Institute and Faculty of Actuaries

PhD studentship output

Funded by the Institute and Faculty of Actuaries

LITERATURE REVIEW ON SURVIVAL ANALYSIS

M.A.H WAHEDALLY

Contents

1. An Overview of Modelling of Medical Data

1.1. Introduction.

Several approach and analytical methods have been proposed in literature when it comes to analysis of medical data, especially for cases when the data is hierarchical(sometimes referred as multi-level). The most common ones utilised for research and analytical purposes are classified into mathematical, statistical and survival modelling. In this introductory chapter we will discuss the use and the arguments for and against utilising these three types of modelling in relation to analysing medical information and data.

1.2. Mathematical Modelling.

Bernoulli (1760) proposed for the first time a mathematical model for epidemiological data (Small Pox Disease data) analysis. In his deterministic approach, the model was based upon a series of differential equations, which is the basis of most mathematical models today (See Bailey 1975, Murray 2003). There exists a large literature regarding applications of mathematical models to medical data. reviews of mathematical epidemiology can be obtained in Bailey (1975) and Anderson and May (1991). Murray (2003) proposed an appropriate definition of a good mathematical model.

To quote Murray (2003): "From a mathematical point of view the art of good modelling relies on : (i) a sound understanding and appreciation of the physical problem; (ii) a realistic mathematical representation of the important physical phenomena; (iii) finding useful solutions, preferably quantitative; and what is crucially important; (iv) a reliable and accurate interpretation of the mathematical results in terms of insights and predictions. The mathematics is dictated by the physical nature of the condition of interest and not vice versa."

1.2.1. *Compartmental Models*.

The most common way to mathematically model data is to allow classification of observation at any time point into different groups based upon several stages in the life cycle of the experiment. Such a process is called compartmental modelling. This allows the developments of equations that aid in determining transition rates between the different groups. Kermack and Menkendrik (1927) suggested a deterministic model that assumes that observations for a medical experiment can be categorised under several well defined states, thereby allowing the development of transition rates from one state to another via differential equations.

However such an approach is limiting in reality due to the huge number of assumptions that have to be made. Adapting and extending the model to accommodate for these assumptions is unrealistic. Similarly such a system is non-dimensional, removing any dependence from physical units of measurements in the model so as to achieve meaningful inferences. Murray (2003) provides a discussion of the method involving differential equations.Also this technique does not take account of random variations in the data because it ignores the effects due to unknown factors. One way to overcome this problem is to input some prior knowledge into the model to

improve the estimates of parameters but this makes the model harder to fit.

An alternative is to apply the compartmental modelling process under a stochastic approach. This method assumes that the different states in which the data can be classified as random variables and defines the transition rates between states through probability distributions (see Bailey 1975). This approach has the advantage of incorporating random variation in the model. It also allows point estimation of parameters and provides information about their variability as well. Moreover the stochastic approach produces more informative predictions due to the inclusion of random variation in the model. For a more detailed explanation on the techniques on stochastic compartmental model, see Bailey (1975).

1.2.2. ***Cellular Automata.***
Cellular automata (CA) offers another route towards modelling certain type of epidemiological data. CA transforms time and space discretely and models the evolution of complex physical systems via local neighbourhood interactions, based on a lattice structure. Generally we assign each cell of the automaton to a specific state, dependent on the application of the model. Transition between states are controlled by a set of rules connected to the state of the local neighbourhood surrounding each cell. For mortality modelling, a simple example would be to represent space as a two-dimensional regular lattice, where each cell represents an individual at risk and takes value zero if the latter survived the medical procedures of interest procedure of interest or one if the individual dies. We then update each cell at each time point (t) according to a time function, f(.) that connects to the number of neighbouring infected cells at the previous time point.

More realistic CA model can be developed by taking into account some real facts about the medical conditions we are interested in while updating the time function and the neighbourhood criteria. These include temporal characteristics of the medical conditions (for example latency periods or length of procedure), covariates or spatial structure and lags in the definition of the local neighbourhood. Similarly applying a prior knowledge of the spatial distribution of the individuals or the the subjects being investigated contributes towards developing more realistic CA models. In addition, extending stochastic approach to CA allows modelling transition rates between states for the cells over time, governed by a probabilistic process. Sirakoulis et al. (2000) proposed an algorithm for CA models that has a fast computational time due to the regular characteristics in the structure of the CA models.

Moreover CA also offers some good computational advantages over models that employs the use of differential equation (for example compartmental model). Variable total population sizes and exposure, complex initial and boundary conditions all cause problems computationally under a a differential equation setting while CA provides relatively straightforward methods due to its regular structure (See Sirakoulis et al. 2000, Fuentes and Kuperman 1999).

1.3. **Statistical Modelling Approach.**

In mathematical modelling, the notion of mathematical laws govern the physical processes that are subject to random influence. Under statistical modelling, we account and provide a method that directly attempts to reduce this random variation. There are several approaches that exist to model medical data statistically. These include simple time series, purely spatial models, space-time approaches and survival modelling.

1.3.1. ***Autoregressive Modelling Average.***
Often measurements in medical field are recorded over discrete time points. These observations are usually dependent, with the degree of dependence across time periods known as the lag. A common method to analyse and model such data is the Autoregressive Modelling Average (ARMA) framework (Box and Jenkins, 1976). This method models the values at each point in the time series via a combination of two independent processes; firstly the autoregressive process (AR) treats the observations as a weighted sum of their values at previous time points and secondly, the moving average (MA) provides a method that accounts for and corrects for the errors in the previous prediction through a weighted linear sum of previous errors. The number of components varies in each case and is dependent on the temporal lag.

Consider an observed time series Y_t where t=1,...,T, an ARMA model with 2 AR and 2 MA components as shown below.

$$Y_t = \mu + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \xi_t - \alpha_1 \xi_{t-1} - \alpha_2 \xi_{t-2} \tag{1.1}$$

Then μ is the common constant intercept while the β's represent the effects of the auto-regression and the α parameters correspond to the MA effect and the error terms (ξ) are independent and identically distributed by a normal distribution with mean zero and variance σ^2. Normality assumption and the requirement that th data should be stationary (Diggle 1990) are the primary theory supporting the ARMA framework. Note that when data is non-normal, Box-Cox transformation can be applied to correct the data while non-stationary data should be transformed into stationary ones by differencing between successive time points. As a result, a new parameter, called the autoregressive integrated moving average (ARIMA), is added to the model. Violation of non-constant variance may be corrected by variance stabilisation transformations.

The model described in (1.1) can easily be extended to accommodate for link functions if the observations can be represented by a particular distribution. For an ARMA framework, these link functions include probit, complementary log-log and logistic functions.Consider for example a medical data, assumed to be binomially distributed. If we define the number of successful operation procedures on different patients as 'success' for a fixed number of observations and I_t for t=1,...,T as observed ordered series for the number of new successes discovered at each time point; ie, $I_t \sim \beta(n_t, p_t)$ where p_t is the probability of success at time t, then we can model the data using a logistic link function (See McCullagh and Nelder, 1989) as shown below:

$$log(\frac{p_t}{1-p_t}) = \eta_t \ for \ t = 1, ..., T \ \& \ 0 \leq p_t \leq 1. \tag{1.2}$$

where η_t is a linear combination of the regression terms that is thought to directly affect the probability of success, p_t at time t.

However despite showing the flexibilities discussed above, the models developed under the ARMA framework do not deal with issues related to seasonal variation (localised trends), cyclical variation (trends over a longer time period) and irregular fluctuations due to unknown factors. In addition, prediction of future evolution requires extrapolation with the time series method under the ARMA framework. Extrapolating method for future prediction in itself has certain modelling difficulties. For an introduction on the time series models under the ARMA framework models, see Chatfield (2004) while Chatfield (2001), Box and Jenkins (1976) and Anderson (1971) provides a deep comprehensive explanation.

1.3.2. ***Spatio-Temporal Point Process.***
In certain studies, observations are taken at the individual level and the probabilistic phenomena of interest are the time and locations of these observations. Such data can be described by a spatio-temporal point process where the main objective is to locate clustering or regularity in recorded events over space and time, estimating and mapping relative risk of the event incidence or identifying clustering around a particular point. The methodology and techniques involved are developed around the understanding that a completely random point pattern will follow a homogeneous Poisson process over space and time.

There exists a spatial analogy of the time series count model if the data can be arranged over space into a set of (regular or irregular) areal units. There are a series of methods to deal with such data and these allow patterns or trends in the relative risks of event incidence to be modelled (Lawson 2001). These methods are widely applied in spatial epidemiology and they have the ability to model autocorrelation between measurements taken at different spatial lags. Good introductory texts on spatial analysis can be obtained in Bailey and Gatrell (1995) and Cressi (1993) while Diggle (2003) provides a comprehensive account of spatial point pattern analysis. Lawson (2001) provides an example of the application of areal modelling techniques to epidemiological data.

The spatio-temporal point process is also often extended to and applied under a Generalised Linear Model (GLM) framework. For instance Bernadinelli et al. (1995) used generalised mixed model to model the disease rates in different areal units for the impact insulin-dependent diabetes in military conscripts in Sardinia between 1936 and 1971. They employed a mixed model that treats the temporal trend and area specific intercept as random effects. However Knorr-Held and Besag (1997) pointed out that the formulation used by Bernardinelli et al. (1995), accounts the temporal trend as linear. To overcome this assumption, Knorr-Held and Besag (1997) extend the dynamic model methodology of West and Harrison (1997) so as to model non-parametrically non-linear temporal trends and spatial variations (Besag et al. 1991). Hence we observe that the spatio-temporal point process offers a good degree of flexibility in terms of extending it to other frameworks to accommodate for random effects.

1.4. **Survival Analysis.**
Survival analysis offers an alternative method to model medical data by directly analysing the time taken for the effects of the medical condition of interest to happen. Such a method offers many advantages compared to mathematical modelling or statistical modelling in the sense that all parameters of interest (for spatial and temporal effects) can be estimated simultaneously as well as allowing predictions of future survival times for individual observations.

Survival models are usually represented by hazard functions. Hazard function is one that uniquely provides a method to define the survival times according to a particular distribution. It also allows the determination of several quantity of interests that answer several questions for a particular research. For example it allows to determine the probability that an individual will survive for a time period τ after the individual has undergone a medical procedure. Similarly hazard function is very flexible in the sense that we can incorporate covariate information into it easily.

The key issue with using and interpreting the results of survival models is censoring. Censored data are those observations that do not contain enough and complete survival information. For example if an individual leaves a study before the end point of the investigation, then we do not have enough information on the outcome of the procedures of interest on that particular individual. This is an example of censored information. There are several types of censoring namely right censoring, left censoring, Type I and II censoring. Several methods have been proposed in literature for survival models to deal with censored data and these will be discussed in details in the next section. In epidemiological field, specification of censored data is of immense importance for models where the time to exposure to a particular medical conditions or procedures changes over time and space.

Furthermore, there exists several medical data that vary across space. There is reasonable argument to advocate that some of these observations do not represent the population at risk to the medical conditions of interest because of their spatial location as some observations will be unlikely exposed to the conditions of interest. Among all the possible solutions proposed to deal with this, all consist of two main facets: firstly modelling space-time dependence in the mean process (first order effects) and secondly through 'stickiness' between neighbouring individuals (second order effects). For instance, space-time covariates is one way to handle the issue of observations not being representative of their population while the use of spatio-temporal random effects that are correlated (also known as frailties in survival literature) deals with the issue of biasness caused by information contributing to the model fitting.

Survival modelling can also be applied under a Bayesian framework. The latter provides a full posterior distribution to estimate parameter and predictors of survival times. Combined with Markov Chain Monte Carlo (MCMC) methodology, the Bayesian framework provides a powerful method to fit complex frailty models. Similarly, issues such as immunity, multiple survival processes, multi-decrement

functions or multiple causes of death can be investigated via extensions to multivariate data and analysis and data that has been aggregated by areas.

1.5. **Conclusion.**

In this introductory section, we have explored and identify different frameworks that have been utilised to model epidemiological and medical data. Medical information can always be categorised into several well defined hierarchical groups. The use of each of these frameworks carry their own advantages and disadvantages, depending on the type of medical data we have.

From a mathematical perspective, we introduced the compartmental model and cellular automata (CA). The former model uses a collection of differential equations in the modelling process but does not account for random variations unless it is employed under a stochastic framework. However CA offers a better alternative mathematical model than compartmental model because of its regular structure which allows solving problems that consist of variable population sizes or complex initial and boundary conditions and that offers good computational advantages.

A statistical approach offers the possibility of modelling the random variations (due to unknown factors) directly. It is very flexible under a stochastic approach and can be easily extended to the GLM framework to explain temporal aspect (model 1.1). However it tends to average the temporal aspect over time across individuals unlike the survival framework. Survival analysis has the ability to accommodate for many features that cannot be assessed under a mathematical or statistical approach. It enables us to deal with changes in the state of the medical condition we are interested in, censoring, immunity, multiple causes of death and the effects of explanatory variables directly on survival time. The ability to directly predict the risk and time of occurrence of a particular event on a local and global level makes survival modelling very attractive compared to models developed under a mathematical or statistical approach.

2. **Survival Analysis**

2.1. **Introduction.**

Survival analysis is one of the main research methods used in many field such as medicine, biology, epidemiology, demography and engineering. The concept of survival analysis arises from the methods used in medical and demographic studies of event history and mortality. Survival modelling is a method used to model the time from the beginning of a follow-up of an individual until a pre-defined event occurs. Usually the event is associated with failure, for example death of a patient who has undergone a particular treatment. Consequently the time for this event to happen is referred as the survival or failure time.

In statistical modelling literature, survival times tend to follow a skewed distribution such as exponential or Weibull distribution for instance. Most common methods discussed in literature to deal with survival modelling have been carried out under a Generalised Linear Model (GLM) framework (Nelder and Wedderburn 1972) which offers an alternative and extension to the classical linear models to

analyse non-normally distributed data. Censoring is a major issue in survival modelling. Aitken and Clayton (1980), Whitehead (1980) and McCullach and Nelder (1989) discussed the traditional survival models that can be viewed as an extension of GLMs that accommodate for censoring by fitting the models using standard techniques such as iterative weighted least-squares. Many of the currently fitting techniques have been developed independently of the GLM framework (See Kalbfeich and Prentice 2002, Collett 2003, Therneau and Grambsch 2000 for example). Hence there is a range of fitting techniques such as the direct application of maximum likelihood or under a Bayesian approach, the use of an iterative sampling algorithm like Markov Chain Monte Carlo (MCMC).

In this section, we will discuss the different modelling strategy that exists under survival modelling as well as different techniques to fit these models under a Frequential or Bayesian approach, depending on the choice of distributional assumptions. Moreover we will explore extensions of the classical survival models that are applicable for mortality analysis. Finally survival analysis is a widely documented topic not only in medical or epidemiology but also in engineering and social science. Since it has such a large literature, this section will concentrate on concepts and techniques that are relevant to mortality analysis. This section is not an exhaustive review of the subject. For a deep comprehensive discussion, see texts by Kalbfeisch and Prentice (2002), Collett (2003), Therneau and Grambsch (2000), Cox and Oakes (1984) or Lee and Wang (2003).

2.2. **Basic Ideas.**

According to Cox and Oakes (1984), to apply survival techniques to different situations, three requirements are needed: firstly determining a time origin that is well-defined; secondly a scale to measure the change in time and thirdly an exact definition of failure. To start, consider a set of homogeneous data where T is a positive random variable representing failure time, with *distribution function*, F(t). We can then define for continuous time T, the *survival function, S(t)* as the probability that an individual survives beyond time t, i.e.

$$S(t) = P(T > t) \ where \ 0 < t < \infty \tag{2.1}$$

Note that $0<S(t)\leq 1$ since $s(0)=1$ while $\lim_{t\to\infty} S(t)=0$. T has a unique distribution defined by the *survivor function* or the *hazard function* or the *probability density function*. The hazard function is defined as

$$\lambda(t) = \frac{f(t)}{1 - F(t)} \tag{2.2}$$

In epidemiology, $\lambda(t)$ is also known as the force of mortality. It can be interpreted as probability that failure occurs in interval $(t,t+\delta t)$ given that the individual survives past time t, i.e,

$$\lambda(t)\delta t \cong \ P(t < T < t + \delta t | T > t) \tag{2.3}$$

Integrating $\lambda(t)$,

$$\int_0^t \lambda(u)du = \int_0^t \frac{f(u)}{1 - F(u)}du = -log(1 - F(t)) = -logS(t) \tag{2.4}$$

which leads to the survival function being expressed in terms of the force of mortality.

$$S(t) = e^{-\int_0^t \lambda(u)du} \tag{2.5}$$

If T is a discrete random variable, then the probability function f(t)=P(T=t) gives the exact probability of failure at time t. Similarly, the hazard function h(t) is defined as the conditional probability of failure at time t given survival to t, i.e.

$$\lambda(t) = P(T = t|T \geq t) = \frac{P(T = t)}{P(T \geq t)} = \frac{P(T = t)}{\sum_{j|t_j \geq t} P(T = t_j)} \tag{2.6}$$

Similarly we can easily define P(T=t) and P(T≥t) in terms of the hazard function by taking the fact that $1 - h(t)$ is the conditional probability of survival at time t given survival to t. It is also possible to have a mixture of continuous and discrete distributional forms in one framework itself as shown in Kalbfleisch and Prentice 2002, Chapter 1.

2.3. **Censoring.**

Censoring is a common characteristic of survival data, especially in the field of medicine and epidemiology. It distinguishes survival modelling from other statistical models. A censored observation is one that does not contain enough or complete survival information. We will consider four main types of censoring.

The most common type of censoring is *right censoring*. It happens when a subject joins an experiment at the start of the study and leaves the investigation without experiencing failure. This usually because either the study finishes before the subject experiences failure or because the subject is lost to follow up or leaves the study before it actually ends. If the study started at time 0, then the subject is said to *right-censored* at time t and known to have survived for a period of [0,t].

Suppose a subject who joined the experiment at the start, experiences failure at a time that is unknown exactly but before time t. The subject is said to be *left censored* and it is known that failure occurs in the period [0,t]. Moreover given that the exact time of failure is unknown but it is known to be between to two points a and b, where $0<a<b<$t, then the subject is said to be *interval censored*. Both right and left censoring are special cases of interval censoring because we only get to know that the random variable time of of failure occurs in an interval.

Type I censoring occurs when the subject encounters failure after a pre-defined length of time while *Type II censoring* occurs when the study stops when a predetermined number of failures have been observed. In this case the remaining individuals are said to be type II censored. In the case where the failures occur independently and at random, then *Type III censoring* is said to occur. The type of censoring affects the form of the likelihood function used to fit survival models. Kalbfleisch and Prentice (2002) discuss several modifications for the likelihood in order to accommodate for Type I and Type II censoring.

Finally we have randomly censored information. If entry of a subject into a study or loss to follow-up of subjects are considered to be random, then random censoring is said to occur. Let C_i be the censoring time and T_i be the random variable of interest for 1<i<n. Then if random censoring occurs, then we can only observe the subject for for Y_i where Y_i is given by $Y_i = \min(T_i, C_i)$.

2.4. **Parametric Modelling.**

2.4.1. ***Introduction.***

There exists several situations where survival data has a known distribution or simply because it is reasonable to assume that the data has a certain parametric specification. In survival literature, there exists various distributions that have been commonly used to fit such data. A selection of these distributions will be discussed in this section. Fitting parametric models to survival data offers some advantages. They have fully specified hazard functions that are dependent on the parameters that determine the overall distributional form. These parameters can be estimated to fit the model at any point in time and can be used to predict the hazard at future time points.

2.4.2. ***Exponential Model.***

From Equation (2.5), if the hazard rate *h(t)* is equal to a constant positive scale parameter λ, then the survival time is said to follow an exponential distribution. The survivor function is given by $S(t)=exp(-\lambda t)$ and the density function is $\lambda exp(-\lambda t)$.

2.4.3. ***The Weibull Model.***

The Weibull model is a generalisation of the exponential model where the hazard function takes the form of $h(t)=\alpha\lambda t^{\alpha-1}$ where the parameter λ and α are both positive. The density function is given by $f(t)=\alpha\lambda t^{\alpha-1} exp(-\lambda t^{\alpha})$ while the survivor function is written as $exp(-\lambda t^{\alpha})$. The Weibull distribution is very flexible because its distributional shape and also because it has scale parameter, hence allowing the density function to take several forms. Inclusion of covariates via link functions in the scale parameter λ allows the model to have a proportional hazard and an accelerated life structure. Such characteristics are only present under the Weibull distribution.

2.4.4. ***The Gamma Model.***

The Gamma model is another generalisation of the exponential model. Its hazard function has a density of the form of $f(t)=\frac{\lambda^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}$ where α and λ are positive parameters. However under the Gamma model, *S(t)* and $\lambda(t)$ does not have a closed form expression. *S(t)* can only be defined as

$$S(t) = 1 - \int_0^t f(u)\, du = 1 - \left(\frac{Incomplete\ gamma\ function}{Complete\ gamma\ function} \right) \tag{2.7}$$

2.4.5. ***The Rayleigh Model.***

Under the Rayleigh model, the hazard function is expressed as a linear risk in the form of $\lambda(t)=\lambda_0+\lambda_1(t)$. However the hazard function can be generalised to polynomials of the form $\lambda(t)=\sum^{p}_{i=0}\lambda_i t^i$. In the case of linear risk, we will obtain

$$\int_0^t \lambda(u)\ du = \lambda_0 t + \frac{1}{2}\lambda_1 t^2 \ \Rightarrow\ S(t) = exp(-\lambda_0 t - \frac{1}{2}\lambda_1 t^2) \tag{2.8}$$

2.4.6. ***The Log-Normal Model***.
If the random variable of interest, T_i is assumed to be log-normally distributed with mean μ and variance σ^2,i.e. logT_i~N(μ,σ^2), then we can obtain the survivor function *S(t)* as shown below. The log-normal model works best for uncensored data. It converts the data into the standard linear model setup.

$$\begin{aligned} S(t) &= 1 - P(T < t) = 1 - P(logT < logt) \\ &= 1 - P\left(\frac{logT - \mu}{\sigma} < \frac{logt - \mu}{\sigma}\right) \\ &= 1 - \phi\left(\frac{logt - \mu}{\sigma}\right) \end{aligned} \tag{2.9}$$

The models discussed above have closed forms for the survivor and hazard functions and easy to work when T_i are continuous. There exists other possible distributions that can be utilised for the survivor and hazard functions. These include the *Pareto distribution, log-logistic, generalised Gamma* and *generalised F*.

2.4.7. ***Discrete T_i***.
Kalbfleisch and Prentice (2002) described how to discretise any continuous survival distribution. This can be done by considering a *discrete random variable* T such that

$$P(T = t) = P(t \leq U < t + 1) \tag{2.10}$$

where U is a continuous random variable with a fully known distribution. As an example, if U follows a Weibull distribution with shape parameter α and scale parameter λ, then following from (3.10), we can write

$$\begin{aligned} P(T = t) &= P(U < t + 1) - P(U < t) \\ &= F(t + 1) - F(t) \\ &= S(t) - S(t + 1) \\ &= exp(-\lambda t^{\alpha}) - exp(-\lambda(t + 1)^{\alpha}) \end{aligned} \tag{2.11}$$

For the above example, we discretise T over over unit time period length. This can be modified depending on the interval chosen.

2.4.8. ***Parametric Model Fitting***.
For parametric models, model fitting is most commonly done via the *Maximum Likelihood Method*. Suppose we have n observations that are randomly censored. For a model consisting of p parameters, $\underline{\theta}$=(θ_1,...,θ_p)$\prime$, the likelihood function is written as

$$\ell(\underline{\theta}) = \Pi_{i=1}^{n}[f(t_i|x_i, \theta_1, ..., \theta_p]^{\delta_i}[S(t_i|x_i, \theta_1, ..., \theta_p)]^{\delta_{1-i}} \tag{2.12}$$

where the binary variable δ_i (for ι=1,...,n) is equal to 1 if the individual experiences failure or 0 if he is right-censored (or left-censored or interval censored). In this way, only censored observations of individuals surviving the interval [0,t] contributes to the likelihood. Then for some specific parameter θ_j, the score equation is given by

$$\frac{\partial log\ \ell(\underline{\theta})}{\partial \theta_j} = \sum_{i=1}^{n} \frac{\partial log\ \ell_{\underline{\theta}}(y_i, \theta_i)}{\partial \theta_j} = 0 \tag{2.13}$$

Solving (2.13) is not a straight forward process. It is usually done on a computer using iterative methods. The most common method to optimise function (2.13) is

the *Newton-Raphson and Method of Scoring.*

Newton-Raphson and Method of Scoring

Let $\ell_i(\underline{\theta})= \ell_{\underline{\theta}}(t_i,\theta_i)$. Hence

$$\frac{\partial logL(\underline{\theta})}{\partial \underline{\theta}}=(\frac{\partial logL(\underline{\theta})}{\partial \theta_1},...,\frac{\partial logL(\underline{\theta})}{\partial \theta_n})'$$

$$\text{and } \frac{\partial^2 log\ \ell(\underline{\theta})}{\partial \underline{\theta}^2}=\begin{pmatrix} \frac{\partial^2 log\ \ell(\underline{\theta})}{\partial \theta_1 \partial \theta_2} & ... & \frac{\partial^2 log\ \ell(\underline{\theta})}{\partial \theta_1 \partial \theta_p} \\ . & . & . \\ . & . & . \\ \frac{\partial^2 log\ \ell(\underline{\theta})}{\partial \theta_p \partial \theta_1} & ... & \frac{\partial^2 log\ \ell(\underline{\theta})}{\partial \theta_p \partial \theta_p} \end{pmatrix}$$

The likelihood equations for *j=1,...,p* are given by (2.13). Let $\underline{\hat{\theta}^0}=(\theta_1{}^0,...,\theta_p{}^0)$ be the initial guess for the solution to (2.13). Usually this first guess is obtained by a simpler method, for instance method of moments. Expanding (2.13) about $\underline{\hat{\theta}^0}$ using the Taylor series, we get

$$\frac{\partial logL(\underline{\hat{\theta}})}{\partial \underline{\theta}} = \frac{\partial logL(\underline{\hat{\theta}^0})}{\partial \underline{\theta}} + \frac{\partial^2 logL(\underline{\hat{\theta}^0})}{\partial \underline{\theta}^2}\left(\underline{\hat{\theta}} - \underline{\hat{\theta}^0}\right) + ... = 0$$

Ignoring second order and higher terms, let $\underline{\theta^1}$ be the solution to the above equation. Then we can write

$$\underline{\hat{\theta^1}} = \underline{\hat{\theta^0}} + \left(-\frac{\partial^2 logL(\underline{\hat{\theta^0}})}{\partial \underline{\hat{\theta^2}}}\right)^{-1}\left(\frac{\partial logL(\underline{\hat{\theta^0}})}{\partial \underline{\hat{\theta}}}\right) \tag{2.14}$$

The vector $\frac{\partial logL(\underline{\hat{\theta^0}})}{\partial \underline{\theta}}$ is known as the *score function* at $\underline{\hat{\theta}^0}$ while the matrix $I(\underline{\hat{\theta}^0})= -\frac{\partial^2 logL(\underline{\hat{\theta^0}})}{\partial \underline{\theta^2}}$ is referred as the *sample information matrix* at $\underline{\hat{\theta}^0}$. Note that taking expectation of $I(\underline{\hat{\theta}^0})$ as shown below, yields in the *Fisher information*, $\underline{I(\theta)}$ for the entire sample.

$$E(I(\underline{\theta^0}) = \left(-E\ \frac{\partial^2 logL(\underline{\hat{\theta}})}{\partial \theta_k \partial \theta_j}\right) = \underline{I}(\underline{\theta}) = \sum_i^n \underline{I}_i(\underline{\theta}) = n\ \underline{I}_i(\underline{\theta})$$

The iterative equation (2.14) is called the Newton-Raphson equation. Substituting the sample information in (2.14) by the Fisher Information gives

$$\underline{\hat{\theta^1}} = \underline{\hat{\theta^0}} + \underline{I}^{-1}(\underline{\hat{\theta}^0})\frac{\partial logL(\underline{\hat{\theta^0}})}{\partial \underline{\theta}} \tag{2.15}$$

which is an iterative scheme known as the method of scoring. Once the estimate $\underline{\hat{\theta}^1}$ is obtained, we then expand (2.13) again using Taylor series about $\underline{\hat{\theta}^1}$ and find the solution $\underline{\hat{\theta}^2}$ that satisfies (2.14). This whole process is repeated until convergence is achieved. Log-likelihood functions are approximately quadratic because asymptotically, normality occurs for many random variables. Thus, the Newton-Raphson and Scoring method is an obvious choice for finding MLEs.

2.4.9. ***Confidence Intervals and Tests for MLE Estimates.***
Under the condition of smoothness, the MLE estimates of $\underline{\hat{\theta}}$ from section (2.4.8) is asymptotically distributed as $\underline{\hat{\theta}} \sim \mathrm{N}(\underline{\theta}, \mathrm{I}^{-1}(\underline{\theta}))$. We can use this information to construct hypothesis tests and confidence intervals for $\underline{\hat{\theta}}$. This asymptotic information allows us to formulate three procedures; the *Wald test*, the *Neyman-Pearson/Wilks likelihood ratio test* and the *Rao statistics*. To explain these methods, consider the hypothesis H_0: $\underline{\theta}=\underline{\theta^0}$.

(i) The *Wald test*
Under H_0,

$$\left(\underline{\hat{\theta}} - \underline{\theta^0}\right)' \underline{I}(\underline{\theta^0}) \left(\underline{\hat{\theta}} - \underline{\theta^0}\right) \sim \chi_p^2$$

(ii) The *Neyman-Pearson/Wilks Likelihood ratio test*
Under H_0,

$$-2log\left(\frac{L(\underline{\theta^0})}{L(\underline{\hat{\theta}})}\right) \sim \chi_p^2$$

(iii) The *Rao statistics*
Under H_0,

$$\left(\frac{\partial log L(\underline{\theta^0})'}{\partial \underline{\theta}}\right) \underline{I}^{-1}(\underline{\theta^0}) \left(\frac{\partial log L(\underline{\theta^0})}{\partial \underline{\theta}}\right) \sim \chi_p^2$$

From above, we notice that the Rao statistics is independent of the MLE estimate of θ, hence this method does not involve any iterative process. Moreover, in order to test estimates of θ or get confidence intervals for them, the calculation of $\hat{\theta}$ is necessary as we need to compute $\underline{\mathrm{I}}(\underline{\theta^0})$ so as to carry out the Wald test.

2.4.10. ***Estimation of Survival Function, S(t) for parametric models.***
The main objective in survival analysis is the determination of the survival function, S(t) which is of the form described in equation (2.5). The survival function allows, for example, to determine the probability that a patient who underwent a hip replacement at age x, to survive for time period,t. S(t) is easily constructed once the MLE estimates of the parameters are determined. Under the exponential model, $\hat{S}(\mathrm{t})$ is of the form $\hat{S}(\mathrm{t})=\mathrm{e}^{-\hat{\lambda}t}$ while the Weibull model has the form $\hat{S}(\mathrm{t})=\mathrm{e}^{(-\hat{\lambda}t)^{\hat{\alpha}}}=\mathrm{e}^{-\hat{\gamma}t^{\hat{\alpha}}}$.

Similarly we can re-parametrise and transform our data such that the problem of estimating our model parameters changes to estimating location and scale parameters. Consider the Weibull distribution for example. We can write

$$\begin{aligned} P(Y > t) &= S(t) = e^{-\lambda t^{\alpha}} \\ &= exp(-esp[\alpha(log\lambda + logt)]) = exp\left\{-exp\left(\frac{logt - \mu}{\sigma}\right)\right\} \end{aligned}$$

where $\mu = -log\lambda$ and $\sigma = \frac{1}{\alpha}$. So we observe that μ and σ are the location and scale parameters of the random variable $logY$. This method can be easily generalised to other distributions and it is known as the *Linear Combination of Order Statistics.*

In most research, the survival time, T often depends on a set of explanatory variables. Let X be the vector representing these variables. Then the survival functions can be written as a linear regression function in the form of $E(T) = \alpha + \beta X$ or a log-linear regression model as $logE(T) = log\alpha + \beta X$, where β is the coefficient of the explanatory variables of the model. The parameters are again estimated via the maximum likelihood method. However the linear model is not ideal as it yields in negative estimate of E(T) when $\hat{\beta}$ is negative. The log-linear model is a commonly known as a precursor of the Cox proportional models, which will be discussed in later section.

2.5. **Non-Parametric Model.**

2.5.1. ***Introduction.***

In previous section, we presented survival models for data that are assumed to follow a particular survival distribution. Non-parametric techniques, in contrast offers the possibility to explore survival data that are not restricted by any particular distributional form assumption. In this section we will discuss the several techniques present in literature for survival analysis on how to estimate the survivor function under a non-parametric framework.

2.5.2. ***Empirical Survivor Function.***

In the absence of censored data, we can use the *empirical survivor function* to estimate the survivor function at time t. This method assumes that the probability that a subject survives beyond a time point,t is proportional to the number of individuals who are still alive after time t. Thus the survival function is given by

$$\hat{S}(t) = \frac{No.\ of\ subjects\ with\ survival\ times > t}{No.\ of\ subjects\ in\ data\ set} \tag{2.16}$$

However, survivor function (2.16) is invalid when censored data are present. Alternatively, we can divide the study period into a set of discrete time intervals where we estimate the survival function at each of these discrete time intervals. In this case the survival estimates are assumed to be proportional to the total number of individuals deemed 'at-risk' in each time interval.

2.5.3. ***Life Table and Actuarial Method.***

In presence of censored data, the classical method of estimating S(t) in epidemiology and actuarial field is via the use of the *life table* and *actuarial*method. To illustrate these methods, we firstly partition the study period into a series of intervals I_i and of length τ_i for i=1,...,k. Let n_i be the number of individuals alive at the start of I_i, d_i be the number of individuals died during I_i, ℓ_i be the number of individuals lost to follow up during I_i, w_i be the number of individuals withdrawn during I_i, P_i be the probability of surviving through I_i given that the subject is alive at the start of I_i and $q_i = 1 - p_i$.

To estimate S(τ_k), the *life table* method only takes into account the observations who are deemed to be at risk during the whole interval of interest, [0,τ_k]. Let

$$n = n_1 - \sum_{i=1}^{k} \ell_i - \sum_{i=1}^{k} w_i,$$
$$d = \sum_{i=1}^{k} d_i,$$
$$then\ \hat{S}(\tau_k) = 1 - \frac{d}{n}$$

Under the *actuarial method*, we define the survival probability function S(τ_k) as a product of probabilities as shown below:

$$\begin{aligned} S(\tau_k) &= P(T > \tau_k) \\ &= P(P(T > \tau_1)\ P(P(T > \tau_2 | T > \tau_1)...P(T > \tau_k | T > \tau_{k-1}) \\ &= \prod_{i=1}^{k} p_i,\ where\ p_i = P(T > \tau_i | T > \tau_{i-1}) \end{aligned}$$

From above actuarial formulation of S(τ_k), we require an estimation of p_i. Under the assumption of no losses or withdrawals, we can use $p_i = 1 - \frac{d_i}{n_i}$. However this is not often the case as ℓ_i and w_i are non-zero. In this case, we estimate p_i by using the *effective sample size* defined as $n_i' = n_i - \frac{1}{2}(\ell_i + w_i)$ instead of n_i so that p_i is given by $p_i = \frac{d_i}{n_i'}$.

Both the life table and actuarial method of estimation for S(t) have certain drawbacks. Firstly the life table method does not take into account of the information the ℓ_i and w_i contains and thus provides a biased (downward) estimate for S(t). Secondly, the actuarial method estimate of S(t) for censored observations depend on n_i'. n_i' loses half of the information provided by ℓ_i and w_i as its computation assumes that on average, lost or withdrawn observations are at risk only for half the interval.

2.5.4. ***Product Limit (PL) estimator (Kaplan-Meier (KM)).***
The Product Limit (PL) estimation method offers an alternative way to estimate S(t) under a non-parametric framework. Unlike the life table or actuarial method, PL produces better estimate of S(t) as it does not have the disadvantages mentioned in previous section. In this section we will describe the PL estimate method for S(t).

Consider a sample of n individuals in an investigation whose time period is divided into different time intervals with variable length. Assuming that failure (or death) occurs at the start of each interval, then we can form a series of intervals that contain only one failure at a time. If there are r≤n failures, then we can have t_j, j=1,...,r as the series of ordered failure times such that the first interval [t_0,t_1] has no failure. If observations are tied, then censoring is accounted for after failure.

Let n_j and d_j be the number of individuals at risk prior to t_j and the numbers of failures at t_j. Assuming independent failures, then survival between t_j and t_{j+1} is

given by $\frac{n_j-d_j}{n_j}$ and thus the KL estimate of S(t) for $t_j \leq t < t_{j+1}$ is given by

$$\hat{S}_{KL}(t) = \prod_{k=1}^{j} \left(\frac{n_j - d_j}{n_j} \right),$$

We observe that $\hat{S}_{KL}(t)$ is a decreasing step-function where $\hat{S}_{KL}(0)=1$ and $\hat{S}_{KL}(t)$ remains unchanged over each time interval $t_j \leq t < t_{j+1}$, j=1,...,r, where $t_{r+1}=\infty$. $\hat{S}_{KL}(t)$ allows the derivation of several quantities of interest such as mean, median, quartiles, associated standard errors, confidence intervals for $\hat{S}_{KL}(t)$, hazard and cumulative hazard functions. It also permits development of plots that provide useful inference about the form of the survival distribution. Collett (2003) and Lee and Wang (2003)provides a detailed text on derivations of these quantities.

2.6. **Semi-Parametric Models.**

The most important issue in survival modelling is to investigate the effect of covariates on survival time. Hence a different approach for this analysis is required. Cox (1972) proposed a model using the hazard function. He suggested that for an individual with a vector of covariates $\underline{x}$, the hazard at time t is made up of firstly, a baseline hazard function in the absence of covariate information and secondly, a parametric function that represents the effects of the covariates on the failure time, over and above the baseline hazard (See Cox and Oakes 1984, Chapter 5). Moreover Cox (1972) introduced this method without having to assume an underlying distribution for the data. The model is defined as

$$h(t, \underline{x}) = h_0(t)\phi(\beta; \underline{x}) \tag{2.17}$$

where $\underline{x}$ is a vector of length p fo explanatory variables, $\phi(.)$ is a parametric function of $\underline{x}$ and $h_0(t)$ is an unspecified baseline hazard function (when $\underline{x} = \underline{0}$). β is also a vector of p parameters. The baseline hazard function in model (2.17) is arbitrary and thus (2.17) is commonly known as a semi-parametric model (or as the proportional hazard (PH) model).

The PH model is famous for various reasons- as formulated by Cox and Oakes (1984). They suggest that it is reasonable to accept the idea that the effect of a covariate is to multiply the hazard by a constant factor. The real strength of the proportional hazards model is that it allows us to model the relationship of survival time, through its hazard function, to many covariates simultaneously. Similarly the model can easily accommodate censored data and the occurrence of multiple failures. Moreover although the underlying survival distribution is unspecified, the model is easily fitted.

2.6.1. *Model Fitting.*

Cox (1972) developed a method known as the *partial likelihood* method to fit the proportional hazard model (PH) in (2.17).This method does not account for actual censored and uncensored survival times. Consider a set of n individuals with r≤n ordered failure times t_j, j=1,...r. The form of *f(.)* is unknown as no distributional form is associated with the data. Hence Cox (1972) formulate the likelihood using the conditional probability that individual i experiences failure at time t_j given

that he survives to t_j and the additional notion of risk-sets. Moreover this method assumes that intervals between successive failures do not contribute any information to the likelihood because conceptually, h_0(t) can be zero in those intervals.

Assuming independent failure times t_j, the following statements hold.

$$\begin{aligned} &P(individual\ i\ fails\ at\ t_j | one\ failure\ at\ t_j) = \\ &= \frac{P(individual\ i\ fails\ at\ t_j\ and\ no\ other\ failure\ occurs)}{P(one\ failure\ at\ t_j)} \\ &= \frac{P(individual\ i\ fails\ at\ t_j\ and\ no\ other\ failure\ occurs)}{\sum_{k \epsilon R(t_j)} P(individual\ k\ fails\ at\ t_j\ and\ no\ other\ failure\ occurs)} \end{aligned} \tag{2.18}$$

If we consider short time interval Δt, then letting Δt$\rightarrow$0, (2.18) can be written as

$$\frac{P(individual\ i\ fails\ in\ [t_j, t_j + \Delta t])/\Delta t}{\sum_{k \epsilon R(t_j)} P(individual\ k\ fails\ in\ [t_j, t_j + \Delta t])/\Delta t} \tag{2.19}$$

Hence if individual i has a vector of covariate x_j, then we transform (2.19) into

$$\frac{h(t_j|x_j)}{\sum_{k \epsilon R(t_j)} h(t_j|x_k)} = \frac{exp(\beta^T x_j)}{\sum_{k \epsilon R(t_j)} exp(\beta^T x_k)} \tag{2.20}$$

using (2.6) where $h(t_j|x_j) = h_0(t_j) exp(\beta^T x_j)$. Therefore for r failures, the partial likelihood is written as

$$L(\beta) = \Pi_{j=1}^{r} \frac{exp(\beta^T x_j)}{\sum_{k \epsilon R(t_j)} exp(\beta^T x_k)} \tag{2.21}$$

The effect of the covariates x_j are estimated by evaluating the β parameters. Cox(1972) demonstrated that inferences on β can be obtained by maximising (2.12), which can otherwise be written as

$$log(L(\beta)) = \sum_{j=1}^{r} \beta^T x_j - \sum_{j=1}^{r} [\sum_{k \epsilon R(t_j)} exp(\beta^T x_k)] \tag{2.22}$$

The score function of (2.22) is similar to the score equation vector described in (2.13) and hence it is solved iteratively using the Newton-Raphson method described in (2.4.8). In addition, this method of estimation permits construction of tests (or more specifically the Wilcoxon statisics) for hypothesis about β's by using the sample information matrix described in (2.4.8).

Tied or Grouped Observations.

The presence of tied observations in our data complicates the derivation of the partial likelihood. Therefore, we need a different technique to construct the partial likelihood in the presence of tied data. In this section we will review the existing methods for finding the likelihood function for tied observations following the Cox model.

The common and standard approaches are Breslow (1972) and Effron (1977) approximations which are simple to implement; see also Therneau and Grambsch (2000). Define T_i for $i = 1, ..., k$ as two tied observations from a data set based on Cox regression model. Let $R_i(t) = Y_i(t)exp(X_i^T\beta)$. The Breslow (Breslow 1972; Peto 1972) approximation is given by

$$\prod_{i=1}^{k} \frac{R_i(T_1)}{R_1(T_1) + R_2(T_1) + ... + R_n(T_1)}$$

while the Effron (Effron 1977) approximation is given by

$$\prod_{i=1}^{k} \frac{R_i(T_1)}{\frac{k-i+1}{k}\sum_{j=1}^{k} R_j(T_1) + R_{K+1}(T_1) + ... + R_n(T_1)}$$

We observe that the Effron approximation uses an average of the k relative-risk terms. Nevertheless both approximations suggested above do not result in expected score functions that are equal to zero and hence produce biased estimates. Breslow estimator for instance, produces estimates that are shrunk towards zero. There are several exact solutions method proposed in literature for partial likelihood or tied observations but these method are computationally extensive, ad-hoc and do not improve the approximations proposed by Effron.

Scheike (2007) proposed an improved approximation method that is easy to implement, more efficient and which is based on the EM-algorithm. Their approach is related to the techniques for interval censored data described in Satten (1996). They tested their proposed approximation using simulated studies and the results obtained suggest that the proposed EM-algorithm is a reliable procedure to use in practice because of its overall well performance across different covariate distributions, estimates, sample sizes, tie sizes and censorship status. For a detailed explanation of the EM-algorithm approximation, see Scheike (2007).

Accelerated Life Model.
An alternate way to the assumption of proportional hazards is to consider the covariates effects directly on failure time via the application of *accelerated life model*. This method models the logarithm of the survival times as a linear combination of the covariates, i.e,

$$log(T) = \beta^T x$$

Unlike the PH model which assumes a multiplicative effects of the covariates on the baseline hazard function that is independent of time, the accelerated life model allows the covariates to directly accelerate or decelerate the failure time. For a more detailed explanation and analysis of this approach, see Cox and Oakes (1984), Therneau and Grambsch (2000) and Kalbfleisch and Prentice (2002).

2.7. **Time Dependent Covariates.**

The survival models discussed so far in previous sections has not accounted for the fact that some covariates are time-dependent for some data set. These models can easily be adapted to accommodate for these time-dependent covariates although this will change the interpretation of the models. Consider a time-dependent covariate

$X_i(t)$ for individual i. If we denote the covariate history up to time t as $X_i(t) = x_i(u); o \leq u \leq t$, then the hazard function for individual i is given by

$$h_i(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T \geq t, X_i(t))}{\delta t}.$$

According to Kalbfleisch and Prentice (2002), there exists two types of time-dependent covariates. The first one is known as the *external covariate* and is defined as the covariate whose future path to any time $t > u$ is independent of the occurrence of a failure at time u. For example the air temperature in a hip replacement study is a possible example.

Secondly we have *internal covariate* which is a set of measurements taken on an individual study subject, leading to a common property of requiring the survival of the individual for its existence. For example, in a study of survival time from a total hip replacement operation, a set of measurements of blood pressure can be an internal covariate.

2.8. **Bayesian Model Fitting.**

There exists a range of techniques to fit the survival models, each having its own advantages and disadvantages. In this section we will focus on the Bayesian approach to fit the survival models. Such approach has a number of advantages, for instance it does not only output the full posterior distributions of the parameters but the predictive distributions of the predicted values. It also permits development of tractable method to fit complex models such as mixed models that consist of random effects. For a detailed explanation on Bayesian methodology, see Gelman et al. (2004), Congdon (2001) and Congdon (2003) while Ibrahim et al. (2001), Hougaard (2000) and Collett (2003) provide a clear account of the application of Bayesian techniques to survival analysis.

2.8.1. *Basic Ideas.*

Under a Bayesian approach, the parameters of interest are treated as random variables. It is assumed that they can be generated from a some probabilistic distribution. The standard Bayesian model for a set of parameters, θ given datat D, is of the form

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_\theta p(\theta|D)p(\theta)} \tag{2.23}$$

(2.23) shows that the conditional posterior distribution for the parameters θ given data D is given by the product of the likelihood of the distribution of D given θ and the prior distribution for θ, up to some normalising constant. (2.23) is easily computed explicitly for simple models but since the denominator consists of integrating over the whole parameter space, computation of (2.23) becomes mathematically intractable for a large set of parameters. Hence a different mechanism is needed to fit (2.23). The most common one is the *Markov Chain Monte Carlo* (MCMC) iterative sampling and this is discussed in the next section.

2.8.2. *Markov Chain Monte Carlo (MCMC) Iterative Method.*
The MCMC method is carried out in two steps. Firstly it includes the Monte Carlo (MC) integration which involves sampling a large quantity of observations from a target distribution and then use these samples to obtain an estimation of several expected values. By the law of large numbers, the estimates become more accurate as sampling size increases. Hence if we obtain large samples from the posterior distribution $p(\theta|D)$, then the MC integration provides a useful method to obtain required quantities of interest from these values.

The next step includes the development of a tractable method that allows sampling from the posterior. The most common method is the Markov Chain. It is a sequence of numbers where the generation of each number depends on the value of the previous number in the chain. Under certain regularity conditions, a Markov Chain always converges to its stationary distribution. If the stationary is similar to the posterior distribution of interest, then we can sample the desired values. Hastings (1970) proposed an algorithm known as the *Metropolis Hastings algorithm (MH)* to construct a Markov Chain whose stationary distribution is identical to the posterior distribution, $p(.)$. Suppose we have a vector of m random variables θ from a (multivariate) distribution, $p(.)$. Then the steps involved in the MH algorithm are:

(1) Set $t = 0$ and let θ_0 be equal to some initial value.
(2) Sample a candidate point θ_c from a proposed distribution, $q(.|\theta_t)$.
(3) Accept θ_c with probability $\alpha(\theta_t, \theta_c)$ where $\alpha(\theta_t, \theta_c) = min\left(1, \frac{p(\theta_c)p(\theta_t|\theta_c)}{p(\theta_c)q(\theta_c|\theta_t)}\right)$.
(4) If θ_c is accepted, set $\theta_{t+1} = \theta_c$, otherwise we let $\theta_{t+1} = \theta_t$.
(5) Set $t = t + 1$ and repeat the process starting from step 2.

Depending on the regularity conditions, the proposal distribution $q(.|.)$ can be of any form and will still converge to the distribution of interest. The rate of convergence of the chain depends on the appropriate choice of the proposal distribution. Moreover the choice of initial values for the parameters is important as it affects the rate at which convergence to the targeted distribution occurs.

Alternately, the vector of parameters θ does not need to be updated as a block. The parameters can be updated individually with corresponding modifications to the proposal distributions. Consequently we have to use a special case of the Metropolis-Hastings algorithm when the full conditional distributions for individual parameters θ_i, $i = 1, ..., m$, given $\theta_{i\_}$, i.e, $p(\theta_i|x, \theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_m)$ are known. Therefore the proposal distribution $q(\theta_{c_i}|\theta_i, \theta_{i-})$ is equal to $p(\theta_{c_i}|\theta_{i-})$ such that the acceptance probability given by step (3) of the MH algorithm is always equal to one. Such an approach is known as the Gibbs sampling (Geman and Geman 1984, Gelfand and Smith 1990).

Combinations of MH algorithm and Gibbs sampling are useful when required and the adaptive-rejection method of sampling suggested by Gilks and Wild (1992) means that even though the distribution is complex and not specified explicitly, Gibbs sampling can still be utilised as long as the conditional distributions of the parameters are log-concave. The techniques discussed here are implemented in WinBUGS (Bayesian inference Using Gibbs Sampling), a free package that can be

downloaded from http://www.mrc-bsu.cam.ac.uk/bugs/.

2.8.3. *Identifiability.*
An important issue in sampling a distribution with suitable parameters is to have *identifiable* probability densities for our distributions. Suppose we have random variable X with distribution function U_α and belonging to a family $\Lambda\{U_\alpha : \alpha \epsilon \Theta\}$ (where Θ is the parameter space) of distribution functions indexed by parameter α. Basu (1983) states that α is *non-identifiable* by X if there is atleast one pair (α, α'), where α and α' belongs to Θ, such that $U_\alpha(x) = U_{\alpha'}(x)$ for all u. In the contrary, we will say that α is *identifiable*.

2.9. **Extensions to Conventional Survival Models.**

In this section we will cover a summary of some extensions to the conventional models discussed previously that deal with different complications associated with the type of data we normally deal with.

2.9.1. *Long Term Survivor Models.*
Usually for certain survival data, it is reasonable to assume that there is a proportion, p, of the total population that are 'immune' or 'cured' from the medical conditions of interest during the investigation period. The conventional method will consider these observations as censored ones. However this intuition is unreasonable because the likelihood contributions from these observations will be incorrect and thus result in biased parameter estimates.

Boag(1949) initially proposed to incorporate an 'immune proportion' to the survival function; though this was later extended by Berkson and Cage (1952) who suggested that the hazard function for 'immune' individuals should reduce to the baseline hazard for the population. The standard model proposed as a result is modelled via the survivor function as

$$S(t) = p + (1-p)S^*(t) \tag{2.24}$$

where $S^*(t)$ is the survivor function for the population who are at risk and p is the proportion considered as 'immune'. Such a model is commonly known as the *long-term survivor* model.

2.9.2. *Mixture Models.*
The long-term survival model in (2.24) assumes that the 'immune' population never experiences failure, an assumption which is unreasonable in the medical field. McLachlan and Peel (2000) proposed a simple generalisation where the population are separated into two groups, each represented by a different survival process i.e.

$$S(t) = pS_0(t) + (1-p)S_1(t) \tag{2.25}$$

where p is the proportion of individuals from group 0, with survivor function $S_o(t)$ and $(1-p)$ is the proportion of individuals from group 1 with corresponding survivor function $S_1(t)$. This is an example of a standard two-group mixture model and it can be generalised to three or more groups.

2.9.3. *Competing Risks Models.*
Often there are situations where failures occur because of more than one cause. Traditional survival analysis do not differentiate between multiple causes of failures. Incorporating this extra information to the survival model is carried out by the *competing risk analysis* as proposed by McLachlan and Peel (2000), Crowder (2001) and Congdon (2001). They propose a model analogous to model in (2.25) i.e.

$$S(t) = pS_0(t) + (1-p)S_1(t) \tag{2.26}$$

where now p represents the probability that failure occurs from cause 0 and $(1-p)$ is that of failing due to cause 1. Unlike the mixture model, here we assume that the causes of failures are to be observed and independent. Ng and McLachlan (1998) explains techniques to accommodate for missing data as well as various extensions to a more generalise form of (2.26) to incorporate more than two causes of failure, and also as part of a long-term survivor model (2.24).

2.9.4. *Multi-state Models.*
The models discussed so far assume that the survival process for an subject is unchanged over time. However there are situations where the risk of failure may change over time. A multi-state framework provides a stochastic process allowing grouping of individuals into discrete states at any time point. This method is flexible and useful for dealing with longitudinal data. Consider Figure below which shows a possible graphical representation of the standard competing risks model (2.26) under a multi-state framework.

Changing from one state to another is referred as a *transition* and this is processed via *transition hazard* (h_0 and h_1 in Figure). One advantage from this method is that interpretation and models fitting become easier as a particular state structure can be formulated in different forms under the multi-state framework. As for the likelihood formulation, it is derived assuming that transition between states is governed by a Markov Process as discussed in Hougaard (2000, 1999) and Commenges (1999).

2.9.5. *Change Point Models.*
The change point model is an extension and similar concept to the multi-state model described previously. Here we assume that the form of the distribution of the survival process changes at one or more time points unlike the multi-state model which considers transition of individuals instead. Ebrahimi et al. (1997) and Chung et al. (2005) developed Bayesian models for a process involving n change points. Chung et al. (2005) defines the hazard under the change point models as

$$h(t) = h_1(t)I(0 \leq t \leq \varsigma_1) + ... + h_n(t)I(\varsigma_{k-1} < t < \varsigma_n) + h_{k+1}(t)I(t > \varsigma_k) \tag{2.27}$$

where ς=$(\varsigma_1, ..., \varsigma_k)$ is the vector containing the change point parameters with $I(U) = 1$ if $x \, \epsilon \, U$ or 0 otherwise.

2.10. **Conclusion.**

In this section, we have seen that survival analysis offers a large variety of options when it come to dealing with a host of different situations, each having their own complications, when investigating failure time data. The conventional survival approach discussed provides a straightforward way to deal with censored information as well as methods to work with data having failure times that are not normally distributed. Moreover we have also introduced several techniques on how to fit these survival models and produce estimation methods for their parameters. In the next section, we will focus on the idea of allowing for random effects and the notion of multi-level modelling in survival analysis.

3. **Survival Models with Frailties**

3.1. **Introduction.**

The idea of frailty provides an appropriate way to introduce random effects, association and unobserved heterogeneity present in survival data. In the simplest form, a frailty is defined as an unseen random proportionality factor that modifies the hazard function of an individual or of related individuals. Fundamentally, the frailty notion goes back to the work of Greenwood and Yule (1920) on "accident proneness". Vaupel et al. (1979) introduced the term frailty in univariate survival models and Clayton (1978) extensively promoted its application to multivariate survival data in a seminal paper (without using the concept of "frailty") on chronic disease incidence in families.

Frailty models are extensions of the proportional hazards model which is best referred as the Cox model (Cox, 1972). Usually in most medical studies, survival analysis basically assumes presence of a homogeneous population. This means that all individuals participating in that study are subject to the same type of exposure (e.g., risk of death, risk of medical condition recurrence). In many cases however, the study population is not homogeneous and must be assumed as a heterogeneous sample, i.e. a combination of individuals with different hazards. For example, in many cases it is not possible to have measurements of all relevant covariates linked to the medical condition of interest, sometimes because of economical reasons or sometimes the essence of some covariates is still unfounded.

The frailty approach is a statistical set up which provides a method to account for heterogeneity caused by unexplained covariates. Statistically speaking, a frailty model is a random effect model for time-to-event data where the effect of the frailty term on baseline hazard function is multiplicative. One can identify two broad types of frailty models:
(1) models with an univariate survival time as endpoint and
(2) models with multivariate survival endpoints (e.g; competing risks, recurrence of events in the same individual, occurrence of a disease in relatives).

Why random effects?

Under standard situation, most clinical research assume that the population being investigated is homogeneous and that the survival data are independent from each other, with independent and identically distributed survival times. However in the medical field, patients differ substantially. The effect of a particular treatment or

the influence of various explanatory variables may vary significantly between subgroups of patients.

The difference between subgroups of patients arise mainly because it is impossible to account for all essential factors on the individual level into the analysis. There are two reasons why this is the case. Firstly sometimes there are too many covariates to account for in the model and secondly the researcher may be unable to obtain measurements of all relevant covariates. These two cases are the sources of two different types of variability in the data: variation due to risk factors that can be measured (which is thus theoretically predictable) and heterogeneity caused by unknown covariates, which is thus theoretically unpredictable. Analysing these two sources of variation separately offers some advantages since heterogeneity in contrast to variability help to understand some "unexpected" results or provide an alternative explanation of some results.

Ignoring a subset of the essential covariates produces biased estimates of both regression coefficients and the hazard rate in proportional hazard model. This is because bias lies in the fact that the time-dependent hazard rate results in changes in the composition of the study population over time with respect to the covariates. Consider for example two groups of patients where some of them have a higher risk of failure, then the remaining exposed individuals tend to form part of a selected group with a lower risk. Estimating the individual hazard rate without accounting for unobserved frailty would thus underestimate the true hazard function and the degree of underestimation would increase as time progresses.

3.2. **Univariate Frailty Models.**

In the univariate case, the frailty model extends the Cox model given by (2.17) such that the hazard of an individual depends in addition on an unseen random variable Z which has a multiplicative effect on the baseline hazard function $h_0(t)$:

$$h(t, Z, \underline{x}) = Z h_0(t) \phi(\beta; \underline{x}) \tag{3.1}$$

Again h_0, β and $\underline{x}$ are as defined in section (2.6) and Z is the unobserved random variable (frailty) fluctuating over the population which lowers ($Z < 1$) or raises ($Z > 1$) the individual risk. Here frailty corresponds to the notion of susceptibility in different settings (Falconer, 1967). The corresponding survival function S for the proportion of individuals surviving to time t in the study is given by

$$S(t|Z, \underline{x}) = exp(-Z \int_0^t h_0(s)ds \; exp(\beta^T \underline{x})) \tag{3.2}$$

So far, the survival function described in (3.2) is at the individual level and this is not observable. Therefore we need to consider the model at the population level. The survival function at the population level is given by the mean of the individual survival functions in (3.2). It can be regarded as the survival function of an observable individual that has been randomly selected. Similarly it is important to note that the observed hazard function will not be similar to the individual hazard rate because what we can observe in the study population is the net result for a number of individuals with different Z. The shape of the individual hazard rate

may be completely different from that of the population.

Frailty models in the univariate case is common and widely applied. Some of the examples which can be reviewed in details are listed here. Firstly Aalen and Tretli (1999) used the compound Poisson distribution to model testicular cancer data. The idea behind such model is that a cluster of individuals showed higher susceptibility to testicular cancer, causing selection over time. Similarly Hougaard (2000) used the data set consisting of patients who underwent radical surgery for skin cancer at the University Hospital of Odense in Denmark to compare the traditional Cox model with a gamma frailty and the three parameter distribution (Hougaard, 1986a) frailty model.

Another example of the application of univariate gamma frailty model is the model proposed by Hougaard (2000) to account for heterogeneity in a data set that deals with the time from insertion of a catheter into dialysis patients until it is removed because of infection. In a more detailed paper, Congdon (1995) investigated the influence of the choice of different frailty distributions (gamma, inverse Gaussian, stable and binary) for a data set consisting of total and all cause-specific mortality for individuals from London (1988-1990).

3.3. **Multivariate Frailty Models.**

Usually data in survival analysis are assumed to be univariate. There exists a set of standard methodology including the Kaplan-Meier plots and Cox analysis to deal with such data set. However there are certain contexts where multivariate survival data arises naturally and therefore poses for ordinary multivariate methods, specially when it comes to dealing with censored multivariate data set.

Typically one can obtain multivariate survival data in two ways. The first case is referred as *recurrent events* and it occurs when an investigator has to record the occurrence of several successive events of the same category for each individual. For example, one may record observations for the same patient each time the latter have a cardiovascular disease. The second one which is called as the *clustered survival data*, consists of several units whose failure are observed and collected in cluster. For example, one may record the death times of patients undergoing similar medical treatments within the same ward. Here ward is considered as one cluster. In such cases, we can no longer assume independence between clustered survival times. Multivariate models enables us to account for the dependence between these event times.

Another issue that is widely discussed in the multivariate frailty models literature is censoring, which takes a new dimension since censored information for individuals depend on previous events for the same individual and possibly on previous events of other individuals. Knowing the nature of censored information is essential to write up likelihood functions and thus statistical analysis. Usually time scales are reshuffled in statistical analysis for censored information, so that the starting point for each individual is zero. However if censoring is influenced by decisions being taken or by the occurrences of other processes, then these influences should follow the actual real time. This creates a complication since reshuffling the censored time

scales can no longer be carried out. Hougaard (2000) provides a detailed section on how to deal with such complication for multivariate survival data while Andersen et al.(1993) and Lee (1989, Chapter 7) provides a general theory concerning likelihood based statistical methods for censored multivariate survival data.

As mentioned previously, assumption of independence no longer holds between clustered survival times. Hougaard (2000) proposed a general approach that is widely used to account for presence of dependence between clustered data in multivariate models. The proposed method specifies independence among observed data conditional on a set of unseen and unobserved or latent variables. The dependence structure for the multivariate model comes from a latent variable in the conditional models for multiple survival times. For example, suppose we have $S(t_1|Z, \underline{x}_1)$ and $S(t_2|Z, \underline{x}_2)$ as the conditional survival functions of two related individuals where $\underline{x}_1$ and $\underline{x}_2$ are the different vector of observed covariates respectively. Then averaging over an assumed distribution for the latent variables (e.g., using a gamma or log-normal or stable distribution) helps in developing a two-dimensional multivariate survival model of the form:

$$S(t_1, t_2) = \int_0^\infty (t_1|Z, \underline{x}_1)(t_2|Z, \underline{x}_2)g(Z)dZ,$$

where g(.) denotes the density of the frailty Z. In this section we showed that multivariate observations do not guarantee independence between clustered survival times. Frailty models for such data are derived under the conditional independence assumption that mainly consists of specifying the latent variables in a way that has a multiplicative effect on the baseline hazard.

3.4. **The shared frailty model.**

The shared frailty model refers to event times of individuals that are related either by sharing a common characteristics or due to experiments producing repeated measurements for each individuals. In this way the individuals sharing that common feature are clustered within the same group. Clayton (1978) introduced the idea of assuming a common frailty Z for individuals in the same cluster. This idea is extensively studied in Hougaard (2000) which provided an important assumption about the conditional independence of the survival times to the common frailty.

To illustrate the idea of shared frailty, consider the example where we have clusters of pairs of individuals, each having bivariate survival times (e.g., event times of parents, twins, wife-husband). Extensions to the multivariate case is straightforward. Conditioning over the common frailty Z, the hazard function for a pair of individual is given by $Zh_0(t)exp(\beta^T \underline{x})$, where Z is the common frailty for both individuals, causing survival times within pairs of individuals to be dependent on each other. Taking the survival times within a pair corresponds to a degenerate frailty distribution ($Z = 1, \sigma^2 = 0$) while for cases when $\sigma^2 > 0$, we obtain a positive dependence by construction of the model. Conditioning on Z, the bivariate survival function is written as

$$S(t_1, t_2|Z) = S_1(t_1)^Z S_2(t_2)^Z$$

For instance, if we assume a Gamma distribution for the frailty with mean equal to 1 and variance σ^2, then the conditional bivariate survival function is of the form

$$S(t_1, t_2) = (S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1)^{1/\sigma^2}$$

Although shared frailty is a nice way to allow for correlation between observations within clusters, it however has some limitations. Firstly it may sometimes wrongly account unobserved factors to be the similar within clusters. This may not be true in reality for some cases. For example, it is not appropriate to assume that all partners in a group share all their hidden risk factors.

Secondly the dependence between survival times within the cluster is established using the marginal distributions of survival times. However for a model with gamma distributed frailty, the dependence parameter and the population heterogeneity are confounded (Clayton and Cuzick, 1985), implying that the joint distribution can be recognized from the marginal distributions (Hougaard, 1986a).

Thirdly, a one-dimensional frailty only generates positive association within clusters. However there exists situations where the survival times for the individuals are negatively associated. For instance, in the medical field, there are situations where the longer the patients wait to receive the appropriate intervention, the less likely the individual is to survive after the intervention. A common example is that of heart transplant. Therefore we observe a negative association between the waiting time and the survival time, which is not detected by one-dimensional frailty.

3.5. **The correlated frailty model.**

Primitively, correlated frailty models were developed to analyse bivariate failure time data, in which dependent random variables are used to outline the frailty effect for each pair. For example, suppose each member of the pair is assigned a random variable such that they do not share a common frailty. Then these two variables are related and have a joint distribution. However knowledge of one of the variable does not compulsorily mean knowing the other. In this case, the type of correlation between the two variables is free from any restriction. Assuming a Gamma distribution for the frailties, Yashin and Iachine (1995) used a correlated Gamma frailty model with a distribution as shown below for bivariate survival data.

$$S(t_1, t_2) = \frac{S_1(t_1)^{1-\rho} S_2(t_2)^{1-\rho}}{(S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1)^{\rho/\sigma^2}}$$

Multivariate frailty models are of numerous types and form part of one of the most important statistical models for survival analysis. McGilchist and Aisbett (1991) used a shared log-normal frailty model to analyse catheter infection data while Pickles et al. (1994) applied the correlated gamma frailty model to a dataset on age of onset of puberty and antisocial behaviour in British twins. Similarly, Dos Santos et al. (1995) applied a shared frailty model with gamma and log-normal distributed frailty for the analysis of recurrence of breast cancer.

We observe in a similar way the correlated gamma frailty model utilised by Yashin and Iachine (1995) and Yashin et al. (1995) to analyse mortality in a population of Danish twins. Zahl (1997) applied different versions of the correlated gamma-frailty model to account for the excess hazard present in a dataset on cancer specific mortality in Norway while Andersen et al. (1999) used a frailty model without specifying the distribution of the frailty to test for the centre effects in a multi-centre survival studies.

Manatunga and Oakes (1999) developed a shared positive stable frailty model by allowing for proportional hazards in the marginal and the conditional model and applied it to the data from the Diabetic Retinopathy Study to examine the effectiveness of laser photo-coagulation in delaying the onset of blindness in patients with diabetic retinopathy. Another example of the use of correlated gamma-frailty model can be observed in Wienke et al. (2001) and Zdravkovic et al. (2002). They applied such frailty model to analyse genetic factors influencing mortality due to coronary heart diseases in Danish twins. Therefore we can conclude that there exists several methods that are extension to the Cox proportional hazard model to accommodate for correlations between survival times in frailty models

3.6. **Choice of frailty distributions.**

In previous sections, we introduce the idea of incorporating random effects in survival models for various reasons as explained before and also provide an overview of the several methods that have been employed to account for random effects. In this section we discuss the various statistical distributions that have been proposed in literature to represent these random effects. Choosing the appropriate frailty distribution is crucial because the frailty distribution contributes to the definition of the dependence contribution arising in the data. Dependence between correlated observations changes over time, thus choice of frailty distribution needs to be cautiously carried out as the latter dictates how dependence in data evolves with time.