Practical Methods for Analysing Dependent Survival Data

John Newell M.Sc.

A Dissertation submitted to the
University of Glasgow
For the degree of
Doctor of Philosophy

Department of Statistics

October, 1999

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

Survival data arises when there is interest in the length of time until a particular event occurs e.g. death due to cancer. Typically, observations are assumed to be statistically independent of each other. This assumption however is violated in many situations which are not as uncommon as one might think. The overall aim of this thesis is to provide practical methods for analysing dependent survival data. All the methods described in this thesis are illustrated using paired survival data from an Orthodontic study and matched survival data from a Melanoma study.

Chapter 1 gives a brief background to survival data, common censoring mechanisms and estimation of the survivor function. A review of some of the standard techniques for summarising survival data is given with particular emphasis on non-parametric estimators of the survivor function.

A discussion of situations where the assumption of independence between observations is likely to be invalid is given in Chapter 2 where Multiple Event and Cluster Survival studies are introduced. This thesis, however, primarily concerns analysing dependent survival data from cluster studies (i.e. where a failure process acts concurrently on individuals in a cluster). Such studies are of two types, namely ***paired*** studies (e.g. time to cataract in left/right eye) and ***matched*** studies where the individuals are matched by design (e.g. comparing time to death). Both matched and paired survival studies will have a pair of observation times recorded which represent

the two 'arms' of the primary variable of interest. In addition to these, additional information may be recorded also in the form of covariates, or prognostic indicators. Matched survival studies will, by definition, have the variables used for the matching present and some additional unmatched covariates, or prognostic indicators, may also be recorded for each individual. Graphical and analytical methods for assessing the quality of matching in matched survival studies were given also. Paired studies, by definition, are unlikely to have any matching variables available but may have 'unit' covariate information recorded for each individual e.g. sex or age.

Two example data sets are introduced (matched survival data from a Melanoma study and paired survival data from an Orthodontic study) which will be used to illustrate the various methods presented in the following chapters.

Chapter 3 presents techniques for graphically displaying dependent survival data, including bivariate survival scatterplots and survival ratio plots. A review of several nonparametric estimators of the bivariate survival function is given with methods for generating reference ranges for such three-dimensional plots. In addition, two methods to graphically assess the independent effect on survival of any continuous covariates are discussed. The first uses a form of kernel estimation to construct an estimator of a percentile of the survivor function as a function of the covariate while the second uses a tree-based approach.

Chapter 4 concerns the comparison of the survival distributions of the two arms of the primary variable (i.e. ignoring all covariates but the primary variable) where a review of several nonparametric paired 'log-rank' tests is given. Two new approaches for

comparing survival in paired/matched survival studies are described and illustrated. The first is a simple test of symmetry based on 'pair performance'. The second is based on estimating the distribution of the (pairwise) difference in survival, using a parametric approach (by providing an interval estimate for the mean difference in survival time) and a nonparametric approach (by providing an interval estimate for an appropriate quantile e.g. the median difference).

Methods for incorporating covariates into the analysis, while at the same time taking the dependency structure of the data into account are presented in chapter 5. A 'covariate adjusted' comparison of the two 'arms' of the primary variable should then be less biased and more precise than a 'covariate free' comparison.

In matched survival studies the matching covariates are available for inclusion in the analysis while in paired studies the covariates representing the degree of similarity for the pair are often unobservable, that is, 'hidden' from the analysis.

A new approach for modelling 'pair performance' which allows for covariates is presented. Regression models for the hazard rate are discussed. Several extensions of the proportional hazards (PH) model to clustered studies are proposed. The conditional PH model ignores the matched structure of the data, however it uses information on the matching to correct inferences made on the primary variable. The justification is that the model assumes conditional independence by forcing in the matching covariates in the final model.

The second extension is similar to the conditional PH in that the regression coefficients are estimated assuming independence. The estimated covariance matrix however is then 'corrected' post fit using a paired-jacknife estimate of the variance.

A further refinement to the PH model for paired/matched survival data is to allow each pair to define a separate stratum. The association within each pair is then considered a fixed effect. An alternative more elaborate procedure introduces a random term for each pair that represents the within-pair association. In a final extension to the PH model a random term corresponding to each pair is introduced into the model. This random pair effect, often termed a 'frailty', generates dependency between the survival times of the individuals in a pair. The random effects represent unobserved covariates. Random effects are assumed to act multiplicatively on the individual's hazard rate. Survival times of all individuals are then assumed to be independent given the random effects (and any observed covariates).

In chapter 6, the results of a large simulation study which compare the different methods proposed for analysing dependent survival data are presented. A range of different degrees of censoring, sample size and primary variable effect size combinations are investigated.

Finally chapter 7 outlines the conclusions and suggests some ideas for further work.

Acknowledgements

I would like to express my sincere thanks, appreciation and respect for my supervisor and friend Tom Aitchison.

I owe a debt of gratitude to the Fulbright Commision of Ireland, the Health Research Board of Ireland and the Department of Statistics, Glasgow University for their financial support.

I would also like to thank all the staff members and postgraduate students of the Department of Statistics in Glasgow, in particular Jim Kay, Keith Humphreys and Agostino Nobile.

Finally, I would like to thank Karen McGuire and my family for their love, support and patience.

Statistics

"Those Platonists are a curse," he said,
"God's fire upon the wane,
A diagram hung there instead,
More women born than men"

W.B.Yeats

Dóibh siúd de'r dhíobh mé

Contents

List of Figures

List of Tables

Chapter 1

Introduction To Survival Analysis

1.1 Introduction

Many studies in medical science involve studying the time taken until a particular event occurs. Survival Analysis is the general term given to describe this type of analysis. If the event of interest is death of an individual the resulting data are literally survival times. However, in medical applications, this event may be the time to development of a disease, response to a treatment, or time to relapse. The term "survival time" therefore is a bit ambiguous as it may not directly involve death as the event of interest and hence is more accurately defined as *time to event*. However, in the context of medical studies the terms survival time and survival data are used in general to represent 'time to event' data and the same convention will be adopted for the remainder of this thesis.

Survival Analysis is not limited to the biomedical field as the methods involved in Survival Analysis are often suitable for applications in industrial reliability, social sciences and business. Examples of survival data in these fields are time to failure of a particular machine, duration of first marriage and length of subscription to a magazine.

There are several special features of survival data that preclude the use of standard statistical procedures used in data analysis such as t-tests, regression analysis, analysis of variance and analysis of covariance.

1.2 Censored Data

The distinguishing feature of survival analysis is that the event times are frequently *censored* where the end-point of interest has not been observed for that individual. One would not want to exclude all of those individuals from the study by declaring them to be missing data, since most of them represent "survival" in the sense that they have not experienced the event of interest yet. Those observations, which contain only partial information are called censored observations, the term censoring being first used by Hald (1949).

A censored observation is one whose value is incomplete due to random factors for each subject. The most commonly encountered form of censoring is one in which some subjects in the study have not experienced the event of interest at the end of the study or time of analysis. As the incomplete nature of this observation occurs in the right tail of the time axis, the observation is termed ***right-censored***. An individual's observation time may be right censored therefore for a variety of reasons: from not having experienced the event by the time the study ended or having been lost to follow-up or having died due to a cause not related to the treatment under study.

A second censoring mechanism that can occur is ***left censoring*** where the event of interest has already occurred when observation begins and the actual event time is

some time less than that observed. One example of a left censored observation could occur in a study relating to time of first cigarette use when an individual has already had a cigarette at some time prior to the study which they cannot remember.

Finally, an observation is ***interval censored*** if all that is known is that the individual experiences the event in some interval of time. For example, the actual time to recurrence of a particular disease may have occurred in the interval between successive consultation visits.

An observation time can thus be one of two types: an event time (e.g. a death time) or a censored time (e.g. left, right or interval censored). In most survival contexts only right-censored data is 'observed'. Lee (1992) gives an excellent description of censoring. The main assumption concerning censoring is that the actual observation time must be independent of any mechanism that causes the individual event time to be censored.

A secondary feature pertaining to the analysis of survival data is that the distribution of event times is generally not symmetrically distributed and often tend to be positively skewed. This feature will, in most cases, rule out using the normal distribution as the underlying distribution from which the data has been generated.

The final assumption for analysing survival data is that the observation times for individuals are mutually independent. In many cases this assumption will be justified but should this assumption not hold many of the standard methods for the analysis of survival data may not be applicable. This independence assumption is violated in

situations which are not as uncommon as might be thought. A discussion of such situations is given later, and the remaining chapters of this thesis deal specifically with analysing so called 'dependent' survival data.

Some necessary background material outlining the standard methods for analysing independent survival data is now given. This will also facilitate in establishing ideas and notation essential for later chapters.

1.3 The Survivor Function and Hazard Function

The survivor function is the complement to the cumulative distribution function. It allows the evaluation of the probability that an individual survives at least to a particular time point, as in most applications interest relates to how long the subjects live rather than to how quickly they die. In practice, population survival distributions must be estimated from (representative) samples of data.

The hazard function is defined as the rate at which an individual is likely to experience the event of interest in the next small time interval given that the individual has survived up to that point. The hazard and survivor functions are formally related to each other as follows: -

Let T be a positive random variable, with distribution function F(t) and density function f(t).

The survivor function is defined as

$$S(t) = Pr(T \geq t) = 1 - F(t) \quad \text{for any } t > 0$$

This is a strictly non-increasing function with a value of 1 at the origin and decreasing to 0 at infinity. Survival distributions are usually skewed and hence the most appropriate 'central' summary of the distribution is provided by the median survival time.

For survival data the q^{th} percentile is defined as the time beyond which q% of the individuals in the population under study are expected to survive such that the survivor function equals q/100. For example, the median or 50^{th} percentile is the time for which the survivor function equals 0.5.

The hazard rate or hazard function is expressed as

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

and is defined as the instantaneous rate of failure, or the probability that an individual experiences the event in the next small interval of time Δt given that he/she has survived to time t. The hazard rate provides information as to the rate of failure of individuals over time. There are many general shapes for the hazard function (e.g. increasing, decreasing, 'bathtub' shaped) where the only restriction on $h(t)$ is that it is non-negative i.e. $h(t) \geq 0$.

The cumulative hazard rate H(t) is written as

$$H(t) = \int_0^t h(t)dt = -\log S(t)$$

indicating the relationship between the survivor function and the hazard function. The word 'cumulative' is used as it represents the "sum total" of the hazard up to time t.

1.4 Estimating the Survivor Function

Two standard approaches are commonly used in estimating the survivor function, namely parametric and non-parametric. The parametric approach involves fitting specific families of distributions to survival data. The most commonly used models are the exponential, Weibull, gamma, log normal, log logistic, Gompertz and generalised gamma distributions.

The non-parametric approach allows a more flexible estimate in that no distributional assumption is made when estimating the survivor function. This approach is now presented.

1.4.1 The Empirical Distribution Estimator

The first method proposed to provide an estimate of the survivor function was the empirical distribution function (EDF) estimator.

Given a sample entirely composed of complete data, the EDF is the simplest estimator of the survival function and is defined as

$$\hat{S}_{EDF}(t) = \frac{\text{number surviving beyond t}}{\text{number in the sample}}$$

The EDF has some good properties as an estimator of S(t); in particular, it is unbiased and consistent for S(t) (Shorack and Wellner 1986). However, in the presence of censoring, the EDF estimate is biased and a modification is needed to allow for censoring.

1.4.2 The Kaplan-Meier Estimator

The most common non-parametric estimate of the survivor function in the presence of right censored data was proposed by Kaplan and Meier (1958).

An indicator function δ is needed here to distinguish between an event time and a censored time and hence to provide 'event type' information for each individual's observation time. Let t_i denote the ***observation time*** for individual i (i=1,..,n) in a sample of n individuals. Further define $\delta_i=1$ if the i^{th} observation time is an ***event*** time, $\delta_i=0$ if the i^{th} observation time is ***right-censored***.

Hence, in survival analysis, each individual provides both an observation time and an 'event type' indicator function such that each individual i contributes (t_i, δ_i) to the dataset.

The Kaplan Meier estimator (KM) is a step function estimator of the survivor function but unlike the EDF it takes into account the fact that the observation times may be right censored.

To define the KM estimator, suppose a sample consists of n observation times $t_1, ..., t_n$ and knowledge as to which of the n observations are censored is provided by censoring indicators $\delta_1, ..., \delta_n$. Denote the subclass of distinct ordered event times in the sample by $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$ where $m \leq n$.

Define for i=1,2, .., m

e_i = number of individuals with an event at time $t_{(i)}$

r_i= number of individuals still 'at risk' at time $t_{(i)}$,

The number of individuals still 'at risk' at time t_i is defined to be the number of individuals present in the data (not having previously died or been censored) at a time just prior to t_i.

The KM estimator is the product of the estimated survival probabilities at each distinct event time i.e.

$$\hat{S}_{KM}(t) = \prod_{j=1}^{i}\left(1-\frac{e_j}{r_j}\right) \qquad \text{for} \qquad t_{(i)} \leq t < t_{(i+1)}.$$

The function defined by $\hat{S}_{KM}(t)$ fulfils the requirement of a distribution function in that it is a right-continuous non-decreasing function however it does not necessarily have total mass 1. The estimator is a step-function with new steps at each observed event time and is not well defined beyond the largest event time. If the largest event time is censored then the survivor function is undetermined beyond this point. Several alternative non-parametric suggestions have been proposed (Efron 1967, Gill 1980, Klein 1991) to account for this indeterminancy. In the absence of censoring $\hat{S}_{KM}(t)$ is simply the EDF.

A vast literature has grown up concerning the theoretical properties of the Kaplan Meier estimator from its conception as a generalisation of earlier actuarial estimators (Breslow 1992) to its practical properties (Andersen 1993).

The KM estimator places probability mass only on the (ordered) event times. The EDF would place mass 1/n at each censored and uncensored observation producing a biased estimate as the mass associated with the censored observations is not 'distributed' correctly amongst the uncensored observations. This leads to the concept of 'redistribution of mass' (Dinse 1985) which is formulated in the following way:- Consider an ordering of the observation times arranged from left (smallest) to right (largest) and initially associate mass 1/n with each observation. Beginning at the far right, move to the left and distribute the mass 1/n of the first censored observation encountered to all the uncensored times to its right (i.e. event times greater than this 'censored observation'), in proportion to the masses already accumulated at those points. This process is continued until the mass of all the censored observations has

been distributed. This resulting distribution of masses, or weights, is precisely the KM estimator (Dinse 1985). As such, the KM estimator can be likened to the EDF but with different weights at the event times where a censored observation time contributes information only to larger event times. A similar procedure involving 'redistribution to the left' was formulated by Efron (1967) for estimating the survivor function for survival data with left and right censoring present. A more detailed discussion of Efron's method will be given in chapter 4 in an appropriate context.

1.4.3 Estimating the Variance of the Kaplan-Meier Estimate

The KM estimator is an estimator of the population survivor function which can be calculated from the sample of observation times. As such, it has an associated variance which represents the precision with which it estimates S(t).

The most common estimate of the variance of the KM estimate is provided by Greenwood's formula (1926) and is defined as

$$\hat{V}_G[\hat{S}(t)] = \hat{S}_{KM}(t)^2 \sum_{j=1}^{i} \frac{e_j}{r_j(r_j - e_j)} \qquad \text{for} \qquad t_{(i)} \le t < t_{(i+1)}.$$

while Aalen and Johnson (1978) proposed an alternative estimator

$$\hat{V}_{AJ}[\hat{S}(t)] = \hat{S}_{KM}(t)^2 \sum_{j=1}^{i} \frac{e_j}{r_j^2} \text{ for } t_{(i)} \le t < t_{(i+1)}.$$

Both $\hat{V}_G[\hat{S}(t)]$ and $\hat{V}_{AJ}[\hat{S}(t)]$ tend to underestimate the true variance of the KM estimator for small to moderate samples, with $\hat{V}_G[\hat{S}(t)]$ coming closest to the true variance (Klein 1991).

An estimate of the survivor function and accompanying variance estimate can be used to provide pointwise confidence intervals for the survivor function at any specified time point. In particular $\hat{V}_G[S(t)]$ can be used to provide an approximate pointwise 95% confidence interval for the true population survivor function S(t), at time t, as follows

$$\hat{S}_{KM}(t) - 1.96\sqrt{\hat{V}_G[\hat{S}_{KM}(t)]} \leq S(t) \leq \hat{S}_{KM}(t) + 1.96\sqrt{\hat{V}_G[\hat{S}_{KM}(t)]}$$

$$\hat{S}_{KM}(t) - 1.96\sqrt{\hat{V}_G[\hat{S}_{KM}(t)]} \leq S(t) \leq \hat{S}_{KM}(t) + 1.96\sqrt{\hat{V}_G[\hat{S}_{KM}(t)]}$$

which is based on assuming that the KM estimate is, for each t, approximately normally distributed in large samples (Breslow and Crowley, 1974).

However, this estimate may lie outside the range [0,1] and several transformations have been suggested to overcome this problem (Borgan and Liestøl 1990) including the log-log and the arcsine-square root.

Based on assuming $\log(-\log(\hat{S}(t)))$ being approximately Normal, the log-log transform provides the following approximate 95% Confidence Interval for S(t),

$$1-(1-\hat{S}_{KM}(t))^{\exp^{-1.96\sqrt{\hat{V}[\log(-\log[\hat{S}_{KM}(t)])]}}} \leq S(t) \leq 1-(1-\hat{S}_{KM}(t))^{\exp^{1.96\sqrt{\hat{V}[\log(-\log[\hat{S}_{KM}(t)]}}}$$

where

$$\hat{V}\left[\log(-\log[\hat{S}_{KM}(t)])\right] = \frac{\sum_{j=1}^{i} e_j / (r_j(r_j - r_j))}{\left[\sum_{j=1}^{i} \log(1 - e_j / r_j)\right]^2} \qquad \text{for} \qquad t_{(i)} \le t < t_{(i+1)}.$$

1.5 Estimating the Cumulative Hazard Function

One estimator of the Cumulative Hazard Function, H(t), proposed by Nelson (1972) and Aalen (1978) is defined as

$$\hat{H}_{NA}(t) = \sum_{j=1}^{i} \frac{e_j}{r_j} \qquad \text{for} \qquad t_{(i)} \le t < t_{(i+1)}.$$

which is a step function that starts at zero and has a step of e_i/r_i at each event time point with variance estimator

$$\hat{V}[\hat{H}_{NA}(t)] = \sum_{j=1}^{i} \frac{e_j}{r_j^2} \qquad \text{for} \qquad t_{(i)} \le t < t_{(i+1)}.$$

This estimate of the cumulative hazard function is used both to provide a crude estimator of the hazard rate h(t) and an estimator (Fleming-Harrington 1991) of the survivor function as

$$\hat{S}_{FH}(t) = e^{-\hat{H}_{NA}(t)}.$$

An estimator of the variance (and resulting confidence interval) of $\hat{S}_{FH}(t)$ can be obtained by substituting $\hat{S}_{FH}(t)$ for $\hat{S}_{KM}(t)$ in Greenwood's formula above.

1.6 The Counting Process Approach to Survival Analysis

An alternative way to motivate Survival Analysis from that used above is to recast the problem in the Counting Process paradigm (Aalen 1975). This method has been the source of many new developments, particularly in terms of proving that the Kaplan-Meier estimator (and several functions of it) are asymptotically normal (Anderson, Borgan, Gill and Kielding 1993).

The main difference between the 'traditional' (as adopted in this thesis) and the Counting Theory approach is that the latter approach records, at each time point, whether the event of interest has happened or not. The three functions central to the Counting Process approach are as follows:

1. The counting process $N_i(t) = I(T_i \leq t, \delta_i=1)$, where I is an indicator function i.e. the process jumps from a 0 to a 1 once an event occurs for subject i . The process $N(t) = \sum_{i=1}^{n} N_i(t)$ is also a counting process and simply counts the number of deaths in the sample at, or prior, to time t.

2. The "at risk" process $Y_i(t) = I(T_i \geq t)$ and indicates whether subject i is still at risk of the event of interest at time t and can be used to provide information on the number of individuals still at risk at time t.

and

3. The intensity process $\lambda(t) = Y(t)\, h(t)$, where $Y(t) = \sum_{i=1}^{n} Y_i(t)$ and h(t) is the hazard function, and can be considered as providing information regarding the "expected" number of events at time t. The total expected number of deaths can be estimated by integrating $\lambda(t)$ over time and is defined as $\Lambda(t)$, the cumulative intensity process.

The counting process therefore provides information on the total number of events while the cumulative intensity process provides information on the expected number of events up to time t. A natural "residual" now emerges, the counting process martingale residual M(t) and is defined as $M(t) = N(t) - \Lambda(t)$. From the definition above, N(t) is a non-decreasing step-function while $\Lambda(t)$ is a smooth process and hence the martingale can be considered as "mean zero noise" (Klein 1997). This property of martingales can be used to check assumptions underlying regression models for survival data and will be returned to in chapter 5.

Regardless of whether the 'traditional' or Counting Process approach is used, one of the main assumptions when analysing survival data is that the survival times for each individual are mutually independent. In many cases this assumption will be justified

but should this assumption not hold many of the standard methods for the analysis of survival data may not be applicable. This assumption may be violated if there is some natural pairing or constructed matching for subjects in the study. A discussion of such situations is given in the next chapter and the remainder of this thesis deals specifically with analysing so called 'dependent' survival data.

1.7 Chapter Summary

Survival Analysis involves studying the time taken until a particular event occurs. It is distinguished from other fields of statistics by the presence of censoring, which is a particular form of incomplete data. A review of the common censoring mechanisms and some of the standard techniques of analysing survival data were given with particular emphasis on non-parametric estimators of the survivor and hazard function.

An important assumption when analysing survival data is that the observations are independent of one another and the next chapter deals specifically with situations where this independence assumption is brought into question.

Chapter 2

Introduction to Matched/Paired Survival Studies

2.1 Introduction

As outlined in the previous chapter, one of the main assumptions when analysing survival data is that the survival times for each individual are mutually independent. This assumption is valid in a variety of studies such as a randomised trial comparing the efficacy of a drug where the survival experiences of two or more independent groups of individuals are compared. Each individual has his/her own "tolerance" that is not influenced by that of any other individual in the trial.

There are however, situations where there is a dependency structure, or degree of similarity, present among some individuals in the study, on occasions by design or sometimes by natural consequences.

2.2 Dependent Survival Studies

Such studies can be broadly categorised into two main categories, namely Multiple Event Studies and Clustered Survival studies, both of which are now introduced.

2.3 Multiple Event Survival Studies

Multiple events arise when episodes of the same failure/disease process act serially on the same individual. Examples of multiple event survival studies include the time to exhaustion in repeated exercise testing, or time to successive asthma attacks in the same individual. The dependency structure often arises from recording several observations on the same individual.

The remainder of this thesis however deals primarily with analysing survival data arising from Clustered Studies. A general introduction to Clustered Studies is now given.

2.4 Clustered Survival Studies

A Clustered Survival Study can be thought of as a situation where a disease mechanism or failure process is acting concurrently on all the individuals in a "cluster". The cluster may have any number of members all of whom are mutually associated but each cluster is assumed independent from all other clusters. Hence there is assumed a dependency structure within the cluster but an independent structure between clusters. Examples of clustered survival studies include situations where the time to possible hereditary/genetic disease onset in a number of members of a family (or litter) are recorded.

The criteria for cluster membership may be a natural consequence (e.g. family membership) and as such may not be measurable. It is assumed that individuals are

similar in that they share some common genetic or environmental characteristics but this degree of association may not be directly measurable.

Clusters however may also be produced by design where individuals are ***matched*** on the basis of certain characteristics to make them as similar as possible in terms of the failure/disease process. The consequence of incorporating this dependency in the analysis is to usually have the effect of increasing the sensitivity of any appropriate statistical tests. In clinical trials, for example, the control of confounding variables through matching can also serve to improve the precision of the comparison of survival distributions across treatments.

In theory, a cluster may have any number of members, with possibly a different number of members in different clusters. This thesis is primarily concerned with clusters involving two observation times, either for two distinct individuals in the cluster providing one observation each or where one unit/individual has two distinct 'survival' observations associated with it. In order to distinguish between these two types of Cluster Studies, the respective terms Matched Survival and Paired Survival Studies are used and both are now defined in more detail.

2.4.1 Matched Survival Studies

A matched survival study is a cluster study where individuals have been matched on certain characteristics to make them as 'similar' as possible in terms of possible survival. The most common framework for a matched survival study is the matched case-control survival study.

A case-control, retrospective or cross-sectional study often concerns the comparison of two distinct groups in terms of a specific characteristic of interest. Many such studies are retrospective involving cases of some relatively rare disease and matched controls. Other such studies may involve matching two 'similar' individuals in advance and then randomising the two individuals to the separate treatments to be compared in a prospective trial.

An important additional issue in retrospective/case-control studies is to consider whether any variables thought likely to influence survival but are unmatched within the 'pair' are indeed significant risk factors for the disease and consequently aid in identifying high risk subgroups of the population.

A matched case-control study therefore has the feature that each case is matched with one or more controls. Variables considered for the matching (i.e. the ***matching variables***) are significant risk factors themselves, an example would be a matched case-control cancer study using age and sex as matching variables.

Note, the "case" and "control" terminology here is somewhat arbitrary in that a control may not represent a 'control' in the true sense of the word in that they may not represent a baseline performance or placebo effect but may instead represent their own treatment group. In most instances (and for the remainder of this thesis) the terms 'case' and 'control' represent the two 'arms' of a ***primary variable*** under study e.g. Treatment A versus Treatment B in a study comparing two treatments.

In summary, matched case-control survival studies usually involve the comparison of the time to event data for two groups of individuals (the cases and controls) which have been matched on certain characteristics i.e. the matching variables. These matching variables are measurable and are available as part of the study design. This will become an important issue in the analysis of such data and will be addressed in more detail later.

2.4.2 Paired Survival Studies

Paired survival studies involve the comparison of time to event data where the study is based on the comparison of the time to the same failure process on two different sites of the same individual or the same process on two distinct family members. For example, the study could involve comparing the time to failure in each of the two kidneys of an individual at risk for renal failure or the time to heart attack for two brothers with family history of cardiac problems.

In paired survival studies usually no matching variables are observed as, in effect, the observations are 'naturally' matched. There may however be some 'unit' covariates recorded for each individual, for example the individual's age or sex or brothers' father's age at time of heart attack.

2.5 Matched/Paired Survival Data

The previous section introduced both matched and paired survival studies in general terms. This section introduces some notation and definitions of the main components involved in both of these study designs.

Both matched and paired survival studies will have pairs of observation times recorded. Each pair of observation times represent the two 'arms' (e.g. treatments) of the primary variable of interest and for simplicity will be referred to from here on as case and control regardless of the study design or actual context.

Matched survival studies will have available by definition the variables used for the matching (i.e. the matching variables), and may have some additional (unmatched) covariates, (i.e. potential prognostic factors) recorded for each pair of observations.

Paired studies, on the other hand, by definition will not have any matching variables available but may have unit (i.e. pair) covariates recorded for each individual of each pair.

2.5.1 Definition of Basic Notation

Let $(t_{ip}, \delta_{ip,}, z_{ip})$ be the observation time, censoring indicator and covariate vector (i.e. primary variable, matching variables and unmatched covariates) respectively for the ith individual of the pth pair (where i=1,2, p=1, .., P) where P is the number of

pairs. Let C be the total number of matching variables and unmatched covariates plus 1 (i.e. the primary variable). Figure 2.1 below depicts a typical matched case-control survival data set with a case-control type identifier variable (Type), two matching variables (Site and Tumour Thickness) and one unmatched covariate (Ulceration Status).

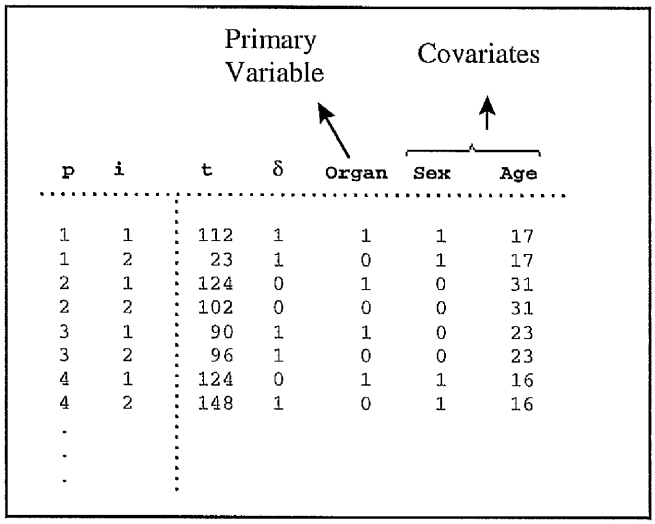
Figure 2.1 Example of a Matched Survival Study Data Set

				Primary Variable	Matching Variables		Unmatched Covariate
p	i	t	δ	Type	Site	Tumour Thickness	Ulcer
1	1	116	0	1	0	1.70	1
1	2	75	1	0	0	1.80	0
2	1	124	0	1	1	2.95	0
2	2	102	0	0	1	2.95	1
3	1	20	1	1	0	3.30	1
3	2	6	1	0	0	4.00	0
4	1	114	0	1	1	5.45	0
4	2	86	0	0	1	6.00	1
.							
.							
.							

Note that all the pairs presented in Figure 2.1 are perfectly matched by the binary variable (Site) while matching for the continuous variable (Tumour Thickness) is taken to some predetermined degree of similarity e.g. to the nearest 10mm in this case. The variable representing the individual in a pair (i.e. "i") is effectively the primary variable but both are presented in this form for completeness.

An example of a typical paired survival data set is given in Figure 2.2 with a primary variable (Organ, e.g. right or left eye for example) and two covariates (Sex and Age).

Figure 2.2 Example of a Paired Survival Study Data Set



Note that for the pairs presented in Figure 2.2 there are no matching variables as each pair represents one individual. Once again the variable representing the individual in a pair (i.e. "i") is effectively the primary variable but both are again included for completeness.

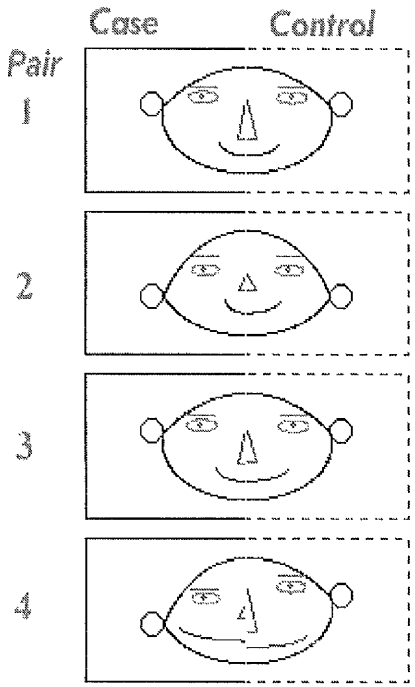
2.5.2 Assessing the Quality of the Matching

Matching variables can be of two types, namely continuous variables (e.g. age) or categorical variables (e.g. sex). The intention in all matched studies is to achieve perfect matching between each case and it's corresponding control but this may not be attainable especially where continuous matching variables are concerned. In this instance matching is usually to a certain predefined interval (e.g. the nearest decade, the nearest 5mm).

In order to visually assess the quality of all the matching variables simultaneously a multivariate icon plot, such as a Chernoff Faces plot, could be used. In the standard version of this plot a separate "face" icon could be drawn for each case and control; relative values of the selected variables for each case and control are assigned to shapes and sizes of individual facial features (e.g., length of nose, angle of eyebrows, width of face). In the present context each matching covariate is represented by a facial attribute and pairs with a good degree of matching will look 'similar' and resemble 'twins'.

However, most matched studies rarely use more than a maximum of four matching variables and a multivariate Chernoff Faces plot may therefore gave an overly optimistic picture of the true quality of the matching. An alternative approach therefore is to display a single face per pair with the left and right sides of the face determined by the separate members of the pair. An example of such a "Chernoff Split-face plot" is given in Figure 2.3 where the degree of matching is good for all but the last pair.

Figure 2.3 Chernoff Split-Face Plot to Assess Matching



Ear level = Binary 1st MC

Eyebrows Height= Binary 2nd MC

Nose Length= Continuous 1st MC

Mouth Length = Continuous 2nd MC

Formally the quality of matching can be assessed by hypothesis tests and confidence intervals based on procedures such as the binomial version of McNemar's test for binary matching variables and paired t-tests for continuous matching variables. These simple ideas will be illustrated when introducing the data sets used throughout the remaining chapters.

2.6 The Illustrative Data Sets

There are two major data sets used in this thesis. The first is an example of a ***matched survival*** study for comparing the survival prospects of melanoma sufferers while the second is an example of a ***paired survival*** study which aims to compare two cement types for bonding orthodontic brackets to teeth.

For simplicity of presentation the melanoma data set will be referred to as the "Melanoma Data" while the orthodontic data set will be referred to as the "Dental Data". These data sets are now introduced.

2.6.1 Melanoma Data

The data used in this study has been supplied by the Scottish Melanoma Group (SMG) Database. The SMG maintains a well-validated database that records clinical and pathological details of all invasive cutaneous melanomas diagnosed in Scotland since 1979. The aim of this particular study was to compare survival prognosis of ***Multiple*** and ***Single*** Melanoma sufferers. An example of a Single Malignant Melanoma is shown in Figure 2.4 overleaf.

Note in this instance a 'control' is classified by the presence of a Single Melanoma and thus should not be confused with the common disease free scenario definition of a control. A Multiple Melanoma is a distinct and aetiologically different type of tumour from a Single Melanoma, and the study therefore should not be in any way be considered as a multiple event scenario. Also note that survival here is compared between the time from 'appearance' of the first melanoma in both types. Obviously Multiple Melanoma cases are not identified until the second melanoma appears which could give rise to some suggestion that 'Single Melanoma' patients may die before the appearance of a second melanoma but because of different aetiology this should not be so.

The cases and controls therefore represent the two 'arms' of the primary variable called ***Tumour Group***.

Figure 2.4. An Example of a Single Melanoma

A matched case-control study of melanoma-associated mortality was undertaken in which each of 108 Multiple Melanoma sufferers (i.e. the cases) from 1976 to 1996 inclusive were matched against a Single primary Melanoma patient (i.e. the controls) controlling for Age (to the nearest 10 years), Sex, Tumour Thickness (to the nearest 10 mm) and Tumour Site (of the first melanoma).

Unmatched covariates which might be potential prognostic indicators in this instance were Level of Invasion of the tumour into the epidermis (as measured by the Clark Level), and Ulceration Status of the first melanoma. Observation time and event type (i.e. dead due to melanoma or censored) were recorded also.

The main interest for this data set is to determine if there is a significant difference in survival time distributions between the populations of Multiple and "equivalent" Single Melanoma sufferers. Summary statistics of recorded observation times for the two tumour groups are given in Table 2.1 where 'complete' (i.e. an event) refers to

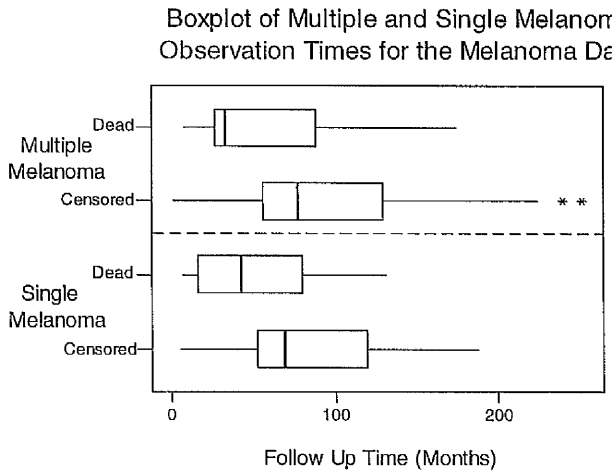
'dead due to melanoma' while censored refers to 'lost to follow up', 'still alive' or dead due to other causes.

Table 2.1 Descriptive statistics for Observation Time by Group for the Melanoma Data

	Sample Sizes	Sample Median Observation Time (months)	Sample Range (months)
Single Melanoma			
Complete	21 (19%)	38	6 - 131
Censored	87 (81%)	69	5 - 188
Multiple Melanoma			
Complete	17 (16%)	32	7 - 174
Censored	91 (84%)	79	0 - 252

There is a similar and high degree of censoring in both the Single and Multiple Melanoma group and the distribution of observation times are similar for both tumour types as displayed in the boxplot below.

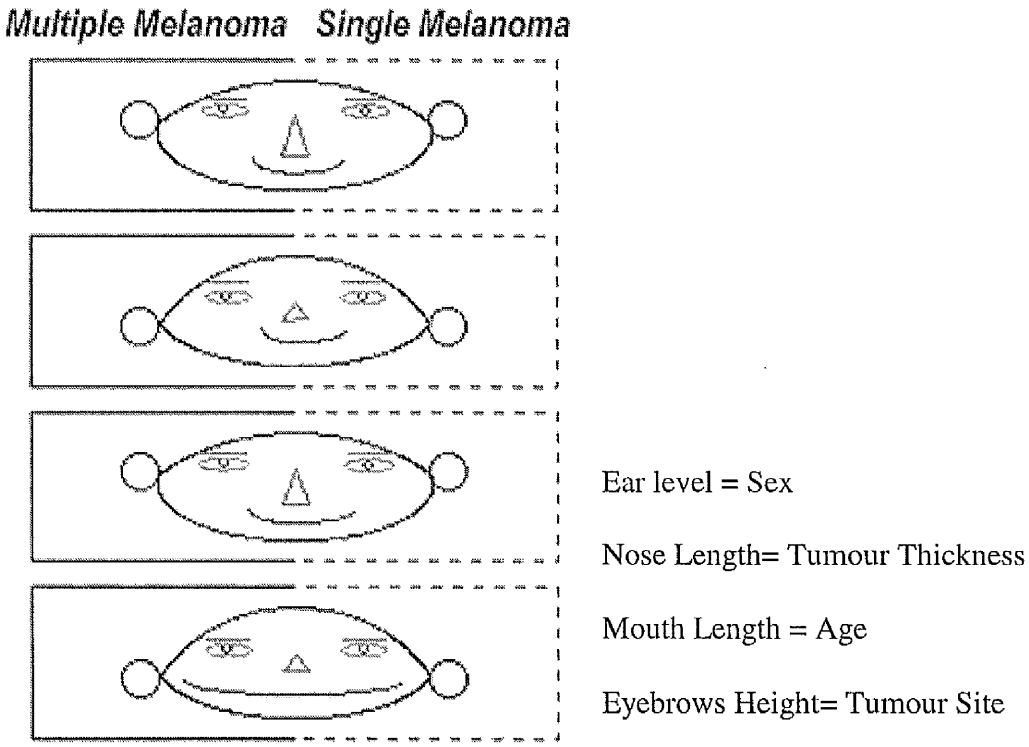
Figure 2.5.



2.7.1.1 Assessing the Quality of Matching

The Multiple/Single Melanoma pairs were matched by Sex, Age, Tumour Thickness and Tumour Site. In general the matching was good as indicated in Figure 2.6 which displays Chernoff faces for a small random selection of Multiple/Single Melanoma pairs.

Figure 2.6 Chernoff Split-Face Plot to Assess Matching for the Melanoma Data



Tables 2.2 and 2.3 show the pairs broken down by Sex and Tumour Site both of which exhibit perfect matching of Single and Multiple Melanoma patients.

Table 2.2 Breakdown of Melanoma Tumour Group Data by Sex for the Melanoma Data

		Single Melanoma	
		Female	Male
Multiple Melanoma	Female	75	0
	Male	0	33

There is a substantially larger number of females than males corresponding to the incidence pattern in the general population of Single Melanomas while the two (composite) Site categories have effectively the same frequency of occurrence.

Table 2.3 Breakdown of Melanoma Tumour Group Data by Site for the Melanoma Data

		Single Melanoma	
		Axial	Extremity
Multiple Melanoma	Axial	53	0
	Extremity	0	55

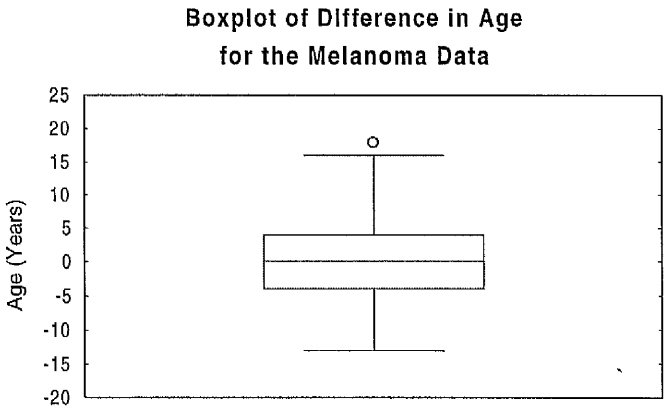
The age distribution is similar for Multiple and Single Melanoma sufferers and similarly Tumour Thickness was matched well for practically all pairs with one or two exceptions (Table 2.4).

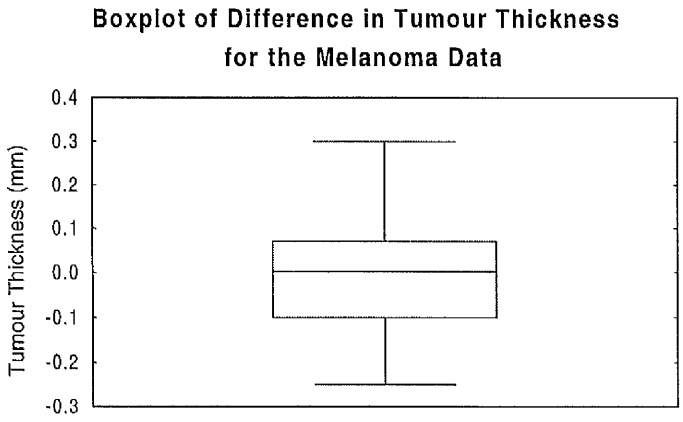
Table 2.4. Mean and 95% Confidence Interval for Age and Tumour Thickness by Tumour group and for Pairwise Difference for the Melanoma Data

Variable	Multiple Melanoma	Single Melanoma	Pairwise Difference (Multiple - Single)
Age (Years)	51.8 (48.5 - 55.1)	52.0 (48.8 - 55.2)	0.05 (-1.0 - 1.1)
Tumour Thickness (mm)	2.2 (1.4 - 3.0)	2.1 (1.4 - 2.8)	0.1 (-0.1 - 0.2)

There was with no significant difference (p=0.72) on average across pairs for Age (Figure 2.6) or Tumour Thickness (p=0.21). Boxplots of the pairwise Age and Tumour Thickness differences are given below (Figures 2.7).

Figure 2.7





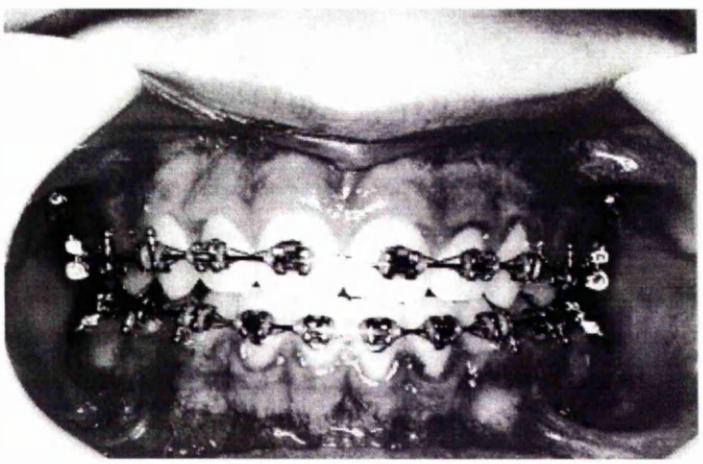
The most extreme pair, in terms of Tumour Thickness, was where the respective Multiple and Single Melanoma Tumour Thickness measurements were 41 and 20 mm

respectively (excluded from the boxplot). Indeed these two tumours were the largest in the sample.

2.6.2 Dental Data

This data relates to a study of bracket bonds in orthodontic practice. Traditionally the fixation of orthodontic brackets to the enamel surface of teeth has been achieved using a chemically-cured cement (Figure 2.8). Despite its universal use this technique has several undesirable consequences including enamel loss and enamel decalcification.

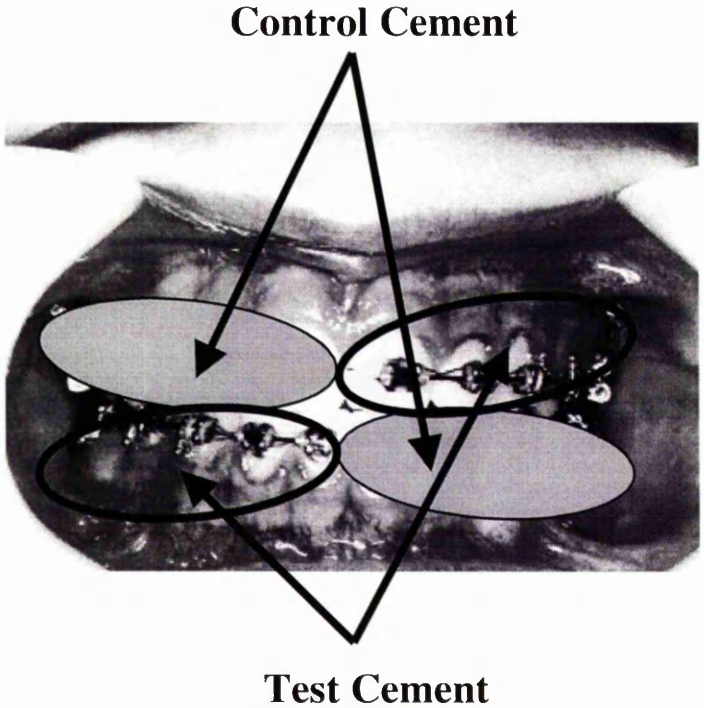
Figure 2.8 An Example of an Orthodontic Bracket



In an attempt to reduce, or even eliminate, these adverse consequences there has been extensive research into developing a replacement adhesive, namely a Glass Ionomer Cement which has been shown not to have the undesirable properties of the chemically-cured cement.

A randomised trial was carried out in order to compare the clinical performance, in terms of the bonding strength, of the two cement types in 41 patients referred to the Department of Orthodontics, Glasgow Dental Hospital between 1995 and 1997. For both arches one of the two bonding cements was randomly assigned contralaterally i.e. upper right and lower left had the same cement and vice versa (Figures 2.9). All brackets used were of the same type and all bonding was carried out by the same operator to ensure a high degree of standardisation.

Figure 2.9 An Example of an Orthodontic Bracket with both Cements Assigned Contralaterally



The time to failure of each bracket was recorded where failure is defined as a bracket having dislodged during treatment. An individual is deemed censored if they were

lost to follow-up during the three year study period or if the bracket had not failed by the end of the study.

The study is an example of a paired survival study with ***Cement Type*** as the primary variable with Test (i.e. Glass Ionomer Cement) and Control (i.e. Chemically-cured cement) as the treatment arms. Each patient's Sex, Age and Malocclusion Type was recorded as potential covariates. An individual's Malocclusion Type gives a description of how the upper and lower teeth come together in terms of whether they protrude inwards or outwards. Malocclusion Type 1 is considered the "best" where the upper and lower teeth come together perfectly.

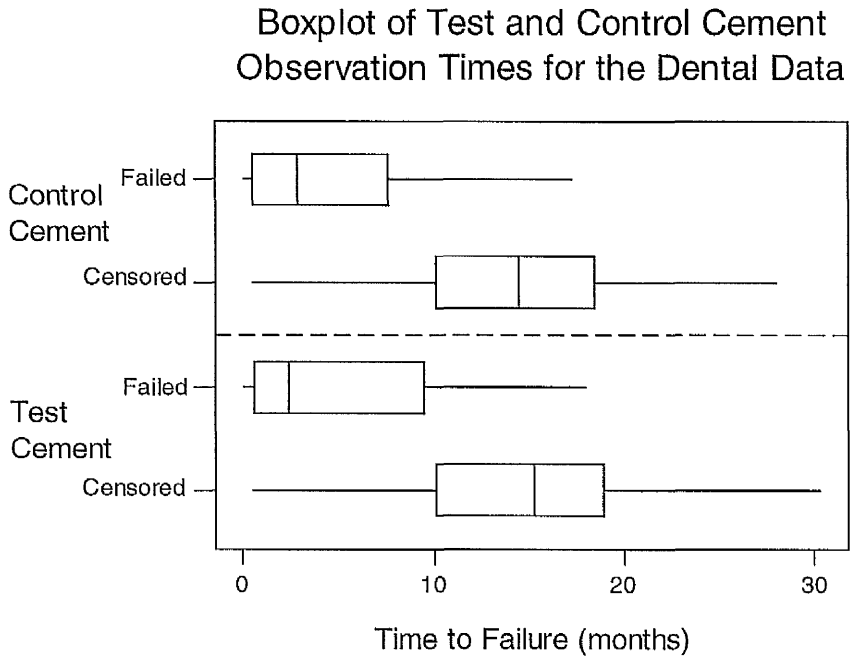
Summary statistics of the failure and censoring times are given in Table 2.5 where it is noted that there is an identically high degree of censoring in both the Control and Test Cement group.

Table 2.5 Descriptive statistics for Observation Time by Cement Type for the Dental Data

	Sample Sizes	Estimated Population Median Observation Time (months)	Sample Range (months)
Control Cement			
Failure	21 (51%)	2.9	0.1 - 17.9
Censored	20 (49%)	14.5	0.5 - 30.3
Test Cement			
Failure	21 (51%)	2.4	0.1 - 17.3
Censored	20 (49%)	15.3	0.5 - 28.1

There is an identical and high degree of censoring in both the Test and Control Cement groups with similar observation time distributions, as displayed in the boxplot below (Figure 2.10).

Figure 2.10



There were 29 females (79%) compared to 12 males (21%) enrolled in the study reflecting the population pattern as orthodontic treatment of this type in general is more frequent for females than males. The sample mean ages were similar for males and females (14.1 and 14.7 respectively), and despite the considerably larger range (Figure 2.11) in age for females there was no significant difference in mean age ($p=0.24$) for males and females.

Figure 2.11

For the individuals in this study the Malocclusion Type distribution for males and females and (Figure 2.12) was quite similar with no significant difference in their distribution for the two cement type groups (χ^2 test, p=0.41).

Table 2.6 Distribution of Malocclusion Type by Sex (with row percentages in brackets) for the Dental Data.

	Malocclusion Type				
	1	2	3	4	Total
Sex					
Male	4 (45%)	1 (17%)	5 (35%)	2 (3%)	12
Female	13 (41%)	5 (15%)	10 (37%)	1 (7%)	29

2.7 Chapter Summary

A general discussion of dependent survival studies was given with particular reference to Cluster survival studies. In general, a Cluster survival study can be one of two distinct types, namely matched and paired survival studies. The main difference between these two types of Cluster survival study is that, in matched studies, the matching is by design while in paired studies there is 'natural' matching between the 'individuals'. Regardless of the study design, both essentially involve a 'case-control' type comparison

Both matched and paired survival studies will provide pairs of observation times and censoring indicators. In addition, matched studies will have by definition some matching variables and possibly some unmatched covariates present. Paired studies on the other hand, may only have 'unit covariates' available for inclusion in any analysis as the individuals within a pair/unit are 'perfectly' matched and thus all matching variables are likely to be 'hidden' from the analysis. Issues relating to the matched case-control study design in terms of assessing the quality of matching were discussed and the example data sets (one matched and one paired) that will be used throughout the thesis were introduced.

Before any formal analysis is undertaken it is essential to graphically investigate all the variables involved in the analysis (in particular the primary variable) in terms of assessing their influence/significance on survival. The next chapter will deal

specifically with graphically displaying survival data, in particular for the matched and paired survival studies introduced in this chapter.

Chapter 3

Methods for Plotting Matched/Paired Survival Data

3.1 Introduction

Before any attempt is made to model data, regardless of the context, it is imperative that clear graphs of the data are provided, not only to allow a subjective impression be made as to the underlying nature and pattern of the data but also to provide clear support for the formal conclusions in reporting the results of a study. The emphasis of this chapter is to suggest methods for graphing the 'survival pattern' across the relevant variables involved in matched and paired survival studies i.e. the primary variable, the matching variables and any additional unmatched covariates recorded.

Chapter 2 introduced notation for the general framework of matched/paired survival data with (t_{ip}, δ_{ip}) used to represent the observation time and censoring indicator respectively for the ith observation in the pth pair where i=1,2, p=1, .., P. For brevity, let in future the first observation in each pair (i=1) be referred to as the 'case' and the second observation (i=2) the 'control'.

This chapter begins with a suggested method for graphing bivariate survival data using a form of scatterplot. Attention then shifts to providing methods for estimating and graphing the bivariate survivor function in order to compare the survival

prospects of the cases and controls. Initially the case and control marginal survivor functions are considered followed by a discussion of various methods proposed for estimating the bivariate survivor function. Interest concentrates on practical applications of these estimators. Following this, methods for graphically assessing the independent effect on survival of each of the matching variables and additional covariates are presented.

3.2 The Bivariate Survival Scatterplot

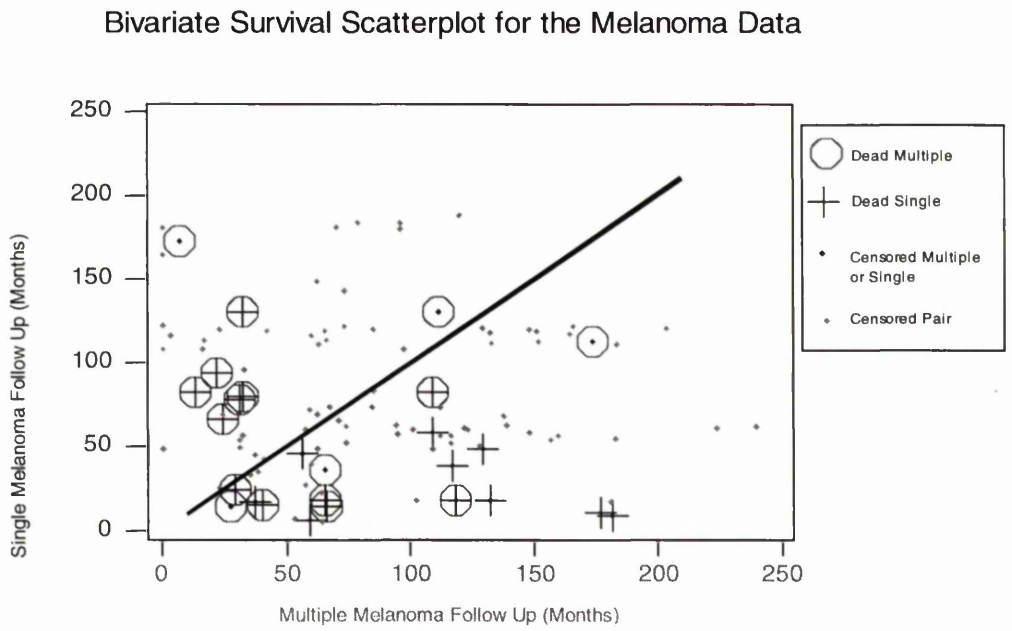
When analysing paired data the first step is usually to examine a scatterplot of the two 'arms' of the primary variable (i.e. the cases and controls). Using this plot, a subjective impression can be made as to whether there is a suggestion of a possible difference in these two variables, using the line of equality as a reference. Assumptions regarding the data (e.g. normality for paired comparisons) may be visually assessed also.

The fact that there is censoring presents a problem for the interpretation of a scatterplot with survival data. That the data are paired or matched provides an additional problem in that a pair can have neither, one or both members censored. If the censoring is ignored the scatterplot represents only pairs with complete information on both arms of the primary variable.

Despite the drawbacks posed by the presence of censoring, a scatterplot can still be a useful and informative tool when analysing paired/matched survival data, if a separate symbol is used to identify the censoring pattern of a pair.

For example, Figure 3.1 below shows the observation times for each case (i.e. Multiple Melanoma) and control (i.e. Single Melanoma) pair labelled by their status. The line of equality is used to investigate informally any suggestion in the data as to whether survival is better in Multiple or Single melanoma (all other factors hopefully being equal).

Figure 3.1



Assuming the case observation time is displayed on the horizontal axis,

- any symbol ***containing*** a + <u>below</u> the line of equality represents a pair where the ***case*** strictly ***outlives*** the control

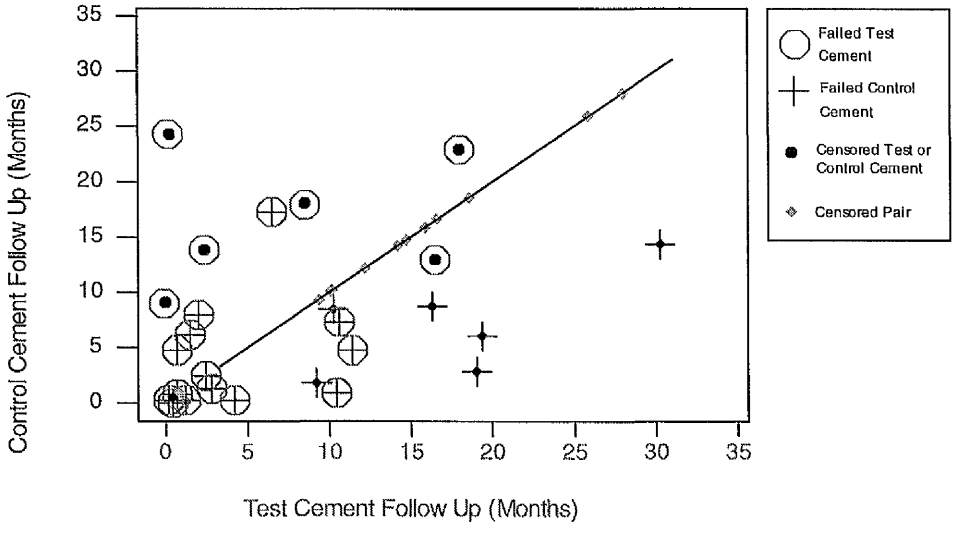
- any symbol ***containing*** a O <u>above</u> the line of equality represents a pair where ***control*** strictly ***outlives*** the case

In this case there is a suggestion of a non-symmetry about the line of equality favouring Multiple Melanoma survival as there is a slight predominance of + below the line compared to the number of O above. Two points are worth noting here, the first is the predominance of doubly censored pairs and the second is the seemingly similar censoring pattern for cases and controls given the symmetry about the line of equality in 'censoring'.

The Bivariate Survival Scatterplot for the Dental Data is given in Figure 3.2 below. There is no suggestion of a clear improvement in bonding strength for the Test cement over the Control cement as there is a similar number of + below and O above the line of equality. There are several doubly censored pairs on the line of equality representing those individuals where neither the brackets bonded with the Test or Control cement failed. The censoring pattern for the Test or Control 'groups' is similar also.

Figure 3.2

Bivariate Survival Scatterplot for the Dental Data



3.3 The Bivariate Survivor Function.

The bivariate survival scatterplot is a useful graphical device to visualise the "collection" of paired observation time points. The graph enables the reader to visually assess whether the assumption of random censoring (for the cases and controls collectively) is valid and it allows the reader to make a subjective impression as to whether there is any suggestion of a difference in the case and control survival prospects.

In survival studies involving the comparisons of independent groups, estimates of the survival functions are usually provided using the Kaplan-Meier estimate. However, in cluster and multiple event studies an estimate of the joint survival distribution is needed due to the dependency structure of the data. When, as in this thesis, matched and paired survival studies are the primary interest, an estimate of the bivariate survival distribution could be useful. Estimating and graphing the bivariate survivor function essentially involves draping a 3D step-function surface plot over the 2D Bivariate Survival Scatterplot where the estimated probability of survival is displayed on the z-axis.

The bivariate survival function is defined as

$$S(t_1,t_2)=S_1(t_1)\, S_2(t_2)\, R(t_1,t_2)$$

where $S_1(t_1)$ and $S_2(t_2)$ are the marginal survival functions for the cases at time t_1 and controls at time t_2 respectively and $R(t_1,t_2)$ can be considered as a measure of

dependence between the cases and controls. The marginal survival functions of the cases and controls, $S_1(t)$ and $S_2(t)$ respectively are naturally estimated using the Kaplan-Meier estimator using the case and control samples separately.

$R(t_1,t_2)$ is considered (van der Laan, 1997) as a cross-ratio of the bivariate survival function over the corners of the rectangle $[0, t_1] \times [0,t_2]$ and is defined as

$$\frac{S(t_1,t_2)S(0,0)}{S_1(t)S_2(t)} \qquad \text{where } S(0,0)=1.$$

It is interpreted in similar fashion to the odds ratio in a 2 x 2 table where if the odds ratio is 1 the interpretation is that the rows and columns are independent while a deviation from 1 indicates positive or negative association. The argument proposed (van der Laan 1997) is that if the mass at each event time point corresponding to $S(t_1, t_2)$, $S(0,0)$, $S(t_1,0)$, $S(0, t_2)$ were observable, then their cross product is a measure of the dependence between the cases and controls.

Before discussing methods for the estimation of the case and control bivariate survivor function consideration is first given to estimating the case and control marginal survivor functions.

3.4 Estimating the Marginal Survivor function

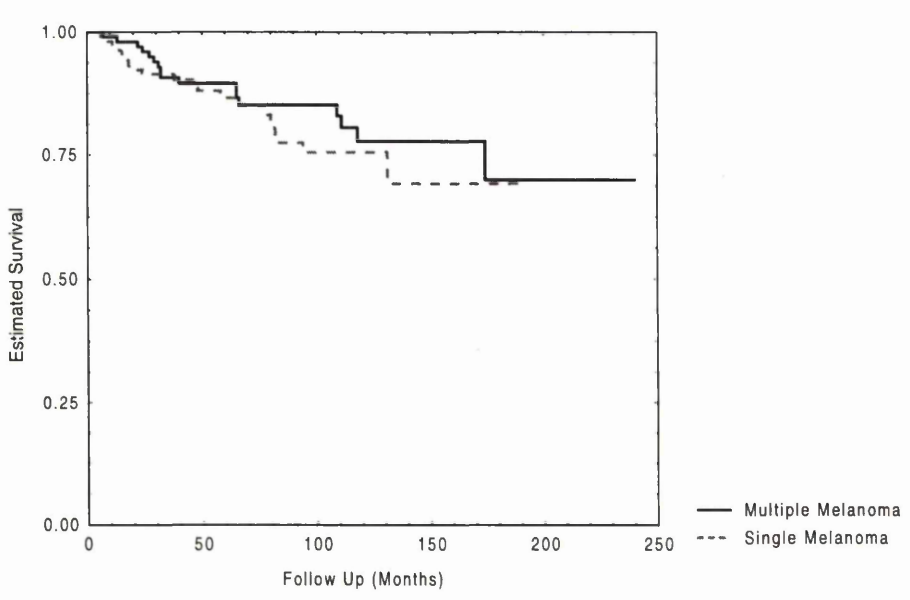
As a first step consider the marginal survivor functions $S_1(t)$ and $S_2(t)$ for the cases and controls respectively. A plot of the estimates of the marginal case and control

population survivor functions, $\hat{S}_1(t)$ and $\hat{S}_2(t)$, can be provided by using the Kaplan-Meier estimates of the respective survivor functions. Note, the Kaplan-Meier approach (c.f. Section 1.4.2) assumes that the observations are independent when calculating the estimated survivor function. In the case of paired data, this assumption is adequate when calculating marginal survival estimates as each pair's first observation time is independent from all other pair's first observation time and similarly for each pair's second observation time.

The estimated marginal survivor functions for the Melanoma and Dental examples are displayed in Figures 3.3 and 3.4 below. From these plots, any large to moderate difference in survival between the groups can be identified by observing whether one plot consistently lies above the other, where it may be thought of as one of the arms of the primary variable having 'better' survival prospects over this period of time.

Figure 3.3

Kaplan-Meier Estimates of the Marginal Survivor Function for the Melanoma Study Primary Variable.

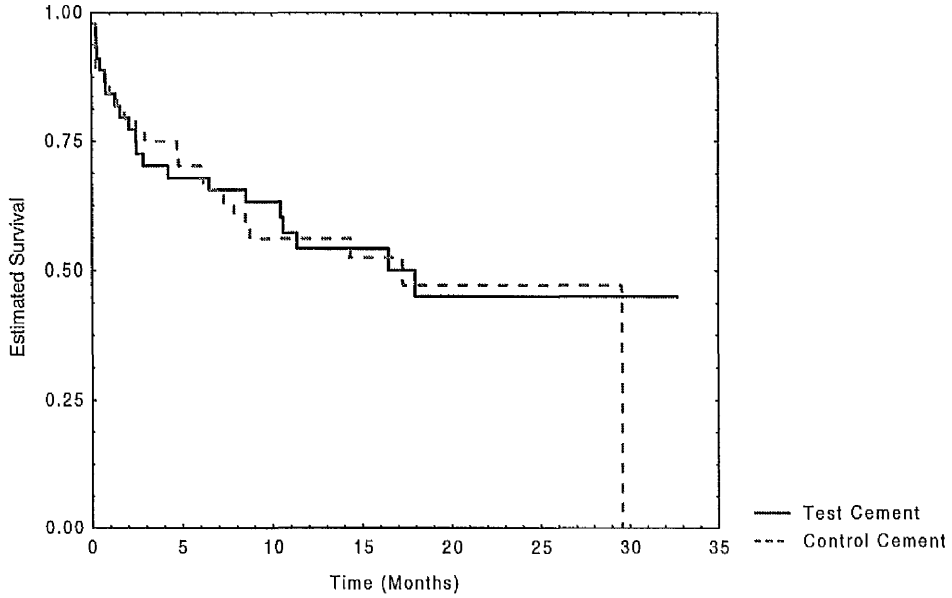


For example, there is a suggestion of slightly improved survival prospects for Multiple Melanoma sufferers compared to Single Melanoma sufferers. The estimated marginal survivor functions for the Multiple Melanoma sufferers descends sharply initially and then tails off gradually to a minimum value of 0.7 at around 16 years. The initial descent is due to the many (early) deaths at the beginning of the study period. The relatively long right tail is a result of many Multiple Melanoma sufferers having long survival times (with respect to melanoma).

The Kaplan-Meier estimated marginal survivor functions for the Test and Control cement types are displayed in Figure 3.4.

Figure 3.4

Kaplan-Meier Estimates of the Marginal Survivor Function for the Dental Study Primary Variable.



From the graph, there is no real suggestion of any (clear) difference between the two cement types where the probability of bracket failure appears similar for both cement types at all time points.

The interpretation of Kaplan-Meier plots is often difficult, i.e. it is unclear exactly where trends lie or hard to identify times where there is a suggestion of a difference in survival prospects due to plots overlapping at certain times. There is the possibility also that *any* observed pattern is a result of sampling variation alone. In most, if not all, applied settings a confidence interval estimate for the estimated survivor function is needed. If confidence intervals for the two 'arms' do not overlap over an 'extensive' time period then there is a strong suggestion of a difference in survival over at least that period.

Methods for calculating pointwise $(1-\alpha)\%$ confidence intervals (at the desired confidence level α) for a single sample were discussed in Chapter 2. The limits of the $(1-\alpha)\%$ pointwise intervals may be joined to form a "confidence band", however the probability that the 'band' contains the true survivor function may be much less than $(1-\alpha)\%$. Simultaneous confidence bands for the estimated survivor function have been proposed however (Hall and Wellner, 1980) but an alternative method proposed here is to produce a confidence band for the ratio $S_1(t)/S_2(t)$ as a function of time.

3.4.1 Marginal Ratio Survival Plots

The estimate of the ratio of the estimated marginal survivor functions can be plotted, along with a unit horizontal reference line. If the ratio lies above this line then, for these times, the survival prospects of the cases appear better than the controls. Conversely, if the ratio lies below the reference line the probability of survival is higher for the controls at that time.

In order to do decide if any trend observed in the ratio plot is due to a real difference, a confidence band for the ratio is needed to indicate the plausible region in which the ratio plot would lie if there was indeed no difference in survival between the cases and controls. Any points on the plot that are outside the reference range would suggest that the pattern is not due to sampling variation and may represent a true effect.

One way of obtaining a suitable reference region would be to estimate the error of the ratio estimate using asymptotic theory. Obviously a first order approximation can be obtained by considering

$$\log(\hat{S}_{KM_1}(t)) - \log(\hat{S}_{KM_2}(t)) \pm 1.96\sqrt{\hat{V}_G[\log(\hat{S}_{KM_1}(t))] + \hat{V}_G[\log(\hat{S}_{KM_2}(t))]}$$

(or some suitable term to 'ensure' simultaneous coverage across all t). Now, if there is 'significant' association between the survival times within a pair, then this interval/band will have an inflated variance term and thus provide intervals for $S_1(t)/S_2(t)$ which are 'too wide' for practical purposes.

A different approach to obtain the reference region is to use the mechanics of ***permutation resampling***.

3.4.2 Reference Regions For Bivariate Survival Data

In this context, the ratio of the marginal survivor functions is estimated from the sample data. The null hypothesis for a paired or matched design is that the both observation times in a pair (across the primary variable) are equally likely for either member of that pair. Based on this hypothesis, suitable 'equally likely' permutations of the observed data can be generated. For each of the permuted data sets, the ratio of the marginal survivor functions can be calculated and the permutation distribution of the ratio constructed. A $100(1-\alpha)\%$ pointwise reference region is then given by $[q(\frac{1}{2}\alpha), q(1-\frac{1}{2}\alpha)]$, i.e. the interval bounded by the $\frac{1}{2}\alpha$ and the $(1-\frac{1}{2}\alpha)^{th}$ quantiles of the permutation distribution. This reference region represents the region where the estimated ratio could fall in if there was indeed no difference in the case and control population survival prospects.

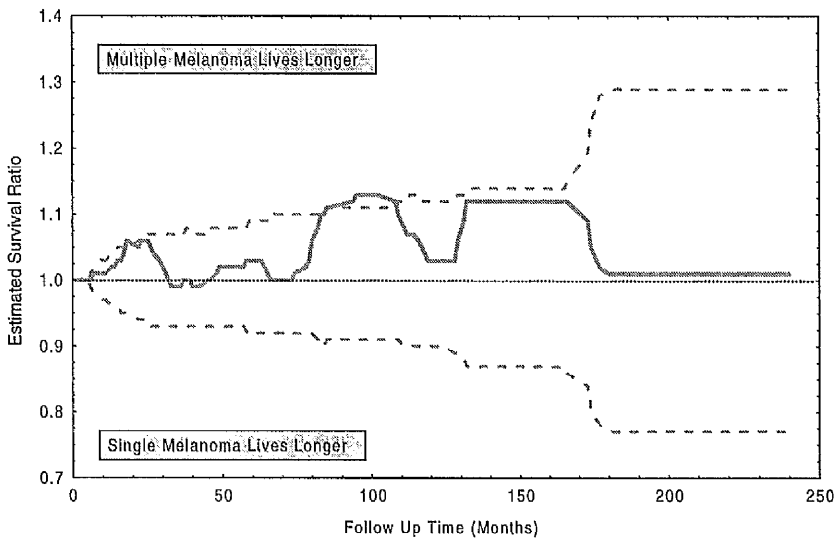
3.4.3 Permutation Envelopes For Ratio Survival Data

In order to calculate an <u>exact</u> reference region all 2^P possible permutations (i.e. for P pairs) must be used. Rather than examining all possible permutations, a substantial reduction in the number of required computations can be achieved by examining a smaller, but representative random sample. This process is termed a "Monte Carlo" simulation. Each 'pass' involves randomly allocating the case and control for the two

observation times in a pair. One way to achieve this is to randomly generate a Bernoulli indicator variable before each pass thus ensuring a random sample of equally likely permutations from all possible permutations is chosen for each pass (for a predetermined number of passes, e.g. 500). For clarity, the term 'permutation envelope' is used to distinguish between an exact and an estimated reference range. For example, an estimated 95% permutation envelope can be calculated by taking upper and lower 2.5% pointwise quantiles of the computed ratios at each observation time point.

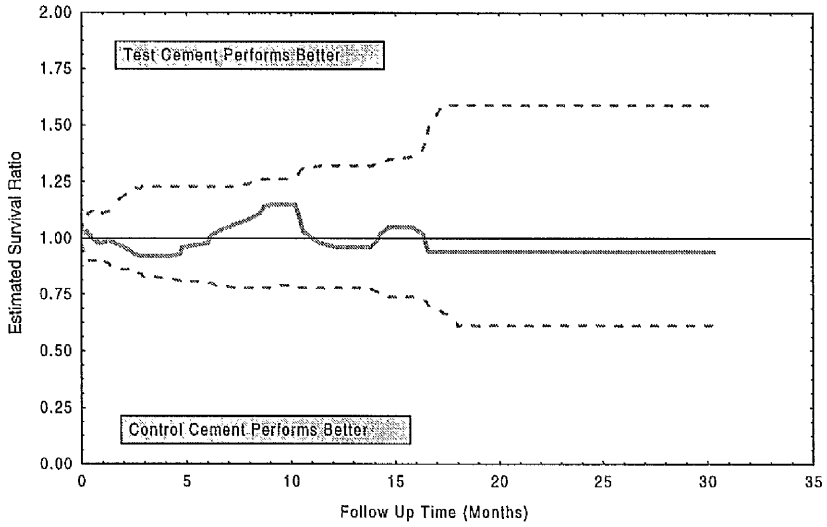
Such permutation envelopes for the ratio of survival for the Melanoma and Dental Examples are given in Figures 3.5 and 3.6 where in each instance 500 simulations were used.

Figure 3.5
Marginal Ratio Survival plots for the Melanoma Study Primary Variable.



There is again a suggestion of improved survival for the Multiple melanoma sufferers particularly for the period between 90 and 110 months.

Figure 3.6
Marginal Ratio Survival plots for the Dental Study Primary Variable.



There is no suggestion of any difference between the performance of the Test and Control cements in terms of the time to bracket failure.

3.5 Estimating The Bivariate Survivor Function.

Various semiparametric estimators (Munoz 1980, Campbell 1982, Langberg and Shaked 1982, Tsai, Leurgans and Crowley 1986, Dabrowska 1988, Pruitt 1991, Prentice and Cai 1992) have been proposed, each of which make little or no assumption about the shape of the function but differ in terms of how the empirical fractions (i.e. the mass provided by each paired observation) are calculated. Recall

that the Kaplan-Meier estimator is a product limit approach where the information provided by the censored observations is 'redistributed' to the right. In the bivariate case this redistribution is more complicated primarily in terms of deciding the contribution of the half censored pairs (i.e. where one member of the pair is censored).

Two review papers are available (Pruitt 1993, van der Laan 1997) which describe and compare the proposed estimators with details of how these estimators are calculated. In both papers the Dabrowska and Pruitt estimator are recommended for general use and, as this chapter is primarily concerned with presenting practical methods for graphing dependent survival studies, only Dabrowska's and Pruitt's estimator will be considered for application. The methods presented however will be applicable to any of the methods referenced above for estimating the bivariate survivor function.

Before considering methods for graphically displaying and interpreting the bivariate survivor function, an intuitive description of both the Dabrowska and Pruitt estimator is now given. Recall that the main 'problem' when estimating a bivariate survivor function involves deciding on what mass contribution each pair type (i.e. non, singly or doubly censored) will provide.

The Dabrowska and Pruitt estimators are similar in terms of the logic proposed for estimating 'mass contribution'. For both estimators, the 'information' from doubly censored pairs is redistributed across an upper right quadrant, a half-censored pair can be considered as contributing information along a horizontal (or vertical) line while a non-censored pair contributes exact information.

Using Dabrowska's approach (1988) the contribution made by pairs where either both members or neither member experience the event follow a similar logic to that proposed in the univariate case while "mass contribution" for the singly and doubly censored pairs is dealt with through a counting process.

Pruitt's approach (1991) on the other hand uses non-parametric smoothing techniques rather than product limit ideas. The estimator initially assigns each contribution mass the value 1/n and the idea is that, by using kernel-density estimators, mass from singly and doubly censored pairs can be redistributed.

A natural approach when comparing paired continuous measurements is to use a scatterplot with a line of equality superimposed. In the case of paired/matched survival studies the bivariate survival function can be used to provide evidence of any departure from symmetry, as assessed by the 'plane' of equality. Improved survival for either the case or control in general will be accompanied by a less steep decline in the surface depicting the bivariate survivor function in one side or other of the 'plane of equality'. The estimated bivariate survivor functions for the Melanoma Data the Dental Data are displayed in Figures 3.7 and 3.8. Note, in both instances both Pruitt's and Dabrowska's estimators gave near identical estimates and hence only one plot is provided for each data set.

Figure 3.7

Surface plot of the Pruitt Estimated Bivariate Survivor Function for the Melanoma Study Primary Variable.

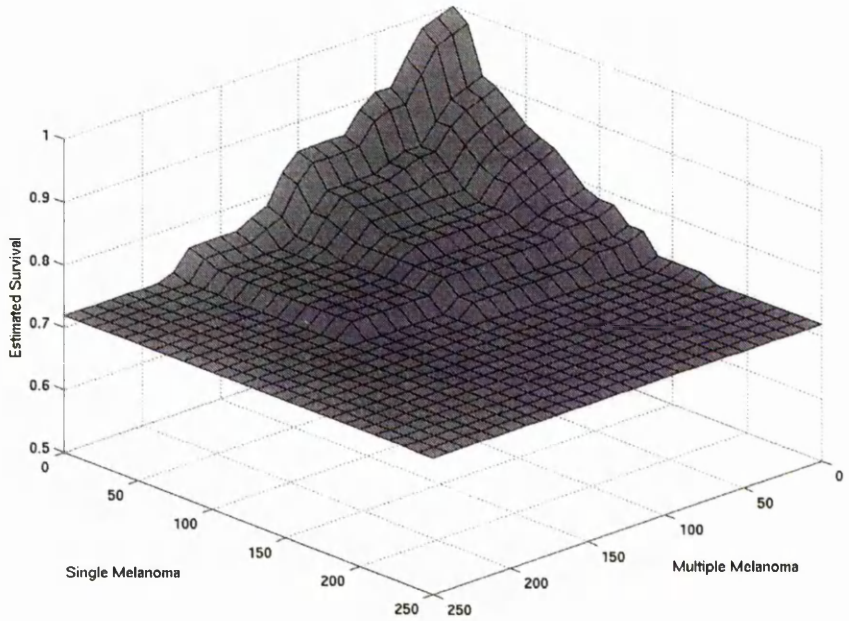
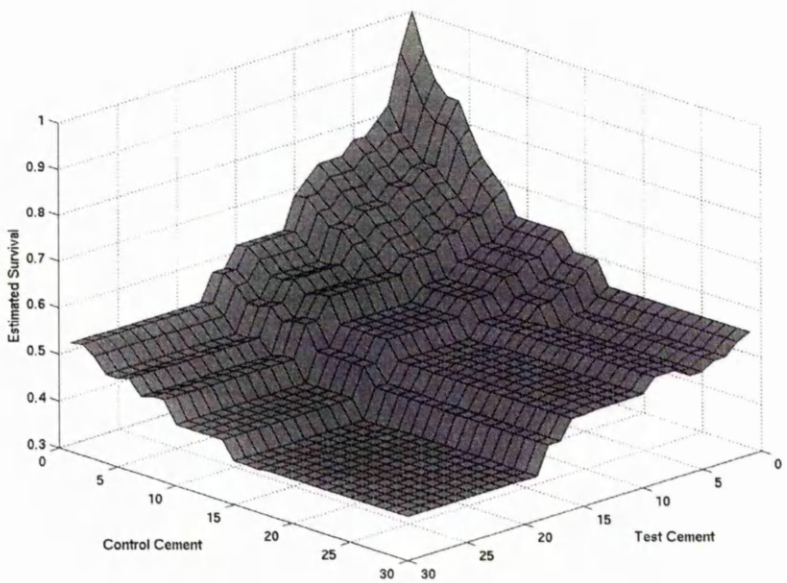


Figure 3.8

Surface plot of the Dabrowska Estimated Bivariate Survivor Function for the Dental Study Primary Variable.



There is some suggestion that the surface has a sharper decline for the Single Melanoma sufferers when compared to the Multiple Melanoma sufferers while there is no suggestion of a 'lack of symmetry' about the (imagined) plane of equality evident in the Dental Data. These suggestions are further confirmed by looking at contour plots (using suitable contours) for the two example data sets (Figures 3.9 and 3.10) where there is a distinct 'leaning' in favour of Multiple Melanoma survival while no clear 'leaning' is evident in the Dental data.

Figure 3.9

Contour Plot of the Pruitt Estimated Bivariate Survivor Function for the Melanoma Study Primary Variable.

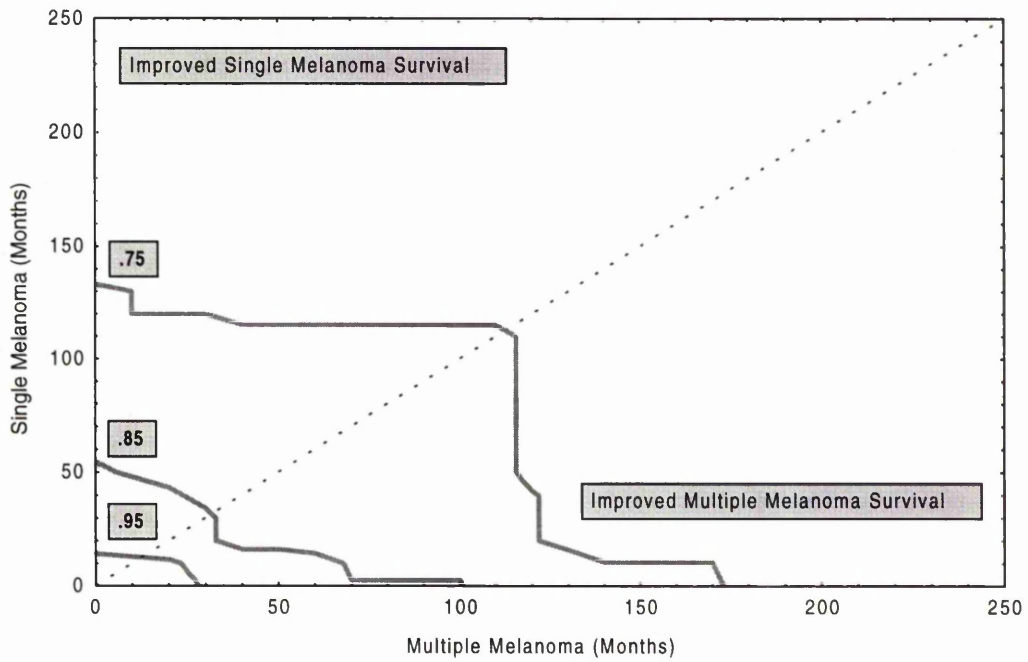
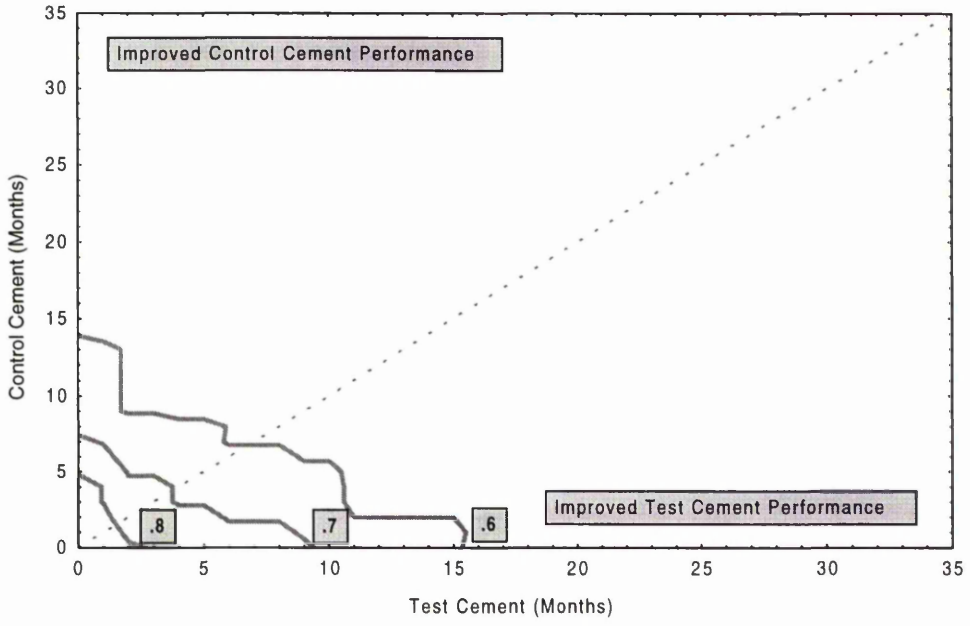


Figure 3.10

Contour plot of the Dabrowska Estimated Bivariate Survivor Function for the Dental Study Primary Variable.



3.5.1 Permutation Envelopes For The Bivariate Survivor Function

In order to determine whether any patterns or differences/asymmetry observed in the estimated bivariate survivor function are real and not due to sampling variation alone, an extension of the permutation test procedure illustrated for the plot of the ratio of survivor functions can be employed. Upper and lower 95% permutation envelopes for the bivariate survivor estimate can be obtained using the same permutation and Monte Carlo methods presented for the plot of the ratio of survivor functions. The problem now involves presenting a method to graphically display the three surfaces simultaneously to determine regions where the estimated bivariate survivor function crosses either the upper or lower reference ranges.

One possible method is to graph the three surfaces simultaneously using a solid colour for the sample estimate and a light colour for the upper and lower reference ranges. Regions where the sample estimate cuts through the references ranges are considered regions where there may be a significant difference in survival (Figures 3.11 and 3.12).

Figure 3.11
Surface plot with Reference ranges for the Melanoma Study Primary Variable.

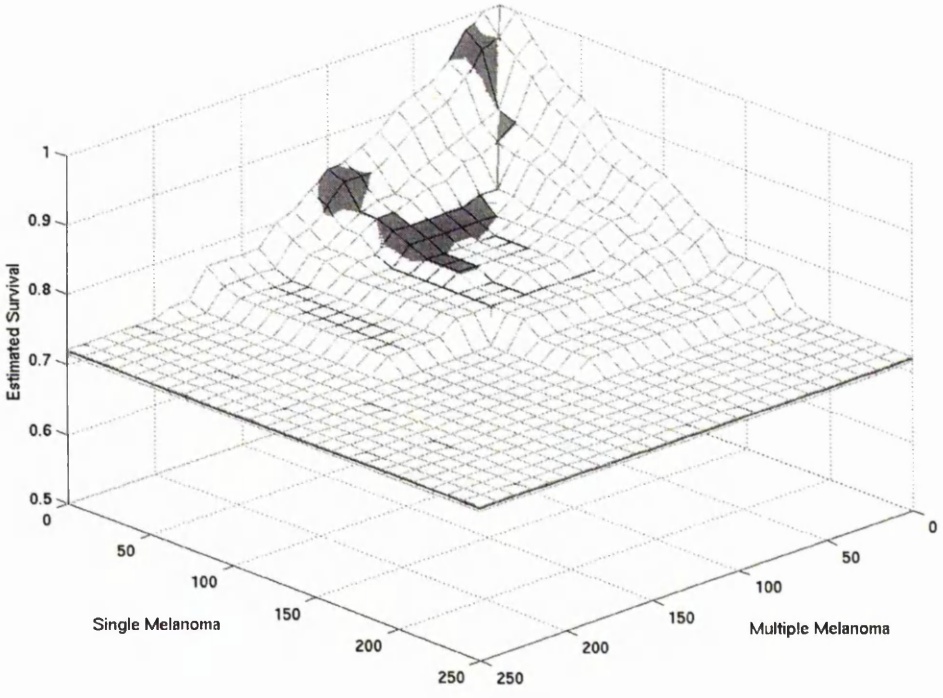
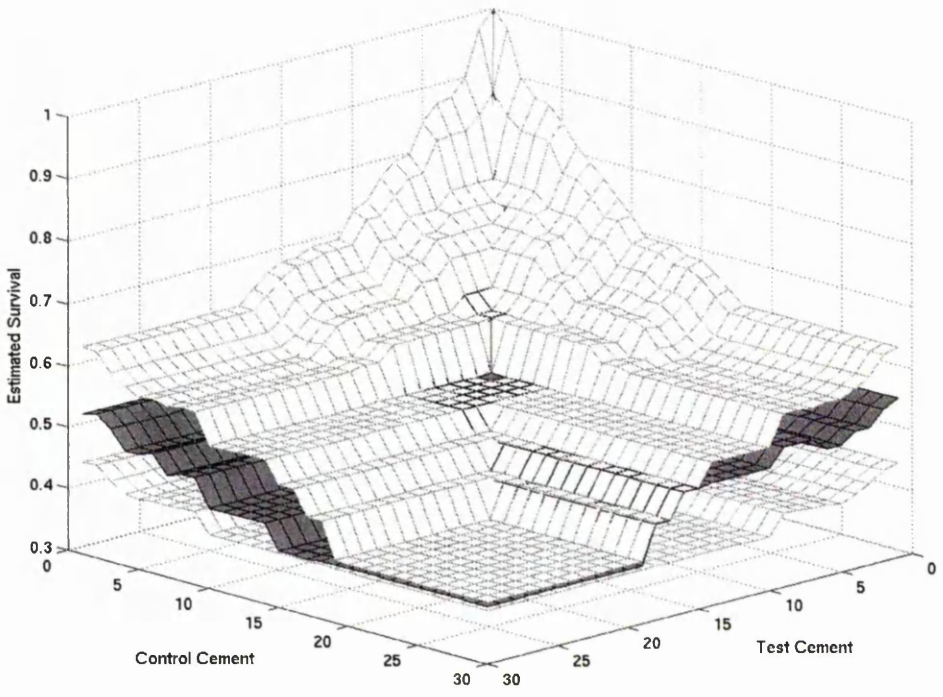


Figure 3.12
Surface plot with Reference ranges
for the Dental Study Primary Variable.



There is again a slight suggestion of a significant difference in the survival prospects of the two Melanoma types in similar epochs/time periods to those suggested in the marginal plot but in general the estimated survivor function is sandwiched within the reference range. Note that the surface is very flat at the edges due to the predominance of censored observations with long follow ups.

One point worth mentioning is that it is quite plausible that the estimated ratio, or surface, could cross the reference ranges (or permutation envelopes) at certain time

intervals when there is actually no difference in survival. However, when the estimated ratio, or surface, are consistently above (or below) the reference ranges (or permutation envelopes) then there is more likely to be a true difference in survival.

For example, despite the fact that the estimated survivor function does appear to cross the upper 95% reference range at a solitary time point, there is no suggestion of any clear difference in the time to failure for either cement type.

An alternative approach (Bowman 1999) to consider is to plot the estimated survivor function alone but to colour code the surface at the times where the estimated survivor function crosses the upper or lower reference range using different colours to depict which reference range (i.e. 90%, 95% and 99%) is crossed.

From the various plots presented in this chapter the following points are evident. There is some slight suggestion of a possible difference in survival favouring Multiple Melanoma compared to their similar Single Melanoma sufferers. However, this difference is evident only over a limited time and therefore may not be strong, or consistent enough to suggest a real significant difference. There is no suggestion either of any difference in the Cement types in terms of their performance of prolonging time to bracket failure.

All of the techniques used so far have effectively assumed that all other factors (i.e. matching variables and additional covariates) were 'equal' for the cases and controls. However, in many analyses (e.g. Analysis of Covariance) it often makes sense to

correct for significant covariates which might distort the findings of an 'unmatched' analysis. Accordingly graphical methods to attempt to allow for this are required.

3.6 Assessing the Individual Effect of Covariates on Survival

As indicated in Chapter 2 the 'recorded' covariates in a matched or paired survival analysis study have different roles. In a matched study, some of the covariates are used to form the matching and are termed matching variables. Earlier in this thesis methods were suggested for assessing the quality of the matching.

Presumably the matching variables have been chosen by virtue of their proven effect on survival. However, it may still be useful to identify the effect the matching variables have on survival in the current study as a way of validating the method used for selecting the controls in terms of how representative they are of the 'control' population in question.

In addition to the matching variables, unmatched covariates are often available which may serve as further useful 'adjusting covariates' when analysing the effect of the primary variable.

Before any modelling approach is considered it is imperative to gain an understanding of the univariate effect of each of the covariates in turn. This section deals primarily with providing methods for graphically assessing the effect of covariates on survival in general, regardless of whether they are matching variables or unmatched covariates.

The first section deals with categorical covariates while the second considers continuous covariates.

3.7 Categorical Covariates

This section is primarily concerned with investigating the effect of categorical covariates on survival. The Kaplan-Meier estimator, as illustrated in Chapter 1, provides an efficient means of estimating the survival function (for right censored data) for each level of the covariate. A graphical comparison of the effect of the covariate on survival can be made by plotting the upper and lower pointwise confidence intervals (e.g. 95%) or confidence bands in addition to the estimated survivor function for each level of the covariate. As previously mentioned, when comparing two or more independent groups of survival data, such a graph can appear quite cluttered, and it is difficult to accurately identify any clear patterns.

One way around this is to modify the argument presented earlier for generating reference ranges for the plot of the estimated ratio of survivor distributions.

3.7.1 Ratio Plots for Independent Survival Data

The marginal ratio survivor plot concerned the estimated ratio of the marginal survivor functions and the reference ranges were generated by virtue of permutation within pairs (in order to remain true to the 'within-subjects' design of such studies). For any binary covariate, the technique introduced in Section 3.4.2 can now be

adapted to the ratio of survivor distributions for the two levels of the covariate (as opposed to case/control).

3.7.2 Reference Range Plots for Binary Covariates in Survival Data

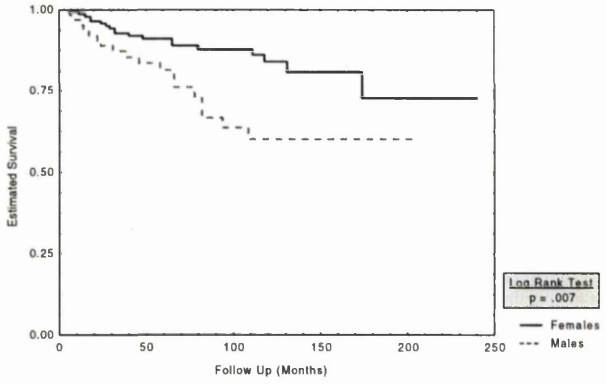
In the case of independent binary covariates, reference ranges can be generated by considering all the possible permutations of the covariate label. If m and n are the number of observations in the two levels of the binary covariate where m+n=P, then there are $\binom{m+n}{m}$ possible permutations of the grouping pattern. There are $\binom{m+n}{m}$ possible data sets therefore which could have been obtained conditional on any subject being equally likely to have arisen from either level of the binary covariate.

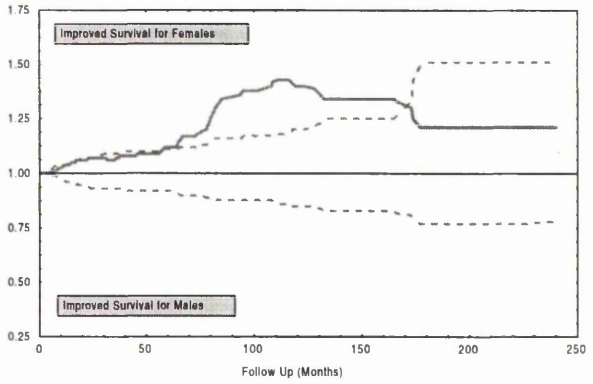
If the number of observations is fairly large this will result in a very large number of possible permutations and a Monte-Carlo simulation is preferable. A selected number (e.g. 500) of simulations is carried out, and the estimated survival ratio is calculated for each permuted data set. Lower and upper 2.5% quantiles can then be computed at each distinct observation time from the simulated set of ratios to generate the permutation envelope as an alternative to the exact reference range. Stratified Kaplan-Meier plots by the binary covariate of the estimated survival functions and accompanying survival ratio plots (with permutation envelopes) for the various categorical covariates recorded in the Melanoma and Dental studies are given overleaf (Figure 3.13) where, in each graph, 500 simulations were used.

Figure 3.13

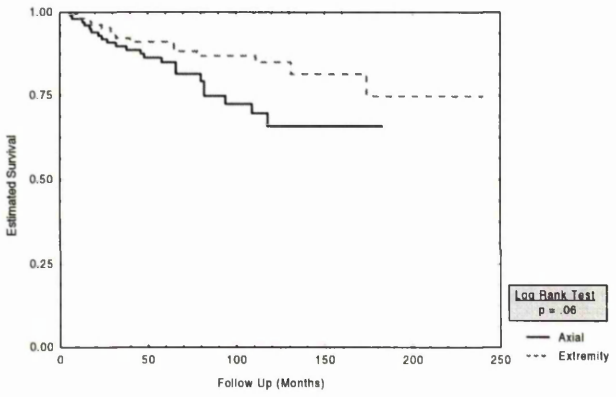
Kaplan-Meier and Ratio Plots For The Melanoma Data

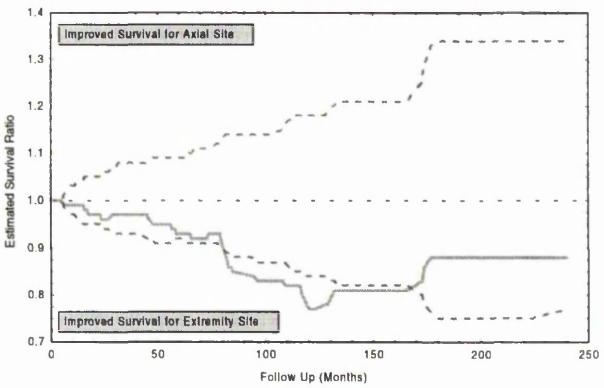
Sex



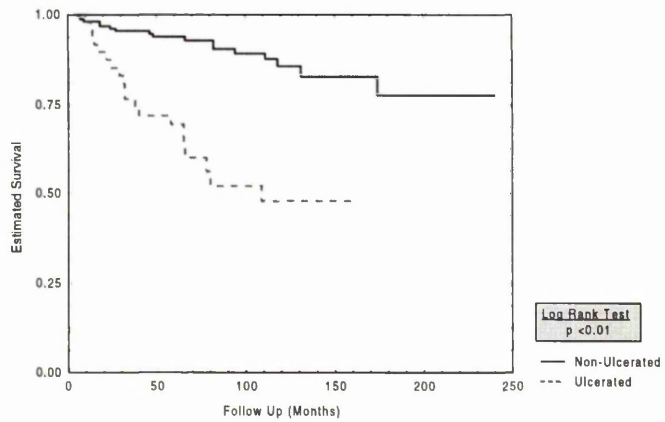


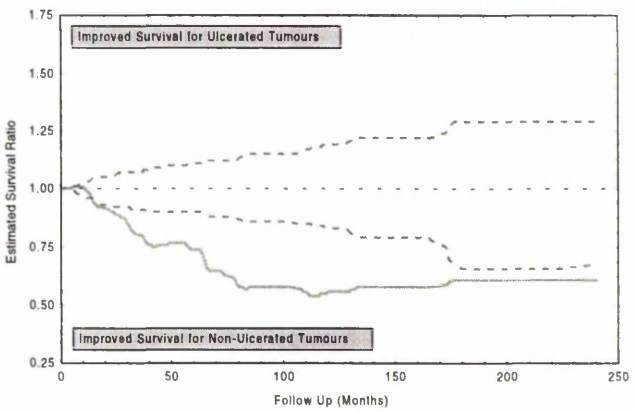
Tumour Site



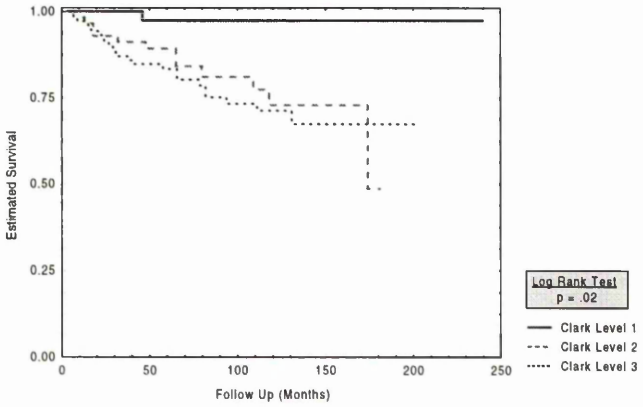


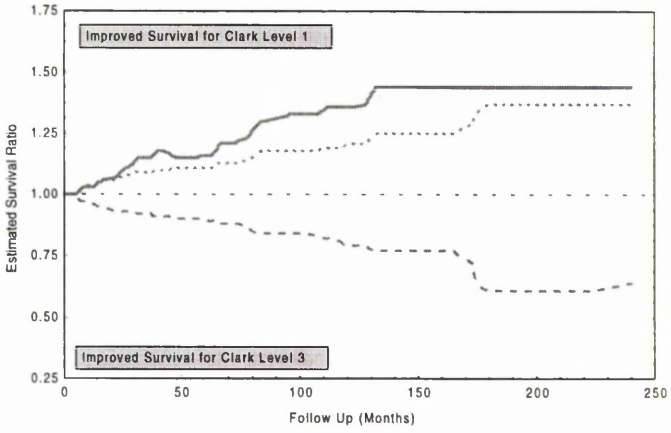
Ulceration Status





Clark Level of Invasion Status





From the plots (and indeed accompanying Log-Rank tests) presented for the Melanoma data it is clear that all the matching variables are significantly associated with survival and therefore the matching was worthwhile.

Poorer survival is associated with males compared to females in general with a suggestion of strongly improved survival for females from 70 to 170 months. Similar survival patterns for both sexes appear from 170 months onwards although this may be a function of the small sample sizes beyond this time. One explanation for this is that in general females are more likely to report possible tumours much earlier than men (Tillman 1998).

Poorer survival is strongly associated with ulcerated tumours with a suggestion of dramatically worse survival prospects for individuals with ulcerated tumours from as early as 10 months onwards.

Individuals with tumours presented on an axial site have poorer survival prospects compared to those with tumours located on an extremity. There is a suggested region (similar to that with sex) corresponding to significantly improved survival for tumours located on an extremity from 80 to 180 months again small sample sizes beyond 180 months may be the cause of this apparent lack of difference after this time.

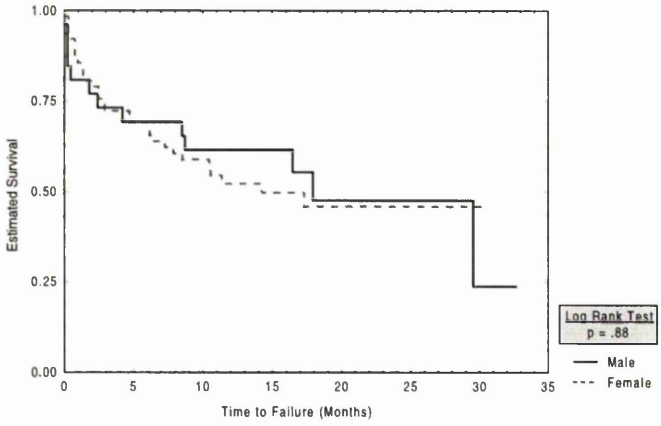
The last two plots investigated the effect of the Clark Level of Invasion covariate on survival. From the stratified Kaplan-Meier plot it is evident that the majority of individuals with Clark Level 1 are censored and poorer survival is evident with increasing Clark Level. For brevity, in order to prepare the ratio plot only the Clark Level 1 and 3 groups were used to highlight the best and worse levels of the covariate. Individuals with Clark Levels of Invasion 2 and 3 exhibit similar survival patterns. Indeed, on the basis of a ratio plot using these two levels (not shown), there was no suggestion of any difference.

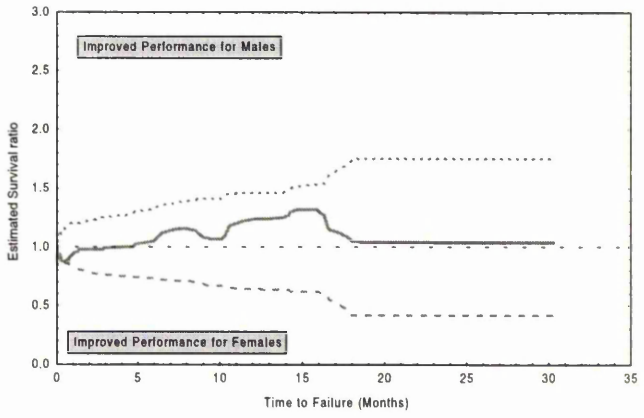
Stratified Kaplan-Meier and ratio plots for the categorical covariates involved in the Dental study are given below (Figure 3.14). There is no suggestion of any sex effect despite males appearing to have slightly longer time to failure in general. Individuals with Malocclusion Type 2 have the poorest performance in terms of time to bracket failure while arguably Malocclusion Type 1 individuals have in general the best performance. A ratio plot was prepared comparing these two levels of the covariate and there was a suggestion of a 'significant' difference from 19 months onwards.

Figure 3.14

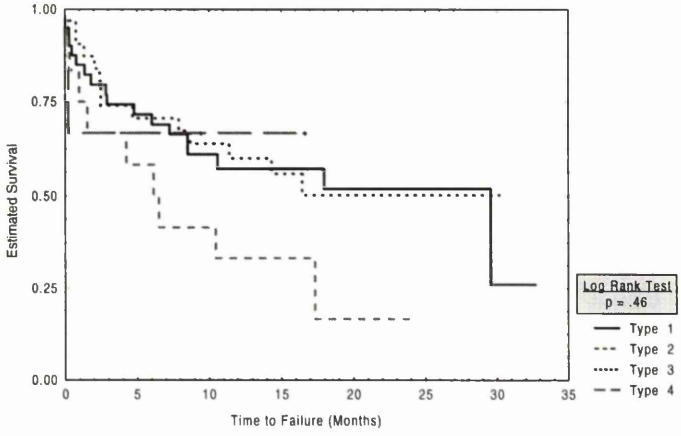
Kaplan-Meier and Ratio Plots For the Dental Data

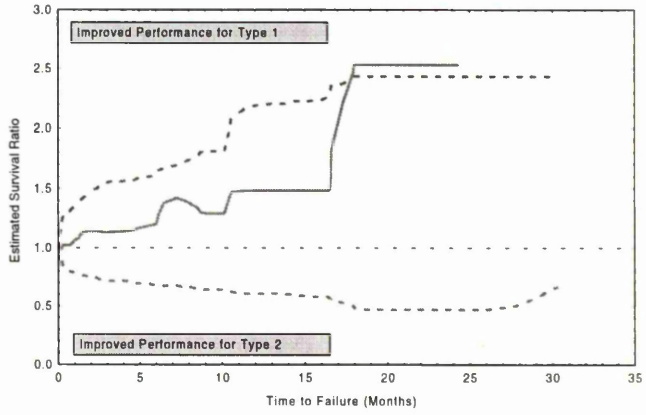
Sex





Malocclusion Type





3.8 Continuous Covariates

In order to assess the effect of a continuous covariate in regression problems, the most natural procedure is to plot the covariate against the response variable. The nature and strength of the relationship between the response variable and the covariate controls the shape of the scatterplot. In survival problems, the shape of the plot often suggests a skew along the time axis due a high proportion of individuals experiencing the event early compared to a smaller number of individuals with large (right) censoring times.

In the absence of censoring the 'true' relationship between the covariate and time is clear, however this is not the case when censoring is present as different information is provided by the complete and censored observations.

The first method presented is a continuation of those presented in the previous section. Simply recode the continuous covariate into a small number of categories chosen to best display the 'true' effect of the covariate. The categorisation process could be based on previous clinical research. For example, previous research in Melanoma (Tillman et al 1991, Aitchison et al 1995) has suggested that identified risk groups for Melanoma survival are, in increasing order of risk, <1.5mm, 1.5-3.5mm and $\geq$ 3.5mm.

If however, such 'classification' information is not available the following tree-based approach could be considered.

3.8.1 Tree-Based Approaches

Tree-based approaches have become increasingly popular in recent years since the publication of CART (Classification and Regression Trees, Breiman et al 1984). In essence, the CART procedure is used to gain a better understanding of the dependence of the response variables on the structure of the relationships of potential explanatory variables (e.g. risk factors) and their combinations, together with their high-order interactions. If the response variable is binary the procedure produces a Classification tree while a Regression tree is produced if the response variable is continuous. The procedure involves successive partitioning of the data set by identifying, at each partition step, which explanatory variable best (and significantly) separates out the data in terms of the response variable on the basis of an appropriate test statistic. This approach allows for 'significant interactions' to be identified in a non-hierarchical manner providing insight and understanding into the structure of the data. A tree-based analysis results in clearly defined steps with easily interpretable splits of individual explanatory variables allowing prediction of the response variable calculated for different 'subgroups' of the explanatory variables. Creation of these "high risk" subgroups, using recursive partitioning, is likely to aid any decision making process.

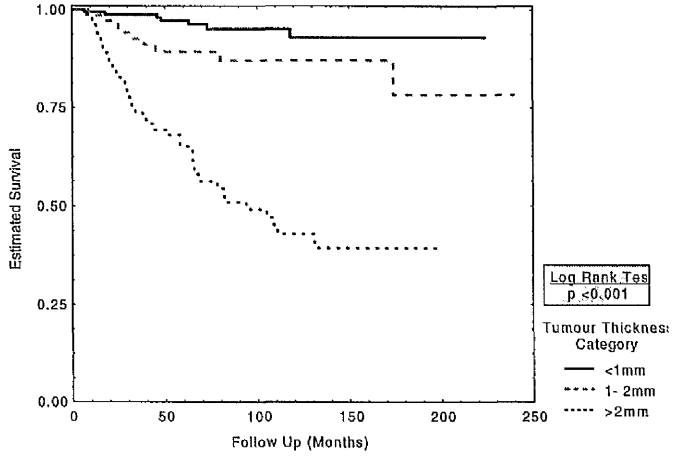
Regression trees for survival data (i.e. continuous data with censoring) have been proposed (Segal 1988) in order to elicit "high risk" subgroups. In regression problems, where there is no censored observations, the two-sample t-statistic is often a natural candidate for use as the splitting criterion. In survival analysis problems however, any member of the Tarone-Ware (1977) class of two-sample statistics (e.g.

the log-rank test) may be considered. The partitioning ceases when no further (useful) splitting of the data can be achieved at any step in the tree, or when the sample size is too small (i.e. ≤ 20 observations), or when the proportion of complete observations is too small.

In general, tree-based techniques are used to identify important prognostic groups but in the application proposed here the tree-based analysis is used to suggest suitable cut-points for recoding the continuous covariate.

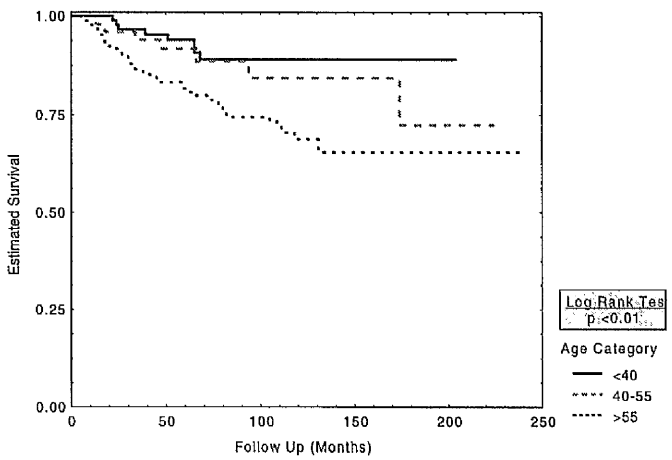
Separate tree-based analysis of the Melanoma data, using the log-rank test as the splitting criterion suggested the following splits for Tumour thickness: ≤1, 1-2 and > 2mm, while the suggested splits for Age were: ≤40, 40-55 and ≥ 55 years. Stratified Kaplan-Meier plots of the effect of Age and Tumour thickness on survival categorised using these cut-points are given in Figure 3.15.

Figure 3.15. Stratified Kaplan-Meier plot of the Effect of Tree-Based Tumour Thickness Risk Groups on Survival for the Melanoma Data.



Note, as indicated in Chapter 2, there were a few pairs with considerably thicker tumours than any other pairs. In order to check whether these pairs had any overly influential effect on determining the categories of tumour thickness used above, the tree-based procedure was repeated ignoring these pairs. There was no suggestion of an alternative re-categorisation than that suggested using all the Tumour thickness data (Figure 3.16)

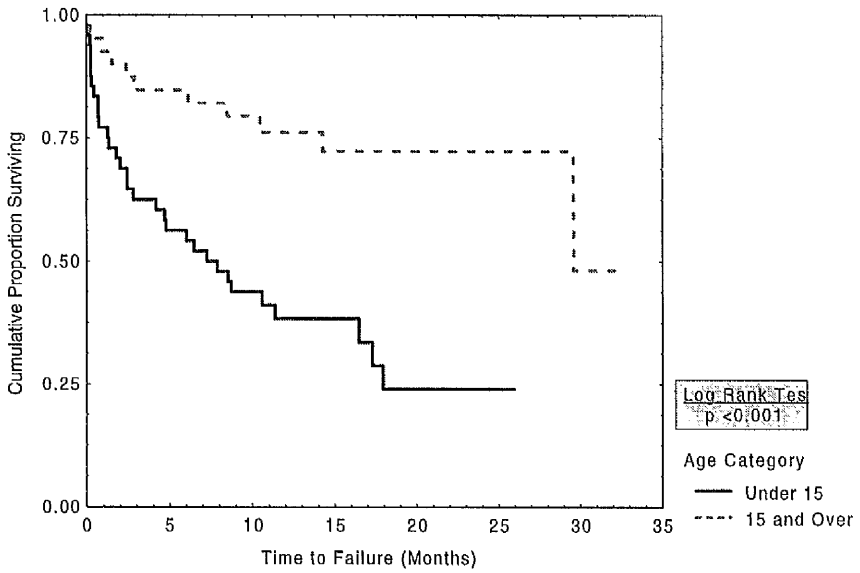
Figure 3.16. Stratified Kaplan-Meier plot of the Effect of Tree-Based Age Risk Groups on Survival for the Melanoma Data.



The only continuous covariate in the Dental study is the subject's age. There is a suggestion that improved performance (i.e. delayed time to breakage) is associated with older subjects (personal communication with the principal experimenter). Older subjects in general, adhere more strictly to the study guidelines in terms of avoiding certain food-types known to adversely effect bonding. A tree-based analysis, again using the log-rank test as splitting criterion, suggested only two age risk groups i.e. under 15 years and older than 15 years of age.

The effect this re-categorisation of Age has on survival is clear from the stratified Kaplan-Meier plot displayed in Figure 3.17.

Figure 3.17. Stratified Kaplan-Meier plot of the Effect of Tree-Based Age Risk Groups on Survival for the Dental Data.



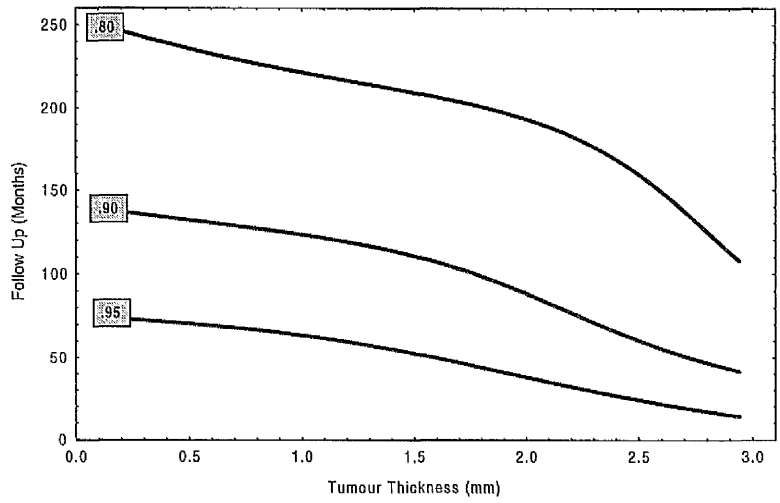
3.8.2 Nonparametric Quantile Regression curves

The method presented in the previous section attempts to present the relationship between the covariate and time on the 'survivor function' scale but it would be useful also to graphically present the relationship, if any, using the original 'observation time' scale.

One approach adopted by several authors (Beran 1981, Bowman and Wright 1998) is to add 'suitable' regression lines to the scatterplot that depict the effect the covariate has on the probability of survival at chosen percentiles. The idea is to 'run' across the covariate and extract bins, or 'windows' of data from which a quantile can be estimated based on the Kaplan-Meier survival curve for the data in that bin. These running quantiles can then be calculated at many covariate values and when plotted should indicate the nature of the relationship between the covariate and observation time.

This procedure is referred to in the literature as nonparametric quantile regression curve estimation. In particular Bowman and Wright (1998) suggest a technique involving a double smoothing process. A weighted Kaplan-Meier estimator is proposed using kernel smoothing using nearest neighbour weighting. The resulting running quantile curve is a step function and a second smoothing process, applying a nonparametric regression procedure to the step positions of the graph, is proposed to yield a smooth quantile curve. As with all smoothing procedures, there is a level of subjectivity as to what is the 'best' value of the smoothing parameter to use. Smoothed nonparametric quantile curves are presented therefore for Tumour Thickness and Age, at suitable quantiles (Figure 3.18) for the Melanoma data example using a value of the smoothing parameter chosen subjectively to avoid 'over' or 'under' smoothing the estimated quantile curve.

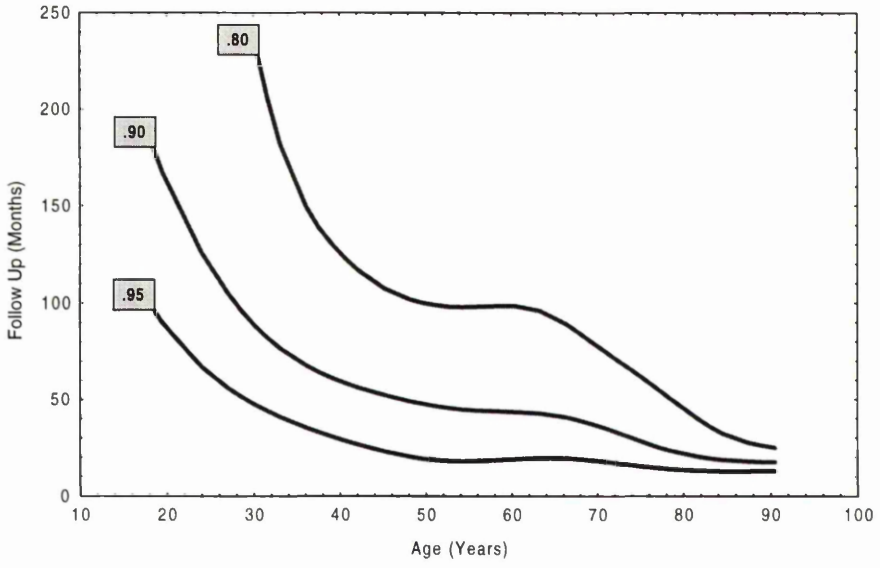
Figure 3.18. Smoothed nonparametric quantile curves for Tumour Thickness for the Melanoma Data.



There is again a suggested effect of Tumour thickness on survival where thicker tumours are associated with poorer survival (for the 0.80, 0.90 and 0.95 quantiles). Once again the influence of two extreme observations were assessed and no undue influence of these was noted. For clarity, the range of tumour thickness is restricted in order to compare the pattern of the smoothed running quantile plots, in terms of suggested turning points, to those identified using the tree based analysis. There is a suggestion also that the 'high risk' tumours thickness categories of <1mm, 1mm - 2mm and ≥ 2mm are plausible (Figure 3.18).

The smoothed nonparametric quantile curves for Age (evaluated at the 0.80, 0.90 and 0.95 quantiles) suggests poorer survival prospects with increasing age (Figure 3.19).

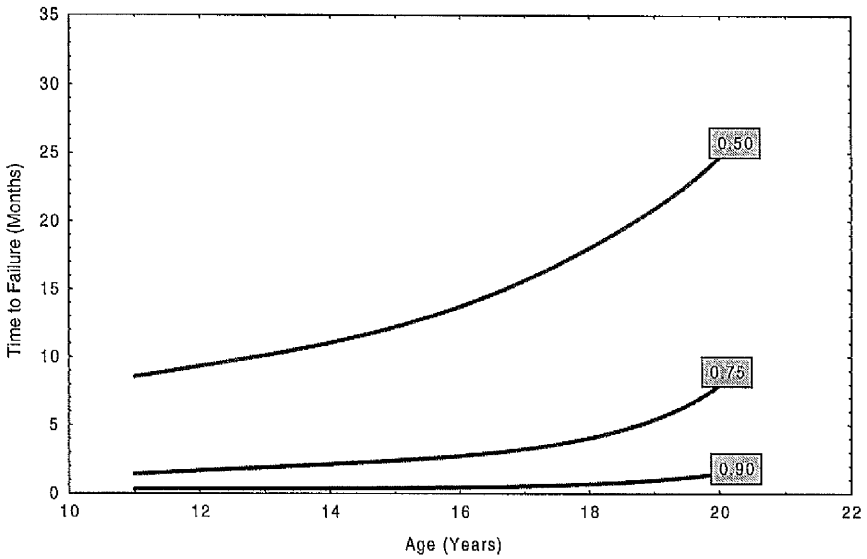
Figure 3.19. Smoothed nonparametric quantile curves for Age for the Melanoma Data.



This appears to be in agreement with the Age risk groups suggested by the tree-based analysis as there are plausibly three steps in the smoothed nonparametric quantile curves corresponding to the splits suggested by the tree-based approach (i.e. 40 for all three quantiles and 60 for the 80% quantile).

As indicated earlier, a patient's age is the only continuous covariate in the Dental study. A tree-based analysis suggested that above and below age 15 years constituted the most significant "at risk" subgroups. The smoothed nonparametric quantile curve (Figure 3.20) also suggests that bracket failure is less likely with increasing age and that the proposed age risk groups suggested by the tree approach seem plausible.

Figure 3.20. Smoothed nonparametric quantile curves for Age for the Dental Data.



Note, if necessary, a permutation test procedure could be adopted here to assess whether the non-linear patterns evident in all the above smoothed nonparametric quantile curves are due to sampling variation or not. If there was no association between the covariate and observation time each covariate is equally likely to have 'occurred' at each observation time. The procedure would therefore involve computing a permutation envelope (similar to those proposed in Section 3.6.1.3 above) by estimating smoothed nonparametric quantile each time using the recorded observation times and a random permutation of the covariate values in the sample.

3.9 Chapter Summary

When analysing matched or paired data a simple scatterplot (with a line of equality superimposed) is the most common method to graphically present such data. This allows the statistician to make a subjective impression as to whether there is an 'average' difference between the 'populations' or indeed if any underlying bivariate pattern is present.

When graphing survival data however, the interpretation of a simple scatterplot is not as straightforward due to the presence of censoring in the data. One way around this is to label each pair by 'joint' censoring status which gives some indication of the 'performance' of each pair and allows a subjective impression to be made as to the pattern of the censoring present.

The most common method for displaying 'independent' survival data from two distinct populations is the Kaplan-Meier estimate. Extensions of this method have been suggested in this chapter for displaying estimates of the marginal survival distributions of the cases and controls in the form of a ratio plot. A permutation test based method for generating 'reference ranges' was proposed which aids in determining whether an observed effect may indeed be real and not due to sampling variation alone. Following this, a review of proposed methods for graphing the bivariate survivor function was given with methods for obtaining suitable reference ranges.

Finally, methods for graphically assessing the independent effect of each of the categorical and continuous matching variables and unmatched covariates on survival were presented. In order to graphically display categorical variables, an extension of the ratio plot was suggested. When graphing continuous variables the first method used a tree-based approach while the second method presented uses a form of kernel estimation to construct an estimator of a percentile of the survivor function as a function of the covariate.

There was a suggestion of a slight improvement in the survival prospects of Multiple Melanoma sufferers when compared to matched Single Melanoma counterparts. This suggested improvement only appeared during certain epochs and therefore it is not clear whether there is an 'overall' improvement. There was a strong suggestion however that a patient's sex, tumour thickness and whether their tumour was ulcerated or not strongly influenced survival prospects for melanoma.

There was no suggestion however of any difference in the time to bracket failure for the Test (i.e. Glass Ionomer) and Control (chemically cured) cement types. The only variable which seemed to affect time to bracket failure was the patient's age.

Having provided mechanisms for displaying data in 'dependent' survival contexts (and hence providing subjective impressions), the obvious next stage in the analysis is to formally test for a difference in survival and this will form the basis for the next chapter.

Chapter 4

Methods for Comparing Matched/Paired Survival Data ignoring Covariates

4.1 Introduction

The previous chapter introduced methods to compare graphically the underlying population distributions of 'time to event' between two distinct populations. These methods are essential in providing mechanisms for obtaining a subjective impression as to whether there is a difference in survival between the two populations or not.

The next stage in any analysis is to formally assess any possible difference in the distribution of survival time between the two populations.

Initially a brief review of formal analytic techniques (mostly in the form of hypothesis tests) for the comparison of survival in two independent groups of survival data will be presented followed by a presentation of formal methods for comparing the survival experience of two groups (i.e. the two arms of the primary variable) of matched or paired survival data while ignoring matching variables and all other covariates. Illustrations of these formal methods are provided using the Melanoma and Dental data sets introduced in Chapter 2.

4.2 Comparing Two Independent Samples of Survival Data

In the rare case of survival data with no censoring, the data could be modelled by some parametric family (such as exponential) or possibly transformed to normality (perhaps using a log transformation) and then the appropriate normal theory analysis could be applied (e.g. a two sample t-test). Alternatively a non-parametric approach such as the Mann-Whitney test could be used. However, the methods used for 'sampling' survival data are such that censoring is almost always inevitable and hence must be accounted for in any analysis.

Most non-parametric procedures involve replacing the actual observations with rank statistics and performing a test on the ranks (e.g. the sign and signed-rank procedures). In the case of survival data the use of rank procedures is attractive for dealing with the asymmetric or heavy tailed nature of the data.

Methods for analysing survival data from two independent groups based on rank statistics were introduced in the literature from as early as 1959. The most important of these are the:

- Cox-Mantel Test (Cox 1959, 1972; Mantel 1966),
- Gehan's Generalised Wilcoxon Test (Gehan 1965),
- Cox's F-test (Cox 1964),
- Peto and Peto's Generalised Wilcoxon test (Peto and Peto 1972)

and the

- Log Rank Test (Mantel and Haenszel 1959, Peto and Peto 1972).

In general, each test has as its basis a test statistic based on the ranks/scores of both samples in a 'combined/pooled' form. The test statistics basically differ only in the choice of such ranks/scores and in the reference distribution used under the null hypothesis of 'no difference'. A full review and comparison of these tests is given by Lee (1992) where the Log Rank test and Peto and Peto's Generalised Wilcoxon test are recommended in general.

This thesis however is primarily concerned with analysing survival data when the assumption of independence is likely to be inappropriate and hence standard hypothesis tests for comparing independent samples of survival data would be inappropriate, if not potentially misleading. The main emphasis of this chapter therefore is to present methods to analyse dependent survival data from matched and paired survival studies.

The first approach presented is a hypothesis test that compares the population survival distributions for the case and control group by assessing each pair in terms of whether a definite decision can be made in terms of pair performance.

The second approach involves a selection of hypothesis tests for the direct comparison of the survival distributions of the case and control groups which allow for the dependency in the data. These tests are similar in nature to the tests referenced above for comparing independent samples of survival data as they are again based on 'translating' each original observation into a score and the test statistic is based upon 'pooling' these scores in some manner.

Finally, methods for modelling the difference between the survival time of a case and the survival time of its matched/paired control (i.e. modelling the difference in a 'typical' pair) are given. Of these methods, the first involves assuming the family of the underlying population distribution of differences is known while the second, a non-parametric approach, needs no specific underlying distributional assumption. Both methods can be used to provide point and interval estimates of any appropriate population quantile (e.g. median) of the difference in survival between the cases and controls.

4.3 Comparing Two groups of Dependent Survival Data

Once again, if censoring was not present statistical tests such as the paired-t test, non parametric sign and signed rank tests could be used to assess whether the population of the differences has zero centre (i.e. the population mean/median difference is zero). As mentioned earlier, most survival data, and in particular matched survival data, has a degree of censoring present so such standard techniques are unsuitable.

Chapter 2 introduced notation for the general framework of matched/paired survival data with (t_{ip}, δ_{ip}) used to represent the observation time and censoring indicator respectively for the ith individual of the pth pair where i=1,2 and p=1, .., P. Once again without any loss of generality let the first individual in the pair (i=1) be referred to as the 'case' and the second individual (i=2) the 'control'.

As this chapter is concerned with comparing the survival distributions of cases and controls, while at present ignoring all covariates, all the methods presented involve a 'pairwise comparison' of observation times. Each pair provides two observation times each with a censoring indicator. In terms of censoring there are four possible 'states' for each pair namely one, both or neither pair member may be censored. This will become an important consideration in later sections.

The bivariate survival scatterplot introduced in Chapter 3 gave a visual impression of the performance of each pair in terms of whether or not one pair member outlived the other member. Effectively exploiting the information in such a graph, the first formal test for deciding if there is a significant difference in pair performance between case and control population is now presented.

4.3.1 The Simple Binomial Test

This test is the most basic for analysing matched/paired survival data where the test statistic, the number of pairs where the case definitely outlives the control, is used to test the Null Hypothesis that P(case outlives control) =½.

Define a score statistic O_p where

$$O_p = \begin{cases} 1 & \text{if } t_{1p} > t_{2p} \quad \text{and} \quad \delta_{2p} = 1 \quad p = 1, .., P \\ 0 & \text{if } t_{1p} < t_{2p} \quad \text{and} \quad \delta_{1p} = 1 \quad p = 1, .., P \\ \text{undefined} & \text{otherwise} \end{cases}$$

and thus is an indicator of a pair's outcome in terms of whether the first member of the pair strictly outlives the second member or not. Hence a score of 1 is given if the case strictly outlives the control while a score of 0 is given if the control strictly outlives the case. Let S ($\leq$ P) be the number of pairs where O_p is defined.

Let $O_T = \sum_{p=1}^{S} O_p$ i.e. the total number of pairs where the case strictly outlives the control. Now if all the O_p can be considered independent, then O_T can be assumed to be distributed as Binomial (S, θ), where, under the null hypothesis of equality of the distribution of the population of the cases and the matched controls, θ = ½. The appropriate hypothesis test is then achieved by referring the observed value of O_T to a Binomial(S, ½) distribution.

Note that the Simple Binomial test uses the overall 'pair performance' as the criterion for testing for a difference in survival between the cases and controls, however not all pairs have definite "outcome" (e.g. both pair members are censored or one pair member is censored at a time less than its matched/paired event time) and therefore have to be excluded.

The next approach also involves comparing case and control survival using a score but in this approach information from all pairs is used. The tests are similar to the Log-Rank procedure in terms of using a rank statistic from a combined/pooled order statistic.

4.3.2 Rank Tests for Matched/Paired Survival Data

As mentioned earlier, a common non-parametric procedure for comparing independent groups of survival data involves replacing the actual observations with rank statistics and carrying out a hypothesis test on the ranks. Adaptations of these procedures have been developed, based on rank statistics, for matched censored problem, notably by

- The Woolsen and Lachenbruch test (1980),
- Gehan's test (Wei 1980),
- The Paired Prentice-Wilcoxon test (O'Brien and Fleming 1987)

and

- The Akritas test (Akritas 1990).

Briefly, the test developed by Woolsen and Lachenbruch is specifically designed for data following a Weibull distribution and incorporates a generalised sign rank test. The procedure for the Gehan, Paired Prentice-Wilcoxon and Akritas tests involves assigning scores to each member of a pair and then computing the difference in score within each pair. For example, in the Gehan test, the score for the i^{th} individual is based on the proportion of individuals with observation time known as less than the observation time of the individual minus the proportion with observation time known as greater. For reasons that will become clear, only a detailed description of the Paired Prentice-Wilcoxon and the Akritas tests are given here.

4.3.2.1 The Paired Prentice-Wilcoxon Test

For the Paired Prentice-Wilcoxon test the following procedure is employed :

1. Order the E case and control event times to obtain $t_{(j)}$, j = 1, ... ,E where $E \leq 2P$.

2. For j=1 to E, let $n_{(j)}$ be the number of cases and controls with observation times greater than or equal to the j^{th} distinct ordered observed death time, $t_{(j)}$.

3. Define $s_e = \prod_{j=1}^{e} \frac{n_{(j)}}{[n_{(j)} + 1]}$, for e= 1, ..., E.

4. Define the Prentice-Wilcoxon score PW_1 for the ***cases*** as follows:

   if the case in the p^{th} pair corresponds to the e^{th} distinct ordered observed event then assign the score

   $$PW_{1p}=1-2s_e$$

while

   if the case in the p^{th} pair is censored at some time between the e^{th} event and the $(e+1)^{st}$ event assign the score

   $$PW_{1p}=1-s_e.$$

5. In the same manner, define the Prentice-Wilcoxon score PW_1 for the ***controls*** as follows:

if the control in the p^{th} pair corresponds to the e^{th} distinct ordered observed event then assign the score

$$PW_{2p}=1-2s_e$$

while

if the control in the p^{th} pair is censored at some time between the e^{th} event and the $(e+1)^{st}$ event assign the score

$$PW_{2p}=1-s_e.$$

6. Let $\Delta_p = PW_{1p} - PW_{2p}$ i.e. the pairwise difference in Prentice-Wilcoxon scores for the p^{th} pair (p=1, .. P).

7. Compute the test statistic $Z_{PPW} = \left(\frac{\sum_{p=1}^{P} \Delta_p}{\left(\sum_{p=1}^{P} \Delta_p^2 \right)^{\frac{1}{2}}} \right)$.

It is assumed that Z_{PPW} approximates in distribution to the standard normal.

Note the above procedure assumes that there are no ties in the observed event times. If ties are present and, for example, several events occur at $t_{(j)}$, these times are then arbitrarily ordered by assigning them distinct values infinitesimally to the left of $t_{(j)}$. The scores are calculated as in steps 1-5 above and then each event originally occurring at $t_{(j)}$ is assigned the average of the corresponding arbitrarily ordered scores.

4.3.2.2 Akritas' Test

The Paired Prentice-Wilcoxon described above was based on calculating scores for each member of a pair based on a 'pooled' order statistic and comparing Z_{PPW} to tables of the standard normal distribution. The Akritas test is similar in nature to the Paired Prentice-Wilcoxon in that a score (using a different procedure but still based on ranks) is assigned to each member of a pair but differs in that a paired *t*-test is then used on these scores. The test is essentially a combination of both parametric and non-parametric approaches in that the original data are 'translated' into ranks (similar to many non-parametric approaches) while a parametric approach is then applied to these ranks.

Akritas (1990) provides a method for estimating the ranks for survival data by defining the rank of a complete observation t_i through the Kaplan-Meier estimator as $n(1-\hat{S}(t_i))$ and the rank of an censored observation as $n\left[1-\frac{1}{2}\hat{S}(t_i)\right]$. The uncensored observation rank is justified by remembering that the survivor function is the complement to the cumulative distribution function. The censored observation rank estimator however is justified by assigning each censored observation the average of the ranks of all observations to its right which, in essence, is a reversal of Efron's (c.f. Chapter 1, section 1.4.2) proposed "redistribution to the right" idea.

Akritas (1990) showed that a paired *t*-test on the ranks is an asymptotically valid test procedure for testing the equality of the case and control survivor functions.

The test mechanism is as follows:

1. Compute $\hat{S}_1(t)$ and $\hat{S}_2(t)$, the Kaplan-Meier estimated case and control survivor functions respectively.

2. Let $\overline{S}(t)$ be the average of $\hat{S}_1(t)$ and $\hat{S}_2(t)$.

3. The rank transformation step involves replacing each observation

$$t_{1p} \text{ by } t_{1p}^* = 2n\left((1-\overline{S}(\bullet))\delta_{1p} + \left[1-\tfrac{1}{2}\overline{S}(\bullet)\right](1-\delta_{1p})\right)$$

and

$$t_{2p} \text{ by } t_{2p}^* = 2n\left((1-\overline{S}(\bullet))\delta_{2p} + \left[1-\tfrac{1}{2}\overline{S}(\bullet)\right](1-\delta_{2p})\right).$$

4. A paired *t*-test is then used on $t_{1p}^* - t_{2p}^*$ to test the equality of the case and control survivor functions (i.e. H_0: $S_1(t)=S_2(t)$).

Note that the Akritas test uses information from all of the pairs while the issue of tied observation times is taken care of in step 1 above and therefore poses no additional problems.

In order to compare the Woolsen and Lachenbruch, Gehan, Paired Prentice-Wilcoxon and Akritas test, Woolson and O'Gorman (1992) carried out a comprehensive Monte Carlo simulation study to compare the size and power of each test. In all simulations 1000 samples of 30 and 100 pairs of observations times were generated with

censoring distributions specified to achieve 30%, 50% and 80% censoring. Survival times were generated for each sample size and censoring configuration initially using an exponential distribution and finally an exponential distribution incorporating outliers. The dependency structure was generated by adding an exponentially distributed random variable to the 'common' survival time for the pair. The results of Woolson and O'Gorman showed that the Akritas test and Paired Prentice-Wilcoxon test are somewhat more powerful than the Gehan statistic for most of the scenarios studied and as such were recommended for general use.

In order to illustrate the Paired Prentice-Wilcoxon and Akritas tests a trivial example of the computational steps in both tests is now given using the matched survival data displayed in Table 4.1.

Table 4.1. Example Data for Paired Prentice-Wilcoxon and Akritas Test Illustrations

p (pair)	**i (individual)**	**t (observation time)**	**δ (censoring indicator)**
1	1	12	1
1	2	7	1
2	1	10	0
2	2	10	1
3	1	16	0
3	2	14	0
4	1	8	1
4	2	18	0
5	1	9	1
5	2	26	1

Following steps 1-3 of the Paired Prentice-Wilcoxon test give the following results:

Table 4.2 Paired Prentice-Wilcoxon Test Illustration for Example Data.

t_e	7	8	9	10	12	26
n_j*	# ≥ 7 10	# ≥ 8 9	# ≥ 9 8	# ≥ 10 7	# ≥ 12 5	# ≥ 26 1
s_e	1*10/11 0.909	0.909*9/10 0.818	0.818*8/9 0.727	0.727*7/8 0.636	0.636*5/6 0.530	0.530*1/2 0.265

allowing the appropriate calculation of R_{PPW} for each observation:

p (pair)	i (1=Case, 2=Control)	t	δ	PW_{ip}
1	1	12	1	1-2*(0.53) = -0.06
1	2	7	1	1-2*(0.909) = -0.818
2	1	10	0	1- 0.636 = 0.364
2	2	10	1	1-2*(.636) = - 0.272
3	1	16	0	1-0.530 = 0.469
3	2	14	0	1-0.530 = 0.469
4	1	8	1	1-2*(0.818) = -0.636
4	2	18	0	1-0.530 = 0.469
5	1	9	1	1-2*(0.727)= -0.454
5	2	26	1	1-2*(0.265) = 0.469

From the table above $\sum_{p=1}^{5} \Delta_p = -0.636$, $\sum_{p=1}^{5} \Delta_p^2 = 1.748$, Z_{PPW} =0.363 and a p-value of 0.72 suggests no significant difference in the population survival distributions of the two groups.

The first step in the Akritas test involves calculating Kaplan-Meier estimates of the marginal case and control survivor functions, $\hat{S}_1(t)$ and $\hat{S}_2(t)$ respectively:

t	$\hat{S}_1(t)$	$\hat{S}_2(t)$
7	1	**0.80**
8	**0.80**	0.80
9	**0.60**	0.80
10	0.60	**0.60**
12	**0.30**	0.60
26	0.30	**0.00**

Using these estimates and the censoring information and t* score can be calculated for each case and control as illustrated in Table 4.3.

Table 4.3 Akritas Test Illustration for Example Data.

p	i	t	δ	$\overline{S}(t)$	t*
1	1	12	1	½(0.3+0.60)=0.45	2*5[1- 0.45] =5.5
1	2	7	1	½(1.0+0.80)=0.90	2*5[1- 0.90] = 1
2	1	10	0	½(0.6+0.6)=0.60	2*5[1-½(0.60)] =7.0
2	2	10	1	½(0.6+0.6)=0.60	2*5[1- 0.60] = 4
3	1	16	0	½(0.3+0.60)=0.45	2*5[1-½(0.45)] = 7.75
3	2	14	0	½(0.3+0.60)=0.45	2*5[1-½(0.45)] = 7.75
4	1	8	1	½(0.8+0.8)=0.80	2*5[1- 0.80] = 6
4	2	18	0	½(0.3+0.60)=0.45	2*5[1-½(0.80)] = 7.75
5	1	9	1	½(0.6+0.8)=0.70	2*5[1- 0.70] = 6.5
5	2	26	1	½(0.3+0)=0.15	2*5[1- 0.15] = 9.25

The result of a paired t-test on the above t* scores gave the following results: t=0.43, df = 4, p-value = 0.68. As with the PPW test this suggests no significant difference in the population survival distributions for cases and controls.

4.3.3 Summary of Proposed Tests

The Simple Binomial test compares the case and control populations by comparing the performance of the each pair in the sample in terms of whether the case or control performed better. No estimate of the marginal survivor functions is used. The Akritas test, and the Paired Prentice-Wilcoxon test to some extent do incorporate marginal survivor function estimates in their testing procedures. The Akritas test uses the average of case and control Kaplan-Meier marginal survival estimates while, in step 3 of the Prentice-Wilcoxon test, a "product limit" type argument, similar to that of the KM estimate, is used.

All three tests provide a mechanism for hypothesis tests of equality of population survival distributions for the case and control groups of matched or paired survival

data which are based on rank statistics rather than the actual survival times directly. The next approach considered here however involves trying to estimate the distribution of the population case and control survivor functions in order to assess whether the population of 'differences' is centred around zero (i.e. effectively no difference between the case and control survival distributions).

4.4 The Differences in Survival Times

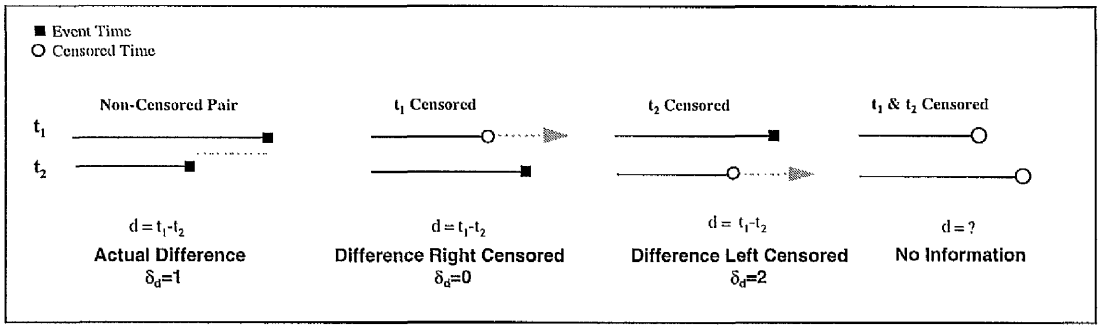
Define the survival time difference d_p for a pair p as

$$d_p = t_{1p} - t_{2p} \quad \text{where } p=1, .., P.$$

As stated earlier, in the absence of censoring some appropriate parametric procedure such as a paired t-test based on the d_p could be used to compare the survival distributions of the two populations. However, it is rarely the case that all observations are 'complete' as censoring is almost invariably present.

For pairs with one member censored the difference in survival time for that pair will be at most or at least the arithmetic difference between the event times in that pair depending on whether it is the first or second member that is the censored observation. In the situation of a doubly censored pair no clear information is available as to what the true difference in survival time might be (see Figure 4.1) although there may be information in each member about the underlying marginal survivor function of each population.

Figure 4.1. Illustrating the Censoring Indicator for the Difference in Survival for a particular Pair.



Using the information provided by the case and control censoring indicators δ_{ip} a 'new' censoring indicator δ_p can be formed for d_p as follows:

$$\delta_p = \begin{cases} 1 & \text{if} \quad \delta_{1p} = \delta_{2p} = 1 \\ 0 & \text{if} \quad \delta_{1p} = 0 \text{ and } \delta_{2p} = 1 \\ 2 & \text{if} \quad \delta_{1p} = 1 \text{ and } \delta_{2p} = 0 \\ \text{undefined} & \text{if} \quad \delta_{1p} = \delta_{2p} = 0 \end{cases}$$

representing an exact, a right censored, a left censored and an undefined difference respectively. The problem now reduces to a univariate estimation problem with both left and right censoring possibly present.

One approach would be to make some distributional assumption about the differences (e.g. Normality) and simply treat the resulting likelihood of "complete" and "censored" differences in a standard manner. Inferences on some simple summary population measure of the distribution of differences could then be made (e.g. mean

population difference). The other approach drops the specific distributional assumption and relies on a non-parametric density estimate for the population of differences. Inference can then be concentrated on an appropriate quantile (e.g. the median).

4.5 Estimating the Distribution of Survival Time Difference

The aim of this section is to model the distribution of the pairwise difference in survival time and to make inferences on an appropriate summary of such a model adopting either a parametric or a non-parametric approach. Then, for example, a formal test of the population median difference being zero can then be carried out based on whether a confidence interval for population median difference contains zero or not.

Note that the null hypothesis for all of the tests presented in section 4.3. is $S_1(t)=S_2(t)$ while the null hypothesis for both tests presented in section 4.5 below is that the population median difference is zero. This is an important distinction to make.

4.5.1 Parametric Approach

Let d_p, p=1, .. P, arise from some distribution of the differences in survival with cumulative distribution $F(d/\theta)$ and probability density function $p(d/\theta)$ for some unknown parameters θ.

The likelihood of such data is

$$L(\theta) = \prod_{\text{all complete } d_p} (p(d_p/\theta)) \prod_{\text{all right-censored } d_p} (1 - F(d_p/\theta)) \prod_{\text{all left-censored } d_p} (F(d_p/\theta))$$

where there is no contribution from any doubly censored pairs. In theory, any parametric form for the probability density function and the cumulative distribution is possible but perhaps the most natural choice is to assume normality.

Assuming that the d_p are normally distributed, interest rests on estimating the population mean (and hence also median) difference μ and population variance σ. The appropriate likelihood is

$$L(\mu,\sigma) = \prod_{\text{all complete } d_p} \left(\phi(\frac{d_p - \mu}{\sigma}) \right) \prod_{\text{all right-censored } d_p} \left(1 - \Phi(\frac{d_p - \mu}{\sigma}) \right) \prod_{\text{all left-censored } d_p} \left(\Phi(\frac{d_p - \mu}{\sigma}) \right)$$

The maximum likelihood estimate of μ, the population mean difference in survival, is then the value $\hat{\mu}$ for which $L(\mu,\sigma)$ is maximised, or more conveniently, the value $\hat{\mu}$ for which the natural logarithm of $L(\mu,\sigma)$ is a maximum. An approximation for the estimated standard error of $\hat{\mu}$, along with appropriate confidence intervals based on asymptotic normality can as usual be based on the information matrix.

The above approach provides an estimate of the mean difference in survival between the cases and controls while incorporating both the paired and censoring structure of the data. This approach can then be used to formally test for a difference in survival

between the two underlying case and control populations by assessing whether the confidence interval for the population mean difference contains zero.

4.5.2 Non-Parametric Approach

The parametric approach above assumes that the underlying family of distributions of the differences in survival times is known. An alternative non-parametric approach for estimating the distribution of the difference in the survival of the two populations is now introduced. This approach relies heavily on the method proposed by Turnbull (1976) which provides a non-parametric estimate of a density function for univariate survival data which consists of complete, right censored and left censored observations.

4.5.2.1 The Self-Consistent Approach of Turnbull

Turnbull (1974) developed an approach for estimating the survivor function for univariate survival data with both left and right censoring present which can be considered a modification of the Kaplan-Meier estimator. In the absence of any left censored observations his approach reduces to the Kaplan-Meier estimate of the survivor function.

Turnbull's algorithm extends the idea of self-consistency initally presented (Efron 1967) for right-censored data, to derive a maximum likelihood estimate of the survivor function for data with both left and right censoring.

Assume a grid of differences $d_0 < d_1 < \ldots < d_m$ (i.e. both actual and censored differences). Note that d_0 may not necessarily equal zero due to possible negative differences but this is immaterial as Turnbull's method is based on ranks.

Let e_j be the number of true differences (i.e. events), r_j be the number of right censored differences and l_j the number of left censored differences at each d_j (j= 0,...,m). The only information provided by a left censored difference l_j is that the true difference is some value $\leq l_j$. Turnbull's approach, using a self-consistent estimator, estimates the probability that this true difference occurred at each possible d_{j*} (where $j^* < j$) based on an initial estimate of the survivor function. Using this estimate the expected number of events (i.e. the true differences) at d_{j*} is calculated which is then used to update the estimate of the survival function. The procedure is continued until the difference in successive estimates of the survivor function is negligible.

4.5.2.2 Summary of Turnbull's Algorithm applied to Paired Difference Problem

1. Calculate an initial estimate of the survivor function, S(d), at each d_j, using the Kaplan-Meier estimator ignoring all left censored observations.

2. Using the current estimate of S(d), estimate

$$p_{jj^*}=P[d_{j-1} < D < d_j \setminus D \leq d_{j^*}] \text{ where } j \leq j^*$$

by

$$\frac{S(d_{j-1}) - S(d_j)}{1 - S(d_{j^*})} \quad \text{for} \quad j \leq j^*.$$

3. Using the results of the previous step, estimate the number of events (i.e. true differences) at d_j by

$$\hat{e}_j = e_j + \sum_{j^*=j}^{m} l_j p_{ij^*}.$$

4. Calculate the Kaplan-Meier estimate based on the estimated right censored data with $\hat{e}_j$ events and r_j right censored observations at d_j ignoring all the left censored observations.

5. If the estimates from successive iterations are close for all d_j (say to within 0.0001) stop the procedure, if not return to Step 2.

In effect, Turnbull's method involves "redistribution to the right" and "redistribution to the left". The right censored data are accounted for in the Kaplan-Meier estimation while the left censored observations are accounted for in Step 2.

The estimator of the survival function based on Turnbull's algorithm can also be derived using a modified EM algorithm (Dempster, Laird and Rubin 1977) approach. Several suggestions have been made to speed up the estimation procedure (Wellner and Zhan 1998). An additional recommendation by Zhou (1997) is to change the smallest and largest observations to event times to make the estimator 'behave' like a true distribution at the tails. Regardless of the approach taken, the estimators are generalised maximum likelihood estimators (Turnbull 1976).

4.5.2.3 Estimating the Variance of $\hat{S}(d)$.

Turnbull (1976) showed that his proposed estimator is a non-parametric maximum likelihood estimator and using maximum likelihood theory presented an estimator for the variance of $\hat{S}(d)$ based on the information matrix as follows:

Define $I=[I_{i,j}]$ by

$$I_{i,i} = \frac{e_i}{[\hat{S}(d_{i-1}) - \hat{S}(d_i)]^2} + \frac{e_{i+1}}{[\hat{S}(d_i) - \hat{S}(d_{i+1})]^2} + \frac{r_i}{\hat{S}(d_i)^2} + \frac{l_i}{[1 - \hat{S}(d_i)]^2}$$

$$\text{for } i = 1, ..m\text{-}1$$

$$I_{m,m} = \frac{e_m}{[\hat{S}(d_{m-1}) - \hat{S}(d_m)]^2} + \frac{r_m}{[\hat{S}(d_m)]^2} + \frac{l_m}{[1 - \hat{S}(d_m)]^2}$$

$$I_{i+1,i} = I_{i,i+1} = -\frac{e_{i+1}}{[\hat{S}(d_i) - \hat{S}(d_{i+1})]^2} \quad \text{for } i = 1, \ldots, m\text{-}1$$

and

$$I_{i,j} = 0 \text{ for } |i - j| \geq 2 \qquad \text{for } i = 1, \ldots, m\text{-}1.$$

The estimated variance covariance matrix $\hat{V}(\hat{S}(d))$ is the inverse of the matrix I and the estimated standard error for $\hat{S}(d)$ is obtained from the appropriate entry in the diagonal of $\sqrt{I^{-1}}$.

An alternative approach to estimating $V(\hat{S}(d))$ is to use the non-parametric bootstrap (Davison and Hinkley 1997, Zhan, 1998) where separate $\hat{S}(d)$ estimates (based on 500 samples for example) are obtained from a repeated sampling (with replacement) of the original difference data. Using this resampling procedure, estimates of the variance of $\hat{S}(d)$ can be obtained from the distribution of the resampled estimates.

4.5.2.4 Estimating the Median of S(d).

Once an estimate of the distribution of differences is available, estimating specific quantiles of the distribution may be of interest. One important summary measure of this distribution is the median. As the estimate of the survivor function is usually a step function, the point estimate of the median survival time is almost always an 'interval' but could be formally defined as the smallest time for which the value of the estimated distribution of differences is less than 0.5.

An approximate 95% confidence interval which can be constructed for the median of S(d) using the asymptotic likelihood properties of $\hat{S}(d)$ is

$$\left\{ d : \hat{S}(d) - 1.96\sqrt{\hat{V}[\hat{S}(d)]} < 0.5 < \hat{S}(d) + 1.96\sqrt{\hat{V}[\hat{S}(d)]} \right\}.$$

This confidence interval can then be used to assess formally for a difference in the survival difference between the two populations by simply considering whether zero lies in the interval or not. Note that the various transformations introduced in section 1.4.3 are applicable here also.

4.6 Examples

Examples of the different approaches outlined in this chapter for testing for a difference in survival distribution for matched and paired survival data (ignoring all covariates) will now be given for both the Melanoma and Dental data sets.

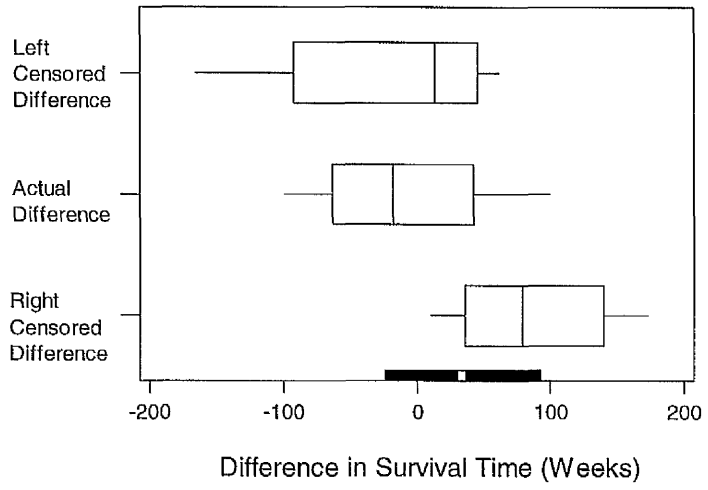
4.6.1 Melanoma Data

From the graphs presented in Chapter 3 there is a suggestion of slightly better survival prospects for the Multiple Melanoma sufferers over those suffering from Single Melanoma. The number of pairs where the Multiple Melanoma sufferer strictly outlived their Single Melanoma counterpart amounted to 15 while there were 8 pairs where the Single Melanoma sufferer strictly outlived the Multiple Melanoma sufferer. Using the Simple Binomial test this suggestion of improved survival for the 15 Multiple Melanoma sufferers proved non-significant ($p= 0.21$).

Neither the Paired PW test nor Akritas test suggested a significant difference in survival for the two Melanoma groups ($p= 0.76$ and 0.75 respectively).

The approach based on the differences in survival for each pair (i.e. Multiple - Single) first requires a plot of the data (Figure 4.2).

Figure 4.2.
Categorised Boxplot of the true and censored Survival Time pairwise differences, estimated mean difference and 95% Confidence Interval for the Melanoma Data.

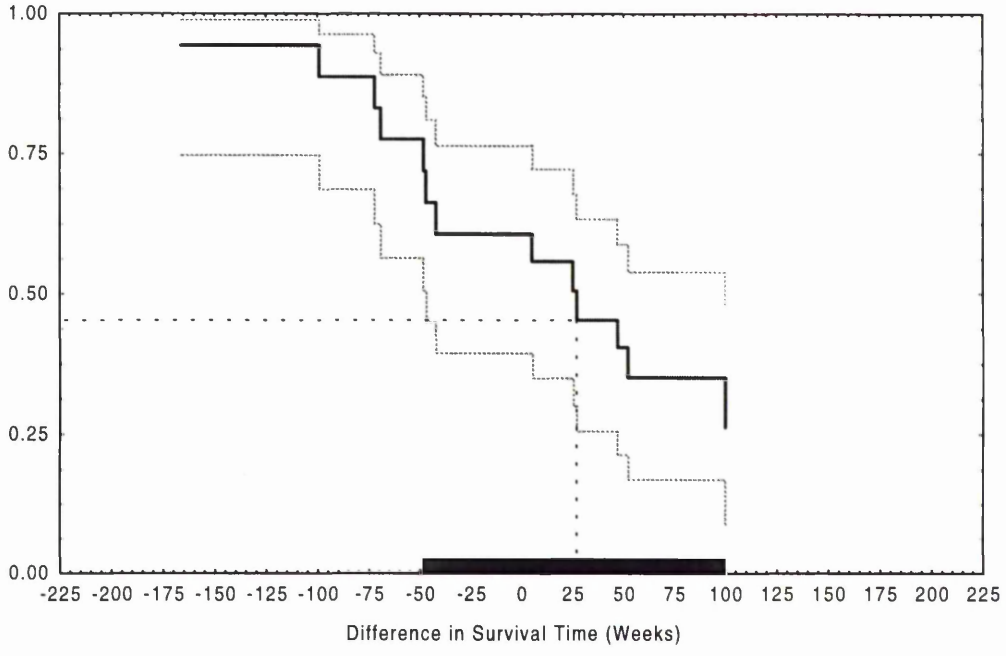


From this boxplot the assumption that the underlying distribution of the differences is normal seems reasonable. Using this normality assumption, an estimate of the population mean difference and 95% confidence interval thereof was obtained as 34.5 weeks and [-21, 91] weeks and is presented graphically in Figure 4.2. As this interval contains zero there is no strong suggestion of improved survival for either the Multiple or Single Melanoma sufferers although the interval is mostly positive agreeing with the conclusions reached in Chapter 3.

The final approach was to estimate the distribution of the difference in survival using Turnbull's method. The estimated survivor function of the difference with approximate pointwise 95% confidence intervals for the differences in Multiple and Single Melanoma survival times is given in Figure 4.3.

Figure 4.3

Estimated Survivor Function and 95% Pointwise Confidence Intervals for the Difference in Survival Times for the Melanoma Data.



The estimated median of the difference in survival times is 27 weeks with a 95% confidence interval of [-47, 100] (highlighted on the horizontal axis in Figure 4.3). Note, using a bootstrap estimate the resulting confidence interval was [-48, 100]. Once again there is no suggestion of a significant difference in survival for the Multiple and Single Melanoma sufferers. This is somewhat wider than the confidence interval based on normality (i.e. [-21, 91]) and is possibly more skewed to negative values perhaps suggesting that the asymmetry seen in the Figure 4.2 for left censored differences is not adequately modelled by normality.

In conclusion, when comparing Multiple and Single Melanoma survival while ignoring all matching covariates and other potential prognostic factors, there was no

strong suggestion of a significant difference in survival for the two types of Melanoma.

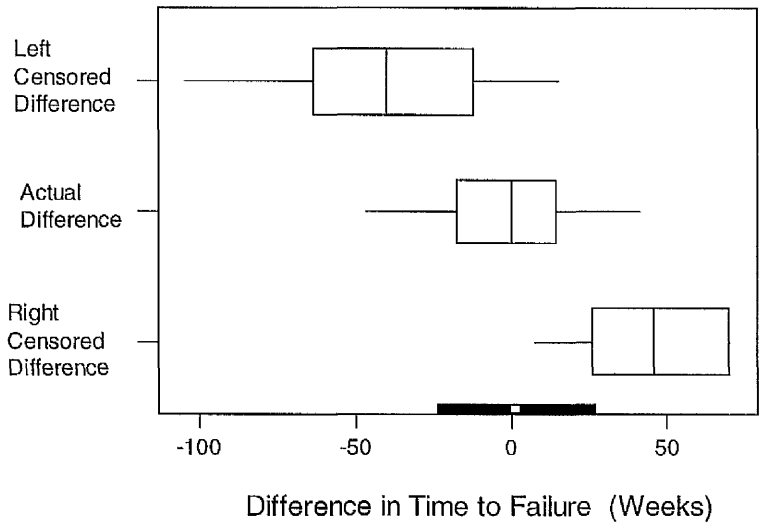
4.6.2 Dental Data

This study aims to compare two different cement types (Test and Control) in terms of their bonding strength. There was no clear suggestion of any difference between the cement types when looking at any of the relevant graphs presented in Chapter 2.

Of the 41 individuals involved in this study the Test material definitely outperformed the Control material in 14 individuals while the Control material outperformed the Test material in 9 individuals. Ignoring the 18 individuals where no clear decision could be made in terms cement performance, a Simple Binomial test suggested no significant difference between the cements (p=0.41). Neither the Paired Prentice-Wilcoxon test nor Akritas test suggested a significant difference in the bonding strengths of the two Cement Types (p=0.73, p=0.75 respectively).

A categorised boxplot of the true and censored differences in failure time along with the estimated mean difference and 95% confidence interval is given in Figure 4.4.

Figure 4.4.
Categorised Boxplot of the True and Censored Pairwise Failure Time Differences, estimated mean difference and 95% Confidence Interval for the Dental Data.

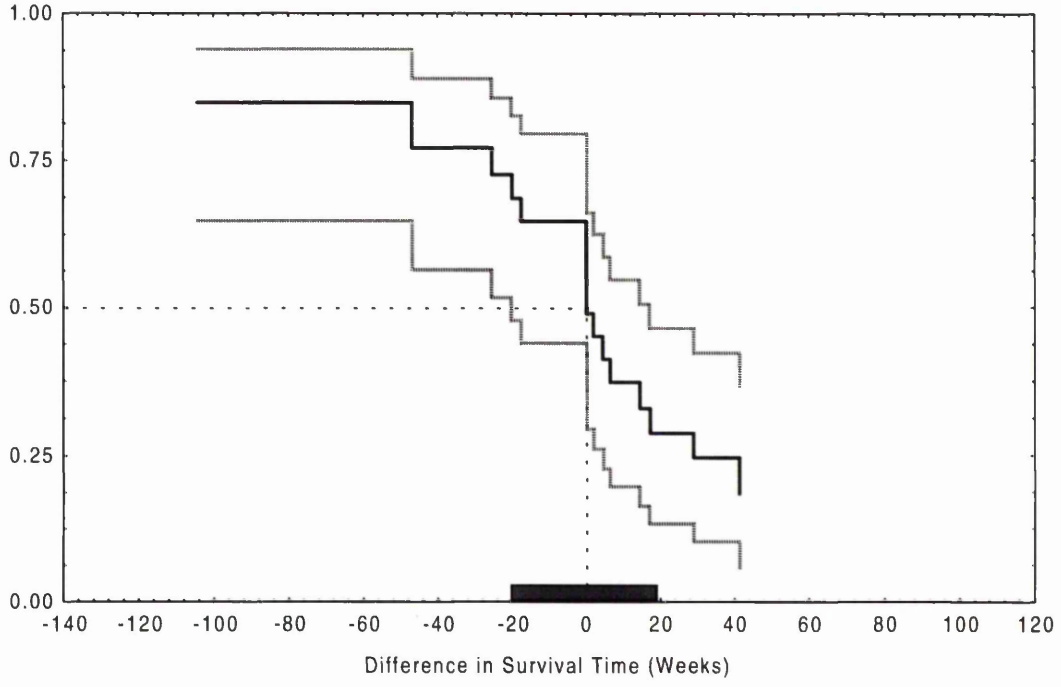


The assumption that the underlying distribution of the differences is normal seems reasonable, except perhaps for the right censored differences. Assuming normality, the estimated mean difference and corresponding 95% confidence interval was 1 week and [-23, 25] weeks respectively as depicted in Figure 4.4. As this interval is centered almost exactely at zero there is again no suggestion of improved survival for either type of bonding Cement bond.

The results of using Turnbull's non-parametric approach is presented in Figure 4.5 whic shows the estimated survivor function of the differences with approximate 95% pointwise confidence intervals.

Figure 4.5.

Estimated Survivor Function and 95% Pointwise Confidence Intervals for the Difference in Survival Times for the Dental Data.



The estimated median difference in time to failure for the two Cement Types is 0 weeks with a 95% interval of [-20, 17] weeks respectively. Note, using a bootstrap estimate the resulting confidence interval was [-20, 14]. Somewhat surprisingly this is a narrower interval than that obtained using normality. As neither interval contains zero there is once again no suggestion of a significant difference in the time to failure of the two Cement Types.

4.7 Chapter Summary

Two distinct approaches were considered in order to formally compare two population survival distributions based on matched/paired survival data. In this chapter all the approaches considered ignore all matching variables and other recorded covariates. The first approach presented a collection of hypothesis tests specifically incorporating the paired nature of the data. The Simple Binomial test does not utilise all the pairs available for analysis, rather only those where a clear decision can be made in terms of improved survival prospects. The Paired Prentice-Wilcoxon and Akritas test on the other hand do utilise all the pairs available in the data.

The second approach involved estimating the actual distribution of the difference in survival time either parametrically or non-parametrically, again using only those pairs where a difference could be defined. From these, point and interval estimates of relevant summary parameters or percentiles can be obtained. Based on the Melanoma and Dental datasets, the non-parametric approach seems more applicable in general as it is difficult to formally justify the normality assumption. A simulation study is required to assess all these approaches under a variety of different conditions and will be considered in a later chapter.

None of the approaches presented in this chapter incorporated the matching variables or the unmatched covariates into the analysis. In general, many of these variables/covariates are likely to influence survival. The next stage in this thesis is to provide methods for the comparison of the two survivor functions incorporating these variables/covariates and this will be the emphasis of the next chapter.

Chapter 5

Methods for Comparing Matched/Paired Survival Data incorporating Covariates

5.1 Introduction

This chapter concentrates on methods for incorporating matching variables and potential unmatched (or unit in the case of paired studies) covariates into the analysis while at the same time taking the dependency structure of the data into account. The comparison of the two 'arms' of the primary variable should then be more precise than a simple 'independent treatments' comparison.

As a first step an extension of the Simple Binomial Test which allows for covariates is presented. This extension uses as its basis a comparison of the pairwise performance of the cases and controls while adjusting for any significant covariate effects.

Following this, a selection of potential regression models for dependent survival data that centre on modelling the hazard rate rather than pair performance are discussed, in particular the proportional hazards model. This model is an attractive choice in terms of the primary goal of this chapter (i.e. assessing the effect of the primary variable on survival while adjusting for any covariate effects). Extensions to this approach that allow for dependent survival data are presented and examples given for both the Melanoma and Dental datasets.

5.2 The Role of Covariates in Dependent Survival Studies

As discussed in Chapter 2, for matched survival studies, the covariate vector contains the primary variable, matching variables and, possibly, some unmatched covariates. For paired survival studies on the other hand, it is not possible to record matching variables hence the covariate vector contains only the primary variable and, possibly, some additional 'unit' covariates (i.e. covariates relevant for both observations in any pair e.g. sex and age).

In both matched and paired studies the unmatched covariates may be considered as 'adjusters' where interest in these covariates matters only insofar as whether they are needed in a final model to adjust the estimates of the survival distribution of the primary variable for any imbalance that may exist across these covariates. For example, in the Melanoma study, Level of Invasion of a patient's tumour and a patient's sex may be needed as 'adjusting' covariates in the comparison of the survival distributions of Single and Multiple Melanoma sufferers.

The selection of appropriate and relevant/significant covariates may be based on partial likelihood ratio hypothesis tests (Klein 1997) and an automated stepwise variable selection procedure (Collett 1994). Stepwise procedures allow identification of subsets of covariates which are significantly related to survival time and thererfore appropriate for inclusion in any final model. Covariates are entered or removed on the results of significance tests based on large sample partial likelihood ratio tests.

As mentioned above, two distinct approaches to modelling matched or paired case-control survival data will be considered in this chapter. The first method uses a direct pairwise performance of the cases and controls and is now introduced.

5.3 Pair Performance Model

The first method for comparing survival between cases and controls is an extension of the Simple Binomial Test introduced in Chapter 4. A pair performance indicator variable was defined in Chapter 4 which categorised pairs by whether or not one member of a pair clearly survived longer that the other member.

Recall, this pair outcome performance indicator was defined in terms of a score statistic O_p where

$$O_p = \begin{cases} 1 \quad \text{if} \quad t_{1p} > t_{2p} \quad \text{and} \quad \delta_{2p} = 1 \quad p = 1,..,P \\ 0 \quad \text{if} \quad t_{1p} < t_{2p} \quad \text{and} \quad \delta_{1p} = 1 \quad p = 1,..,P \\ \text{undefined otherwise} \end{cases}$$

Note, as discussed in Chapter 4, that for some pairs no value of O_p can be obtained (i.e. both members of the pair are censored or one member censored before the other member died) and hence all such pairs are effectively excluded from this form of analysis. By considering only those S pairs where O_p is defined, an O_p score equal to 1 represents the case outliving the control while an O_p score of 0 represents a pair where the control strictly outlives the case.

The Simple Binomial Test uses as a test statistic the number of pairs where the case outlives its matched control (i.e. number of "successes") which, under the null hypothesis of identical survival distributions for cases and controls, should behave as a Binomial(S, ½) distribution. In order to incorporate any imbalance that may exist in the covariates (both matching and non matched covariates) an extension to the approach used in the Simple Binomial Test needs to be provided.

A natural extension in order to incorporate covariates is to calculate a 'difference' covariate z^*_{pc} for each pair with a "well-defined" value of O_p as follows:

$$z^*_{pc} = z_{1pc} - z_{2pc}$$

where c=2,....C (i.e. excluding the primary variable) and p=1, ...P.

Each new "difference" covariate z^*_{pc} is the pairwise-difference between the case value and the corresponding control value for the appropriate covariate over the S pairs where O_p is defined.

This so called "Pair Performance Model" is now presented.

5.3.1 Estimation in the Pair Performance Model

This 'data' can be viewed as a form of logistic regression where the response variable is the pair performance indicator. Techniques for modelling binary data are well established (McCullagh and Nelder 1989, Hosmer and Lemeshow 1989, Cox 1989).

By modelling the probability of the case outliving the control as a logistic regression on the covariate differences i.e.

$$P(O_p = 1 \mid z^*_{pc}) = \frac{\exp(\beta_0 + \beta' z^*_{pc})}{1 + \exp(\beta_0 + \beta' z^*_{pc})}$$

the likelihood for all S pairs where O_p is defined can be written as

$$L_{PP}(\beta_0, \beta) = \prod_{p=1}^{S} \left[\frac{\exp(\beta_0 + \beta' z^*_{pc})}{1 + \exp(\beta_0 + \beta' z^*_{pc})} \right]^{O_p} \left[\frac{1}{1 + \exp(\beta_0 + \beta' z^*_{pc})} \right]^{1-O_p}.$$

The constant term β_0 in this model is the probability of the case outliving the control when all other covariates in the final model are equal for a pair (i.e. $z^*_{pc}=0$ for all c=2,..,C). Stepwise procedures may be used to determine the "best" final model in terms of which "difference covariates" are found relevant for inclusion.

Maximum likelihood estimates $\hat{\beta}$ of the regression parameters can be provided using a Newton-Raphson approach while the asymptotic covariance matrix $\hat{V}$ may be estimated by the inverse of the negative of the matrix of second partial derivatives evaluated at the maximum likelihood estimates.

The role of the case and control survival times in the Pair Performance Modelling approach is purely to calculate the performance score (i.e. the response variable) for

each pair to be used in the logistic regression. The second series of approaches, presented below, will involve regression methods where the hazard rate is modelled and the full survival information from each pair is utilised.

5.4 Regression Models for Independent Survival Data

Before discussing these approaches to modelling 'dependent' survival data it is useful to revise regression techniques for 'independent' survival data. This will serve two purposes, firstly to establish notation and secondly to introduce concepts that will later be refined to incorporate the specific features of the problem at the heart of this thesis i.e. the analysis of dependent survival data.

Due to the nature of survival data and the general properties of the survivor function (as discussed in Chapter 1), standard regression techniques are not applicable. Two general approaches to the modelling of covariate effects on survival have been proposed. The first approach is analogous to classical linear regression but modelling some transformation f (e.g. natural log) of the survival time t in terms of the linear model

$$f(\mathrm{t}) = \mathrm{g}(\mathrm{Z}) + \varepsilon$$

where g(Z) is some pre-specified transformation of the covariate vector Z and ε represents the additive error distribution. There is a variety of choices for the error distribution. By letting the error distribution follow a standard normal distribution the resulting model is a log-normal regression model, the extreme value distribution

yields a Weibull regression while choosing a logistic distribution results in a log-logistic regression model. This family of regression models is commonly called the Accelerated Failure-Time (ACF) models and, as outlined above, a parametric assumption must be made regarding the error distribution on the linear model.

The second approach is to model the event *rate* in terms of the risk an individual has of experiencing the event of interest. The model is intrinsically linked to the hazard function and hazard rate by quantifying how a set of covariates influence the hazard rate for a particular individual. By specifying a model through the hazard function, specific questions such as how survival is related to the primary variable and other covariates under study can be addressed.

One of the most common such regression models is to allow the hazard function to be multiplied by a "risk score" for each individual. The hazard function is therefore a product of two functions: the underlying 'baseline' hazard function $h_0(t)$ which characterises how the hazard function changes as a function of time, and the risk score which characterises how the hazard function changes as a function of the covariates. The most popular choice for the risk score is $\exp(\beta' z)$ (Cox 1972) and the model is written as

$$h(t \mid z) = h_0(t)\exp(\beta' z)$$

where $h_0(t)$, represents the baseline hazard function (i.e. the hazard function for the "standard" individual and z represents a vector of covariates suitably centered).

Using the relationship between the survivor function and the hazard function discussed in Chapter 2, the proportional hazards model can be written in terms of the survivor function as follows

$$S(t \mid z) = S_0(t)^{\exp(\beta' z)}$$

where $S_0(t)$ is the baseline survivor function (i.e. the survivor function for a "standard" individual).

The principal assumptions underlying this model are:

1. the assumption of proportional hazards

   i.e. the ratio of hazard functions for two individuals

   with different covariates does not vary with time.

2. the relationship between the covariates and the hazard function

   should be linear in the log space.

3. i) the survival times for each individual should be independent,

   ii) the survival times and censoring times are independent

and

   iii) the censoring is not affected by the covariates.

The most widely used proportional hazards regression model in survival analysis involves leaving the parametric form of the baseline hazard function unspecified (Cox 1972) resulting in a very flexible model. This model is primarily used when the

emphasis is on estimating the relative effect of the covariates on the survival prospects of an individual rather than modelling the survivor function itself (which is often assumed to be of secondary interest).

5.5 The Cox Proportional Hazards Model (PH)

The proportional hazards model was first introduced by Cox in 1972 in his seminal paper entitled 'Regression models and life tables' which incorporated the idea of partial likelihood. The model proposed defines the hazard function $h(t \mid z)$ of a continuous random variable T for an individual with covariate vector z, as

$$h(t \mid z) = h_0(t)\exp(\beta' z)$$

where β is a vector of unknown parameters and $h_0(t)$ is the (unknown) baseline hazard rate.

The proportional hazards model formulation has the specific distinction that no parametric form is assumed for the underlying baseline hazard function. In essence the model can be considered as a semi-parametric model as a parametric form is only assumed for the covariate modelling.

It is assumed that the covariates themselves are not functions of time (or indeed change through time), although generalisation to incorporate time-dependent covariates in the model can be relatively straightforward (Cox 1975).

5.5.1 Estimating the regression coefficients β

Based on the notation already developed, $t_1, t_2, \ldots, t_n$ are the n observation times with δ_i as the corresponding indicator variables which is zero if the i^{th} survival time t_i is right-censored, and unity otherwise. Also let z_{ic} is the cth covariate (c=1, .. , C) associated with the individual whose observation time is t_i.

When there are no ties among the survival times, Cox formulated the partial likelihood L_{PH} for his proportional hazards model (Cox 1972) as follows

$$L_{PH}(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta' z_i)}{\sum_{j \in R(t_i)} \exp(\beta' z_j)} \right]^{\delta_i}$$

where $R(t_i)$ is the risk set at time t_i (i.e. the set of all individuals who are still at risk just prior to t_i).

The corresponding log-likelihood function is given by

$$\ln(L_{PH}(\beta)) = \sum_{i=1}^{n} \delta_i \left\{ \beta' z_i - \ln \sum_{j \in R(t_i)} \exp(\beta' z_j) \right\}.$$

Although this is not a genuine likelihood, Cox (1975) justified the form of L_{PH} within the framework of partial likelihood, and has shown that standard large sample maximum likelihood results may be applied with regard to resulting estimators. The

estimator of β has a large sample C-variate normal distribution with a mean of β and variance V, so that Wald, score or likelihood ratio hypothesis tests for global and local inference about β can be constructed.

Essentially the likelihood is a comparison of those individuals who experience an event at a particular time to those available to experience an event at that time. The estimates of β do not depend on the exact times at which the events occur, but rather on the rank ordering of the event times. The model adjusts for censored individuals by deletion of individuals from the risk sets. When ties occur in the survival times, two modifications of the conditional likelihood are available (Breslow 1974, Efron 1977).

Point estimation of β can be achieved by using the methods of maximum likelihood. First and second derivatives of $\ln(L_{PH})$ are easily obtained and iterative use of these through a Newton-Raphson numerical technique will yield parameter estimates. The asymptotic covariance matrix $\hat{V}$ may be estimated consistently by the inverse of the negative of the matrix of second partial derivatives evaluated at the maximum likelihood estimates.

In any regression model the estimated coefficient for a covariate represents the rate of change of a function of the response variable per unit change in the covariate. In the PH model a regression parameter can be interpreted as the logarithm of the *ratio* of the hazard of death for a particular individual with these covariate values to the

baseline hazard. These "ratios" play an analogous role in interpreting the results of a CPH model as odds ratios play in logistic regression.

If the covariate is a continuous variable (e.g. tumour thickness) then its estimated β coefficient from a proportional hazards model is the estimated change in the logarithm of the hazard ratio when the value of the covariate is increased by 1 unit. If, on the other hand, the covariate is a categorical variable with 2 levels (e.g. sex) then the estimated β coefficient is interpreted as the change in the logarithm of the hazard ratio from the first level to the second. In this case therefore, if $\exp(\beta)$ is larger than 1 the second level of the covariate has a higher risk of death while, if $\exp(\beta)$ is less than 1, the first level of the covariate has a higher risk of death.

In any modelling procedure it is important to investigate whether interactions (in particular with the primary variable in this context) are deemed necessary for inclusion in the model. One approach is to adopt a 0/1 coding system for the primary variable (where the case is coded as 0) and all categorical variables (i.e. matching and variables and unmatched covariates) in order to facilitate the ease of interpretation of the estimated regression coefficients. This coding scheme is particularly attractive when fitting interactions (which are the product of the original variables in the model) and for creating design variables when comparing specific contrasts. The significance of each separate interaction can be assessed (using design variables where necessary) by adding it to the main effects model and using partial likelihood ratio tests.

In the last few years the theoretical basis for the model has been solidified by formulating the model in terms of a counting process and invoking aspects of Martingale theory (Anderson 1993). Structuring the model in terms of a counting process essentially involves counting the number of events each subject experiences up to a specific and fixed time. This allows the formulation of a Martingale residual which can be considered as a difference between the observed number of events for an individual and the expected number given the current model (Barlo and Prentice 1988, Therneau et al 1990). Other residuals, whose primary function is in checking the model assumptions, are derived from the Martingale residuals (Therneau et al 1990) are discussed later.

5.6 The Independence Assumption and the PH model

The primary assumption of all the regression models for survival data discussed so far is that the survival times for each observation are independent. This limits the use of such models in the analysis of dependent survival data as the main assumption is automatically violated. However, one could still fit such models to dependent survival data ignoring the dependency. This would be analogous to carrying out a two-sample analysis (e.g. a two sample t-test) on paired data where the proper test would be a paired sample t-test on differences. Clearly if the dependency is 'weak' this might be 'perfectly' acceptable but if the dependency is 'strong' then the 'independent' analysis is likely to be very conservative in nature and therefore inferior to (and less powerful than) an appropriate 'dependent' analysis.

Adaptations to the Cox PH model to allow for dependent data, specifically involving matched and paired studies are now introduced.

5.7 Adapting the Cox PH Model to Clustered Data.

As discussed in Chapter 2, studies in Survival Analysis involving clustered survival data are primarily of two types, namely matched or paired survival studies.

In matched survival studies the matching variables are available for inclusion in the analysis while in paired studies the variables representing the 'degree of similarity' for the case and control are often unobservable and are thus 'hidden' from the analysis. In a genetic study for example, family members are considered similar but it may be impossible to accurately measure their degree of similarity to include in the analysis.

The main component of both of these types of dependent survival study is the 'pair' structure. It will often be reasonable to assume that each pair can be considered independent of all other pairs while individuals within a pair are 'clearly' dependent. A crucial point in the analysis is the assumption that pairs themselves can be considered nuisance parameters in that no estimate of pair effects are of any real importance. The main emphasis is to estimate the regression coefficients of the PH model and, principally, the 'difference in survival between the members of a pair i.e. the primary variable comparison.

The next sections develop extensions of the Cox PH model for both matched and paired survival studies.

5.8 Extensions of the Cox PH for Matched Survival Data

Extensions to the Cox PH model for matched survival data (i.e. where information is available on the matching variables) are introduced below.

5.8.1 The Conditionally Independent Cox PH Model (CPH)

In these contexts the actual values of the matching variables are available so the simplest approach here is simply to

i) ignore the matching structure,

ii) force the matching variables as covariates into the analysis

and

iii) include any relevant unmatched covariates chosen through some variable selection procedure.

A standard Cox model PH model is then applied to the data with the key purpose of the analysis the estimation of the effect of the primary variable.

Note that the use of ii) above will to some extent correct for any lack of 'quality in the matching' as it will allow for the actual values of the matching variables in any pair.

The basis of the conditional model therefore is that even though the matched structure of the data is ignored, information on the matching is used to adjust, or correct, the inferences made on the primary variable. The main justification of this model is that the assumption of conditional independence is balanced out by the use of the matching variables in the model.

The CPH model therefore is identical to the Cox PH model except that the strategy for variable selection is confined only to the unmatched covariates (i.e. the adjusters) as all matching covariates are forced into the model although their 'significance' should be noted and reported to the 'client' for any future studies involving matching in the particular context.

The hazard rate for the i^{th} individual in the p^{th} pair with covariate vector z_{ipc} (c=1,.., C) is modelled as

$$h_{ip}(t \mid z_{ipc}) = h_0(t)\exp(\beta' z_{ipc})$$

where the covariate vector contains the primary variable, all matching covariates and, possibly, some unmatched covariates are included.

Procedures for estimating β and its corresponding asymptotic covariance matrix V for the CPH model is identical to that for the Cox PH model.

5.8.2 The Marginal Cox PH Model (MPH)

The CPH model introduced in the previous section assumes that the lack of independence is controlled for by the matching variables alone. An extension of this approach is to model the data, initially ignoring the dependency structure, producing the usual estimates of the regression parameters and then correct the covariance structure of the regression parameter estimates to 'allow', to some extent, for the matching.

The model is identical to the CPH and the regression parameters β are once again estimated by maximising the partial likelihood $L_{PH}(\beta)$. The regression coefficients are estimated assuming independence (i.e. dropping the pair subscript and treating all observations as independent) while the estimated covariance matrix is then 'corrected' to account for the matching.

The marginal model was first proposed (Lee 1992) in the context of dependent survival studies involving clusters (e.g. litters, families etc.). When the number of observations in a cluster is small in comparison to the number of clusters Lee et al (1992) proposed a method to adjust the usual covariance matrix for the possible association within each cluster (this association measure being assumed the same across all clusters). Using large sample theory they proved that the estimator $\hat{\beta}$ is consistent for the underlying regression parameter β. If the survival times within each cluster are indeed independent, the corrected covariance matrix reduces to the covariance matrix calculated for fully independent data (Lin & Wei 1989). The corrected covariance matrix, defined as $\tilde{V}$, can be thought of as the summation of the

squares of all independent contributions to the partial likelihood (King 1996). A different derivation of $\tilde{V}$ is proposed by Therneau (Therneau 1997) by using a paired-jacknife estimate of the variance where the change in $\hat{\beta}$ on removing the pairs one at a time is used. In the context of jackknife estimation, the removal of each pair can be considered as the removal of an independent 'observation' from the data and forms the basis of the adjustment matrix. This approach has been shown to give the same results as that proposed by Lee et al (Lipsitz and Parzen 1996).

Lee's variance estimator applied to matched/paired data is calculated as follows. Let $Y_{ip}(t)$ indicate, by values 1 or 0, whether or not the i^{th} individual in the p^{th} pair is at risk at time t.

Define

$$M_0(t) = \sum_{p=1}^{P} \sum_{i=1}^{2} Y_{ip}(t) \exp(\hat{\beta}' z_{ip})$$

and

$$M_{1c}(t) = \sum_{p=1}^{P} \sum_{i=1}^{2} Y_{ip}(t) z_{ipc} \exp(\hat{\beta}' z_{ip})$$

for i=1,2, p=1, ..., P and c= 1, ... C.

Let

$$W_{ipc} = \delta_{ip}\left[z_{ipc} - \frac{M_{1c}(t_{ip})}{M_0(t_{ip})} \right] - \sum_{w=1}^{P} \sum_{q=1}^{2} \frac{\delta Y_{qw}(t_{qw}) \exp(\hat{\beta}' z_{qw})}{M_0(t_{qw})} \left[z_{iqw} - \frac{M_{1c}(t_{qw})}{M_0(t_{qw})} \right].$$

Define the $a_{b,c}$ element of the $c \times c$ adjustment matrix A as

$$a_{b,c} = \sum_{p=1}^{P} \sum_{i=1}^{2} W_{ipb} W_{ipc}$$

resulting in the adjusted estimator $\tilde{V}$ of the covariance matrix of $\hat{\beta}$

$$\tilde{V} = \hat{V} A \hat{V}$$

where $\hat{V}$ is the standard asymptotic covariance matrix estimate of $\hat{\beta}$ from the PH model (section 5.5.1).

Note that the MPH model allows for the matching by correcting the variance estimates post-fit and thus, unlike the CPH model, the matching variables need not be forced into the (final) model used. This might well change the other parameter estimates and in particular the estimates of the regression coefficients for the primary variable of interest.

5.9 Extensions of the Cox PH Model for Paired Survival Data

In such contexts the actual values of the 'matching variables' are unavailable so the approach here is exploit the paired structure of the data .

5.9.1 The Stratified Proportional Hazards Model (SPH)

A further refinement to the PH model for paired survival data is to allow each cluster to define a separate stratum and employ a PH model within each stratum. In the paired framework each pair forms a separate stratum with its own distinct arbitrary "baseline" hazard function.

It is further assumed that the effect of other covariates on the hazard function for a particular pair p is constant across pairs/strata and thus the model can be written as

$$h_{ip}(t \mid z_{ipc}) = h_{0p}(t)\exp(\beta' z_{ipc}), \quad i=1,2,\ p = 1, \ldots, P,\ c=1, \ldots, C$$

i.e. the regression parameters are assumed to be identical for each pair but each pair has a different baseline hazard function. Note, the covariate vector now includes the primary variable and relevant 'unit' covariates (as found relevant by an appropriate variable selection procedure based on this stratified model).

Estimation and hypothesis testing are the same as described previously for the PH model where the partial likelihood function is now given by

$$L_{SPH}(\beta) = \prod_{p=1}^{P} [L_p(\beta)]$$

where the partial likelihood's $L_p(\beta)$ for each pair (stratum) are of the standard PH model form but using only the data from the pth (p=1, .., P) pair.

This model has appeal for paired survival studies when the matching variables are not estimable (e.g. studies involving an individual's eyes) as the dependency structure can be incorporated into the analysis by stratifying by pair. If the matching variables are available (e.g. a matched case-control study) it is not clear whether using the matching variables purely to define a cluster is sensible as this may result in a large degree of inefficiency in terms of estimating the effect of the primary variable of interest. In the matched case-control scenario sensible regression parameter estimates may not be possible due to there only being two observation time points in each pair. It is also not clear what asymptotic properties L_{SPH} has since each pair adds a new nuisance parameter to the model and therefore the number of parameters tends to infinity with the number of pairs.

There is a strong similarity between the Pair Performance model (PP) and the SPH model. Kalbfleisch and Prentice (1980) show that by assuming a proportional hazards model and forming a likelihood based on the 'pair rank' (i.e. by looking at which member failed first) they arrive at the same likelihood as that presented for the PP model (Section 5.3.1 above).

Therefore, despite the fact that the Pair Performance model is essentially a logistic regression approach while the SPH model is an extension of the proportional hazards model, the interpretation of the regression coefficients in both models should be the same (Kalbfleisch and Prentice 1980).

In summary, the stratified PH model incorporates the paired structure of the data by considering the association within each pair as a fixed effect. An alternative and

computationally more elaborate procedure is to introduce a random effect term for each pair that represents the within-pair association. This idea is expanded in the following section.

5.9.2 The Random Effects PH Model

The final extension to the PH model considered here for analysing paired survival data is to introduce a random effect into the model corresponding to each pair. This random 'pair' effect, often termed a *'frailty'*, generates dependency between the survival times of the individuals in a pair. The frailty terms represent covariates that are unaccounted for in the model. In the PH model these are assumed to act multiplicatively on any individual's hazard rate. Survival times of all individuals are assumed to be independent given the frailty values (and any observed covariates).

The term 'frailty' originates from the original use of such a survival regression model where the random effect was considered to represent an (unobservable) measure of an individual's 'proneness to failure' (Vaupel 1979). Some individuals may be more prone than others to a particular disease due to some genetic or environmental conditions that may not be directly measurable (e.g. they come from the same family and are related genetically, or they are exposed to the same environmental conditions). The random effects PH model formulates the overall variability of the survival times of the individuals as having two components: the first is 'natural' variability which is modelled by the hazard function, and the second is variability which is common to individuals in the same pair as modelled by the frailty.

Several authors have addressed the estimation of such frailty models for the covariate-free bivariate case (Clayton 1978, Oakes 1982, Lee and Klein 1989, Hougaard 1984, Hougaard 1986a, Hougaard 1986b). Estimation of the frailty in the presence of covariates has been considered (Clayton and Cuzick 1985) using a modified EM algorithm (Dempster, Laird and Rubin, 1977).

In the context of paired survival studies, a random effect or frailty can be incorporated into a proportional hazards model but in order to 'estimate' the 'frailty' an assumption must be made with regard to its underlying distribution. The most common distribution assumed for the 'frailty' is the gamma distribution (Clayton 1978, Oakes 1982, Clayton and Cuzik 1985, Klein 1992, Nielsen 1992) because it is strictly positive and provides a tractable solution to the problem of parameter estimation (Aalen, 1994). The assumption that the gamma distribution is valid for the random effect has been addressed by Lawless (1982) and Conoway (1990) where they conclude that the gamma distribution is quite flexible in that it provides a wide variety of shapes for the frailty distribution.

Parameter estimation for the Cox PH model with gamma frailties is now described.

5.9.2.1 Estimation for the Gamma Frailty PH Model (GPH)

According to the random effects PH model the hazard for the i^{th} individual in the p^{th} pair, given a frailty ω_p and covariate vector z_{ipc}, is

$$h_{ip}(t \mid z_{ipc}, \omega_p) = h_0(t)\omega_p \exp(\beta' z_{ipc})$$

where β is the regression vector and $h_0(t)$ is the assumed common baseline hazard function as before. The GPH model further assumes that the frailty terms, ω_p, p=1,...., P, are independent and identically distributed observations from a gamma distribution with density function

$$g(\omega) = \frac{\omega^{(1/\theta - 1)} \exp(-\omega/\theta)}{\Gamma[1/\theta]\theta^{1/\theta}}$$

with mean 1 and variance θ. Large values of θ represents strong association within pairs. Oakes (1982) and Klein (1997) have shown that the parameter θ is closely related to Kendall's coefficient of rank correlation τ where the expected value of τ (i.e. the measure of association between pairs) equals $\theta/\theta+2$.

The frailties for each pair can be considered as missing data, and the EM algorithm, a standard approach to parameter estimation in missing data problems, can be used. Alternative approaches for parameter estimation that have been considered for the GPH model are a partial likelihood approach (Klein 1992), a counting process approach (Neilsen 1992) and a penalised likelihood approach (Therneau 1997). For continuity with earlier sections the partial likelihood approach will be presented here.

The partial likelihood approach involves writing the full likelihood in terms of the observed survival times and the unobserved frailties. The expectation of this likelihood with respect to the observable data is carried out in the E-step. A partial likelihood is constructed for estimating the regression parameters using a profile

likelihood technique (Johansen 1983) in the M-step. The algorithm iterates between these two steps until convergence. In order to estimate the parameters θ and β a modified EM algorithm is used. Following the derivation by Klein (1992), the complete data log-likelihood can be written as

$$\ln(L(\theta,\beta,H_0,\omega_1,\ldots,\omega_P)) = L_1(\theta) + L_2(\beta,H_0)$$

where H_0 is the baseline cumulative hazard function for each individual. Let E_p be the number of events in the p^{th} pair then

$$L_1(\theta) = -P[(1/\theta)\ln\theta + \ln\Gamma[1/\theta]] + \sum_{p=1}^{P}\{(1/\theta + E_p - 1)\ln\omega_p - \omega_p/\theta\}$$

and

$$L_2(\beta,H_o) = \sum_{p=1}^{P}\sum_{i=1}^{2}[\delta_{ip}[\beta' z_{ip} + \ln h_o(t_{ip})] - \omega_p H_o(t_{ip})\exp(\beta' z_{ip})].$$

In order to implement the EM algorithm initial estimates for β, θ, and $H_o(t)$ are needed. 'Obvious' initial estimates for β and H_0 can be provided by fitting a Cox PH model ignoring the frailty i.e. letting $\omega_p = 1$ for all p=1, .. ,P, while an initial estimate for θ of 0.25 has been proposed (Klein 1992, Therneau 1997, Hosmer and Lemeshow 1998).

To apply the E step of the algorithm it can be shown (Klein 1992, Therneau 1997) that, conditional on the observed data, the ω_p are independent gamma random variables with shape parameter

$$A_P = [1/\theta + E_p]$$

and scale parameter

$$B_p = [1/\theta + \sum_{i=1}^{2} H_0(t_{ip})\exp(\beta' z_{ip})] \ .$$

Thus

$$E(\omega_p \mid Data) = \frac{A_p}{B_p}$$

and, after some algebra,

$$E(\ln \omega_p \mid Data) = [\psi(A_p) - \ln B_p]$$

where $\psi(\cdot)$ denotes the digamma function.

The resulting expectation of $L(\theta, \beta, H_0, \omega_1, ..., \omega_P)$ given the data and the current values of A_p and B_p is

$$L_1(\theta) = -P[(1/\theta)\ln\theta + \ln\Gamma[1/\theta]] + \sum_{p=1}^{P}\left\{[1/\theta + E_p - 1][\psi(A_p) - \ln B_p] - \frac{A_p}{B_p\theta}\right\}$$

and

$$L_2(\beta, H_o) = \sum_{p=1}^{P} \sum_{i=1}^{2} \left[\delta_{ip} [\beta' z_{ip} + \ln h_o(t_{ip})] - \frac{A_p}{B_p} H_o(t_{ip}) \exp(\beta' z_{ip}) \right]$$

The M-Step of the algorithm involves maximisation of $L_1(\theta)$ and $L_2(\beta, H_o)$ with respect to the unknown parameters θ and β in order to provide updated estimates of β and H_0 (while including the estimated frailty term $\hat{\omega}$) in the E step. Maximisation of θ involves maximisation of $L_1(\theta)$ only while maximisation of $L_2(\beta, H_o)$, which contains the nuisance "parameter" $H_o(t)$, is required to obtain the updated estimate of β.

A non-parametric estimate (Klein 1992) of $H_0(t)$ (in this case the cumulative baseline hazard function which includes the frailty term) is

$$\hat{H}_o(t) = \sum_{t(k) \le t} h_{ko}$$

where

$$h_{ko} = \frac{e_{(k)}}{\sum_{j \in R(t(k))} \hat{\omega}_j \exp(\beta' z_j)}$$

where $t_{(k)}$ is the k^{th} smallest event time, regardless of the pair, $e_{(k)}$ is the number of deaths at $t_{(k)}$ for k=1, ... e, $R(t_{(k)})$ is the risk set of individuals at $t_{(k)}$ and $\hat{\omega}_j$ and z_j are the expected value of the frailty, given the data, and the covariate value associated with the j^{th} individual in the sample respectively.

In summary therefore, initial estimates of β and H_0 are provided using a standard Cox PH model and are subsequently updated using the M step.

Note that an alternative method for maximising $L_2(\beta, H_o)$ is to fit a standard Cox PH model while including $\ln(\hat{\omega})$ as a model covariate with a fixed coefficient of 1 i.e. as an offset term.

This completes the M-step.

The full implementation of the EM algorithm is thus:

Initialisation:

1) provide starting values for β and $H_0(t)$ from, for example, a Cox PH model ignoring the frailty term

2) provide a starting value for θ, the variance parameter of the gamma frailty, for example θ=0.25

Estimation Step:

using the current values of β, θ and $H_0(t)$ compute A_p and B_p and hence $\hat{\omega}_p$.

Maximisation Step:

i) Update the estimate of θ based on $L_1(\theta)$,

ii) Update the estimate of β and $H_o(t)$ using $L_2(\beta, H_o)$

Iterate between the Estimation and Maximisation steps until convergence is obtained.

Significance tests for regression coefficients can be performed using Wald tests based on the observed information matrix for $(\hat{\theta}, \hat{\beta})$ (Klein 1992, Anderson, Klein et al 1997), or by a likelihood ratio test (Nielsen, Gill and Andersen 1992) based on comparing the partial log-likelihood from fitting the 'full' model to one where the covariate has been excluded.

Several methods for assessing the significance of the frailty parameter θ have been suggested including a score test (Klein and Moeschberger 1997) a likelihood ratio test (Nielsen, Gill and Andersen 1992) and a bootstrap procedure (Therneau 1997). In particular, under the null hypothesis that θ=0, the likelihood ratio test statistic is $2\left[L(\hat{\theta},\hat{\beta}) - L(0,\hat{\beta})\right]$ which has an approximate chi-square distribution with 1 degree of freedom.

The gamma frailty proportional hazards model provides a mechanism for modelling the dependency structure in paired data by estimating and modelling the unmeasureable 'covariates' that represent the dependency structure in a pair through a proportional hazards model (Kieding et al 1997). The model is very attractive for paired studies where the dependency is often unobservable (e.g. familial studies). One obvious question for such a model (which will be addressed later) is how effective a frailty model is if one incorporates known matching variables in the model.

5.10 Model Summary

Each of the models presented above are distinctly different in their approach to analysing dependent survival data where some are more suited to matched studies and others to paired studies. The best model chosen for any specific analysis will be likely to be influenced by the nature or design of the study and therefore a summary of all the models considered in this chapter is given in Table 5.1 indicating the role of the matched variables and unmatched covariates in each modelling approach.

Examples of each of the regression techniques described in this chapter are now presented for the Melanoma and Dental datasets.

5.11 Examples

5.11.1 Melanoma Tumour Group Data

Survival for both Multiple and Single Melanoma is likely to be influenced by a number of covariates such as tumour thickness, sex, ulceration etc. As mentioned in Chapter 2, individuals were matched by Sex, Age, Tumour Thickness and Tumour Site, while the unmatched covariates, or potential prognostic indicators, were Level of Invasion and Ulceration status.

Table 5.1 Model Summary.

Matched Studies	***Matching Variables***		***Unmatched Covariates***	
Model	**Role**	**Criteria for Inclusion**	**Role**	**Criteria for Inclusion**
Pair Performance Model	difference in values of non-perfectly matched covariates in a pair used to adjust for Pair Performance comparison	Forced into Model	difference in values of unmatched covariates in a pair used to adjust for comparison of primary variable	Variable Selection Procedure
Conditional PH Model	used to control for dependency in the comparison of primary variable	Forced Into Model	used to adjust for comparison of primary variable	Variable Selection Procedure
Marginal PH Model	used to control for dependency in the comparison of primary variable and later used to "correct" variance estimates	Variable Selection Procedure	used to adjust for comparison of primary variable and later used to "correct" variance estimates	Variable Selection Procedure

Paired Studies	***Matching Variables***		***Hidden Covariates***		***Unit Covariates***	
Model	**Role**	**Criteria for Inclusion**	**Role**	**Criteria for Inclusion**	**Role**	**Criteria for Inclusion**
Stratified PH Model	used to define strata	Not Available	Ignored	Ignored	used to adjust for comparison of primary variable	Variable Selection Procedure
Gamma PH Model	used to control for dependency in the comparison of primary variable	Not Available	used to control for dependency in the comparison of primary variable	Implicitly forced into model	used to adjust for comparison of primary variable	Variable Selection Procedure

From the plots of the data presented in Chapter 3 there is a suggestion of slightly improved survival for Multiple Melanoma sufferers while tumour thickness, ulcerated tumours and a patient being male all suggest a detrimental effect on an individual's survival prospects.

In order to assess the effect of the primary variable (i.e. Single/Multiple Melanoma) on survival while adjusting for both the dependency structure produced by the matching and any imbalance in the covariates (both matched and unmatched) all of the models presented earlier in this chapter were fitted to this data.

As outlined above, the matching covariates have a distinct role in each of the modelling approaches while any unmatched covariates may be included in the final model if they proved useful on the basis of a variable selection approach. A summary of the steps taken and the results of the final fitted model for each approach are now presented.

5.11.1.1 Pair Performance Model (PP)

The Simple Binomial test applied to this data (see Chapter 4) suggested no significant advantage in survival for Multiple Melanoma sufferers over Single Melanoma sufferers (p=0.21) although there is a larger proportion of pairs (i.e. 15 against 8) where the Multiple Melanoma sufferer survived longer than the Single. This test ignored all covariates except the primary variable (i.e. Tumour group).

In fitting a Pair Performance model the role of the covariates was as follows:

1. ***Tumour Thickness*** and ***Age*** were both included as neither was perfectly matched;
2. Site and Sex were redundant as both were perfectly matched;
3. No other covariates or two-way interactions were deemed necessary for inclusion in the final model on the basis of both forward and backward stepwise variable selection procedures.

The results for the final model are given in Table 5.2.

Table 5.2 Results of a Pair Performance Model for the Melanoma Data.

Variable	**Regression Coefficient** $\hat{\beta}$	**ese**$(\hat{\beta})$	**Exp**$(\hat{\beta})$ **(95% C.I.)**	**p-value**
Tumour Group (Multiple/Single)	0.48	0.46	1.62 (0.6 - 4.2)	0.29
Tumour Thickness	0.19	0.37	1.22 (0.6 - 2.5)	0.59
Age	-0.10	0.01	0.90 (0.8 - 1.1)	0.24

Again there is no significant difference in terms of pair performance for Multiple Melanoma sufferers over Single Melanoma sufferers while adjusting for any imbalance in the matching with similar p-values. The resulting effect of tumour group is again in favour of Multiple Melanoma having (marginally) better survival.

The Tumour thickness and Age regression coefficients are both used to adjust the primary variable comparison for any imbalance present in the matching of these two variables. Since both are non-significant it could be argued that the quality of matching (to the nearest mm and 10 years respectively) was adequate to remove the actual (known) effects of Tumour Thickness and Age on melanoma survival.

Note, the above model used 23 pairs of observation (15 Multiple, 8 Single) while no decision in terms of pair performance could be made for the other 85 pairs (79% of the available data) due to the high degree of censoring in both tumour groups.

5.11.1.2 Conditional Proportional Hazards (CPH) Model

A CPH model was fitted which ignored the specific pairing but included all the matching variables and whichever of the unmatched covariates were found to be significant prognostic factors by means of variable selection techniques.

The following steps indicate how the CPH model was fitted for this context :

1. The primary variable (i.e. ***Tumour group***) and all matching variables (i.e. ***Sex***, ***Age***, ***Site*** and ***Tumour Thickness***) were forced in the model;
2. ***Ulceration*** was the only unmatched covariate found necessary for inclusion in the final model on the basis of both forward and backward stepwise variable selection procedures;

3. No other covariates or two-way interactions were deemed necessary for inclusion in the final model on the basis of the stepwise variable selection procedures.

The results of this approach in terms of the final model are presented in Table 5.3.

Table 5.3. Results of the final CPH Model for the Melanoma Data.

Variable	**Regression Coefficient** $\hat{\beta}$	**ese**$(\hat{\beta})$	**Exp**$(\hat{\beta})$ **(95% C.I.)**	**p-value**
Tumour Group (Multiple/ Single)	0.39	0.30	1.48 (0.8 - 2.7)	0.19
Sex (Female/Male)	1.01	0.30	2.76 (1.5 - 4.9)	<0.001
Age	0.01	0.01	1.01 (0.9- 1.03)	0.22
Tumour Site (Extremity/Axial)	0.51	0.33	1.67 (0.8 - 3.2)	0.12
Tumour Thickness	0.07	0.02	1.08 (1.0 - 1.1)	<0.001
Ulceration (Non-ulcerated/Ulcerated)	1.02	0.28	2.77 (1.6 - 4.8)	<0.001

On the basis of these results one can see that survival is significantly poorer

(i) the *greater* the *Tumour Thickness*;

(ii) for *ulcerated* over *non-ulcerated* tumours;

and

(iii) for *males* over *females*.

However, there was still no clear evidence of a significant difference in survival between Multiple and Single melanoma sufferers but the p-value was marginally reduced, this time to 0.19 compared to 0.29 from the Pair Performance model and 0.21 for the Simple Binomial Test.

5.11.1.3 Marginal Proportional Hazards (MPH) Model

As described earlier the MPH model is identical in approach to fitting a CPH model except that the MPH model accounts for the matching by adjusting the variance of the regression coefficients post-fit and hence the matching variables need not be forced into the final model. However, in order to compare the MPH model to the CPH model in terms of the effect of using the adjusted covariance matrix, the same estimates as presented above in 5.10.1.2 was fitted and the results are given in Table 5.4 below.

Table 5.4 Results of a Marginal Proportional Hazards Model for the Melanoma Data.

Variable	**Regression Coefficient** $\hat{\beta}$	**ese**$(\hat{\beta})$	**Exp**$(\hat{\beta})$ **(95% C.I.)**	**p-value**
Tumour Group (Multiple/Single)	0.39	0.28	1.48 (0.9 - 2.6)	0.16
Sex (Female/Male)	1.01	0.32	2.76 (1.5 - 5.2)	<0.001
Age	0.01	0.01	1.01 (0.9- 1.04)	0.32
Tumour Site (Extremity/Axial)	0.51	0.31	1.67 (0.9 - 3.1)	0.10
Tumour Thickness	0.07	0.02	1.08 (1.0 - 1.1)	<0.001
Ulceration (Non-ulcerated/Ulcerated)	1.02	0.35	2.77 (1.4 - 5.5)	<0.001

The estimated regression coefficients are of course identical but there is a slight reduction in the estimated standard error (ese) of the primary variable and several other covariates reflecting the dependency within the data. The results for both models suggest that matching by Age and Tumour Site was not fully justified on the basis of this data.

When fitting the MPH model there is no necessity to force the matching variables into the final model and therefore the primary variable is the only variable that needs to be forced into the model while stepwise procedures determine the inclusion or not of matching variables and unmatched covariates alike.

The strategy for variable inclusion in this approach for the MPH model is as follows:

1. The primary variable (i.e. ***Tumour group***) was forced in the model;
2. ***Tumour Thickness*** and ***Sex*** were the only matching variables and ***Ulceration*** the only unmatched covariate found significant for inclusion in the final model on the basis of a stepwise variable selection procedure;
3. No other covariates or two-way interactions were found to significantly influence survival and thus no other terms were included in the final model.

On the basis of this strategy, the final model resulting from fitting a MPH model to the Melanoma data is given in Table 5.5.

Table 5.5 Results of a Final Marginal Proportional Hazards Model for the Melanoma Data.

Variable	**Regression Coefficient** $\hat{\beta}$	**ese**$(\hat{\beta})$	**Exp**$(\hat{\beta})$ **(95% C.I.)**	**p-value**
Tumour Group (Multiple/Single)	0.38	0.28	1.47 (0.9 - 2.5)	0.17
Sex (Female/Male)	1.22	0.30	3.39 (1.9 - 6.1)	<0.001
Tumour Thickness	0.07	0.02	1.08 (1.0 - 1.1)	<0.001
Ulceration (Non-ulcerated/Ulcerated)	1.17	0.30	3.23 (1.8 - 5.8)	<0.001

This final model had a slightly lower estimated standard error (ese) for the estimated regression coefficient of Tumour Group and consequently a lower p-value in comparison to the final CPH model.

When comparing the final MPH model to that presented in Table 5.4 the estimated regression coefficients for the variables common to the two models have obviously changed with the estimated effect on survival of Ulceration and Sex being increased and their estimated standard errors being reduced.

5.11.1.4 Stratified Proportional Hazards (SPH) Model

The SPH model seems more suited to paired studies but for completeness the approach will be applied to the Melanoma data for comparison purposes.

Initially, a SPH model was fitted with the "non-perfectly matched" variables forced into the model to adjust for any imbalance in these variables that may exist in each pair.

The results of this modelling approach are given in Table 5.6.

Table 5.6 Results of a Stratified Proportional Hazards Analysis for the Melanoma Data.

Variable	**Regression Coefficient** $\hat{\beta}$	**ese**$(\hat{\beta})$	**Exp**$(\hat{\beta})$ **(95% C.I.)**	**p-value**
Tumour Group (Multiple/Single)	0.22	0.35	1.26 (0.6 - 2.5)	0.52
Tumour Thickness	-0.06	0.11	0.94 (0.8 - 1.2)	0.59
Age	0.04	0.05	1.04 (0.9 - 1.2)	0.44

As explained in section 5.9.1 the results are similar to those presented for the PP model in terms of magnitude of the regression parameters and their overall effect on survival (see Table 5.2 for comparison). It appears that adjusting for both Tumour Thickness and Age is unnecessary due to the high quality of matching (though not perfect) in these variables.

As presented earlier the SPH model controls for the dependency of the data by fitting separate baseline hazard functions for each pair. In effect, the matching variables may not be needed as covariates in this model. A further SPH model fitted in this section

therefore used a variable selection procedure to determine the inclusion of all matching variables and unmatched covariates.

A summary of fitting this strategy through an SPH model in terms of the covariates chosen and their respective role was :

1. ***Tumour Group***, the primary variable, was forced into the model:
2. No other matching variables or unmatched covariates or their two-way interactions were found necessary for inclusion in the final model.

The final model resulting from fitting this strategy to selection of an SPH model is given in Table 5.7 below.

Table 5.7 Results of a Stratified Proportional Hazards Analysis for the Melanoma Data.

Variable	**Regression Coefficient** $\hat{\beta}$	**ese**$(\hat{\beta})$	**Exp**$(\hat{\beta})$ **(95% C.I.)**	**p-value**
Tumour Group (Multiple/Single)	0.30	0.34	1.35 (0.7 - 2.6)	0.38

Once again there is no strong suggestion of a significant tumour group difference when stratifying by pair and indeed, if anything, the estimated effect of Tumour Group seems somewhat diluted by this approach compared to the other approaches (Note using Tumour Group alone marginally 'increases' the effect of the Tumour Group coefficient of Table 5.6).

5.11.1.5 Random Effects (GPH) Model

The gamma frailty model described in section 5.9.2.1 was fitted to the Melanoma data. For continuity, the ' full' CPH model used in 5.10.1.2 was fitted with a gamma frailty term and the results are given in table 5.8 below.

Table 5.8 Results of a Random Effects Proportional Hazards Analysis for the Melanoma Data.

Variable	**Regression Coefficient** $\hat{\beta}$	**ese**$(\hat{\beta})$	**Exp**$(\hat{\beta})$ **(95% C.I.)**	**p value**
Tumour Group (Multiple/Single)	0.43	0.30	1.54 (0.8 - 2.8)	0.16
Sex (Female/Male)	1.03	0.32	2.81 (1.5 - 5.3)	0.001
Age	0.01	0.01	1.01 (0.9- 1.04)	0.19
Tumour Site (Extremity/Axial)	0.54	0.35	1.71 (0.9 - 3.4)	0.13
Tumour Thickness	0.08	0.02	1.08 (1.0 - 1.1)	0.001
Ulceration (Non-ulcerated/Ulcerated)	1.01	0.30	2.73 (1.5 - 4.9)	<0.001

There was a slight reduction in the estimated standard error of the estimated Tumour Group regression coefficient when compared to the CPH model, reflecting the dependency of the data. However, once again, this reduction did not change the overall conclusion in terms of identifying a significance difference in the survival prospects of individuals in the two tumour groups. Including the frailty term added little to the model also and indeed would be thought of as unnecessary on the basis of

the Likelihood ratio test of the significance of the frailty parameter ($\hat{\theta}$=0.32, ese($\hat{\theta}$)=0.31, p=0.30, see section 5.9.2.1 for details of the test).

Since the preferred strategy for GPH models would be to only include matching variables if necessary (i.e. any matching variables included should account for aspects of survival not accounted for by the frailty terms), a further fit of the GPH model was obtained allowing matching variables and unmatched covariates into the final model only through 'significance' in a variable selection procedure. A summary of the resulting GPH model is given in Table 5.9.

Table 5.9 Results of a Final Random Effects Proportional Hazards Model for the Melanoma Data

Variable	**Regression Coefficient** $\hat{\beta}$	**ese($\hat{\beta}$)**	**Exp($\hat{\beta}$) (95% C.I.)**	**p value**
Tumour Group (Multiple/Single)	0.41	0.30	1.51 (0.8 - 2.7)	0.17
Sex (Female/Male)	1.23	0.29	3.44 (1.9 - 6.2)	0.001
Tumour Thickness	0.08	0.02	1.08 (1.0 - 1.1)	0.001
Ulceration (Non-ulcerated/Ulcerated)	1.14	0.29	3.14 (1.8 - 5.6)	<0.001

The results are similar in terms of variables included and indeed the estimated coefficients to those found in the second (and preferred) MPH model. Again the effects of Sex and Ulceration are increased relative to the model fitted in Table 5.8

while the effect of Tumour Group is also slightly increased (from 1.47 in the MPH model to 1.51 here) although the level of significance remains around 0.17.

5.11.2 Melanoma Data Summary

After fitting the models described in this chapter there was no strong suggestion from any of these that there was a significant difference in survival for Multiple and Single Melanoma sufferers while adjusting for the matched structure of the data and any imbalances that existed in either matched variables or unmatched covariates. The only variables significantly effecting the survival of melanoma patients were the Tumour Thickness, Sex and Ulceration status of the melanoma sufferer.

A summary of the results of each final model in terms of estimating the effect of Tumour Group on survival is given in Table 5.10.

Table 5.10 Tumour Group Effect for each Model fitted to the Melanoma Data.

Model	**Regression Coefficient** $\hat{\beta}$	**ese**$(\hat{\beta})$	**Exp**$(\hat{\beta})$ **(95% C.I.)**	**p-value**
Pair Performance	0.48	0.46	1.62 (0.6 - 4.2)	0.29
Conditional PH	0.39	0.30	1.48 (0.8 - 2.7)	0.19
Marginal PH	0.38	0.28	1.47 (0.9 - 2.5)	0.17
Stratified PH	0.30	0.34	1.35 (0.7 - 2.6)	0.38
Gamma Frailty PH	0.41	0.30	1.51 (0.8 - 2.7)	0.17

The Pair Performance (PP) model provided the largest Tumour Group estimated standard error. This larger variation is probably due to the considerable number of doubly censored pairs excluded from the PP analysis.

In general, all of the proportional hazards models gave similar estimates of the regression parameter for Tumour Group except the SPH model which was quite a bit smaller as well as having a larger ese. This is probably due to the SPH model being not particularly suitable if any (sensible) matching variables are available for inclusion, as in this example with Sex and Tumour Thickness.

The Marginal PH and Gamma Frailty PH models provided the same final model in terms of 'included covariates'. The Marginal PH approach resulted in a slight decrease in the ese associated with the Tumour Group estimated regression coefficient compared to the Conditional PH model, highlighting the dependency structure in the data. The Gamma Frailty PH suggested that no further matching variable was needed to adjust for the dependency structure of the data when the matching variables were already in the model which could be entirely due to over-fitting through the frailty terms.

The estimated regression coefficients for each of the other covariates for all models considered here are given in Table 5.11.

Once again the proprotional hazards models are similar in terms of the magnitude and interpretation of the estimated regression coefficients of the covariates. The Level of Invasion of the tumour did not prove significant in any of the models proposed.

Table 5.11. Estimated regression coefficients (with estimated standard error in brackets) for each matching variable and unmatched covariate included in final preferred model of each type.

	Variable	PP	CPH	MPH	SPH	GPH
Matching Variables	Sex (Male/ Female)	–	1.01 (0.30)	1.22 (0.30)	–	1.23 (0.29)
	Age	-0.10 (0.01)	0.01 (0.01)	–	–	–
	Tumour Thickness	0.19 (0.37)	0.07 (0.02)	0.07 (0.02)	–	0.08 (0.02)
	Tumour Site (Axial/ Extremity)	–	0.51 (0.33)	–	–	–
Unmatched Covariates	Ulceration (Ulcerated/ Non-ulcerated)	–	1.02 (0.28)	1.17 (0.30)	–	1.14 (0.29)
	Level of Invasion	–	–	–	–	–
	Deprivation Category	–	–	–	–	–

Note : a blank space signifies that this variable was not included in the final model.

Remembering that the sample size is moderate (108 pairs with a high degree of censoring) one is left with the conclusion that, although, on the basis of the data and any of the models used, there is no significant difference in survival between Single and Multiple Melanoma sufferers there is a lingering suspicion that patients with Multiple Melanomas have moderately better survival prospects than those with Single Melanoma which may well be identified in a larger study. However the 108 Multiple Melanoma patients used in this study were the only such cases in Scotland between 1979 and 1997 so a larger catchment area or a much larger time interval would have to be used to achieve a sufficiently large enough sample size. Whether this would be

worthwhile either in terms of including other 'components of variability' (across regions for example) or identifying the moderate improvement in survival for Multiple Melanoma sufferers is an open question at present.

5.11.3 Dental Data

As discussed in Chapter 2, the Orthodontic study aims to compare two different cements which may be used for bonding orthodontic brackets to teeth. The main interest in the study is the effect of the primary variable, Cement type, on the time to breakage of the bracket. As both cements are used on each individual the study is a paired survival study and hence no matching variables are available for inclusion in the analysis. Several "unmatched" or "unit" covariates are available, namely a patient's Sex, Age and Malocclusion Type.

The initial impression of the performance of the two cement bonds (see Chapter 3) suggests a slightly improved performance of the Test bond over the Control bond, in particular for older patients, males and patients with improved Malocclusion status.

In order to assess the effect of the primary variable (i.e. Cement Type) on survival while adjusting for both the dependency structure and any imbalance in the covariates each of the models presented earlier in this chapter were fitted.

5.11.3.1 The Pair Performance (PP) Model

As this is a paired study all 'difference' covariates will be identically zero and will be redundant in any Pair Performance model. The only approach available is to fit a PP model estimating the primary variable only.

Recall from Chapter 4, there are 23 pairs where a definite pair performance score is obtained (14 Test material, 9 Control material) while no decision in terms of pair performance could be made for the other 22 pairs (48% of the available data) due to the high degree of censoring in both treatment groups. The Simple Binomial test applied to this data (see Chapter 4) suggested no significant disadvantage (p=0.41) in bracket failure for patients treated with the Test material over those treated with the Control material.

The results of fitting the Pair Performance model are given in table 5.12 below.

Table 5.12 Results of the final Pair Performance Model for the Dental Data.

Variable	**Regression Coefficient** $\hat{\beta}$	**ese**$(\hat{\beta})$	**Exp**$(\hat{\beta})$ **(95% C.I.)**	**p-value**
Cement Type (Test/Control)	0.44	0.42	1.55 (0.7 - 3.6)	0.31

The estimated regression coefficient for Cement Type is precisely ln(14/9) in agreement with the Simple Binomial test. Note that the Simple Binomial Test is an exact test, hence the discrepancy in the p-values. As with the Simple Binomial test

however, there is again no strong suggestion of a significant difference in the distribution of bracket failure times between the two cement types.

5.11.3.2 Conditional Proportional Hazards (CPH) Model

Due to the paired nature of this study the CPH model is not appropriate as no matching variables are available for inclusion in the final model. However, as an exercise to compare the perfromances of the models described in this chapter a CPH model was fitted based on a variable selection procedure applied to those available "unit" covariates (i.e. Age, Sex and Malocclusion type).

A summary of the fitted "CPH" model is as follows :

1. The primary variable (i.e. ***Cement Type***) was forced in the model;
2. ***Age*** was the only unmatched covariate found essential for inclusion in the final model on the basis of forward and backward stepwise variable selection procedures;
3. No other covariates or two-way interactions were found significant for inclusion in the final model.

The results of this approach in terms of the final model are presented in table 5.13.

Table 5.13. Results of the final "CPH" Model for the Dental Data

Variable	**Regression Coefficient** $\hat{\beta}$	**ese**$(\hat{\beta})$	**Exp**$(\hat{\beta})$ **(95% C.I.)**	**p-value**
Cement Type (Test/Control)	0.05	0.31	1.05 (0.6 - 1.9)	0.87
Age	-0.24	0.07	0.78 (0.7 - 0.9)	0.001

On the basis of these results one can see that the risk of breakage decreases with increasing age and that there was no significant difference in survival for cement type.

5.11.3.3 Marginal Proportional Hazards (MPH) Model

The estimated standard errors (ese) of the regression coefficients estimates from the CPH model fitted above are likely to be incorrect as the paired structure of the data was ignored. The next approach taken therefore was to fit an MPH model in order to correct the standard errors of the regression coefficients by accounting for the paired structure.

Model selection was identical to that described for the CPH model where:

1. The primary variable (i.e. ***Cement Type***) was forced in the model;

2. ***Age*** was the only unmatched covariate found essential for inclusion in the final model on the basis of stepwise variable selection procedures;

3. No other covariates or two-way interactions were found essential for inclusion in the final model on the basis of stepwise variable selection procedures.

The results of the final MPH model fitted are given in Table 5.14.

Table 5.14. Results of the final CPH Model for the Dental Data.

Variable	**Regression Coefficient** $\hat{\beta}$	**ese**$(\hat{\beta})$	**Exp**$(\hat{\beta})$ **(95% C.I.)**	**p-value**
Cement Type (Test/Control)	0.05	0.26	1.05 (0.6 - 1.7)	0.86
Age	-0.24	0.08	0.78 (0.7 - 0.9)	0.005

The effect of fitting the MPH model was to reduce the ese of the primary variable regression coefficient but did not change the overall conclusion in terms of identifying a significant difference in the time to bracket failure between the two cement types.

5.11.3.4 Stratified Proportional Hazards (SPH) Model

A SPH model was fitted to the Dental data where, unlike the Melanoma data example, no matching variables are available for inclusion. A summary of the strategy for fitting such a model is as follows :

1. ***Cement Type***, the primary variable was forced into the model:
2. No other covariates or two-way interactions were found essential for inclusion in the final model on the basis of stepwise variable selection procedures.

The strategy for fitting the GPH model was as follows:

1. The primary variable (i.e. ***Cement Type***) was forced in the model;
2. ***Age*** was the only unmatched covariate found essential for inclusion in the final model on the basis of stepwise variable selection procedures;
3. No other covariates or two-way interactions were found essential for inclusion in the final model on the basis of stepwise variable selection procedures.

The results of the final GPH model fitted are given in Table 5.16.

Table 5.16. Results of the final GPH Model for the Dental Data.

Variable	**Regression Coefficient** $\hat{\beta}$	**ese**($\hat{\beta}$)	**Exp**($\hat{\beta}$) **(95% C.I.)**	**p-value**
Cement Type (Test/Control)	0.11	0.32	1.11 (0.6 - 2.1)	0.74
Age	-0.27	0.09	0.78 (0.6 - 0.9)	0.002

There was a slight increase in the ese of the Cement Type estimated regression coefficient while the respective p-value decreased when compared to the MPH model. There was a significant contribution to the overall model fit by including the frailty term ($\hat{\theta}$=0.87, ese($\hat{\theta}$)=0.37, Likelihood Ratio Test, p=0.02) suggesting that a random effect is needed in the model and that an analysis based on a 'simple' stratification

(i.e. the SPH) model may not be adequate. This may also explain the considerable difference in the p-values for the Cement Type effect between the SPH and GPH models. Adding this frailty term did not however change the conclusion of the study that there is no real suggestion of a significant difference in the breaking times for the test and control cement compounds.

5.11.4 Dental Data Summary

After fitting the models described in this chapter there was ***no strong suggestion*** of a significant difference in breakage time for the ***Test and Control*** cement compounds while adjusting for the paired structure of the data and any imbalances that existed in any of the unmatched covariates. The only variable significantly affecting the time to breakage of the bracket was the individual's ***age***. The study suggests that the Test material can be considered 'equally' as good as the Control cement in terms of time to bracket failure, and given that the Test cement has less side effects (e.g. reduction in decalcification) it would appear to be the preferred cement.

A summary of the results of each model in terms of assessing the effect of Cement Type on time to breakage is given in Table 5.17.

Table 5.17 Cement Type Effect for each Model fitted to the Dental Data.

Model	Regression Coefficient $\hat{\beta}$	ese($\hat{\beta}$)	Exp($\hat{\beta}$) (95% C.I.)	p-value
Pair Performance Model	0.44	0.42	1.55 (0.7 - 3.6)	0.31
Conditional PH Model	0.05	0.31	1.05 (0.6 - 1.9)	0.87
Marginal PH Model	0.05	0.26	1.05 (0.6 - 1.7)	0.86
Stratified PH Model	0.33	0.36	1.38 (0.7 - 2.8)	0.37
Gamma Frailty PH Model	0.11	0.32	1.11 (0.6 - 2.1)	0.74

The Pair Performance and SPH model are quite similar in terms of their estimated regression parameter both without correcting for the effect of Age. The results of fitting the PP, SPH and GPH models seems to suggest that incorporating a "pair-specific" effect has more of an effect on the estimated regression coefficient when analysing paired rather than matched survival data. Recall that the MPH procedure provides identical estimated regression coefficients as the CPH model and differs only in terms of how the respective estimated standard errors are obtained. Indeed, the benefit of fitting an MPH model to control for the dependency structure of the data (when compared to the clearly inappropriate CPH model) was clear as there was a considerable reduction in the estimated standard error for Cement Type under this model. The GPH model suggested that a frailty term was needed for the analysis and hence an SPH model may not be suitable, further highlighting the dependency structure due to the paired nature of the data.

The effects of the other covariates for each of the models are given in Table 5.18 below.

Table 5.18. Estimated Regression parameter (with ese in brackets) for each matching variable and unmatched covariate when included in final model.

Variable		PP	CPH	MPH	SPH	GPH
Unmatched Covariates	Sex	–	–	–	–	–
	Malocculsion Type	–	–	–	–	–
	Age	–	-0.24 (0.07)	-0.24 (0.08)	–	-0.27 (0.09)

Note : a blank space signifies that this variable was not included in the final model.

The CPH, MPH and GPH models are similar in terms of the magnitude and interpretation of Age, the only covariate significantly associated with the time to bracket failure. Neither of the covariates representing an individual's Sex or Malocculsion Type were deemed necessary for inclusion in any of the models proposed.

5.12 Assessing Goodness-of-Fit

A brief discussion is now given to methods for checking the model assumptions and goodness-of-fit for the various models presented in this chapter.

The Pair Performance (PP) model is a binary logistic regression approach to analysing dependent survival data. The logistic model is well understood and a good reference for logistic modelling is Van Houwelingen (1988). Several types of residuals can be

obtained from binary logistic model fits (Hosmer and Lemeshow 1989, Collett 1991). These residuals may be used to assess the influence of individual pairs on the fit or to assess how each 'difference' covariate may be transformed to linearity on the log odds scale. Assessment of fit for the PP model is generally determined using one of several global tests for goodness of fit (Hosmer and Lemeshow 1989).

The results of fitting the Goodness-of-Fit tests outlined earlier to the Melanoma Data gave the following results:

Table 5.19. Goodness-of-Fit Tests for the Melanoma Data.

Method	Chi-Square	DF	P
Pearson	22.3	20	0.32
Deviance	27.6	20	0.12
Hosmer-Lemeshow	8.8	8	0.36

suggesting that the PP model is an adequate fit for the data. While as the Dental data contained no terms except for the primary variable no goodness-of-fit tests are needed.

The assumptions for the Proportional Hazards (PH) model can be grouped into those relating to the functional form of the regression components (i.e. is it adequately described by a linear function) and those relating to proportional hazards.

Residuals in classical regression models are the differences between the observed value and that predicted by the model. In survival analysis however, due to censoring, the definition of a residual is not as clear, and this has led to the development of several different types of residuals each tailored to checking a specific assumption underlying the proportional hazards model.

In summary, these are

1. Martingale residuals (Barlow and Prentice 1988) defined as

$$M_{ip} = \delta_{ip} - \hat{H}_0(t_{ip}) \exp(\beta' z_{ipc})$$

which represent the difference, for an individual i, between the observed number of events and the expected number of events given the model. Their primary use is to test the functional form of each covariate.

2. Schoenfeld residuals (Schoenfeld, 1982) which are defined as

$$SF_{ip} = z_{ip} - \bar{z}(t_{ip})$$

where $\bar{z}(t)$ is the mean of z weighted by $\exp(\hat{\beta}' z)$ for all those individuals still in the risk set at time t. These are used as a method of formally testing the proportional hazards assumption (Grambsch and Therneau 1994, Pettitt and Bin Daud 1990) for both categorical and continuous variables. If the proportional hazards assumption is valid, a plot and 'lowess smooth' of the Schoenfeld residuals against time (for each covariate) should have non-zero slope. This forms the basis of the formal test provided by Grambsch and Therneau (1994).

3. Deviance and score residuals (Schoenfeld 1982), refinements of the Martingale and Schoenfeld residuals respectively, may be used to assess which observations are

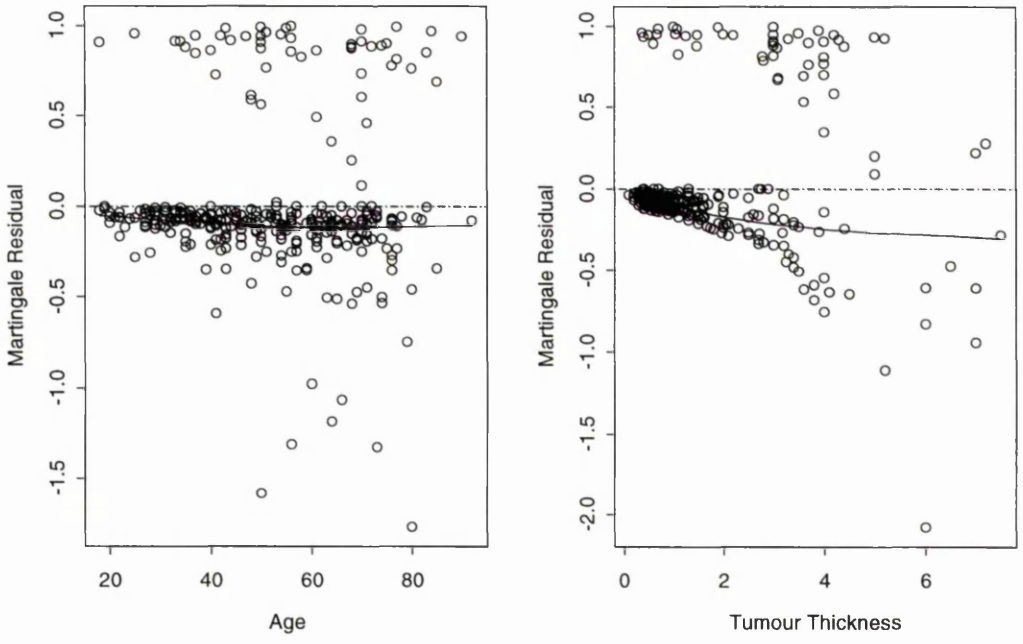
not well-fitted by the model as well as to measure the influence of individual observations (Cain and Lange 1984, Reid and Crepeau 1985, Storer and Crowley 1985).

In addition to the residual based tests, there are a number of graphical and analytical methods for assessing the proportional hazards assumption (Harrell 1986, Muenz 1983, Arjas 1988, Gore 1984). The most basic assessment of the proportional hazards assumption for a binary covariate is provided by a plot of $\log(-\log(\hat{S}(t)))$ against time for each level separately which will yield 'parallel curves' if the hazards are proportional across the two levels of the covariate. Recently an additional test for assessing the proportional hazards assumption (for both categorical and continuous variables) with respect to being 'non-constant' over time has been proposed (Quantin, et al 1996) which amounts to including an interaction term representing the covariate being tested and log time. This is formally assessed using a Wald test.

5.12.1 Melanoma Data

When considering the Melanoma data, the linearity assumption of each continuous covariate in the final model was assessed by scatterplots of the Martingale residuals against each covariate in turn (Figure 5.1). All of these plots suggest that the assumption of linearity (in the log hazard function) is reasonable for all of the covariates considered.

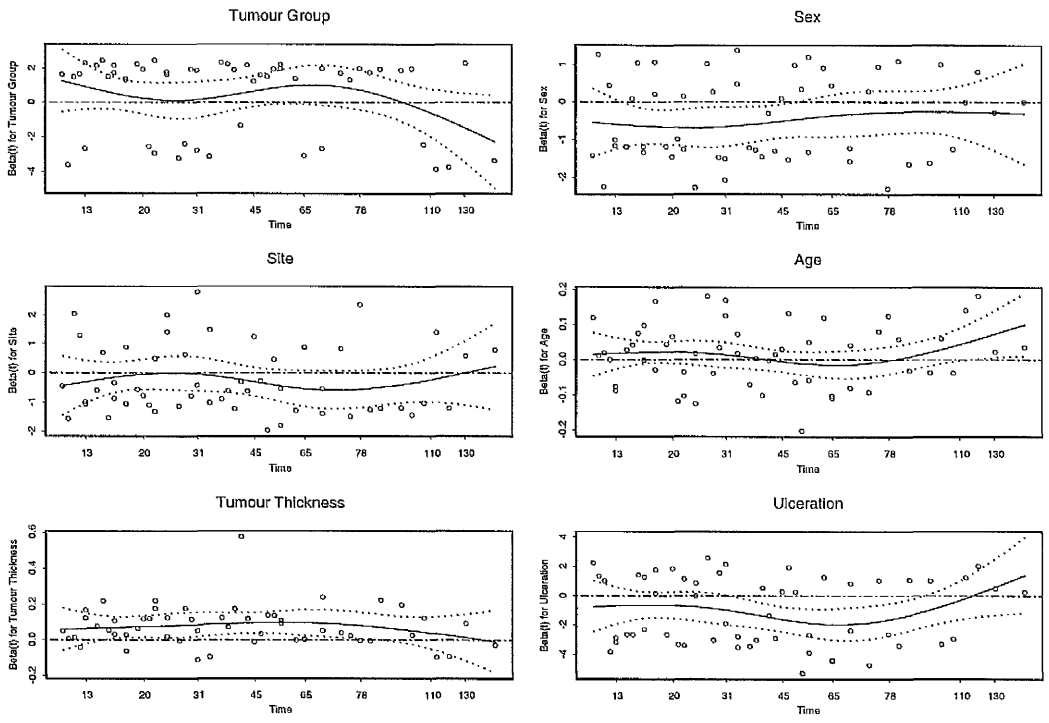
Figure 5.1. Plot of Martingale Residuals for Age and Tumour Thickness in the final CPH model for the Melanoma Data.



In order to assess the proportional hazards assumption a plot of the Schoenfeld residuals plot (with 'lowess smooth') for each covariate over time (Figure 5.2) is needed. If the proportional hazards assumption is valid, a plot of $\beta(t)$ for each covariate would be horizontal. By virtue of these plots, the proportional hazards assumption seems plausible for all the variables except perhaps the primary variable Tumour Group.

Figure 5.2

Plot of Schoenfeld Residuals over Time for each covariate included in the final CPH model for the Melanoma Data.



In order to formally test the proportional hazards assumption the p-values for the Grambsch and Therneau (1994) and Quantin test (1990) are as follows:

Table 5.20. Grambsch and Therneau and Quantin test p-values for the Melanoma Data.

	Grambsch and Therneau	Quantin
Tumour Group	0.28	0.36
Sex	0.36	0.21
Age	0.71	0.62
Tumour Site	0.76	0.71
Tumour Thickness	0.76	0.58
Ulceration	0.93	0.32

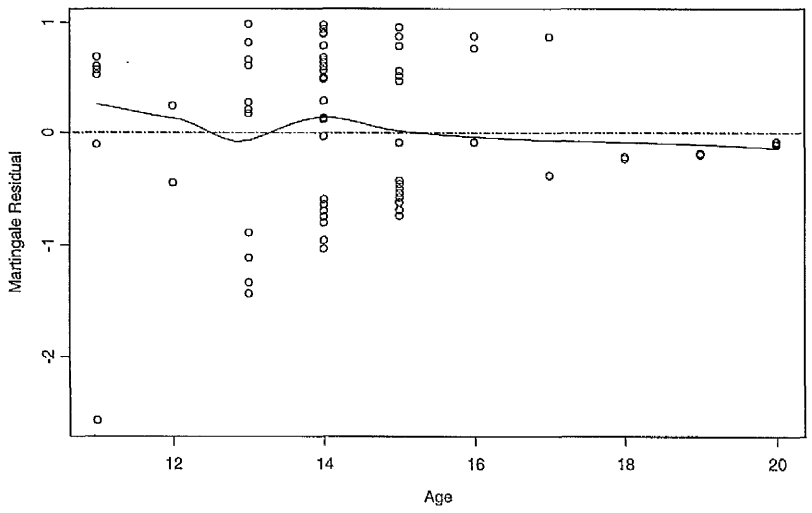
suggesting that the hazard function may be proportional for each these variables.

5.12.2 Dental Data

When considering the Dental data, as there was no matching variables available for inclusion, all model checking relates to the final Marginal Proportional Hazards model (MPH). The covariate representing an individual's Age was the only continuous covariate in the final MPH model.

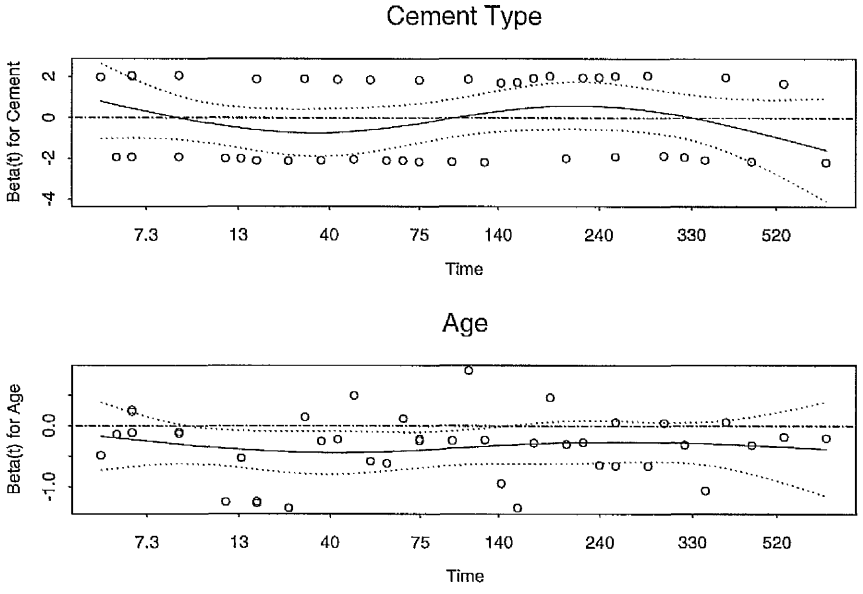
The linearity assumption of the effect of Age was confirmed by looking at a scatterplot of the Martingale residuals where Age demonstrated a reasonable degree of linearity (Figure 5.3).

Figure 5.3. Plot of Martingale Residuals for Age in the final MPH model for the Dental Data



In order to assess the proportional hazards assumption a plot of the Schoenfeld residuals for both of these covariates is given in Figure 5.4.

Figure 5.4. Plot of Schoenfeld Residuals against Time for Cement Type and Age for the final MPH Model Dental Data



The proportional hazards assumption seems reasonable for both Cement Type and Age by virtue of the plots and the results of both the Grambsch and Therneau (1994) or Quantin tests (1990) displayed in Table 5.21.

Table 5.21. Grambsch and Therneau and Quantin test p-values for the Dental Data.

	Grambsch and Therneau	Quantin
Cement Type	0.84	0.57
Age	0.99	0.51

5.13 Chapter Summary

A selection of methods was presented for modelling dependent survival data in the presence of covariates. Separate approaches were outlined for matched and paired scenarios where the approaches differ in how adjustment is made not only for the matching structure itself but also any imbalance that may exist in either the matching variables or any unmatched covariates.

The first method presented was an extension of the Simple Binomial test introduced in Chapter 4. Following this, a discussion of regression methods for survival data was presented with emphasis on the proportional hazards model in particular. One of the main assumptions in the proportional hazards approach is that survival times for each individual are independent of each other which is clearly violated in both matched and paired survival studies. The remainder of the chapter concentrated on extending the basic framework of the proportional hazards model to account for this dependency structure in the data.

Several extensions of the proportional hazards (PH) model to clustered studies are proposed. The Conditional PH model ignores the matched structure of the data but it uses information on the matching variables to correct inference made on the primary variable. In the Marginal PH model the regression coefficients are estimated assuming independence in a manner similar to the CPH model while the estimated covariance matrix is then 'corrected' post fit. A further refinement to the PH model for paired/matched survival data is to allow each pair to define a separate stratum while the final extension to the PH model presented in this chapter involved

introducing a random term, corresponding to each pair, into the model. This random pair effect, often termed a 'frailty', generates dependency between the survival times of the individuals in a pair and can be considered to represent unobserved, or 'hidden', covariates.

Given the results of this chapter, a natural question to ask is which of these models is, in general, best suited to Matched and which of these is best suited to Paired survival studies. In order to ascertain which of these suggested models might be 'best' in common practice a simulation study across a range of realistic scenarios is described in the following chapter.

Chapter 6

Analysing Matched/Paired Survival Data:

A Simulation Study

6.1 Introduction

The previous two chapters described methods for analysing matched and paired survival data, first comparing only the two levels of the primary variable but then incorporating matching variables and unmatched covariates into this comparison. The performance of all these methods was illustrated on the two example data sets.

The aim of this chapter is to investigate the performance of all these methods through simulations covering an extensive range of underlying scenarios intended to cover a wide range of potential real-life 'dependent' survival data problems. One key ingredient will, of course, be whether the data arise from a matched survival study or a paired survival study. The desired outcome of this simulation study is to provide general guidelines, if not specific recommendations, as to which of the methods covered in Chapters 4 and 5 are the most appropriate for these two types of 'dependent' survival study with guidelines as to how the level of dependency (i.e. the degree of association present) affect the general results of the simulation study.

6.2 The Aim of the Simulation

The matched and paired studies under consideration in this thesis relate specifically to cluster studies involving two measurements per cluster (i.e. a pair). The goal of each study is to assess the effect of a primary variable (with two levels) while controlling for the dependency structure of the data. The level of dependency is not of primary interest, the central issue is that it is present. Therefore, one cannot assume independence, and this should be accounted for in the analysis.

In this simulation study, the dependency structure is provided through the design of the study (through the matching variables and unmatched covariates) with no distributional assumption specifically made regarding the dependency structure and indeed no interest in estimating such. The assumption of independence of all observations is not valid by virtue of the design; thus any methods that assume independence are primarily incorrect and therefore likely to be flawed.

The mechanism for simulating matched and paired survival data in this thesis is introduced in Section 6.3. In order to 'mimic' a matched scenario, all the covariates were available when fitting the models with the covariate vector containing the primary variable, all matching variables and unmatched covariates. However, in order to 'mimic' a paired scenario, the matching variables were 'hidden' when fitting models for such with the covariate vector containing only the primary variable and unit covariates in such instances.

In summary, the dependency structure is present by design through the matching variables, the study design (for both matched and paired scenarios) is a case control cluster survival study and the aim of the simulation is to suggest which models are best suited for each scenario in terms of 'best' estimating the effect of the primary variable.

6.3 The Simulated Data

In both the matched and paired scenarios, a proportional hazards model was used to generate the survival times for both individuals in each pair, while a uniform distribution was used to generate censoring times. Four additional variables were simulated: the primary variable, two matching variables and an unmatched covariate. As will become clear later, the dependency structure is a direct consequence of how the matching variables are generated.

The PH model is the 'natural' model to use in simulating both matched and paired survival study data as all but one of the models under investigation assume proportional hazards. This then removes any suggestion that poor performance may be attributed to the proportional hazards assumption being invalid.

Recall from Section 5.4 that the proportional hazards (PH) model can be written as

$$S(t_{ip} \mid z) = S_0(t_{ip})^{\exp(\beta' z_{ipc})}$$

where the survival time for the i^{th} individual (i=1,2) in the p^{th} pair (p=1, .., P) is a product of an underlying baseline survival function $S_o(t)$ and a "risk score" (i.e. $\exp(\beta' z_{ipc})$) which is a function of that individual's covariate values (i.e. z_{ipc}).

Based on the PH model, a simple method for introducing 'dependency' is through the covariate vector. Consider the paired scenario first. Clearly, if the values of the matching variables are equal for each case and control pair the resulting survival times generated through the PH model will be correlated, the larger the number of matching variables the 'higher' the degree of association. In matched survival studies the emphasis on the matching is to make the two individuals in each pair as similar as possible in terms of survival. However, as indicated in Chapter 2, for continuous variables matching is usually only possible to within a certain range the width of which will directly affect the degree of association present.

Once a baseline survivor function and covariate vector are specified, survival times t_{ip} can be generated for each individual in each pair using the PH model. If there was no censoring these survival times are all event times. However, in a survival study, censoring is inevitable and for that reason a mechanism is needed to simulate censored observations (i.e. right censored).

One method to do this is to make use of the fact that the only information provided by a censored observation is that its true event time is some time in the future. Using this rationale, a binomial distribution can be used to randomly 'select' the specified proportion of cases and controls which are to be considered censored. Once an

'individual' with event time t_{ip} has been deemed censored their censoring time can then be simulated as $U[0, t_{ip}]$. This censoring mechanism insures that censoring is conditionally independent of survival time and enables the degree of censoring (i.e. proportion censored) to be similar for the cases and controls.

In order to investigate the effect the degree of censoring has on the performance of the methods presented in this thesis the following three proportions were considered: 0%, 30% and 60% censoring. These proportions were chosen in order to mirror survival studies with no, moderate and large amounts of censoring.

A myriad of choices now present themselves for the simulation strategy in terms of how many matching variables should be used and how 'accurate' the matching is likely be. A discussion of the strategy used to choose the matching variables for the matched and paired scenarios is now given.

6.3.1 Simulated Data for Matched Survival Studies

In order to simulate survival data using the PH model several decisions have to be made i.e. the components of the covariate vector (i.e. the effect each covariate has on survival) and the choice of baseline survivor function.

The covariate vector z_{ipc} consists of the primary variable, two matching variables (one categorical, one continuous) and a single unmatched variable. A description of the covariates in terms of how they were simulated and the relative effect they had on

survival (i.e. did they increase or decrease the risk of 'death') is given in Table 6.1 below.

Table 6.1. Description of the mechanism for generating the primary variable, matching and unmatched covariates for Matched Survival Data.

Variable	Case Value	Control Value	Relative Effect
Primary Variable	0	1	β_T=0, 1, 3
Matching variables:			
Categorical	Bin(1,½)	identical to Case value (i.e. perfect matching)	β_{MC1}=1
Continuous	N(½,¼)	U(Case value -0.05, Case value +0.05) (i.e. 'interval' matching)	β_{MC2}=1
Unmatched covariate	N(½,¼)	N(½,¼) (i.e. independent of the Case)	β_{UC}=1

In order to mirror the most 'real life' matched survival studies the matching is chosen to be perfect for the categorical matching variable and to within a specified range (i.e. effectively each pair are matched to within 1/20 of the range) for the continuous matching variable.

In order to get an impression of the effect of the matching criterion on the correlation between the case and control for the continuous matching variable, consider two random variables X and Y where $X \sim N(\mu_x, \sigma^2_x)$ and $Y = X + U(-a, a)$. It can be shown that

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_x}{\sqrt{\sigma^2{}_x + \frac{a^2}{3}}}.$$

Using a matching criterion of a=0.05 (as in Table 6.1) results in high correlation (ρ =0.99) between the case and control values for the simulated continuous matching variable. In the matched scenario therefore, there is perfect correlation between the case and control for the categorical matching variable and nearly perfect correlation for the continuous matching variable thereby introducing dependency into the design (albeit possibly weakened somewhat by the presence of an unmatched covariate) and subsequently making the independence assumption invalid.

Note that one method to 'quantify' the degree of association between the cases and controls in the simulated data is to estimate θ (the variance parameter of the gamma frailty model) by fitting a GPH model. The approximation between θ and Kendall's τ (i.e. $E(\tau)=\theta/\theta+2$, see section 5.9.2.1) can then be used. This will be returned to later in this chapter when a discussion on the effect of the degree of dependency is given.

From Table 6.1, three different values for the coefficient β_T of the primary variable were chosen, namely 0, 1 and 3, in order to mirror survival studies where the primary variable has no effect, a moderate effect and a considerable effect. The matching variables and the unmatched covariate were all chosen to have the same β coefficient (β = 1) representing a 'moderate' effect on survival.

Recall that the PH model estimates the relative effect of the covariates on an unspecified baseline hazard function. As the primary variable is coded using 0 for the case and 1 for the control, the control will always have poorer survival prospects than the case for positive values of β_T. Note also from Table 6.1 that the relative effect of the matching variables and unmatched covariate are all positive resulting in poorer survival prospects for increasing values of the matching variables and unmatched covariate.

Recall also that the PH model is based on analysing ranks, and therefore the observation times may be recorded as days, months or years without affecting the analysis. In survival models, the exponential distribution is the simplest choice for the baseline survivor function and so will be used here. By virtue of the choice of regression coefficients in this simulation study, the covariate vector will always deflate the baseline hazard function. Given this fact, a suitable large value of 400 for the mean parameter of the exponential distribution was chosen so as not to deflate the median case and control survivor times to be close to zero. Care was taken also to avoid round off error in order to avoid an unnecessarily large number of tied survival times.

6.3.2 Simulated Data for Paired Survival Studies

A summary of the components of the covariate vector, in terms of how they were simulated and their relative effect on survival, is given in Table 6.2.

Table 6.2. Description of the mechanism for generating the primary variable, matching and unmatched covariates for Paired Survival Data.

Variable	Case Value	Control Value	Relative Effect
Primary Variable	0	1	β_T=0,1,3
Matching Variables: Categorical Continuous	 Bin(1,½) N(½,¼)	 identical to Case Value identical to Case Value	 β_{MC1}=1 β_{MC2}=1
Unit Covariate: Continuous	N(½,¼)	identical to Case Value	β_{UC}=1

In order to mimic a Paired survival study the case and control covariates are identical for each pair (i.e. perfect correlation) and were again all chosen to have the same coefficient (β = 1) but of course will be 'hidden' in the subsequent analysis. The values for the coefficient of the primary variable were chosen in the same manner as when simulating data for the Matched study scenario.

Once again the observation times for each pair were generated using a proportional hazards (PH) model, where $S_0(t)$ was again generated using an exponential distribution with mean 400. The desired degree of censoring and the censoring times were generated in exactly the same manner as for the Matched study simulations.

6.3.3 Simulation Configurations

To investigate the performance of the competing models under each of the two approaches, 1000 simulations of each of a number of configurations were carried out.

The configurations were defined by the following quantities:

i) The number of pairs of subjects, P, taken as ***25***, ***100*** and ***250***

ii) The percentage of pairs censored, % Censored, taken as ***0%, 30%*** and ***60%***

iii) The true relative effect of the primary variable, β_T, taken as ***0***, ***1*** or ***3***

For illustrative purposes a sample matched survival simulated data set is given in Table 6.3 with censoring status coded as 1 for complete observations and 0 for censored observations.

Note, from Table 6.3, the near perfect matching of the continuous variable, and the fact that, although 30% censoring was desired, the actual censoring proportions in this simulation were 36% for the cases and 24% for the controls of course, over all the simulations the average proportion censored was 30% for both levels of the primary variable.

Table 6.3. An example of a simulated 25 pairs matched survival data set using 30% censoring and β_T=1.

Pair	Individual	Primary Variable (i.e. Case/Control)	Continuous Matching Variable	Categorical Matching Variable	Unmatched Covariate	Observation Time	Censoring Status
1	1	0	0.29	1	1	1	0
1	2	1	0.31	1	0.85	27	1
2	1	0	0.30	0	0.30	84	0
2	2	1	0.29	0	0.57	0	0
3	1	0	0.65	1	0.93	36	1
3	2	1	0.64	1	0.61	18	1
4	1	0	0.23	1	0.83	25	0
4	2	1	0.22	1	0.90	25	1
5	1	0	0.45	1	0.35	29	0
5	2	1	0.47	1	0.79	20	0
6	1	0	0.13	0	0.50	247	1
6	2	1	0.12	0	0.47	14	1
7	1	0	0.52	0	0.42	21	1
7	2	1	0.52	0	0.60	37	1
8	1	0	-0.16	1	0.24	40	0
8	2	1	-0.15	1	0.19	119	1
9	1	0	0.60	0	0.15	21	1
9	2	1	0.56	0	0.76	4	1
10	1	0	0.97	1	0.90	56	1
10	2	1	0.95	1	0.13	14	1
11	1	0	0.24	0	0.41	450	1
11	2	1	0.28	1	0.02	15	0
12	1	0	0.46	0	0.47	27	1
12	2	1	0.43	0	0.28	77	1
13	1	0	0.54	0	0.21	109	1
13	2	1	0.51	1	0.43	42	1
14	1	0	0.29	0	0.44	24	0
14	2	1	0.32	0	0.58	22	0
15	1	0	0.55	1	0.40	69	1
15	2	1	0.53	1	0.90	17	1
16	1	0	0.14	0	0.34	1608	1
16	2	1	0.09	0	0.67	3	1
17	1	0	1.14	0	0.52	25	1
17	2	1	1.15	0	0.51	68	1
18	1	0	0.94	1	0.07	178	1
18	2	1	0.90	1	0.44	12	1
19	1	0	0.67	1	0.60	23	0
19	2	1	0.67	1	0.35	80	1
20	1	0	0.67	1	0.24	43	1
20	2	1	0.69	1	0.24	3	1
21	1	0	0.52	1	0.27	31	1
21	2	1	0.50	1	0.03	17	1
22	1	0	0.54	0	0.49	169	1
22	2	1	0.57	0	0.27	14	1
23	1	0	0.57	1	0.61	21	0
23	2	1	0.53	1	0.73	13	0
24	1	0	0.33	0	0.67	82	0
24	2	1	0.35	0	0.51	18	1
25	1	0	0.75	1	0.90	9	1
25	2	1	0.79	1	0.49	14	0

In order to illustrate the typical data sets in the paired survival simulations, Kaplan-Meier plots of the estimated case and control survivor functions are displayed in Figures 6.1, 6.2 and 6.3 for a selection of configurations. Note from these plots the effect of simulating data using an exponential baseline distribution and a proportional hazards model as well as the effect of the primary variable, matching variables and unit covariate on the baseline hazard function.

Figure 6.1
Sample Paired Simulated Data
with P=25, 0% censoring and $\beta_T=0$.

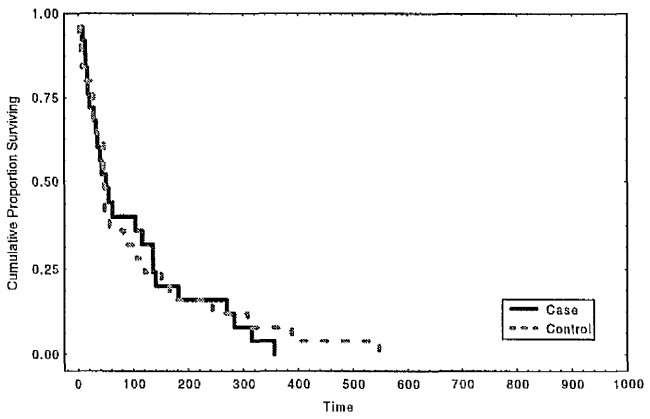


Figure 6.2
Sample Paired Simulated Dataset
with P=100, 30% censoring and $\beta_T=1$.

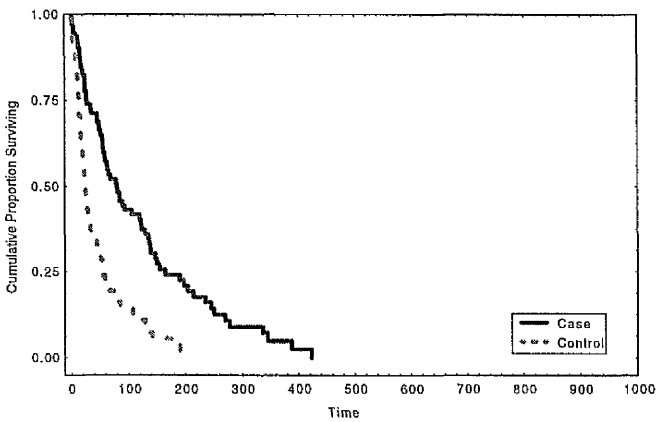
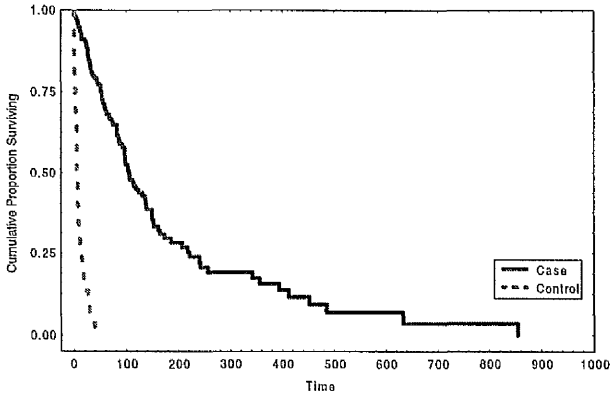


Figure 6.3
Sample Paired Simulated Dataset
with P=250, 60% censoring and β_T=3.



6.4 Model Performance Indicators

For each simulation a set of data based on one of the above configurations was generated and the performance of each of the various modelling approaches was investigated based on a number of statistical criteria described below.

6.4.1 Methods involving the Primary Variable Alone

This section aims to compare the methods outlined in Chapter 4 which compared the case and control survival prospects while ignoring all covariates except the primary variable.

For each simulation, test statistics were calculated for the Simple Binomial, Paired Prentice-Wilcoxon and Akritas tests. An investigation of the parametric and non-parametric approaches (involving the pair-wise differences in survival time) in terms of their use as hypothesis tests was made also. Recall that the 'pair-wise differences'

procedures have the added feature of estimating the distribution of the difference in survival time (in units of time) between the case and control populations but for this simulation study only their use as hypothesis tests will be under consideration.

As all of the tests presented in Chapter 4 are two-sided tests and all use a normal approximation, critical values from a normal distribution for nominal 5% and 1% significance levels were used to compare the test's performance. The performance of each of these tests was assessed across all the simulations based on the size (i.e. whether the test achieved the nominal significance level) which is illustrated by the $\beta_T=0$ configuration results and power of the test (i.e. the probability of rejecting the null hypothesis when it is false) which is illustrated by the $\beta_T=1$ and $\beta_T=3$ configuration results.

6.4.2 Methods incorporating matching variables and covariates

This approach considers all the methods outlined in Chapter 5 which compare the case and control survival prospects while incorporating all covariates. For each simulation a point and interval estimate of the primary variable regression coefficient was calculated for each of the models outlined in Chapter 5. Unlike section 6.4.1, the emphasis here is on interval estimates rather than test statistics with the performance of each of these models being assessed (across all the simulations) on:

i) ***bias***, the long run average of $\hat{\beta}_T$ minus β_T (the 'average' estimated minus true value of the primary variable regression coefficient)

ii) ***coverage rate***, the long run proportion of occasions when the (nominally 95% confidence) interval estimate contains the true regression coefficient

and

iii) the average ***width*** of this interval estimate.

6.5 Simulation Results

The simulation results are in two sections, the first section deals with the methods presented in Chapter 4 (involving hypothesis tests on the primary variable alone) while the methods proposed in Chapter 5 (involving interval estimates of the effect of the primary variable incorporating the effects of the matching variables and unmatched or unit covariates) are the subject of the second section.

6.5.1 Methods involving the Primary Variable Alone

Results are presented first for the various 'covariate free' methods presented in Chapter 4 for the ***matched*** data simulation configurations. Essentially these are tests of $\beta_T=0$ where β_T is the true 'effect' of the primary variable.

6.5.2 Matched Survival Simulation Study

The performance of each of the following:

- the Simple Binomial (SB) test,
- Paired Prentice-Wilcoxon (PPW) test,
- Akritas test (AKR) test,
- parametric (D_p) 'differences' test

and the

- non-parametric (D_{np}) 'differences' test

in terms of whether each test made the correct decision in terms of rejecting the null hypothesis of $\beta_T=0$ for all the various pair-size, censoring and effect-size configurations are displayed in Table 6.4 and Figure 6.4.

The graphs are arranged as follows: a categorised scatterplot of the performance of each test (in terms of the percentage of times each test correctly rejects the null hypothesis) by the size of the true effect and the degree of censoring is displayed for each critical level.

Table 6.4.
Proportion of times each test made the correct decision for all Sample Size, % Censoring and Effect Size configurations for the Matched Data Simulations.

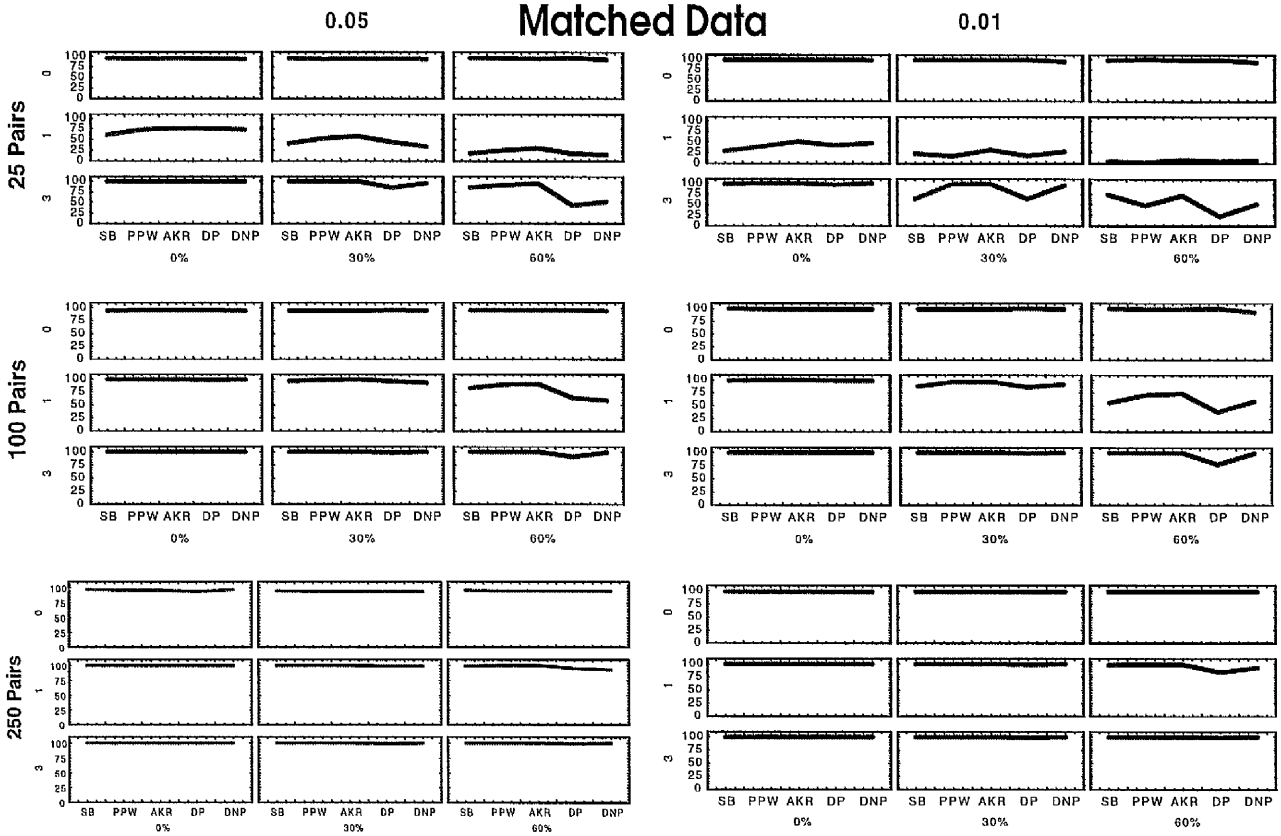
β_T=0	25 Pairs					100 Pairs					250 Pairs				
0.05	SB	PPW	AKR	D_P	D_{NP}	SB	PPW	AKR	D_P	D_{NP}	SB	PPW	AKR	D_P	D_{NP}
0%	96	95	96	95	94	95	96	96	96	95	97	96	96	94	97
30%	96	94	95	95	94	95	95	94	96	95	95	94	94	94	94
60%	97	95	95	96	93	96	96	96	96	94	96	95	95	95	95
0.01															
0%	99	99	99	99	98	100	99	99	99	99	99	99	99	99	99
30%	99	99	99	99	94	99	99	99	100	99	99	99	99	99	99
60%	98	99	99	99	93	99	99	99	100	93	99	99	99	99	99

β_T =1	25 Pairs					100 Pairs					250 Pairs				
0.05	SB	PPW	AKR	D_P	D_{NP}	SB	PPW	AKR	D_P	D_{NP}	SB	PPW	AKR	D_P	D_{NP}
0%	61	74	77	76	74	100	100	100	99	100	100	100	100	100	100
30%	41	53	58	44	34	97	99	100	96	93	100	100	100	99	99
60%	18	25	30	17	15	83	90	91	64	59	99	100	100	95	93
0.01															
0%	29	40	52	42	48	99	100	100	99	99	100	100	100	100	100
30%	23	17	32	18	28	88	97	97	86	92	100	100	100	99	100
60%	5	3	9	5	7	56	71	74	38	59	98	99	99	84	93

β_T =3	25 Pairs					100 Pairs					250 Pairs				
0.05	SB	PPW	AKR	D_P	D_{NP}	SB	PPW	AKR	D_P	D_{NP}	SB	PPW	AKR	D_P	D_{NP}
0%	100	100	100	99	100	100	100	100	100	100	100	100	100	100	100
30%	100	100	100	84	95	100	100	100	99	100	100	100	100	99	100
60%	85	91	94	43	52	100	100	100	91	99	100	100	100	99	100
0.01															
0%	99	100	100	97	100	100	100	100	100	100	100	100	100	100	100
30%	74	99	99	64	95	100	100	100	99	100	100	100	100	99	100
60%	64	48	72	22	52	100	100	100	78	99	100	100	100	99	100

Key:
SB – Simple Binomial Test, **PPW** - Paired Prentice-Wilcoxon Test, **AKR** - Akritas Test,
D_P - Parametric Differences Test, D_{NP} - Non-Parametric Differences Test

Figure 6.4
Performance of each test for each Matched simulation configuration.



For the smallest sample size first (i.e. ***25 pairs***) all the tests performed well by achieving the nominal significance level when there was no simulated effect (i.e. β_T=0) regardless of the censoring. However, all the tests performed poorly with increasing censoring for the β_T=1 simulated effect with particularly low power at the α= 0.01 level. All the tests performed well for β_T=3 except when there was 60% censoring where the 'interval based' tests performed poorly.

In general only the PPW and AKR tests maintained their power for a 0.01% significance level test when compared to their performance for a 0.05% significance level test again excluding the moderate effect size (β_T=1) and high censoring configuration where all tests performed poorly. The AKR test is recommended as, although it sometimes performs poorly, it consistently performs better than all the other proposed tests.

When considering the ***100 pair*** simulations, all the tests performed well for all configurations with the AKR test showing consistent but slight improvements in power over the others. The effect of increasing censoring and a moderate effect size (β_T=1) is again evident but is not as severe as that exhibited for the configurations with 25 pairs. The Simple Binomial test (SB) performed well when compared to the mathematically more "complex" tests at both the nominal significance levels.

When the number of simulated pairs increased to ***250***, there is effectively nothing to choose between any of the tests except for one configuration (namely β_T=1 and 60% censoring) where the D_P test performed slightly worse. Once again the SB test performed well for all configurations.

6.5.3 Paired Survival Simulation Study

The results of how the 'covariate free' methods presented in Chapter 4 performed when analysing ***paired*** survival data for the simulation configurations described in section 6.3.3. are now given.

Following the same convention as used with the matched survival data simulations, a table (Table 6.5) and categorised scatterplot (Figure 6.5) of the performance of each of the tests are displayed. Once again, performance was assessed by the proportion of times the test made the 'correct' decision in terms of correctly rejecting, the null hypothesis of $\beta_T=0$ (representing no primary variable effect).

For the configurations with ***25 pairs***, all the tests performed well (as with the Matched simulation) when $\beta_T=0$ regardless of the censoring and significance levels (i.e. they achieved the nominal significance levels).

Again, all the tests performed ***poorly*** with increasing censoring for the case of $\beta_T=1$ with particularly low power at the 0.01 significance level. The D_P test performed poorest of all while the SB test and D_{NP} based test had comparable performances for most simulation configurations. The AKR test performed best in terms of maintaining power at both the 5% and 1% nominal significance levels although, as to be expected, showed poor power for the combination of moderate effect size ($\beta_T=1$) and high censoring (60%).

Table 6.5
Proportion of times each test made the correct decision for all Sample Size, % Censoring and Effect Size Configuration for Paired Data Simulations.

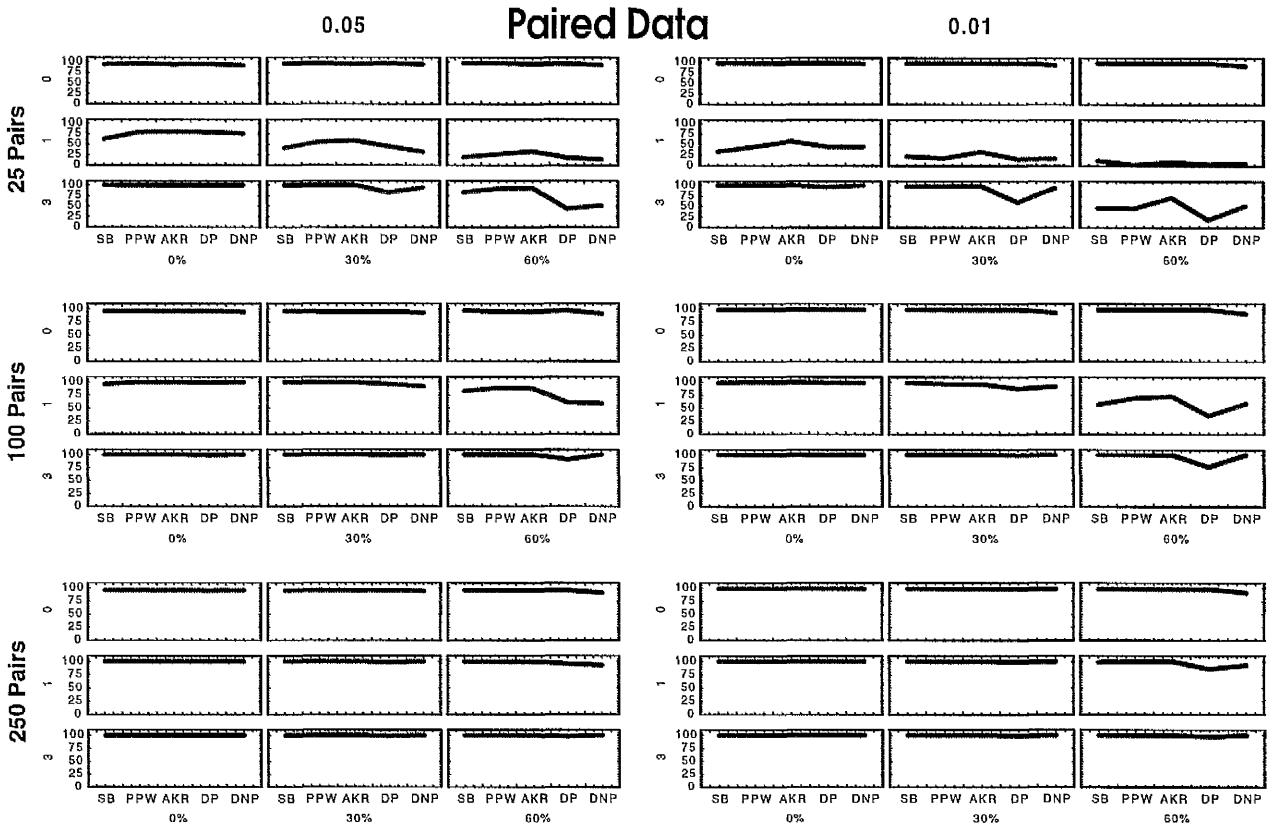
β_T =0	25 Pairs					100 Pairs					250 Pairs				
0.05	SB	PPW	AKR	D_P	D_{NP}	SB	PPW	AKR	D_P	D_{NP}	SB	PPW	AKR	D_P	D_{NP}
0%	94	96	94	95	93	96	96	96	96	95	96	96	96	95	96
30%	94	96	95	96	93	96	95	95	95	93	95	96	95	95	94
60%	94	96	95	96	93	96	95	95	95	91	95	96	95	95	91
0.01															
0%	98	99	99	99	97	99	99	99	99	99	99	99	99	99	99
30%	99	99	99	99	94	99	99	99	99	93	99	99	99	99	99
60%	99	99	99	99	92	99	99	99	99	91	98	98	98	98	91

β_T =1	25 Pairs					100 Pairs					250 Pairs				
0.05	SB	PPW	AKR	D_P	D_{NP}	SB	PPW	AKR	D_P	D_{NP}	SB	PPW	AKR	D_P	D_{NP}
0%	63	79	80	79	76	99	100	100	99	100	100	100	100	100	100
30%	41	55	59	45	32	99	99	99	96	92	100	100	100	99	100
60%	19	27	33	18	14	82	89	88	61	59	100	100	100	95	93
0.01															
0%	33	45	58	45	45	99	100	100	99	99	100	100	100	100	100
30%	22	18	33	16	17	99	97	97	87	92	100	100	100	99	100
60%	12	2	9	4	5	57	70	73	36	59	99	100	100	86	93

β_T =3	25 Pairs					100 Pairs					250 Pairs				
0.05	SB	PPW	AKR	D_P	D_{NP}	SB	PPW	AKR	D_P	D_{NP}	SB	PPW	AKR	D_P	D_{NP}
0%	100	100	100	100	100	100	100	100	99	100	100	100	100	100	100
30%	100	100	100	82	94	100	100	100	99	100	100	100	100	99	100
60%	83	92	93	44	52	100	100	100	91	100	100	100	100	98	100
0.01															
0%	100	100	100	96	100	100	100	100	99	100	100	100	100	100	100
30%	98	98	99	60	94	100	100	100	99	100	100	100	100	99	100
60%	47	46	71	19	52	100	100	100	77	100	100	100	100	97	100

Key:
SB - Pair Performance Test, **PPW** - Paired Prentice-Wilcoxon Test, **AKR** - Akritas Test,
D_P - Parametric Differences Test, D_{NP} - Non-Parametric Differences Test

Figure 6.5
Performance of each test for each Paired simulation configuration.



When the number of pairs increased to ***100***, all the tests performed well for all configurations with the PPW and AKR test performing consistently as the best. The effect of increasing censoring and a moderate effect size (β_T=1) is again evident but once more the SB, D_P and D_{NP} tests performed well for all but the high censoring configurations.

Finally, when considering the simulations with ***250 pairs***, the only configuration where there was any substantial difference in performance between any of the tests was the β_T=1 and 60% censoring configuration where the D_P test had slightly poorer performance.

6.5.4 Conclusion

Not surprisingly all the tests perform better (in terms of power) as the number of pairs increased, and as the relative effect of the primary variable increased. The degree of censoring proves important especially for small samples and for moderate effects.

For both matched and paired survival studies, the AKR test is recommended in general as this test appears to perform "best" in terms of power for all simulated sample size, effect size and censoring configurations.

All the tests performed poorly when detecting a moderate effect in small sample sizes with a high degree of censoring. In addition, all of the newly proposed 'tests' (i.e. the SB, D_P and D_{NP} test) had comparable performance when the number of pairs was 100 and larger and there was no or moderate censoring.

In addition, all of the newly proposed tests had comparable performance to the AKR and PPW tests in general except when detecting a small difference in the presence of a high degree of censoring. This discrepancy in performance is presumably due to the potentially large number of doubly censored pairs excluded from these new tests. It should be noted that the "interval based" procedures (D_P and D_{NP}) are based on a null

hypothesis that the median difference in survival is zero while the AKR and PPW tests have as null hypothesis that the survival distributions are identical for the two levels of the primary variable.

Further it should be noted that the D_P and D_{NP} tests are interval estimation driven procedures which provide an estimate of the difference in survival between the cases and controls and consequently the hypothesis tests are really a by-product. A more specific simulation study may be carried out as future work in order to compare these 'interval based" procedures more comprehensively, specifically in terms of their ability to estimate median difference.

6.6 Methods involving the Matching Variables and Unmatched Covariates

Across 1000 simulations for each of the configurations described in 6.3.3, the following approaches to estimation of the effect of the primary variable correcting for all available matched variables and unmatched covariates were compared in terms of a number of criteria.

These approaches were:

- Pair Performance (PP),
- Conditionally Independent Proportional Hazards (CPH),
- Marginal Proportional Hazards (MPH),
- Stratified Proportional Hazards (SPH)

and

- Gamma Proportional Hazards (GPH) Models

The criteria used to compare the performance of each of the approaches, as outlined in section 6.4.2, were ***bias***, ***coverage*** rate and the average ***width*** of the interval estimate.

6.6.1 Matched Survival Study Simulation

The median of the bias, the percentage coverage and the median interval width for nominally 95% confidence interval estimate of the primary variable coefficient β_T for each model across 1000 simulations for each pair size, censoring and effect size configuration are given in Table 6.6. Recall that the 'bias' for a simulation is calculated as $\hat{\beta}_T$ minus the true value, β_T so negative values represent instances when that particular model under-estimated the true regression coefficient. The "best" model will be one that

1. has ***no*** 'average' bias
2. gets ***close*** to nominal 95% coverage

and

3. has the ***smallest*** interval width 'possible.

6.6.1.1 Comparing Models in terms of Bias

In order to compare the models in terms of bias, tables and boxplots of the bias for each model for the 25, 100 and 250 pairs simulation configurations are given in Figure 6.6 and Table 6.6.

As one would expect the bias in general decreases with increasing sample size while the variance of the bias increases with increasing censoring. Note that the bias distributions for the CPH and MPH models are identical as both methods use the same point estimate. From the results relating to the ***25 pairs***, there is no suggestion of any real bias in any of the configurations based on the models fitted when β_T=0. The CPH, MPH and GPH all perform in a similar manner where there is no suggestion that any of these methods are biased. The PP and SPH models exhibit larger variability than these, and this is particularly evident as the level of censoring increases. In particular, the PP model showed the largest variability of $\hat{\beta}_T$ of all the simulation configurations for the 25 pairs, 60% censoring configuration.

In many of the simulations with 60% censoring, the PP and SPH model provided dubious estimates for the coefficients as the estimation procedure did not converge resulting in skewed β_T estimates (with extremely large estimated standard errors) as evidenced in the lower panels of Figure 6.6. There is a suggestion that the estimate of β_T from the PP model may be biased as the median bias appears to increase with increasing magnitude of β_T.

Figure 6.6
Boxplot of Bias for 25 pair Matched simulation configurations.

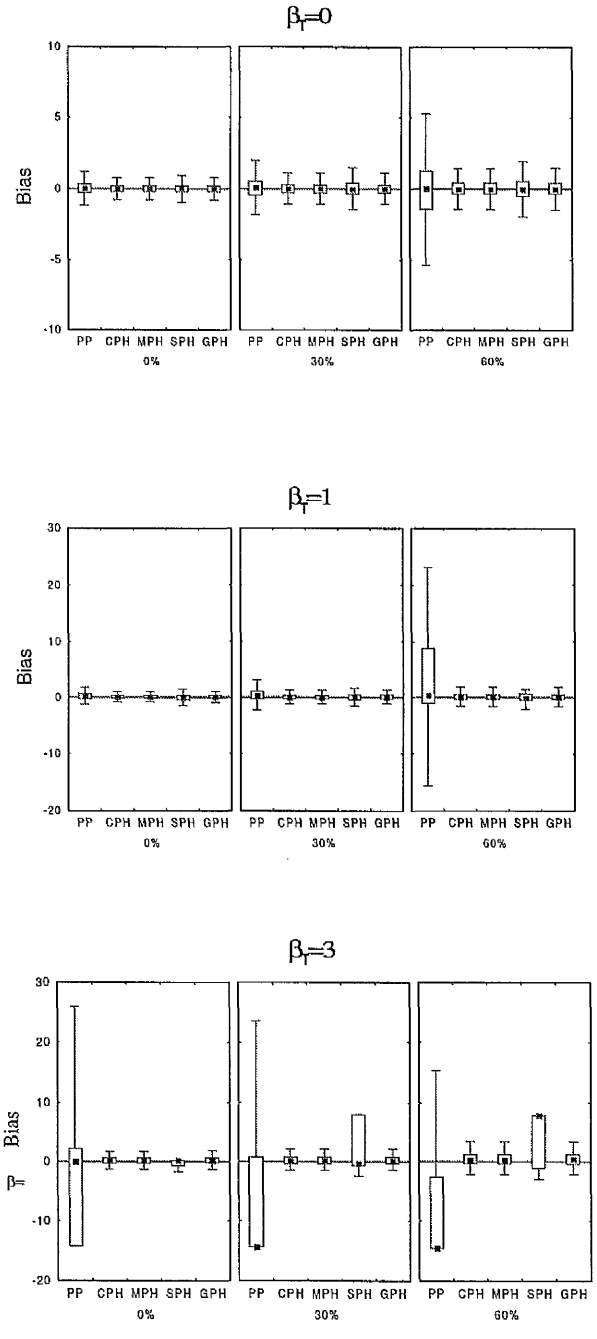


Table 6.6. Median Bias, % Coverage (% Cov) and Median Interval Width (IW) for each Matched Data Simulation Configuration.

			25 Pairs			100 Pairs			250 Pairs		
β_T	% Censored	Model	Median Bias	% Cov	Median IW	Median Bias	% Cov	Median IW	Median Bias	% Cov	Median IW
0	*0*	PP	0.029	96	1.75	0.014	95	0.81	0.000	96	0.51
		CPH	-0.002	93	1.19	0.009	95	0.57	0.000	95	0.35
		MPH	-0.002	91	1.15	0.009	94	0.56	0.000	95	0.35
		SPH	0.080	97	1.58	0.020	95	0.78	0.000	97	0.49
		GPH	-0.001	92	1.19	0.011	94	0.57	0.000	95	0.35
	30										
		PP	0.075	98	2.46	0.004	95	1.05	0.006	95	0.65
		CPH	0.024	93	1.45	-0.001	95	0.68	0.000	95	0.42
		MPH	0.024	91	1.41	-0.001	94	0.66	0.000	95	0.42
		SPH	-0.003	96	2.03	-0.003	95	0.99	0.009	95	0.63
		GPH	0.025	92	1.19	0.001	95	0.68	0.000	95	0.42
	60										
		PP	0.026	99	5.95	0.003	96	1.57	-0.016	96	0.94
		CPH	-0.017	94	2.03	-0.007	95	0.91	0.000	96	0.56
		MPH	-0.017	92	1.96	-0.007	95	0.89	0.000	96	0.55
		SPH	0.008	99	2.99	-0.008	96	1.43	-0.008	96	0.89
		GPH	-0.017	93	2.03	-0.009	95	0.91	0.000	96	0.56
1	*0*	PP	0.121	98	2.05	0.036	96	0.93	0.017	95	0.58
		CPH	0.052	93	1.30	0.015	95	0.62	0.000	94	0.39
		MPH	0.052	91	1.23	0.015	95	0.60	0.000	95	0.38
		SPH	-0.059	97	1.75	-0.045	95	0.87	-0.039	94	0.55
		GPH	0.066	93	1.46	0.025	94	0.62	0.006	95	0.39
	30										
		PP	0.264	98	3.04	0.055	95	1.21	0.022	96	0.74
		CPH	0.088	93	1.60	0.018	94	0.74	0.000	95	0.46
		MPH	0.088	92	1.53	0.018	93	0.72	0.000	95	0.45
		SPH	-0.033	97	2.26	-0.037	95	1.12	-0.044	95	0.70
		GPH	0.100	92	1.67	0.030	93	0.74	0.008	95	0.46
	60										
		PP	0.343	99	24.02	0.082	98	1.85	0.0365	94	1.10
		CPH	0.077	93	2.22	0.008	96	0.99	0.000	94	0.62
		MPH	0.077	91	2.13	0.008	96	0.97	0.000	94	0.61
		SPH	-0.053	97	3.28	-0.053	96	1.59	-0.069	94	1.00
		GPH	0.090	92	2.24	0.012	96	0.99	0.006	94	0.62
3	*0*	PP	0.092	99	21.07	0.349	99	2.42	0.204	96	1.35
		CPH	0.185	96	2.13	0.016	96	0.97	0.001	95	0.61
		MPH	0.185	92	1.99	0.016	94	0.94	0.001	95	0.59
		SPH	0.169	95	4.00	-0.065	94	1.80	-0.172	89	1.08
		GPH	0.201	95	5.37	0.038	94	0.98	0.011	95	0.61
	30										
		PP	-14.36	99	161.75	0.422	99	3.68	0.269	99	1.87
		CPH	0.172	97	2.62	0.026	95	1.19	0.000	95	0.74
		MPH	0.172	92	2.45	0.026	94	1.14	0.000	94	0.71
		SPH	-0.266	95	4.06	-0.111	93	2.31	-0.186	90	1.42
		GPH	0.178	96	5.45	0.044	95	1.20	0.016	95	0.74
	60										
		PP	-14.56	99	236.24	0.107	99	12.85	0.297	99	3.03
		CPH	0.391	98	3.83	0.034	95	1.61	0.000	95	0.99
		MPH	0.391	81	3.21	0.034	94	1.56	0.000	95	0.97
		SPH	7.845	94	320.07	-0.026	92	3.96	-0.290	88	2.00
		GPH	0.426	98	3.84	0.048	95	1.61	0.009	95	0.99

There was a reduction in the variability of $\hat{\beta}_T$ for all configurations when the sample size increases from 25 to ***100 pairs*** (Figure 6.7). All of the models performed well for the 100 pair configurations β_T=0, with the 'best performance', in terms of smallest median bias, for the CPH and MPH models. The PP model had the largest variability, especially for the β_T=3 and 60% censoring configuration, with the SPH model the next most variable.

A continuing decrease in bias occurred for the scenario of ***250 pairs*** of observations (Figure 6.8). Here, there was no suggestion of any bias for the CPH, MPH and GPH models across all the configurations. There was a suggestion however that the PP model overestimated β_T while the SPH model under-estimated the effect of the primary variable for configuration with the largest primary variable effect (i.e. β_T=3).

Figure 6.7
Boxplot of Bias for 100 pair Matched simulation configurations.

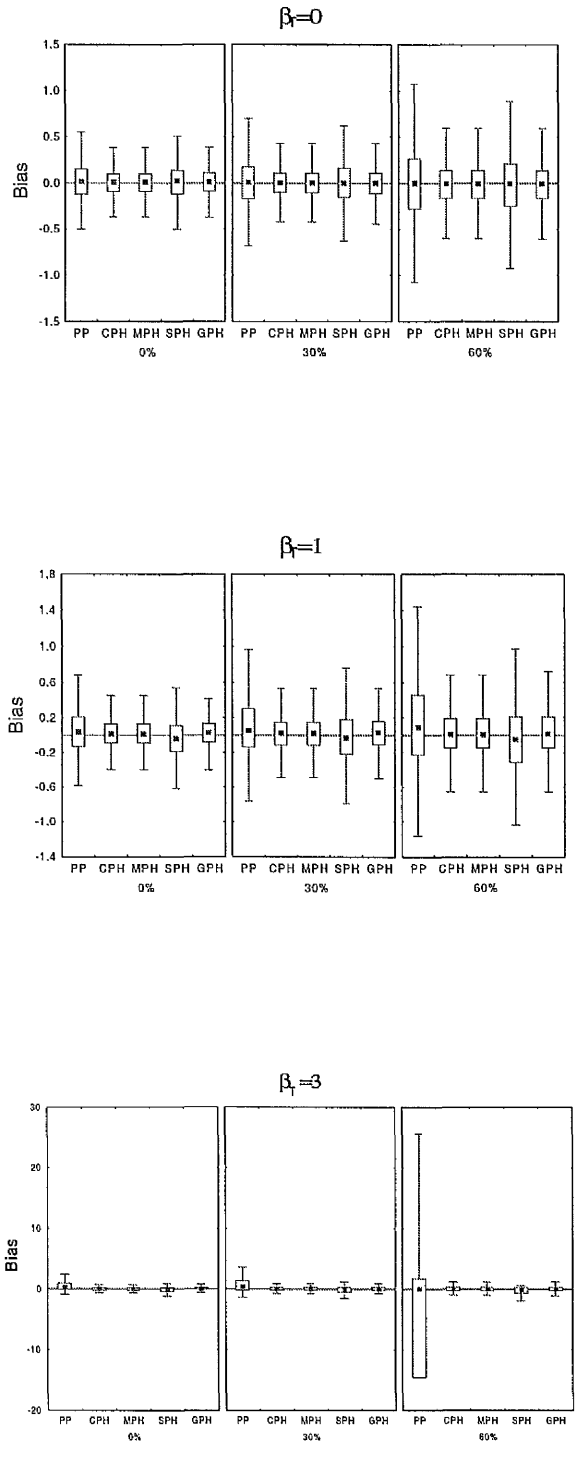
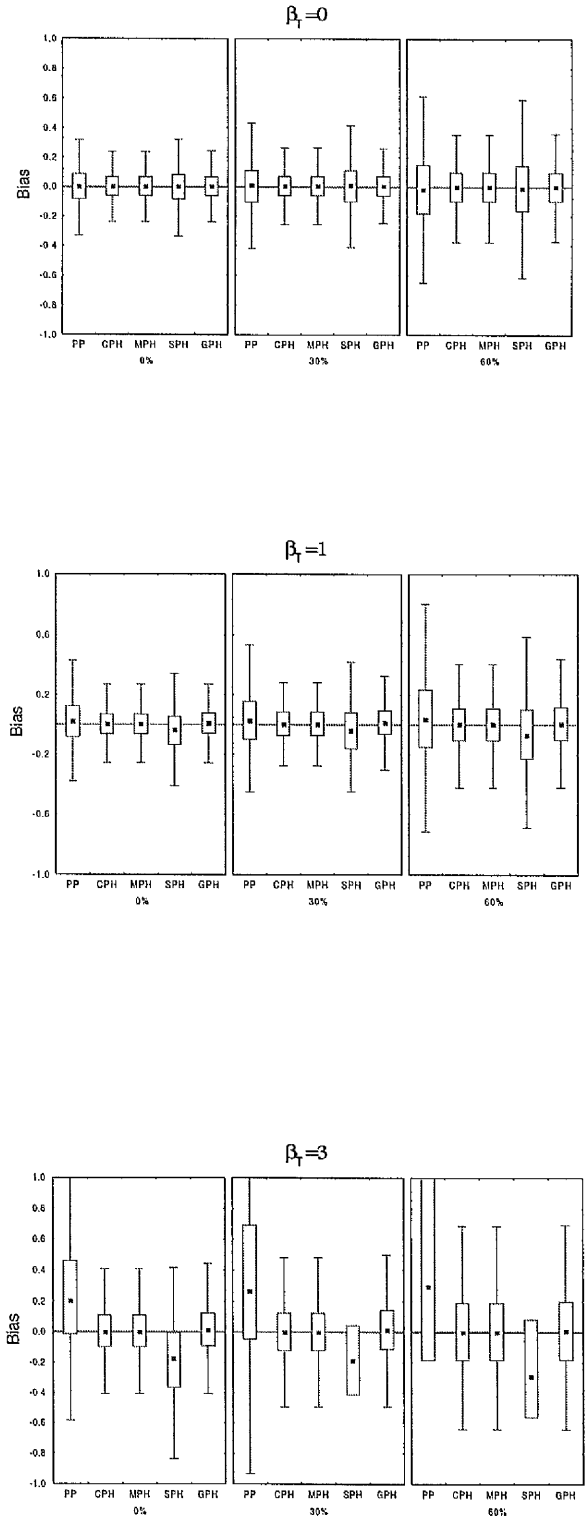


Figure 6.8
Boxplot of Bias for 250 pair Matched simulation configurations.



6.6.1.2 Comparison of Coverage Rates and Confidence Interval Widths

The percentage coverage (i.e. estimated confidence coefficient) for the interval estimates based on each model respectively across the underlying simulation configurations are given in Table 6.5. The percentage coverage represents the proportion of times in the 1000 simulations that the (nominal 95% confidence) interval estimate captured the true value β_T. Scatterplots of the median interval width against percentage coverage for each simulation configuration labelled by model are given in Figures 6.9,6.10 and 6.11. The best model is that which has attains nominal coverage in addition to having the smallest median interval width.

As expected, the median interval width decreases with increasing sample size and increases slightly with increasing censoring. From these plots the general pattern appears to be that the CPH, MPH and GPH models behave similarly and distinctly better that the SPH and PP models which have poorer pattern in both coverage and interval width.

The CPH model appears to perform best for the 25 pair configuration where the MPH model appears best for the 100 and 250 pair configurations. As the number of simulated pairs increases however, the MPH model performs best in terms of having good coverage with the narrowest intervals. The GPH model has comparable performance to the CPH model for most configurations, with a suggestion of wider intervals for the 25 pairs, β_T=3 configuration.

The PP and SPH models perform poorly in terms of having extremely wide intervals (in comparison to the other models) accompanied, not surprisingly, with good coverage. The large intervals are particularly noticeable in the case of the high censoring configurations.

In conclusion, the CPH model performs best when considering all the simulation configurations with the MPH model a relatively close second. The MPH model however appears to perform best when the number of pairs is large.

Figure 6.9
Scatterplot of % Coverage by Median Interval Width for 25 Pair Matched Simulation Configurations.

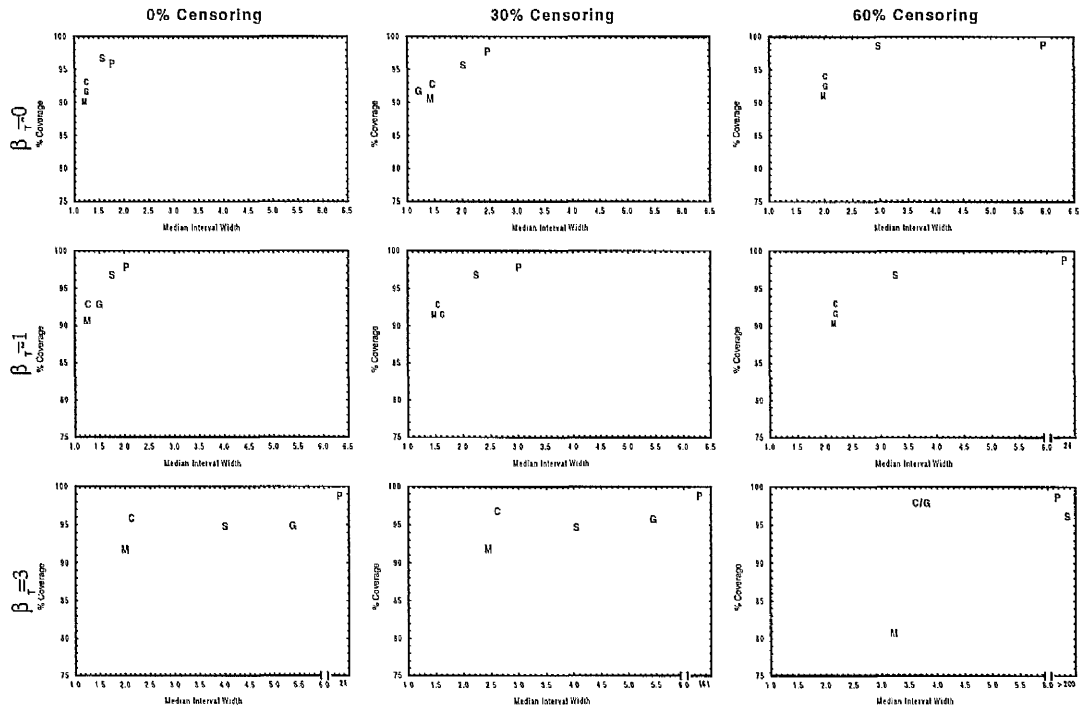


Figure 6.10
Scatterplot of % Coverage by Median Interval Width for 100 Pair Matched Simulation Configurations

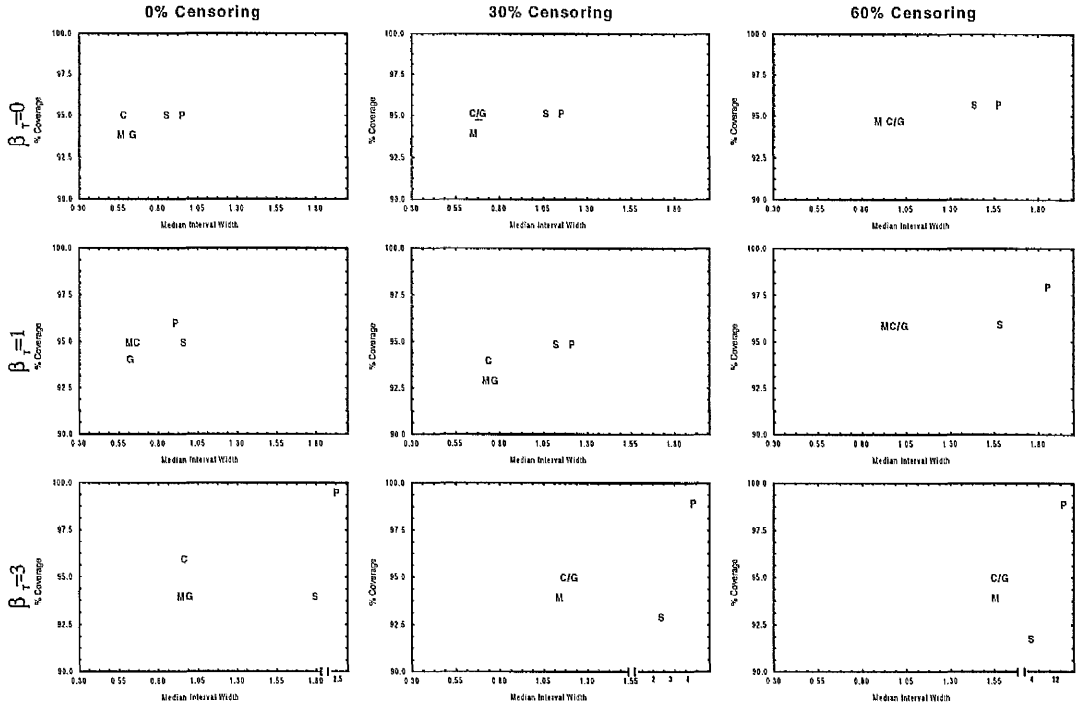
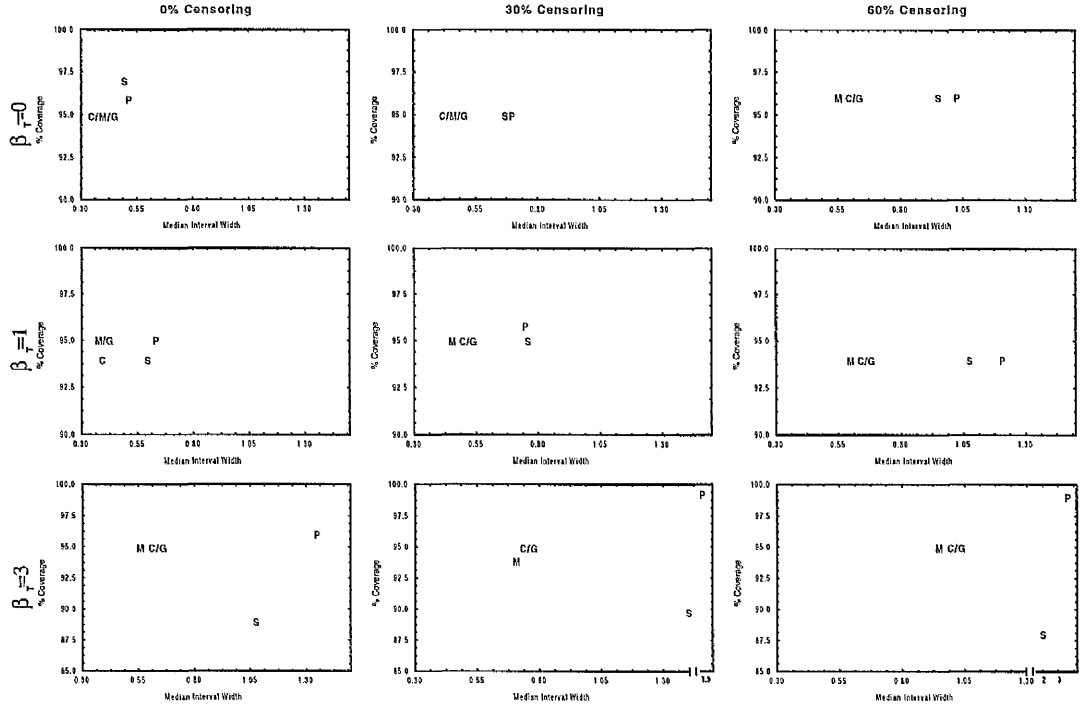


Figure 6.11
Scatterplot of % Coverage by Median Interval Width for 250 Pair Matched Simulation Configurations.



6.6.1.3 Assessing the Performance of the Estimated Standard Errors for the Primary Variable Coefficient Estimate.

A further important area of comparison of the models is in terms of the (estimated) standard error of the estimates of the 'regression' coefficients which is achieved by investigating how well the estimated distribution of the standard error approximates the 'true' sampling variability of the estimates. This is achieved by direct comparison of the 'sample' standard deviation of $\hat{\beta}_T$ across the 1000 simulations with the 'mean' of the individual estimated standard errors of $\hat{\beta}_T$ from each of the 1000 simulations.

The estimated standard deviation of $\hat{\beta}_T$, and the mean of the estimated standard errors of the $\hat{\beta}_T$ as well as the ratio of these two quantities are given in Tables 6.7, 6.8 and 6.9 for all configurations. Note, any value of the ratio greater than 1 suggests that the model is under-estimating the 'true' sampling variability of $\hat{\beta}_T$ while a value less than 1 suggests that the model is over-estimating the 'true' sampling variability of $\hat{\beta}_T$.

The results from Table 6.7 for the ***25 pair*** configurations suggest that all the models in general underestimate the 'true' sampling variability of $\hat{\beta}_T$. The PP and SPH models have the poorest performance in particular for the 60% censoring and β_T=3 configuration.

When the number of pairs increases to ***100*** there is, as expected, better agreement between the estimated and the 'true' sampling variability of the estimates for all models, with the CPH, MPH and GPH displaying the best agreement.

Table 6.7. $\sqrt{\text{var}(\hat{\beta}_T)}$ and mean estimated standard error for all Matched simulation configurations with 25 Pairs.

		0% Censoring			30% Censoring			60% Censoring		
β_T	Model	$\sqrt{\text{var}(\hat{\beta}_T)}$	mean ese($\hat{\beta}_T$)	$\frac{\sqrt{\text{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$	$\sqrt{\text{var}(\hat{\beta}_T)}$	mean ese($\hat{\beta}_T$)	$\frac{\sqrt{\text{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$	$\sqrt{\text{var}(\hat{\beta}_T)}$	mean ese($\hat{\beta}_T$)	$\frac{\sqrt{\text{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$
0										
	PP	0.257	0.217	1.18	49.591	2.772	17.89	55.945	5.694	9.83
	CPH	0.111	0.094	1.18	0.172	0.141	1.22	0.337	0.280	1.20
	MPH	0.111	0.085	1.31	0.172	0.130	1.32	0.337	0.267	1.26
	SPH	0.164	0.165	0.99	0.314	0.280	1.12	3.332	11.848	0.28
	GPH	0.115	0.094	1.22	0.180	0.141	1.28	0.354	0.282	1.26
1										
	PP	13.059	0.774	16.87	342.614	40.50	8.46	545.332	253.668	2.15
	CPH	0.139	0.112	1.24	0.215	0.173	1.24	0.519	0.353	1.47
	MPH	0.139	0.101	1.38	0.215	0.158	1.36	0.519	0.328	1.58
	SPH	0.236	0.214	1.10	1.196	1.553	0.77	11.455	191.563	0.06
	GPH	0.146	0.113	1.29	0.227	0.174	1.30	0.532	0.356	1.49
3										
	PP	871.473	434.609	2.01	1369.207	1212.732	1.13	1855.541	1354.658	1.37
	CPH	0.535	0.369	1.45	1.688	1.293	1.31	0.519	0.353	1.47
	MPH	0.535	0.279	1.92	1.688	1.266	1.33	0.519	0.328	1.58
	SPH	12.972	168.973	0.08	18.938	856.489	0.02	11.455	191.563	0.06
	GPH	1.132	15.946	0.07	0.345	0.299	1.15	0.534	0.356	1.50

Table 6.8. $\sqrt{\text{var}(\hat{\beta}_T)}$ and mean estimated standard error for all Matched simulation configurations with 100 Pairs.

		0% Censoring			30% Censoring			60% Censoring		
β_T	Model	$\sqrt{\text{var}(\hat{\beta}_T)}$	mean ese($\hat{\beta}_T$)	$\frac{\sqrt{\text{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$	$\sqrt{\text{var}(\hat{\beta}_T)}$	mean ese($\hat{\beta}_T$)	$\frac{\sqrt{\text{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$	$\sqrt{\text{var}(\hat{\beta}_T)}$	mean ese($\hat{\beta}_T$)	$\frac{\sqrt{\text{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$
0										
	PP	0.043	0.044	0.98	0.075	0.074	1.01	0.193	0.169	1.14
	CPH	0.021	0.021	1.00	0.029	0.030	0.97	0.052	0.054	0.96
	MPH	0.021	0.020	1.05	0.029	0.029	1.00	0.052	0.052	1.00
	SPH	0.038	0.040	0.95	0.063	0.065	0.97	0.131	0.034	3.85
	GPH	0.021	0.021	1.00	0.030	0.030	1.00	0.053	0.055	0.96
1										
	PP	0.060	0.058	1.03	0.130	0.105	1.24	13.309	0.416	31.99
	CPH	0.025	0.025	1.00	0.040	0.037	1.08	0.062	0.066	0.94
	MPH	0.025	0.024	1.04	0.040	0.034	1.18	0.062	0.062	1.00
	SPH	0.049	0.050	0.98	0.088	0.083	1.06	0.182	0.176	1.03
	GPH	0.026	0.025	1.04	0.043	0.036	1.19	0.064	0.066	0.97
3										
	PP	16.416	1.484	11.06	125.566	15.251	8.23	574.487	232.412	2.47
	CPH	0.061	0.063	0.97	0.097	0.095	1.02	0.191	0.176	1.09
	MPH	0.061	0.058	1.05	0.097	0.088	1.10	0.191	0.168	1.14
	SPH	0.505	0.326	1.55	2.503	2.946	0.85	10.154	76.509	0.13
	GPH	0.069	0.064	1.08	0.103	0.096	1.07	0.196	0.178	1.10

Table 6.9. $\sqrt{var(\hat{\beta}_T)}$ *and mean estimated standard error for all Matched simulation configurations with 250 Pairs.*

		0% Censoring			30% Censoring			60% Censoring		
β_T	Model	$\sqrt{var(\hat{\beta}_T)}$	mean $ese(\hat{\beta}_T)$	$\frac{\sqrt{var(\hat{\beta}_T)}}{\text{mean } ese(\hat{\beta}_T)}$	$\sqrt{var(\hat{\beta}_T)}$	mean $ese(\hat{\beta}_T)$	$\frac{\sqrt{var(\hat{\beta}_T)}}{\text{mean } ese(\hat{\beta}_T)}$	$\sqrt{var(\hat{\beta}_T)}$	mean $ese(\hat{\beta}_T)$	$\frac{\sqrt{var(\hat{\beta}_T)}}{\text{mean } ese(\hat{\beta}_T)}$
0										
	PP	0.016	0.017	0.94	0.027	0.027	1.00	0.059	0.059	1.00
	CPH	0.008	0.008	1.00	0.011	0.012	0.92	0.020	0.021	0.95
	MPH	0.008	0.008	1.00	0.011	0.011	1.00	0.020	0.020	1.00
	SPH	0.015	0.016	0.94	0.025	0.026	0.96	0.052	0.052	1.00
	GPH	0.008	0.008	1.00	0.011	0.012	0.92	0.020	0.021	0.95
1										
	PP	0.023	0.022	1.05	0.035	0.037	0.95	0.090	0.082	1.10
	CPH	0.010	0.010	1.00	0.013	0.014	0.93	0.026	0.025	1.04
	MPH	0.010	0.010	1.00	0.013	0.013	1.00	0.026	0.024	1.08
	SPH	0.019	0.020	0.95	0.029	0.032	0.91	0.065	0.067	0.97
	GPH	0.010	0.010	1.00	0.014	0.014	1.00	0.027	0.025	1.08
3										
	PP	0.147	0.132	1.11	1.405	0.351	4.00	60.872	7.315	8.32
	CPH	0.025	0.024	1.04	0.037	0.036	1.03	0.073	0.067	1.09
	MPH	0.025	0.023	1.09	0.037	0.034	1.09	0.073	0.064	1.14
	SPH	0.074	0.079	0.94	0.151	0.136	1.11	1.371	1.018	1.35
	GPH	0.025	0.024	1.04	0.038	0.036	1.06	0.075	0.066	1.14

Finally, for the ***250 pair*** configuration results (Table 6.9) there is again better agreement due to the larger sample sizes where, once again, all the models tend to overestimate the variability of $\hat{\beta}_T$ except perhaps the SPH model.

The MPH and CPH models appear to have the best 'asymptotic' performance with the MPH model best for the 0% and 30% censoring configurations while the CPH model best for the 60% censoring configurations. The PP and SPH models continue to perform worst of all the methods presented, in particular for the high censoring, large effect size configuration.

6.6.1.4 Matched Study Conclusion

In general across the range of configurations, the CPH and MPH models appear to perform best across all the criteria. The CPH model performs well across all the simulation configurations. The MPH model however has, on average, smaller estimated standard errors (and subsequently narrower intervals) while still maintaining effective 95% coverage/confidence.

The GPH model exhibited comparable if slightly poorer performance in terms of bias and coverage to the CPH and MPH models but it is computationally more complex to fit than each of these and there is no strong suggestion that it is of real practical value for matched studies in conditions similar to the configurations investigated here.

6.6.2 Paired Survival Simulation Study

As indicated earlier, a second simulation study was carried out to compare the performance of the models when analysing ***paired*** survival data. The mechanism for generating the data was similar to that for the matched study except that in this instance both (the perfectly matched) matching variables are 'hidden' from the analysis in order to mimic a real life paired study.

Again, the basis for model comparison will be the bias distribution, coverage, interval width distribution and asymptotic performance of each model under each simulation configuration. The median bias, percentage coverage and median interval width for

each model for the various pair size, censoring and effect size configurations are given in Table 6.10.

6.6.2.1 Comparing Models in terms of Bias

As in the matched simulation, bias tends to decrease with increasing sample size while increasing as the censoring gets larger. Boxplots of the bias distribution for each model for all the simulation conditions are given in Figures 6.12, 6.13 and 6.14.

The overall pattern from these graphs is that the 'bias' appears normally distributed for all models. There is a strong suggestion that the CPH and MPH models tend to under estimate β_T for all configurations except $\beta_T = 0$ and this under estimation tends to increase as β_T increases. For example, for the 25 pairs configurations with no censoring, the CPH model tends to under estimate the true value by 10-15% on average when $\beta_T = 1$.

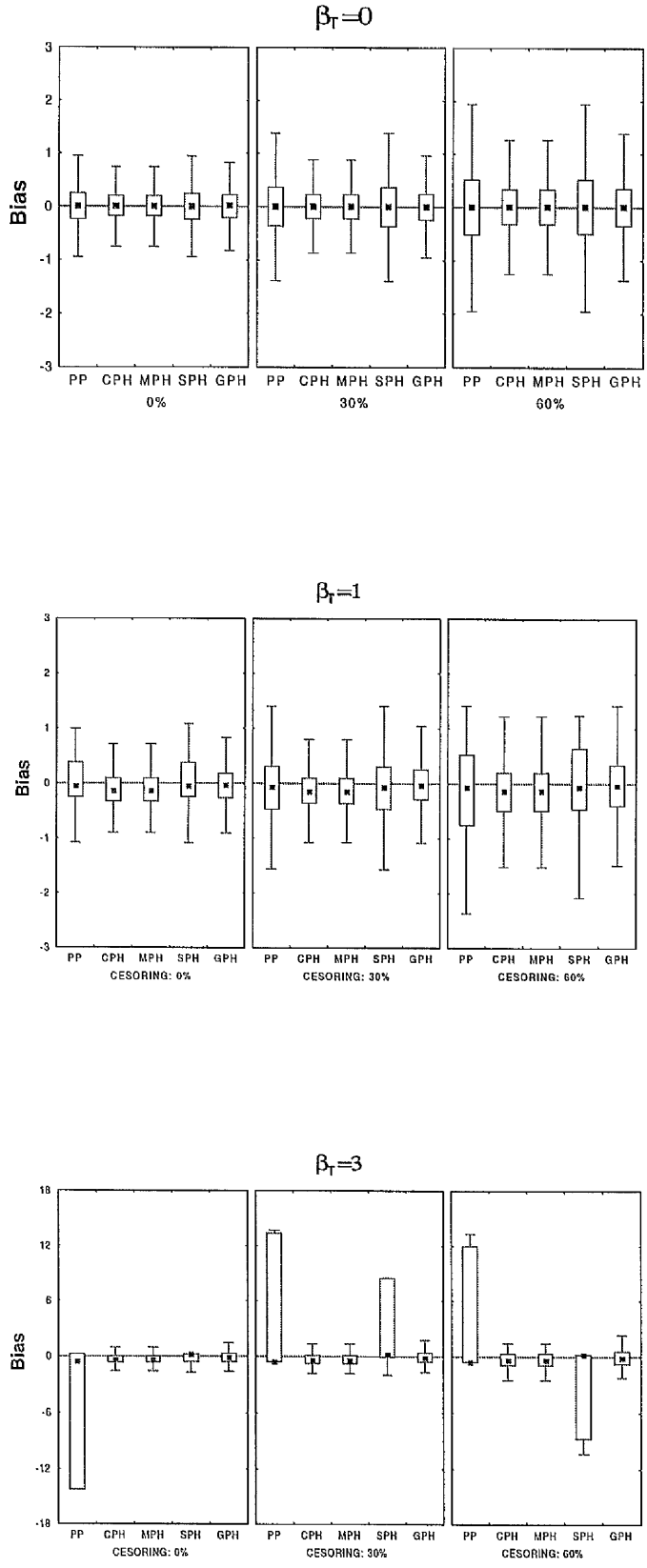
The SPH, PP and to some extent the GPH model do not appear to be biased across the simulation configurations, if indeed biased at all.

When considering the ***25 pair*** simulations, there is no real suggestion of bias for any of the models fitted when estimating in the case of $\beta_T = 0$. The CPH, MPH and GPH have similar distributions to each other while the PP and SPH models exhibit larger variability.

Table 6.10. Median Bias, % Coverage (%Cov) and Median Interval Width (IW) for the Paired Data Simulation Configurations.

			25 Pairs			100 Pairs			250 Pairs		
β	% Cen	Model	Median Bias	% Cov	Median IW	Median Bias	% Cov	Median IW	Median Bias	% Cov	Median IW
0	***0***										
		PP	0.000	94	1.60	0.000	95	0.79	-0.008	95	0.50
		CPH	0.005	96	1.15	-0.002	98	0.56	-0.003	98	0.35
		MPH	0.005	92	1.03	-0.002	95	0.50	-0.003	95	0.31
		SPH	0.000	94	1.58	0.000	95	0.78	-0.008	96	0.50
		GPH	0.011	95	1.18	-0.001	96	0.58	-0.004	96	0.37
	30										
		PP	0.000	96	2.06	0.000	95	1.01	0.000	95	0.64
		CPH	-0.003	97	1.40	0.006	96	0.67	0.000	97	0.42
		MPH	-0.003	93	1.26	0.006	93	0.61	0.000	96	0.38
		SPH	0.000	96	2.03	0.000	95	1.00	0.000	95	0.63
		GPH	0.001	95	1.43	0.009	95	0.70	0.002	97	0.44
	60										
		PP	0.000	99	2.99	0.000	96	1.43	0.000	96	0.91
		CPH	-0.002	96	1.91	0.009	97	0.90	0.002	96	0.56
		MPH	-0.002	93	1.76	0.009	96	0.84	0.002	96	0.91
		SPH	0.000	99	2.99	0.000	97	1.43	0.000	96	0.90
		GPH	-0.002	95	1.95	0.001	96	0.93	0.002	95	0.58
1	***0***	PP	-0.056	96	1.76	0.005	94	0.89	0.008	95	0.56
		CPH	-0.140	93	1.22	-0.172	81	0.60	-0.172	56	0.37
		MPH	-0.140	89	1.08	-0.172	75	0.53	-0.172	49	0.34
		SPH	-0.056	97	1.75	-0.005	94	0.88	-0.016	95	0.55
		GPH	-0.037	93	1.28	-0.070	90	0.63	-0.066	86	0.40
	30										
		PP	-0.061	96	2.32	0.005	95	1.14	0.012	96	0.72
		CPH	-0.144	94	1.49	-0.171	85	0.72	-0.172	67	0.45
		MPH	-0.144	89	1.33	-0.171	81	0.66	-0.172	63	0.41
		SPH	-0.061	95	2.26	-0.002	95	1.13	-0.015	96	0.71
		GPH	-0.031	94	1.54	-0.070	92	0.75	-0.063	90	0.47
	60										
		PP	-0.064	99	3.39	0.009	96	1.63	0.009	95	1.03
		CPH	-0.137	94	2.04	-0.168	90	0.95	-0.172	80	0.60
		MPH	-0.137	90	1.88	-0.168	87.1	0.90	-0.172	77	0.56
		SPH	-0.061	97	3.28	0.004	96	1.59	-0.013	95	1.01
		GPH	-0.034	94	2.10	-0.067	94	0.99	-0.070	91	0.62
3	***0***	PP	-0.558	98	3.99	0.157	98	1.20	0.161	96	1.26
		CPH	-0.365	86	1.88	-0.413	55	0.89	-0.422	19	0.55
		MPH	-0.365	81	1.70	-0.143	52	0.83	-0.422	17	0.53
		SPH	0.178	95	4.00	-0.156	95	1.80	-0.088	91	1.12
		GPH	-0.155	90	1.98	-0.171	83	0.96	-0.182	71	0.60
	30										
		PP	-0.563	53	143.97	0.160	99	2.79	0.161	98	1.62
		CPH	-0.364	88	2.33	-0.406	67	1.08	-0.423	33	0.67
		MPH	-0.364	81	2.08	-0.406	64	1.02	-0.423	30	0.64
		SPH	0.180	99	4.06	-0.054	96	2.31	-0.080	94	1.42
		GPH	-0.155	92	2.45	-0.179	85	1.15	-0.179	76	0.72
	60										
		PP	-0.563	73	203.61	0.163	100	3.95	0.160	98	2.29
		CPH	-0.363	89	3.26	-0.410	76	1.46	-0.416	55	0.90
		MPH	-0.363	72	2.67	-0.410	73	1.37	-0.416	53	0.86
		SPH	0.183	66	320.07	-0.059	91	3.96	-0.093	94	2.00
		GPH	-0.158	92	3.23	-0.181	89	1.53	-0.181	85	0.94

Figure 6.12
Boxplot of Bias for 25 pair Paired simulation configurations.



There is a suggestion that the CPH and MPH models are biased when looking at the ***100 pair*** simulations and this bias tends to increase as β_T increases (Figure 6.13). There is a suggestion that the GPH model is also underestimating the true effect whereas the SPH and PP models appear relatively unbiased with a suggestion that the PP model underestimates the true effect when there was 60% censoring. This could be a consequence of the maximisation procedure not converging when fitting the PP model (as evidenced by the skewed distribution in Figure 6.13) due to the number of observations being small due to censoring.

When considering the performance of all the models for the ***250 pair*** simulations the CPH and MPH model again tend to underestimate the true effect (Figure 6.14). The GPH model appears to perform marginally better in terms of bias while the SPH and PP models do not appear to be biased across any of the simulations. The variability in the bias distribution is similar across the CPH, MPH and GPH models while the SPH and PP model have similarly large variability to each other.

Figure 6.13
Boxplot of Bias for 100 Pairs Paired simulation configurations.

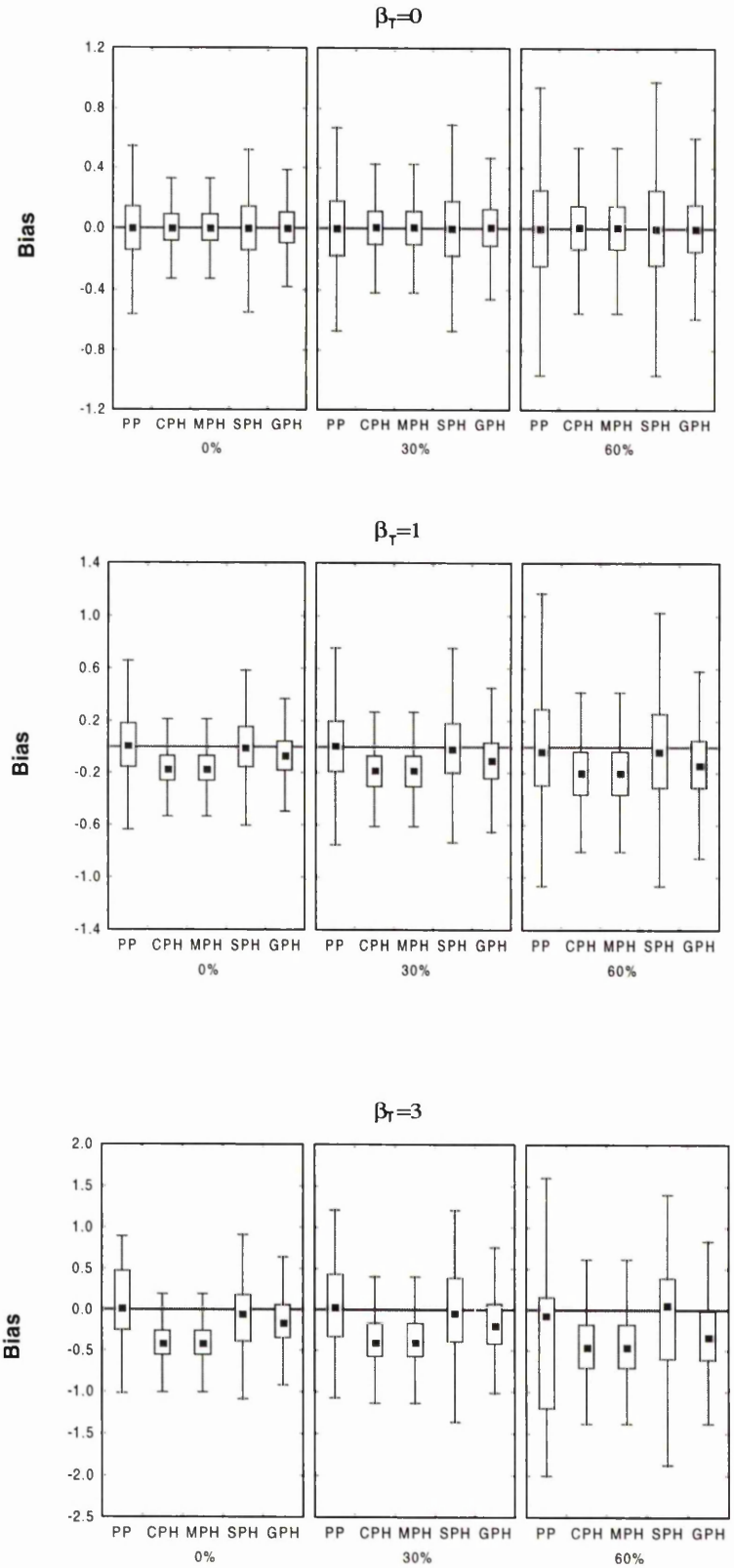
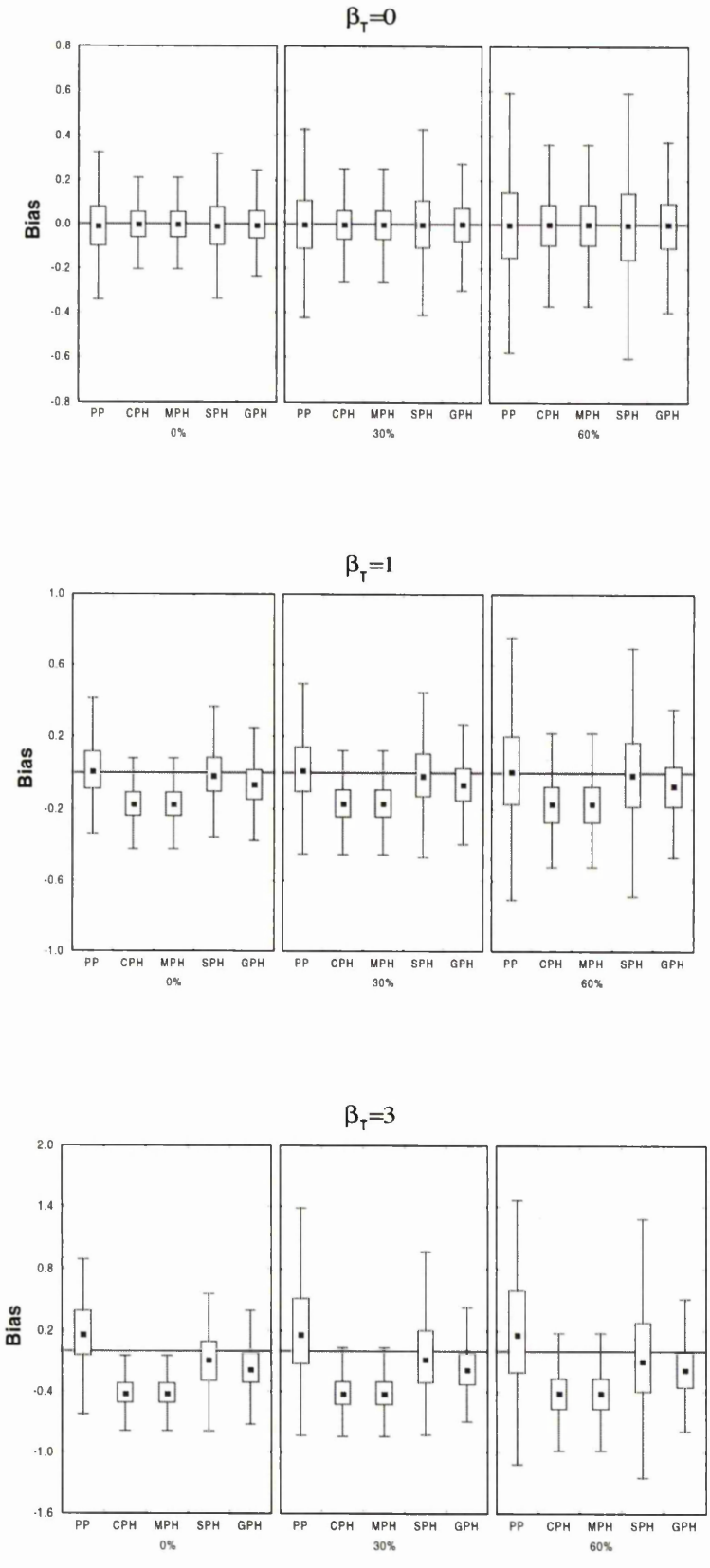


Figure 6.14
Boxplot of Bias for 250 Pairs Paired simulation configurations.



6.6.2.2 Comparison of Confidence Interval Widths and Coverage Rates

In order to determine the 'best' model (i.e. the model that has the highest coverage coupled with the smallest interval width), labelled scatterplots of the median interval width against percentage coverage for each simulation configuration are given in Figures 6.15,6.16 and 6.17.

There is again a strong suggestion that the CPH and MPH models are not the best candidates to consider when analysing paired data as both models have poor coverage despite relatively small interval widths. There is a suggestion that the GPH model underestimates β_T but is clearly an improvement on both the CPH and MPH as its coverage is, in general, much closer to the nominal 95%.

The SPH and PP models have consistently high coverage (except perhaps the PP model for the 25 pairs, high censoring configuration). This high coverage is achieved by considerably wider intervals with, for example, the median interval width for the GPH model being nearly half that of the SPH and PP across most of the configurations.

Figure 6.15.
Scatterplot of % Coverage by Median Interval Width for 25 Pairs
Paired Simulation Configurations.

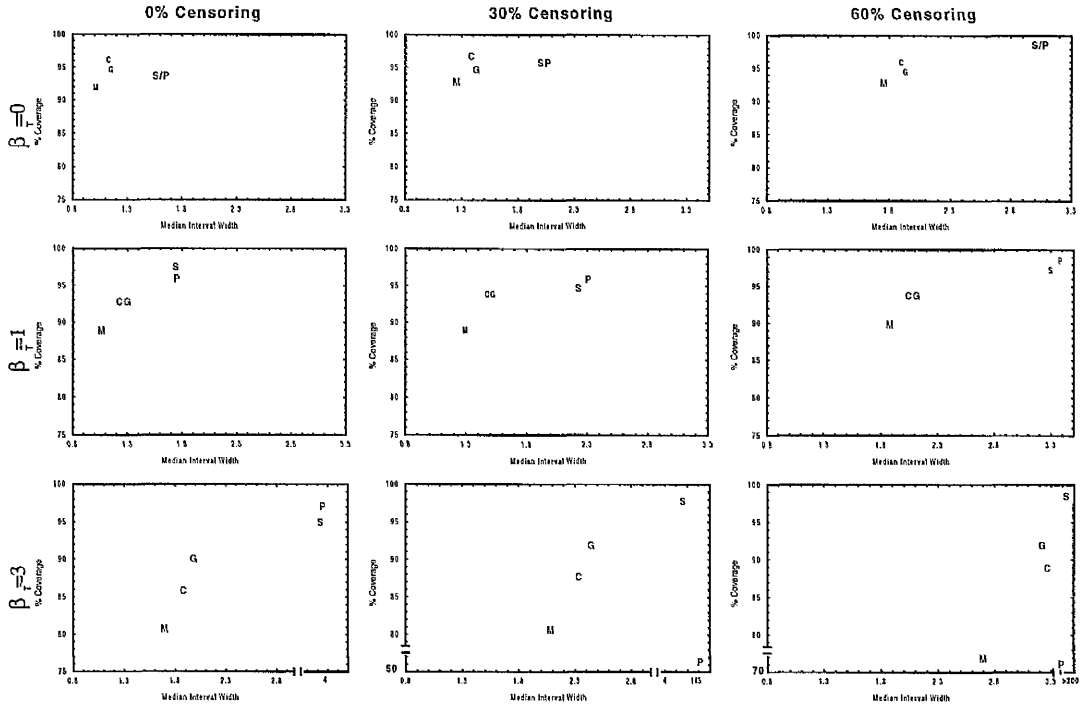


Figure 6.16.
Scatterplot of % Coverage by Median Interval Width for 100 Pairs
Paired Simulation Configurations

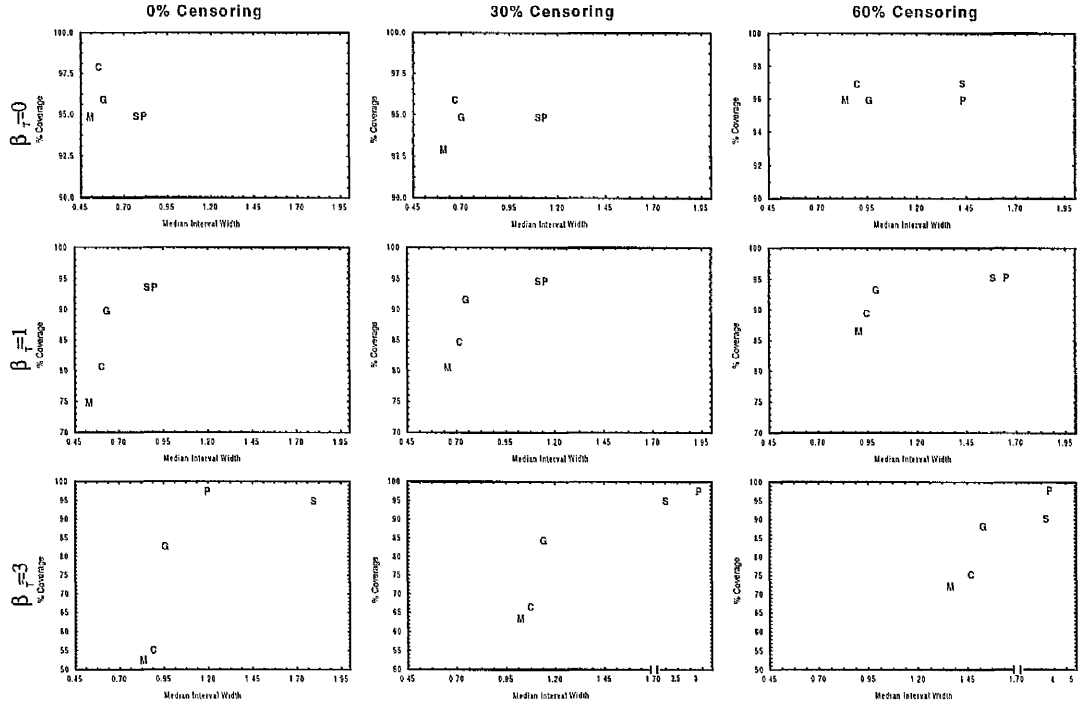
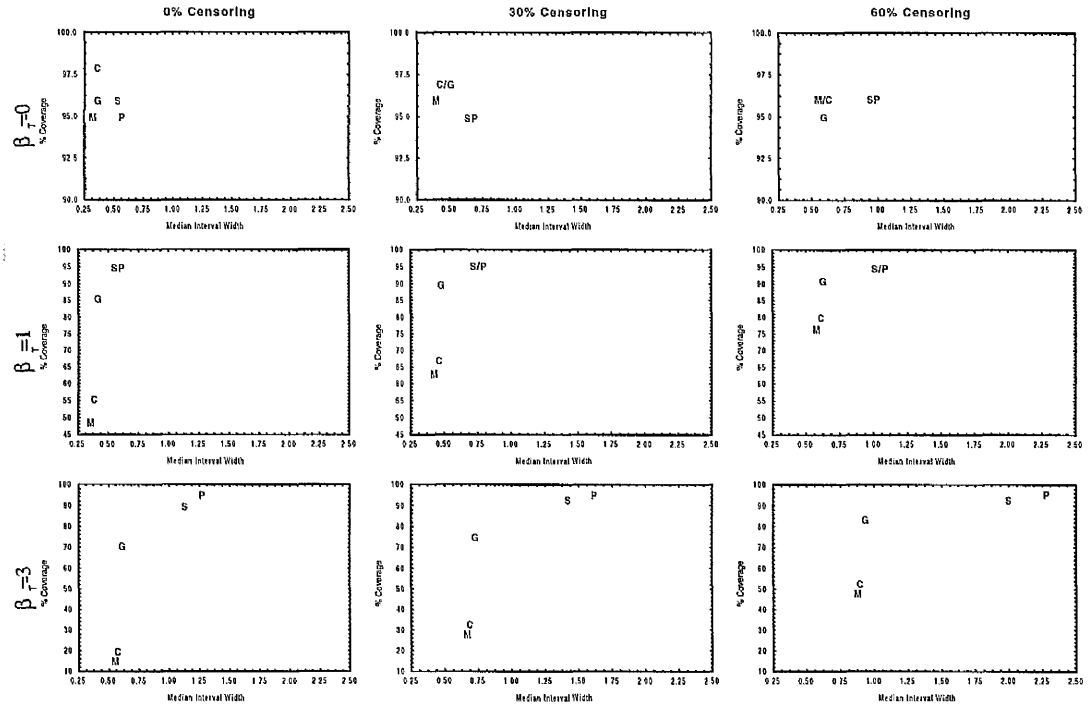


Figure 6.17.
Scatterplot of % Coverage by Median Interval Width for 250 Pairs Paired Simulation Configurations.



6.6.2.3 Assessing the Performance of the Estimated Standard Errors for the Primary Variable Coefficient Estimate.

The 'sample' standard deviation across all 1000 estimates of $\hat{\beta}_T$ and the mean of the estimated standard errors of the $\hat{\beta}_T$ as well as the ratio of these two quantities are illustrated in Tables 6.11, 6.12 and 6.13. This allows a comparison of how well each model performs in terms of the estimated standard error compared to the 'true' sampling variability of the estimates for the various sample configurations.

Table 6.11. $\sqrt{\text{var}(\hat{\beta}_T)}$ *and mean estimated standard error for all Paired simulation configurations with 25 Pairs.*

		0% Censoring			30% Censoring			60% Censoring		
β_T	**Model**	$\sqrt{\text{var}(\hat{\beta}_T)}$	mean $\text{ese}(\hat{\beta}_T)$	$\frac{\sqrt{\text{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$	$\sqrt{\text{var}(\hat{\beta}_T)}$	mean $\text{ese}(\hat{\beta}_T)$	$\frac{\sqrt{\text{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$	$\sqrt{\text{var}(\hat{\beta}_T)}$	mean $\text{ese}(\hat{\beta}_T)$	$\frac{\sqrt{\text{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$
0										
	PP	0.191	0.169	1.13	0.316	0.289	1.09	3.839	8.338	0.46
	CPH	0.085	0.087	0.98	0.119	0.129	0.92	0.263	0.246	1.07
	MPH	0.085	0.069	1.23	0.119	0.105	1.13	0.263	0.207	1.27
	SPH	0.186	0.166	1.12	0.306	0.283	1.08	3.649	15.843	0.23
	GPH	0.103	0.092	1.12	0.147	0.137	1.07	0.305	0.261	1.17
1										
	PP	0.263	0.226	1.16	2.778	1.761	1.58	18.282	98.819	0.19
	CPH	0.094	0.099	0.95	0.147	0.148	0.99	0.388	0.321	1.21
	MPH	0.094	0.077	1.22	0.147	0.119	1.24	0.388	0.243	1.60
	SPH	0.244	0.217	1.12	1.668	2.382	0.70	11.503	197.906	0.06
	GPH	0.121	0.107	1.13	0.185	0.162	1.14	0.630	5.153	0.12
3										
	PP	41.870	116.30	0.36	47.103	498.231	0.09	34.29	1746.590	0.02
	CPH	0.553	0.326	1.70	1.228	0.860	1.43	6.662	19.546	0.34
	MPH	0.553	0.214	2.58	1.228	0.334	3.68	6.662	5.061	1.32
	SPH	13.333	180.106	0.07	18.839	897.247	0.02	19.223	3767.409	0.01
	GPH	2.048	7.587	0.27	4.600	237.723	0.02	30.457	8652.114	0.00

Table 6.12. $\sqrt{\text{var}(\hat{\beta}_T)}$ *and mean estimated standard error for all Paired simulation configurations with 100 Pairs.*

		0% Censoring			30% Censoring			60% Censoring		
β_T	**Model**	$\sqrt{\text{var}(\hat{\beta}_T)}$	mean $\text{ese}(\hat{\beta}_T)$	$\frac{\sqrt{\text{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$	$\sqrt{\text{var}(\hat{\beta}_T)}$	mean $\text{ese}(\hat{\beta}_T)$	$\frac{\sqrt{\text{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$	$\sqrt{\text{var}(\hat{\beta}_T)}$	mean $\text{ese}(\hat{\beta}_T)$	$\frac{\sqrt{\text{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$
0										
	PP	0.043	0.041	1.05	0.069	0.067	1.03	0.133	0.137	0.97
	CPH	0.017	0.021	0.81	0.027	0.029	0.93	0.045	0.052	0.87
	MPH	0.017	0.016	1.06	0.027	0.024	1.13	0.045	0.046	0.98
	SPH	0.042	0.040	1.05	0.066	0.065	1.02	0.130	0.134	0.97
	GPH	0.022	0.022	1.00	0.032	0.032	1.00	0.052	0.056	0.93
1										
	PP	0.056	0.052	1.08	0.092	0.087	1.06	0.360	0.211	1.71
	CPH	0.022	0.023	0.96	0.031	0.033	0.94	0.054	0.060	0.90
	MPH	0.022	0.019	1.16	0.031	0.029	1.07	0.054	0.053	1.02
	SPH	0.054	0.051	1.06	0.088	0.084	1.05	0.297	0.220	1.35
	GPH	0.031	0.026	1.19	0.044	0.037	1.19	0.068	0.065	1.05
3										
	PP	2.354	0.487	4.83	17.105	5.833	2.93	45.152	90.959	0.50
	CPH	0.052	0.052	1.00	0.083	0.078	1.06	0.166	0.148	1.12
	MPH	0.052	0.046	1.13	0.083	0.070	1.19	0.166	0.131	1.27
	SPH	0.503	0.335	1.50	2.581	3.281	0.79	11.096	97.629	0.11
	GPH	0.089	0.061	1.46	0.129	0.089	1.45	0.212	0.162	1.31

Table 6.13. $\sqrt{\mathrm{var}(\hat{\beta}_T)}$ and mean estimated standard error for all Paired simulation configurations with 250 Pairs.

β_T	Model	0% Censoring			30% Censoring			60% Censoring		
		$\sqrt{\mathrm{var}(\hat{\beta}_T)}$	mean $\mathrm{ese}(\hat{\beta}_T)$	$\frac{\sqrt{\mathrm{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$	$\sqrt{\mathrm{var}(\hat{\beta}_T)}$	mean $\mathrm{ese}(\hat{\beta}_T)$	$\frac{\sqrt{\mathrm{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$	$\sqrt{\mathrm{var}(\hat{\beta}_T)}$	mean $\mathrm{ese}(\hat{\beta}_T)$	$\frac{\sqrt{\mathrm{var}(\hat{\beta}_T)}}{\text{mean ese}(\hat{\beta}_T)}$
0										
	PP	0.015	0.016	0.94	0.026	0.026	1.00	0.052	0.054	0.96
	CPH	0.006	0.008	0.75	0.009	0.012	0.75	0.019	0.021	0.90
	MPH	0.006	0.006	1.00	0.009	0.010	0.90	0.019	0.018	1.06
	SPH	0.015	0.016	0.94	0.026	0.026	1.00	0.051	0.053	0.96
	GPH	0.008	0.009	0.89	0.011	0.013	0.85	0.023	0.022	1.05
1										
	PP	0.021	0.021	1.00	0.031	0.034	0.91	0.073	0.071	1.03
	CPH	0.008	0.009	0.89	0.012	0.013	0.92	0.021	0.023	0.91
	MPH	0.008	0.007	1.14	0.012	0.011	1.09	0.021	0.021	1.00
	SPH	0.021	0.020	1.05	0.030	0.033	0.91	0.073	0.071	1.03
	GPH	0.014	0.010	1.40	0.017	0.015	1.13	0.067	0.068	0.99
3										
	PP	0.120	0.107	1.12	0.635	0.215	2.95	11.53	2.45	4.71
	CPH	0.020	0.020	1.00	0.029	0.029	1.00	0.052	0.054	0.96
	MPH	0.020	0.019	1.05	0.029	0.027	1.07	0.052	0.050	1.04
	SPH	0.091	0.084	1.08	0.236	0.160	1.48	1.422	1.112	1.28
	GPH	0.043	0.024	1.79	0.054	0.034	1.59	0.079	0.060	1.32

As with the matched simulation results, all of the models tend to under-estimate the 'true' variability of $\hat{\beta}_T$. However in these simulations there is better agreement than for the matched scenario as the sample size increases. The PP, SPH and GPH models have poor performance when the number of pairs is small, the censoring high and the primary variable effect large.

The MPH and GPH models appear to best estimate the 'true' standard error when considering the 250 pair results in comparison to the SPH and PP models which tended to be slightly under estimate the 'true' standard error. The reduction in the mean of the estimated standard deviation for the MPH model when compared to the CPH model is evident and possibly due to the dependency structure.

6.6.2.4 Paired Study Conclusions

The CPH and MPH models are shown to be poor candidates for analysing paired survival study data as both are substantially biased (possibly due to model mis-specification given the 'hidden' covariates) and consequently have poor coverage.

The GPH model performs considerably better than both the CPH and MPH models in terms of increased coverage while retaining similar interval estimate widths. The SPH and PP models, which address the dependency by considering each pair as a separate stratum, consistently maintain the nominal 95% coverage but their interval estimates are rather wide, in particular for small sample sizes with 'large' censoring.

In order therefore to determine the best model there is a trade off between the SPH and PP model which have good coverage but wide interval estimates and the GPH model which has considerably narrower interval estimates (and subsequently slightly poorer coverage), and is not unduly affected by large censoring.

6.7 Assessing the Effect of The Degree of Association within a 'Pair'

The primary aim of this chapter is to identify which of the methods proposed in this thesis are best suited for matched or paired survival studies in order to provide guidelines for their general use. An additional question of interest is what role the degree of association plays in this comparison. In order to answer this, a strategy for simulating matched and paired survival data for predefined degrees of dependency is needed.

One such simulation strategy for specifically comparing 'naturally matched' cluster studies (i.e. familial or litter studies with possibly varying numbers in each cluster) with different levels of dependency uses a multivariate survival model (Oakes 1982, Clayton 1978, Clayton and Cuzick 1985) to simulate the data.

The model is defined as

$$S(t_1,\ldots,t_m)=\left(\sum_{i=1}^{m} S_i^{1-\phi}(t_i)-(m-1)\right)^{-\frac{1}{\phi-1}}$$

where $S(t_1, \ldots, t_m)$ is the joint survival function for the m members in any cluster and

$$S_i(t_i)=\exp\left\{\int_0^{t_i}\lambda_0(s)\exp(\beta' z_i)ds\right\}$$

is the marginal survival function for i^{th} individual in a cluster where z_i is the vector of covariates for an individual having an event at time t_i. The level of dependence between survival times of members in a pair is measured by ϕ. Note that in the multivariate survival time model, independence is represented with $\phi=1$ while in the GPH model a $\theta=0$ represents independence (recall that in the GPH model θ is the variance of the gamma distributed random effect). No guidelines are available in the literature however as to what represents low, medium and high dependency although as mentioned earlier in section 5.9.2.1, the estimate of θ in a GPH model can be 'translated' into an estimate of Kendall's τ (Oakes 1982 and Klein 1997).

One simulation study using this multivariate survival model (King et al 1996) was carried out to specifically compare the CPH and MPH models for differing levels of dependency. King's simulations strictly concerned cluster studies with varying number of individuals in a cluster. This simulation investigated the effect of analysing such dependent survival data under the false assumption of independence using a PH model compared to using methods (e.g. the Marginal PH model) specifically designed to analyse dependent data. In King's study the level of dependency was chosen to as $\phi=1$, 3 and 5 which, according to the authors, represented independence, moderate and finally high dependence.

In order to compare the results and guidelines arising from the simulation study presented in this chapter to those suggested by King (1996), an indication of the level of dependency from using the strategy outlined in section 6.3 (where the dependency is a direct consequence of the matching variables) is needed. For the matched data

simulations the estimate of θ from the GPH model across all configurations ranged from 0.71 to 1.21 (corresponding to an estimated Kendall's τ of 0.26 to 0.38) while for the paired simulations the estimate of θ ranged from 0.82 to 1.74 (corresponding to an estimated Kendall's τ of 0.29 to 0.47) across all configurations. These estimates suggest that the degree of dependency for both the matched and paired simulations can be considered 'moderate' where the lower level of dependency in the matched compared to the paired simulations is probably due to the effect of the unmatched covariate.

As the results in this thesis relate, in some sense, to matched and paired survival studies with moderate dependency a suggestion as to how the various models might perform with increasing (and no) dependency can be ascertained from King's study.

In summary, King's simulation study showed that both the PH and MPH models provide unbiased estimates of the effect of the primary variable but with estimated standard errors under-estimated as the level of dependence increased. In particular, the PH model showed increasing Type 1 errors as the dependency structure increased suggesting possible model mis-specification. The MPH model had high efficiency (in terms of the standard deviation of the estimated coefficients across the simulations being similar to the mean estimated standard errors of the coefficient across the simulations) for $\beta_T=0$ but this efficiency decreased as β_T increased. In conclusion, King suggested that the MPH model should be recommended over the PH model when analysing cluster survival studies.

The simulation study undertaken in this thesis suggested that the MPH model is more suited to matched than to paired studies. One reason as to why the MPH model performed poorly for the paired study may be that, given the moderate degree of dependency present, model mis-specification arises from the 'hidden' matching variables having a greater 'influence' on the estimation of the primary variable than the modelling of the dependency alone. This appears to consolidate the results from King's study where the MPH performed well when there was low to moderate dependency present but showed evidence of poor performance with increasing dependency.

The clear suggestion from both studies however, is that the PH model should not be used to analyse dependent survival data as it has no 'mechanism' for dealing with such, although the PH model will perform quite well when there is a small degree of dependency (King 1996).

An additional reason for using the GPH or MPH model over the PH model is that both reduce to the PH model when modelling independent survival data thus eliminating the issue of the magnitude of the degree of dependency. Furthermore, King's results suggested the somewhat obvious fact that there is no loss in power or efficiency in using the MPH and GPH models when modelling independent survival data.

In conclusion, given King's results and the results of the simulations in this thesis, the MPH model appears the best suited approach to analysing matched data while the GPH model appears the best suited for paired survival studies.

6.8 Availability of Proposed Models in Statistical Software

Some final points worth addressing in this chapter are the 'availability' of each model in 'standard' statistical software packages, and the time taken (as assessed by the Central Processing Unit) to fit each model.

None of the tests presented in Chapter 4 are available directly in commercial software while an indication of the availability of the models in commonly available software is given in Table 6.14.

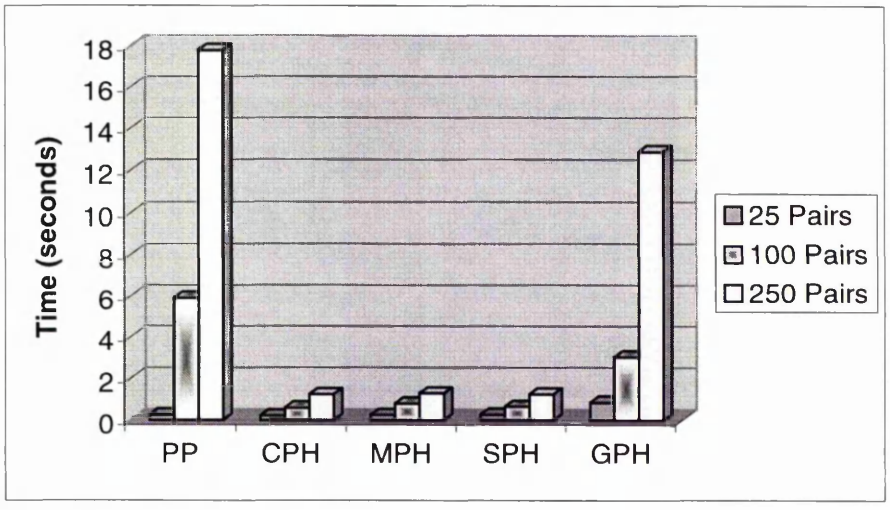
Table 6.14 Availability of Models for major software packages

	PP	CPH	MPH	SPH	GPH
Minitab (v 12)	✓				
Splus (v 4.5)	✓	✓	✓	✓	*
SAS	✓	✓	✓	✓	
Statistica (v 99)	✓	✓		✓	
STATA (v 6.0)	✓	✓		✓	
SPSS (v 8.0)	✓	✓		✓	

** Frailty Models (i.e. GPH) will be available in Splus 2000*

The simulations described in this chapter were carried out on a Sun Ultra workstation. Figure 6.18 displays the average CPU time (in seconds) taken (across all simulations configurations) to fit each of the models considered in this thesis. Not surprisingly the average CPU increases for all the models with increasing sample sizes. The CPH, MPH and SPH take the smallest amount of time to fit while the PP and GPH models are the slowest, markedly so for the 250 pair configurations. However, the mean time taken to fit any of the models was under 20 seconds and therefore of no great concern for practical use in a one-off analysis.

Figure 6.18.
Mean time taken to fit each
model across all simulation configurations.



6.9 Chapter Summary

The primary goal of this chapter was an investigation of the various methods proposed for analysing dependent survival data from matched and paired studies. A strategy was proposed to simulate matched and paired survival data using a proportional hazards model across a variety of configurations defined by the number of pairs, the effect size of the primary variable and the proportion of censored observations. The PH was used to underpin the simulations as all of the models suggested in Chapter 5 assumed proportional hazards and thus removed any suggestion that poor performance may be attributed to this assumption being invalid.

The first half of the chapter dealt with comparing the performance of the 'covariate free' procedures (Chapter 4) for testing a difference in survival between the two arms

of the primary variable (i.e. the case and control). For both the matched and paired scenarios there was a strong suggestion that Akritas' test performed best.

The second half of the chapter compared the various modelling approaches presented in Chapter 5 for comparing survival between the two arms of the primary variable while adjusting for 'known' covariates and hence the dependency structure of the data. The Marginal Proportional Hazards model appeared best suited to analysing matched survival data while the Gamma Frailty Proportional Hazards model appeared best suited to analysing paired survival problems.

In order however to fully understand the behaviour of the various approaches presented in this study, in particular the models presented in Chapter 5, additional simulation studies should be conducted and these will be outlined as further work in the concluding chapter.

Chapter 7

Conclusions and Further Work

7.1 Conclusions

The main aim of this thesis is to present methods for the analysis of dependent survival data, primarily from cluster studies.

Chapter 1 gave a brief introduction to Survival Analysis and presented background information for analysing independent survival data in order to provide a basis for subsequent chapters. The emphasis of Chapter 2 was to introduce survival designs where the common assumption of independence between observations is likely to be invalid. Two illustrative examples, involving data from Melanoma and Dental studies, were used to illustrate the methods presented in this chapter. The Melanoma study, a matched survival study, involved a comparison of the survival prospects of a sample of matched Multiple and Single Melanoma sufferers. The Dental study was an example of a paired survival study, with the aim of the study being to compare the time to orthodontic bracket failure under two different types of bonding cements.

Before any formal analysis is carried out in this, and indeed any context, a graphical representation of the data must be carried out not only to check the validity of underlying assumptions but to allow some subjective impression of the magnitude of the effect on survival of the primary variable and indeed for presentation of results. A

variety of approaches for displaying matched and paired survival data were reviewed or introduced in Chapter 3 including a new non-parametric approach for generating reference ranges.

The emphasis of Chapter 4 was to provide formal methods for analysing matched and paired survival data where emphasis is on the comparison the primary variable alone. A review of the methods available in the current literature and several new approaches were presented based on comparing 'pair performance' as well as the estimation of the distribution of the (pairwise) difference in survival.

A natural extension to the approach adopted in Chapter 4 is to incorporate any additional covariates (in the form of matching variables and/or unmatched covariates) into the analysis in order to give a more 'balanced' assessment of the effect of the primary variable on survival. An introduction to the various approaches available for modelling survival data, in particular through the proportional hazards model, was given in Chapter 5 followed by an extension of the "pair performance" approach and several extensions of the proportional hazards (PH) model to analyse survival data from clustered studies.

In order to compare the various methods presented respectively in Chapters 4 and 5 simulation studies were carried out and the results were presented in Chapter 6. These suggested that when analysing matched and paired survival data using the primary variable alone, the Akritas test is recommended for general use. For the methods presented in Chapter 5 (i.e. incorporating the matching variables and unmatched covariates), the Marginal proportional hazards models appeared best

suited to matched survival studies while the Gamma frailty proportional hazards models appeared best suited to paired survival studies with respect to estimating the effect of the primary variable.

The conclusions reached from the applications of the methods in this thesis to the two example data sets were that there was a slight but non-significant difference in survival between the Multiple and Single Melanoma sufferers, and no significant difference in the time to bracket failure between the Glass Ionomer and chemically-cured cements.

7.2 Further Work

Methods for graphing and analysing matched and paired survival data using a variety of techniques were described in Chapter 3. The following considerations could be the subject of further work:

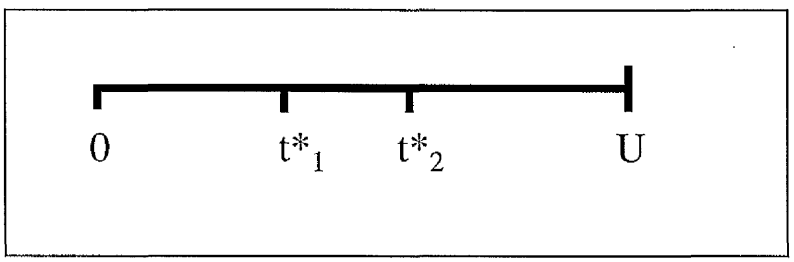
- The methods presented for generating reference ranges in Chapter 3 involved Monte Carlo simulation in order to 'estimate' the reference range rather than to calculate it 'exactly' using all possible permutations. The estimation procedure was based on estimating (using 2.5% percentiles) the upper and lower reference ranges from a random sample of all possible ratios. An alternative reference range that could be investigated is one which uses the maximum and minimum respectively rather than percentiles.

- The Simple Binomial and interval based tests presented in Chapter 4 ignore all contributions made by doubly censored pairs. This can have serious consequences, as illustrated in Chapter 6, when there is a high degree of censoring and a small sample size. In particular, when considering the tests based on estimating the pairwise difference in survival two approaches worth further investigation are:

1) Recast the problem as an interval censored estimation problem by defining an upper bound for the possible maximum survival time. For example, in a survival study involving human lifetimes a sensible upper bound might be 100 years of life. Consider a doubly censored pair with observation times t^*_1 and t^*_2 respectively (Figure 7.1) where a subjective upper bound has been chosen to be U.

Figure 7.1.

Defining an Upper Bound for the Difference in Survival Time.



Let T_1 and T_2 represent the actual (and unknown) event times for the two individuals in the pair respectively. Hence, $t^*_1 \leq T_1 \leq U$ and $t^*_2 \leq T_2 \leq U$ allowing an interval estimate

$$t^*_1 - U \leq T_1 - T_2 \leq U - t^*_2$$

of the true doubly censored difference to be obtained.

An estimate of the survivor function of the differences (and hence an interval estimate of the median difference) can then be based on the event times, singly censored differences and interval censored differences (derived from the doubly censored pairs) using Turnbull's approach (1974). Following this, a sensitivity analysis could be carried out to investigate the upper bound on the interval estimate.

2) Use a kernel density estimation approach to provide a smoothed estimate of the distribution of differences. The normal distribution is an obvious choice for the kernel for actual differences while a positive distribution (e.g. gamma or exponential) or negative distribution (e.g. negative gamma or negative exponential) might be a good choice for the kernel for right and left censored differences respectively. The choice of kernel for doubly censored differences might involve a uniform or normal distribution.

- The simulation study compared the performance of the models proposed using data simulated from a proportional hazards model. A natural question to consider is how well these models behave if the proportional hazards is not valid and what alternative models could be used.

References

Aalen, O.O, Johnson, S (1978). An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*, **5**:141-150.

Aalen, O.O. (1994). The effects of frailty in survival analysis. *Statistical Methods in Medical Research*, **3**, 227-243.

Aitchison, TC, Sirel JM, Watt DC, Mackie_RM. (1995). Prognostic trees to aid prognosis in patients with cutaneous malignant-melanoma. *British Medical Journal*, 1995, vol.311, no.7019, pp.1536-1539.

Akritas, M.G. (1990). Rank Transformed Statistics with Censored Data. *Statistics & Probability Letters*, , 1992, Vol.13, No.3, pp.209-221.

Altshuler, B. (1970). Theory for Measurement of Competing Risks in Animal Experiments. *Mathematical Biosciences*, **6**, 1-11.

Andersen, P.K., Borgan Ø., et al (1993) *Statistical Models Based On Counting Processes*. New York: Springer-Verlag.

Andersen, P.K., Klein, J.P., Knudsen, K.M., Palacios, R.T.Y. (1997). Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics*, **53**, No.4, pp.1475-1484.

Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *Journal of the American Statistical Association*, **83**, 204-212.

Barlow, W.E., Prentice, R.L. (1988). Residuals for relative risk regression. *Biometrika*, **75**, 65-74.

Beran, R. (1981). Regression with Randomly Censored Survival Data. Technical Report. University of California, Berkley.

Borgan, Ø, Liestøl, K. (1990). A Note on Confidence Intervals and Bands for the Survival Curve based on Transformations. *Scandinavian Journal of Statistics*, 17:35-41.

Bowman, A. W., Wright, E.M. (1998). Graphical exploration of covariate effects on survival data through Nonparametric Quantile Curves. Technical Report No.98-11. University of Glasgow.

Bowman, A. W. (1999). Three-dimensional nonparametric model plots. Technical Report No.98-14. University of Glasgow.

Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*, **30**, 89-99.

Breiman, L., Friedman, J. H., et al. (1984). *Classification and Regression Trees.* Belmont, CA: Wadsworth.

Breslow, N.E., Crowley, J.J. (1974). A Large Sample Study of the Life Table and Product Limit Estimates under Random Censorship, *Annals of Statistics*, **2**, 437-453.

Breslow, N.E. (1992). Introduction to: Kaplan and Meier (1958) 'Nonparametric Estimation from Incomplete Observations'. *Breakthrough in Statistics, Vol II.* Kotz and Johnson (Eds), Springer-Verlag, New York.

Cain, K.C., Lange, N.T. (1984). Approximate case-influence for the proportional hazards model with censored data. *Biometrics*, **40**, 493-499.

Campbell, G., Foldes, A. (1982). Large sample propoerties of nonparametric bivariate estimators with censored data. In *Nonparametric Statistical Inference, Colloquia Mathematica-Societatis Janos Bolyai.* North Holland, Amsterdam.

Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiology studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141-151.

Clayon, D.G, Cuzick, J. (1985). Multivariate generalisations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A*, **148**, 82-117.

Collet, D. (1994). *Modelling Binary Data.* Chapman and Hall, London.

Collet, D. (1991). *Modelling Survival Data in Medical Research.* Chapman and Hall, London.

Commenges, D., Andersen, P.K. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis*, **1**, 145-160.

Cox, D.R (1959). The Analysis of Exponentially Distributed Life-times with Two types of Failures. *Journal of the Royal Statistical Society*, Series B, **21**, 411-421.

Cox, D.R (1964). Some Applications of Exponentially Distributed Life-Times with Two Types of Failures. *Journal of the Royal Statistical Society*, **26**, 103-110

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

Cox, D.R. (1975). Partial Likelihood. *Biometrika* **62**, 269-276.

Cox, D.R., Snell, E.J. (1989). *The Analysis of Binary Data.* Chapman and Hall, London, 2nd Edition.

Dabrowska, D.M. (1988). Kaplan-Meier estimates on the plane. *Annals of Statistics*, **16**, 1475-1489.

Dempster, A.P, Laird, N.M, Rubin, D.R. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

Dinse, G.E. (1985). An Alternative to Efron's Redistribution-Of Mass Construction of the Kaplan-Meier Estimator, *American Statistician*, **39**, 299-300.

Drum, M., McCullagh, P. (1993). Comment on regression models for discrete longitudinal responses by G.M. Fitzmaurice, N.M. Laird and A.G. Rotnitzky. *Statistical Science*, **8**, 300-301.

Efron, B. (1967). The Two Sample Problem with Censored Data. *Proceedings of the Fifth Berkely Symposium on Mathematical Statistics and Probability.* New York:Prentice-Hall, (1967): **4**, 831:853.

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

Fleming, T.R., Harrington, D.P.A. (1991). *Counting Process and Survival Analysis.* New York: John Wiley and Sons.

Gehan, E.A. (1965). A Generalised Wilcoxon test for Comparing Arbitrarily Singly-Censored Samples. *Biometrika*, **52**, 650-653.

Gill, R.D. (1980) Censoring and Stochastic Integrals. *Mathematical Centre Tracts.* Amsterdam: Mathematisch Centrum, 124.

Gore, S.M., Poocock, S.J., Kerr, G.R. (1984). Regression models and non-proportional hazards in the analysis of breast cancer survival. *Applied Statistics*, **33**, 176-195.

Grambsch, P.M., Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika,* **81**:515-526.

Greenwood, M (1926) The Natural Duration of Cancer. *Reports on Public Health and Medical Subjects 33.* London: His Majesty's Stationery Office,1-26.

Hald, A. (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. *Skandinavisk Aktuarietidskrift*, 1949, 119-134.

Harrell, F.E., Lee, K.L. (1986). Verifying assumptions of the Cox proportional hazards model. *In Proceeding of the 11th Annual SAS Users Group International Conference, 823-828, Cary, NC, SAS Institute, INC.*

Hosmer, D.W., Lemeshow, S. (1989). *Applied Logistic Regression.* Wiley, New York.

Hosmer, D.W., Lemeshow, S. (1998). *Applied Survival Analysis*. Wiley, New York.

Van Houwelingen, J.C., le Cessie, S. (1988). Logistic Regression, a review. *Statistica Neerlandica,* **42**:215-232, 1988.

Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika* **71**, 75-83.

Hougaard, P (1986a). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387-396.

Hougaard, P (1986b). A class of multivariate failure time distributions. *Biometrika*, **73**, 671-678.

Johansen, S (1983). An extension of Cox's regression model. *International Statistical Review*, **51**, 258-262.

Kalbfleisch, J.D, Prentice, R.L. (1980). *The Statistical analysis of Failure Time Data*. Wiley, New York.

Kaplan, E.L., Meier, P (1958) Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53**: 457-481.

Keiding, N., Andersen, P.K., 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**, No.1-3, pp.215-224.

Kendall, M.G. (1962). *Rank Correlation Methods*, London: Griffin.

King, T.M., Beaty, T.H., Liang, K. (1996). Comparison of methods for survival analysis of dependent data. *Genetic Epidemiology*, **13**, 139-158.

Klein, J.P (1991) Small Sample Moments of some Estimators of the Variance of the Kaplan-Meier and Nelson-Aalen Estimators. *Scandinavian Journal of Statistics*, **18**:333-40.

Klein, J.P. (1992). Semi-parametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 795-806.

Klein, J.P., Moeschberger, M.L. (1997). *Survival Analysis, techniques for censored and truncated data.* Springer Verlag, New York.

van der Laan, M.J. (1997). Nonparametric estimators of the bivariate survival function under random censoring. *Statistica Neerlandica*, Vol. 51, nr. 2, pp 178-200.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* John Wiley & Sons, Inc. New York.

Langberg, S., Shaked, M. (1982). On the identifiability of multivariate life distribution functions. *Annals of Probability*, **10**, 773-779.

Lee, S., Klein, J.P. (1989). Statistical method for combining laboratory and field data based on the random environmental stress model. *In Recent Developments in Statistics and Their Applications, J.P. Klein and J.C. Lee (eds), 87-117. Soeul: Freedom Academy Publishing Company.*

Lee E,L. (1992) *Statistical Methods For Survival Data Analysis.* Wiley, New York.

Lee, E.W., Wei, L.J., Amato, D.A. (1992). Cox type regression analysis for large numbers of small groups of correlated failure time observations. *In J.P. Klein and P.K. Goel (eds) Survival Analysis: State of the Art, NATO ASI, 237-247. Kluwer Academic Publishers Boston.*

Lipsitz, S., Parzen, D. (1996). A jackknife estimator of variance for Cox regression for correlated survival-data. *Biometrics*, **52**, no.1, pp.291-298.

Lin, D.Y., Wei, L.J. (1989). Robust inference for the Cox proportional hazard model. *Journal of the American Statistical Association*, **84**, 1074-1078.

Mantel, N., Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute*, **22**, 719-748.

Mantel, N. (1966). Evaluation of Survival Data and Two New Rank Order Statistics Arising in its Consideration. *Cancer Chemotherapy Reports*, **50**, 163-170.

McCullagh, P., Nelder, J.A. (1989). *Generalised Linear Models*. 2nd Edition, London: Chapman and Hall.

Muenz, L.R. (1983). Comparing survival distributions, a review for non-statisticians. *II, Cancer Investigation*, **1**, 537-545.

Munoz , A. (1980). Nonparametric estimation from censored bivariate observations. Technical Report 60, Stanford University.

Nelson, W (1972) Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics* **14**: 945-965

Nielsen, G., Gill, R.D., Andersen, P.K. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics,* **19**, 25-44.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society, Series B*, **44**, 414-422.

O' Brien, P.C., Fleming, T.R. (1987). A paired Prentice-Wilcoxon test for censored paired data. *Biometrics*, **43**, 169-180.

Peto, R., Peto, J. (1972). Asymptotically efficient rank invariant procedures. *Journal of the Royal Statistical Society, Series A*, **135**, 185-207

Pettitt, A.N., Bin Daud, I. (1990). Investigating time dependence in Cox's proportional hazards model. *Applied Statistics*, **39**, 313-329.

Prentice, R.L., Marek, P. (1979). A Quantative Discrepancy between Censored data Rank Tests. *Biometrics, 35, 861-867.*

Prentice, R.L. Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data, *Biometrika* **79**, 495-512.

Pruitt, R. (1991). Strong consistency of self-consistent estimators: general theory and an application to bivariate survival analyis. Technical Report nr. 543, University of Minnesota.

Pruitt, R.C. (1993). Small Sample Comparison of Six Bivariate Survival Curve Estimators. *Journal of Statistical Computation and Simulation*, **45**, 147-167.

Reid, N., Crepeau, H. (1985). Influence functions for proportional hazards regression. *Biometrika*, **72**, 1-9.

Quantin, C., Moreau, T., Asselain, B (1996). A regression survival model for testing the proportional hazards hypothesis. *Biometrics, 52: 874-885.*

Rodgers, W.H. (1993). Regression standard errors in clustered samples. *Stata technical Bulletin*, STB-13, 19-23.

Segal, M.R. (1988). Regression trees for censored data. *Biometrics* **44,** 35-47.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239-241.

Shorack, G.R., Wellner, J.A. (1986) *Empirical Processes with Applications to Statistics*, Wiley, New York.

Storer, B.E., Crowley, J. (1985). A diagnostic for Cox regression and general conditional likelihood. *Journal of the American Statistical Association*, **80**, 139-147.

Therneau, T.M., Grambsch, P.M., Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika*, **77**, 216-218.

Therneau, T.M., (1997). Penalised Cox models and Frailty. *Technical Report.*

Tillman, D.M., Aitchison, T., Watt, D.C., MacKie, R.M. (1991). Stage II Melanoma in the West of Scotland, 1976-1985: Prognostic Factors for Survival. *European Journal of Cancer.* Vol 27, No.7, 870-876.

Tsai, W-Y., Leurgans, S., Crowley, J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring. *Annals of Statistics* **14**, 1351-1365.

Turnbull, B.W. (1974). Nonparametric Estimation of a Survivorship Function with Doubly Censored Data. *Journal of the American Statistical Association*, **69**, 169-173.

Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B,* **38**, 290-295.

Vaupel, J.W., Manton, K.G., Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439-454.

Wei, L.J. (1986) A Generalised Gehan and Gilbert test for paired observations that are subject to arbitrary right censorship. *Journal of the American Statistical Association,* **75,** *634-637.*

Woolson, R. F., Lachenbruch, P.A.. (1980). Rank Tests for censored Matched Pairs. *Biometrika*, **67**, 597-606.

Woolson, R. F., O'Gorman, W. (1992). A Comparison of Several Tests for Censored Paired Data. *Statistics in Medicine*, Vol. 11, 193-208.

Wellner, J.A. and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of American Statistical Association.* **92**, 945-959.

Zhan, Y. (1998). "Induce: A nonparametric MLE Survival Analysis Module". *Technical Report of MathSoft Seattle WA 98109.*

Zhou (1997). Computing the NPMLE of distribution function from doubly censored data. *Technical Report Department of Statistics, University of Kentucky Lexington, KY 40506-0027.*