up until that point. It thus measures the instantaneous death

rate for an individual surviving to time t (Collet, 1994). The hazard function is defined by

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T \leq t + \delta t | T \geq t)}{\delta t}. \tag{1.2}$$

By using conditional probability laws, and the mathematical definition of derivatives, the above equation can be rewritten in the following manner

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \frac{P(t \leq T \leq t + \delta t)}{\delta t P(T \geq t)} \\ &= \lim_{\delta t \to 0} \left[\frac{F(t + \delta t) - F(t)}{\delta t} \right] \frac{1}{P(T \geq t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{1.3}$$

Thus the following relationships arise from the above equation, namely,

$$\begin{aligned} h(t) &= -\frac{\mathrm{d}}{\mathrm{d}t} \{\log S(t)\} \\ S(t) &= \exp\left\{ -\int_0^t h(u) \mathrm{d}u \right\} \\ H(t) &= \int_0^t h(u) \mathrm{d}u. \end{aligned} \tag{1.4}$$

The function $H(t)$ is known as the integrated or cumulative hazard function, and is sometimes denoted $\Lambda(t)$. The mathematical relationship between $H(t)$ and $S(t)$ is given below as

$$H(t) = -\log S(t). \tag{1.5}$$

Both the survivor function and the hazard function can be estimated from the given survival data. The methods of estimation can be broadly grouped into parametric and nonparametric methods. Other methods such as the semi-parametric approach (namely the Cox proportional hazards model) have also been developed. These methods will be explored in later chapters.

1.3 Types of Survival Distributions

Survival data are usually right skewed or skewed to the right, thus symmetric models are not useful in analysis. Typical asymmetric distributions are the exponential, Weibull and log-logistic distributions. Only the exponential and Weibull models will be briefly discussed in this section. The aim is to derive some basic relationships when specific survival distributions are adopted.

1.3.1 Exponential Distribution

The exponential distribution is characterised by the following probability density function (p.d.f.)

$$f(t;\lambda) = \lambda e^{-\lambda t}, \qquad t > 0.$$

The cumulative distribution function (c.d.f.) is given by

$$F(t) = 1 - e^{-\lambda t}$$

and the survivor function is

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= e^{-\lambda t}. \end{aligned}$$

The hazard function is thus

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} \\ &= \lambda \qquad \text{(constant)}. \end{aligned}$$

From this it can be seen that the exponential distribution has a constant hazard, which means that the risk of death is independent of time, which is an unrealistic assumption,

because intuitively the risk of death may increase or decrease as an individual ages, for example.

An important property of the exponential distribution is the lack of memory property. Suppose that the random variable T is associated with survival time, and is exponentially distributed with parameter λ. Consider the probability that an individual survives for a time greater than t_1, given that he or she has survived up until time t_0. Then

$$\begin{aligned} P(T > t_1 | T > t_0) &= \frac{P(T > t_1 \text{ and } T > t_0)}{P(T > t_0)} \\ &= \frac{P(T > t_1)}{P(T > t_0)} \\ &= \frac{S(t_1)}{S(t_0)} \\ &= \frac{e^{-\lambda t_1}}{e^{-\lambda t_0}} \\ &= e^{-\lambda(t_1 - t_0)}. \end{aligned}$$

This can be interpreted in the following manner. Given survival to time t_0, the excess life beyond t_0 still has the exponential distribution with parameter λ. This result also explains why the exponential distribution is not a realistic distribution for time-to-event data. However, since the equation is simple and the calculations relatively easy, this model can be appealing in certain circumstances.

1.3.2 Weibull Distribution

The two-parameter p.d.f. for the Weibull function is given by

$$f(t; \gamma, \delta) = \delta \gamma t^{\gamma - 1} e^{-\delta t^{\gamma}}, \quad t > 0.$$

Here γ is known as the shape parameter, and δ is the scale parameter. Note that when $\gamma = 1$ the Weibull distribution reduces to the exponential distribution with parameter δ. The c.d.f. of the Weibull distribution is given by

$$F(t) = 1 - e^{-\delta t^{\gamma}}, \quad t > 0$$

and thus using Eq.(1.1), the corresponding survivor function is

$$S(t) = e^{-\delta t^{\gamma}} \tag{1.6}$$

and hence the hazard function is

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \delta\gamma t^{\gamma-1}. \end{aligned}$$

Clearly for $\gamma \neq 1$, the hazard is not constant, in contrast to the exponential distribution. The hazard function takes a different shape depending on the shape parameter γ, as summarised in Table 1.1.

Table 1.1: Hazard function for different values of γ

Values of γ	Shape of $h(t)$
$0 < \gamma < 1$	Exponential decay
$\gamma = 1$	Constant ($h(t) = \delta$)
$\gamma = 2$	Straight line
$\gamma > 2$	Exponential growth

1.4 Nonparametric Procedures

The advantage of using nonparametric estimation methods is that they do not require assumptions to be made about the underlying distribution of the time-to-event data. Using such methods the survival and hazard function can be estimated, and various hypotheses about them can be tested. What follows in this section is a brief discussion on a few procedures that are based mainly on the Kaplan-Meier (1958) estimation method.

1.4.1 Estimating the Survival Function

A crude estimate of the survival function does not take into account censored observations in its calculations. Because the survival function, $S(t)$, represents the probability that an individual survives for a time greater than t, the function can be estimated by the following expression

$$\tilde{S}(t) = \frac{\text{Number of individuals with survival times} \geq t}{\text{Number of individuals in the sample}}.$$

This is assumed to be constant between two adjacent death times. Plotting $\tilde{S}(t)$ against t results in a step function starting at 1, and thereafter decreasing after every time step corresponding to the distinct death times.

Kaplan-Meier estimate of the survivor function

Suppose there are k survival times, of which m are uncensored and n are censored. Suppose also that there are m_i deaths at time t_i. Then $m_1 + m_2 + \ldots + m_k = m$ where $t_1 \leq t_2 \leq \ldots \leq t_k \leq t$. Suppose r_i patients are at risk during the time interval t_{i-1} to t_i, and at time t_i, m_i of these individuals experience an event considered to be an end-point of the study (for example, death), and c_i individuals are censored. Altshuler's estimate, also derived in Fleming and Harrington (1991), of the survivor function is given by

$$\hat{S}(t) = \exp(-\sum_{i \leq t} \frac{m_i}{r_i}) = \prod_{i \leq t} \exp\left(-\frac{m_i}{r_i}\right).$$

The Kaplan-Meier (1958) estimate is obtained by noting that for large r_i,

$$\exp\left(-\frac{m_i}{r_i}\right) = 1 - \frac{m_i}{r_i} + \frac{m_i^2}{2!r_i^2} - \frac{m_i^3}{3!r_i^3} + \ldots$$

The above series is a Maclaurin series, which is a Taylor series expansion around 0 (Stewart, 1997). Ignoring the higher order terms, the above expression can be simplified to

$$\exp\left(-\frac{m_i}{r_i}\right) \approx 1 - \frac{m_i}{r_i}.$$

Thus the Kaplan-Meier Product Limit estimator is given by

$$\hat{S}_{KM}(t) = \prod_{i \leq t} \frac{r_i - m_i}{r_i} = \prod_{i \leq t} \left(1 - \frac{m_i}{r_i}\right).$$

Life table or actuarial estimate of the survivor function

Given a sample of N individuals, suppose that the survival time is partitioned into I intervals, $i = 0, 1, \ldots, I-1$, where the length of the i^{th} interval is $t_{i+1} - t_i$. For the i^{th} interval let k_i' be the number of individuals entering the i^{th} interval, c_i the number of censored observations in that interval, and m_i the number of failures. The number of individuals entering the $(i+1)^{th}$ interval is then $k_{i+1}' = k_i' - c_i - m_i$.

When there are no censored observations, the estimate of the conditional probability of death, q_i, in the interval $[t_i, t_{i+1})$ would be m_i/k_i'. However if censoring is present this quantity will underestimate q_i (Marubini and Valsecchi, 1995). To define the number at risk during the i^{th} interval assume that on average censoring occurs at the midpoint of the interval. Every individual with censored time corresponds to half an individual alive at the beginning of the interval and exposed to risk for the full interval. Therefore those at risk, denoted by r_i, is obtained by subtracting half of those with censored observations from k_i'. The actuarial estimator of q_i is then

$$\begin{aligned} \hat{q}_i &= \frac{m_i}{r_i} \\ &= \frac{m_i}{k_i' - \frac{1}{2}c_i}. \end{aligned}$$

Once q_i has been estimated, the conditional probability of surviving the i^{th} interval, given by $\hat{p}_i = 1 - \hat{q}_i$, can be estimated. The cumulative probability of surviving all intervals preceding interval i is then given by the product of the conditional probabilities of surviving each interval before i

$$\hat{P}_i = \hat{p}_0 \times \hat{p}_1 \times \ldots \times \hat{p}_{i-1}, \qquad i = 1, \ldots, I.$$

The survival function can thus be constructed in this manner.

1.4.2 Estimating the Hazard Function

The hazard function measures the instantaneous death rate for an individual surviving until time t. Thus an intuitive method of estimating the hazard function is to take the ratio of the number of deaths occurring at a given time to the number of individuals at risk at that time. Assuming the hazard to be constant between successive times at which deaths occur, the hazard per unit time can be found by dividing the number of deaths by the total time individuals are exposed or at risk of death. So if there are m_i deaths and r_i are at risk at time t_i, the hazard function for the interval from t_i to t_{i+1} can be estimated by

$$\hat{h}(t_i) = \frac{m_i}{r_i \tau_i}$$

where $\tau_i = t_{i+1} - t_i$ is the time interval. This estimate is known as the Kaplan-Meier estimate of the hazard function.

1.4.3 Estimating the Cumulative Hazard Function

The cumulative hazard is the integral of the hazard function, and from Eq.(1.5) we know that

$$H(t) = -\log S(t).$$

If $\hat{S}(t)$ is the Kaplan-Meier estimate of the survivor function, then $\hat{H}(t) = -\log \hat{S}(t)$ is a suitable estimate of the cumulative hazard to time t (Collett, 1994). Substituting the Kaplan-Meier estimate for $S(t)$ gives

$$\hat{H}(t) = -\sum_{i=1}^{k} \log\left(\frac{r_i - m_i}{r_i}\right). \tag{1.7}$$

By using the following mathematical result, namely,

$$\begin{aligned} \log(1-x) &= -x - \frac{x^2}{2} + \ldots \\ &\approx -x \end{aligned}$$

and by ignoring higher order terms, Eq.(1.7) can be reduced to

$$\hat{H}(t) \approx \sum_{i=1}^{k} \frac{m_i}{r_i}$$

which is the cumulative sum of the estimated probabilities of death from the first time interval to the k^{th} (Collett, 1994).

1.4.4 Comparison of Two Groups of Survival Data

One common problem in the analysis of time-to-event data or survival time data is that of comparing the experience of two groups. A good example is if one wishes to compare the survival times of patients on treatment to those of patients on a placebo.

The easiest way of comparing two groups of survival data is to plot their survival functions on the same set of axes. However, sometimes it is difficult to draw concrete conclusions from these graphs, and hypothesis tests need to be performed. A variety of parametric and nonparametric significance tests can be used to assess observed differences in empirical survival curves. One common nonparametric test is that based on the log-rank statistic, which has an appealing heuristic derivation described below.

The log-rank test

The log-rank test tests the null hypothesis that the risk or hazard of death is the same in the two groups. In other words, if the study was a clinical trial whereby two different treatments were being tested to see which of the two is more effective, the null hypothesis would be that there is no difference between the effects of the two treatments. The test is described in detail below.

Suppose the 2 groups are denoted by A and B, and that there are k distinct death times, $t_1 < t_2 < \ldots < t_k$, across the two groups. The test uses a conditioning argument based on the numbers at risk of failing just prior to each observed failure time. Suppose that at time t_i there are m_i deaths and r_i at risk in total, with m_{iA} and m_{iB} deaths and r_{iA} and r_{iB} at

risk in groups A and B respectively such that $m_{iA} + m_{iB} = m_i$ and $r_{iA} + r_{iB} = r_i$. At each distinct death time t_i such data can be summarised as in Table 1.2 below.

Table 1.2: Number of deaths at time t_i

Group	No. of deaths	Number survived	Total
A	m_{iA}	$r_{iA} - m_{iA}$	r_{iA}
B	m_{iB}	$r_{iB} - m_{iB}$	r_{iB}
Total	m_i	$r_i - m_i$	r_i

Except for tied survival times, $m_i = 1$, and either m_{iA} or m_{iB} have the values 0 or 1. If a subject is censored at time t_i then that subject is considered to be at risk at that time and is included in r_i. If the null hypothesis is true, then the number of deaths at any time is expected to follow the hypergeometric distribution, and therefore

$$E(m_{iA}) = e_{iA} = \frac{r_{iA} m_i}{r_i}$$
$$Var(m_{iA}) = \frac{m_i(r_i - m_i) r_{iA} r_{iB}}{r_i^2 (r_i - 1)}.$$

The difference between m_{iA} and e_{iA} is the basis for the test statistic for testing the null hypothesis. The log-rank test is the combination of these differences over all death times. Summing the various measures over the death times gives

$$\begin{aligned} O_A &= \sum_i m_{iA} \\ E_A &= \sum_i e_{iA} \\ V_A &= \sum_i Var(m_{iA}) \end{aligned}$$

where E_A can be seen as the expected number of deaths occurring in Group A over the entire time period. The test statistic is then given by

$$\chi_1^2 = \frac{(O_A - E_A)^2}{V_A}$$

which, under H_0, is χ^2 distributed with 1 degree of freedom. If the calculated value is larger than the table value corresponding to the χ^2 distribution at a significance level of α, then the null hypothesis of no group differences is rejected and one can conclude that the risk of death is different in the two groups.

Alternatively, assuming the deviations $m_{iA} - e_{iA}$, $i = 1, \ldots, k$, are independent,

$$Q = \frac{O_A - E_A}{\sqrt{V_A}}$$

should have an approximately standard normal distribution, and the null hypothesis is rejected for large values of Q.

The ratios $\frac{O_A}{E_A}$ and $\frac{O_B}{E_B}$ are referred to as the relative death rates and they measure the ratio of the death rate in each group to the death rate among both groups combined. The ratio of these two relative rates estimates the death rate in group A relative to the death rate in group B.

The log-rank test can be generalised to test for s treatment group differences. The test statistic, with $(s-1)$ degrees of freedom, would then be given by

$$\chi^2_{s-1} = \frac{(O_A - E_A)^2}{V_A} + \frac{(O_B - E_B)^2}{V_B} + \frac{(O_C - E_C)^2}{V_C} + \ldots$$

If the calculated value exceeds the table value at α significance level, we reject the null hypothesis of no group differences in survival times.

Some important remarks in the derivation of the log-rank statistic are stated below. First the vector of observed-minus-expected failures does not in fact have independent components and the central limit theorem usually applied to prove asymptotic normality fails. Further still, differences between observed and expected failures are given equal weight, regardless of the risk set (namely numbers of cases still under observation) at observed failure times. Such weighting will have implications for the operating function of Q. These more delicate aspects of significance tests are studied in the book on counting processes and survival analysis by Fleming and Harrington (1991).

Chapter 2

Parametric Regression Modeling of Survival Data

2.1 Introduction

In most medical studies additional data such as age and sex are recorded on each individual at the beginning and sometimes at the end of the study. It is reasonable to assume that certain variables may have an impact on the survival experience of a patient, and thus in order to incorporate them into the analysis the use of statistical modeling is necessary.

Regression modeling of the relationship between an outcome variable and independent explanatory variable is a common approach because biologically plausible models can be easily fitted, evaluated and interpreted. However, specification of the model requires choosing systematic and error components that are relevant to the problem at hand (Hosmer and Lemeshow, 1999). Choosing the systematic component involves assessing the relationship between an "average" of the dependent variable and the independent variables, which can be based on an exploratory analysis of the data, and past experience. The choice of an error component involves specifying the statistical distribution of what remains to be explained after the model has been fitted, namely the residuals (Hosmer and Lemeshow, 1999).

Model selection can be based on the data that is being analysed. The general starting

point is to use a model with a linear systematic component and normally distributed errors, in other words, the usual linear regression model. Suppose though, that the response variable is a dichotomous variable (taking on only two values, for example 1 = Yes, 0 = No) and measures of risk such as odds-ratios need to be calculated. In this case the logistic regression model is the obvious choice. This has a systematic component that is linear in the log-odds and has binomial or Bernoulli distributed errors. For general non-Gaussian distributed errors the theory of generalised models is used (McCullagh and Nelder, 1989).

These two models are commonly used for most data-based approaches. However, when it comes to time-to-event data, or survival time data, they are not particularly useful. One of the reasons for this is that some of the survival times may only be partially observed as observations may be censored. An important goal of statistical analysis is to yield estimates that are easily interpreted and results that are statistically plausible. The first step in any analysis is to a conduct an exploratory analysis, or univariate analysis of the data to obtain a clear sense of the distributional characteristics of our outcome variable as well as all possible predictor variables (Hosmer and Lemeshow, 1999). However, the outcome variable, which is the survival time, may be incomplete for censored observations, which poses a problem for conventional univariate measures such as the mean, standard deviation and median. If the censored observations are treated as survival times, then the resulting sample statistics are not estimates of the survival time distribution. Rather they are estimators of a combination of the survival time distribution and that of a second distribution that depends on survival time as well as statistical assumptions about the censoring mechanism (Hosmer and Lemeshow, 1999). Thus a numeric value for the mean is not interpreted as the average survival time, but as the survival time being at least that value.

A scatterplot of the data must also be carefully interpreted when the data is of the survival analysis type. Generally, a scatterplot can confirm whether the straight line

assumption is valid as well as whether there are any outliers that may affect the validity of a linear regression model. If the censored observations are ignored, the same problem of whether the arithmetic mean is the "true" mean would still persist. The straight line would not tell us that a point on that line is the mean at that point; only that the mean is at least as large as that point(Hosmer and Lemeshow, 1999). However, a scatterplot is still a useful tool with censored survival time data, especially if there is a single continuous variable plotted against log$T$, although it becomes more difficult to interpret when there are many covariates to be considered.

2.2 Exponential Model

In a linear regression model the basic shape of the scatterplot is controlled by the nature and strength of the relationship between the outcome and independent variables and the fact that the residuals follow a normal distribution (Hosmer and Lemeshow, 1999). With survival data the shape of the plot is also determined by the nature of the systematic relationship between the natural log of the time variable and the covariate; but the distribution of errors is typically skewed to the right (Hosmer and Lemeshow, 1999). A simple statistical distribution with this characteristic is the exponential distribution. Also, the dependent variable, which is time, is strictly positive. For these reasons, many of the curves in scatterplots of survival data can be described by an equation of the form $t = e^{-x}$. The combination of an exponential systematic component and exponentially distributed errors suggests an exponential regression model as a starting point (Hosmer and Lemeshow, 1999). Assuming only one covariate, this model can be expressed as

$$T = e^{\beta_0+\beta_1 x} \times \varepsilon \tag{2.1}$$

where T denotes the survival time, and ε is exponentially distributed with parameter equal to 1. This model is not linear in its parameters, but can be linearised by taking the natural log on both sides of Eq.(2.1). This results in the equation

$$Y = \beta_0 + \beta_1 x + \theta \tag{2.2}$$

where $Y = \ln T$ and $\theta = \ln \varepsilon$. This model was termed as the location-scale model for $\ln T$ by Lawless (1982). The above equation looks like the usual linear regression model, but in this case the errors, θ, are not normally distributed and have a distribution independent of x (Lawless, 1982). These errors follow the standard extreme minimum value distribution, often referred to as the Gumbel distribution or double exponential distribution (Allison, 1995). The mode of this distribution is 0, its scale parameter is 1, and the model is denoted Gumbel(0,1). The extreme minimum value distribution is derived by considering the minimum value from a simple random sample of observations. As the sample size increases the distribution of the minimum value can be shown to tend towards Gumbel(0,1) (Hosmer and Lemeshow, 1999). Using this distribution is analogous to using the standard normal distribution, however, practically the errors generally do not have unit variance, but are assumed to be constant. An additional parameter may be introduced into Eq.(2.2) by multiplying θ by σ. Then Eq.(2.2) becomes

$$y = \beta_0 + \beta_1 x + \sigma \times \theta. \tag{2.3}$$

The distribution of $\sigma \times \theta$ is denoted Gumbel(0, σ). The location-scale model can be applied to exponential, Weibull, log-normal and generalised gamma models (Lawless, 1982).

The Gumbel distribution has p.d.f. $f(\varepsilon) = \exp(\varepsilon - \exp(\varepsilon))$. This is a unimodal distribution defined on the real line, and is slightly skewed to the left (Allison, 1995). If the data follows the exponential model then this implies that the hazard is constant (proven in the previous chapter). The departure from this assumption can be tested in SAS, which

includes a Lagrange multiplier statistic in the output, which is χ^2 distributed. This statistic tests the null hypothesis that the hazard is constant over time, and thus it is possible to assess whether the exponential model is suitable for the particular problem or not.

2.3 Log-Normal Model

Consider a generalisation of Eq.(2.3) of the form

$$\ln T_i \quad = \quad \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \sigma\varepsilon_i \tag{2.4}$$

where $x_{i1}, \ldots, x_{ik}, i = 1, \ldots, n$ are the values of the k covariates, ε_i is the random error term, and the β_j's and σ are the parameters to be estimated. Exponentiating the above equation on both sides results in the following equation

$$T_i \quad = \quad \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \sigma\varepsilon_i).$$

The log transformation of T is to ensure that the predicted values of T are positive, regardless of the values of $\mathbf{x}$ and $\boldsymbol{\beta}$ (Allison, 1995).

The log-normal model implies that T has a log-normal distribution. This comes about from the assumption that ε_i has a normal distribution, which means that $\log T$ is normally distributed. Further, ε_i has a mean and variance that is constant over i, and the ε's are independent across observations (Allison, 1995). The hazard for the log-normal distribution has an inverted U-shape in that it starts at 0 when $t = 0$, rises to a peak and then declines to 0 as $t \to \infty$.

2.4 Weibull Model

This model is one step up from the Exponential model in that ε is still assumed to have a Gumbel distribution, however with the exception that σ can vary. In Table 2.1 below the shape of the hazard function, which depends on the value of σ, is described.

Table 2.1: Shape of the hazard function according to the shape parameter

σ	Shape of hazard function
$\sigma > 1$	decrease with time
$0.5 < \sigma < 1$	increasing at a decreasing rate
$0 < \sigma < 0.5$	increasing at an increasing rate
$\sigma = 0.5$	straight line, origin at 0

2.4.1 Assessing the Suitability of a Weibull Model

The assumption that the survival data has a Weibull distribution can be tested by an appropriate method described below. The survival function associated with the Weibull distribution, given in Eq.(1.6), can be manipulated to give the following relationships.

$$\begin{aligned} -\log S(t) &= \delta t^{\gamma} \\ \log[-\log S(t)] &= \log\delta + \gamma \log t \end{aligned}$$

A plot of $\log[-\log S(t)]$, also known as the log cumulative hazard, against $\log t$ should result in a straight line graph with slope γ and intercept $\log\delta$ if a Weibull model is appropriate. In order to apply this method for assessing the suitability of a Weibull model, the survivor function needs to be estimated.

2.5 Log-Logistic Model

The log-logistic model assumes that ε has a logistic distribution, implying that $\ln T$ has a logistic distribution. Thus from the transformation, T is log-logistic distributed. In this case, the p.d.f. of ε is given by

$$f(\varepsilon) = \frac{e^{\varepsilon}}{(1+e^{\varepsilon})^2}.$$

The shape of the hazard function varies according to the shape parameter, σ. When $\sigma < 1$ the hazard has a shape similar to the log-normal hazard, namely, it starts at 0 when $t = 0$,

rises to a peak and then declines to 0 as $t \to \infty$. For $\sigma > 1$ the hazard starts at infinity and declines to 0 as t increases, and when $\sigma = 1$ the hazard takes a value of λ_0 when $t = 0$, and declines to 0 as $t \to \infty$.

Recall that $S(t)$ is the probability of surviving to time t. The quantity $\frac{S(t)}{1-S(t)}$ is the odds of surviving to time t. Thus the log-logistic model is a member of a general class of models known as proportional odds models. The assumption for these models is that

$$\frac{S_i(t)}{1 - S_i(t)} = \phi_{ij} \left(\frac{S_j(t)}{1 - S_j(t)} \right), \qquad \forall t$$

where ϕ_{ij} is some constant that is specific to the pair of observations for individual i and individual j.

2.6 Generalised Gamma Model

There are two types of gamma models. These are the standard 2-parameter gamma model, and the generalised 3-parameter model. The advantage of using the gamma model is that the hazard function can take on a variety of shapes, rendering it a less restrictive model to use. The exponential, Weibull, standard gamma and log-normal models all become special cases of the generalised gamma model, a fact which can be used in assessing the models to find the one of best fit (Allison, 1995). The p.d.f. of the generalised gamma model is given as

$$f(t) = \frac{\lambda\beta(\lambda t)^{k\beta-1}e^{-(\lambda t)^{\beta}}}{\Gamma(k)}. \tag{2.5}$$

When $\beta = 1$ the resulting distribution is the standard 2-parameter gamma distribution, when $k = 1$ Eq.(2.5) becomes the p.d.f. of the Weibull distribution, when $\beta = 1$ and $k = 1$ it reduces to the exponential distribution, and when $k \to \infty$ the log-normal distribution arises.

The reason why the gamma model is not always used is that the formula for the hazard function, $h(t)$, is complicated and it may be difficult to assess the shape from the estimates.

In addition the model is difficult and computationally extensive, and can thus take much longer to compute, especially in the case of large data sets. There may also be convergence problems associated with it.

A good summary of the assumptions on ε and the resulting distribution of T is given in Table 2.2 below.

Table 2.2: Summary of distribution of ε and T

Distribution of ε	Distribution of T
Extreme value (2 parameter)	Weibull
Extreme value (1 parameter)	Exponential
Log-gamma	Gamma
Logistic	Log-logistic
Normal	Log-normal

2.7 Model Fitting

When some of the observations are censored, fitting the above models becomes problematic. For normal linear regression, the least squares method of estimation is often employed. This method has advantages in that it yields estimates that are asymptotically normally distributed with obtainable variances and covariances, and the t-distribution and F-distribution can be used to test certain hypotheses concerning individual parameters and overall model significance respectively. For data that has censored observations, the method of Maximum Likelihood Estimation (M.L.E.) and the methodology associated with large sample procedures can be adapted to fit the above models (Lawless, 1982). This allows us to test hypotheses and form confidence intervals for individual parameters, and to assess overall model significance with relative ease (Hosmer and Lemeshow, 1999).

Assume that continued observation of a subject is controlled by two completely independent time processes - actual survival time associated with the disease of interest, and

the length of time until a subject is lost to follow-up. Assume that both of these processes are under observation and that the recorded time represents the time to the event which occurred first. Two variables are used to characterise a subject's survival time; the actual observed time, T, and a censoring indicator variable, C. When $C = 1$, T measures the actual survival time and when $C = 0$, T measures the time until follow-up ends for reasons other than death from the disease of interest. These values, and a value for a measured covariate X are denoted by lower case letters in the triplet (t, c, x), where t is the observed survival time, c is the value of the indicator variable, and x denotes the value of the covariate of interest.

The likelihood function yields a quantity similar to the probability of the observed data under the model, and can be derived in order to obtain estimates of the unknown parameters. Suppose that the distribution of survival time for a subject with covariate x and who has the disease of interest can be described by the cumulative distribution function (c.d.f.), denoted by $F(t, x, \beta)$. This c.d.f. can be interpreted as the proportion of individuals with covariate x expected to die in less than t time units. Recall that the survivorship function is given by $S(t, x, \beta) = 1 - F(t, x, \beta)$. This can be interpreted as the probability that an individual with covariate x will survive to at least t time units. Also the probability that the survival time for an individual with covariate x is exactly t is given by the p.d.f. $f(t, x, \beta)$, corresponding to $F(t, x, \beta)$.

In the construction of the likelihood function the contribution of the triplets $(t, 1, x)$ and $(t, 0, x)$ are considered separately. For the triplet $(t, 1, x)$ it is known that the contribution of survival time is exactly t time units, which is given by $f(t, x, \beta)$. For censored individuals with triplets in the form $(t, 0, x)$, the contribution to survival time is known to be at least t time units, which is given by the survivor function, $S(t, x, \beta)$. Thus assuming that the n

observations are independent, the likelihood function is given by

$$L(\beta) = \prod_{i=1}^{n}\{[f(t_i, x_i, \beta)]^{c_i} \times [S(t_i, x_i, \beta)]^{1-c_i}\}. \quad (2.6)$$

In general, it is easier to maximise the log-likelihood, so taking the natural log on both sides of Eq.(2.6) yields

$$\ell(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n}\{c_i \ln(f(t_i, x_i, \beta)) + (1 - c_i) \ln(S(t_i, x_i, \beta))\}. \quad (2.7)$$

In order to obtain estimated values of the unknown parameters, Eq.(2.7) is differentiated with respect to the unknown parameters, set equal to 0, and then solved for β. If these equations are non-linear in the unknown parameters to be estimated, then iterative techniques such as Newton-Raphson and Fisher Scoring algorithms can be used.

2.8 Choosing the Best Model to Fit the Data

The likelihood-ratio statistic can be used to compare nested models, where the first model is a special case of the second model. An example of this is when restrictions are placed on the parameters of one model to produce another. The likelihood-ratio statistic is χ^2 distributed under the null hypothesis that the two models are equal. The test statistic can be simply stated as

$$G = -2[LL_A - LL_B]$$

where LL_A and LL_B are the log-likelihoods of the models A and B respectively; where model A is nested within model B. The degrees of freedom associated with G is $df_B - df_A$, where df_B is equal to the number of parameters in model B, and df_A is equal to the number of parameters in model A.

In the case of time-to-event data, the exponential model is nested within the Weibull and standard gamma models. Thus clearly the exponential, Weibull, standard gamma and log-normal models are all nested within the generalised gamma model by imposing restrictions

on certain parameters. The likelihood-ratio test thus tests the null hypothesis that the particular restriction is true.

Chapter 3

Parametric Data Analysis

3.1 Introduction

In this chapter three different sets of time-to-event data, which were obtained from various sources are briefly described then analysed using the parametric method of analysis. Each data set presents a specific feature of its own hence providing a wide understanding of the methods. All analysis in this chapter was done in SAS. The survival functions and hazard functions for the various data sets were done using proc lifetest, and the parametric regression models were fitted using proc lifereg where the distribution was specified in the options of the model statement.

3.2 Old Order Amish Community Data

This data was obtained from the Harvard School of Public Health. It is time-to-event data clustered by family, and the data set was used in a publication on the comparison of methods for survival analysis of dependent data (King *et al.*, 1996). The data is from the Fisher family history, a genealogy of the old order Amish community based in Lancaster County, PA. The population was a closed and stable one, and had maintained extensive genealogical records.

The data set consists of 2860 individuals in 458 sibships (where individuals with the

same sibship index are siblings, that is, brother and sister). There were 594 observations which were censored (20.8 %). Gender and birth year are the only covariates available in the data. Table 3.1 shows the distribution of gender in the data.

Table 3.1: Distribution of sex in Amish data

Sex	Frequency	Percent(%)
Male	1480	51.75
Female	1380	48.25
Total	2860	100

The sibship size ranged from 1 to 20 with an average sibship size of 6.24. Table 3.2 describes how many families had a certain sibship size, for example, 38 families have a sibship size of 5.

Table 3.2: Sibship size for the families

Sibship size	1	2	3	4	5	6	7	8	9	10
Frequency	32	44	46	45	38	42	43	46	35	28
Sibship size	11	12	13	14	15	16	17	18	19	20
Frequency	30	11	10	2	4	1	-	-	-	1

The variables in the data set are

- id (from 1 to 2860 individuals)

- age (age at death in years)

- dlt (event indicator, 1=death; 0=censored)

- sib (sibship indicator, from 1 to 458)

- sex (gender, 1=Male; 2=Female)

- byr (birth year)

Looking at a univariate analysis of the response variable, age, it was found that the mean was 52.899 years (with a standard deviation of 31 years), and the median was 64 years. The variable age ranged from 0 to 108 years at death. The birth year ranged from 1801 to 1921.

3.2.1 The Survival Curve and Hazard function

Using proc lifetest in SAS, the survival curve and hazard function for the Amish family data was estimated. The method chosen for estimating these curves was the life-table or the actuarial method (Lee, 1992). It is also possible to choose the Kaplan Meier method of estimation in the options proceeding the proc lifetest statement.

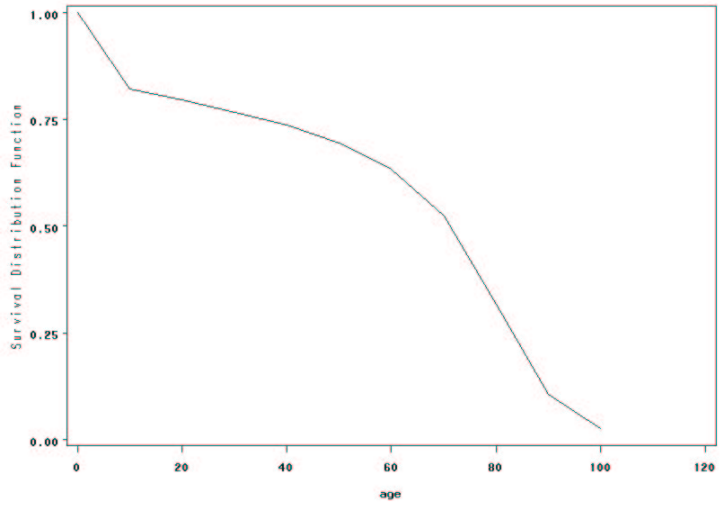

Figure 3.1: Survival function for Amish family data set

The survival function is clearly decreasing with age. There is a steep decrease between ages 0 and 10, indicating a high child mortality. Thereafter it decreases less rapidly until about age 70, where it decreases slightly faster until age 100.

The hazard function shown in Figure 3.2 shows a very slowly increasing hazard from around age 15 to about age 60. It then increases very rapidly, indicating that a person's

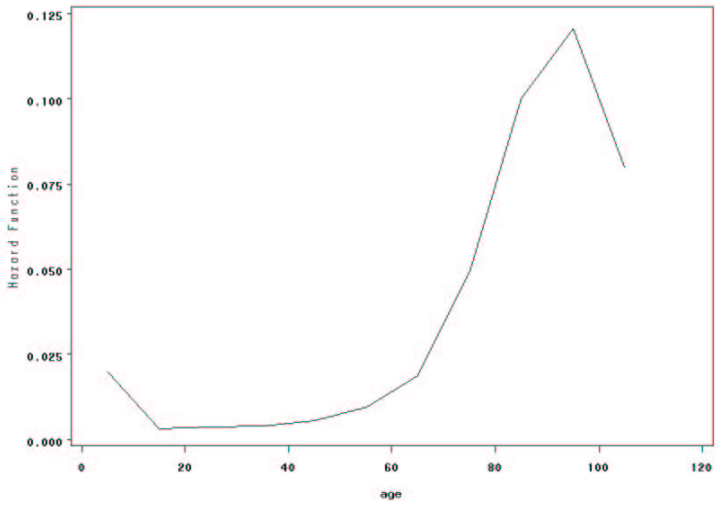

Figure 3.2: Hazard function for Amish family data set

chance of dying is highly increased after age 60.

Using proc lifereg in SAS, various parametric regression models were fitted and then compared to determine the best-fitting model. The 5 models described in the previous chapters were fitted and later compared. This is by no means exhaustive, other models could also be fitted.

3.2.2 The Log-Normal Model

The log-normal model has the assumptions that the errors for an individual time-to-event observation follow a normal distribution with a mean and variance that is constant over that individual, and that the errors are independent across observations (Allison, 1995). The log-normal model is part of a very general group of models known as the accelerated failure time (AFT) models. In its most general form, the AFT model describes a relationship between the survivor functions of any two individuals and implies that what makes any two individuals different from each other is the rate at which they age (Allison, 1995).

The clustering variable, sib, has not been included in any of the parametric models.

Intracluster dependence is accounted for elsewhere when applying the Cox proportional hazards model (Cox, 1972). The model fitted was

$$age = \exp(\beta_0 + \beta_1 x_{sex} + \beta_2 x_{byr} + error). \tag{3.1}$$

The number of observations used in the analysis was 2602, with 594 censored values. The algorithm was said to have converged. The type III analysis of effects are summarised in Table 3.3.

Table 3.3: Type III analysis of effects for the log-normal model

Effect	DF	Wald χ^2	Pr $> \chi^2$
sex	1	3.3470	0.0673
byr	1	45.5621	<0.0001

Birth year is a significant variable in the regression, whereas sex is just insignificant at a 5% significance level. The intercept was found to be significant with a p-value of < 0.0001. The analysis of parameter estimates are presented in the Table 3.4. The estimates can be

Table 3.4: Parameter estimates for the log-normal model

Parameter	DF	Estimate	Std Error	χ^2	Pr $> \chi^2$	e^{β}
Intercept	1	-10.2385	2.1157	23.42	<0.0001	-
sex	1	-0.0917	0.0501	3.35	0.0673	0.9124
byr	1	0.0076	0.0011	45.56	<0.0001	1.0076
scale	1	1.2169	0.0194	-	-	-

interpreted in the following manner. For a categorical variable such as sex one examines the transformed parameter $e^{\beta} = e^{-0.0917} = 0.9124$. The reference category is female. This can be interpreted in much the same way as an odds ratio, that the expected death time for males is approximately 0.9 that of females, holding all other variables constant. This intuitively makes sense; it implies that women live longer than men. For a quantitative variable such as the birth year, one examines the expression $100(e^{\beta} - 1) = 0.76$. This means

that every additional year later a person was born is associated with a 0.76% increase in expected time to death, holding all other variables constant. This implies that the more recently you are born, the longer you will live. This makes sense as one may assume health care and awareness are better in more modern times than in the 1800's say. The log-likelihood is -3619.8282 which is a useful measure in comparing nested models in order to find a model of best fit.

3.2.3 The Exponential Model

The exponential model specifies that the errors have an extreme-value or Gumbel distribution. It is a unimodal distribution which is skewed to the left.

The model fitted in SAS is the same as in Eq.(3.1). Again, 2602 observations were analysed, and 594 were reported censored. The type III analysis of effects are given in Table 3.5. In this case the covariate sex is clearly not significant, while birth year is significant at

Table 3.5: Type III analysis of effects for the exponential model

Effect	DF	Wald χ^2	Pr $> \chi^2$
sex	1	0.8566	0.3547
byr	1	79.0952	<0.0001

a 5% significance level. The algorithm converged and the log-likelihood was -3183.831466. The specific parameter estimates are reflected in Table 3.6.

Table 3.6: Parameter estimates for the exponential model

Parameter	DF	Estimate	Std Error	χ^2	Pr $> \chi^2$
Intercept	1	-10.7870	1.6991	40.31	<0.0001
sex	1	-0.0413	0.0447	0.86	0.3547
byr	1	0.0081	0.0009	79.10	<0.0001
scale	0	1.0000	0.0000	-	-
weibull shape	0	1.0000	0.0000	-	-

What is also given in the SAS output is a Lagrange multiplier statistic for scale, which is a χ^2 statistic. The value is given as 3709.6236, with an associated probability of <0.0001. This statistic tests the null hypothesis that the hazard is constant over time or that the scale parameter, σ, is equal to one. In this case the hypothesis is clearly rejected, suggesting that the hazard is not constant over time, and thus that the exponential model is not a suitable model to use.

3.2.4 The Weibull Model

The Weibull model is a modification of the exponential model. The errors are still extreme-value distributed, but the hazard is not constant over time, in other words, the assumption that $\sigma = 1$ is relaxed. Depending on the value of σ, the hazard may either increase or decrease. The Weibull model is a popular model since it has an easy form for the survivor function, it is an accelerated failure time model, and it is also a proportional hazards model (which means that the coefficients can be interpreted as relative hazard ratios).

The model fitted in SAS is the same as in Eq.(3.1). Again, 2602 observations were analysed, and 594 were reported censored. The type III analysis of effects are given in Table 3.7.

Table 3.7: Type III analysis of effects for Weibull model

Effect	DF	Wald χ^2	Pr > χ^2
sex	1	0.9578	0.3277
byr	1	74.9111	<0.0001

Once again, sex is not a significant variable, but birth year is, at a 5% significance level. The algorithm was reported to have converged. The analysis of parameter estimates are reflected in Table 3.8

Since $\hat{\sigma}$ is between 0.5 and 1.0, the hazard is increasing at a decreasing rate. The coefficient for log$t$ in the log-hazard model is given by $(1/\sigma) - 1 = 0.7519$. This can be

Table 3.8: Parameter estimates for Weibull model

Parameter	DF	Estimate	Std Error	χ^2	Pr > χ^2	e^{β}
Intercept	1	-4.1993	0.9813	18.31	<0.0001	-
sex	1	-0.0250	0.0255	0.96	0.3277	0.9753
byr	1	0.0045	0.0005	74.91	<0.0001	1.0045
scale	1	0.5708	0.0115	-	-	-
Weibull shape	1	1.7518	0.0354	-	-	-

interpreted to mean that a 1% increase in the time since birth produces a 0.7519% increase in the hazard for death. The log-likelihood was -2869.19018. The hazard ratio for sex is 0.9753, which means that the expected time to death for males is 0.9753 that of females. For birth year, $100(e^{\beta} - 1) = 0.45$, implying that every additional year later a person is born is associated with a 0.45% increase in expected time to death; which is lower than the log-normal model estimates.

3.2.5 The Log-Logistic Model

The log-logistic model assumes that the errors have a logistic distribution. Sex is not a significant variable in the analysis, but birth year is highly significant at a 5% significance level.

Table 3.9: Type III analysis of effects for the log-logistic model

Effect	DF	Wald χ^2	Pr > χ^2
sex	1	0.7830	0.3762
byr	1	42.2548	<0.0001

When $\sigma < 1$, as is the case here ($\sigma = 0.5285$), the log-logistic distribution hazard is similar to the log-normal (or AFT) hazard in that it starts at zero, rises to a peak, and then declines to zero. When $\sigma > 1$ the hazard behaves like the decreasing Weibull hazard. It starts at infinity and declines to zero and when $\sigma = 1$ the hazard takes a value λ_0 at

Table 3.10: Parameter estimates for the log-logistic model

Parameter	DF	Estimate	Std Error	χ^2	Pr > χ^2
Intercept	1	-5.5311	1.4924	13.74	0.0002
sex	1	-0.0312	0.0353	0.78	0.3762
byr	1	0.0052	0.0008	42.25	<0.0001
scale	1	0.5285	0.0107	-	-

time $t = 0$ and declines to zero as t goes to infinity (Allison, 1995). The log-likelihood was -3305.325522.

3.2.6 The Gamma Model

Recall that there are two gamma models which can be fitted, the standard 2-parameter model and the generalised 3-parameter model. Proc lifereg in SAS fits the generalised model, which allows the hazard function to take on a variety of shapes by including an extra parameter. When fitting the generalised gamma model to the data, problems of convergence and long computation times were not experienced. The standard gamma model is not simple to fit in proc lifereg and involves finding a common value for the scale and shape parameters that maximise the likelihood.

Table 3.11: Type III analysis of effects for the gamma model

Effect	DF	Wald χ^2	Pr > χ^2
sex	1	15.1242	0.0001
byr	1	117.1372	<0.0001

The main difference between the output of this model and the other models is that now sex is a significant covariate. The log-likelihood is also smaller than for the other models with a value of -2294.33199.

In SAS, due to the reparameterization of the generalised gamma p.d.f., when the shape parameter is 0, the result is the log-normal distribution, and when it is 1 it becomes the

Table 3.12: Parameter estimates for the gamma model

Parameter	DF	Estimate	Std Error	χ^2	Pr $> \chi^2$
Intercept	1	1.0748	0.3178	11.43	0.0007
sex	1	-0.0414	0.0107	15.12	0.0001
byr	1	0.0018	0.0002	117.14	<0.0001
scale	1	0.1674	0.0096	-	-
shape	1	4.4144	0.2663	-	-

weibull distribution. Obviously here this is not the case, since the shape parameter is closer to 4.5. The shape and scale parameters are very different, thus a generalised gamma model is better suited to the data, and it is not necessary to fit a standard gamma model.

3.2.7 Choosing the Best Model to Fit the Data

The log-normal, exponential, Weibull and log-logistic model all gave the same conclusion in that birth year was significant and gender was not. However, the generalised gamma model found gender to be a significant variable in the model as well. Thus it is important to find a model which best fits the data given that there may be different conclusions as to which variables are important.

Recall that the test statistic for testing which model best fits the data is given by

$$G = -2[LL_A - LL_B].$$

The log-likelihoods associated with the various fitted models are reflected in Table 3.13 below. Unfortunately it is not possible to evaluate the log-logistic model against the other models, as it is not a nested model of any of the other fitted models. By inspection it appears that the generalised gamma model may be a better fit than any other since it's log-likelihood is by far the smallest in absolute magnitude. The actual test statistics and their associated probabilities are presented in Table 3.14. Clearly the generalised gamma model is far superior to any other models compared to it and thus is probably the best

Table 3.13: Log-likelihoods for the fitted models

Model	Log-Likelihood
Log-Normal (AFT)	-3619.8282
Exponential	-3183.831466
Weibull	-2869.19018
Log-Logistic	-3305.325522
Generalised Gamma	-2294.33199

Table 3.14: Deviance for comparisons of various models

Comparison	G	$\Pr > \chi^2$
Exp vs. Weibull	629.282	< 0.0001
Exp vs. G.Gamma	1778.999	< 0.0001
Weibull vs. G.Gamma	1149.7164	< 0.0001
Log-normal vs. G.Gamma	2650.994	< 0.0001

model to use in this instance.

Graphical method of testing for linearity

Another method of testing for the best model is a graphical method. If a distribution is assumed to be exponential, say, then a plot of $-\log \hat{S}(t)$ against t should be a straight line with an origin at 0.

Clearly the graph in Figure 3.3 is not linear, and thus the exponential model is not suitable. Similarly, to test for linearity in the Weibull model, a graph of $\log[-\log \hat{S}(t)]$ against $\log t$ is plotted. This is shown in Figure 3.4. Once again, this graph is meant to be a straight line, and clearly it is not, thus the Weibull model is not suitable. It is possible to use a similar approach to evaluate the log-logistic model and log-normal model by plotting $\log\left[\frac{1-\hat{S}(t)}{\hat{S}(t)}\right]$ against $\log t$, and $\Phi^{-1}[1-\hat{S}(t)]$ against $\log t$ respectively; where $\Phi(\cdot)$ is the c.d.f. of a standard normal variable and Φ^{-1} is the inverse. It is clear from the plots in Figures 3.5 and 3.6 that the log-normal model and the log-logistic model are also not suitable.

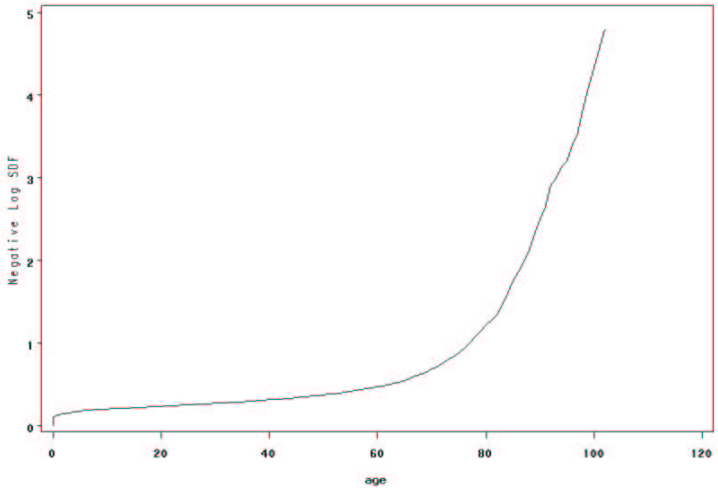

Figure 3.3: Log-survivor plot for Amish family data set

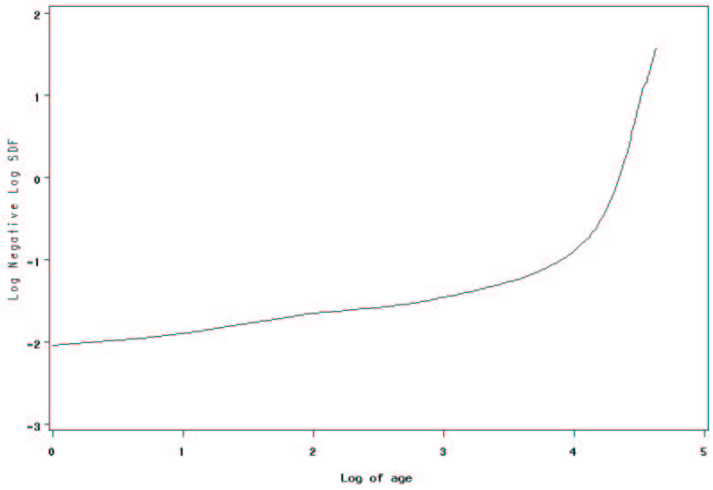

Figure 3.4: Log-log survivor plot for Amish family data set

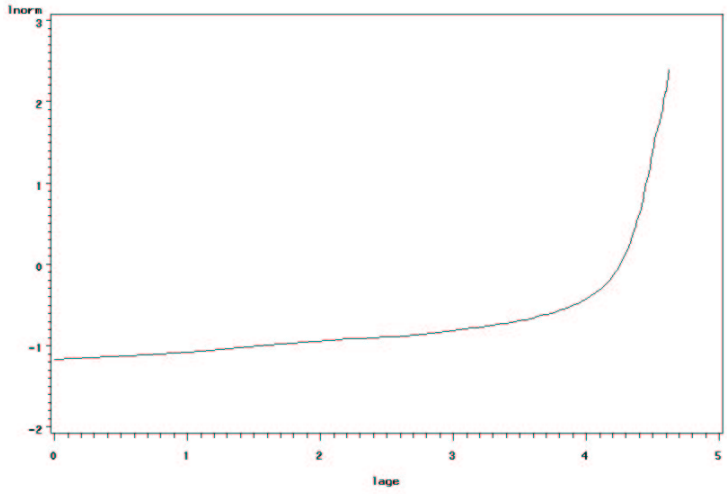

Figure 3.5: Plot for evaluating log-normal model for Amish family data set

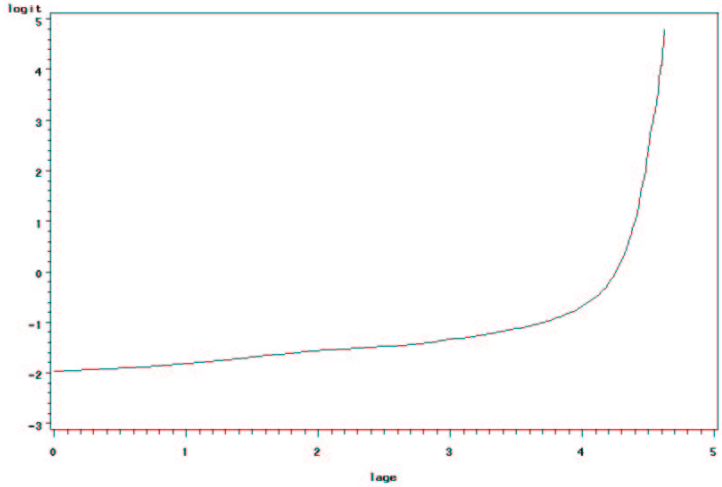

Figure 3.6: Plot for evaluating log-logistic model for Amish family data set

3.3 Lung Cancer Data

This data set was used in Gray (1995), and was obtained from a randomised control trial in which two treatments for lung cancer were being compared. The two treatment arms were 'CAV-HEM' and 'CAV'. CAV-HEM is a new form of chemotherapy, whereas CAV is the standard chemotherapy given to patients. There are 579 observations (285 from the CAV-HEM group and 294 from the CAV group). Only 1.727% of the observations are censored. The variables in the data set are listed below.

- survival time
- event indicator (0=censored, 1=experienced the event)
- institution number
- treatment (0=CAV, 1=CAV-HEM)
- performance status(1=performing well, 0=not performing well)
- liver metastases (1=yes, 0=no)
- bone metastases (1=yes, 0=no)
- weight loss (1=yes, 0=no)

There were 31 institutions, which could have as few as 1 patient, or as many as 56 patients. In Gray's paper, the institutions with less than 4 patients were deleted from the data set, leaving 570 observations (281 on CAV-HEM, 289 on CAV) and 26 institutions. All statistical analysis done using this data set shall exclude the deleted observations. The average number in each institution was 21.9 (standard deviation = 13.73) and the median institution size was 18.5.

Below is a table describing the characteristics of the study population.

Table 3.15: Characteristics of study population

	Yes	No
Liver Metastases	229 (40.2%)	341 (59.8%)
Bone Metastases	177 (31.1%)	393 (69.0%)
Weight loss	337 (59.1%)	233 (40.9%)

Approximately 40% of the patients experienced liver metastases, 31% experienced bone metastases and 59% experienced weight loss. Table 3.16 looks at these characteristics according to treatment regimen. From this table it can be seen that there is not much

Table 3.16: Characteristics according to treatment regimen

	CAV (n=289)	CAV-HEM (n=281)
Liver Metastases		
Yes	121 (41.9%)	108 (38.4%)
No	168 (58.1%)	173 (61.6%)
Bone Metastases		
Yes	89 (30.8%)	88 (31.3%)
No	200 (69.2%)	193 (68.7%)
Weight Loss		
Yes	172 (59.5%)	165 (58.7%)
No	117 (40.5%)	116 (41.3%)

difference in the distribution of the variables according to which treatment regimen the patients are on.

Using the lifetest procedure in SAS, it is possible to test for differences between groups; that is the difference between the two treatments arms in this case. Table 3.17 shows the results of testing for equality over strata. Since the associated probabilities for all statistics is less than 0.05, the null hypothesis of no difference between groups can be soundly rejected, implying that there is a difference in the two treatments.

Table 3.17: Test of equality over strata

Test	χ^2	DF	Pr> χ^2
Log-rank	11.5497	1	0.0007
Wilcoxon	4.5993	1	0.0320
-2Log(LR)	10.9319	1	0.0009

3.3.1 The Survival Curve and Hazard Function

The survival functions are given in Figure 3.7, and the hazard functions in Figure 3.8.

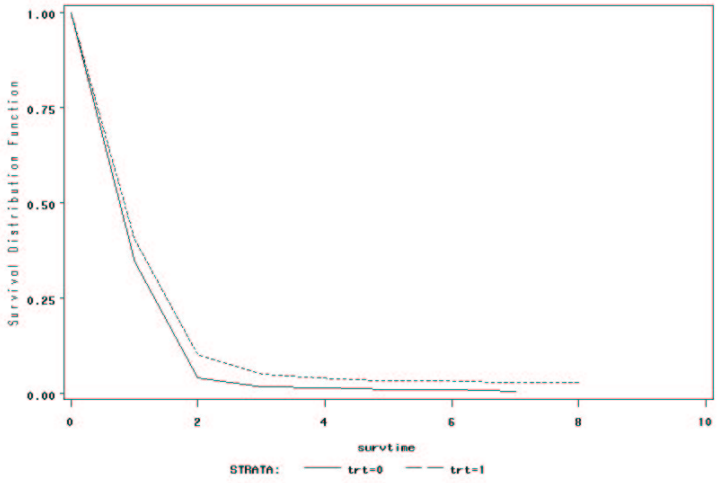

Figure 3.7: Survival function for lung cancer data set

From Figure 3.7 it can be seen that the patients on the CAV-HEM treatment (trt=1) have a higher survival probability than the patients on the CAV treatment. This is supported by the hazard functions in that it appears that patients on CAV have a consistently higher hazard than patients on CAV-HEM, particularly at the end of the survival time interval. Looking at the basic trend of the survival function it appears that the probability of survival rapidly decreases until the second year, and then the curve flattens out thereafter. There is a sharp increase in the hazard function until just before year 2,

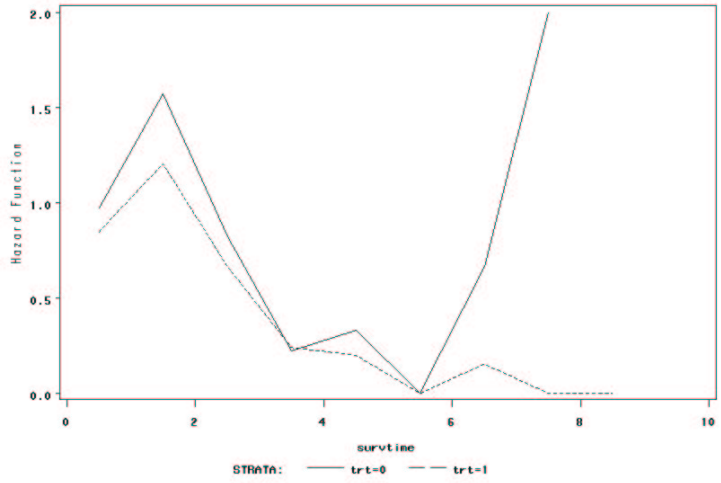

Figure 3.8: Hazard function for lung cancer data set

after which the hazard decreases. This implies that the hazard of death is high in the first 2 years, thereafter it decreases. The extreme difference in hazard functions after 6 years is most likely due to the fact that the estimates are based on very small numbers at this point (only 10 (1.75%) have survival times greater than 6 years), and this can lead to estimates with large uncertainty.

3.3.2 The Log-Normal Model

Proc lifereg was used to fit the log-normal model. The model fitted is given by the following equation,

$$survtime = \exp(\beta_0 + \beta_1 x_{trt} + \beta_2 x_{perfstat} + \beta_3 x_{liver} + \beta_4 x_{bone} + \beta_5 x_{weightloss} + error) \tag{3.2}$$

where all the independent variables are binary variables, taking on the values 0 or 1. In the model the variables survtime, trt, perfstat, liver, bone and weightloss respectively mean survival time, treatment arm, performance status, liver metastases, bone metastases and

weight loss. All the variables are included in the model to determine whether they affect survival time. The number of observations used in the analysis was 570, only 10 of which were censored (1.7544%). The algorithm was said to converge, and the log-likelihood was -729.4114671. The various statistics and parameter estimates are presented in Tables 3.18 and 3.19. Perfstat, liver and bone are significant variables at a 5% significance level, while

Table 3.18: Type III analysis of effects for the log-normal model

Effect	DF	Wald χ^2	Pr> χ^2
trt	1	3.2287	0.0724
perfstat	1	68.4041	< 0.0001
liver	1	15.2102	< 0.0001
bone	1	4.5625	0.0327
weightloss	1	3.4499	0.0633

trt and weightloss are just insignificant. The intercept is highly significant as well. The

Table 3.19: Parameter estimates for the log-normal model

Parameter	DF	Estimate	Std Error	χ^2	Pr> χ^2	e^{β}
Intercept	1	-0.6870	0.1079	40.51	< 0.0001	-
trt	1	0.1312	0.0730	3.23	0.0724	1.140196
perfstat	1	0.7454	0.0901	68.40	< 0.0001	2.107284
liver	1	-0.2974	0.0762	15.21	< 0.0001	0.742747
bone	1	-0.1731	0.0810	4.56	0.0327	0.841054
weightloss	1	-0.1394	0.0751	3.45	0.0633	0.86988
Scale	1	0.8693	0.0261	-	-	-

interpretation of the statistic e^{β} is as for categorical variables as there are no quantitative variables. For trt the interpretation is that the expected survival time for patients on treatment regimen 1 (CAV-HEM) is 14% higher than for patients on treatment 0 (CAV only). For people with a performance indicator of 1, their expected survival time is approximately double that of patients with a performance indicator of 0. The survival time for patients that have liver metastases is 75% that of patients who don't have liver

metastases. The survival time for patients with bone metastases is about 84% that of patients who do not have bone metastases, and for patients with weightloss, their survival time is about 87% that of patients who do not suffer from weightloss.

3.3.3 The Exponential Model

The same model as specified in Eq.(3.2) was fitted using an exponential model. Again 570 observations were analysed, 10 of which were censored. The algorithm was said to converge, and the log-likelihood was -762.3956347. The output is slightly different to the log-normal

Table 3.20: Type III analysis of effects for the exponential model

Effect	DF	Wald χ^2	Pr> χ^2
trt	1	8.3532	0.0039
perfstat	1	19.8390	< 0.0001
liver	1	8.4557	0.0036
bone	1	3.4419	0.0636
weightloss	1	3.9208	0.0477

model, in that now trt is a significant variable at a 5% significance level, where it was not before. Also the variable bone is not significant anymore, whereas weightloss is now a significant variable in the regression. Perfstat and liver are still significant at a 5% level. From Table 3.21 it can be seen that the intercept is no longer significant. The Lagrange

Table 3.21: Parameter estimates for the exponential model

Parameter	DF	Estimate	Std Error	χ^2	Pr> χ^2
Intercept	1	-0.1895	0.1166	2.64	0.1040
trt	1	0.2458	0.0850	8.35	0.0039
perfstat	1	0.4604	0.1034	19.84	< 0.0001
liver	1	-0.2598	0.0893	8.46	0.0036
bone	1	-0.1729	0.0932	3.44	0.0636
weightloss	1	-0.1732	0.0875	3.92	0.0477
Scale	0	1	0.0000	-	-
Weibull shape	0	1.0000	0.0000	-	-

multiplier statistic for scale is 77.3574 which has an associated probability of < 0.0001. The null hypothesis of a constant hazard over time is clearly rejected, implying that the hazard is not constant over time, and that the exponential model may not be good model for this data.

3.3.4 The Weibull Model

The model fitted is the model given in Eq.(3.2). In the analysis 570 observations were used, 10 were censored, and the algorithm was said to converge. The log-likelihood was -738.9177051. In this model, all the variables are significant at a 5% significance level,

Table 3.22: Type III analysis of effects for the Weibull model

Effect	DF	Wald χ^2	Pr> χ^2
trt	1	15.6652	< 0.0001
perfstat	1	25.4291	< 0.0001
liver	1	10.5664	0.0012
bone	1	5.9492	0.0147
weightloss	1	6.8108	0.0091

Table 3.23: Parameter estimates for the Weibull model

Parameter	DF	Estimate	Std Error	χ^2	Pr> χ^2	e^{β}
Intercept	1	-0.1000	0.0919	1.18	0.2764	-
trt	1	0.2721	0.0688	15.67	< 0.0001	1.3127
perfstat	1	0.4206	0.0834	25.43	< 0.0001	1.5229
liver	1	-0.2358	0.0725	10.57	0.0012	0.7899
bone	1	-0.1827	0.0749	5.95	0.0147	0.8330
weightloss	1	-0.1846	0.0707	6.81	0.0091	0.8314
Scale	1	0.8039	0.0237	-	-	-
Weibull shape	1	1.2439	0.0367	-	-	-

however the intercept is not significant. The scale parameter (σ) is between 0.5 and 1.0, thus implying that the hazard is increasing at a decreasing rate. The coefficient for log$t$ is given by $(1/\sigma) - 1 = 0.2439$, which can be interpreted to mean that a 1% increase in the

time since follow-up produces a 0.2439% increase in the hazard for death. The treatment effect is more pronounced in this model than in the log-normal model in that the expected survival time for patients on the CAV-HEM arm is approximately 1.3 times higher than that for patients on the CAV arm.

3.3.5 The Log-Logistic Model

The model fitted is the same as in Eq.(3.2), where 570 observations were used, of which 10 were censored. The algorithm was said to converge and the log-likelihood was -687.8800281.

Table 3.24: Type III analysis of effects for the log-logistic model

Effect	DF	Wald χ^2	Pr> χ^2
trt	1	3.5897	0.0581
perfstat	1	42.5353	< 0.0001
liver	1	18.0966	< 0.0001
bone	1	4.5050	0.0338
weightloss	1	2.333	0.1266

Table 3.25: Parameter estimates for the log-logistic model

Parameter	DF	Estimate	Std Error	χ^2	Pr> χ^2
Intercept	1	-0.5059	0.0985	26.38	< 0.0001
trt	1	0.1175	0.0620	3.59	0.0581
perfstat	1	0.5486	0.0841	42.54	< 0.0001
liver	1	-0.2740	0.0644	18.10	< 0.0001
bone	1	-0.1475	0.0695	4.51	0.0338
weightloss	1	-0.0976	0.0639	2.33	0.1266
Scale	1	0.4378	0.0159	-	-

Here trt was just insignificant at a 5% significance level. The intercept, perfstat, liver and bone were significant at a 5% significance level, and weightloss was not significant. Since $\sigma < 1$, ($\sigma = 0.4378$), the log-logistic hazard is similar to the log-normal hazard in that it starts at 0, rises to a peak and then declines to 0.

3.3.6 The Gamma Model

The generalised gamma model was first fitted to the model described in Eq.(3.2). There were no convergence problems, and the computations did not take very long. The standard gamma model was also then fitted to check whether it would be a plausible model for the data.

When fitting the generalised gamma model, 570 observations were analysed, 10 were censored. The algorithm was said to converge, and the log-likelihood was -711.399885. All

Table 3.26: Type III analysis of effects for the generalised gamma model

Effect	DF	Wald χ^2	Pr> χ^2
trt	1	7.1511	0.0075
perfstat	1	39.0555	< 0.0001
liver	1	15.5566	< 0.0001
bone	1	4.9754	0.0257
weightloss	1	4.5742	0.0325

Table 3.27: Parameter estimates for the generalised gamma model

Parameter	DF	Estimate	Std Error	χ^2	Pr> χ^2
Intercept	1	-0.3724	0.1068	12.15	0.0005
trt	1	0.1839	0.0688	7.15	0.0075
perfstat	1	0.5434	0.0869	39.06	< 0.0001
liver	1	-0.2809	0.0712	15.56	< 0.0001
bone	1	-0.1672	0.0750	4.98	0.0257
weightloss	1	-0.1498	0.0701	4.57	0.0325
Scale	1	0.8069	0.0254	-	-
Shape	1	0.4466	0.0718	-	-

the variables are significant at a 5% significance level. When the shape parameter is close to 0, the distribution is similar to the log-normal model. Here the shape parameter is 0.4466, which is not that close to 0. Even though the scale and shape parameters are not highly similar, it is worth looking at a standard gamma model as both parameters are less than

one. To fit the model in proc lifereg, it is necessary to fix the scale and shape parameters at specific values. To find good estimates of these values it is necessary to try different values, and use the values that maximise the likelihood. Since the shape parameter is 0.4466 and the scale parameter is 0.8069, a good idea would be to start at the lower end of the range with 0.44, and increasing by 0.02 until a value is found which maximises the likelihood. Table 3.28 shows the various parameter estimates used. It appears that the best value for the

Table 3.28: Searching for values of the shape and scale parameters

Estimate	Log-Likelihood
0.44	-1015.433516
0.46	-962.3539106
0.48	-917.8719016
0.50	-880.567308
0.52	-849.2936768
0.54	-823.1179377
0.56	-801.2749617
0.58	-783.1329683
0.60	-768.1669266
0.62	-755.9379119
0.64	-746.076945
0.66	-738.2722383
0.68	-732.2590559
0.70	-727.8115934
0.72	-724.7364334
0.74	-722.8672378
0.76**	-722.0604175**
0.78	-722.1915811
0.80	-723.1526065
0.82	-724.8492148

shape and scale parameter which maximises the log-likelihood is 0.76. The final output for the standard gamma model is given in Tables 3.29 and 3.30. All the explanatory variables as well as the intercept are significant at a 5% significance level. When compared to the generalised gamma model, the standard errors appear to be slightly less in the standard

Table 3.29: Type III analysis of effects for the standard gamma model

Effect	DF	Wald χ^2	Pr> χ^2
trt	1	13.4578	0.0002
perfstat	1	33.3139	< 0.0001
liver	1	14.2711	0.0002
bone	1	6.0886	0.0136
weightloss	1	6.6122	0.0101

Table 3.30: Parameter estimates for the standard gamma model

Parameter	DF	Estimate	Std Error	χ^2	Pr> χ^2
Intercept	1	-0.1881	0.0885	4.52	0.0336
trt	1	0.2364	0.0644	13.46	0.0002
perfstat	1	0.4531	0.0785	33.31	< 0.0001
liver	1	-0.2558	0.0677	14.27	0.0002
bone	1	-0.1742	0.0706	6.09	0.0136
weightloss	1	-0.1704	0.0663	6.61	0.0101
Scale	0	0.7600	0.0000	-	-
Shape	0	0.7600	0.0000	-	-

gamma model. The standard errors are probably slight underestimates because they do not take into account the sampling distribution in the scale and shape estimate (Allison, 1995). The log-likelihood is slightly larger than in the generalised gamma model, and there are slight changes in the parameter estimates. A statistic defined by $K = 1/\delta^2$ is useful in determining the shape of the hazard function. When $K > 1$, the hazard is 0 at time 0, and increases thereafter. When $0 < K < 1$, the hazard is infinite at time 0 and then decreases, and when $K = 1$ the hazard is constant (namely the exponential model) (Allison, 1995). In this case $K = 1.7313019 > 1$, thus there is evidence that the hazard increases with time.

3.3.7 Choosing the Best Model to Fit the Data

The Weibull model, generalised gamma and standard gamma model all give the same conclusion regarding the significance of the variables. Treatment group, performance status, liver metastases, bone metastases and weightloss were all found to be significant. The log-normal and log-logistic model both found treatment group and weightloss to be insignificant in the model, which is an important discrepancy particularly since the effectiveness of the two different treatments are trying to be assessed. The exponential model found the treatment group, performance status, liver metastases and weightloss to be significant variables in the model.

The likelihood ratio statistic shall be used to compare nested models once again, thereby testing the following null hypothesis

$$H_0 : \text{The two models are the same}$$

$$H_1 : \text{The two models are not the same.}$$

Recall the likelihood ratio statistic

$$G = -2[LL_A - LL_B].$$

The exponential, Weibull, log-normal and standard gamma models are nested within the

generalised gamma models and can be compared against it; and the exponential model can be evaluated against the weibull and standard gamma models. The likelihoods associated with the various fitted models are given in Table 3.31. At a first glance it appears that the

Table 3.31: Log-likelihoods for the fitted models

Model	Log-Likelihood
Log-Normal (AFT)	-729.4114671
Exponential	-762.3956347
Weibull	-738.9177051
Log-Logistic	-687.8800281
Standard Gamma	-722.0604175
Generalised Gamma	-711.399885

log-logistic model is the best fit, however it cannot be evaluated against the other models. From Table 3.32 it can be seen that the generalised gamma model is the best-fitting model,

Table 3.32: Deviance for comparisons of various models

Comparison	G	Pr> χ^2
Exp vs. Weibull	46.95584	< 0.0001
Exp vs. Std Gamma	80.67042	< 0.0001
Exp vs. G. Gamma	101.99148	< 0.0001
Weibull vs. G. Gamma	55.03566	< 0.0001
Log-normal vs. G. Gamma	36.02318	< 0.0001
Std. Gamma vs. G. Gamma	21.32103	< 0.0001

as the null hypothesis is rejected in every instance at a 5% significance level.

Graphical method of testing for linearity

The log-survivor plot in Figure 3.9 to test linearity in the exponential model appears to be non-linear; and thus the exponential model is not suitable for this data. This supports the hypothesis against constant hazards done earlier.

The log-log survivor plot in Figure 3.10 appears to be more linear than the log-survivor

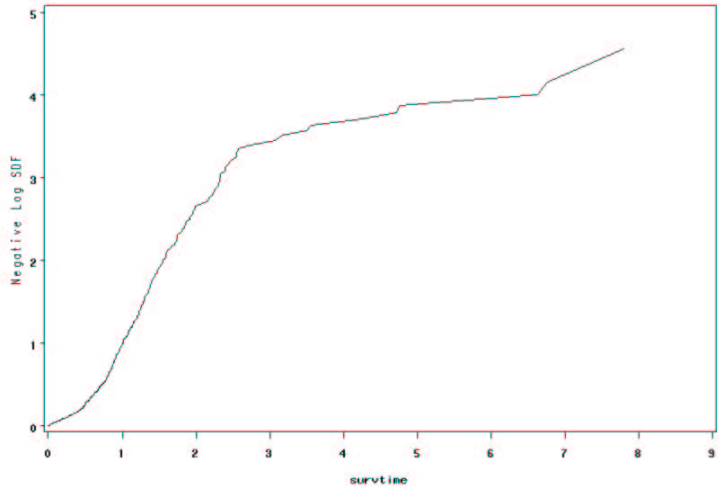

Figure 3.9: Log-survivor plot for lung cancer data set

plot, except for a few deviations from linearity. Thus the Weibull model may be a suitable model for the lung cancer data.

The plot for evaluating the log-normal model shown in Figure 3.11 shows a curve in the middle of the plot, which is more pronounced than in the Weibull case. Thus there does appear to be a deviation from linearity, and the log-normal model may not be the most suitable model to use in this case.

The plot for evaluating the log-logistic model in Figure 3.12 also shows a slight curve in the middle of the plot, but this does not seem to be as pronounced as in the log-normal case. There is some evidence of deviation from linearity, but it is not too severe.

3.4 Warfarin Data

This data set was obtained from the Medical Research Council of South Africa, with permission to use the data set from Buchanan-Lee (2002) of the Groote Schuur Hospital in Cape Town. The study was a five-year prospective randomized double-blind study, whose aim was to compare the efficacy and safety of a predetermined, individualised fixed-dose

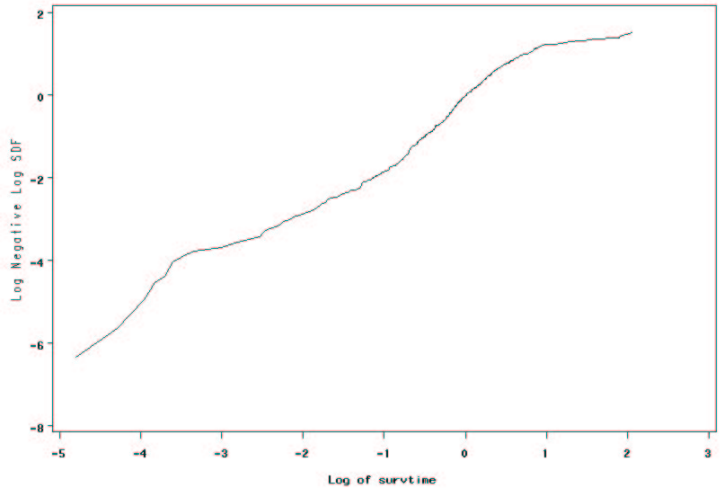

Figure 3.10: Log-log survivor plot for lung cancer data set

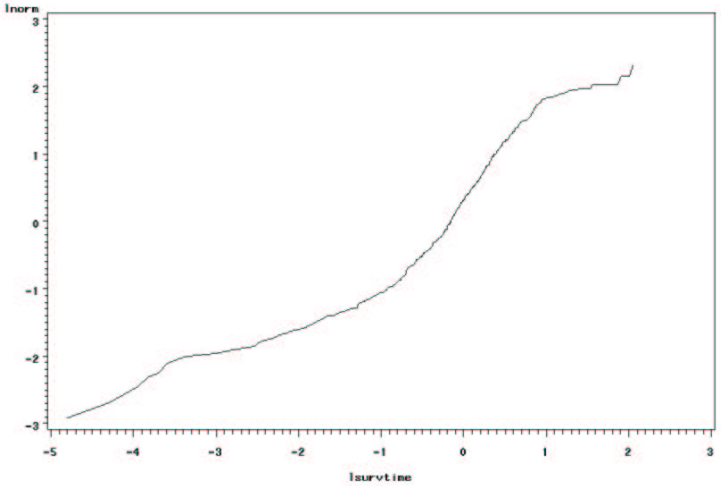

Figure 3.11: Plot for evaluating log-normal model for lung cancer data set

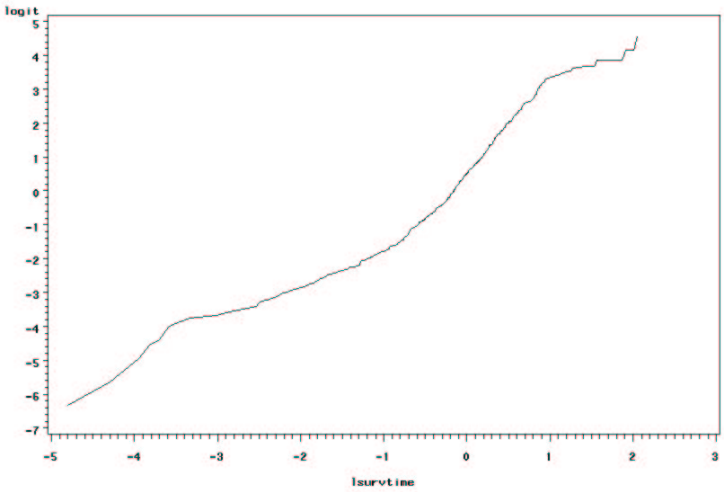

Figure 3.12: Plot for evaluating log-logistic model for lung cancer data set

versus adjusted dose warfarin, in which 296 patients with mechanical heart valves were randomised after an initial dose-finding phase to either fixed-dose or adjusted dose warfarin. The patients were young, geographically dispersed and socio-economically deprived. The results of the study have been published (Buchanan-Lee *et al.*, 2002).

There were 12 events that could have possibly been recorded, where each patient may have had more than one event. Table 3.31 below is a listing of the events, and the frequency in each group of each occurring event. In the analysis which will follow in the next sections, some assumptions regarding the data were made in order to make the data easier to analyse. Firstly only the first 5 events in the above table were considered "events"; and the remaining 7 events were grouped together. The reason for doing this is that the first 5 events are considered in a medical sense to be major events, and the remaining events were minor events. The end date for the second group was taken to be when the participant was censored, in other words ignoring the events which occurred before this date. For events 1 to 5, the time to *first* event was recorded. Thus events occurring after the first event were ignored. After reorganising the data in this manner, the events in Table 3.34 were

Table 3.33: Table of frequency of events for each group

Event	Group 1(Fixed)	Group 2(Adjusted)	Total
Death	7	5	12
Blocked Valve	5	1	6
Major Thrombotic	8	3	11
Intracranial haemorrhagic	2	5	7
Major Bleed	10	7	17
Minor Bleed	8	1	9
INR < 1.3(low)	27	20	47
INR > 6(high)	4	4	8
Medical end pt	10	17	27
Lost	0	2	2
Violation	1	0	1
Censored	140	144	284
Total	222	209	431

then ready for analysis. The assumptions were simplified further so that the 5 events were

Table 3.34: Table of frequency of combined events for each group in new approach

Event	Group 1(Fixed)	Group 2(Adjusted)	Total
Death	6	5	11
Blocked Valve	4	1	5
Major Thrombotic	7	1	8
Intracranial haemorrhagic	1	4	5
Major Bleed	9	7	16
Censored	119	134	253
Total	146	152	298

further grouped together, so that the resulting binary indicator for each participant was either “experienced an event” or “censored”. Thus 27 participants experienced an event in the fixed-dose group, and 18 participants experienced an event in the adjusted-dose group, a total of 45 events in all. In the fixed-dose warfarin group 119 participants were censored, and 134 participants in the adjusted-dose warfarin group were censored, leaving a very large proportion of censored observations (84.9%).

The baseline characteristics were age, weight and sex. For the variable age, only 4 data points were missing. The average age was 41 years (standard deviation = 13.47), and the median was 40. The oldest participant was 77 years old and the youngest was 14 years old. Unfortunately, more than half of the data points were missing for the variable weight. Only 143 of the 298 participants had weight recorded at baseline. Thus for this reason weight was not included as a covariate in the analysis. Nonetheless the average weight was 61.7 kgs (standard deviation = 13.157) with a median of 60 kgs. The maximum weight recorded was 94 kg, and the minimum was 34 kg. Just more than half of the participants were female (59.12%), and 40.88% were male. The time to an event was measured in months.

3.4.1 The Survival Curve and Hazard Function

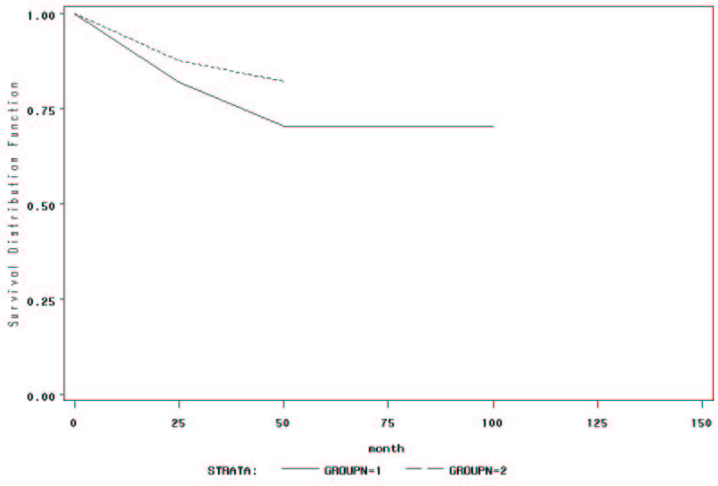

Figure 3.13: Survivor function for the warfarin data set

Group 1 is the fixed dose warfarin group and group 2 represents the adjusted dose warfarin group. From the survivor functions in Figure 3.13 the adjusted dose warfarin group have a higher survival plot than the fixed dose warfarin, implying that the patients on adjusted dose warfarin have a longer survival time. However, there are no patients after

time 50, so it is not possible to know what may happen beyond this point.

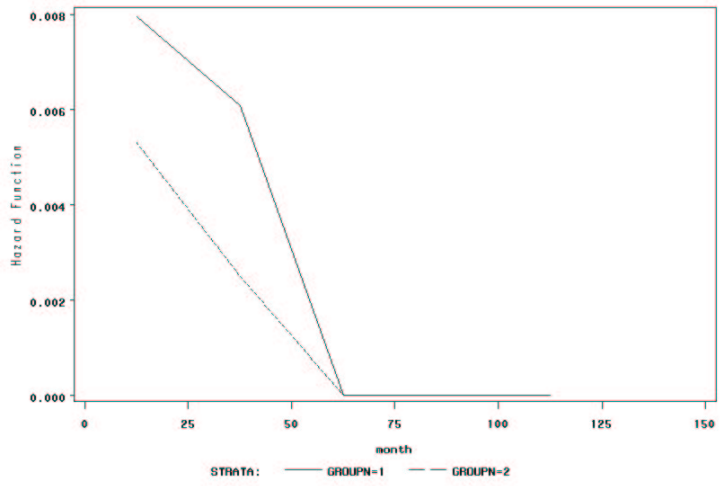

Figure 3.14: Hazard function for the warfarin data set

The hazard function in Figure 3.14 shows clearly that the fixed dose group have a higher hazard at the beginning, but then both groups decline fairly rapidly until about time 60, and then the hazard for group 1 levels off, while the hazard for group 2 stops at this point.

3.4.2 The Log-Normal Model

In all the parametric models fitted, 293 observations were used, and the algorithm was said to converge. The model fitted was

$$time \quad = \quad \exp(\beta_0 + \beta_1 x_{group} + \beta_2 x_{age} + \beta_3 x_{sex} + error)$$

where x_{group}, x_{age} and x_{sex} are variables representing the treatment arm, age and sex of the individual in question.

In the log-normal model, none of the variables were significant. The model output is presented in Tables 3.35 and 3.36.

The log-likelihood was -171.9325. As can be seen from the tables, none of the variables are significant, except for the intercept. To interpret the estimates, if they were significant,

Table 3.35: Type III analysis of effects for the log-normal model

Effect	DF	Wald χ^2	Pr> χ^2
group	1	1.4019	0.2364
age	1	0.6429	0.4227
sex	1	0.9063	0.3411

Table 3.36: Parameter estimates for the log-normal model

Parameter	DF	Estimate	Std Error	χ^2	Pr> χ^2	e^{β}
Intercept	1	4.2559	0.9424	20.39	< 0.0001	-
group	1	0.5153	0.4352	1.40	0.5364	1.6741
age	1	0.0128	0.0160	0.64	0.4227	1.0129
sex	1	0.4130	0.4338	0.91	0.3411	1.5113
Scale	1	2.4132	0.2924	-	-	-

one could say that the adjusted dose warfarin group's survival time was 167% of that of the fixed dose warfarin, and that the females' survival time was 150% that of the males. For age, the interpretation would be that for a 1 year increase in age, survival time increases by 1.29%, which is interesting because it would imply that older patients have higher survival times if this variable was significant.

3.4.3 The Exponential Model

The exponential model does not yield different results from the log-normal model. The parameter estimates are presented in Tables 3.37 and 3.38.

Table 3.37: Type III analysis of effects for the exponential model

Effect	DF	Wald χ^2	Pr> χ^2
group	1	2.4162	0.1201
age	1	0.3559	0.5508
sex	1	0.0567	0.8118

The χ^2 value of the Lagrange multiplier statistic for scale is 2.5908, which has an

Table 3.38: Parameter estimates for the exponential model

Parameter	DF	Estimate	Std Error	χ^2	Pr> χ^2
Intercept	1	4.1589	0.6259	44.16	< 0.0001
group	1	0.4783	0.3077	2.42	0.1201
age	1	0.0067	0.0112	0.36	0.5508
sex	1	0.0733	0.3081	0.06	0.8118
Scale	0	1.0000	0.0000	-	-
Weibull shape	0	1.0000	0.0000	-	-

associated probability of 0.1075. The null hypothesis of a constant hazard is not rejected, and thus the exponential model may be a viable model to use. The log-likelihood is -174.5121745.

3.4.4 The Weibull Model

The results for this model are contained in Tables 3.39 and 3.40. From Table 3.40 it is clear that all the variables are still insignificant, except for the intercept. The log-likelihood was -172.7999. The interpretation of the hazard ratio for group is that the survival time for the adjusted dose group is 1.79 times that of the fixed dose group. For age, every additional year increases the survival time by 0.86%. The females' survival time is 113% that of the men.

Table 3.39: Type III analysis of effects for the Weibull model

Effect	DF	Wald χ^2	Pr> χ^2
group	1	2.1714	0.1406
age	1	0.3592	0.5490
sex	1	0.0986	0.7536

Table 3.40: Parameter estimates for the Weibull model

Parameter	DF	Estimate	Std Error	χ^2	Pr> χ^2	e^{β}
Intercept	1	4.3539	0.8052	29.24	< 0.0001	-
group	1	0.5827	0.3954	2.17	0.1406	1.7909
age	1	0.0086	0.0143	0.36	0.5490	1.0086
sex	1	0.1227	0.3908	0.10	0.7536	1.1305
Scale	1	1.2675	0.1707	-	-	-
Weibull shape	1	0.7889	0.1062	-	-	-

3.4.5 The Log-logistic Model

From Table 3.42 it is again clear that none of the variables were significant at a 5% significance level. The log-likelihood was -172.5608. Since the scale parameter is greater than one the hazard behaves like the decreasing Weibull hazard in that it starts at infinity and declines to zero as t tends to infinity. This is consistent with the graphical plot of the hazard function in Figure 3.14.

Table 3.41: Type III analysis of effects for the log-logistic model

Effect	DF	Wald χ^2	Pr> χ^2
group	1	1.9859	0.1588
age	1	0.4399	0.5072
sex	1	0.2386	0.6252

Table 3.42: Parameter estimates for the log-logistic model

Parameter	DF	Estimate	Std Error	χ^2	Pr> χ^2
Intercept	1	4.0238	0.8519	22.31	< 0.0001
group	1	0.5727	0.4064	1.99	0.1588
age	1	0.0097	0.0147	0.44	0.5072
sex	1	0.1971	0.4035	0.24	0.6252
Scale	1	1.1945	0.1592	-	-

3.4.6 The Gamma Model

The results for fitting the generalised gamma model are shown in Tables 3.43 and 3.44. None of the variables other than the intercept are significant. The shape and scale parameters are quite different, and thus it is not necessary to fit a standard gamma model. The log-likelihood was -171.8356.

Table 3.43: Type III analysis of effects for the generalised gamma model

Effect	DF	Wald χ^2	Pr> χ^2
group	1	1.0099	0.3149
age	1	0.7307	0.3927
sex	1	0.8765	0.3492

Table 3.44: Parameter estimates for the generalised gamma model

Parameter	DF	Estimate	Std Error	χ^2	Pr> χ^2
Intercept	1	3.9746	1.3812	8.28	0.0040
group	1	0.4695	0.4672	1.01	0.3149
age	1	0.0155	0.0181	0.73	0.3927
sex	1	0.5874	0.6274	0.88	0.3492
Scale	1	2.9852	1.3183	-	-
Shape	1	-0.5501	1.4055	-	-

3.4.7 Choosing the Best Model to Fit the Data

The different models all have the same conclusion with respect to the significance of the variables. None of the variables in any of the models were found to be significant.

The likelihood ratio statistic, also known as the deviance, was used to assess the model of best fit. In this instance the choice of the best model is not too critical as they all show similar results. The statistic used is

$$G = -2[LL_A - LL_B].$$

The log-likelihoods for the various models are reflected in Table 3.45. Table 3.46 shows the

Table 3.45: Log-likelihoods for the fitted models

Model	Log-Likelihood
Log-Normal (AFT)	-171.9325
Exponential	-174.5122
Weibull	-172.7999
Log-Logistic	-172.5608
Generalised Gamma	-171.8356

comparisons made and the probability associated with the test statistic, G. From this table

Table 3.46: Deviance for comparisons of various models

Comparison	DF	G	Pr> χ^2
Exp vs. Weibull	1	3.4246	0.0642
Exp vs. G. Gamma	2	5.3532	0.0688
Weibull vs. G. Gamma	1	1.9286	0.1649
Log-normal vs. G. Gamma	1	0.1938	0.659773

it can be seen that the simplest model, namely the exponential model, is as good as the most complicated model, the generalised gamma model. In other words, we fail to reject the exponential model in favour of any other model at a 5% significance level. In this case it is best to take the least complicated model. Thus the final model recommended is the exponential model, and this result is reinforced by the fact that the assumption of constant hazards was not rejected when the separate model analyses were carried out. However, the log-logistic model does not fit into the nested scheme, and so cannot be evaluated against the other models. It appears to have a comparable log-likelihood, and so it also fits the data quite well.

Graphical method of testing for linearity

The log-survivor plot in Figure 3.15 to test linearity in the exponential model appears to be linear, except for slight deviations near the top of the line. Thus the exponential model may be a suitable for this data. This reinforces the conclusion drawn earlier.

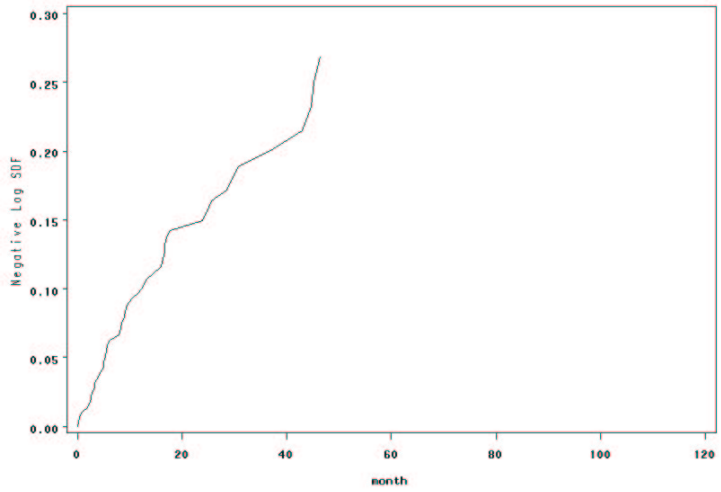

Figure 3.15: Log-survivor plot for warfarin data set

The log-log survivor plot in Figure 3.16 also appears to be linear, except for the small deviation near the beginning of the line. Thus the Weibull model may be a suitable model for the warfarin data.

The plot for evaluating the log-normal model shown in Figure 3.17 shows some non-linearity in the beginning of the graph, thereafter it appears to be fairly linear. Thus the log-normal may be a suitable model to use.

The plot for evaluating the log-logistic model (Figure 3.18) also shows a slight deviation from linearity at the beginning of the plot, but does appear to be linear thereafter.

Thus all models appear to satisfy the linearity requirement, and thus the decision to use the exponential model is still a valid conclusion. It is the most parsimonious model among

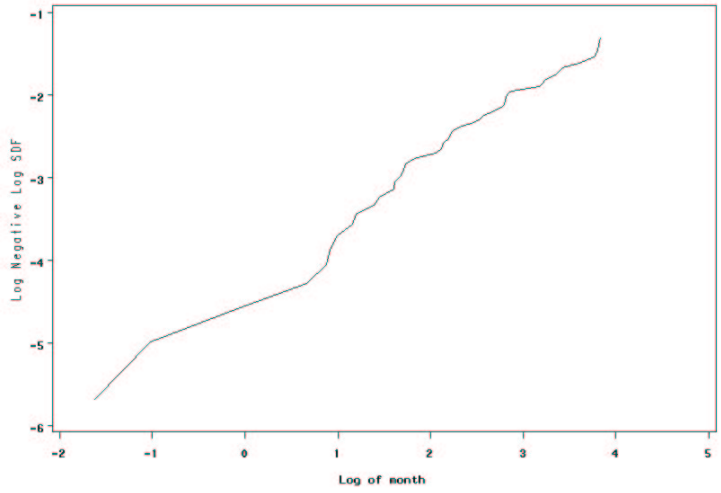

Figure 3.16: Log-log survivor plot for warfarin data set

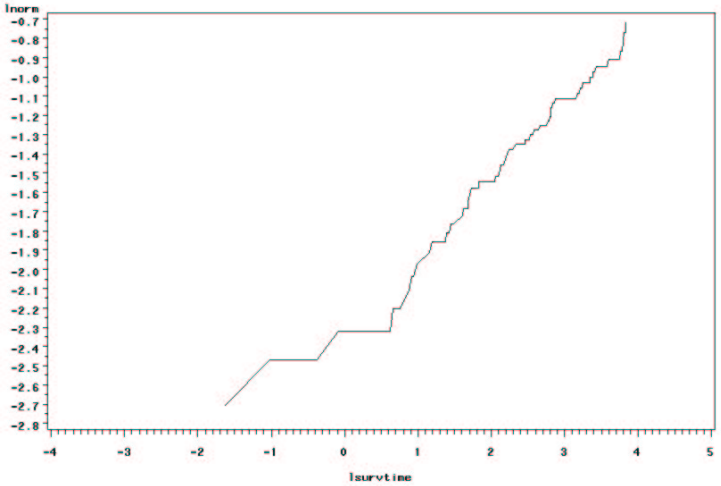

Figure 3.17: Plot for evaluating log-normal model for warfarin data set

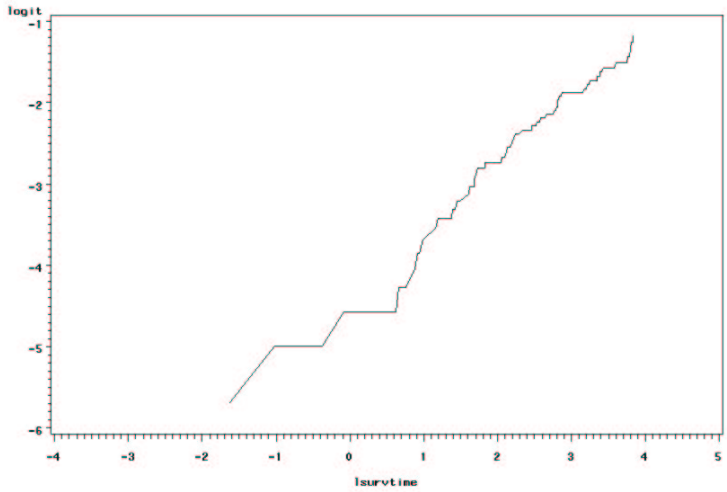

Figure 3.18: Plot for evaluating log-logistic model for warfarin data set

equally comparable models.

Chapter 4

Semi-Parametric Models for Survival Distributions

4.1 Introduction

The two main reasons for modeling survival data is to determine which explanatory variables affect the form of the hazard function, and to obtain an estimate of the hazard function for an individual (Collett, 1994). The distribution of the survival time variable, T, can be specified in either of the two following ways. These are through modeling the density function of a parametric distribution for T, or to model the hazard function as a function of risk factors (Hosmer and Lemeshow, 1999). Since the focus of survival analysis is the risk of death, it makes sense to model the hazard function directly (Collett, 1994). An advantage of modeling the hazard function is that the aging process can be addressed directly which may be preferable in the situation where there is more than one group whose survival experience may be compared. However, a disadvantage of using the hazard function is that the use of scatterplots is not useful to motivate regression models.

4.2 The Cox Proportional Hazards Model

The proportional hazards model was proposed by Cox (1972), and has come to be known as the *Cox regression model*. It is known as a semi-parametric model because no assumptions

are made about the nature of the baseline hazard. However, it is based on the assumption of proportional hazards.

Suppose, for simplicity, that x is the single known covariate, and β the corresponding unknown coefficient. Recall that the hazard function, defined in Eq.(1.2), is the probability that an individual dies at time t, given that they have survived up until that point. In general, the hazard function can be specified as a function of time and the covariates in the form

$$h(t, x, \beta) = h_0(t) r(x, \beta) \tag{4.1}$$

where $r(x, \beta)$ is a function of the covariates. The above equation should be strictly positive. Here $h_0(t)$ characterises how the hazard function changes as a function of survival time and is also known as the baseline hazard since $h(t, x, \beta) = h_0(t)$ when $x = 0$. The term $r(x, \beta)$ explains how the hazard function changes as a function of specific covariates.

Consider the simplest case where patients randomly allocated to two groups are compared. For simplicity, assume the two groups being compared are a control group and a treatment group. Let the hazard for the control group be $h(t, x_0, \beta)$ and for the treatment group $h(t, x_1, \beta)$. Suppose that the ratio of the two hazard functions for the treatment and control group is

$$\psi = \frac{h(t, x_1, \beta)}{h(t, x_0, \beta)}.$$

Under Eq.(4.1) this becomes

$$\begin{aligned} \psi &= \frac{h_0(t) r(x_1, \beta)}{h_0(t) r(x_0, \beta)} \\ &= \frac{r(x_1, \beta)}{r(x_0, \beta)}. \end{aligned}$$

If the hazard ratio, ψ, is easily interpreted then the actual form of the baseline hazard function is of little importance (Hosmer and Lemeshow, 1999). The ratio ψ measures the

risk of death at time t for an individual on treatment relative to a person on the control. If $\psi < 1$ the hazard for an individual on treatment is said to be smaller than for an individual on the control, and the treatment is thus an improvement. If $\psi > 1$ then the hazard is smaller for a person in the control group than for a person on the treatment and it cannot be concluded that the treatment is effective in increasing survival time.

From the proportional hazards assumption the following important relationship between the survivor function of the treatment group and the survivor function of the control group emerges

$$\begin{aligned} S_1(t) &= e^{-H(t,x_1,\beta)} \\ &= \exp\{-\int_0^t h(u,x_1,\beta)\mathrm{d}u\} \\ &= \exp\{-\int_0^t \psi h_0(u) r(x_0,\beta)\mathrm{d}u\} \\ &= \exp\{-\psi\int_0^t h_0(u) r(x_0,\beta)\mathrm{d}u\} \\ &= [S_0(t)]^{\psi}. \end{aligned}$$

Cox (1972) proposed a model that uses $r(x,\beta) = e^{x\beta}$. The proportional hazards model assumption then becomes

$$h(t,x,\beta) = h_0(t)e^{x\beta}$$

and the hazard ratio for the simple case of comparing two groups is

$$\psi = e^{\beta(x_1-x_0)}.$$

In the case where the single covariate x is dichotomous, in other words taking values 0 or 1, the hazard ratio can be seen as a type of "relative-risk" ratio (Hosmer and Lemeshow, 1999). Hence if $\beta = \ln(2)$, then those with $x = 1$ are dying at twice the rate of those with $x = 0$.

Under Cox's regression model the survivor function can be rewritten as

$$S_1(t) = [S_0(t)]^{\exp(x\beta)}. \tag{4.2}$$

4.2.1 The General Proportional Hazards Model

Suppose now that the hazard of death at a particular time depends on the values $x_1, x_2, \ldots, x_p$ of p explanatory variables $X_1, X_2, \ldots, X_p$, where $x_1, x_2, \ldots, x_p$ are assumed to be recorded at the outset of the study for a given individual. Let $x_{i1}, x_{i2}, \ldots, x_{ip}$ denote the measured values of the p covariates for individual i. Thus the set of variable values can be denoted by the vector $\mathbf{x}_i$. Let $h_0(t)$ be the hazard function for an individual whose set of covariates, $\mathbf{x}_i$, are equal to zero, that is $h_0(t)$ gives the baseline hazard function. The hazard function for the i^{th} individual can then be written as

$$\begin{aligned} h_i(t|\mathbf{x}_i) &= \lim_{\delta t \to 0} \frac{P(t \leq T \leq t + \delta t | T \geq t; \mathbf{x}_i)}{\delta t} \\ &= \psi(\mathbf{x}_i) h_0(t) \end{aligned}$$

where $\psi(\mathbf{x}_i)$ is a function of the explanatory variables for the i^{th} person. Now $\psi(\mathbf{x}_i)$ can be interpreted as the hazard at time t for an individual whose vector of explanatory variables is $\mathbf{x}_i$, relative to the baseline hazard. In mathematical terms this relationship can be written as

$$\psi(\mathbf{x}_i) = \frac{h_i(t|\mathbf{x}_i)}{h_0(t)}.$$

Extending Cox's definition for the proportional hazards model, the hazard ratio can be expressed as follows

$$\begin{aligned} \psi(\mathbf{x}_i) &= \exp(\eta_i) \\ &= \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p) \\ &= \exp(\sum_{j=1}^{p} x_{ij}\beta_j) \\ &= \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \end{aligned}$$

where $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ is a vector of regression coefficients and η_i is the linear component of the model, where η_i is also known as the risk score or prognostic index for the i^{th} individual (Collett, 1994). The general proportional hazards model then becomes

$$h_i(t) \quad = \quad \exp(x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p)h_0(t). \tag{4.3}$$

This can be linearised by dividing through by the baseline hazard function and taking logs on both sides of Eq.(4.3)

$$\log\left[\frac{h_i(t)}{h_0(t)}\right] \quad = \quad \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}. \tag{4.4}$$

The structure of the model can be viewed in the context of the logistic regression model, the difference being that in the logistic regression model there is a constant term, β_0, whereas in the above model there is no constant term.

4.2.2 Fitting the Proportional Hazards Model

To fit the proportional hazards model, the unknown coefficients $\beta_1, \beta_2, \ldots, \beta_p$ need to be estimated. Let $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \ldots, \beta_p)$ denote a vector of unknown coefficients corresponding to the p known covariates. In some cases it may be necessary that the baseline hazard $h_0(t)$ be estimated. The most common approach to estimating the regression coefficients is that of the method of maximum likelihood.

Suppose that there are n individuals, each with the triplet $(t_i, \mathbf{x}_i, c_i)$, where t_i is the observed survival time, $\mathbf{x}_i$ is the covariate and c_i is the censoring indicator variable for individual i. Suppose that there are r ordered distinct death times such that $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$. This implies that there are $n - r$ right-censored survival times. The treatment of ties is not considered here, but will be discussed later. Suppose that the set of individuals that are at risk at time $t_{(j)}$ are denoted by $R(t_{(j)})$, which is also known as the risk set. The risk set consists of all subjects with survival or censored times greater than or equal to the

specified time. Now, consider the result derived earlier from Eq.(1.3), namely,

$$f(t, \mathbf{x}, \boldsymbol{\beta}) \quad = \quad h(t, \mathbf{x}, \boldsymbol{\beta}) \times S(t, \mathbf{x}, \boldsymbol{\beta}). \tag{4.5}$$

Recall also that the likelihood function from the regression models given by Eq.(2.6) is

$$L(\beta) \quad = \quad \prod_{i=1}^{n} \{[f(t_i, \mathbf{x}_i, \beta)]^{c_i} \times [S(t_i, \mathbf{x}_i, \beta)]^{1-c_i}\}. \tag{4.6}$$

Substituting Eq.(4.5) into Eq.(4.6) yields

$$\begin{aligned} L(\beta) \quad &= \quad \prod_{i=1}^{n} \{[h(t_i, \mathbf{x}_i, \beta) \times S(t_i, \mathbf{x}_i, \beta)]^{c_i} \times [S(t_i, \mathbf{x}_i, \beta)]^{1-c_i}\} \\ &= \quad \prod_{i=1}^{n} \{[h(t_i, \mathbf{x}_i, \beta)]^{c_i} \times [S(t_i, \mathbf{x}_i, \beta)]\}. \end{aligned}$$

Substituting $h(t_i, \mathbf{x}_i, \boldsymbol{\beta}) = h_0(t_i) e^{\mathbf{x}_i^T \boldsymbol{\beta}}$ and $S(t_i, \mathbf{x}_i, \boldsymbol{\beta}) = [S_0(t)]^{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}$ into the above equation results in

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \{[h_0(t_i) e^{\mathbf{x}_i^T \boldsymbol{\beta}}]^{c_i} \times [S_0(t_i)]^{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}\} \tag{4.7}$$

$$\ln[L(\boldsymbol{\beta})] = \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{c_i \ln[h_0(t_i)] + c_i \mathbf{x}_i^T \boldsymbol{\beta} + e^{\mathbf{x}_i^T \boldsymbol{\beta}} \ln[S_0(t_i)]\}. \tag{4.8}$$

The maximum likelihood estimation method requires that Eq.(4.8) be maximised with respect to the unknown parameters, $\boldsymbol{\beta}$, and a parametric model for the baseline hazard be specified. However, the proportional hazards model is adopted in order to avoid explicitly defining the baseline hazard function.

Cox (1972) constructed a partial likelihood function (depending only on the parameters of interest) that can be maximised in order to obtain estimates for the unknown parameters. He showed that the resulting parameter estimators from the partial likelihood function would have the same distributional properties as the full maximum likelihood estimators. Suppose that $\mathbf{x}_{(j)}$ is the vector of the covariates for a subject with observed ordered survival time $t_{(j)}$. Then the partial likelihood which can be derived using the counting process

approach (Fleming and Harrington, 1991) is given by

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\mathbf{x}_{(j)}^T \boldsymbol{\beta})}{\sum_{l \in R(t_{(j)})} \exp(\mathbf{x}_l^T \boldsymbol{\beta})}. \tag{4.9}$$

Its derivation is based on a conditional probability argument. The partial likelihood given by Eq.(4.9) is for participants who experience an event. However, if c_i is the censoring variable that takes the value 0 when an observation is censored and 1 when the observation is not censored, then the above likelihood can be rewritten to account for censoring in the form

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{\sum_{l \in R(t_{(j)})} \exp(\mathbf{x}_l^T \boldsymbol{\beta})} \right]^{c_i}.$$

Since it is simpler to maximise the natural log of the likelihood (as is the usual practise) the resulting equation to be maximised is

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} c_i \left[\mathbf{x}_i^T \boldsymbol{\beta} - \ln \sum_{l \in R(t_{(j)})} \exp(\mathbf{x}_l^T \boldsymbol{\beta}) \right].$$

To maximise this log likelihood we differentiate it with respect to the unknown coefficients, $\boldsymbol{\beta}$, to obtain p equations given by

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_k} &= \sum_{j=1}^{r} \left[x_{(jk)} - \frac{\sum_{l \in R(t_{(j)})} x_{lk} \exp(\mathbf{x}_l^T \boldsymbol{\beta})}{\sum_{l \in R(t_{(j)})} \exp(\mathbf{x}_l^T \boldsymbol{\beta})} \right] \\ &= \sum_{j=1}^{r} \{ x_{(jk)} - \bar{x}_{wjk} \} \end{aligned}$$

where

$$\bar{x}_{wjk} = \sum_{l \in R(t_{(j)})} w_{jl} x_{lk} \tag{4.10}$$

and

$$w_{jl} = \frac{\exp(\mathbf{x}_l^T \boldsymbol{\beta})}{\sum_{l \in R(t_{(j)})} \exp(\mathbf{x}_l^T \boldsymbol{\beta})}. \tag{4.11}$$

Here $x_{(jk)}$ is the value of the covariate x_k for a subject with observed ordered survival time $t_{(j)}$. The estimator is obtained by setting the derivative equal to 0 and solving for $\boldsymbol{\beta}$. In general, an iterative technique needs to be employed in order to solve for the unknown parameters.

The variance of the estimator of $\boldsymbol{\beta}$ is obtained by taking the inverse of the negative of the second derivative of the log partial likelihood at the value of the estimator. Namely,

$$\hat{var}(\hat{\boldsymbol{\beta}}) = I(\hat{\boldsymbol{\beta}})^{-1}$$

where

$$I(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}.$$

The diagonal elements of the information matrix are given by

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_k^2} = -\sum_{j=1}^{r} \sum_{l \in R(t_{(j)})} w_{jl}(x_{lk} - \bar{x}_{wjk})^2$$

and the off-diagonal elements are

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_h} = -\sum_{j=1}^{r} \sum_{l \in R(t_{(j)})} w_{jl}(x_{lk} - \bar{x}_{wjk})(x_{lh} - \bar{x}_{wjh})$$

where $\bar{x}_{wjk}$ and w_{jl} are defined as in Eq.(4.10) and Eq.(4.11) respectively.

4.2.3 Tests of Significance

The significance of the estimated coefficients needs to be assessed, and it is usual practise to form confidence intervals for these estimates. The three tests commonly used are the partial likelihood ratio test, the Wald test and the score test.

Partial likelihood ratio test

The partial likelihood ratio test is the easiest test to compute, and the best of the three above-mentioned tests for assessing the significance of the fitted model (Hosmer and

Lemeshow, 1999). The test statistic is given by

$$G = 2\{\ell_p(\hat{\boldsymbol{\beta}}) - \ell_p(0)\}$$

where $\ell_p(\hat{\boldsymbol{\beta}})$ is the log partial likelihood evaluated at $\hat{\boldsymbol{\beta}}$, and $\ell_p(0) = -\sum_{i=1}^{m} n_i$, where n_i is the number of subjects in the risk set at observed survival time $t_{(i)}$. This statistic tests whether all coefficients are equal to 0 versus that at least one is non-zero. Under the null hypothesis, $H_0 : \boldsymbol{\beta} = \mathbf{0}$, G follows the χ^2 distribution with degrees of freedom equal to the number of parameters estimated in the model.

Score test

As opposed to the partial likelihood ratio test, the Wald and score tests require matrix formulations and calculations (Hosmer and Lemeshow, 1999). Let the vector of the first order partial derivatives of Eq.(4.9) be denoted by $\mathbf{u}(\boldsymbol{\beta})$. Under the null hypothesis that all coefficients are equal to 0, the vector of scores, $\mathbf{u}(\mathbf{0}) = \mathbf{u}(\boldsymbol{\beta})\rfloor_{\boldsymbol{\beta}=\mathbf{0}}$ will be multivariate normally distributed with mean vector $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix given by the information matrix evaluated at $\boldsymbol{\beta} = \mathbf{0}$, $I(\mathbf{0}) = I(\boldsymbol{\beta})\rfloor_{\boldsymbol{\beta}=\mathbf{0}}$. The score statistic is then

$$\mathbf{u}^T(\mathbf{0})[I(\mathbf{0})]^{-1}\mathbf{u}(\mathbf{0})$$

which is, under H_0, approximately χ^2 distributed with degrees of freedom equal to the number of parameters in the model. In the case of one covariate, the score test is given by

$$z^* = \left.\frac{d\ell_p/d\beta}{\sqrt{I(\beta)}}\right\rfloor_{\beta=0}$$

and under H_0, $z^* \sim N(0,1)$ or $(z^*)^2 \sim \chi^2(1)$.

Wald test

The Wald test statistic is obtained from the theory which states that under the null hypothesis the estimator of the coefficients, $\hat{\boldsymbol{\beta}}$, will be asymptotically normally distributed

with mean vector $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix estimated by $\hat{var}(\hat{\boldsymbol{\beta}}) = I(\hat{\boldsymbol{\beta}})^{-1}$ (Hosmer and Lemeshow, 1999). The multiple variable Wald test statistic is given by

$$\hat{\boldsymbol{\beta}}^T I(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}$$

which is, under H_0, χ^2 distributed with degrees of freedom equal to the number of parameters fitted in the model. In the case of a single covariate the square root of the test statistic reduces to

$$z = \frac{\hat{\beta}}{se(\hat{\beta})}$$

where $se(\hat{\beta}) = \sqrt{var(\hat{\beta})}$. Thus z is standard normally distributed under H_0, or $(z)^2 \sim \chi^2(1)$.

The confidence interval of $\hat{\beta}$ is based on the Wald statistic and can be found from the usual expression

$$\hat{\beta} \pm Z_{\alpha/2} se(\hat{\beta}).$$

In practise $\sqrt{G}, z, z^*$ should all be quite similar, resulting in the same conclusion. However, the partial likelihood ratio test is the preferred choice. An advantage of using the score test is that the statistic can be computed without evaluating the maximum partial likelihood estimates of the parameters. It is useful as a test to use in model building applications in which evaluation of the estimator is computationally intensive (Hosmer and Lemeshow, 1999).

4.3 Fitting the Proportional Hazards Model with Tied Survival Times

The partial likelihood function methods described previously are based on the assumption that there are no tied survival times among the observed survival times. Most applied settings are likely to have some tied observations. In these cases the partial likelihood function needs to be modified. The exact expression for the modified partial likelihood

function is derived by Kalbfleish and Prentice (1980) and approximations are due to Breslow (1974) and Efron (1977).

For simplicity assume only one covariate. The basis for construction of the exact partial likelihood is to assume that the d ties at a particular survival time are due to lack of precision in measuring the survival time, such as recording the times in days ignoring the fractional days. Thus the tied survival times could actually have been observed in any one of the $d!$ possible arrangements of their values (Hosmer and Lemeshow, 1999). The exact partial likelihood is obtained by modifying the denominator of

$$L_p(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)}\beta}}{\sum_{j \in R(t_{(j)})} e^{x_j \beta}}$$

to include each of these arrangements. Approximations derived by Breslow (1974) and Efron (1977) are designed to provide expressions that are easier to compute than the exact partial likelihood, and yet still account for the fact that ties are present among the observed values of survival time (Hosmer and Lemeshow, 1999). The Breslow (1974) approximation uses as the partial likelihood

$$L_{p1}(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)+}\beta}}{[\sum_{j \in R(t_{(j)})} e^{x_j \beta}]^{d_i}} \tag{4.12}$$

where d_i denotes the number of subjects with survival time $t_{(i)}$, $x_{(i)+}$ is the sum of the covariate over the d_i subjects, namely, $x_{(i)+} = \sum_{j \in D(t_{(j)})} x_j$, where $D(t_{(j)})$ represents subjects with survival times equal to $t_{(i)}$. The upper limit m is the number of distinct survival times.

The Efron (1977) approximation yields a slightly better approximation to the exact partial likelihood than the Breslow (1974) approximation, with the partial likelihood given as

$$L_{p2}(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)+}\beta}}{\prod_{k=1}^{d_i} [\sum_{j \in R(t_{(j)})} e^{x_j \beta} - \frac{k-1}{d_i} \sum_{j \in D(t_{(j)})} e^{x_j \beta}]}. \tag{4.13}$$

The modified partial likelihood function for β in the presence of ties is obtained in the same manner as in the non-tied data case, with the exception that derivatives are taken with respect to the unknown parameters in the natural log of either Eq.(4.12) or Eq.(4.13). The variance of the estimated coefficient is obtained from the second partial derivative evaluated at the estimated value of the parameter. In many applied settings there will be little practical difference between the estimators obtained from the two estimators, thus usually the Breslow (1974) approximation is used because it is simpler.

4.4 Estimating the Survivor Function of the Proportional Hazards Regression Model

Recall that the expression for the survivorship function of the proportional hazards model can be expressed as

$$S(t, \mathbf{x}, \boldsymbol{\beta}) = [S_0(t)]^{\exp(\mathbf{x}^T \boldsymbol{\beta})}.$$

Thus once the regression coefficients have been estimated, all that is needed is an estimator of the baseline survivorship function. A likelihood-based approach, which assumes that the hazard is constant between observed survival times, is the foundation of the method (Hosmer and Lemeshow, 1999). There is also a derivation based on counting processes (Fleming and Harrington, 1991). The idea of the likelihood approach is to follow the same argument that lead to the Kaplan-Meier estimator of the survivorship function.

Recall $\hat{\alpha}_i = 1 - \frac{d_i}{n_i}$ is the conditional survival probability at time $t_{(i)}$, namely,

$$\hat{\alpha}_i = \frac{S(t_{(i)})}{S(t_{(i-1)})}.$$

Define the conditional baseline survival probability as

$$\alpha_{i0} = \frac{S_0(t_{(i)})}{S_0(t_{(i-1)})}.$$

Then the conditional survival probability can be written as

$$\begin{aligned}\frac{S(t_{(i)},\mathbf{x},\boldsymbol{\beta})}{S(t_{(i-1)},\mathbf{x},\boldsymbol{\beta})} &= \frac{[S_0(t_{(i)})]^{\exp(\mathbf{x}^T\boldsymbol{\beta})}}{[S_0(t_{(i-1)})]^{\exp(\mathbf{x}^T\boldsymbol{\beta})}} \\ &= \left[\frac{S_0(t_{(i)})}{S_0(t_{(i-1)})}\right]^{\exp(\mathbf{x}^T\boldsymbol{\beta})} \\ &= \alpha_{i0}^{\exp(\mathbf{x}^T\boldsymbol{\beta})}.\end{aligned}$$

Maximum likelihood methods are employed conditional on the partial likelihood estimator of the regression coefficients in the model, $\hat{\boldsymbol{\beta}}$. Let $\hat{\theta}_l = \exp(\mathbf{x}^T\hat{\boldsymbol{\beta}})$. Then the estimator of the conditional baseline survival probability is obtained by solving for $\hat{\alpha}_{i0}$ in the equation

$$\sum_{l\in D_i}\frac{\hat{\theta}_l}{1-\alpha_{i0}^{\hat{\theta}_l}} = \sum_{l\in R_i}\hat{\theta}_l \tag{4.14}$$

where R_i denotes the subjects in the risk set at ordered observed survival time $t_{(i)}$ and D_i denotes the subjects in the risk set with survival times equal to $t_{(i)}$. If there are no tied survival times, D_i contains one subject and the solution to Eq.(4.14) is

$$\hat{\alpha}_{i0} = \left[1-\frac{\hat{\theta}_i}{\sum_{l\in R_i}\hat{\theta}_l}\right]^{\hat{\theta}_i^{-1}}.$$

If there are tied survival times , the solution to Eq.(4.14) is obtained by using iterative methods. The estimator of the baseline survivorship function is the product of the individual estimators of the conditional baseline survival probabilities (Hosmer and Lemeshow, 1999)

$$\hat{S_0(t)} = \prod_{t_{(i)}\le t}\hat{\alpha}_{i0}.$$

An approximation to the solution due to Breslow (1974) is obtained by replacing $\alpha_{i0}^{\hat{\theta}_l}$ on the left hand side of Eq.(4.14) with the approximation

$$\alpha_{i0}^{\hat{\theta}_l} = 1+\hat{\theta}_l\ln(\alpha_{i0}).$$

The solution is then

$$\tilde{\alpha}_{i0} = \exp\left[\frac{-d_i}{\sum_{l\in R_i}\hat{\theta}_l}\right]$$

and

$$\hat{S}_0(t) = \prod_{t_{(i)} \leq t} \tilde{\alpha}_{i0}.$$

The estimator of the survivorship function in $S(t, \mathbf{x}, \boldsymbol{\beta}) = [S_0(t)]^{\exp(\mathbf{x}^T \boldsymbol{\beta})}$ is obtained by substituting the estimators of the baseline survivorship function and the estimators of the parameters using covariate values of interest (Hosmer and Lemeshow, 1999).

The estimator of the baseline hazard function is given by

$$\hat{h}_0(t_{(i)}) = 1 - \hat{\alpha}_{i0}.$$

The individual pointwise estimators of the baseline hazard function will typically be too unstable to use themselves, however the use of smoothing methods can be employed to get an idea of the shape of the underlying baseline hazard function (Hosmer and Lemeshow, 1999). The estimator of the cumulative hazard function is more practical, and can be obtained from the following relationship

$$\begin{aligned} \hat{S}_0(t) &= e^{-\hat{H}_0(t)} \\ \Rightarrow \hat{H}_0(t) &= -\ln[\hat{S}_0(t)] \end{aligned}$$

and the estimator of the cumulative hazard function for a specific set of covariate values is

$$\begin{aligned} \hat{H}(t, \mathbf{x}, \hat{\boldsymbol{\beta}}) &= -\ln[\hat{S}(t, \mathbf{x}, \hat{\boldsymbol{\beta}})] \\ &= -e^{\mathbf{x}^T \hat{\boldsymbol{\beta}}} \ln[\hat{S}_0(t)]. \end{aligned}$$

When plotted against a variable of time, the cumulative hazard function may provide a useful graphical descriptor of the risk experience (Hosmer and Lemeshow, 1999).

Chapter 5

Application of the Cox Proportional Hazards Regression Model

5.1 Introduction

Cox regression is a form of semi-parametric regression modeling which therefore makes it a more robust method than parametric regression methods (Cox, 1972; Allison, 1995). It is known as the proportional hazards model because the hazard for any individual is a fixed proportion of the hazard for any other individual. A remarkable feature of this method is that the baseline hazard does not need to be specified in order to estimate the model parameters, and partial likelihood estimates are dependent only on the ranks of event times, not their numerical values (Breslow, 1974; Allison, 1995). In the sections that follow, the three data sets introduced in Chapter 3 are re-analysed using the Cox regression model.

5.2 Old Order Amish Community Data

The model fitted was

$$h(t) = h_0(t)\exp(\beta_1 x_{sex} + \beta_2 x_{byr})$$

where x_{sex} and x_{byr} are notations for the sex and birth year of the individual in question. All the 2860 observations were used in the analysis, and the convergence criterion was said to be satisfied. The method used for handling ties was the Breslow (1974) method. The

Table 5.1: Model fit statistics

Criterion	Without Covariates	With Covariates
-2LogL	31850.097	31788.014
AIC	31850.097	31792.014
SBC	31850.097	31803.466

Table 5.2: Testing global null hypothesis: $\boldsymbol{\beta} = 0$

Test	χ^2	DF	Pr $> \chi^2$
Likelihood Ratio	62.0834	2	< 0.0001
Score	64.9409	2	< 0.0001
Wald	64.7585	2	< 0.0001

global null hypothesis is that all coefficients are 0. Table 5.2 indicates that this hypothesis is rejected at a 5% significance level, concluding that at least one coefficient is not equal to 0, and that the model is significant. From Table 5.3 it can be seen that there is no intercept

Table 5.3: Analysis of maximum likelihood estimates

Variable	DF	Parameter Estimate	Std. Error	χ^2	Pr $> \chi^2$	Hazard Ratio
sex	1	0.10928	0.04218	6.7109	0.0096	1.115
byr	1	-0.00679	0.0008776	59.9098	< 0.0001	0.993

estimate, which is a characteristic feature of partial likelihood estimation. The χ^2 tests are Wald tests for the null hypothesis that each coefficient is equal to 0, and are calculated simply by the following formula

$$\text{Wald } \chi^2 \quad = \quad \left(\frac{\text{Parameter Estimate}}{\text{Std. Error}}\right)^2.$$

Both the variables sex and birth year are significant at a 5% significance level, which is the same conclusion that was reached using a generalised gamma parametric model. The hazard ratio is simply e^{β}. The interpretation is as follows. For the variable sex, men are about 1.12 times more likely to die than women, or the hazard of death for men is 112% that of the hazard for women. For birth year, which is a quantitative variable, a useful statistic is $100(e^{\beta} - 1) = 100(0.993 - 1) = -0.7$. The interpretation of this is that for each one-year increase in birth year, the hazard of death goes down by an estimated 0.7%.

Recall that the Amish community data possessed an extra family cluster variable, denoted *sib*. However, in the above analysis, the cluster effect of family is not accounted for in the model. A way to do this in SAS is to use the approach by Lee *et al.* (1992) by estimating the regression parameters in the Cox model by the partial likelihood method under an independent working assumption and then use a robust sandwich covariance matrix estimate to account for the intracluster dependence. This is achieved by adding the option COVS(AGGREGATE) in the proc phreg statement, and adding an ID statement for sibship after the model statement. The SAS code to implement this additional feature is given by

```
PROC PHREG COVS(AGGREGATE);
MODEL AGE*DLT(0) = SEX BYR;
ID SIB;
RUN;
```

Running this program in SAS produces results that are similar to those obtained before. The model based score and Wald statistics and the likelihood ratio statistic are identical to those obtained before, and Table 5.4 is the same as before except for the addition of two extra statistics signifying the correlation structure for individuals in the same cluster. The parameter estimates, and thus the hazard ratios in Table 5.5 remain unchanged, however there is a change in the standard errors in that they are slightly larger than before. The probability associated with the significance of sex as a variable is larger than before, but

Table 5.4: Testing global null hypothesis: $\boldsymbol{\beta} = 0$ (clustering included)

Test	χ^2	DF	Pr > χ^2
Likelihood Ratio	62.0834	2	< 0.0001
Score (model-based)	64.9409	2	< 0.0001
Score (sandwich)	44.5145	2	< 0.0001
Wald (model-based)	64.7585	2	< 0.0001
Wald (sandwich)	40.7163	2	< 0.0001

Table 5.5: Analysis of maximum likelihood estimates (clustering included)

Variable	DF	Parameter Estimate	Std. Error	χ^2	Pr > χ^2	Hazard Ratio
sex	1	0.10928	0.04316	6.4109	0.0113	1.115
byr	1	-0.00679	0.00119	32.3639	< 0.0001	0.993

sex is still a significant covariate at a 5% significance level.

5.3 Lung Cancer Data

The model fitted was

$$h(t) = h_0(t)\exp(\beta_1 x_{trt} + \beta_2 x_{perfstat} + \beta_3 x_{liver} + \beta_4 x_{bone} + \beta_5 x_{weightloss}).$$

The number of observations used was 570 of which 10 (1.75%) were censored. The convergence criterion was satisfied, and the Breslow (1974) method for handling ties was used. From Table 5.7 it is clear that the global null hypothesis that all coefficients are equal to 0 is strongly rejected at a 5% significance level, implying that at least one coefficient is not equal to 0. It is clear from Table 5.8 that all the variables are significant at a 5% significance level, which is what was found when fitting the generalised gamma model to the data. The interpretation of the hazard ratios are as follows. The hazard for participants on treatment 1 (the CAV-HEM arm) is 77% that of the hazard for participants on treatment 0 (the CAV arm). The hazard for participants with a performance status of 1 is about

Table 5.6: Model fit statistics

Criterion	Without Covariates	With Covariates
-2LogL	6061.813	5983.467
AIC	6061.813	5993.467
SBC	6061.813	6015.107

Table 5.7: Testing global null hypothesis: $\beta = 0$

Test	χ^2	DF	Pr> χ^2
Likelihood Ratio	78.3456	5	< 0.0001
Score	82.9221	5	< 0.0001
Wald	80.7884	5	< 0.0001

Table 5.8: Analysis of maximum likelihood estimates

Variable	DF	Parameter Estimate	Std. Error	χ^2	Pr> χ^2	Hazard Ratio
trt	1	-0.25672	0.08598	8.9151	0.0028	0.774
perfstat	1	-0.60648	0.10440	33.7437	< 0.0001	0.545
liver	1	0.41800	0.09031	21.4240	< 0.0001	1.519
bone	1	0.23242	0.09337	6.1956	0.0128	1.262
weightloss	1	0.20270	0.08762	5.3521	0.0207	1.225

50% that of the hazard for participants with a performance status of 0. The hazard for a participant with liver metastases is 1.5 times that for a participant without liver metastases. The hazard for a participant with bone metastases is about 1.3 times that for a participant without bone metastases. A participant who experiences weight loss will have a hazard about 1.2 times that of a participant who does not experience any weightloss.

In the above analysis, the clustering effect of institution has been ignored. This can be incorporated into the model using a robust sandwich covariance matrix estimate to account for the intracluster dependence. These results are displayed in Tables 5.9 and 5.10. Including this clustering effect results in the standard errors being slightly different to before, and now bone metastases is no longer a significant variable in the regression. The model fit statistics are identical to the previous analysis.

Table 5.9: Testing global null hypothesis: $\boldsymbol{\beta} = 0$ (clustering included)

Test	χ^2	DF	Pr $> \chi^2$
Likelihood Ratio	78.3456	5	< 0.0001
Score (model-based)	82.9221	5	< 0.0001
Score (sandwich)	19.1155	5	0.0018
Wald (model-based)	80.7884	5	< 0.0001
Wald (sandwich)	123.8384	5	< 0.0001

Table 5.10: Analysis of maximum likelihood estimates (clustering included)

Variable	DF	Parameter Estimate	Std. Error	χ^2	Pr $> \chi^2$	Hazard Ratio
trt	1	-0.25672	0.10615	5.8494	0.0156	0.774
perfstat	1	-0.60648	0.11972	25.6637	< 0.0001	0.545
liver	1	0.41800	0.08601	23.6171	< 0.0001	1.519
bone	1	0.23242	0.14878	2.4404	0.1182	1.262
weightloss	1	0.20270	0.10309	3.8664	0.0493	1.225

5.4 Warfarin data

The model fitted was

$$h(t) = h_0(t)\exp(\beta_1 x_{group} + \beta_2 x_{age} + \beta_3 x_{sex}).$$

The number of values used was 294, with 85.03% of the data censored. The convergence criterion was said to have been satisfied. The method for handling ties was the Breslow (1974) method.

Table 5.11: Model fit statistics

Criterion	Without Covariates	With Covariates
-2LogL	461.276	458.144
AIC	461.276	464.144
SBC	461.276	469.497

Table 5.12: Testing global null hypothesis: $\boldsymbol{\beta} = 0$

Test	χ^2	DF	Pr $> \chi^2$
Likelihood Ratio	3.1316	3	0.3718
Score	3.1377	3	0.3709
Wald	3.0863	3	0.3785

The global null hypothesis that all coefficients are zero in Table 5.12 is not rejected at a 5% significant level. This result is in perfect agreement to what was found under the parametric models, namely that the covariates are not significant in the model.

Table 5.13: Analysis of maximum likelihood estimates

Variable	DF	Parameter Estimate	Std. Error	χ^2	Pr $> \chi^2$	Hazard Ratio
Group	1	-0.47206	0.30861	2.3397	0.1261	0.624
Age	1	-0.00647	0.01126	0.3303	0.5655	0.994
Sex	1	0.11348	0.30775	0.1360	0.7123	1.120

From Table 5.13 it can be seen that none of the covariates are significant. However if one

wishes to interpret the hazard ratios it would be as follows. Individuals on adjusted dose warfarin have a hazard approximately 62% that of the hazard for individuals on fixed dose warfarin. The hazard of experiencing an event for males is 112% that of the females. For age, $100(e^{\beta}-1) = -0.6$ is calculated because it is a quantitative variable. The interpretation is that for a one-year increase in age, the hazard of experiencing an event goes down by an estimated 0.6%.

Chapter 6

Frailty Models

6.1 Introduction

There may be times when the proportional hazards model does not adequately describe the distribution of the survival time. The deviations from the proportional hazards model may sometimes be explained by unaccounted random heterogeneity, otherwise known as frailty (Keiding *et al.*, 1997). A possible cause of this is when covariates that are important in describing the survival of an individual are omitted. If standard methods are applied to this data (such as the Cox proportional hazards model) the resulting estimates will be biased. To account for frailty in a model, an unmeasured random effect is incorporated in the hazard function, under the assumption that frailty is independent of any censoring that may take place. The frailty term acts multiplicatively on the hazard function. The following introduction to frailty is based on the 1979 paper by Vaupel, Manton and Stallard.

Let $h_i(t, \mathbf{x}, z)$ be the hazard function for an individual in population group i, with a vector of covariates $\mathbf{x}$, at some time t, and with a 'frailty' of z. The definition of frailty, as defined by Vaupel *et al.* (1979), states that the ratio of the hazards for two different individuals in population group i is equal to the ratio of their frailties. Mathematically this is expressed as

$$\frac{h_i(t, \mathbf{x}, z)}{h_i(t, \mathbf{x}, z')} = \frac{z}{z'}$$

or

$$h_i(t, \mathbf{x}, z) = z h_i(t, \mathbf{x}, 1) \tag{6.1}$$

where an individual with a frailty of 1 might be viewed as a 'standard' individual. If an individual has a frailty of 2, then that person is twice as likely to die at any particular age, at any particular time, than a standard individual. On the other hand, a person with a frailty of 0.5 is only half as likely to die. In other words, if $z > 1$ then an individual is more 'frail' than a standard individual, if $z < 1$ the subject is less 'frail' than an average individual. Thus the frailties can be interpreted as relative risks.

The above definition of frailty assumes that each individual maintains a constant level of frailty, from birth to death. However, it does not imply that individuals with the same frailty are identical. Also, it is more convenient to define frailty in terms of the hazard, rather than the age-specific probability of death, q_x for the following reasons.

1. q_x is bounded above by one (because it is a probability) and thus the range of the frailty would also be bounded above.

2. q_x is a nonlinear function of the size of the age interval used.

For simplicity, let $h_i(t, \mathbf{x}, z)$ and $h_i(t, \mathbf{x}, 1)$ be denoted by $h(z)$ and h respectively. Then Eq.(6.1) can be rewritten as

$$h(z) = zh.$$

The following relationships for the cumulative hazard and hence the survivor function clearly follow

$$\begin{aligned} H(z) &= zH \\ S &= e^{-H} \\ \Rightarrow S(z) &= S^z \end{aligned} \tag{6.2}$$

where $S = S(t, \mathbf{x}, 1)$ for some vector $\mathbf{x}$ and time t.

6.2 The Distribution of Frailty

Let $\bar{h_i}(t, x)$ be the hazard for a cohort of individuals from a population group i at age x at time t. For simplicity assume only one covariate is measured, in this case age, denoted by x. Note that $\bar{h_i}(t, x)$ is analogous to the average hazard in a group of individuals. Then

$$\bar{h_i}(t, x) = \int_0^\infty h_i(t, x, z) f_x(z) \mathrm{d}z$$

where $f_x(z)$ is the p.d.f. of the frailty at age x among the surviving individuals in the cohort. Average frailty in the cohort, $\bar{z}$, is defined by

$$\bar{z_i}(t, x) = \int_0^\infty z f_x(z) \mathrm{d}z.$$

It follows from Eq.(6.1) that

$$\begin{aligned} \bar{h_i}(t, x) &= \int_0^\infty h_i(t, x, z) f_x(z) \mathrm{d}z \\ &= \int_0^\infty z h_i(t, x, 1) f_x(z) \mathrm{d}z \\ &= h_i(t, x, 1) \int_0^\infty z f_x(z) \mathrm{d}z \\ &= h_i(t, x, 1) \bar{z_i}(t, x) \end{aligned}$$

or, in simpler notation, $\bar{h} = h\bar{z}$.

Frail individuals with high values of z will tend to die first. This implies that $\bar{z}$ (which is the average frailty of the surviving cohort) will decline with age. The equation $\bar{h} = h\bar{z}$ also indicates that the hazard for individuals increases more rapidly than for the cohort in which the individuals belong (in other words individuals "age faster" than cohorts). The relationship between the individual and cohort aging depends on the distribution of frailty among individuals. Many papers and texts (Nguti, 2003; Zuma and Lurie, 2005; Bolstad

and Manda, 2001) assume that frailty is Gamma distributed with p.d.f.

$$f(z) = \frac{\lambda^k z^{k-1} e^{-\lambda z}}{\Gamma(k)}$$

where λ and k are the scale and shape parameters respectively. The mean and variance are given by

$$\begin{aligned} \bar{z} &= \frac{k}{\lambda} \\ \sigma_z^2 &= \frac{k}{\lambda^2}. \end{aligned}$$

It is common to set the mean equal to 1 so that $\lambda = k$ and $\sigma^2 = 1/k$.

There are a few reasons why the Gamma distribution is chosen for the frailty. It is analytically tractable, readily computable, and is one of a few distributions that is used to model variables that are positive, and since frailty cannot be negative it is thus suitable. It is also a flexible distribution that can take on a variety of shapes as k varies. When $k = 1$ it simplifies to the exponential distribution, and when k becomes large it assumes a bell-shaped distribution similar to that of the normal distribution. As k increases, and thus as variability in frailty decreases, mortality rates for standard individuals become more like the observed cohort rates. There are two useful mathematical results noted that arise from the assumption that frailty at birth is Gamma distributed:

1. Frailty among those that have not yet died is Gamma distributed with the same value of shape parameter as at birth, but now

$$\lambda(x) = \lambda + H(x)$$

and the mean frailty is then

$$\bar{z}(x) = \bar{z}\frac{k}{k + \bar{z}H(x)}$$

where $\bar{z}$ is the average frailty of the cohort at birth. When $k = 1$ and $\bar{z} = 1$ the mean frailty reduces to

$$\bar{z}(x) = \frac{1}{1 + H(x)}.$$

It is obvious from the above equation that as the cumulative hazard, $H(x)$, increases the average frailty of the remaining cohort decreases.

2. Frailty among those who die at any age x is also Gamma distributed, with the same scale parameter $\lambda(x)$ as among those surviving to age x, but with shape parameter $k + 1$. This implies that the mean frailty of those who die at age x, denoted by $\bar{z}'(x)$, is greater than the mean frailty of the survivors

$$\bar{z}'(x) = \bar{z}(x)\frac{k+1}{k}.$$

Thus Vaupel *et al.* (1979) concluded that ignoring frailty in a survival model may lead to biased estimates. Individual aging rates, past and future progress in reducing mortality, and mortality differentials between populations may be underestimated, and current life expectancy and potential gains in life expectancy from averting specific causes of death may be overestimated.

6.3 Univariate Semi-Parametric Frailty Models

In the previous chapters of this thesis, all the statistical models used to describe the distribution of survival time have assumed that the hazard function is completely specified by the baseline hazard function and the covariate values. However, there may be factors other than the measured covariates that significantly affect the distribution of survival time, a condition often referred to as heterogeneity of subjects (Hosmer and Lemeshow, 1999). In the early 1980s, a series of studies showed evidence of a possible bias in the estimated treatment effect when important covariates were omitted (Keiding *et al.*, 1997). Struthers

and Kalbfleisch (1986) showed that if one of two important covariates is omitted, then the effect of the other is underestimated.

A method of accounting for the heterogeneity due to omitted covariates is frailty modeling, whereby an unmeasured random effect in the hazard function is incorporated in the model. The proportional hazards frailty model assumes that for a given frailty variable z_i and covariates $\mathbf{x}_i$, individual i has a hazard function given by

$$\begin{aligned} h_i(t|z_i, \mathbf{x}_i) &= h_0(t)e^{\mathbf{x}_i^T\boldsymbol{\beta}+w_i} \\ &= z_i h_0(t)e^{\mathbf{x}_i^T\boldsymbol{\beta}} \end{aligned}$$

where $z_i = e^{w_i}$, and w_i is the random effect for the i^{th} individual. The participants who experience an event contribute the product of their conditional hazards function and conditional survival function to the likelihood whereas those who do not experience an event, implying that they are right censored, contribute only their conditional survival function to the likelihood. The conditional survival function is given by

$$\begin{aligned} S(t|z_i, \mathbf{x}_i) &= \exp[-H(t|z_i, \mathbf{x}_i)] \\ &= \exp[-z_i\Lambda_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta})] \end{aligned}$$

where $\Lambda_0(t)$ is the integrated or cumulative baseline hazard function at time t.

6.4 Multivariate Semi-Parametric Frailty Models

6.4.1 Multivariate Survival Data

Generally, when modeling time-to-event data, the underlying assumption is that the event times among the individuals are independent. However, this may not always be the case as the failure times of certain individuals may be correlated, for example, individuals from the same family or community may have correlated failure times. Thus the independence assumption is violated, and these data are referred to as multivariate survival data. Time-to-event data that are not correlated are known as univariate survival data.

Parallel and longitudinal data are the two main types of multivariate survival data. Parallel data consist of different clusters which have a number of items or individuals contained in them. Longitudinal data are a result of a stochastic process of events, namely an individual experiences a number of the same event over time, which results in recurrent data. The cluster is now the individual, and within that individual the events are observed. In both types of data, events within a cluster are correlated. It is thought that there are unobserved risk factors that explain the dependence. These factors are generally assumed to be constant over time, and using the standard approach of modeling survival data (for example using Cox's proportional hazards model) may lead to biased estimates.

6.4.2 The Shared Frailty Model

The shared frailty model assumes that all individuals within the same cluster or group have the same frailty. To account for the heterogeneity among groups a random effect (or the frailty effect) is included in the hazard function to account for correlation of failure times within a cluster. The frailty model can be seen as a linear mixed effects model with a frailty term acting multiplicatively on the hazard function. The frailties are assumed to be independent between or across clusters, whilst the failure times of individuals within a cluster are dependent. However, conditional on frailties, the failure times are independent.

In contrast, for univariate data, when there is no clustering effect, a frailty term is assumed for each individual, and is thought to represent unmeasured covariates. Adding frailty effects for each individual is thought to induce heterogeneity among individuals after taking into account any measured covariates.

6.4.3 Model Formulation

Suppose that there are n individuals, assigned to I groups, where the i^{th} group has n_i individuals such that $\sum_{i=1}^{I} n_i = n$. Suppose that the number of events experienced by the i^{th} group is given by $D_i = \sum_{j=1}^{n_i} \delta_{ij}$, where δ_{ij} is the censoring indicator which takes on

the value 1 when an event occurs and 0 when it does not. Then the hazard for the j^{th} individual from the i^{th} group is given by

$$h_{ij}(t) = h_0(t)\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta} + w_i) \tag{6.3}$$

where $\mathbf{x}_{ij}$ is a vector of p covariates for individual j in group i, $h_0(t)$ is the baseline hazard, and w_i is the random effect for the i^{th} group. The w_i's are an independently and identically distributed random sample from a density $f_W(\cdot)$. The model can be rewritten in the following manner

$$\begin{aligned} h_{ij}(t) &= h_0(t)\exp(w_i)\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta}) \\ &= z_i h_0(t)\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta}) \end{aligned} \tag{6.4}$$

where $z_i = \exp(w_i)$ is the frailty term. The z_i's are independent, and are assumed to have common density $f_Z(\cdot)$. The two commonly used densities for the frailties typically chosen are

1. The zero-mean normal density for W which transforms to the log-normal density for Z, that is,

$$f_Z(z) = \frac{1}{z\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\log z)^2}{2\sigma^2}\right)$$

with mean $e^{\sigma^2/2}$ and variance $e^{\sigma^2}(e^{\sigma^2}-1)$.

2. The one-parameter gamma density for Z, with density given by

$$f_Z(z) = \frac{\alpha^\alpha z^{\alpha-1}e^{-\alpha z}}{\Gamma(\alpha)} \tag{6.5}$$

which corresponds to a log-gamma density for W. The mean and variance for Z are given by

$$\begin{aligned} E[Z] &= 1 \\ Var[Z] &= \frac{1}{\alpha}. \end{aligned}$$

Since z_i in Eq.(6.4) can be thought of as a mixing term, its corresponding density, $f_Z(\cdot)$ is also referred to as a mixing distribution. When using the log-normal density for $f_Z(\cdot)$, $Var[Z] = \sigma_z^2$ is used to describe the heterogeneity among the groups, whereas $Var[Z] = 1/\alpha$ is used when a gamma density is assumed for the frailties. For generality, assume the heterogeneity can be described by a parameter θ, meaning σ_z^2 for the log-normal density and $1/\alpha$ for the gamma density. If $1/\alpha$ is small then the gamma and log-normal distributions are similar (Kalbfleish and Prentice, 1980).

The gamma distribution is extensively used in frailty modeling for many reasons, described in the section on univariate frailty modeling. However the log-normal distribution is more flexible than the gamma in creating correlated frailties, making it a very useful distribution when modeling multivariate frailty models. Other distributions that can be used for the frailties are the stable distribution and the power variance functions (Hougaard, 2000). The power variance function is a large family of distributions that include the gamma and positive stable distributions, thus rendering it a less restrictive function to use. However, the calculations are more difficult for this function, resulting in it being used less frequently.

6.5 Estimation in the Frailty Model

In Eq.(6.4), the baseline hazard, $h_0(t)$ can be specified explicitly, or left unspecified. Under a parametric assumption for $h_0(t)$, parameters in the resulting model can be estimated using maximum likelihood estimation (M.L.E.) procedures. However, if $h_0(t)$ is left unspecified, then the unknown parameters in the shared frailty model have to be estimated by various approaches or methods such as

1. Expectation Maximization (EM) algorithm (Klein, 1992)

2. Penalized Partial Likelihood (PPL) approach (Therneau and Grambsch, 2000)

3. Markov Chain Monte Carlo (MCMC) methods (Vaida and Xu, 2000)

4. Monte Carlo EM (MCEM) approach (Ripatti *et al.*, 2002)

5. Different methods using Laplace approximation (Ripatti and Palmgren, 2000, Cortinas Abrahantes and Burzykowski, 2004).

The choice of estimation method depends largely on the choice of frailty distribution. When a gamma frailty is assumed, the EM algorithm can be used. However, when a log-normal frailty is used, the estimation procedures are based on numerical integration methods such as the Laplace approximation methods. This thesis shall focus on the EM algorithm, Penalized Partial Likelihood approach, and MCMC methods as a means of estimation.

6.5.1 The Expectation-Maximisation Algorithm

Introduction

Suppose that we have a probability density function $f(\mathbf{x}|\Theta)$ that has a set of parameters Θ, and data set of size N drawn from this distribution. Assuming that the data are independent and identically distributed (*iid*), the resulting likelihood function is

$$\begin{aligned} L(\Theta|X) &= f(X|\Theta) \\ &= \prod_{i=1}^{N} f(x_i|\Theta). \end{aligned}$$

The likelihood is thought of as a function of the parameters Θ where the data X are fixed (Bilmes, 1998). We wish to find a set of parameters that maximises the likelihood, $L(\Theta|X)$. Generally, the log-likelihood is maximised instead of the likelihood because it is analytically easier.

Depending on the form of $f(\mathbf{x}|\Theta)$, the problem of maximisation may be easy or difficult. In some cases it is sufficient to take the derivative of the log-likelihood and equate it to zero, and the parameter estimates can be extracted exactly. However, in many cases it is not possible to find an analytical expression for the parameters, and other methods need to be employed.

The basic EM algorithm

The EM algorithm is a general method of finding the maximum likelihood estimates of the parameters of an underlying distribution from a given data set when the data is incomplete or has missing values (Bilmes, 1998).

Assume that X is the observed data set, and is generated by some distribution. Then X is known as the incomplete data set. Suppose that a complete data set exists, and is denoted by Z, where $Z = (X, Y)$. Assume a joint density function,

$$f(\mathbf{z}|\Theta) = f(\mathbf{x}, \mathbf{y}|\Theta) = f(\mathbf{y}|\mathbf{x}, \Theta) f(\mathbf{x}|\Theta)$$

by conditional probability arguments. From this joint density function a new likelihood function can be defined as

$$L(\Theta|Z) = L(\Theta|X, Y) = f(X, Y|\Theta).$$

This likelihood is the complete-data likelihood. The original likelihood, $L(\Theta|X)$ is known as the incomplete-data likelihood. $L(\Theta|X, Y)$ can be thought of as a random variable since the missing information Y is unknown, random, and comes from an underlying distribution. In the context of this thesis, the missing information Y is the frailty, and the underlying distribution is generally assumed to be the gamma or log-normal distribution.

The EM algorithm first finds the expected value of the complete-data log-likelihood $\log f(X, Y|\Theta)$ with respect to the unknown data Y, given the observed data X and the current parameter estimates (Bilmes, 1998). Define

$$Q(\Theta, \Theta^{(i-1)}) \quad = \quad E[\log f(X, Y|\Theta)|X, \Theta^{(i-1)}] \tag{6.6}$$

where $\Theta^{(i-1)}$ are the current parameter estimates used to evaluate the expectation, and Θ are the new parameters that are optimized to increase Q. Here X and $\Theta^{(i-1)}$ are constants, Θ is a normal variable which we wish to adjust, and Y is a random variable governed by

the distribution $f(\mathbf{y}|X,\Theta^{(i-1)})$. From the following result

$$E[h(Y)|X=x] \quad = \quad \int_y h(y) f_{Y|X}(y|x)\mathrm{d}y$$

the right side of Eq.(6.6) can be re-written as

$$E[\log f(X,Y|\Theta)|X,\Theta^{(i-1)}] \quad = \quad \int_{\mathbf{y}\in\Upsilon} \log f(X,\mathbf{y}|\Theta) f(\mathbf{y}|X,\Theta^{(i-1)})\mathrm{d}\mathbf{y} \tag{6.7}$$

where Υ is the space of values that $\mathbf{y}$ can take. Note that $f(\mathbf{y}|X,\Theta^{(i-1)})$ is the marginal distribution of the unobserved data and is dependent on both the observed data x and on the current parameters. This marginal distribution may be a simple analytical expression of the assumed parameters $\Theta^{(i-1)}$ and perhaps the data. However this density may be very difficult to obtain. Sometimes the density actually used is $f(\mathbf{y},X|\Theta^{(i-1)}) = f(\mathbf{y}|X,\Theta^{(i-1)})f(X|\Theta^{(i-1)})$. This does not affect subsequent steps of the EM algorithm, because the extra factor $f(X|\Theta^{(i-1)})$ does not depend on Θ.

The evaluation of this expectation is called the E-step of the algorithm. The second step, called the M-step of the EM algorithm is to maximise the expectation computed in the first step. In other words, we find

$$\Theta^{(i)} \quad = \quad \mathrm{argmax}_{\Theta} Q(\Theta,\Theta^{(i-1)})$$

These two steps are repeated as often as necessary until the convergence criterion are reached. Each iteration is guaranteed to increase the log-likelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood function (Bilmes, 1998). A modified form of the M-step is to find some $\Theta^{(i)}$ such that $Q(\Theta^{(i)},\Theta^{(i-1)}) > Q(\Theta,\Theta^{(i-1)})$ instead of maximising $Q(\Theta,\Theta^{(i-1)})$. This is known as the Generalised EM (GEM) algorithm, and is also guaranteed to converge (Bilmes, 1998). For more detailed properties and proofs of the EM algorithm, Dempster *et al.* (1977) can be referred to.

The EM algorithm applied to gamma frailty models

For simplicity consider a univariate analysis with the following hazard function for individual i

$$h_i(t|z_i, \mathbf{x}_i) = z_i h_0(t) e^{\mathbf{x}_i^T \boldsymbol{\beta}}.$$

Suppose that the baseline hazard is constant, $h_0(t) = h_0$. Also assume that frailty is Gamma(α, α) distributed. The individuals that experienced an event contribute the product of their hazard and survival function, whereas those individuals who are censored contribute only the survival function to the likelihood. From the relationship between the survival function and hazard function given in Eq.(1.4), the survival function can be found to be

$$\begin{aligned} S(t) &= e^{-\int_o^t h(u)\mathrm{d}u} \\ &= e^{-\int_0^t z_i h_0 e^{\mathbf{x}_i^T \boldsymbol{\beta}} \mathrm{d}u} \\ &= e^{-z_i h_0 t e^{\mathbf{x}_i^T \boldsymbol{\beta}}}. \end{aligned}$$

The complete-data likelihood is then given by

$$\begin{aligned} L_i(z_i, t_i; \alpha) &= f(z) \times [S_i(t_i) h_i(t_i)]^{\delta_i} [S_i(t_i)]^{1-\delta_i} \\ &= \frac{\alpha^\alpha}{\Gamma(\alpha)} z_i^{\alpha-1} e^{-\alpha z_i} \\ &\times [e^{-z_i h_0 t_i e^{\mathbf{x}_i^T \boldsymbol{\beta}}} z_i h_0 e^{\mathbf{x}_i^T \boldsymbol{\beta}}]^{\delta_i} [e^{-z_i h_0 t_i e^{\mathbf{x}_i^T \boldsymbol{\beta}}}]^{1-\delta_i}. \end{aligned} \tag{6.8}$$

The associated complete-data log-likelihood is then

$$\begin{aligned} \ell_i(\alpha; z_i, t_i) &= \alpha \ln \alpha - \ln \Gamma(\alpha) + (\alpha - 1) \ln z_i - \alpha z_i \\ &+ \delta_i [-z_i h_0 t_i e^{\mathbf{x}_i^T \boldsymbol{\beta}} + \ln z_i + \ln h_0 + \mathbf{x}_i^T \boldsymbol{\beta}] \\ &- (1 - \delta_i) z_i h_0 t_i e^{\mathbf{x}_i^T \boldsymbol{\beta}}. \end{aligned}$$

The observed data likelihood is attained by integrating out unobserved data from the likelihood given in Eq.(6.8) (Zuma and Lurie, 2005).

$$\begin{aligned} L_{obs,i}(t_i;\alpha) &= \int_0^\infty \frac{\alpha^\alpha}{\Gamma(\alpha)} z_i^{\alpha-1} e^{-\alpha z_i} \\ &\times [e^{-z_i h_0 t_i e^{\mathbf{x}_i^T\boldsymbol{\beta}}} z_i h_0 e^{\mathbf{x}_i^T\boldsymbol{\beta}}]^{\delta_i} [e^{-z_i h_0 t_i e^{\mathbf{x}_i^T\boldsymbol{\beta}}}]^{1-\delta_i} \mathrm{d}z_i \\ &= \frac{\alpha^\alpha}{\Gamma(\alpha)} \int_0^\infty z_i^{\alpha-1} e^{-\alpha z_i} e^{-\delta_i z_i h_0 t_i e^{\mathbf{x}_i^T\boldsymbol{\beta}}} \\ &\times z_i^{\delta_i} h_0^{\delta_i} e^{\delta_i \mathbf{x}_i^T\boldsymbol{\beta}} e^{-(1-\delta_i) z_i h_0 t_i e^{\mathbf{x}_i^T\boldsymbol{\beta}}} \mathrm{d}z_i \\ &= \frac{\alpha^\alpha}{\Gamma(\alpha)} h_0^{\delta_i} e^{\delta_i \mathbf{x}_i^T\boldsymbol{\beta}} \int_0^\infty z_i^{\alpha+\delta_i-1} e^{-z_i(\alpha + h_0 t_i e^{\mathbf{x}_i^T\boldsymbol{\beta}})} \mathrm{d}z_i \end{aligned}$$

The integral here appears to be the kernel of a Gamma$(\alpha+\delta_i, \frac{1}{\alpha + h_0 t_i e^{\mathbf{x}_i^T\boldsymbol{\beta}}})$ function. The resulting integrand of the above equation is then

$$L_{obs,i}(t_i;\alpha) = \frac{\alpha^\alpha (h_0 e^{\mathbf{x}_i^T\boldsymbol{\beta}})^{\delta_i} \Gamma(\alpha+\delta_i)}{\Gamma(\alpha)(\alpha + h_0 t_i e^{\mathbf{x}_i^T\boldsymbol{\beta}})^{\alpha+\delta_i}}.$$

To estimate the parameters, the log-likelihood of the observed data likelihood needs to be maximised

$$\begin{aligned} \ell_{obs,i}(\alpha;t_i) &= \alpha \ln \alpha + \delta_i \ln h_0 + \delta_i \mathbf{x}_i^T\boldsymbol{\beta} + \ln \Gamma(\alpha+\delta_i) \\ &- \ln \Gamma(\alpha) - (\alpha+\delta_i) \ln(\alpha + h_0 t_i e^{\mathbf{x}_i^T\boldsymbol{\beta}}). \end{aligned}$$

This log-likelihood is difficult to maximise as it contains the unspecified baseline hazard; thus the EM algorithm needs to be employed to solve for the unknown parameters. In order to run the EM algorithm the complete-data log-likelihood and the marginal distribution of the unobserved data is needed, as is indicated by Eq.(6.7). The marginal distribution of the unobserved data is found to be

$$\begin{aligned} g(z_i|x_i;\alpha) &= \frac{L_i(z_i,t_i;\alpha)}{L_{obs,i}(t_i;\alpha)} \\ &= \frac{(\alpha + h_0 t_i e^{\mathbf{x}_i^T\boldsymbol{\beta}})^{\alpha+\delta_i}}{\Gamma(\alpha+\delta_i)} Z_i^{\alpha+\delta_i-1} e^{-z_i(\alpha + h_0 t_i e^{\mathbf{x}_i^T\boldsymbol{\beta}})} \end{aligned}$$

after some simplification of the expression. This is a two parameter gamma distribution with parameters $(\alpha + \delta_i, \alpha + h_0 t_i e^{\mathbf{x}_i^T \boldsymbol{\beta}})$. The EM algorithm can now be used to solve for the unknown parameters. Note that if the baseline hazard is not constant, then

$$\int_0^t h_0(u)\mathrm{d}u \quad = \quad \Lambda_0(t).$$

6.5.2 The Penalized Partial Likelihood Approach

Penalized partial likelihood estimation originates from cubic splines regression in the Cox proportional hazards model (Nguti, 2003). When using the penalized partial likelihood approach for estimation, the random effects w_i are used rather than the frailties, z_i. Once again, for simplicity, assume the univariate frailty model, with the corresponding equations

$$\begin{aligned}
h_i(t) &= h_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta} + w_i) \\
S_i(t) &= \exp[-\Lambda_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta} + w_i)] \\
L_i(w_i, t_i; \alpha) &= f_w(w_i) \times h_i(t)^{\delta_i} S_i(t) \\
\ell_i(\alpha; w_i, t_i) &= \ln f_w(w_i) + \delta_i \ln h_i(t) + \ln S_i(t).
\end{aligned}$$

Substituting into the full data log-likelihood gives the contribution for individual i as

$$\begin{aligned}
\ell_i(\alpha; w_i, t_i) &= \ln f_w(w_i) + \delta_i \left(\ln[h_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta} + w_i)]\right) - \Lambda_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta} + w_i) \\
&= \ln f_w(w_i) + \delta_i(\ln h_0(t) + \mathbf{x}_i^T\boldsymbol{\beta} + w_i) - \Lambda_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta} + w_i).
\end{aligned}$$

The full data log-likelihood can be written as

$$\tilde{\ell}_{full}(\boldsymbol{\beta}, \alpha, h_0) \quad = \quad \tilde{\ell}_{full,1}(\boldsymbol{\beta}, h_0) + \tilde{\ell}_{full,2}(\alpha)$$

where

$$\begin{aligned}
\tilde{\ell}_{full,1}(\boldsymbol{\beta}, h_0) &= \sum_{i=1}^{I} \left[\delta_i(\ln h_0(t) + \mathbf{x}_i^T\boldsymbol{\beta} + w_i) - \Lambda_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta} + w_i)\right] \\
\tilde{\ell}_{full,2}(\alpha) &= \sum_{i=1}^{I} \ln f_w(w_i)
\end{aligned}$$

where $\tilde{\ell}_{full,2}(\alpha)$ can be seen as the penalty term, and the mean for the $w_i's$ is 0. For $w_i << 0$ or $w_i >> 0$, $f_w(w_i)$ is small, and thus $\log f_w(w_i)$ takes a large negative value which in turn decreases the likelihood, in other words it acts like a penalty. We therefore take

$$\tilde{\ell}_{full,2}(\alpha) \quad = \quad -\ell_{pen}(\alpha)$$

with

$$\ell_{pen}(\alpha) \quad = \quad -\sum_{i=1}^{I} \ln f_w(w_i).$$

In order to apply semi-parametric ideas, consider the w_i's in $\tilde{\ell}_{full,1}(\boldsymbol{\beta}, h_0)$ as 'parameters' with corresponding covariates similar to that of a design matrix in the equation

$$Y \quad = \quad X\boldsymbol{\beta} + Z\mathbf{w}$$

where Z is the design matrix (Janssen, 2005). Using partial likelihood ideas, $\tilde{\ell}_{full,1}(\boldsymbol{\beta}, h_0)$ is replaced by

$$\ell_{part}(\boldsymbol{\beta}, w) \quad = \quad \sum_{i=1}^{I} \delta_i \left[\eta_i - \ln \left(\sum_{qw \in R(t_i)} \exp(\eta_{qw}) \right) \right]$$

where

$$\eta_i \quad = \quad \mathbf{x}_i^T \boldsymbol{\beta} + w_i.$$

Thus in order to make inference on the parameters $\boldsymbol{\beta}$ and α, the following penalized partial likelihood is used

$$\ell_{ppl}(\beta, \alpha, w) \quad = \quad \ell_{part}(\boldsymbol{\beta}, w) - \ell_{pen}(\alpha, w).$$

Application to loggamma density

If the frailties are assumed to be gamma distributed, then the random terms, the w_i's, are loggamma distributed, with probability density function

$$f_w(w) \quad = \quad \frac{\alpha^\alpha (\exp(w))^\alpha \exp[-\alpha \exp(w)]}{\Gamma(\alpha)}.$$

Taking the natural log of this p.d.f. results in

$$\ln f_w(w) \quad = \quad \alpha(w - \exp(w)) - (\alpha \ln \alpha + \ln \Gamma(\alpha)) \, .$$

Hence $\ell_{pen}(\alpha)$ is given by

$$\ell_{pen}(\alpha) \quad = \quad -\sum_{i=1}^{I} (\alpha(w_i - \exp(w_i))) + I \, (\alpha \ln \alpha + \ln \Gamma(\alpha)) \, .$$

The maximisation of the penalized partial likelihood consists of an inner and outer loop (Nguti, 2003). In the inner loop the rule is that given a provisional value for α, the Newton Raphson procedure is employed to maximise $\ell_{ppl}(\boldsymbol{\beta}, \alpha, \mathbf{w})$ for $\boldsymbol{\beta}$ and $\mathbf{w}$ to obtain the best linear unbiased predictors (BLUP's). In the outer loop, a log-likelihood similar to $\ell_{obs}(\cdot)$ is maximised for α as in the case of the EM algorithm. Let ℓ denote the outer loop index, and k the inner loop index. Let $\alpha^{(\ell)}$ be the estimate for α at the ℓ^{th} iteration of the outer loop. Then $\boldsymbol{\beta}^{(\ell,k)}$ and $\mathbf{w}^{(\ell,k)}$ are the estimates and predictions for $\boldsymbol{\beta}$ and $\mathbf{w}$ at the k^{th} iterative step, given $\alpha^{(\ell)}$. The starting value for $\boldsymbol{\beta}$ is obtained from the estimates from fitting a normal cox model, and for starting values of $\mathbf{w}^{(1,0)}$ and $\alpha^{(1)}$ the k^{th} iterative step is given by

$$\begin{bmatrix} \boldsymbol{\beta}^{(\ell,k)} \\ \mathbf{w}^{(\ell,k)} \end{bmatrix} \quad = \quad \begin{bmatrix} \boldsymbol{\beta}^{(\ell,k-1)} \\ \mathbf{w}^{(\ell,k-1)} \end{bmatrix} - V^{-1} \begin{bmatrix} \mathbf{0} \\ \left[\alpha^{(\ell)}\right]^{-1} \mathbf{w}^{(\ell,k-1)} \end{bmatrix} + V^{-1} \begin{bmatrix} X & Z \end{bmatrix} \frac{\mathrm{d}\ell_{part}(\boldsymbol{\beta}, \mathbf{w})}{\mathrm{d}\boldsymbol{\eta}}$$

where

$$\begin{aligned} V \quad &= \quad \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \\ &= \quad \begin{bmatrix} X^T \\ Z^T \end{bmatrix} \left(\frac{-\partial^2 \ell_{part}(\boldsymbol{\beta}, \mathbf{w})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \right) \begin{bmatrix} X & Z \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \left[\alpha^{(\ell)}\right]^{-1} I_I \end{bmatrix} \end{aligned}$$

where $X = [\mathbf{x_{11}}, \ldots, \mathbf{x_{In_I}}]^T$ is an $n \times p$ covariate matrix with $n = \sum_{i=1}^{I} n_i$, $Z = \mathrm{diag}(\mathbf{1}_{n1}, \ldots, \mathbf{1}_{nI})$ with $\mathbf{1}_{ni}$ as a column vector of size n_i with all entries one, and $\boldsymbol{\eta} = X\boldsymbol{\beta} + Z\mathbf{w}$, such that $\boldsymbol{\eta}^T = (\boldsymbol{\eta}_{11}, \ldots, \boldsymbol{\eta}_{In_I})$ (Nguti, 2003).

Once the Newton Raphson procedure has converged for the current value of $\alpha^{(\ell)}$, the procedure moves to the outer loop of the algorithm. In the outer loop of the algorithm, a golden section search (Press *et al.*, 1992), described in Appendix A, is applied to a modified version of the log-likelihood in order to update the estimate of α. This likelihood is

$$\begin{aligned}\ell_{part,obs}(\boldsymbol{\beta}, \mathbf{w}) &= \ell_{part}(\boldsymbol{\beta}, \mathbf{w}) \\ &+ \sum_{i=1}^{I}\left[\ln\left(\frac{\Gamma(D_i+\alpha)}{\Gamma(\alpha)}\right) + \alpha\ln\left(\frac{\alpha}{\Lambda_i+\alpha}\right) - D_i\ln(D_i+\alpha+D_i\right]\end{aligned}$$

and the details of how this is derived are available in Nguti (2003) and Therneau and Grambsch (2000).

The algorithm continues until the stopping criterion given by

$$|\ell_{part,obs}(\hat{\beta}^{(\ell)}, \alpha^{(\ell)}, \hat{w}^{(\ell)}) - \ell_{part,obs}(\hat{\beta}^{(\ell-1)}, \alpha^{(\ell-1)}, \hat{w}^{(\ell-1)}| < \varepsilon*$$

is reached.

6.5.3 The Bayesian Approach

Introduction to Bayesian inference

Bayesian analysis involves explicitly using probability for quantifying uncertainty in inferences based on statistical data analysis. Gelman *et al.* (1995) lays out 3 steps to follow when performing Bayesian data analysis. These are

1. Setting up a full probability model
2. Conditioning on observed data
3. Evaluating the fit of the model and the implications of the resulting posterior distribution.

The first step involves finding a joint probability distribution for all observable and unobservable quantities. The second step involves calculating and interpreting the

appropriate posterior distribution. After the model has been evaluated in the third step, it may be necessary to change or adapt the model. The difficult part of Bayesian analysis is knowing which models to use and what assumptions to make, which is why it is important to evaluate the fitted model.

Suppose that the parameter θ denotes the unobservable vector quantities or population parameters of interest, let y denote the observed data and $\tilde{y}$ denote the unknown, but potentially observable quantities. Let $f(\cdot|\cdot)$ denote a conditional probability density, and $f(\cdot)$ a marginal distribution.

Bayesian statistical conclusions about a parameter θ, or unobserved data $\tilde{y}$ are made in terms of probability statements which are conditional on the observed value of y and on known values of any covariates, x (Gelman *et al.*, 1995). These are written as $f(\theta|y)$ or $f(\tilde{y}|y)$.

The basis of all Bayesian analysis is Bayes rule. It states that the joint probability distribution can be written as a product of the prior distribution and the sampling distribution

$$f(\theta, y) = f(\theta)f(y|\theta).$$

The posterior density is found by conditioning on y, yielding Bayes Theorem,

$$\begin{aligned} f(\theta|y) &= \frac{f(\theta, y)}{f(y)} \\ &= \frac{f(\theta)f(y|\theta)}{f(y)} \end{aligned} \tag{6.9}$$

where $f(y) = \sum_\theta f(\theta)f(y|\theta)$ in the discrete case, and $f(y) = \int_\theta f(\theta)f(y|\theta)\mathrm{d}\theta$ in the continuous case. An equivalent form of Eq.(6.9) omits $f(y)$ since it does not depend on θ; and with fixed y, $f(y)$ is a constant. In this case Eq.(6.9) can be written

$$f(\theta|y) \propto f(\theta)f(y|\theta).$$

The main task of any application is to develop the joint probability distribution, $f(\theta, y)$, and then to perform the necessary computations to summarize $f(\theta|y)$ in appropriate ways (Gelman *et al.*, 1995).

Introduction to monte carlo methods

Monte carlo methods can be generally viewed as statistical simulation techniques. In most applications of MC methods, all that is required are the p.d.f.'s that describe the system as the process is simulated directly. Thus there is no need to write down the sometimes complex differential equations that describe the behaviour of the system. Once the p.d.f.'s are known, MC simulation proceeds by random sampling from the p.d.f.'s (Drakos, 1994). Usually many simulations are performed and the estimated parameters are taken to be an average over the number of observations. If the variance of the resulting parameter estimate can be predicted, an estimate of the number of MC trials can be calculated.

In the context of this thesis, the focus of the simulation techniques is on Markov Chain Monte Carlo methods. The basic difference is that the sequence of generated points is similar to a random walk, and the probability of jumping from one point to another depends only on the last point (D'Agostini, 2003).

Suppose a sequence $\{X_0, X_1, X_2, \ldots\}$ of random variables is generated such that at each time $t \geq 0$ the next state X_{t+1} depends only on the current state of the chain. That is, X_{t+1} is sampled from a transition kernel $P(X_{t+1}|X_t)$ of the chain, and the state X_{t+1} does not depend on the rest of the chain, $\{X_0, X_1, X_2, \ldots, X_{t-1}\}$ (Gilks *et al.*, 1996). The resulting sequence is known as a markov chain.

The initial state of the chain is eventually 'forgotten' under certain regularity conditions, and $P^{(t)}(.|X_0)$ will eventually converge to a unique stationary distribution. This distribution does not depend on t or X_0 (Gilks *et al.*, 1996). After a sufficiently long burn-in period of say m iterations, the points $\{X_t : t = m + 1, \ldots, n\}$ will be dependent samples from the

stationary distribution (Gilks *et al.*, 1996).

Bayesian modeling of frailty data

Once again, for simplicity, assume a univariate frailty model. Suppose that the formulation of the model is the same as for the EM algorithm approach. In other words, the hazard function can be written as

$$h_i(t|z_i, \mathbf{x}_i) = z_i h_0(t) e^{\mathbf{x}_i^T \boldsymbol{\beta}}.$$

The corresponding survivor function is then

$$S_i(t) = \exp[-z_i \Lambda_0(t) \exp(\mathbf{x}_i^T \boldsymbol{\beta})]$$

and the time to death density is given by

$$f_i(t) = S_i(t) \times h_i(t|z_i, \mathbf{x}_i).$$

The likelihood is then

$$\begin{aligned} L_i(z_i, t_i; \theta) &= (f_i(t))^{\delta_i} \times (S_i(t))^{1-\delta_i} \\ &= z_i^{\delta_i} (h_0(t) e^{\mathbf{x}_i^T \boldsymbol{\beta}})^{\delta_i} \exp(-z_i \Lambda_0(t) e^{\mathbf{x}_i^T \boldsymbol{\beta}}). \end{aligned}$$

To fully specify the model from a Bayesian perspective, prior distributions need to be assigned to the fixed effects, the random effect and the hyperparameter for the distribution of frailty. Following the procedure of Bolstad and Manda (2001), the prior for the fixed effects, $\boldsymbol{\beta}$, is normally distributed with mean vector $\mathbf{d_0}$ and diagonal covariance matrix $D_0 = cI$, where c is a number. The prior mean is assumed to be 0 because the fixed effects represent logarithms of relative risks (Bolstad and Manda, 2001).

The frailty effect represents a relative risk, and thus should have a mean equal to 1. It is modelled as an independent draw from a gamma distribution with both parameters equal to α. The hyperparameter α is modelled as an independent draw from a gamma distribution that has parameters υ and κ (Bolstad and Manda, 2001).

A good visual description of the parameters and hyperparameters can be seen in a directed graph. Each parameter node is represented by a circle, and each fixed node is represented by a square. Data and prior constants are examples of fixed nodes. Arrows show direction from a parent node to a child node (Bolstad and Manda, 2001). Factors that affect a node directly are known as parent nodes, and nodes that directly depend on a node are called children nodes. This graph represents conditional independence between the nodes. In other words, for any node, if the parent nodes are known then no other node contains information about that particular node other than it's children nodes (Zuma, 2005).

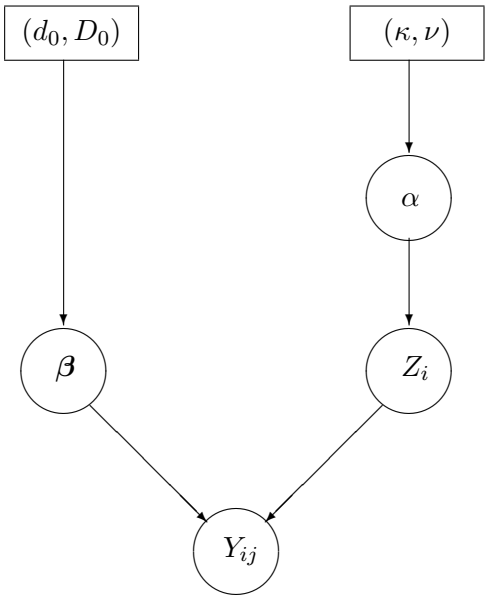

Figure 6.1: The directed acyclic graph representation of a frailty model

The joint distribution of the data and the parameters is given by

$$\begin{aligned} f(data, \boldsymbol{\beta}, z_i, \alpha) &= f(\boldsymbol{\beta}) f(\alpha) \times \prod_{i=1}^{I} f(z_i|\alpha) \times L_i(z_i, t_i; \theta) \\ &= (2\pi)^{-p/2} |D_0|^{-1/2} \exp[-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{d_0})^T D_0^{-1} (\boldsymbol{\beta} - \mathbf{d_0})] \\ &\times \frac{\upsilon^\kappa}{\Gamma(\kappa)} \alpha^{k-1} e^{-\upsilon\alpha} \\ &\times \prod_{i=1}^{I} \frac{\alpha^\alpha}{\Gamma(\alpha)} z_i^{\alpha-1} e^{-\alpha z_i} \left(z_i h_0(t) e^{\mathbf{x_i}^T \boldsymbol{\beta}} \right)^{\delta_i} \exp(-z_i \Lambda_0(t) e^{\mathbf{x}_i^T \boldsymbol{\beta}}). \end{aligned}$$

The analytic solution of the problem requires determining the posterior distribution of the parameters and hyperparameters, given the observed data. In other words, parameters not of interest need to be integrated out of the joint distribution to obtain the marginals

$$\begin{aligned} f(\beta_1|data, z_i, \alpha) &= \int \dots \int f(data, \boldsymbol{\beta}, z_i, \phi) \partial\beta_2 \dots \partial\beta_k \partial z_i \partial\alpha \\ &\vdots \\ f(\beta_k|data, z_i, \alpha) &= \int \dots \int f(data, \boldsymbol{\beta}, z_i, \phi) \partial\beta_1 \dots \partial\beta_{k-1} \partial z_i \partial\alpha \\ f(z_i|data, \beta, \alpha) &= \int \dots \int f(data, \boldsymbol{\beta}, z_i, \phi) \partial\beta_1 \dots \partial\beta_k \partial\alpha \\ f(\alpha|data, \beta, z_i) &= \int \dots \int f(data, \boldsymbol{\beta}, z_i, \phi) \partial\beta_1 \dots \partial\beta_k \partial z_i. \end{aligned}$$

This cannot be done analytically, and is impractical to do numerically, thus the approach is to find a Markov chain that has the posterior distribution as its long-run distribution (Bolstad and Manda, 2001). Thus a sample is drawn from the joint posterior density and the empirical version of the marginals of the posterior density are obtained. This is done using the Metropolis algorithm and Gibbs sampler. These two algorithms are described in Appendix B and C. The estimates $\hat{\beta}_1, \dots, \hat{\beta}_k, \hat{\alpha}$ are then taken as the medians of the empirical densities (Janssen, 2005).

Details of how the relevant Gibbs conditionals are directly computed from the joint distribution can be found in Zuma and Lurie (2005) and Bolstad and Manda (2001).

Chapter 7

Application of Frailty Modeling

Even though both the EM algorithm and penalized partial likelihood (PPL) approach are included in the theory in chapter 6, only the penalized partial likelihood and the Bayesian approach are used to model frailty. It has been shown that the EM algorithm and PPL method produce the same estimates, and since there was no readily available software to fit frailty using the EM algorithm, only the PPL approach was used. It is further known that the EM algorithm can take as much as ten times longer to compute than the PPL approach, making it far less efficient. The software used for fitting the frailty models using the PPL approach was Stata 9, and WINBUGS14 was used for the Bayesian approach. These estimates were then obtained and compared. Only the first two data sets are used, the old order Amish family data and the lung cancer data. The reason for this is that they have cluster variables to demonstrate the shared frailty model. Univariate frailty modeling on the warfarin data set is not performed in the analysis.

7.1 Penalized Partial Likelihood Approach

7.1.1 Old Order Amish Data

The model fitted is

$$h_{ij}(t) = h_0(t)\exp(\beta_1 x_{sex} + \beta_2 x_{byr} + w_i)$$

where $i = 1, \ldots, 458$, $j = 1, \ldots, n_i$, $n = 2860$, and w_i is the random effect for sibship i.

The commands given in Stata to fit the cox proportional hazards frailty model are

```
stset age, failure(dlt)
set matsize 500
stcox sex byr, shared(sib) nohr.
```

The option 'nohr' requests that parameter estimates instead of hazard ratios are outputted. The hazard ratios are easily obtained by exponentiating the parameter estimates. Here sibship is the clustering variable. The data set is very large resulting in 459 cluster groups plus covariates. The default for matsize in Stata is 200, but can be changed upwards or downwards, the maximum value being 800. It took 3 iterations to estimate the frailty variance, and 5 iterations to estimate the final model. The final log-likelihood was -13758.044. Stata uses the Breslow (1974) method for handling ties as a default. A gamma shared frailty for the clustering variable sibship was assumed. The parameter estimates for the model are given in Table 7.1, where θ is the estimate of the frailty distribution variance.

Table 7.1: Parameter estimates for gamma shared frailty model

Variable	Parameter Estimates	Std. Error	z	Pr>z	Hazard Ratio
Sex	0.1053385	0.0487993	2.16	0.031	1.111
byr	-0.0103334	0.0016479	-6.27	< 0.0001	0.9897
θ	0.2859868	0.040405			

The likelihood-ratio test of $\theta = 0$ has a χ^2 statistic = 142.72, which yields a probability of < 0.0001. Thus the random effect is highly significant, and should be included in the analysis. The other variables are significant even when a frailty term is included. The frailty variance is given as 0.2859868, indicating that there is some variation in the frailty between the clusters. It is also possible to obtain the log-frailty estimates for each cluster, however as there are 458 clusters, not all the estimates are shown here. In order to demonstrate how

one would interpret the results, consider Table 7.2. For sibship 34, the log-frailty is given as

Table 7.2: Estimate of log-frailty for cluster 34 and 400

id	age	dlt	sib	sex	byr	log-frailty
⋮	⋮	⋮	⋮	⋮	⋮	⋮
216	1	1	34	1	1897	0.46984433
217	35	1	34	2	1888	0.46984433
218	83	1	34	2	1884	0.46984433
219	40	0	34	2	1891	0.46984433
220	56	1	34	2	1905	0.46984433
221	52	1	34	2	1886	0.46984433
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2507	33	0	400	1	1861	-0.30439511
2508	1	1	400	2	1871	-0.30439511
2509	91	1	400	1	1865	-0.30439511
2510	82	1	400	1	1862	-0.30439511
2511	57	1	400	2	1868	-0.30439511
2512	39	0	400	2	1867	-0.30439511
⋮	⋮	⋮	⋮	⋮	⋮	⋮

0.46984433. The frailty is then $e^{0.46984433} = 1.5997$. Since this value is greater than one, the individuals in this cluster are more frail than the standard individual. In sibship 400, the log-frailty is -0.30439511, and hence the frailty is 0.737569. Thus the individuals in cluster 400 are less frail than the standard individual.

7.1.2 Lung cancer data

The covariates used in the analysis are treatment, performance status, liver metastases, bone metastases and weight loss. The model is given by

$$h_{ij}(t) \quad = \quad h_0(t)\exp(\beta_1 x_{trt} + \beta_2 x_{perfstat} + \beta_3 x_{liver} + \beta_4 x_{bone} + \beta_5 x_{weightloss} + w_i)$$

where $i = 1, \ldots, 26$, $j = 1, \ldots, n_i$, $n = 570$, and w_i is the random effect for institution i.

The commands to fit the model in Stata are

stset survtime, failure(event)

set matsize 500

stcox trt perfstat liver bone weightloss, shared(sib) nohr.

The Breslow (1974) method for handling ties was used, and a gamma shared frailty was assumed. The clustering variable is the institution in which the patients were treated. It took 0 iterations to obtain the frailty variance, and 3 iterations to fit the final model. The final log-likelihood was -2991.7337. The parameter estimates are presented in Table 7.3. The estimate for the frailty variance is so small that it is very close to 0. This implies that

Table 7.3: Parameter estimates for gamma shared frailty model

Variable	Parameter Estimates	Std. Error	z	Pr>z	Hazard Ratio
trt	-0.2567204	0.0859798	-2.99	0.003	0.774
perfstat	-0.6064778	0.1044048	-5.81	0.000	0.54527
liver	0.4179981	0.0903075	4.63	0.000	1.5189
bone	0.2324197	0.0933749	2.49	0.013	1.2616
weightloss	0.2027015	0.0876182	2.31	0.021	1.2247
θ	6.87e-19	7.54e-15			

there is not much variation between the institutions. The estimated log-frailties for the clusters are all 0, implying that all people between institutions have the same frailty. The likelihood ratio test of $\theta = 0$ gives a χ^2 statistic of 5.4e-10; with associated probability of 0.5. Thus the random effect is not significant, and in this case a model without random effects may be a better model to use. Thus it seems that there is not enough evidence to suggest a significant heterogeneity between institutions in survival time. All the other variables included in the analysis are significant.

7.2 Bayesian Approach

7.2.1 Old Order Amish Data

The model fitted is in a different format to the PPL approach, given by the following equation

$$h_{ij}(t) \quad = \quad h_0(t) z_i e^{\mathbf{x}_{ij}^T \boldsymbol{\beta}}.$$

The baseline hazard was modelled as Gamma(μ,c) and the frailty effect was modelled as Gamma(α, α). The hyperparameter α was modelled as Gamma(0.001,0.001) and the fixed effects were modelled as N(0,0.000001). Three parallel chains were run and 20 000 iterations were computed. All the fixed effects parameters, a subset of the random effects, the variance of the random effects and baseline hazard were monitored for convergence. The first 1999 iterations were discarded for the burn-in. The Gelman-Rubin statistic, which was in a graphical format, was used to determine convergence, as well as the history plots that are produced in WINBUGS. The use of multiple chains allows one to monitor convergence of the parameters. The Gelman-Rubin statistic is based on the ratio of the variance between all chains to the variance within a single chain. If the chains have converged, both estimates are unbiased, but if the ratio is larger than one, convergence has not been reached. The diagram that WINBUGS plots has three components; the Gelman-Rubin statistic, the variance within a single chain and the variation between all chains. There is convergence if the Gelman-Rubin statistic is about 1, and both the variances stabilise around the same value. Convergence appeared to be satisfied in this case, although it is difficult and subjective to determine whether convergence has definitely been reached. Another measure of convergence is whether the ratio of the MC error and the standard deviation of the parameter estimates is less than 5%. The estimates of the parameters were taken to be the median instead of the mean, as it is a more stable measure. Table 7.4 shows the results obtained. The frailty variance estimate is 0.5005 with a 95% confidence interval

Table 7.4: Bayesian parameter estimates for family data

Parameter	Mean	SD	MC Error	Median	MCE/SD×100
sex	-0.1239	0.04555	0.0008363	-0.124	1.836%
byr	-0.002364	0.0005855	0.00002905	-0.002474	4.962%
θ	0.501	0.03519	0.0005216	0.5005	1.48%

of (0.4338; 0.5708), indicating that there is variation between the clusters. The estimate of the frailty for cluster 34 is 1.223, which is slightly smaller than the estimate from the PPL approach; and the estimate of frailty in cluster 400 is 0.7368, which is very similar to the estimate obtained before. The reference level for *sex* is now men instead of women. The hazard ratio is 0.883 and the interpretation is that the hazard for women is 80% that for the men. In order to compare results, the hazard ratio can be inverted so that women are the reference category, and the hazard ratio is then 1.132. It is also important to note that the ratio of the MC Error and the standard deviation is less than 5% for each variable; which indicates convergence. The other monitors to measure convergence are included in the disk in the back of this thesis.

7.2.2 Lung cancer data

The baseline hazard was modelled as Gamma(μ,c), the frailty was modelled as Gamma(α, α), the hyperparameter α was modelled as Gamma(0.001,0.001) and the fixed effects were modelled as N(0,0.000001). Three parallel chains were run and 5000 iterations were computed. All the fixed effects parameters, a subset of the random effects, the variance of the random effects and baseline hazard was monitored for convergence. The burn-in was taken to be the first 1000 iterations. The model converged relatively quickly compared to the family data model. Table 7.5 contains all the estimates of the parameters. The frailty variance is quite small, indicating that there is not much variation between the institutions. All the monitors for the parameter estimates indicated that convergence had been satisfied.

Table 7.5: Bayesian parameter estimates for lung cancer data

Parameter	Mean	SD	MC Error	Median	MCE/SD×100
trt	-0.248	0.08653	0.001387	-0.2477	1.603%
perfstat	-0.6023	0.1056	0.002777	-0.6008	2.630%
liver	0.4182	0.09103	0.001208	0.4189	1.327%
bone	0.2339	0.09314	0.00114	0.2342	1.224%
weightloss	0.2067	0.08934	0.001646	0.2053	1.842%
θ	0.07179	0.04094	0.001785	0.06162	4.36%

The ratio of the MC error and the standard deviation was less than 5% for all parameter estimates. Using the Wald χ^2 statistic to test for significance of parameters it was found that the variables are all significant.

7.3 Comparison of the Different Methods

7.3.1 Old Order Amish Data

The models compared are the Cox proportional hazards model without random effects, and the frailty models. The two different methods for fitting the frailty models are compared. From Table 7.6 it can be seen that the parameter estimates of gender are very similar for the different methods, as well as the standard errors. There is some discrepancy in the estimate of birth year, although this is not immediately clear. In the Cox model each later birth year has a 0.7% reduction in the hazard of death, and this is 1% for the PPL approach and 0.2% for the Bayesian approach. The effect between two birth years 10 years apart is a 10% change in the PPL approach compared to a 2% change in the Bayesian approach, with the Cox estimate somewhere in-between at 7%. The parameter estimate for sex in the Bayesian approach was transformed so that all three methods have the same reference category. When examining the Wald χ^2 test for significance of the parameter estimates in the Bayesian model it was seen that the fixed effects are significant. What is noticeably different in the results is the estimate of frailty variance. It is larger using the Bayesian method than the PPL

method, thus showing that the PPL approach exhibits downward bias compared to the Bayesian approach. This bias reflects the incapability of the PPL approach to take into consideration the uncertainty due to the estimation of parameters describing random effect distributions (Ripatti and Palmgren, 2000). However both estimation methods show that the survival experience of the participants vary considerably across families.

Table 7.6: Comparing the parameter estimates for the different approaches

	Cox PH		PPL		Bayes	
Parameter	Estimate	SD	Median	SD	Median	SD
sex	0.10928	0.04218	0.10533	0.0487993	0.124	0.04555
byr	-0.00679	0.0008776	-0.01033	0.0016479	-0.002364	0.0005855
θ	-	-	0.28599	0.040405	0.5005	0.03519

The effect of including a random effect in the Cox proportional hazards model does not appear to affect the estimates or significance of the fixed effects in this data set. However, there does seem to be variability between the different sibships, and thus it is important to include the frailty term in the model. The problem with the PPL approach is that there is possible downward bias in the frailty variance, however the model fitting in WINBUGS can be very computationally extensive and may take a very long time to run.

7.3.2 Lung cancer data

The estimates of the fixed effects parameters are highly similar in all three models and the standard errors are not very different either. The main difference is in the estimate of the frailty effect variance in the PPL model and the Bayesian model. In the PPL model it is much smaller in dimension than in the Bayesian model, reinforcing the fact that the PPL approach has downward bias in the estimate of the frailty variance. It is still quite small in the Bayesian approach though, suggesting that there is not much variation between the institutions. The standard deviation of the frailty variance is also much larger in the

Bayesian model than in the PPL model. The results are shown in Table 7.7. In this case the Cox proportional hazards model is adequate, and a frailty term is not needed in the analysis.

Table 7.7: Comparing the parameter estimates for the different approaches

	Cox PH		PPL		Bayes	
Parameter	Estimate	SD	Median	SD	Median	SD
trt	-0.25672	0.08598	-0.2567204	0.0859798	-0.2477	0.08653
perfstat	-0.60648	0.10440	-0.6064778	0.1044048	-0.6008	0.1056
liver	0.41800	0.09031	0.4179981	0.0903075	0.4189	0.09103
bone	0.23242	0.09337	0.2324197	0.0933749	0.2343	0.09314
weightloss	0.20270	0.08762	0.2027015	0.0876182	0.2053	0.08934
θ	-	-	6.87e-19	7.54e-15	0.07179	0.04094

Chapter 8

Conclusion

In this thesis, many aspects of survival analysis have been explored. What type of analysis one does on a certain data set should be specific to that problem. Non-parametric statistics are useful in estimating and determining the shape of the survival and hazard function, and the log-rank statistic can be used to test for treatment effect between two groups.

Parametric regression modeling may be a viable choice for certain data sets, however they are limited by the assumptions that are placed on survival time of the individuals, for example an assumption of constant hazards when using the exponential regression model. The different types of parametric models are fairly easily fitted in readily available software such as SAS and Stata, and testing certain assumptions about the models can be done, such as the graphical test for linearity in the exponential model.

A well known method of modeling survival data is the Cox proportional hazards model. This is based on the assumption of proportional hazards, and does not require any assumptions to be made regarding the distribution of survival time. The parameter estimates can be transformed into hazard ratios and easily interpreted. This model is readily fitted in many software packages.

Recently much work has been done on frailty modeling, which can be thought of as an extension of the Cox proportional hazards model. Including a frailty term in the model is thought to account for individual random heterogeneity as well as for any clustering that

has occurred in the data set. The frailty term is simply modelled as a random effect, and an estimate of the frailty variance can be obtained.

There are different types of frailty modeling. The frailty term may be at an individual level, where every individual is assumed to have a different frailty due to unmeasured covariates. This is known as univariate frailty modeling. In multivariate frailty modeling each cluster is assigned a frailty term, in other words all individuals within a cluster are assumed to have the same frailty. The frailty term is thought to account for the correlated nature of the data. For example different communities may be assumed to have the same frailty because of the geographical location of the community. Thus it is thought that people may share similar experiences in one community, for example lack of running water, and thus have the same frailty.

Some of the different methods of estimation in frailty models were discussed in Chapter 6. However in the analysis only the penalized partial likelihood approach and the Bayesian approach were done. The reason for this is that there does not seem to be any software that fits the models using the EM algorithm, whereas Stata 9 readily fitted the frailty models using the PPL approach. It has been shown that the EM algorithm and PPL approach produce the same estimates, thus it did not seem necessary to use both methods. Only shared frailty modeling may be done in Stata 9, thus the warfarin data set was not used in the frailty modeling as there was no clustering variable available.

The Bayesian model was computed in WINBUGS14, which requires setting up the model manually. The disadvantage of the Bayesian approach is that it may take a very long time to run many iterations. For the family data set the 20 000 iterations took approximately 27 hours to run, whereas the same model took only minutes in Stata. However it has been mentioned in the literature that there is downward bias in the estimation of the frailty variance using the PPL approach, and there may also be downward bias in the estimates of the standard errors of the fixed effects. A limiting factor of the PPL approach in Stata

9 is that only the gamma frailty model may be assumed, whereas there is more freedom in choosing a distribution for the frailty in WINBUGS. Univariate frailty modeling may also be done in WINBUGS.

Another type of frailty modeling, which was not addressed in this thesis is multi-level frailty modeling. A multi-level frailty model is when more than one frailty term is used in the model. To illustrate this idea, consider the lung cancer data set. If one were to model the institution as a random effect, as well as to assign a frailty for each individual to account for unidentified random heterogeneity, this would be an example of a multi-level frailty model. This could be done in WINBUGS, but there is no readily available software for using the PPL approach to fit these types of frailty models.

It is also possible to include a frailty term in fully parametric regression models. This can be done in Stata 9 and SAS, at an individual level as well as to account for clustering. For example, if a Weibull model is a better fit than the Cox proportional hazards model, it would then be possible to include a frailty term in the Weibull model. Another method of deriving results in survival analysis is the counting process approach. This was not done in this thesis. However, relevant references include Fleming and Harrington (1991), and a derivation of the PPL approach using counting processes is described in Therneau and Grambsch (2000).

To conclude, including frailty in a survival model is an important consideration, and is especially useful in situations where clustering needs to be accounted for. When comparing the Cox proportional hazards model to the frailty models, it was found that the estimates of the parameters for the fixed effects were highly similar. What may happen in some cases is that including a frailty term may increase the estimate of the standard error for the fixed effects and may make a result that was significant in the Cox model insignificant in the frailty model.

Appendices

Appendix A

The Golden Section Method

Typical bisection methods involve finding roots of functions in one dimension, where the root is bracketed in some interval (a, b). The function is evaluated at some point x, where $(a < x < b)$, and a new, smaller bracketing interval is obtained (a, x) or (x, b). This process continues until the bracketing interval is acceptably small.

The golden section search is an analog of bisection methods, where the problem is now to find a minimum in a given interval (Press *et al.*, 1992). In order to bracket a minimum, 3 points are needed, $a < b < c$ such that $f(b) < f(a)$ and $f(b) < f(c)$. We then choose a new point x, which lies either between a and b, or b and c.

Suppose that b is a fraction w of the way between a and c, namely,

$$\begin{aligned} \frac{b-a}{c-a} &= w \\ \frac{c-b}{c-a} &= 1-w. \end{aligned}$$

Also suppose that the next trial point x is an additional fraction z beyond b, where

$$\frac{x-b}{c-a} = z.$$

Thus the next bracketing interval will be either of length $w + z$ relative to the current interval, or of length $1 - w$. In order to minimize the worst case scenario, z is chosen such

that

$$\begin{aligned} w + z &= 1 - w \\ \Rightarrow z &= 1 - 2w. \end{aligned} \tag{A.1}$$

Thus x will be symmetric to b in the original interval, namely with $|b - a|$ equal to $|x - c|$, which implies that x lies in the larger of the two segments (Press *et al.*, 1992). Also x should be the same fraction of the way from b to c as b was from a to c, implying that

$$\frac{z}{1 - w} = w. \tag{A.2}$$

Solving Eq.(A.1) and Eq.(A.2) simultaneously results in the following quadratic equation in w

$$w^2 - 3w + 1 = 0$$

the solution of which (discarding the solution where $w > 1$) is

$$\begin{aligned} w &= \frac{3 - \sqrt{5}}{2} \\ &\approx 0.38197. \end{aligned}$$

Thus the optimal bracketing interval (a, b, c) has it's middle point b a fractional distance 0.38197 from one end (say from a), and 0.61803 from the other end (say c). These fractions are those of the golden mean or golden section (Press *et al.*, 1992), and thus the optimal method of function minimization is called the golden section search, which is as follows (stated in Press *et al.*, 1992):

> Given, at each stage, a bracketing triplet of points, the next point to be tried is that which is a fraction 0.38197 into the larger of the two intervals, measuring from the central point of the triplet.

The golden section search guarantees that each new iteration will bracket the minimum to an interval just 0.61803 times the size of the preceding interval, and the convergence is linear. If the original interval is not in golden ratios, the procedure of choosing successive points at the golden mean point of the larger segment will quickly converge to proper, self-replicating ratios (Press *et al.*, 1992).

Appendix B

Metropolis-Hastings Algorithm

Suppose there is a candidate-generating density, denoted $q(x, y)$, where $\int q(x, y)\mathrm{d}y = 1$. The density depends on the current state of the process, and can be interpreted to mean that when a process is at point x, the density generates a value y from $q(x, y)$.

The reversibility condition states that

$$\pi(x)q(x, y) \quad = \quad \pi(y)q(y, x)$$

where $\pi(\cdot)$ is the invariant density, which is the target distribution from which samples are desired. To generate samples from $\pi(\cdot)$, the process is started at an arbitrary starting point and iterated a large number of times; after which the distribution of the observations generated from the simulation is approximately the target distribution (Chib and Greenberg, 1995).

If $q(x, y)$ satisfies the reversibility condition then y is accepted as the next point in the set, however this does not usually happen. Instead, the following alternative given by

$$\pi(x)q(x, y) \quad > \quad \pi(y)q(y, x) \tag{B.1}$$

may apply, but this implies that the process moves from x to y too often, and from y to x too rarely. In order to correct this, a transition probability, $\alpha(x, y) < 1$ is introduced (Chib and Greenberg, 1995). If the move to y is not made, then x is returned as the next value

from the target distribution. Thus transitions from $x \rightarrow y$ are made according to

$$p_{MH}(x,y) \equiv q(x,y)\alpha(x,y), \qquad x \neq y.$$

Since movement from $y \rightarrow x$ is not made often enough, $\alpha(y,x)$ is defined to be as large as possible. Since $\alpha(y,x)$ represents a probability, the largest value it can take on is 1, thus $\alpha(y,x)$ is set to be 1. The value p_{MH} must satisfy the reversibility condition, so that

$$\begin{aligned} \pi(x)q(x,y)\alpha(x,y) &= \pi(y)q(y,x)\alpha(y,x) \\ &= \pi(y)q(y,x). \end{aligned}$$

Thus from the above equation,

$$\alpha(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}.$$

If Eq.(B.1) were reversed, and the process moved from y to x too often, then set $\alpha(x,y) = 1$ and derive $\alpha(y,x)$ as above. Thus

$$\begin{aligned} \alpha(x,y) &= \min\left[\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right] \qquad \text{if } \pi(x)q(x,y) > 0 \\ &= 1 \qquad\qquad\qquad\qquad\qquad\quad \text{otherwise.} \end{aligned}$$

Consider the probability that the process stays at x. Then this probability is given by

$$r(x) = 1 - \int_{R^d} q(x,y)\alpha(x,y)\mathrm{d}y.$$

The transition kernel of the Metropolis-Hastings chain is

$$P_{MH}(x,\mathrm{d}y) = q(x,y)\alpha(x,y)\mathrm{d}y + \left[1 - \int_{R^d} q(x,y)\alpha(x,y)\mathrm{d}y\right]\delta_x \mathrm{d}y.$$

The algorithm can thus be summarised as follows:

- Repeat for $j = 1, 2, \ldots, N$
- Generate y from $q(x^{(j)}, \cdot)$ and u from $U(0,1)$, where $U(0,1)$ is the uniform distribution.

- If $u \leq \alpha(x^{(j)}, y)$ set $x^{(j+1)} = y$.
- Else set $x^{(j+1)} = x^{(j)}$
- Return $\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$

The draws are seen as a sample from $\pi(x)$, the target density, only after a sufficient 'burn-in' period has passed, and those values discarded. Doing this ensures that the effect of the fixed starting value has become so small that it can be ignored (Chib and Greenberg, 1995). In fact, convergence to the invariant distribution occurs under mild regularity conditions; namely irreducibility and aperiodicity. Main concerns are what the burn-in should be, and how long the sampling should be run. Another problem, pointed out in D'Agostini (2003), is that each point in the chain has some correlation with the points which immediately precede it.

If $q(x, y)$ is symmetric, then $q(x, y) = q(y, x)$, and the probability of a move simplifies to

$$\alpha(x, y) \quad = \quad \frac{\pi(y)}{\pi(x)}$$

thus if $\pi(y) > \pi(x)$ the chain moves to y otherwise it moves with probability $\alpha(x, y)$ (Chib and Greenberg, 1995). The acceptance probability is now

$$\alpha(x, y) \quad = \quad \min\left(\frac{\pi(y)}{\pi(x)}, 1\right).$$

This is known as the Metropolis algorithm. In the Bayesian context, the candidate-generating density is the prior distribution, and the target density is the posterior distribution.

Appendix C

The Gibbs Sampler

In some cases the marginal density is difficult or complex to calculate directly. In cases such as these the Gibbs sampler can be employed to generate the random variables from the marginal distribution (Casella and George, 1992). This technique is based on elementary properties of Markov chains.

Suppose that the joint density is given by $f(x, y_1, \ldots, y_p)$, and the marginal distribution, $f(x)$ is required. In order to obtain the marginal distribution, the other variables need to be integrated out

$$f(x) \quad = \quad \int \ldots \int f(x, y_1, \ldots, y_p) \mathrm{d}y_1 \ldots \mathrm{d}y_p.$$

These integrals may be too difficult to perform and the Gibbs sampler offers an alternative approach to obtain $f(x)$. Thus a sample $X_1, X_2, \ldots, X_m \sim f(x)$ is generated, without explicitly requiring $f(x)$ (Casella and George, 1992). For a large enough sample, any characteristics of $f(x)$, such as the mean and variance, can be calculated to the desired degree of accuracy (Casella and George, 1992). The algorithm is explained in the case of two variables below.

Suppose there are two variables, X, Y, and the interest is in generating a sample from the target distribution $f(x)$. Using the Gibbs sampler, the sample is generated from $f(x|y)$ and $f(y|x)$ and the resulting sequence is known as a Gibbs sequence. Such a sequence can

be represented by

$$Y'_0, X'_0, Y'_1, X'_1, Y'_2, X'_2, \ldots, Y'_k, X'_k$$

where Y'_0 is a specified initial value, and the rest of the sequence is obtained iteratively by alternately generating values from

$$X'_j \sim f(x|Y'_j = y'_j)$$
$$Y'_{j+1} \sim f(y|X'_j = x'_j).$$

The distribution of X'_k converges to the target distribution, $f(x)$, as $k \to \infty$, under reasonably general conditions (Casella and George, 1992). Thus for k large enough, the final observation, $X'_k = x'_k$ can be viewed as a sample point from $f(x)$. This fact can be exploited to obtain an approximate sample from $f(x)$. Gelfand and Smith (1990) suggest generating m independent Gibbs sequences of length k and using the final value X'_k from each sequence to obtain an approximate *iid* sample from $f(x)$. Another method would be to allow the chain an appropriate burn-in period, discarding the first half of the sequence, and focussing attention on the second half. In this approach, however, one must be aware of the possibility of correlated values on the sequence. In order to overcome this Gelman *et al.* (1995) mention using only every i^{th} simulation draw, where i is between 10 and 50, in order to have approximately independent draws from the target distribution.

Bibliography

[1] Allison, P.D. 1995. *Survival Analysis using SAS®: A Practical Guide.* SAS Institute Inc., Cary, NC, USA.

[2] Bilmes, J.A. 1998. A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Guassian Mixture and Hidden Markov Models. ICSI TR-97-021, U.C. Berkeley, USA.

[3] Bolstad, W.M. and Manda, S.O. 2001. Investigating Child Mortality in Malawi using Family and Community Effects: A Bayesian Analysis. *Journal of the American Statistical Association,* **96**, 453:12-19.

[4] Brent, R.P. 1973. *Algorithms for Minimization without Derivatives.* Prentice-Hall, Englewood Cliffs, NJ.

[5] Breslow, N.E. 1974. Covariance Analysis of Censored Survival Data. *Biometrics,* **30**, 89-100.

[6] Buchanan-Lee, B., Levetan, B.N., Lombard, C.J. and Commerford, P.J. 2002. Fixed-Dose Versus Adjusted-Dose Warfarin in Patients with Prosthetic Heart Valves in a Peri-Urban Impoverished Population. *The Journal of Heart Valve Disease,* **11**, 583-593.

[7] Casella, G. and George, E.I. 1992. Explaining the Gibbs Sampler. *The American Statistician,* **46**, 3:167-174.

[8] Chib, S. and Greenberg, E. 1995. Understanding the Metropolis-Hastings Algorithm. *The American Statistician,* **49**, 4:327-335.

[9] Collett, D. 1994. *Modelling Survival Data in Medical Research.* Chapman & Hall, London.

[10] Cortinas Abrahantes, J. and Burzykowski, T. 2004. A Version of the EM Algorithm for Proportional Hazard Model with Random Effects. *Biometrical Journal,* **47**, 6:847-862.

[11] Cox, D.R. 1972. Regression Models and Life Tables. *Journal of the Royal Statistical Society, Series B,* **74**, 187-220.

[12] D'Agostini, G. 2003. Bayesian Inference in Processing Experimental Data Principles and Basic Applications. *Reports on Progress in Physics,* **66**, 1383-1419.

[13] Dempster, A.P., Laird, N.M. and Rubin, D.B. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological),* **39**, 1:1-38.

[14] Drakos, Nikos. 1994. *Introduction to Monte Carlo Methods.* Computer Based Learning Unit, University of Leeds. Source: electronic book.

[15] Efron, B. 1977. The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association,* **72**, 557-565.

[16] Fleming, T.R. and Harrington, D.H. 1991. *Counting Processes and Survival Analysis.* Wiley, New York.

[17] Gelfand, A.E. and Smith, A.F.M. 1990. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association,* **85**, 398-409.

[18] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. 1995. *Bayesian Data Analysis.* Chapman & Hall, London.

[19] Gilks, W.R., Richardson, S. and Speigelhalter, D.J. 1996. *Markov Chain Monte Carlo in Practise.* Chapman and Hall, London.

[20] Gray, R. 1995. Tests for Variation Over Groups in Survival Data. *Journal of the American Statistical Association,* **90**, 198-203.

[21] Hosmer, D.W. and Lemeshow, S. 1999. *Applied Survival Analysis Regression Modeling of Time to Event Data.* John Wiley & Sons, Inc., United States of America.

[22] Hougaard, P. 2000. *Analysis of Multivariate Survival Data. Statistics for biology and health.* Springer-Verlag, New York.

[23] Janssen, P. 2005. Parametric Frailty Models. Workshop on Frailty Models. 52^{nd} Annual Conference of the South African Statistical Association. 31 October - 4 November 2005, Grahamstown, South Africa.

[24] Kalbfleisch, J.D. and Prentice, R.L. 1980. *The Statistical Analysis of Failure Time Data.* John Wiley & Sons, Inc., New York.

[25] Kaplan, E.L. and Meier, P. 1958. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association,* **53**, 457-481.

[26] Keiding, N., Andersen, P.K. and Klein, J.P. 1997. The Role of Frailty Models and Accelerated Failure Time Models in Describing Heterogeneity due to Omitted Covariates. *Statistics in Medicine,* **16**, 215-224.

[27] King, T.M., Beaty, T.H. and Liang, K.Y. 1996. Comparison of Methods for Survival Analysis of Dependent Data. *Genetic Epidemiology,* **13**, 139-158.

[28] Klein, J.P. 1992. Semiparametric Estimation of Random Effects using the Cox Model Based on the EM Algorithm. *Biometrics,* **48**, 795-806.

[29] Lawless, Jerald F. 1982. *Statistical Models and Methods for Lifetime Data.* John Wiley & Sons, Inc., New York.

[30] Lee, E.T. 1992. *Statistical Methods for Survival Data Analysis.* Wiley, New York.

[31] Lee, E., Wei, L. and Amato, D. 1992. Cox-Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations. *Netherlands: Kluwer Academic Publishers,* 237-247.

[32] Marubini, E. and Valsecchi, M.G. 1995. *Analysing Survival Data from Clinical Trials and Observational Studies.* John Wiley & Sons Ltd, England.

[33] McCullagh, P. and Nelder, J.A. 1989. *Generalized Linear Models (Second Edition).* Chapman and Hall, London.

[34] Nguti, R. 2003. *Random Effects Survival Models Applied to Animal Breeding Data.* Unpublished Ph.D dissertation, Limburgs Universitair Centrum, Belgium.

[35] Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. 1992. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, UK.

[36] Ripatti, S. and Palmgren, J. 2000. Estimation of Multivariate Frailty Models using Penalised Partial Likelihood. *Biometrics,* **56**, 1016-1022.

[37] Ripatti, S., Larsen K. and Palmgren, J. 2002. Maximum Likelihood Inference for Multivariate Frailty Models using an Automated Monte Carlo EM Algorithm. *Lifetime Data Analysis,* **8**, 349-360.

[38] Stewart, J. 1997. *Calculus Concepts and Contexts.* Brooks/Cole Publishing Company, USA.

[39] Struthers, C.A. and Kalbfleisch, J.D. 1986. Misspecified Proportional Hazards Models. *Biometrika,* **73**, 363-369.

[40] Therneau, T.M. and Grambsch, P.M. 2000. *Modeling Survival Data: Extending the Cox Model.* Springer-Verlag New York, Inc., New York.

[41] Vaida, F. and Xu, R. 2000. Proportional Hazards Model with Random Effects. *Statistics in Medicine,* **19**, 3309-3324.

[42] Vaupel, J.W., Manton, K.G. and Stallard, E. 1979. The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. *Demography,* **16**, 439-454.

[43] Zuma, K. 2005. Interval Censored Data: Application to HIV/AIDS. Workshop on Frailty Models. 52^{nd} Annual Conference of the South African Statistical Association. 31 October - 4 November 2005, Grahamstown, South Africa.

[44] Zuma, K. and Lurie, M.N. 2005. Application and Comparison of Methods for Analysing Correlated Interval-censored Data from Sexual Partnerships. *Journal of Data Science,* **3**, 241-256.