Non-Ignorable Missing Covariate Data in Parametric Survival Analysis

Katherine Boyd

A Thesis presented for the degree of
Doctor of Philosophy

THE UNIVERSITY OF
WARWICK

Department of Statistics
University of Warwick
England

May 2007

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank everyone who has helped me throughout my PhD. Most specifically, my supervisor Prof Jane Hutton, whose patience and support have been invaluable. I would also like to thank all other members of the department and, in particular, Dr Karla Hemming for helping with the data. Thank you also to Paula, Julia, and Sue who are always on hand in a crisis! And to all the other PhD students; good luck! To Stephen and Chris, thanks for joining me on coffee breaks and I hope everything goes well for you in the future.

I would also like to thank my parents for putting up with the embarrassment of having to tell people that their daughter is still at university (yes - it has been a long time!) and also for their love, support, and encouragement.

Thank you also to Billy who has put up with me, particularly in the last few months. I really appreciate the help.

Finally, I acknowledge the help of the EPSRC who funded my time as a PhD student.

Declaration

I declare that this thesis is my own work, except where explicitly stated, and has not been submitted elsewhere.

Abstract

Within any epidemiological study missing data is almost inevitable. This missing data is often ignored; however, unless we can assume quite restrictive mechanisms, this will lead to biased estimates. Our motivation are data collected to study the long-term effect of severity of disability upon survival in children with cerebral palsy (henceforth CP). The analysis of such an old data set brings to light statistical difficulties. The main issue in this data is the amount of missing covariate data. We raise concerns about the mechanism causing data to be missing.

We present a flexible class of joint models for the survival times and the missing data mechanism which allows us to vary the mechanism causing the missing data. Simulation studies prove this model to be both precise and reliable in estimating survival with missing data. We show that long term survival in the moderately disabled is high and, therefore, a large proportion will be surviving to times when they require care specifically for elderly CP sufferers. In particular, our models suggest that survival from diagnosis is considerably higher than has been previously estimated from this data.

This thesis contributes to the discussion of possible methods for dealing with NMAR data.

Dedicated to

my family and friends.

Chapter 1

Introduction

This thesis is concerned with two main issues. The first, the motivating problem, is the analysis of data from a cohort study of cerebral palsy sufferers. Cerebral palsy is the name given to the mental and physical impairments caused by complications with brain development or brain injury. These complications normally arise during pregnancy or childbirth but can be the result of postnatal trauma. The severity of the condition varies immensely and, therefore, an individual's requirements and their impact on services changes vastly. For this reason good survival estimates are imperative in order to plan the distribution of various resources including schooling, medical facilities, and equipment for the home. There are several well established cohorts within the UK, and also abroad, looking at both severity and survival to try and establish both the prevalence and level of the condition but none can provide information on long term survival as they have not been running for a long enough period. We consider data collected by a paediatrician in the Bristol area prior to the start of these cohorts. We are, therefore, able to look at survival rates at older ages and

the dependence of survival upon level of disability. This is particularly important as we are now seeing much better survival in older sufferers due to better medical care and so a whole new area of resource is required. We need to know what proportion of sufferers are expected to survive into old age and what level of disability they are likely to suffer from.

The second issue considered in this thesis is the statistical difficulties presented by this analysis. Due to the nature of the data collection we have had to adapt usual survival analysis models. The main problem is the amount of missing covariate information. Our interest is in log-term survival prediction from baseline disability levels but a proportion of this baseline information is missing. This information was collected as early back as the 1930's and we are concerned that data is not randomly missing but was not recorded because it was too complicated or impossible to collect. For example, mental impairment is measured via the use of an IQ test. In order to conduct such a test a certain level of ability is required. We do not know if when a child fell below this level it was recorded so if this is not taken into account and survival is dependent upon the level of disability then standard survival estimates may be biased.

We start this thesis by describing the established theory of survival or time-to-event analysis. In Chapter 2 we discuss the issues that arise, the functions of interest, and the different models that might be used to describe them. Our data is subject to both truncation and censoring both of which are discussed and compared. Non-parametric, semi-parametric, and fully parametric models are presented and their implementation discussed.

Within this chapter we also discuss the main issues with missing data analysis. The hierarchy of modelling assumptions first presented by Ru-

bin (Rubin 1976) and commonly used in the literature to accommodate missing data are presented. We discuss the formulation and refinement of these assumptions and introduce the notation and definitions that we will use for the rest of the thesis.

After introducing the setting for modelling with missing covariate data we can examine the standard methods used. These are presented in Chapter 3. We start with basic case deletion methods, commonly used in practice, and then we look at more complicated likelihood based methods and imputation techniques. The advantages and disadvantages of each of these methods are commented on.

We then focus on the issue of missing data in survival analysis. A full literature review focuses on the issue of missing covariate information, as this is the issue with our motivating cerebral palsy study, but we also consider the possibility of missing outcome in a competing risks setting. We identify the particular issues raised with missing data in survival analysis and look at the different approaches to analysis. Previous research has looked at fully parametric models but focus has been on the popular Cox proportional hazards model. This research is discussed in detail and then summarised, identifying some of the open questions.

Chapter 4 introduces the motivating data. It is important to understand the problems surrounding cerebral palsy and also some of the medical background in order to focus on the correct issues. We look at previous work on both shorter term survival with cerebral palsy and other work on this same data. We also identify the main questions that we are going to investigate in the remainder of the thesis. We then summarise the available

data and look at the missing data pattern.

In Chapter 3 we discuss standard methods for analysing data with missing values and in Chapter 4 we implement these. We discuss the possible missing data mechanisms and the level to which these are accommodated by each of the simpler techniques. These methods include simple case deletion and multiple imputation. Under these methods we look at non-parametric survival to try and increase our understanding of the relationship between severity of disability and survival. We then adapt one of the likelihood based methods discussed in the literature review specifically designed to cope with missing data in survival analysis to allow for fully parametric models. Using this model we look at the effect of severe disability on survival and compare estimates to those from both the complete case and imputed data. This method uses the missing at random assumption. However, we also show why we have doubts about the validity of this assumption in our data.

Having established that the mechanism behind the missing data is likely to be complicated we consider how we can model it. We highlight the similarities between missing data analysis and other statistical issues including measurement error and selection bias. It is in this chapter that we introduce the main model of our analysis; a joint model for the survival time and the missing data mechanism. The missing data mechanism is modelled through the use of a latent variable. Models of this type were first fully investigated in the economics literature. We show its full construction and the derivation of the full likelihood function allowing for both right censoring and left truncation. We present the model for a variety of different distributions commonly used in survival analysis showing the flexibility of

this joint model.

We then discuss and implement this model, using it to investigate the changes in survival model estimates as we relax the missing data assumptions, our focus throughout being the estimation of survival from baseline disability. As with any model it is important to consider the sensitivity of results to the assumptions made and we do this by carefully considering what we know about the data collection methods and missing data mechanisms that may result from these.

In Chapter 6 we present results from a simulation study designed to investigate the reliability of the joint model and to compare it to results from the more standard missing data methods. If the missingness mechanisms acting on the data are such that simpler methods can be used to provide accurate parameter estimates then it is important that our model also performs precisely and efficiently and the simulation study shows that this is the case. However, it does highlight that the efficiency of the model is dependent upon the quantity of missing data.

Finally, in Chapter 7, we extend the univariate model to a multivariate setting. This raises several issues, particularly concerning the covariate model and the structure of the missing data mechanism. We discuss how these issues may be tackled and implement multivariate models to further look at the effect of the level of disability upon survival. We also discuss the inclusion of continuous covariates, informative truncation or censoring, and suggest further work that might follow from this thesis.

Chapter 2

Theoretical Framework

The aim of this chapter is to discuss the main ideas already established in statistical literature that will be used throughout this thesis. We discuss the concepts behind survival analysis, and the implementation of standard analysis techniques, with particular reference to the issue of left-truncation. Also considered are the issues around missing data. We review the assumptions behind any analysis involving missing covariate data and introduce standard techniques for handling this issue.

2.1 Survival Analysis

Survival analysis is an area of statistics widely used primarily in both medicine and biology, but it is also used in economics and engineering. The main aim is to investigate the time until some end-point or event from a particular origin. Within a medical setting, this time origin could be a patient's birth, the enrolment of an individual into a clinical trial, or the time of diagnosis with the illness under investigation, for example. The only restriction

on this starting point is that it must be well defined but it does not need to be the same for each individual in the study. The event of interest could be death in which case the resulting (non-negative) time is quite literally a survival time. However, it may also be the curing of the disease or a reoccurrence of symptoms for example. Another aim is to study the effect of covariate information upon this survival time. Details of survival analysis can be found in several books including those by Cox & Oakes (1984) and Collett (1999). A nice review of survival analysis techniques was also given by Oakes (1982).

2.1.1 Introduction to Survival Analysis

We define the starting time for an individual to be t_0 and the event of interest then occurs at time x measured from this origin.

Censoring

One of the main issues in studying survival is that times are often censored. By this we mean that we do not observe an exact event time, x, but know only that it falls into a set of times, A. For example, in a clinical trial, it is possible that at the end of the study not all the individuals will have experienced the event of interest. An individual who is observed for a set length of time without failure must have a survival time that is longer than this period but this precise time is not recorded. We can record only a censoring time c. This is an example of right-censoring. In right-censoring $A = (c, \infty)$. Other possibilities are left-censoring, when an event is known only to occur before an observed censoring time $A = (0, c)$, and interval censoring, when an event is known to occur between two observed cen-

soring times, c_1 and c_2. It is usually assumed that the censoring times and failure times are independent given the covariates.

In this work we will only be considering right-censoring. We can formally describe this restriction by defining x, the true survival time, and c, the censoring time. Then note, that we only observe a time $t = \min(x, c)$ and a censoring indicator $\delta = I(x < c)$.

Applications of Survival Analysis

Survival analysis techniques are used widely in the areas of medicine and biology. Many examples of their application can be found and several are discussed in the many books published on the subject. These include those by Cox & Oakes (1984) and Collett (1999) mentioned earlier as well as those by Lawless (2003), Klein & Moeschberger (1997), and Fleming & Harrington (1991). Some specific areas of application that have been discussed include the comparison of a particular treatment (the drug 6-mercaptopurine) to placebo for the treatment of leukemia patients (Freireich 1963), the time to first exit-site infection in patients with renal insufficiency (Nahman 1992), and the prognosis of women with breast cancer (Leathem & Brooks 1987).

Applications also occur in the engineering literature such as in the work of Nelson (1970) who test the failure of electrical appliances. This falls into the area of reliability data which is discussed in data by Crowder et al. (1991).

2.1.2 The Survival and Hazard Functions

When studying survival analysis there are two main functions of interest, the survival function and the hazard function. We can regard the actual survival time, x, of an individual to be a realization of a random variable X. If this random variable has an underlying density function, $f_X(x)$, then the cumulative distribution function is given by

$$F_X(x) = P(X < x) = \int_0^x f_X(u)\,\mathrm{du},$$

which represents the probability that an individual has a survival time of at most x.

The survival function is defined as

$$S_X(x) = P(X \geq x) = 1 - F_X(x).$$

This is the probability that an individual has of surviving to at least time x. This function is of great interest as we can use it to look at median and mean survival times.

The hazard function can be defined as

$$h_X(x) = \lim_{\delta x \to 0} \frac{P(x \leq X < x + \delta x \mid X \geq x)}{\delta x}.$$

It is the probability that, given an individual has survived to a time just before x, they fail at time x. This function shows us when an individual is most at risk.

It is possible to show that

$$h_X(x) = \frac{f_X(x)}{S_X(x)}.$$

If we define the cumulative hazard function

$$H_X(x) = \int_0^x h_X(u)\,\mathrm{du}$$

then we can also go on to show that

$$S_X(x) = \exp\{-H_X(x)\}.$$

When analyzing survival data we can estimate both the survival and hazard functions. We can do this by specifying a parametric probability density function $f_X(x)$ or by using non-parametric methods. These methods are described in the following sections.

2.1.3 Non-Parametric Analysis

Perhaps one of the simplest approaches to modelling survival data is via non-parametric methods. These may be simply an initial investigation into the data or may be used as the complete analysis. In particular, we can estimate the survival function for a set of survival times and then compare these functions over different groups of individuals. Here we present methods for analysing right censored data.

Life-tables

We calculate the life-table (or actuarial) estimate of the survival curve is based upon by dividing the full observation period into a series of intervals. We define these m intervals and assume that in the kth of these intervals time extends from time t_k to time t_{k+1}. In this period, count up the number of recorded deaths and censoring times and define these as d_k and c_k respectively. Also, define the number of individuals at risk at the start of the period as n_k. We then assume that the censoring process over each interval is such that the right censored survival times falling into the interval occur uniformly. Now, assuming this uniform censoring the average number of individuals at risk during the period is

$$n'_k = n_k - \frac{c_k}{2}.$$

Calculating the survival probability in each interval as $(n'_k - d_k)/n'_k$ means that we can estimate the survival curve as

$$S(t) = \prod_{k=1}^{j} \left(\frac{n'_k - d_k}{n'_k} \right)$$

for the interval $t_j \leq t < t_{j+1}$, j=1,...,m. The initial probability of survival is, of course, unity. This results in a step-function when plotted against time.

Kaplan-Meier Curves

Kaplan & Meier (1958), presented an extension to the life-table where the intervals are determined by the r ordered recorded distinct death times, $t_{(1)}, ..., t_{(r)}$ which are taken to occur at the start of each period. This is also referred to as the product limit estimate. By making the assumption that

the deaths occur independently we arrive at an estimated survival function similar in form to that of the life-table. The Kaplan-Meier estimate is

$$\hat{S}(t) = \prod_{k=1}^{j} \left(\frac{n_k - d_k}{n_k} \right)$$

for the interval $t_{(j)} \leq t < t_{(j+1)}$, j=1,...,r where n_k is the number of individuals at risk just before the death that occurs at time $t_{(k)}$, d_k is the number who die at time $t_{(k)}$, and $t_{(r+1)}$ is defined to be ∞. Also note that $\hat{S}(t) = 1$ for time $t < t_{(1)}$.

We can calculate the standard error for the Kaplan-Meier estimate via Greenwood's formula. It is given by

$$\text{s.e.}\left\{\hat{S}(t)\right\} \approx \left[\hat{S}(t)\right] \left\{ \sum_{k=1}^{j} \frac{d_k}{n_k(n_k - d_k)} \right\}^{\frac{1}{2}}$$

for $t_{(j)} \leq t < t_{(j+1)}$. This expression can be used to calculate confidence intervals for the estimated survival function. Standard $100(1-\alpha)\%$ confidence intervals are of the form

$$\hat{S}(t) \pm z_{\alpha/2}\text{s.e.}\hat{S}(t)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-point of the standard normal distribution. However, this can lead to confidence intervals that fall outside the region (0,1). We can transform $\hat{S}(t)$ to lie in the range $(-\infty, \infty)$ by using, for example, the log-log transform $\log\{-\log S(t)\}$ and calculate an interval for this transformed value. Using this transform leads to the confidence interval $100(1-\alpha)\%$

$$\hat{S}(t)^{\exp[\pm z_{\alpha/2}\text{s.e.}[\log\{-\log \hat{S}(t)\}]]}.$$

A Taylor series approximation can be used to calculate the Var $[\log\{-\log S(t)\}]$

Examples show that using Greenwood's formula for the standard error sometimes underestimates the confidence region. An alternative was suggested by Peto *et al.* (1977). They propose that the standard error should be calculated as

$$\text{s.e.}\{\hat{S}(t)\} = \frac{\hat{S}(t)\sqrt{\{1-\hat{S}(t)\}}}{\sqrt{(n_j)}}$$

for $t_{(j)} \leq t < t_{(j+1)}$. However, this expression slightly overestimates the standard error so confidence intervals will tend to be slightly larger than the should be.

Hall & Welner (1980) show how to eliminate this incorrect estimation. They propose the confidence bands

$$\hat{S}(t) \pm D_n \hat{S}(t)[1 + C_n(t)] \quad t \leq t_{(n)},$$

where D_n is the value of the Kolmogorov-Smirnov statistic at significance level α and

$$C_n(t) = n\frac{\text{var}\left[\hat{S}(t)\right]}{\hat{S}(t)^2}.$$

Comparing Survival Curves

We may wish to compare estimated survival curves for two or more groups of individuals in order to establish the effect of discrete covariates upon survival. Here we consider two non-parametric tests that can be used to test the null hypothesis that there is no difference between two survival curves.

The first of these is the log-rank test. This is based upon the hypergeometric distribution. Consider two groups, group 1 and group 2, for

whom we wish to compare survival. Again, label the distinct death times across both groups, $t_{(1)} < ... < t_{(r)}$, as we did when considering the Kaplan-Meier estimate for the survival curve. Similarly, define d_k and n_k, $k = 1, ..., r$, as before. Further, suppose there are n_{ik} individuals at risk and d_{ik} deaths at time k in group i, $i = 1, 2$. If we assume that the marginals (i.e. d_k and n_k) are fixed and have the null hypothesis that the survival is independent of the group then d_{1k} has a hypergeometric distribution. We then consider the statistic

$$U_L = \sum_{k=1}^{r}(d_{1k} - e_{1k}),$$

the difference between the observed and expected numbers of deaths. e_{1k} is the mean of the hypergeometric random variable d_{1k} and is given by $e_{1k} = n_{1k}d_j/n_j$. We can now calculate the variance of U_L,

$$\text{var}(U_L) = \sum_{k=1}^{r} \frac{n_{1k}n_{2k}d_k(n_k - d_k)}{n_k^2(n_k - 1)} = \sum_{k=1}^{r} v_{1k} = V_L.$$

It can be shown that U_L has an approximately normal distribution so therefore, $U_L/\sqrt{V_L} \sim N(0, 1)$.

This implies that

$$\frac{U_L^2}{V_L} \sim \chi_1^2.$$

We can then compare this value to critical values for the chi-squared distribution. The larger the value the greater the evidence against the null hypothesis.

The second test is the Wilcoxon test. This is conducted in a similar man-

ner to that of the log-rank test. However, here we calculate the statistic

$$U_W = \sum_{k=1}^{r} n_k(d_{1k} - e_{1k}),$$

where the notation is the same as defined in the previous paragraph. As you can see we are weighting deviations from the expectation by the size of the risk set. This means that the Wilcoxon test is less sensitive to small sample sizes, i.e. towards the end of the study. The variance of U_W is given by $\text{var}(U_W) = \sum_{k=1}^{r} n_k^2 v_{1k} = V_W$ and the Wilcoxon test statistic by

$$W_W = \frac{U_W^2}{V_W} \sim \chi_1^2.$$

These tests can be easily generalised to the situation when we have more than two groups. It is important to note that the log-rank test is more suitable when the alternative hypothesis is that of proportional hazards as it uses this assumption although smaller deviations from proportionality have a minor impact. The Wilcoxon test is more recommended when the alternative hypothesis is not that of proportional hazards.

2.1.4 Semi-Parametric and Parametric Models

In many studies there are covariates including individual characteristics and treatments whose effect on survival is of primary interest. This leads us to consider parametric regression models.

The Cox Proportional Hazards Model

One of the most commonly used models for survival data is the Cox proportional hazards model (Cox 1972). This is a semi-parametric model which

uses the assumption that covariates have a multiplicative effect upon the hazard function. In section 2.1.2 we introduced the hazard function, $h(x)$, which can be thought of as the instantaneous risk of failure conditional upon survival until that point. The proportional hazards assumption is that

$$h_i(x) = \exp\{\beta^T z_i\} h_0(x) \qquad \text{i=1,...,n,}$$

where z_i is the covariate vector (possibly containing dummy variables in the case of factors) for individual i, $h_i(x)$ is the hazard function for individual i, and $h_0(x)$ is a baseline hazard. Assume, β is a set of model parameters. Note, that we have made no assumptions concerning the form of the baseline hazard. We can allow for different types of covariates, including those that are time-dependent, and can calculate maximum likelihood estimates for model parameters through the use of a partial likelihood function. Details of this partial likelihood can be found in Cox (1972). We can maximise the partial likelihood using Newton-Raphson procedures and can approximate the covariance matrix of the parameter estimates using the inverse of the information matrix.

The Weibull Model

We can still make this proportional hazards assumption but now we can allow the baseline hazard to have a Weibull distribution. The density function for the Weibull distribution is

$$f(x) = \lambda\gamma x^{\gamma-1} \exp(-\lambda x^{\gamma}), \qquad \lambda, \gamma > 0 \text{ and } 0 \leq x < \infty.$$

A special case of the Weibull distribution is the exponential distribution which arises when $\gamma = 1$. If this assumption is valid then models based upon this distribution will arrive at more precise parameter estimates. Again we can fit this model using Newton-Raphson or other numerical procedures, although now we can obviously calculate a standard likelihood function, details of which are discussed below.

The Accelerated Failure Time Models

An alternative assumption to that of proportional hazards is that of accelerated failure. Here, the covariates act multiplicatively on the time scale. The assumption that we make in this general form of model is that

$$h_i(x) = \exp\{\beta^T z_i\} h_0\left(\exp\{\beta^T z_i\} x\right), \qquad \text{i=1,...,n.}$$

The Weibull distribution can also be used to model the survival times under this assumption. However, we can also use other distributions including:

- the log-logistic $f(x) = \frac{\theta\xi x^{\xi-1}}{\left(1+\theta x^{\xi}\right)^2}$,
- the gamma distribution $f(x) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}x^{\alpha-1}\mathrm{e}^{-\beta x}$ and,
- the log-normal $f(x) = \frac{1}{\sigma\sqrt{2\pi}}x^{-1}\exp\left\{-(\log x - \mu)^2/2\sigma^2\right\}$.

We should note that we can write both Weibull proportional hazards models and accelerated failure models in log-linear form with suitable choices of residual distribution.

Calculating the Likelihood

As discussed, when considering the Cox proportional hazard model we must use a partial likelihood as we are making no assumption about the

distribution of survival times. However, when using a fully parametric model we can use a normal likelihood function. The main issue to note though is that, due to censoring, we do not always observe the true survival time. This means that the contributions from censored and uncensored individuals to the likelihood are going to be different. Here, we consider only right-censoring. If we consider the full data set (t_i, δ_i, z_i), for $i = 1, ..., n$, consisting of observed times, censoring information, and covariate values then the likelihood can be calculated as follows...

$$L(\Theta|t, \delta, z) = \prod_U f(t_i|z_i, \Theta) \prod_C S(t_i|z_i, \Theta) = \prod_{i=1}^{n} h(t_i|z_i, \Theta)^{\delta_i} S(t_i|z_i, \Theta)$$

where Θ is the full set of model parameters. We can see that the contribution to the likelihood from the set of uncensored individuals (U) is as expected but from right-censored individuals (C) it equates to the survival function as we know only that failure occurs after the observed censoring time.

Maximum likelihood estimates for the p unknown parameters, $\hat{\Theta} = (\hat{\Theta}_1, ..., \hat{\Theta}_p)$, are the values of $(\Theta_1, ..., \Theta_p)$ that maximise the likelihood function. They are found by solving the score equations...

$$\frac{\partial}{\partial \Theta_j} \log L(\Theta|t, \delta, z)\Big|_{\hat{\Theta}} = 0, \qquad \text{j=1,...,p.}$$

The covariance matrix of the maximum likelihood estimates can be approximated via the inverse of the observed Fisher information matrix,

$$\text{var}(\hat{\Theta}) \approx I^{-1}(\hat{\Theta}).$$

The observed information matrix is $I(\Theta) = -H(\Theta)$ where $H(\Theta)$ is the $(p \times$

p) Hessian matrix with (i,j)th entry

$$H(\Theta)_{i,j} = \frac{\partial^2 \log L(\Theta)}{\partial\Theta_i \partial\Theta_j}$$

for $i,j = 1,...,p$. Variances for $\hat{\Theta}$ can then be found along the diagonal of the information matrix, i.e. the standard error of $\hat{\Theta}_j$ is the square root of the (j,j)th entry of $I(\hat{\Theta})$. We can use these standard errors to calculate confidence intervals for the parameter estimates and hence use these to decide upon the inclusion of covariates into the survival model and check for a significant impact on survival.

2.1.5 Model Selection

We can use the log likelihood ratio to choose between nested models and to informally compare non-nested models the AIC as discussed below. As previously mentioned, we can decide upon covariates to include firstly through the estimated parameter standard errors. However, we may wish to compare alternative models. As in any model fitting situation we may not wish to fit the most accurate model but the most parsimonious model (i.e. the model that best fits the data with the least number of parameters). In order to do this we consider the maximum log-likelihood $\log \hat{L}$ (the log-likelihood function evaluated at its maximum likelihood estimates) or, more conveniently, $-2\log \hat{L}$. $\hat{L}$ is a product of conditional probabilities and hence is less than unity, so the smaller the value of $-2\log \hat{L}$ the better the model fits the data.

If we are comparing two nested models, say Model 1 nested in Model 2, with maximized log-likelihoods $\log \hat{L}_{(1)}$ and $\log \hat{L}_{(2)}$ respectively, then a large difference between $-2\log \hat{L}_{(1)}$ and $-2\log \hat{L}_{(2)}$ would lead to the con-

clusion that the additional covariates in Model 2 do improve the adequacy of the model. The difference between the two maximized log-likelihoods can be written as

$$-2\log\frac{\hat{L}_{(1)}}{\hat{L}_{(2)}}$$

and is called the likelihood ratio statistic. It can be shown that this has an asymptotic chi-squared distribution under the null hypothesis that the additional covariate coefficients are zero with degrees of freedom equal to the number of additional parameters in Model 2. Therefore, we can use it to assess the need for the additional terms.

Sometimes, we do not have nested models and therefore, need an alternative method for model selection. One method for choosing a model in this case is Akaike's Information Criterion (AIC). The AIC statistic for a single model is defined as

$$AIC = -2\log\hat{L} + \alpha q$$

where q is the number of unknown covariate coefficients and α is a predetermined constant usually approximately 3 as this is approximately equivalent to using a 5% significance level in judging the difference between two nested models differing by up to three parameters. When there are no subject specific reasons for a particular model choice a suitable model can be identified as that with the lowest AIC.

2.2 Left-Truncation

Survival data can also be subject to truncation. This is a slightly different idea to that of censoring. Truncation occurs when sample values larger

(right-truncation) or smaller (left-truncation) than a fixed value, y, are not recorded. This need not be the same point for each subject. For example, if individuals enter a study when they first see a specialist then they have to have survived until their first appointment in order to be included in the data set. If they die before this point they do not enter the study and no data are recorded. This is an example of left-truncation. So now our observed data for individual i is of the form $(t_i, y_i, \delta_i, z_i)$. Note that truncation can lead to the complete absence of an individual from the data set while censoring leads to the inclusion of partial information. Ignoring truncation can lead to biased survival estimates as highest risk individuals are more likely to fail before first being observed. We consider independent left-truncation when truncation times are independent from survival times.

Figure 2.1 shows the different mechanisms that can act on the survival times if we allow right censoring and left truncation. This figure covers the types of data we will encounter within this thesis.

The first lifetime shown (subject A) is an typical survival time which is observed for its whole survival period and for whom an exact failure time is observed. Subject B is also observed from the moment it enters the population of interest but is censored at the end of the study as it has not experienced the failure event. Individuals C and D are truncated data. C is included in the study as failure has not occurred before they could be included in the study but is censored before the end of the study period. However, subject D has failed before the study period starts so is an example of a truncated time. It cannot be included in the study. The last example, individual E, is missed by the study organizers despite being in the population of interest during the study period.

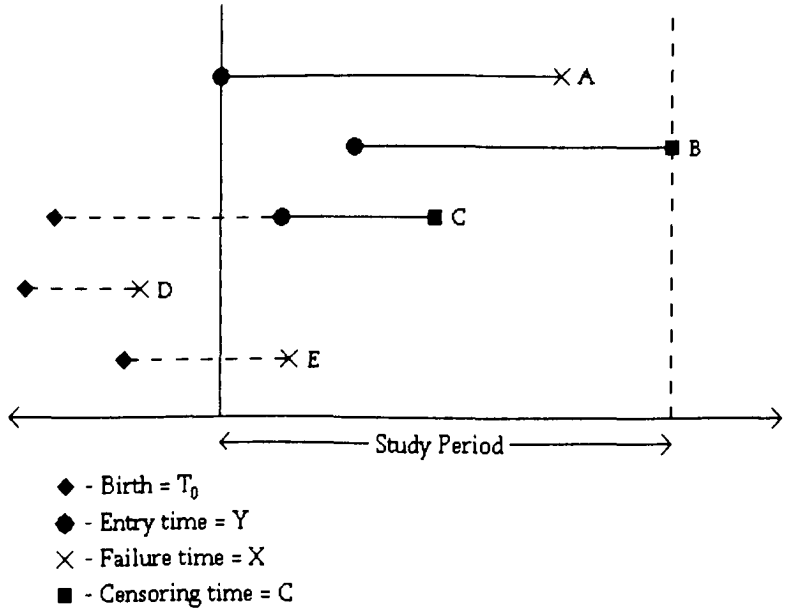


Figure 2.1: Truncated and censored survival data

2.2.1 Left-Truncated Kaplan-Meier

We can extend the techniques of Sections 2.1.3 and 2.1.4 to allow for left-truncated data. A modified estimate similar to that of Kaplan and Meier was first introduced in statistical literature by Woodroofe (1985) although previously, Lynden-Bell (1971) had derived a non-parametric maximum likelihood estimate within his work in astronomy. However, Woodroofe's estimate does not allow for censoring. In 1987, Tsai et al. presented asymptotic results for an analogue to the Kaplan-Meier curve where risk sets are adjusted at each failure time to account for the delayed entry. Using the same notation as that used in the earlier discussion of the Kaplan-Meier curve, we arrive at the same estimate Equation 2.1.3 except the risk set n_k,

i.e. those individuals at risk just before time t_k, is defined as

$$n_k = \sum_{i=1}^{n} I(y_i \leq t_k \leq t_i),$$

where I is the usual indicator function. They also present an extended Greenwood's formula for calculating the standard error of the estimated survival function. Note that issues raised by Pan & Chappell (1998) do not present an issue here as truncation mainly occurs at a young age and the majority of censoring at considerably older ages. They highlight the fact that the standard technique described can underestimate survival at very early times for left truncated and right censored data.

2.2.2 Likelihood under Left-Truncation

We again have to condition on entry time when considering the construction of the likelihood function. For uncensored individuals the contribution is equal to $P(T = t_i \geq y_i, z_i)$ and for censored individuals it is equal to $P(T > t_i \geq y_i, z_i)$. Therefore, using Bayes theorem we can deduce that the likelihood function is of the form

$$L_T(\Theta|t, y, \delta, z) = \prod_U \frac{f(t_i|z_i, \Theta)}{S(y_i|z_i, \Theta)} \prod_C \frac{S(t_i|z_i, \Theta)}{S(y_i|z_i, \Theta)}. \tag{2.1}$$

2.2.3 Applications with Left-Truncation

Censoring is a very common issue in survival time studies. Truncation, while it arises less often, also features in many studies. Within a medical setting there are several examples. Struthers & Farewell (1989) present a model to investigate the development of AIDS (Acquired Immune Defi-

ciency Syndrome) from the point of infection with HIV (Human Immunodeficiency Virus). The left-truncation occurs due to the lag between infection and diagnosis. There is also an issue with right-censoring as the point of progression is often uncertain. Analysis of this same issue is also considered by Lui et al. (1986) and Medley et al. (1987). Another example comes from the work by Hyde (1980) which discusses data collected from the Channing House retirement centre in California. The data here is left-truncated because individuals must survive to an old enough age to enter the centre before they can be included in the data set. This excludes those who die at a young age leading to a *length biased sample*.

2.3 Missing Data

Missing data is an issue that often occurs in studies. It is particularly common within medical and survey settings where data collection may be difficult. Censoring and truncation can fall under the heading of incomplete or coarsened data but we can also have unobserved covariate data. There has been much research into the problem of missing data and the standard approaches are discussed in Chapter 3.

2.3.1 The Missing Data Mechanism

In order to analyse data with missing observations we must first consider the missing data mechanism acting upon the data set. The role of this mechanism was widely overlooked until the idea was formalized by Rubin. This is fully discussed in Little & Rubin (2002). He introduced notation based upon the concept of treating missing data indicators as random variables.

Assume, for simplicity, that the same mechanism applies to the whole

data set. We define the complete true data as $D = (d_{ij}) \in \mathbb{D}$ where $\mathbb{D}$ is the family of all possible observable data sets given the sampling method. This is, in reality, not entirely observed. With regard to survival analysis we can consider $D = (T, Y, \delta, Z)$ where T, Y, δ, and Z are the observed survival times, the entry times, the censoring indicator, and the recorded covariates respectively. Rubin introduced a missing data indicator matrix $M \in \mathbb{M}$. We construct $M = (m_{ij})$, of the same dimension as D, where $m_{ij} = 1$ if d_{ij} is missing and $m_{ij} = 0$ if d_{ij} is observed. The missing data mechanism can then be characterized by the conditional distribution of M given D,

$$P(M = m|D = d, \Phi) = f(m|d) \text{ for all m} \in \mathbb{M} \text{ and d} \in \mathbb{D}$$

where Φ are unknown parameters.

The most restrictive missing data mechanism is defined to be when the probability of missingness does not depend on any of the values in D and is called the *missing completely at random* (MCAR) assumption. This occurs if

$$f(m|d, \Phi) = f(m|\Phi) \text{ for all } d \in \mathbb{D} \text{ and } \Phi.$$

A slightly less restrictive mechanism is in operation if the data are *missing at random* (MAR). Here, missingness is allowed to depend upon the observed values of D but not on the unobserved values. Let D_{obs} denote the observed entries of D and D_{mis} indicate the unobserved. We can then define the MAR assumption as

$$f(m|d, \Phi) = f(m|d_{obs}, \Phi) \text{ for all } d_{mis} \text{ and } \Phi. \tag{2.2}$$

Note that under the MAR assumption we can have different mechanisms

for different subgroups or even different individuals.

Finally, if missingness is allowed to depend on both the observed and unobserved data (i.e. the full data, d) then the data is said to be *not missing at random* (NMAR).

This hierarchy of missing data mechanisms and the corresponding notation is now widely used in missing data literature and whilst the concept of the three mechanisms is clear there is a subtlety in the notation that can cause confusion. This confusion arises with the definition of D_{obs} and D_{mis}.

The notation could imply that in the separation of D into its two components both values and their positions are maintained i.e. D_{obs} and D_{mis} have the same dimension as D with the relevant entries missing. If this is indeed the case then, within the MAR definition, conditioning on D_{obs} means that we can fully determine the value of M as we merely look at the missingness structure of D_{obs}. This is something we cannot do when conditioning on D and so the equality above, Equation 2.2, does not hold. By definition, we can also fully determine M from D_{mis}.

Alternatively, if we take the notation to imply that we retain only the relevant matrix entries, and not their location within D, in the construction of D_{obs} and D_{mis} a problem arises in how we condition the distribution of the object M with known dimension on an object with unknown structure.

In order to avoid this confusion we introduce a slight variation to Rubin's original notation. Note first that the definitions and notation for the MCAR and NMAR mechanisms is clear and the issue only arises when concerned with data that is MAR. Again, let

$$P(M = m|D = d) = f(m, d), \text{ m} \in \mathbb{M} \text{ and d} \in \mathbb{D}$$

Instead of defining the two components D_{obs} and D_{mis} consider the class of matrices $\mathbb{D}^*$ in which each matrix shares observed entries with D but has alternative values for those that we do not observe i.e.

$$\mathbb{D}^*(m,d) = \{d^* \in \mathbb{D} : d^*_{ij} = d_{ij} \text{ for all i,j with } m_{ij} = 0\}.$$

We can then rewrite the definition for MAR data as

$$f(m|d,\Phi) = f(m|d^*,\Phi) \text{ for all } d \in \mathbb{D},\ d^* \in \mathbb{D}^*(m,d) \text{ and } \Phi.$$

This implies that missingness depends only on the observed values of D whilst avoiding the notational confusion discussed above.

Methods for analysing data with missing observations are discussed in Chapter 3. There we focus upon, in particular, missing covariate data in survival analysis. In Chapter 5 we look at recent research in NMAR missing data. Recent summaries of missing data research include Little (1992), Rubin (1996), Schafer (1999), and Schafer & Graham (2002).

Chapter 3

Literature Review

3.1 Early Historical Development

The first methods used for dealing with missing data were editing methods. These are described in Section 3.1.1. They were the primary idea used up until the 1970's and are still commonly used today. The formulation of the EM (Expectation-Maximization) algorithm (Dempster et al. 1977) first made it possible to compute maximum likelihood estimates in the presence of missing data (see Section 3.1.3). This meant that instead of deleting or filling in incomplete cases we can treat the missing data as random variables which can be integrated out of the likelihood function as if they were never sampled. More recently, Markov chain monte carlo (MCMC) methods have also been used in likelihood based approaches. In 1987, Rubin (1987) introduced the idea of multiple imputation. This is discussed in Section 3.1.2. In this method, each missing value is replaced with $D \geq 5$ simulated values prior to analysis and computed parameters averaged over the D complete data sets.

3.1.1 Editing Methods

The first, and most commonly used method, is complete case analysis. This involves removing observations with one or more of the variables missing.

Complete Case Analysis

There are both advantages and disadvantages to using complete case analysis. The most obvious disadvantage is the, often large, reduction in the quantity of data so it is not a recommended method when there is a high proportion of missing data. This loss of data results in a potential loss of information in two respects. Firstly, a loss of precision, and secondly bias caused by the data not being MCAR (missing completely at random) or the complete cases not being a representative sample of all the cases. One strategy for partially adjusting the bias in complete-case analysis is to assign each individual a weighting. There are also advantages to complete-case analysis, however, which cause it to be the most commonly used missing data technique and the default in many computer software packages. The most obvious is its simplicity and lack of need for extensive data manipulation. Another advantage is that univariate statistics are all calculated on the same sample meaning that direct comparison is justified. In this survival analysis it also allows us to look at multivariate models involving all the variables that are available to us.

Available Case Analysis

Another method for dealing with missing data is available case analysis. This time we can look at subsets of the data and remove only the missing values in those variables.

The main disadvantage of available-case analysis is that the sample base changes from variable to variable according to the pattern of missing data. Problems are also caused in multivariate analysis as more data is lost and variable selection becomes more difficult, as was the issue with complete case analysis.

Available case analysis is more useful than complete case analysis when there is a larger amount of missing data, particularly when conducting a simple analysis. However, in multivariate analysis the available cases must be recalculated for each new model hence slightly increasing computation time and also meaning we can not use the log likelihood ratio or AIC to compare models.

3.1.2 Imputation Methods

It is important to note that both complete-case and available-case analysis make no use of cases with one variable missing when estimating either the marginal distribution of that variable or measures of covariation between that variable and others. This leads to a loss of important information. One method to regain some of this lost information is to impute, or fill-in, the missing data. Imputation is a flexible method for handling missing data but it does have problems. There are a variety of ways in which the missing data can be imputed. However, we must be careful with imputation methods as they can be dangerous. For example, once we have imputed the data we can start to believe that we are dealing with a complete data set which is obviously not the case.

We can use either explicit or implicit modelling to impute the data. In explicit models, the predictive distribution from which we draw imputa-

tions is based on a formal model and hence the assumptions are explicit. In implicit models the imputation is based on an algorithm which implies an underlying model.

Single Imputation

In single imputation we impute only one value to substitute for each missing value. This imputation can be done in several ways, some of which are discussed below. Mean and regression imputation are explicit methods while the hot and cold deck and substitution methods are implicit.

- Mean - One of the basic methods for continuous data is single mean imputation. Missing data in the continuous variables are replaced with either a unconditional or conditional mean value.
- Regression - This method replaces missing values with predicted values from a regression of the covariate containing the missing data on the variables observed for the individual. This can also be done stochastically if we include a residual drawn to reflect the uncertainty in the predicted value.
- Hot deck - Here, draws are based on an implicit model and replace missing values by values from similar responding units in the sample. This can involve very elaborate schemes for unit selection.
- Cold deck - Missing observations are replaced by constant values from some external source. For example, an earlier study.
- Substitution - Replacement of unit if missing values occur. This is many used in survey data where such cases can be substituted at the fieldwork stage.

The main problem with single imputation methods is that it is difficult to assess the uncertainty in the final results as we impute only one fixed value. Bootstrap and Jackknife methods can be used to calculate standard errors for the imputed data parameters. These both involve resampling methods. Alternatively, we could use multiple imputation.

Multiple Imputation

Multiple imputation could also be used to impute values for the missing data. This method is Bayesian and involves replacing each missing value by a vector of $J \geq 2$ imputed values. The J values are ordered in the sense that J completed data sets can be created from the vectors of imputations. Standard complete data methods can then be used to analyze the data sets.

Multiple imputation shares the advantages of single imputation but also rectifies some of the disadvantages. The resulting complete data analyses can be easily combined to create an inference that validly reflects sampling variability because of the missing values. The resulting estimates are often more efficient than in MCAR analysis. The only disadvantage of multiple imputation over single imputation is that it takes more work to create the imputation and analyze the results.

Model choice and variable selection in the analysis of multiple imputed data sets is an issue as alternative models and variables may appear to be best fitting in the different data sets. The same analysis must be carried out on each data set so that the parameters can be combined.

After the multiple data sets are imputed and standard analysis completed on each one parameter estimates and variances need to be combined. This is done as proposed by Rubin (1987). The combined estimate for each

parameter θ calculated from J repeated imputations is

$$\overline{\theta} = \frac{1}{J}\sum_{j=1}^{J}\hat{\theta}_j$$

where $\hat{\theta}_j$ is the estimate from imputation j and the variability associated with this estimate is

$$V = \frac{1}{J}\sum_{j=1}^{J}V_j + \frac{J+1}{J}\left\{\frac{1}{J-1}\sum_{j=1}^{J}\left(\hat{\theta}_j - \overline{\theta}\right)^2\right\}$$

where V_j is the variance of estimate θ_j. Note that this combined variance is the sum of the within and between estimate variances. Rubin's rules for combining estimates require underlying normality of the estimator.

3.1.3 Maximum Likelihood Methods

More complicated approaches to missing data analysis involve likelihood procedures based upon explicit modelling assumptions. Imputation techniques also involve modelling assumptions in a more implicit way. For general patterns of missing data maximum likelihood estimates cannot be calculated explicitly utilizing factorizations of the likelihood. If this is the case, and if the closed-form solutions for the score functions cannot be found, then iterative procedures can be used to maximise the likelihood.

The first of these is the Newton-Raphson algorithm. Assume that we have a function $f(x)$ which has a root x_r such that $f(x_r) = 0$. We can use the second degree approximation from Taylor's series to approximate this root. If we set an initial estimate for the root as x_0 then we can use the

iteration

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

to find the estimated root for the $n + 1$st iteration. Here, x_n is the estimate for the nth iteration and $f'(x_n)$ is the differential of $f(x)$ with respect to x evaluated at x_n. The estimates converge to the root. In the case of multiple roots the root located depends upon the arbitrarily chosen initial value. We are considering the maximisation of the likelihood function so if this is unimodal and concave then the sequence of estimates converges to the maximum likelihood estimate if the function f is taken to be the score function (i.e. the differential of the likelihood, or log-likelihood, function). The main issue with this method is that the matrix of second derivatives of the log-likelihood needs to be calculated which can be of high dimension and complicated functions of the parameters.

An alternative method, that does not require the calculation of second derivatives is the Expectation-Maximisation (EM) algorithm. The earliest reference to the algorithm seems to be that of McKendrick (1928). Several other authors then used the algorithm in differing circumstances and Orchard & Woodbury (1972) first noted the general applicability of the method calling it the missing information principle. The EM algorithm was formalized by Dempster et al. (1977). Since then further work has been done regarding its convergence (Wu 1983). It is a very general iterative algorithm for maximum likelihood estimation in incomplete-data problems. It consists of two steps.

- The E-step - Finds the expected complete-data log-likelihood given

the current estimate of the parameters, $\theta = \theta^{(r)}$:

$$Q(\theta|\theta^{(r)}) = \int l(\theta|Y) f(Y_{mis}|Y_{obs}, \theta = \theta^{(r)}) dY_{mis}.$$

- The M-step - The M-step determines $\theta^{(r+1)}$ by maximising this expected log-likelihood:

$$Q(\theta^{(r+1)}|\theta^{(r)}) \geq Q(\theta|\theta^{(r)}).$$

Standards errors are generally harder to calculate through the EM algorithm than when using multiple imputation.

3.1.4 Markov Chain Monte Carlo (MCMC) Methods

Markov chain Monte Carlo (MCMC) is essentially Monte Carlo integration using Markov chains and provides great assistance in statistical modelling (Gamerman 2002). Monte Carlo integration evaluates the expectation of a function, $f(x)$, by drawing samples $\{x_k, k = 1, ..n\}$ from the posterior distribution or likelihood, $\pi(.)$, and then approximating

$$E[f(x)] \approx \frac{1}{n} \sum_{k=1}^{n} f(x_k).$$

So the population mean is estimated by a sample mean. The difficulty in this arises due to the general infeasibility of drawing independent samples from the posterior distribution. However, the $\{x_k\}$ need not be independent so can be drawn from a Markov chain with stationary distribution $\pi(.)$.

The Gibbs sampler is a special case of the Metropolis-Hastings algo-

rithm. The Metropolis-Hastings algorithm is used to construct the Markov Chain with stationary distribution, $\pi(.)$. At each time t, the next state, X_{t+1}, is chosen by firstly sampling a candidate point Y from a proposal distribution $q(.|X)$ which may depend on the current X_t. The candidate point is then accepted with probability

$$\alpha(X,Y) = \min\left(1, \frac{\pi(Y)q(X|Y)}{\pi(X)q(Y|X)}\right).$$

If the candidate point is accepted the next state is $X_{t+1} = Y$. Otherwise, $X_{t+1} = X_t$. The proposal distribution, $q(.|.)$ can have any form and the stationary distribution of the chain will be $\pi(.)$. Instead of updating the whole of X at once it is often more convenient to divide X into components $X = \{X_{.1}, .., X_{.h}\}$ and update these components one by one. Define $X_{.-i} = \{X_{.1}, .., X_{.i-1}, X_{.i+1}, .., X_{.h}\}$. For the Gibbs sampler, the proposal distribution for updating the ith component of X is

$$q_i(Y_{.i}|X_{.i}, X_{.-i}) = \pi(Y_{.i}|X_{.-i})$$

where $\pi(Y_{.i}|X_{.-i})$ is the full conditional distribution.

3.2 Missing Data in Survival Analysis

The previous methods discussed can be applied to many types of missing data. However, they are crude and rely heavily on some strong and untestable assumptions. They can also lead to very biased results if these assumptions are not true. As discussed, more recently methods have been developed which approach the problem of missing data through a likelihood approach. This has led, in particular, to the introduction of methods

based specifically on survival models.

3.2.1 Missing Data Problems

Firstly, we can think about the types of missing data problems that may arise in survival analysis studies. Research focuses on two main missing data patterns. The first is when there is missing data on the covariates that are collected in the studies and used to model survival. In this case we must usually have complete information on the survival times and the censoring indicator. Several approaches to this problem are described in Section 3.2.2. The second pattern, discussed in Section 3.2.3, is when we have complete data on any covariates but there is missing data on the survival times or censoring indicator. This is often looked at in a competing risks setting. We can also consider event times to be missing for censored individuals and an imputation method using this idea is discussed in Section 3.2.3.

The published research looking at these two problems focus on maximum likelihood approaches, particularly using the Cox proportional hazards profile likelihood. However, some multiple imputation ideas are looked at briefly.

3.2.2 Approaches to Missing Covariate Data in Survival Analysis Problems

We start by considering the issue of missing covariate data. We assume full observation of T i.e. the censoring or failure time, and the right-censoring indicator δ.

Parametric Models

One of the earliest references to missing data in survival analysis is that of Schluchter & Jackson (1989). The method has three parts. They start by constructing a multinomial model based on the discrete covariates. The multinomial model arises by using the categorical covariates to form a contingency table so that each fully observed individual falls in to only one cell. The distribution of the hazard function conditional on the values of the covariates is then described using a log-linear model. Schluchter and Jackson use a stepwise constant function for the hazard (i.e. piecewise exponential survival). The resulting likelihood is then maximized over the missing data using the EM or Newton-Raphson algorithm as described in Section 3.1.3. When fitting an unsaturated log-linear model for the hazard the M-step of the algorithm also requires the application of one step of the IPF (Iterative Proportionality Fitting) algorithm (Bishop et al. 1975).

Several assumptions are made in the formulation of this model. Firstly, only categorical covariates can be incorporated. Assumptions also need to be made concerning the censoring and missing value mechanisms. The censoring must be independent of the true survival time given the covariates. This is the usual assumption of independent censoring and most methods that will be discussed require this. Also, the mechanism causing covariates to be missing must be ignorable (Rubin 1976) i.e. the data is both MAR and distinct.

This model leaves many avenues for extension. The stepwise log-linear model for the hazard is quite restrictive but it can be extended to simple parametric models. Extensions will be discussed in Section 4.6. It also requires the assumption that the data is MAR which is usually hard to verify.

A later model by Baker (1994) tries to relax the MAR assumption made by Schluchter and Jackson. It again only allows for categorical covariates, in this case only a single covariate is included. It also allows for a non-ignorable censoring mechanism.

Its main approach is to group survival into discrete times and then model discrete time hazards for both failure and censoring. This makes the analysis tractable but obviously causes a loss of information and hence precision. The formulation of the model considers four random variables: an indicator of a missing covariate, censoring time, failure time, and the true covariate stratum. Using Bayes Theorem, the joint probabilities needed for the construction of the likelihood are decomposed.

The model is then able to consider a variety of possible parameterizations for the hazard function and it is this function that is of primary interest. It again uses a log-linear model, as in Schluchter and Jackson, which can describe both proportional and non-proportional hazards. A model for the hazard of censoring is also required and it is at this point that the method can allow for non-ignorable censoring as we can include a term dependent on the covariate stratum. Again a log-linear model is used. Similarly, log models are used for the missing data mechanism and these can be made to depend on the covariate stratum to allow for non-ignorable missing data.

The likelihood formed using these models is then maximized using a composite linear model as discussed in Baker (1994).

This approach is more complex and time consuming than that of Schluchter and Jackson but it is very useful in investigating the appropriateness of the assumptions regarding the censoring and missing data mechanism. However, it still does not allow for continuous covariates and the use of discrete

times again causes a loss of information.

As discussed in Section 3.1.3, the EM algorithm is now used often in missing data problems to maximise incomplete data likelihoods. The first method described in this section, that of Schluchter and Jackson, uses this algorithm. However, it is quite simple to implement in this case, as is the Newton-Raphson procedure which is also described. An extension to this algorithm is the EM algorithm by the method of weights, as described by Ibrahim (1990). It is shown that, under some very general conditions, the E-step of the EM algorithm can be written as a weighted complete data log-likelihood for any generalized linear model (GLM), nonlinear regression model, or time series. In a survival setting, denote the covariates as z, the survival times as t, and the missing observations for individual i as $z_{mis,i}$. Thus the E-step for a general regression problem at the $(r+1)$st iteration can be written as

$$Q(\theta|\theta^{(r)}) = \sum_{i=1}^{n} \sum_{z_{mis,i}} w_{i(r)} l(\theta; z_i, t_i, \delta_i) \tag{3.1}$$

where, given $\theta = (\alpha, \beta)$,

$$\begin{aligned} l(\theta; z_i, t_i, \delta_i) &= \log\{P(t_i, \delta_i|z_i, \beta)\} + \log\{P(z_i|\alpha)\} \quad \text{and} \\ w_{i(r)} &= P(z_{mis,i}|z_{obs,i}, t_i, \delta_i, \theta^{(r)}) \end{aligned}$$

Here α and β are parameter vectors. We must again assume that the covariates are categorical and that the responses (in our case, the survival times and censoring indicator) are complete.

Using this EM algorithm by the method of weights means that we do not have to calculate the incomplete data likelihood so it can be used when it is not feasible to use Newton-Raphson directly.

This algorithm was first applied to missing covariates in survival data by Lipsitz & Ibrahim (1996). They show, however, that the method of Ibrahim (1990) calculates many nuisance parameters and that by proposing a conditional model for the covariate distribution they can reduce the number of such parameters as if a covariate is fully observed its distribution need not be estimated. This conditional model for the p covariates denoted $z = (z_1, ..z_p)$ is

$$P(z_1, .., z_p|\alpha) = P(z_p|z_1, .., z_{p-1}, \alpha_p)...P(z_2|z_1, \alpha_2)P(z_1|\alpha_1)$$

which could be fitted using a series of logistic regression models in the case of dichotomous covariates. This idea is useful for any parametric model of a response given discrete covariates.

We can also use these method to incorporate continuous covariates that have no missing values. As discussed, if a covariate has no missing values we do not need to estimate its distribution given the other covariates in the above model thus allowing it to be continuous. Lipsitz and Ibrahim also demonstrate numerically that the parameter estimates for the final model and the resulting test statistics are not sensitive to the order of conditioning in the conditional model for the covariates.

Lipsitz and Ibrahim introduce a liver cancer survival analysis data set as an example. In order to model the survival function they decide on a piecewise exponential model as in Schluchter & Jackson (1989).

Obvious extensions to the EM algorithm method by weights described by Ibrahim can be seen. Continuous covariates with missing data still need to be allowed for and we need to look at the details of using more complex parametric survival models when there are missing covariate data. Recent research has focused on these problems.

In a subsequent paper by Ibrahim et al. (1999) the technique of Monte Carlo EM (as discussed by Wei & Tanner (1990)) was applied to tackle the issue of incorporating continuous covariates. The motivating example is a right-censored survival analysis data set although this technique can be used with a variety of models.

The Monte Carlo E-step for the EM algorithm is derived. For missing continuous covariates, the usual E-step for the ith observation can be written as

$$Q(\theta|\theta^{(r)}) = \int l(\theta; z_i, t_i, \delta_i) P(z_{mis,i}|z_{obs,i}, t_i, \delta_i, \theta^{(r)}) dz_{mis,i}. \qquad (3.2)$$

We can compare Equation 3.2 to Equation 3.1 and note the obvious integration caused by the inclusion of continuous covariates. It can be shown that

$$P(z_{mis,i}|z_{obs,i}, t_i, \delta_i, \theta^{(r)}) \propto P(t_i, \delta_i|z_i, \theta^{(r)}) P(z_i|\alpha^{(r)}),$$

and therefore the product on the right can be used for sampling from the distribution $[z_{mis,i}|z_{obs,i}, t_i, \delta_i, \theta^{(r)}]$. For the ith observation a sample of size m_i, $z_{i,1}, ..., z_{i,m_i}$, is taken from the distribution of $[z_{mis,i}|z_{obs,i}, t_i, \delta_i\theta^{(r)}]$ via the Gibbs sampler in conjunction with the adaptive rejection algorithm.

The E-step for all observations, at the $(t+1)$st iteration, is given by

$$Q(\theta|\theta^{(r)}) = \sum_{i=1}^{n} \left\{ \sum_{h=1}^{m_i} \frac{1}{m_i} l(\theta; z_{i,h}, z_{obs,i}, t_i, \delta_i) \right\}.$$

This method for estimating the model parameters when there are missing categorical, continuous, and mixed covariates can be used for arbitrary parametric regression models. However, the maximum likelihood method proposed requires the specification of a parametric distribution for the covariates and thus introduces the possibility of misspecification.

Cho & Schenker (1999) look at the log-F accelerated failure model. The log-F AFT model includes models with extreme value, logistic, normal, and log-gamma errors. This means that many parametric models, including the Weibull, gamma, log-logistic, and log-normal, can be fitted by adjusting the number of degrees of freedom in the log-F distribution. They assume that the missing covariate mechanism is ignorable (i.e. MAR and distinct) and that censoring is random and non-informative. They do, however, develop an extension that allows the censoring mechanism to depend on missing covariate values. Covariates can be continuous. They take a Bayesian approach and utilize MCMC (Markov Chain Monte Carlo) techniques.

Denote the vector of categorical covariates as $\mathbf{U}$ and the vector of continuous covariates as $\mathbf{V}$. The AFT survival model for survival times, T, is given by

$$\log(t) = \beta_0 + \beta_1'\mathbf{U} + \beta_2' g(\mathbf{V}) + \beta_3' h(\mathbf{U}, \mathbf{V}) + \sigma\epsilon$$

where $g(\mathbf{V})$ is a vector representing the main effects and interactions of $\mathbf{V}$, $h(\mathbf{U}, \mathbf{V})$ is a vector of the selected interactions between $\mathbf{U}$ and $\mathbf{V}$, and ϵ is a random error variable. Assume that $\epsilon \sim \log F(2a, 2b)$. Cho and Schenker

then suggest models for the covariates and the censoring. The model for the covariates is the general location model. This is the model used by Schluchter & Jackson (1989) and Lipsitz & Ibrahim (1998). It consists of a multinomial model for the contingency tables formed by the categorical covariates and then has a multivariate normal distribution for the continuous covariates within each cell. There can be restrictions on this model as it can have many parameters. The model suggested for the censoring mechanism is an exponential regression model. They use the Gibbs sampler to estimate posterior distributions for the model parameters.

Meng & Schenker (1999) also investigated the problem of missing data on continuous covariates. They again make the assumptions that data is MAR and that censoring is non-informative and random. They also assume that the censoring distribution does not depend on any predictors that are missing. They look at log-linear regression models of survival time on the covariates. This includes AFT models. They restrict the error variable to have a standard normal distribution.

Instead of using MCMC techniques, as in Cho & Schenker (1999), Meng and Schenker simply construct the likelihood and maximise it over the missing data using the EM algorithm. Alternatively, the Newton-Raphson algorithm could be used although it tends not to converge without a good initial estimate for the parameters.

Cox Proportional Hazards Models

Section 3.1 looked at some of the earliest methods for missing data in survival analysis and also some later more general approaches. As discussed

in Section 2.1.4 the most commonly used model in survival analysis is the Cox proportional hazards model (Cox 1972). Hence, there has been significant research into using this model in the presence of missing data. The main problem with the Cox model is that standard likelihood methods can not be used as no parametric distribution is used for the baseline hazard so a partial likelihood is used instead.

Let (x_i, c_i, z_i), for $i = 1, .., n$, be n independent replicates of (X, C, Z) where X is the failure time, C is the censoring time, and Z the vector of covariates. If we take $t_i = \min(x_i, c_i)$, $\delta_i = I(x_i \leq c_i)$ and $w_i(t) = I(t_i \geq t)$ then the complete data partial score function for β_0 is

$$U(\beta) = \sum_{i=1}^{n} \delta_i \{z_i(t_i) - \bar{z}(\beta, t_i)\}, \tag{3.3}$$

where

$$\bar{z}(\beta, t) = \frac{\sum_{l=1}^{n} w_l(t) \exp\{\beta^T z_l(t)\} z_l(t)}{\sum_{l=1}^{n} w_l(t) \exp\{\beta^T z_l(t)\}}.$$

Note that $\bar{z}(\beta, t)$is the conditional expectation of $Z_l(t)$ on $\{l : \ t_l \geq t\}$. The MPLE $\hat{\beta}$ is defined as the solution to the score equation $\{U(\beta) = 0\}$. It can be solved using Newton-Raphson.

The earliest methods for incorporating the Cox proportional hazards model into analysis with missing data often required quite strict assumptions. In particular, some of the first research of Lin & Ying (1993) requires the data to be missing completely at random (MCAR). However, they do allow the use of time-dependent continuous covariates.

They derive an estimating function for the vector of regression param-

eters which is an approximation to the standard partial likelihood score function. Firstly, they estimate the conditional expectation, $\bar{z}(\beta, t)$, from the subjects who have complete data at time t. The sum over the uncensored failure times of the observed value of $z_i(t_i)$ minus its estimated $\bar{z}(\beta, t_i)$ can then be used as an estimating equation for β_0. For those uncensored individuals with missing observations, the equivalent components of z_i are merely excluded from the summation.

To construct the estimating equation they consider two random variables, $\{H_{0i}(.), \mathbf{H_i}(.)\}$, where $H_i(.)$ is a p x p matrix (p = number of covariates) with diagonal elements $\{H_{1i}(.), ..., H_{pi}(.)\}$. Also, denote $H_{ji}(t) = I(z_{ji}(t)$ observed) and $H_{0i}(t) = I(H_{ji}(t) = 1)$ for all $j = 1, ...p$. It is now that the MCAR assumption is required as this corresponds to the assumption that the missing indicators $\{H_{ji};\ j = 1, .., p\}$ are independent of all other random variables. Now they introduce the following notation...

$$S^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^{n} H_{01}(t) y_i(t) \exp\{\beta^T z_i(t)\} z_i(t)^{\otimes r},$$

$$E(\beta, t) = S^{(1)}(\beta, t)/S^{(0)}(\beta, t),$$

where for a vector a, $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa^T$.

The approximate partial likelihood estimator (APLE) can then be written as

$$\tilde{U}(\beta) = \sum_{i=1}^{n} \delta_i H_i(t_i)\{z_i(t_i) - E(\beta, t_i)\}.$$

This APLE is shown to have only slightly reduced efficiency in comparison to the MPLE with full covariate measurements unless there is a large quantity of missing data.

A later paper by Zhou & Pepe (1995) also looks at an estimated partial likelihood estimator for the Cox regression model but makes use of auxiliary covariate data which are considered to be informative about the data but which are not part of the regression model.

They differentiate between individuals with complete data, which they call the validation sample V, and those with missing data. Those with missing data must have observed auxiliary data. For those in the validation sample they look at the relative risk and for those not in the sample they look at the expectation of the relative risk given the observed auxiliary data. The association between the covariates in the regression model and the auxiliary covariates is left unspecified and is estimated nonparametrically.

Although the induced relative risks are unknown they can be estimated using the data in the validation sample. Then we can look at the intuitive expectations of these relative risks given the auxiliary data, which are just weighted sums, and use these values in the sum over individuals that forms the approximate partial likelihood. The estimated partial likelihood score function is then solved using Newton-Raphson.

This method does have some limitations. Firstly, whilst the potentially unobserved covariates in the proportional hazards model are allowed to be continuous the auxiliary covariates are not. Secondly, if the dimension of the auxiliary covariates is large, validation subsets within each distinct category may be small causing unstable induced relative risks. It is also difficult to see if the validation sample is a simple random sample of the whole data set. This assumption is required for the analysis.

An alternative to the methods of Lin & Ying (1993) and Zhou & Pepe

(1995) is described by Paik & Tsai (1997). Their method still requires the data to be MAR. Denote those covariates that are completely observed for individual i as $z_{1,i}(t)$ and those covariates that may be missing as $z_{2,i}(t)$. The partial likelihood score function for the Cox proportional regression model can be rewritten as

$$\begin{aligned} U_f(\beta) &= \sum_{i=1}^{n} \delta_i \left\{ \begin{pmatrix} z_{1,i}(t_i) \\ z_{2,i}(t_i) \end{pmatrix} - \frac{\sum_{j=1}^{n} w_j(t_j) \begin{pmatrix} z_{1,j}(t_j) \\ z_{1,j}(t_j) \end{pmatrix} e^{\beta_1^T z_{1,j}(t_i) + \beta_2^T z_{2,j}(t_j)}}{\sum_{j=1}^{n} w_j(t_j) e^{\beta_1^T z_{1,j}(t_i) + \beta_2^T z_{2,j}(t_j)}} \right\} \\ &= \begin{pmatrix} O_1 \\ O_2 \end{pmatrix} - \begin{pmatrix} E_1 \\ E_2 \end{pmatrix} \end{aligned}$$

using the same notation as in Equation 3.3.

The first term in the above equation is a sum of "observed" covariates from failed study subjects, and the second term is a sum of "expected" covariate values given the prior information. Paik and Tsai propose two estimating functions. In the first they impute the "expected" term only. As in the previous method of Lin and Ying (1993) the contribution to the score function is discarded if the failed study subject has missing covariates. This imputation yields consistent estimators of the "expected" term under a restricted MAR assumption. This assumption is that missingness can depend on observed covariates but not on missing covariates or the corresponding failure or censoring times. The second method imputes both the "expected" and "observed" terms. This method yields consistent estimators under the normal MAR assumption. If some of the fully observed covariates are continuous a smoothing technique is required. The imputed values can then be used in an imputed partial likelihood score function which can be solved via Newton-Raphson to obtain the proposed estimators for the regression parameters.

A slightly different approach to the problem of missing covariate data in the Cox proportional hazards model is taken by Pugh et al. (1993). Instead of using an approximate partial likelihood by estimating the relative risk for those individuals with missing values they weight the score equation from the complete case analysis to remove the bias caused by data not being MCAR. The subject-specific weights are proportional to the reciprocal of the probability of having complete data. Although these weights are generally not known they can be estimated from the data using a binary regression model such as the logistic or probit model. Like the previous two approaches they allow continuous covariates.

If there is independent censoring then the resulting parameters for the complete case analysis are consistent to the true parameters. Pugh et al. introduce a weighted score equation which yields unbiased estimates under less restrictive assumptions.

There is a problem in this approach: we have to estimate the probabilities that individuals are fully observed given their covariate values. There is the danger of misspecification in this model which may lead to bias.

A common unattractive feature of the methods discussed so far is that the variance formulas are very complicated. Multiple imputation methods provide estimates whose variances can be easily computed by adding between-imputation and within-imputation variances. Paik (1997) looks at three multiple imputation methods for the Cox proportional model with missing covariate data. Two of the imputation methods provide estimates that are asymptotically equivalent to the earlier results of Zhou and Pepe (1995) and Paik and Tsai (1997). The third is a modified version of the sec-

ond of these for which estimates and standard errors can be calculated using standard software, and time varying covariates can be incorporated.

In drawing random samples from the observed data, we need to incorporate variability. One method for doing this is the Approximate Bayesian Bootstrap (ABB) (Rubin, 1987, p124). This method has two steps. Firstly, sample with replacement from the observed observations, then draw imputes from this bootstrap sample.

The first two imputation methods described are completed by drawing imputes of the missing data using ABB from the observed covariates in each risk set given the fully observed covariates and then using these imputations in the estimating equations of Zhou and Pepe (1995) and Paik and Tsai (1997). Repeating the imputation D times leads to a series of parameter estimates, $\hat{\beta}^{(1)}, ..., \hat{\beta}^{(D)}$, which we can then average over. Although the variances of the regression parameters for these two imputation methods are easier to compute than those from the methods on which they are based, they still can not be calculated using standard software so a modification to one of the methods is discussed. The difficulty in implementing the first method occurs as the imputed values for the missing covariates should be updated for every iteration of the Newton-Raphson algorithm as they depend on the regression parameters that we are trying to find. The modification changes the form of these imputed values so that they do not depend on the parameters. Again, a smoothing technique would be needed in any of these methods in the presence of continuous covariates.

Using simulation methods, the authors compare their imputation methods to the methods on which they are based and also complete case analysis results and results from the methods of Lin and Ying (1993) and Pugh et al. (1993). They conclude that the imputation methods are more efficient than

the estimates based only on complete data and inference can be drawn using standard software. They also have the advantage of simple variance calculation but there is a loss in efficiency over the other discussed methods.

Another approach is taken by Chen & Little (1999). Unlike their predecessors, who look at approximations to the partial likelihood, these authors look at a nonparametric maximum likelihood approach. They do, however, use an EM type algorithm to solve the maximisation problem. Again the method requires the MAR assumption. Simulation results suggest that this nonparametric method is more efficient than approximate partial likelihood methods and complete case analysis.

The main feature of the model is a discretization of the likelihood which can then be maximized simply. When there are no missing data the Cox partial likelihood is obtained as the profile likelihood under this discretization. This method requires specification of the distribution of the missing covariates. The likelihood of the Cox proportional hazards model is (up to a constant factor):

$$L(\theta|z_{obs}, t, \delta) = \prod_{i=1}^{n} \{h_0(t_i)\}^{\delta_i} \int \{\exp(\beta^T z)\}^{\delta_i} \exp(-H_0(t_i)\exp(\beta^T z)) f(z, \theta)\, dz$$

where the integration is over the possible covariate values for individual i. The maximum over the parameter space for θ does not exist so the likelihood is modified by discretization of the cumulative baseline hazard, $H_0(t_i)$, as a step function with jumps at the observed failure times. This can then be maximized using an EM type algorithm.

However, this model requires a specification of the distributional form of the covariates. This feature is not present in the imputation method of Paik and Tsai (1997) or the approximate profile likelihood method of Zhou and Pepe (1995).

A paper by Martinussen (1999) follows this method. Again the covariates are assumed to be missing at random. Non-informative censoring, as usual, is also assumed. The method relies on the nonparametric maximum likelihood interpretation of the Nelson-Aalen estimator in the Cox regression setting (see Anderson et al. (1989)) whereas the method of Chen and Little (1999) uses the Breslow estimator.

Again Martinussen specifies the complete data likelihood and maximizes it using an EM type algorithm. This is the same EM algorithm by weights described by Ibrahim (1990) and used in Lipsitz & Ibrahim (1996). While the method is described for categorical covariates this is only a technical assumption not a conceptual one. This method can also be extended to non-ignorable non-response if we can model the missing data mechanism. Simulations again suggest that this method is more efficient than the imputation method of Paik and Tsai (1997).

MCMC methods of analysis can be used in missing data problems. They are used in the paper of Lipsitz & Ibrahim (1998). They propose a set of estimating equations for the parameters in a Cox model which they suggest are solved using MCMC, as an approximation to the EM algorithm, due to their computational intensity. They require the MAR assumption and restrict their analysis at first to categorical covariates. Their work

can be considered as an extension to the likelihood method of Lipsitz and Ibrahim (1996) to the Cox model. They take a semiparametric approach to specifying the joint distribution of the survival times and the covariates. A fully parametric distribution is given for the covariates and the conditional distribution of the survival times given the covariates is given by assuming proportional hazards.

To specify the distribution of the covariates they use a multinomial distribution similar to that of Schluchter and Jackson (1989) as the covariates are categorical. An obvious choice of model is then a log-linear model with regression parameters α. As the data is assumed to be MAR a consistent estimator of the parameters, θ, can be obtained by using the conditional expectation of the complete data score vector. Therefore they use the standard forms for the score equations for the parameters in the proportional hazards model. Thus,

$$\begin{aligned} U^*(\theta) &= E[U(\theta)|z_{obs}] \\ &= E\left\{ \left[\begin{array}{l} \sum_{i=1}^n \int_0^\infty \{z_i - \bar{z}(s,\beta)\} dN_i(s) \\ \sum_{i=1}^n \left\{ dN_i(t) - \lambda_0(t) y_i(t) e^{\beta^T z_i} \right\} \\ \sum_{i=1}^n \partial \log p(Z_i|\alpha)/\partial\alpha \end{array} \right] \middle| \text{ observed data} \right\} \end{aligned}$$

where $\theta = (\beta, \lambda_0, \alpha)$ and the score equations for β (the survival model covariates) and λ_0 (the baseline hazard) are written in counting process notation (Fleming & Harrington 1991). Since the missing covariates are categorical we can remove the conditionality from the score function and multiply by the conditional probabilities for specific values summing over all possible values.

Their MCMC approximation to the EM algorithm is then summarized

as follows:

- Obtain an initial estimate of $\theta = \theta^{(1)}$, say, by complete cases. At the kth step we have $\theta^{(k)}$.
- Using this calculate the posterior probabilities for the multinomial model.
- Fixing these, sample missing data from its conditional multinomial distribution. Repeat this L times to get L complete data sets.
- Find the mean of the score functions for the L data sets and set to 0. Solve for $\theta = \theta^{(m+1)}$.
- Iterate until convergence.

This is similar to the method of Wei and Tanner (1990). This method is computationally feasible and can be easily implemented in standard software packages.

MCMC is also used in follow-up papers by Herring & Ibrahim (2001), Leong et al. (2001), and Herring et al. (2004). The first of these, that of Herring and Ibrahim (2001) extends the method described above to continuous covariates. The MAR and non-informative censoring assumptions are again required. They implement a Monte Carlo version of the weighted EM algorithm along with the Gibbs sampler which is different to the method of Lipsitz and Ibrahim (1998). It is more computationally feasible to this earlier approach. When covariates are categorical it leads to similar results to the method of Chen and Little (1999).

The results of Lipsitz and Ibrahim (1998) are extended to allow for non-ignorably missing covariate data in the paper by Leong et al. (2001). Covariates are still required to be categorical.

Recall the estimating equations described earlier. We now include a model for the missingness mechanism and, hence, a corresponding score function

$$U_\phi(\phi) = \sum_{i=1}^{n} \frac{\partial[\log\{p(m_i|t_i, \delta_i, z_i, \phi)\}]}{\partial\phi}$$

where m_i is a vector of indicator values for the missingness of covariates for individual i, t_i is the survival time, δ_i is the censoring indicator, z_i are the covariate values, and ϕ are the parameters in the model for the missingness mechanism. The method uses the same Monte Carlo EM algorithm as that of Lipsitz and Ibrahim (1998).

The covariates can be modelled using a saturated linear model and the number of nuisance parameters can be reduced using the idea of Lipsitz and Ibrahim (1996) where we write the distribution of the covariate vector as a product of one-dimensional conditional distributions. It is suggested that the missing data mechanism is also modelled by a sequence of one-dimensional conditional distributions. Since all components of m_i are binary a sequence of logistic regressions is an obvious choice for the model form. The exact form of these logistic regression models can be determined by using Akaike's information criterion or the likelihood ratio. We need to be careful not to include too many factors in the model as it would soon become unidentifiable.

This method is again extended by Herring et al. (2004) to include con-

tinuous covariates. Missingness is still allowed to be non-ignorable. It is a similar approach to that of Leong, Lipsitz, and Ibrahim (2001).

When the covariates are continuous the E-step in the EM algorithm includes an integral instead of a sum. This integral generally has no closed form. However, it has the form of an expectation with respect to the missing data given the observed data and current parameter estimates. This means that it can be evaluated via the Monte Carlo EM algorithm of Ibrahim et al. (1999). Samples are taken using the Gibbs sampler. In the model for the covariates logistic or normal linear regression models can be used.

3.2.3 Approaches to Missing Survival Time Data in Survival Analysis Problems

This issue is not going to be a focus in this thesis but it is interesting to look at previous methods to handle the problem and we include the discussion here for completeness. This problem often occurs, particularly in a competing risks framework. It is interesting to consider the comparison between censoring and complete missingness, both of which are examples of coarsened data. This is just a small sample of the relevant literature.

Multiple Imputation

As described in Section 3.1.2 we can use multiple imputation in dealing with missing data problems. If it is the covariates that contain missing data and the survival times and censoring indicator are complete standard imputation techniques can easily be used. Problems arise when the missing data occurs within the survival time variable. Taylor et al. (2002) discuss three methods for the imputation of missing survival times. The missing

data that they discuss are the true event times that go unrecorded for censored individuals.

The first of these methods is risk set imputation (RSI). In this approach we impute a survival time and censoring indicator from those individuals in the same risk set (i.e. those individuals still alive at the time of censoring). If the last observed event is censored then it retains its value since the risk set contains no possible donors.

The second method is Kaplan-Meier imputation (KMI). This method draws an event time from a Kaplan-Meier estimator of the event times among those at risk. A KM survival curve is estimated from those individuals in the corresponding risk set to each individual with missing data.

The final imputation method is Bootstrap imputation (BI). The RSI and KMI methods alone do not incorporate the uncertainty in the imputes. Consider a bootstrap sample selected with replacement from the original data set. The imputing risk set then consists of those individuals in the bootstrap sample who are still at risk at the censoring time for the relevant individual. RSI or KMI can then be used.

Imputations can be calculated J times and then standard multiple imputation techniques can be used.

Competing Risks Models

One form of multistate survival model is the competing risks model. This model is relevant when there are several types of failure so instead of a binary censoring indicator we have a failure type indicator. So far we have focused on the problem of missing covariate data but it may occur, particularly, in competing risks models, that the failure type indicator is missing.

Censored survival times often preclude observation of the censoring indicator. There has been less investigation into this area.

Dinse (1982) considered this issue. Data consist only of a survival time and failure indicator for each subject, there are no covariates. The constructed model looks at the following four possible individual missing data patterns.

1. Known survival time and failure type.
2. Survival time right-censored and failure type unknown.
3. Survival time right-censored and failure type observed.
4. Known survival time but missing failure type.

The model does not look at the situation when the survival time is completely unknown. Censoring is assumed to be non-informative and data is MAR. A non-parametric likelihood can then be constructed. The joint distribution of the survival time and failure type indicator is estimated via the EM algorithm. This method can also be used in a traditional survival analysis setting when there is one covariate with missing values.

The non-parametric MLE of the survival function is discrete (Kaplan & Meier 1958). The E-step of the EM algorithm only involves estimating the (unobserved) number of failures at time t_k who have an observed failure type, j, and a right-censored survival time. Therefore, it is quite simple to compute.

Less restrictive forms of incomplete observations would permit further extension of this technique. For example, we may observe a union of intervals on the positive real line for the survival time. Left and interval-censoring are special cases of this. Alternatively, we may know that the

failure type was one of a subset of the complete set. This would alter the contributions of the four possible missing data patterns but the EM algorithm would still provide maximum likelihood estimates. However, computation would be more complicated.

It has been shown that nonparametric maximum likelihood estimators are inconsistent when failure indicators are missing. Van Der Laan & McKeague (1998) introduce a sieved non-parametric maximum likelihood estimator and show that it is efficient. The assumption of MAR is again made. They do not strictly work with competing risks as there is no failure type indicator only the standard censoring indicator which can be missing.

Their approach is to find the nonparametric maximum likelihood estimator (NPMLE) of the survival function based on reduced data produced by a discretization of the observed (possibly censored) survival times. This discretization is done by interval censoring the survival times of those individuals for whom the censoring indicator is unobserved. This method provides consistent results.

The methods of Dinse and Van Der Laan and McKeague looked at nonparametric models for the survival function. A natural progression is to look at the Cox proportional hazards model. Goetghebeur & Ryan (1995) base their method on the solution to estimating equations. MAR and noninformative censoring assumptions are again made. The method also allows for the inclusion of time-dependent covariates.

Their approach is developed in two steps. They work with two possible failure types. Firstly they assume the baseline cause-specific hazards for

the two types of failure are proportional and that the proportionality constant is known. This could be dependent on time. They construct a partial likelihood which leads to an score equation estimator that reduces to the Cox proportional hazards model when there is no missing data. Then, secondly, they allow for estimation of the proportionality constant. However, in this second approach standard results are not achieved if the method is used on complete data. Score tests and cumulative hazard estimators are also derived.

3.2.4 Summary

As we have discussed missing data is a very complicated issue within statistical theory. Research has focussed on the issue of missing at random (MAR) data, in particular within the Cox proportional hazards model as the semi-parametric nature of this model presents specific issues. We have considered a range of these approaches. There is some consideration of the NMAR assumption again mainly in the Cox model.

We are left with some obvious areas for further study. In particular, the use of fully parametric models under the NMAR assumption. In the following chapter we will present our motivating data and then go on to consider how we might fit such models under this assumption.

One important thing to note is that we have barely touched on the vast literature on the use of multiple imputation. The flexible nature of the approach means that it is not discussed exclusively with the survival literature but is able to stand alone. There is considerable interest in the use of multiple imputation as its flexible nature means that it is very useful in standard analyses. Increasing software is being developed to conduct mul-

tiple imputation and this growing availability means that it is becoming slowly more popular in applied research. There are of course many difficult questions that need to be resolved including how to best simulate the complete data and how to conduct model selection.

Having investigated the literature we can now go on to look at our data. We will implement some of the simpler techniques and then develop flexible methods for modelling under less restrictive missing data mechanism assumptions.

Chapter 4

Motivating Data

4.1 Cerebral Palsy

Cerebral palsy (henceforth CP) is a condition which affects many physical and mental characteristics. It is due, either, to a failure of part of the brain to develop properly or an injury that damages sections of it. It is usually acquired at a very young age, during pregnancy or labour, and is commonly diagnosed in the first years of childhood. Current beliefs suggest that the condition affects one in every four hundred children (*Scope: About Cerebral Palsy* 2006). There are several possible causes of CP and it is often difficult to identify the relevant incident in any one child. However, some of the known causes are infection in pregnancy, abnormal development, a difficult or premature birth, genetic factors, or infection or injury in childhood.

It is not a new disorder but the medical profession did not begin to study cerebral palsy as a distinct medical condition until the late 19th century.

CP is nether progressive or communicable. It is also not curable al-

though therapies and technology can be highly beneficial to individuals affected by the condition. There are three main types of CP: spastic, athetoid, and ataxic CP. The most common form of CP, affecting approximately 80% of sufferers (*Cerebral Palsy - Ask the Doctor* 2006), is spastic CP which causes the muscles to stiffen and decreases the possible range of movement. If only one side of the body is affected the condition is referred to as 'hemiplegia'. If legs are affected but arms are unaffected or only slightly affected this is known as 'diplegia' and if both are equally affected, then the term used is 'quadriplegia'. With athetoid CP muscles switch rapidly between tense and loose hence causing involuntary movements. These movements often interfere with skills other than obvious motor functioning including swallowing and speech. Ataxic CP is a rare form of the condition and causes difficulties in balance and coordination. It is possible to have mixed forms of CP. Other symptoms associated with CP include epilepsy, poor sight and hearing, spatial awareness problems, and learning difficulties. Risk factors include mother's and father's age and position in family.

4.1.1 The Effect of Severity on Survival

The effect of the severity of physical, cognitive, and sensory disability on the survival of people with cerebral palsy (CP) has been described previously (Evans et al. 1990, Hutton et al. 1994, 2000, 2002, Strauss et al. 1998a, 1998b, and Blair et al. 2001). Research strongly suggests that the severity of disability has a highly significant effect on the expected survival. Those individuals with less severe disabilities can live well into adulthood. Indeed, estimates suggest that over 80% of children diagnosed with early impairment CP survive beyond their 30th birthday (Hutton & Pharoah 2002).

However, this data differs markedly from other available information and the true proportion may be lower. Physical disabilities, and in particular the lack of primary functional skills, are considered the most indicative of poor future survival. Strauss et al. (1998) investigate this in detail, looking at the ability of the child to roll and sit independently. Severe learning disabilities are also significant when considering survival but there is some belief that this is due to a high correlation with physical disability (Blair et al. 2001). More recently, Hutton & Pharoah (2002) have shown that severe sensory disability is also predictive of poor survival. A summary of recent research can be found in Katz (2003).

Interestingly, and perhaps counterintuitively, birth weight and gestational age are less predictive of survival and a low birth weight actually increases survival expectations. This is attributed to the likelihood that the most at risk babies of low birth weight die before a diagnosis of CP can be made. As neonatal care improves we might expect to see this change as more severely disabled children of low birth weight survive until CP is recognized. Currently, levels of severe disability are lower in low birth weight groups which suggests that cohorts are losing a large number of undiagnosed individuals (Hutton et al. 2000).

Nearly all studies look at short term survival in child cohorts and use information obtained at or close to diagnosis to model future survival. Strauss & Shavelle (1998) consider long-term adult prognosis and conclude that this may not be reliably deduced from a follow-up of children in the original same condition as a change in disability level can occur and this affects survival. Further work found that excess mortality risk in comparison to the general population decreased with age. Strauss et al. (2004) show that levels of severe mobility disability increase in adults over the age of 60

years. They also observed poorer survival in this age bracket. Their work is based upon a large cohort from California, USA. Work in the UK (Hutton et al. 2000) concludes that the hazard functions here vary from those in their cohort but that this difference is not due to differing rates of severe disabilities.

As with all survival research the Cox proportional hazards model (Cox 1972) is commonly used to model survival and estimate hazard ratios. This appears to be done without investigation of the proportional hazards assumption. Hutton & Pharoah 1994, 2002, instead of using this approach, consider parametric models. In particular, they use the log-logistic accelerated failure model (Collett 1999) which they conclude is most appropriate for their data.

There are two main issues that occur in nearly all the previous research into cerebral palsy survival. These involve ascertainment bias and missing covariate information. Problems with ascertainment are widespread and most research attempts to tackle the issues that are believed to affect the data. In particular, the Californian cohort discussed above is collected from information regarding those who receive care in the state. This means that they may miss some individuals with low disability who require little or no care. However, they believe that any bias that this causes is small as their survival estimates for low disability individuals are very close to that of the general population. Also, Hutton & Pharoah (2002) investigate possible ascertainment bias caused by the part retrospective nature of their data from Merseyside, UK by comparing survival from entry to the cohort and survival conditional on survival until two years. They conclude that there is little difference between the two.

All the studies suffer from missing covariate data but there is generally

little discussion concerning the mechanism behind the missingness. Complete case analysis is sometimes used but Hutton et al. (1994) point out that missing disability covariate information is unlikely to be independent of the severity level. In this case, using complete cases, a method only suitable for missing completely at random data, will lead to bias in survival model estimates. This is the problem at the foundation of this thesis.

4.2 The Bristol Data

The motivating data come from a part retrospectively and part prospectively ascertained 1930s to 1960s birth cohort based on consultant paediatrician Dr Grace Wood's case referral in the Bristol region of the UK. Each individual was diagnosed with cerebral palsy (CP). From 1951 to 1964, all cases under the care of the paediatrician were recorded on professionally designed punch cards. This later became the subject of her MD thesis (Woods 1957). The cohort claims to see all cerebral palsied children from Bristol and the surrounding area.

The information held on the punch cards was subsequently compiled into a database. Details of this method can be found in Hemming et al. (2006). Individuals were included if they met certain criteria and could be clearly diagnosed with CP. Only those with early impairment CP were included i.e. if there was mention of a postnatal event after 28 days the child was excluded. Inevitably some cases were excluded as there was not enough information to allow for diagnosis.

The data consists of information on birth weight, gestational age, the mother's age at birth, and several disability covariates. These include levels of ambulation (leg movement), manual dexterity (hands and arms), vision,

and IQ (intelligence quotient). All can be grouped into severe and non-severe groups. See Section 4.2.1 for the precise definitions of levels. Previous research (Hutton et al. 2000) suggests that this distinction provides the greatest significant difference in survival. Information is also available on date of birth, date of death (where appropriate) and, the age at first assessment. For those individuals in the study who are still alive, lifetimes are defined as timed from birth until the censoring date, December 2005. Deaths are flagged via the National Health Service Central Register (NHSCR) of the Office for National Statistics.

4.2.1 The Variables

The data consists of information on gestational age, mother's age at birth, and several disability covariates. Information is also available on date of birth, date of death (where appropriate) and, the age at first assessment.

- Birth characteristics

  Gestational age Length of the gestational period of the child (in weeks).

  Birth weight Weight at birth (in grams).

  Mother's age Age of the child's mother at birth (in years).

- Disability variables

  Ambulation Level of ambulatory disability (1 - none/mild (lowest), 4 - wheelchair dependent (highest)).

  Manual dexterity Level of manual disability (1 - none/mild (lowest), 4 - unable to feed or dress themselves (highest)).

  IQ Intelligence Quotient measurement.

Vision Level of visual disability (1 - non-severe (lowest), 2 - registered blind or attend school for partially sighted (highest)).

- Lifetime outcome data

Age Age (in days) at death or censoring.

Age of assessment Age at first assessment (in days).

Dead Censoring indicator (0-censored, 1-dead).

Little is known about how this data were collected, particularly the disability variables. Specifically, it is not known which method(s) were used to calculate IQ. However, the simple categorical structure of each of the physical impairments means that measurement error is unlikely. For example, opinion on whether a child is dependent upon a wheelchair is unlikely to differ. As mentioned we will split the disability covariates into two levels: severe and non-severe. The severe group for each variable will include only those in the highest level. This means that those coded as severe for the ambulation and manual dexterity are those at level 4 (wheelchair dependent / unable feed or dress themselves). IQ is a continuous variable, we define a severe IQ to be less than 50. Vision is only recorded as a binary covariate anyway. These definitions have been used previously by Hutton et al. (2000). Using these binary covariates again minimises the effect of measurement error.

4.2.2 The Work of Hemming et al. (2006)

This data has already been the motivation for work by Hemming et al. (2006). Their paper focused on two main aims. The first was to investigate the long-term survival in adults with CP and compare it to the general

population. They also examined the cause of death. We are mainly interested in the first of these here. Survival is analyzed by birth characteristics and severity of disability conditional on survival until 20 years (and 2 years for a subset of the data). Conditioning on this 20 year survival, 85% of the cohort survived for another 30 years, compared to 95% for the general population. Indeed, expected survival for the CP cohort is consistently lower than for the general population. However, the outlook for survival is generally good. Intellectual ability is shown to be particularly associated with survival.

In general, findings are consistent with the Californian long-term investigation into adult survival Strauss & Shavelle (1998) discussed earlier in Section 4.1.1. It was found in both that excess risk of death over the general population decreased with age. However, the Bristol cohort exhibited an increase in relative risk for females over 50 years of age that is not found in the American study. Both studies found a significant difference between male and female survival. An observation that is not found in any of the childhood CP studies.

There are limitations to this study caused by the nature of the data collection. Its retrospective nature and reliance on case referral have implications with regard to survival estimate biases. Despite the fact that all individuals in the study were first seen before their 20th birthday this does not mean that we have full ascertainment after this time. This is a particular issue with regards to those with less severe disabilities.

There is also a proportion of missing data on each of the covariates. There is some debate about the mechanism behind this missing data. These issues will all be discussed later in this chapter.

4.2.3 Identification of relevant cohorts

As discussed the data were originally collected between 1951 and 1964. However, there are obvious issues with the data that mean we cannot simply consider the remaining data as a representative cohort. The data were collected part-retrospectively, as at the start of the study period Dr Woods looked at all children placed under her care. Some of whom were already of a reasonable age. Due to this retrospective nature of the data, survival times are subject to left truncation and will not be representative of the population as some of the severest cases will have died before they could be seen. We will therefore consider two cohorts. Both conditional on survival until a certain point but one with and one without the issue of left truncation. It should first be noted that we will not include those born in the 1960's as we leave a 5-year notification lag. This is because we have very low levels of data in this period implying we have clear under-ascertainment.

For the first cohort, to eliminate the issue of left truncation, we consider those who survived longer than 22 years and model survival conditional on first reaching this age. We choose the age 22 because all recorded first assessments are done by this time and therefore we will assume that the cohort beyond this age is complete. For this first "adult" cohort we need to examine the ascertainment of individuals as there are clearly smaller numbers in the 1930's and 1940's than in the later decades.

Table 4.1 presents the individuals included in the whole data set by decade of birth and age at first assessment (in years). As we would expect, those born in the 1930's are all seen some time after 10 years. As we have discussed, data collection did not begin until 1951 so it would be im-

Decade of birth	Age at first assessment					Total
	0-4	5-9	10-14	15+	Missing	
1930's	0	0	10	18	3	31
1940's	41	85	40	2	6	174
1950's	168	36	3	1	6	214
1960's	45	2	0	0	5	52

Table 4.1: Number of cases by age of first assessment and decade of birth (n=471)

possible to see children born in the 1930's before they reached a later age. This pattern continues in later decades, with the average age of first assessment decreasing in later decades. From those born in the 1930's and 1940's we clearly observe only children who survived long enough to be seen. However, the question is whether these individuals are representative of children surviving until these later ages.

There is no way of seeing if we have indeed found a representative sample from our target population. However, we can decide whether the group appears to be representative of what we would expect the cohort to look like. To do this we can consider the survival pattern and covariate structure. Our main concern stems from those born in the 1930's as there is very clear under-ascertainment in this decade. Table 4.2 shows the levels of disability by birth decade conditional upon survival until 22 years. As discussed, we consider this restricted subcohort because all individuals in the data set have entered by the age of 22 and, therefore, we will not have to consider the left-truncation issue for this cohort.

Looking at Table 4.2 we can see that there are higher levels of severe ambulation, manual dexterity, and IQ in the earlier decades than those col-

Decade of Birth	Ambulation			Manual Dexterity		
	Non-Severe	Severe	Missing	Non-Severe	Severe	Missing
1930's	0.76	0.17	0.07	0.83	0.10	0.07
1940's	0.86	0.06	0.08	0.88	0.03	0.09
1950's	0.77	0.02	0.21	0.77	0.01	0.22

Decade of Birth	Vision			IQ		
	Non-Severe	Severe	Missing	Non-Severe	Severe	Missing
1930's	0.90	0.00	0.10	0.55	0.38	0.07
1940's	0.92	0.03	0.06	0.73	0.20	0.07
1950's	0.86	0.03	0.11	0.58	0.05	0.37

Table 4.2: Proportions of severe disability and missingness structure by decade of birth (n=368)

lected during the study period. However, there are also lower levels of missing data so it may be that in the 1950's we are missing more data on those with severe disability. This increased level of missingness may be because it is harder to collect this information on younger children. If we look at the survival patterns within each decade we see that they are quite similar (see Figure 4.1). Again this is conditional on survival until age 22. One important issue is going to be the low level of observed severe vision in all decades. This is going to mean we will have little power to estimate any model.

Considering this evidence we decide to include all individuals born between 1930 and 1959 inclusively who survived longer than 22 years in our first "adult" cohort. We expect to lose the most severely disabled, as severity of disability has already been associated with survival. There are 368 individuals in this first sub-cohort.

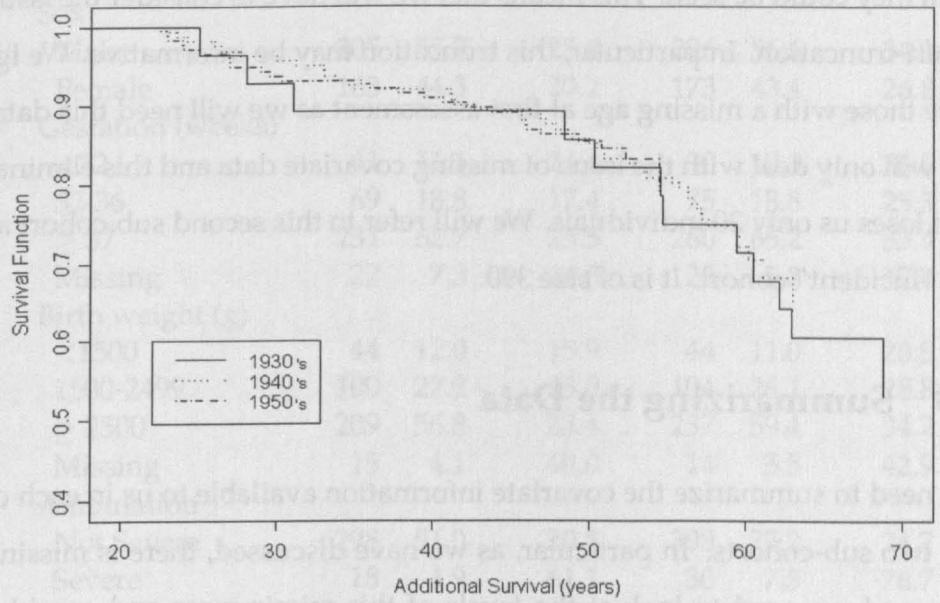


Figure 4.1: Additional survival by decade of birth conditional on survival until 22 years

The second sub-cohort looks at those again born between 1930 and 1959 conditional on survival until two years. From Table 4.1 we see that after the study period began in 1951 Dr Woods started to see individuals in the very first years of life but we condition on two year survival as she will inevitably have missed children who died very early on and did not survive long enough to be referred. We also need to consider if the children in this second cohort are representative cases. We suspect that those with the most severe disabilities, and hence most obvious diagnosis, would have been referred quickly and seen by Dr Woods early on whilst those with less severe disabilities would be see later. However, because these later individuals are

the less disabled they are likely to have longer lifetimes and hence survive until they could be seen. This means that we will have to consider the issue of left-truncation. In particular, this truncation may be informative. We ignore those with a missing age at first assessment as we will need this data. We will only deal with the issue of missing covariate data and this elimination loses us only 20 individuals. We will refer to this second sub-cohort as the "incident" cohort. It is of size 390.

4.3 Summarizing the Data

We need to summarize the covariate information available to us in each of the two sub-cohorts. In particular, as we have discussed, there is missing data and we need to look at the levels of this missingness and consider the missing data mechanism behind it. The adult cohort consists of 368 individuals of which 85 (23%) have a recorded death. In comparison, the incident cohort is 399 individuals with 126 (32%) deaths before December 2005. Note that we have very high levels of censoring. This leads to less precision in survival model estimates but is common in epidemiological studies particularly those with such potentially long survival times.

Section 4.2.1 presented the available covariates for the data. Table 4.3 summarizes the data for each of the two sub-cohorts. It can also be noted that the mean values of gestational age, in weeks, are 37.7 for the adult cohort and 37.8 for the incident cohort. The corresponding means for birth weight are 2690g and 2725g respectively. There are slightly more men in each of our sub-cohorts than there are women and they have a slightly higher death rate. We can see that the majority of individuals have normal gestational lengths ($\geq$ 37 weeks) and normal birth weight ($\geq$ 2500 grams).

	Adult Sub-cohort			Incident Cohort		
	No.	%	% dead	No.	%	% dead
Sex						
Male	205	55.7	25.4	226	56.6	34.1
Female	163	44.3	20.2	173	43.4	26.8
Gestation (weeks)						
<32	41	11.1	24.4	30	10.3	26.8
32-36	69	18.8	17.4	75	18.8	25.3
≥ 37	231	62.7	25.5	260	65.2	35.4
Missing	27	7.3	14.8	23	5.8	17.4
Birth weight (g)						
<1500	44	12.0	15.9	44	11.0	20.5
1500-2499	100	27.2	23.0	104	26.1	28.8
≥ 2500	209	56.8	23.4	237	59.4	34.2
Missing	15	4.1	40.0	14	3.5	42.9
Ambulation						
Not Severe	298	81.0	20.5	308	77.2	24.7
Severe	18	4.9	61.1	30	7.5	76.7
Missing	52	14.1	25.0	61	15.3	44.3
Manual dexterity						
Not Severe	303	82.3	22.1	317	79.4	26.5
Severe	9	2.4	55.6	17	4.3	76.5
Missing	56	15.2	23.2	65	16.3	44.6
Vision						
Not Severe	327	88.9	22.0	348	87.2	27.3
Severe	10	2.7	30.0	14	3.5	71.4
Missing	31	8.4	32.3	37	9.3	56.8
IQ						
Not Severe	236	64.1	17.4	244	61.2	19.7
Severe	52	14.1	42.3	67	16.8	62.7
Missing	80	21.7	27.5	88	22.1	40.9
Number severe dis.						
0	208	56.5	16.3	216	54.1	18.5
1	31	8.4	45.2	35	8.8	54.3
2	6	1.6	50.0	7	1.8	71.4
3	5	1.4	40.0	7	1.8	57.1
4	2	0.5	100.0	7	1.8	100.0
Missing	116	25.9	38.1	127	31.8	40.2

Table 4.3: Birth characteristics and levels of disability for two cohort groups with cerebral palsy

We can also seen that there seems to be little association between either of these covariates and survival outcome although this will be investigated further in the following section. Levels of missingness on these two variables are quite low. This is probably because this information is routinely recorded on medical records so is easily found. If we consider the disability covariates however, we see higher levels of missing data (8-22% for the adult cohort and (10-22% for the incident cohort). We also see greater association between severity of disability and death within the follow-up period although there are particulary low proportions with recorded severe disability particularly with regards to manual dexterity and vision. In previous research, (Hutton & Pharoah 2002) suggest that the number of severe disabilities is also associated with outcome and we can see this here although data is sparse. If we consider the incident cohort then we see that there is still a low observation of severe disabilities although there are slightly more observed than in the adult cohort. This is as we would expect because in the adult cohort we expect to have lost some of the more disabled individuals because they do not survive until 22 years. Hemming et al. (2005) present data on the proportion of children with severe disability in a selection of British studies which we can see is around $2 - 4\%$ so this is similar to the levels we observe. Interestingly, we see higher levels of severe IQ. We note that the percentages of individuals with at least one disability covariate missing is high and therefore it is possibly not sensible to consider the number of severe level disabilities as having a possible effect on survival although they are not that much higher than for the IQ variable. However, we observe so few individuals at some levels that estimating survival at these levels would be very difficult.

We can see the levels of missingness on the covariates increase in the

incident cohort. This increase is greater in the disability covariates than the birth characteristics suggesting missingness may be dependent on survival time or entry time.

4.4 Available Case Survival Analysis

Our first step is to consider non-parametric survival analysis. In Section 2.1.3 we discussed the Kaplan-Meier survival function estimate and in Section 2.2.1 how to adapt it to left truncated data. We now use these methods to investigate the effect of the covariates on survival based upon the available case data and present summarized life table data.

	Survival Time			
	30y	40y	50y	p*
Total	0.95 (0.93-0.97)	0.91 (0.88-0.94)	0.86 (0.82-0.89)	
Gender				
Male	0.95 (0.92-0.98)	0.90 (0.86-0.94)	0.84 (0.79-0.89)	0.27
Female	0.95 (0.92-0.99)	0.93 (0.89-0.97)	0.89 (0.84-0.94)	
Gestation (weeks)				
<32	0.93 (0.83-1.00)	0.93 (0.85-1.00)	0.82 (0.70-0.95)	0.45
32-36	0.97 (0.93-1.00)	0.96 (0.91-1.00)	0.92 (0.85-0.98)	
≥ 37	0.94 (0.91-0.97)	0.89 (0.85-0.93)	0.84 (0.80-0.89)	
Birth weight (g)				
<1500	0.98 (0.93-1.00)	0.98 (0.93-1.00)	0.88 (0.79-0.99)	0.53
1500-2499	0.94 (0.89-0.99)	0.93 (0.88-0.98)	0.90 (0.84-0.96)	
≥ 2500	0.94 (0.91-0.98)	0.89 (0.84-0.93)	0.84 (0.79-0.89)	
Ambulation				
Not Severe	0.96 (0.94-0.99)	0.94 (0.91-0.97)	0.88 (0.84-0.92)	<0.001
Severe	0.88 (0.68-1.00)	0.67 (0.42-0.88)	0.56 (0.37-0.84)	
Manual dexterity				
Not Severe	0.96 (0.93-0.98)	0.92 (0.89-0.95)	0.87 (0.84-0.91)	0.03
Severe	0.78 (0.55-1.00)	0.67 (0.42-1.00)	0.56 (0.31-1.00)	
Vision				
Not Severe	0.96 (0.94-0.98)	0.92 (0.90-0.95)	0.87 (0.84-0.91)	0.25
Severe	0.90 (0.59-1.00)	0.70 (0.47-1.00)	0.70 (0.47-1.00)	
IQ				
Not Severe	0.98 (0.97-1.00)	0.96 (0.94-0.99)	0.92 (0.88-0.95)	<0.001
Severe	0.89 (0.80-0.98)	0.77 (0.66-0.89)	0.69 (0.58-0.83)	

p* - Wilcoxon test p-value

Table 4.4: Estimated survival percentages (95 percent confidence intervals) for the adult cohort

Firstly, if we look at Table 4.4, we can see survival life tables for the adult sub-cohort. The full sub-cohort has a 50 year survival rate of 86%. We can see that survival is very strongly associated with severe levels of ambulation, manual dexterity, and IQ ($p<0.01$). There does not seem to be any association between poor vision and survival although Table 4.3 shows that we observe only three deaths with severe disability.

We repeat the analysis for the incident data. Recall that this is now conditional upon survival until age two but we now need to allow for the left truncation of some of the survival times. Results are presented in Table 4.5.

	Survival Time				
	10y	20y	30y	50y	p*
Total	0.94 (0.91-0.96)	0.90 (0.87-0.93)	0.85 (0.81-0.88)	0.76 (0.72-0.80)	
Gender					
Male	0.94 (0.90-0.96)	0.90 (0.85-0.93)	0.84 (0.79-0.88)	0.74 (0.68-0.79)	0.19
Female	0.95 (0.90-0.97)	0.90 (0.85-0.94)	0.85 (0.79-0.90)	0.79 (0.73-0.85)	
Gestation (wks)					
<32	1.00 (1.00-1.00)	0.98 (0.84-1.00)	0.90 (0.76-0.96)	0.80 (0.63-0.89)	0.08
32-36	0.93 (0.85-0.97)	0.89 (0.80-0.95)	0.87 (0.76-0.93)	0.81 (0.71-0.88)	
≥ 37	0.93 (0.89-0.96)	0.89 (0.84-0.92)	0.82 (0.77-0.87)	0.73 (0.67-0.78)	
Birth weight (g)					
<1500	1.00 (1.00-1.00)	0.95 (0.83-0.99)	0.93 (0.80-0.98)	0.83 (0.68-0.92)	0.02
1500-2499	0.93 (0.86-0.97)	0.92 (0.85-0.96)	0.86 (0.78-0.92)	0.82 (0.74-0.89)	
≥ 2500	0.93 (0.89-0.96)	0.87 (0.82-0.91)	0.81 (0.76-0.86)	0.72 (0.66-0.77)	
Ambulation					
Not Severe	0.97 (0.95-0.99)	0.95 (0.92-0.97)	0.91 (0.87-0.94)	0.83 (0.78-0.87)	<0.001
Severe	0.80 (0.61-0.90)	0.63 (0.44-0.78)	0.50 (0.31-0.66)	0.33 (0.17-0.50)	
Manual dext.					
Not Severe	0.97 (0.94-0.98)	0.95 (0.92-0.97)	0.90 (0.86-0.93)	0.82 (0.77-0.86)	<0.001
Severe	0.82 (0.55-0.94)	0.59 (0.33-0.78)	0.41 (0.19-0.63)	0.29 (0.11-0.51)	
Vision					
Not Severe	0.97 (0.94-0.98)	0.93 (0.90-0.96)	0.89 (0.85-0.92)	0.81 (0.77-0.85)	<0.001
Severe	0.71 (0.41-0.88)	0.57 (0.28-0.78)	0.35 (0.13-0.59)	0.29 (0.08-0.59)	
IQ					
Not Severe	0.98 (0.96-0.99)	0.97 (0.94-0.99)	0.95 (0.91-0.97)	0.88 (0.84-0.92)	<0.001
Severe	0.88 (0.78-0.94)	0.73 (0.61-0.82)	0.61 (0.48-0.72)	0.46 (0.34-0.58)	

p* - Wilcoxon test p-value

Table 4.5: Estimated survival percentages (95 percent confidence intervals) for the incident cohort

In this analysis we again used the Wilcoxon test to compare the sur-

vival across covariate strata as this test is appropriate when the alternative hypothesis to the null hypothesis of no difference in the hazard function is that of non-proportional but different hazards. Note that we must allow for the left truncation when estimating the survival curves and calculating the Wilcoxon test statistic as it is well known that the Kaplan-Meier underestimates survival in the presence of truncation (Pan & Chappell 1998). This extension to the Wilcoxon test was programmed in S-Plus.

Again we see that severe disability is highly associated with survival. In particular, a severe visual impairment is now highly significant ($p<0.001$), a relationship that was not apparent in the adult cohort due to a lack of severe observations. Gestational age and birth weight can also be seen to be associated with survival. If we compare the survival proportions at 30 and 50 years for the adult cohort and the incident cohort we see that the decreasing trend in survival over an increase in birth weight is more clearly defined than before, and this is reflected by the p-values of the Wilcoxon test. This is despite not seeing vastly different proportions of severe cases to the adult cohort.

This analysis highlights one of the major problems with complete and available case analysis. By ignoring observations with a missing covariate we reduce our sample size considerably and hence our ability to extract information.

4.5 Considering the Missing Data Mechanism

We have seen that there is a reasonable amount of missing covariate data within our data set. Whenever we wish to conduct analysis in the presence of missing data we need to consider the mechanism behind the observa-

tion process. The possibilities for this were discussed in Section 2.3.1. We can think about the missing data mechanism behind the unobserved data on the disability covariates. Intuitively, this information is possibly more likely to be missing if the lifetime is very short as an individual is more likely to have left the study before all their information was recorded. Table 4.6 shows that those, for the incident cohort, those with a failure or censoring time of less than six years have very high levels of missing data and that these probabilities decrease as survival time increases. However, survival is good at young ages so numbers failing early are small.

Disability	Survival time (years)		
	0-5 (n=10)	6-10 (n=15)	11+ (n=374)
Ambulation	0.60	0.20	0.14
Manual dexterity	0.60	0.27	0.15
Vision	0.60	0.20	0.07
IQ	0.8	0.20	0.21

Table 4.6: Proportions of missing covariate data for the disability covariates in the incident cohort by length of lifetime

We do not know if attempts to record data were continued over the individuals lifetime or just at the first assessment. If data was only recorded at the first assessment it is possible that an early entry into the study increases the probability of missing data as it more difficult to gather from young children or, if attempts were continued over the study period, it is possible that those entering in the late 1950's are more likely to have data missing. Note, however, that we have allowed for a five year lag so this is likely to counteract this second possibility.

We can also use logistic regression methods to look at the effect of sur-

vival time upon the missing data mechanism. We construct an vector, Y, of Bernoulli indicator variables for each of the four disability covariates whose entries show whether a value is observed or missing. We define $\pi_i = \mathrm{P}(Y_i = 1)$ (i.e. the probability the covariate value for the ith individual is missing) and use a logistic link function to construct the model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 t_i.$$

Here, t_i is the observed survival time of individual i and β_0 and β_1 are parameters. We can fit this model via maximum likelihood methods. Having fitted this model for each of the four disabilities we can then look at the significance of the survival time in each model by comparing the fit of each model to that of the null models i.e we look to see if $\beta_1 = 0$. In Table 4.7 we present p-values from a series of univariate analyses of deviance using the χ^2 distribution to do this comparison.

Disability	χ^2 p-value	
	Adult cohort	Incident cohort
Ambulation	0.022	<0.001
Manual dexterity	0.131	<0.001
Vision	0.072	<0.001
IQ	<0.001	<0.001

Table 4.7: Analysis of deviance to consider the effect of survival time on the probability of missing disability data

From Table 4.7 we see that the effect of survival time on the missing data mechanism is highly significant in the incident cohort. As we are including in this cohort those children who died very young this is unsurprising as the child may not have been in follow up long enough to collect the data. In the adult cohort we see a significant effect on the missingness of ambu-

lation and IQ. Children start to walk a different ages so they would have to survive to an old enough age in order to determine if they have a true disability or just have not started to walk yet. Conversely, a disability in the hands and arms could be detected earlier. To measure IQ a child would have to survive until an age when they had the language skills to take the required test. There is a smaller effect of survival time upon the missingness of the vision data. However, referring back to Table 4.3 recall that this variable had the smallest proportion of missing data making an effect harder to detect.

Decade of Birth	Ambulation				Manual Dexterity			
	Age at first assessment				Age at first assessment			
	0-4	5-9	10-14	15+	0-4	5-9	10-14	15+
1930's	-	-	0.00	0.06	-	-	0.00	0.06
1940's	0.12	0.01	0.08	0.50	0.10	0.04	0.13	0.50
1950's	0.28	0.08	0.33	0.00	0.26	0.19	0.33	0.00

Decade of Birth	Vision				IQ			
	Age at first assessment				Age at first assessment			
	0-4	5-9	10-14	15+	0-4	5-9	10-14	15+
1930's	-	-	0.00	0.11	-	-	0.00	0.06
1940's	0.07	0.01	0.08	0.00	0.03	0.08	0.08	0.00
1950's	0.15	0.11	0.00	0.00	0.44	0.14	0.00	0.00

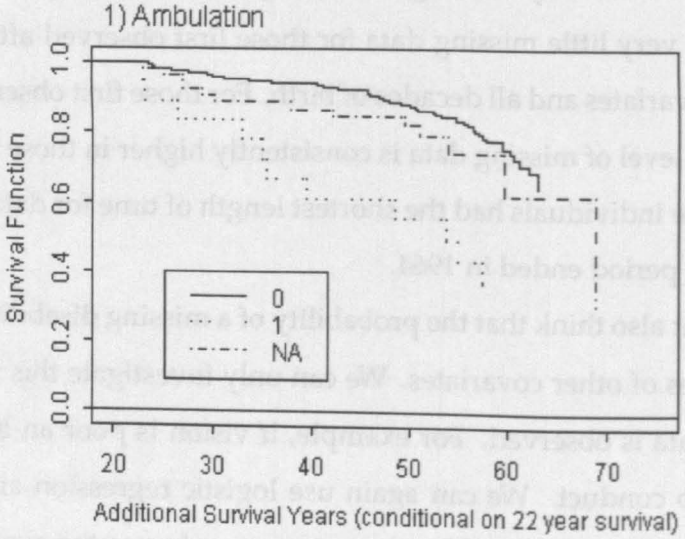
Table 4.8: Proportions of missing covariate data for the disability covariates in the incident cohort

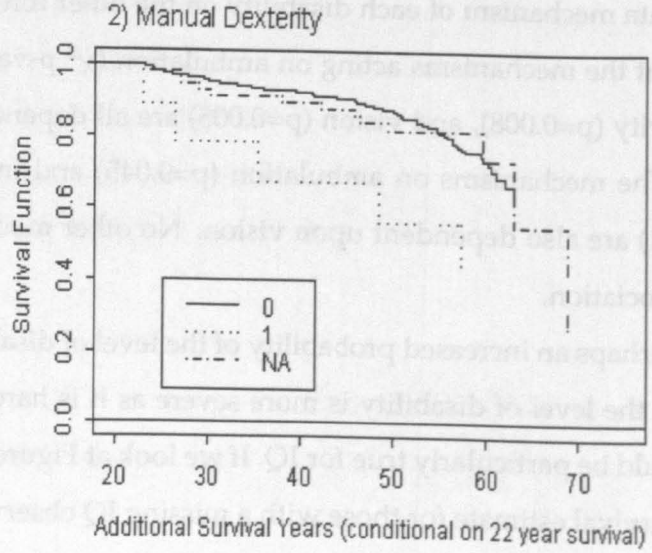
Table 4.8 presents the proportion of missing data for each of the four disability covariates by decade of birth and age at first assessment. One obvious feature is that the levels of missingness are, in general, lowest for those born in the 1930's and highest for those born in the 1950's. It should

be noted, from Table 4.1, that there are only two individuals born in the 1940's first observed after the age of 15 so, with a missing proportion of 0.05 we are in fact only missing one data point. Therefore, it also seems that there is very little missing data for those first observed after 10 years across all covariates and all decades of birth. For those first observed before 10 years the level of missing data is consistently higher in those born in the 1950's. These individuals had the shortest length of time for data collection as the study period ended in 1964.

We might also think that the probability of a missing disability depends on the values of other covariates. We can only investigate this if the other disability data is observed. For example, if vision is poor an IQ test may be harder to conduct. We can again use logistic regression and analysis of deviance to look at this. However, we can only use the available data. We consider each pairwise univariate model looking at the dependence of the missing data mechanism of each disability on the other three. If we do this we see that the mechanisms acting on ambulation (χ^2 p-value=0.047), manual dexterity (p=0.008), and vision (p=0.005) are all dependent on the value of IQ. The mechanisms on ambulation (p=0.045) and manual dexterity (p=0.001) are also dependent upon vision. No other model shows a significant association.

There is perhaps an increased probability of the level of disability being unobserved if the level of disability is more severe as it is harder to measure. This would be particularly true for IQ. If we look at Figure 4.2 we can see that the survival estimate for those with a missing IQ observation have a survival rate at shorter lifetimes similar to those with an observed severe covariate. For the manual dexterity covariate the survival curve for those with missing data seems very similar to those with non-severe disability.


1) Ambulation
Survival Function
0.0
0.2
0.4
0.6
0.8
1.0
0
1
NA
20
30
40
50
60
70
Additional Survival Years (conditional on 22 year survival)


2) Manual Dexterity
Survival Function
0.0
0.2
0.4
0.6
0.8
1.0
0
1
NA
20
30
40
50
60
70
Additional Survival Years (conditional on 22 year survival)

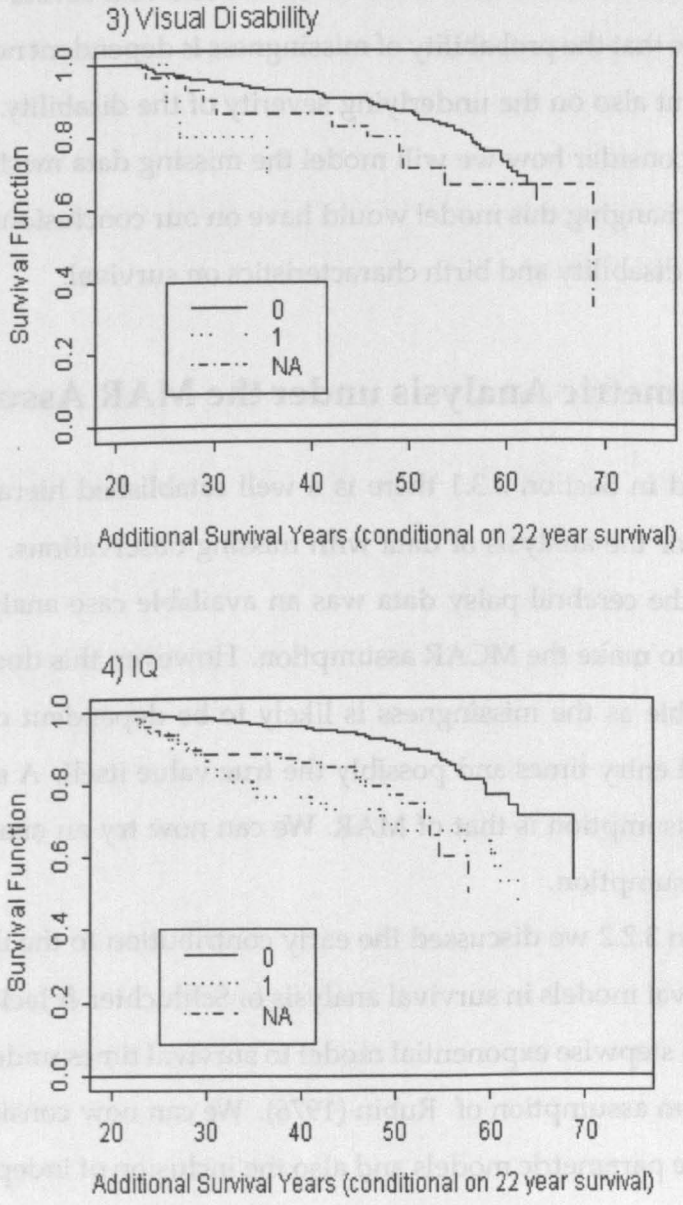


Figure 4.2: Survival by level of disability for the adult cohort including those with missing covariate data

For the remaining disabilities the survival curve for those with missing data lies somewhere between the curves for severe and non-severe disabilities. It is plausible that the probability of missingness is dependent not only time covariates but also on the underlying severity of the disability. Therefore, we need to consider how we will model the missing data mechanism and what effect changing this model would have on our conclusions regarding the effect of disability and birth characteristics on survival.

4.6 Parametric Analysis under the MAR Assumption

As discussed in Section 2.3.1 there is a well established hierarchy of assumptions for the analysis of data with missing observations. Our initial analysis of the cerebral palsy data was an available case analysis so this required us to make the MCAR assumption. However, this does not seem at all plausible as the missingness is likely to be dependent on both the survival and entry times and possibly the true value itself. A slightly less restrictive assumption is that of MAR. We can now try an analysis based upon this assumption.

In Section 3.2.2 we discussed the early contribution to the use of parametric survival models in survival analysis of Schluchter & Jackson (1989). They fitted a stepwise exponential model to survival times under the missing at random assumption of Rubin (1976). We can now consider the use of alternative parametric models and also the inclusion of independent left truncation.

4.6.1 Introduction to the MAR Model

If we start with complete data we can describe a model for the underlying true time to failure. Assume the complete data consist of (t_i, δ_i, z_i) for $i = 1, ..., n$, ignoring the late entry for the moment. As in Section 2.1.1, t_i is the observed survival, δ_i the censoring indicator, and z_i a vector of p categorical covariates. The covariates define a contingency table with $M = I_1 \times ... \times I_p$ cells, where I_j is the number of levels for the jth covariate. Also define θ_m to be the probability associated with cell m, $(m = 1, ..., M)$, such that $\sum_{m=1}^{M} \theta_m = 1$. Let $\mathbf{E}_i = (E_{i1}, ..., E_{iM})'$ be a multinomial indicator vector whose mth component is 1 if subject i belongs to cell m and is 0 otherwise. Note that $\mathbf{E}_i$ is only known if no data are missing for i. We must also define a vector $\mathbf{W}_i = (W_{i1}, ..., W_{iM})'$ which indicates which cells of the contingency table i could possibly be given the actual observed covariate information.

The model is written using complete data. To describe the distribution of survival times, conditional on the covariates, let $\lambda_m(t)$ denote the hazard function of a subject belonging to the mth cell defined by the covariates. Schluchter and Jackson assume this hazard to be a stepwise function on K disjoint time interval defined by arbitrarily chosen cut points $0 = T_0^* < T_1^* < ... < T_K^* = \infty$:

$$\lambda_m(t) = \lambda_{km}, \quad T_{k-1}^* < t \leq T_k^*.$$

A log-linear parametrization for the hazard function can be adopted.

Schluchter and Jackson then go on to construct the likelihood function for the observed data. Let b_{ki} be the amount of exposure time contributed

by subject i to the kth time interval:

$$\begin{aligned} b_{ki} &= 0 && \text{If } T_i < T^*_{k-1}, \\ &= T_i - T^*_{k-1} && \text{if } T^*_{k-1} < T_i \leq T^*_k, \text{ or} \\ &= T^*_k - T^*_{k-1} && \text{if } T^*_k < T_i. \end{aligned}$$

We define S_{im} to be the probability that a subject in cell m will survive up to time T_i, for $i = 1, ..., n$ and $m = 1, ..., M$. Therefore,

$$S_{im} = \exp\left(-\mathrm{H}(t_i)\right) = \exp\left(-\sum_{k=1}^{K} \lambda_{km} b_{ki}\right),$$

if we define H(t) to be the cumulative hazard function. If the censorship or failure for subject i occurred in the kth time interval, then, using that $p(t, z|\lambda, \theta) = p(t|z, \lambda)p(z|\theta)$ and the form of the likelihood for right censored data discussed in Section 2.1.4, the contribution to the likelihood for subject i is proportional to

$$l_i = \sum_{m=1}^{M} W_{im}\theta_m S_{im}\lambda_{km}^{\delta_i}. \tag{4.1}$$

We can now move on to maximizing this full log-likelihood (the sum of the individual contributions in Equation 4.1). Firstly, the log-likelihood for the hypothetical complete-data set $(T_i, \delta_i, \mathbf{E}_i)$, $i = 1, ..., n$ can be shown, except for an additive constant, to be

$$l^* = \sum_{k=1}^{K}\sum_{m=1}^{M} \{D_{km}\log(\lambda_{km}) - \lambda_{km}U_{km}\} + \sum_{m=1}^{M} V_m \log(\theta_m), \tag{4.2}$$

where D_{km} is the number of failures that occurred in the kth time interval amongst individuals belonging to the mth cell, U_{km} is the equivalent ex-

posure time, and V_m is the number of subjects belonging to cell m. That is

$$\begin{aligned} D_{km} &= \sum_{i=1}^{n} E_{im}\delta_i I(T^*_{k-1} < T_i \leq T^*_k), \\ U_{km} &= \sum_{i=1}^{n} E_{im}b_{ki}, \\ V_m &= \sum_{i=1}^{n} E_{im}. \end{aligned}$$

The paper describes two methods for maximizing this log-likelihood, the EM algorithm and the Newton-Raphson algorithm. We look at the EM algorithm.

The E-step involves computing the conditional expectation of the log-likelihood, Equation 4.2, given the observed data $(T_i, \delta_i, \mathbf{W}_i)$, $i = 1, ..., n$. Therefore, we need to calculate the expectation of D_{km}, U_{km}, and V_m. It is seen that these are equal to

$$\begin{aligned} T^{(1)}_{km} &= \sum_{i=1}^{n} \delta_i I(T^*_{k-1} < T_i \leq T^*_k) P_{im}, \\ T^{(2)}_{km} &= \sum_{i=1}^{n} b_{ki} P_{im}, \\ T^{(3)}_{m} &= \sum_{i=1}^{n} P_{im}, \end{aligned} \tag{4.3}$$

where $P_{im} = Pr(E_{im} = 1 | T_i, \delta_i, \mathbf{W}_i)$ is the posterior probability that subject i belongs to cell m given the observed failure and censoring information, and the observed covariate information.

If failure or censoring occurs in the kth time interval for a subject i, then an equation for P_{im} is

$$P_{im} = \frac{\lambda_{km}^{\delta_i} S_{im} P(E_{im} = 1|\mathbf{W}_i)}{\sum_{l=1}^{M} \lambda_{kl}^{\delta_i} S_{il} P(E_{il} = 1|\mathbf{W}_i)}, \tag{4.4}$$

where

$$P(E_{im} = 1|\mathbf{W}_i) = \frac{W_{im}\theta_{im}}{\sum_{l=1}^{M} W_{il}\theta_l}.$$

When fitting the saturated log-linear model, the updated estimates of the cell probabilities and hazard parameters obtained in the M-step of the algorithm are simply

$$\tilde{\theta}_m = \frac{T_m^{(3)}}{n} \tag{4.5}$$

and

$$\tilde{\lambda}_{km} = \frac{T_{km}^{(1)}}{T_{km}^{(2)}}. \tag{4.6}$$

The algorithm alternates between the E-step, Equations 4.3 and 4.4, and the M-step, Equations 4.5 and 4.6, until convergence. This is taken to be when the log-likelihood changes by < 0.0001. This value is arbitrary but must be sufficiently small. When fitting an unsaturated model the M-step of the algorithm also requires the application of one step of the IPF (Iterative Proportionality Fitting) algorithm to the counts contained in $T_{km}^{(1)}$. Standard errors can be calculated using the information matrix, details of which are presented in the paper's appendix.

4.6.2 Parametric Extension to the Model

One of the most obvious questions with regards to this model is can we adapt it to allow parametric hazards which are likely to be more realistic

in survival analysis? Also, given that our motivating data is subject to left-truncation, how do we construct the likelihood to allow for this?

Shao & Zhou (2004) discussed the use of the Burr XII distribution for the analysis of survival data with long-term survivors. This general distribution, suggested by Burr (1942), has the Weibull and log-logistic distributions as special cases. These are distributions commonly used in survival analysis. The Burr XII distribution function is given by

$$f_B(t|\lambda,\alpha,\beta) = \alpha\lambda t^{\alpha-1}\left\{1+\beta\lambda t^{\alpha}\right\}^{-\left(1+\frac{1}{\beta}\right)} \quad \lambda,\alpha, \text{ and } \beta > 0$$

with survival and hazard functions

$$\begin{aligned} S_B(t|\lambda,\alpha,\beta) &= \left\{1+\beta\lambda t^{\alpha}\right\}^{-\frac{1}{\beta}}, \quad \text{and} \\ h_B(t|\lambda,\alpha,\beta) &= \alpha\lambda t^{\alpha-1}\left\{1+\beta\lambda t^{\alpha}\right\}^{-1}. \end{aligned}$$

The Weibull distribution occurs as $\beta \to 0$ and the log-logistic distribution when $\beta = 1$. A criterion can be used to derived to test if $\beta = 0$ using results from Vu & Zhou (1997). It can also be noted that the Burr XII also has the Pareto distribution as a special case if $\alpha \to \infty$ and $\lambda \to 0$ with $\alpha\lambda$ fixed (or tends to a limit).

The standard form of the likelihood function under the assumption of independent left-truncation was given in Section 2.2.2.

Using the Burr XII distribution the log-likelihood allowing for indepen-

dent right-censoring and left-truncation is given by

$$l(\lambda,\alpha,\beta,\theta) = \prod_{i=1}^{n}\left[\sum_{m=1}^{M} w_{im}\theta_{m_i}\frac{h_B(t_i|\lambda_{m_i},\alpha_{m_i},\beta_{m_i})^{\delta_i}S_B(t_i|\lambda_{m_i},\alpha_{m_i},\beta_{m_i})}{S_B(y_i|\lambda_{m_i},\alpha_{m_i},\beta_{m_i})}\right]$$
$$= \prod_{i=1}^{n}\left[\sum_{m=1}^{M} w_{im}\theta_{im}\left(\alpha_{m_i}\lambda_{m_i}t_i^{\alpha_{m_i}-1}\{1+\beta_{m_i}\lambda_{m_i}t_i^{\alpha_{m_i}}\}^{-1}\right)^{\delta_i}\right.$$
$$\left.\times\{1+\beta_{m_i}\lambda_{m_i}t_i^{\alpha_{m_i}}\}^{-\frac{1}{\beta_{m_i}}}\{1+\beta_{m_i}\lambda_{m_i}y_{i_{m_i}}^{\alpha}\}^{\frac{1}{\beta_{m_i}}}\right] \tag{4.7}$$

where y_i is the left-truncated time of first assessment. Note that we are using the same model structure for the categorical covariates as just described where θ_m is the probability of lying in cell m. This general log-likelihood can be used to fit a separate Burr XII distribution to the survival times in each cell of the covariate defined contingency table. We can show that the Burr XII distribution can be used under the accelerated failure distribution assumption. This assumption keeps the shape parameters (α, β) fixed across the levels and changes the scale parameter, λ. Therefore, if we wish to fit an AFT model we must make α_m and β_m constant over m. The likelihood functions for fitting Weibull and Log-Normal models can also be calculated. Using the parameterizations of these distributions in Section 2.1.4 these are as follows.

Weibull distribution

$$l(\lambda,\kappa,\theta) = \prod_{i=1}^{n}\left[\sum_{m=1}^{M} w_{im}\theta_{m_i}(\lambda_{m_i}\kappa_{m_i}t_i^{\kappa_{m_i}-1})^{\delta_i}\frac{\exp(\lambda_{m_i}t_i^{\kappa_{m_i}})}{\exp(\lambda_{m_i}y_i^{\kappa_{m_i}})}\right]$$

Log-Normal distribution

$$l(\mu,\sigma,\theta) = \prod_{i=1}^{n}\left[\sum_{m=1}^{M} w_{im}\theta_{m_i}\left(\frac{1}{t\sigma\sqrt{2\pi}}\exp\left(-\{\log t-\mu\}^2/2\sigma^2\right)\right)^{\delta_i} \times\left\{1-\Phi\left(\frac{\log t-\mu}{\sigma\sqrt{2}}\right)\right\}\right].$$

The S-Plus (Insightful Corporation 2002) code used to implement these models via the Newton Raphson algorithm is included in Appendix A. We assume non-informative left truncation for the analysis of the incident cohort. This means that we must modify the likelihood by conditioning upon survival until the observed point of entry as in Equation 2.1.

We could also consider allowing for informative left-truncation. In which case, we have to model the distribution of entry times. This model can be extended to a multivariate setting using continuous covariates.

4.7 Application of the MAR model to Cerebral Palsy Data

We can now fit these parametric models to the two sub-cohorts identified in Section 4.2.3. We consider only the univariate case and investigate the effect of severe disability upon survival. Table 4.9 gives the maximum likelihoods for the different distributions. During calculation difficulties were encountered with the convergence of the Burr XII model when the true distribution appears to be the Weibull. This is because the Weibull occurs as a limiting case of the Burr distribution.

By comparing the maximum likelihoods we can choose the best fitting parametric model. If we consider models fitted to both sub-cohorts we see

1) Adult cohort

Disability	Model distribution			
	Burr XII	Log-logistic	Weibull	Log-normal
Ambulation	*	-563.37	-562.13	-567.74
Manual dext.	*	-539.20	-537.66	-543.75
Vision	*	-545.01	-543.31	-549.98
IQ	-629.18	-629.70	-628.84	-632.43

2) Incident cohort

Disability	Model distribution			
	Burr XII	Log-logistic	Weibull	Log-normal
Ambulation	*	-843.23	-826.69	-832.39
Manual dext.	*	-816.41	-799.99	-806.54
Vision	*	-809.98	-793.33	-799.03
IQ	-897.53	-897.63	-882.01	-835.73

* - Failure to converge

Table 4.9: Maximum log-likelihood values for univariate accelerated failure models over different distributions under the MAR assumption

that all disabilities are best modeled using the Weibull distribution. We will consider later the impact this has on the estimates hazard functions. As discussed, and as can be seen in the table, convergence can be difficult to obtain with the Burr XII distribution. Hutton & Monaghan (2002) discuss the impact of misspecification of parametric survival models. Accelerated failure models are reasonably robust to misspecification because of their log-linear structure.

We can fit the appropriate survival models and compare to available case estimates. Figures 4.3- 4.6 show the comparison of the optimal models under the MAR assumption to the equivalent model calculated by available case analysis for the adult cohort. Recall that this cohort is conditional upon

survival until 22 years.

Available case analysis appears to lead to very similar estimates to the MAR model for the physical disabilities. Figure 4.6 shows that in the available case analysis we are overestimating survival for the severely intelligently impaired. Table 4.3 shows that of the four disability covariates IQ had the highest level of missing data in the adult cohort as well as the highest numbers of observed severe. Figure 4.2 showed that the non-parametric survival for the individuals with missing IQ was closer to that of the severe level individuals than for the other covariates and so it is sensible that allowing for a less restrictive missing data mechanism lowers survival. We also see a drop of approximately 6 years in the median survival ($\approx 67.1 - 60.8$ years). We can also consider survival for the incident cohort which is conditional upon survival until 2 years.

Figures 4.7- 4.10 show the estimated survival curves for the effect of severity on survival for the four different disability covariates and compares them to their available case equivalents. We see that restriction to the MAR assumption increases the survival for non-severe impairment but decreases the estimated survival for those with severe disability. The difference is much greater than in the adult cohort. As we are now only conditioning upon survival until 2 years we expect to pick up more deaths and hence, survival to be lower.

Indeed, we see that survival in the early years is dramatically different between those with severe and non-severe disabilities.

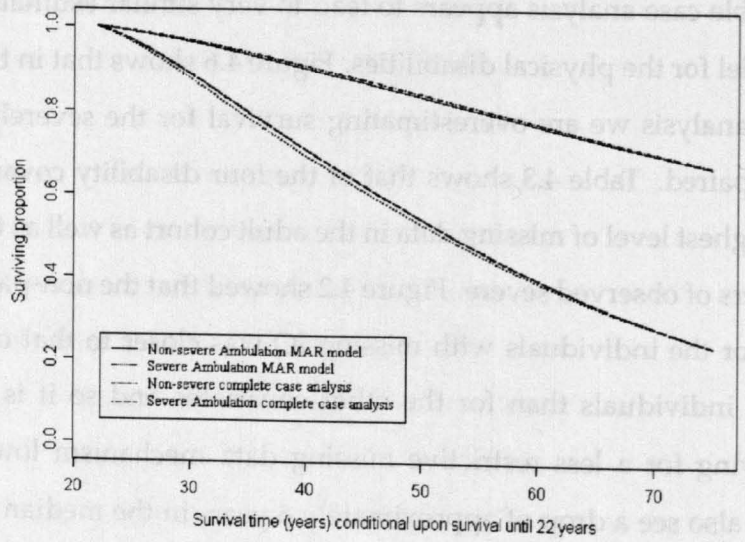


Figure 4.3: Survival by severity of ambulation for the adult cohort under the MAR assumption

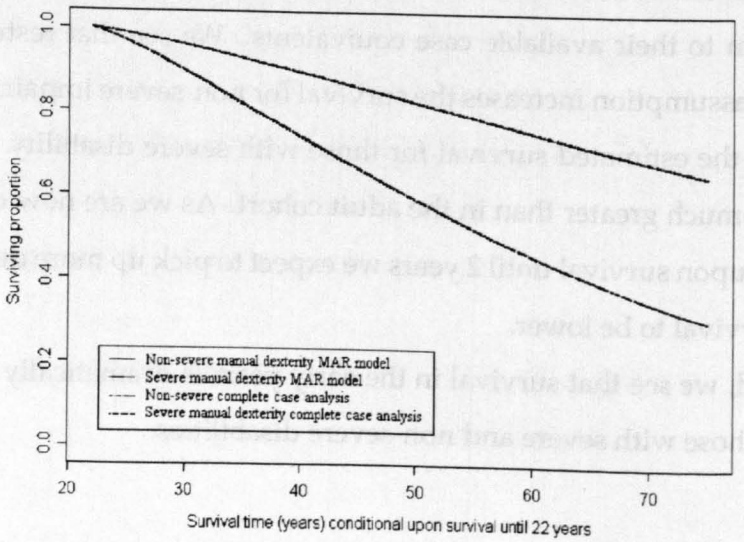


Figure 4.4: Survival by severity of manual dexterity for the adult cohort under the MAR assumption

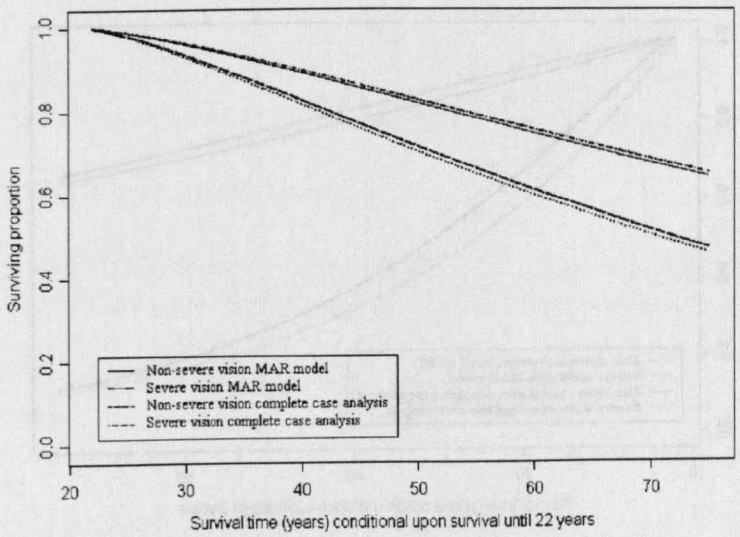


Figure 4.5: Survival by severity of visual impairment for the adult cohort under the MAR assumption

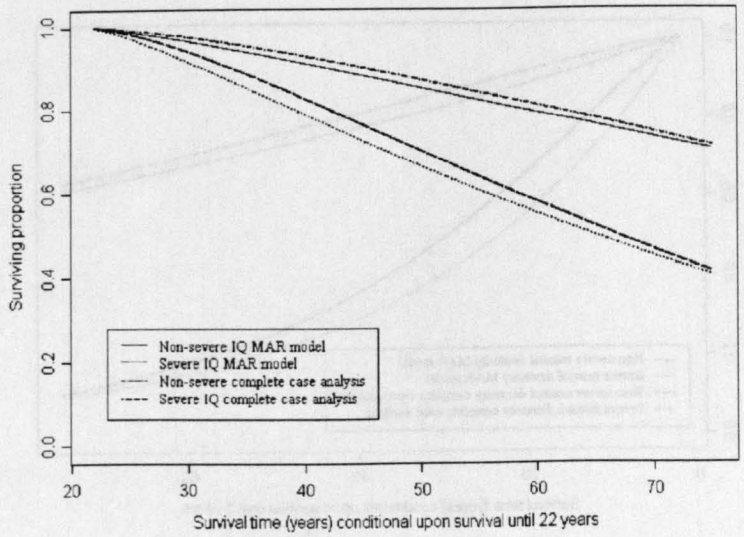


Figure 4.6: Survival by severity of IQ for the adult cohort under the MAR assumption

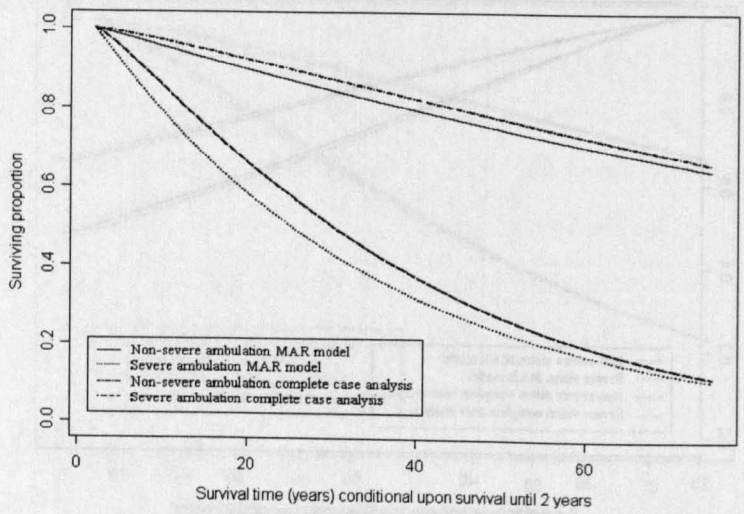


Figure 4.7: Survival by severity of ambulation for the incident cohort under the MAR assumption

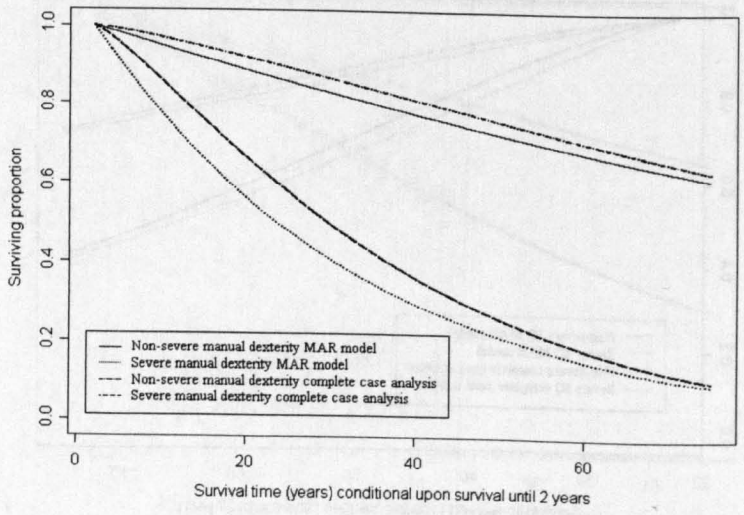


Figure 4.8: Survival by severity of manual dexterity for the incident cohort under the MAR assumption

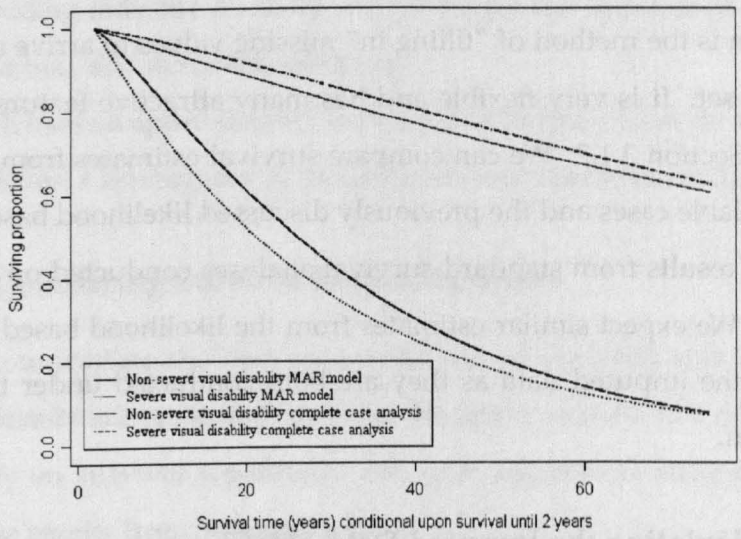


Figure 4.9: Survival by severity of visual impairment for the incident cohort under the MAR assumption

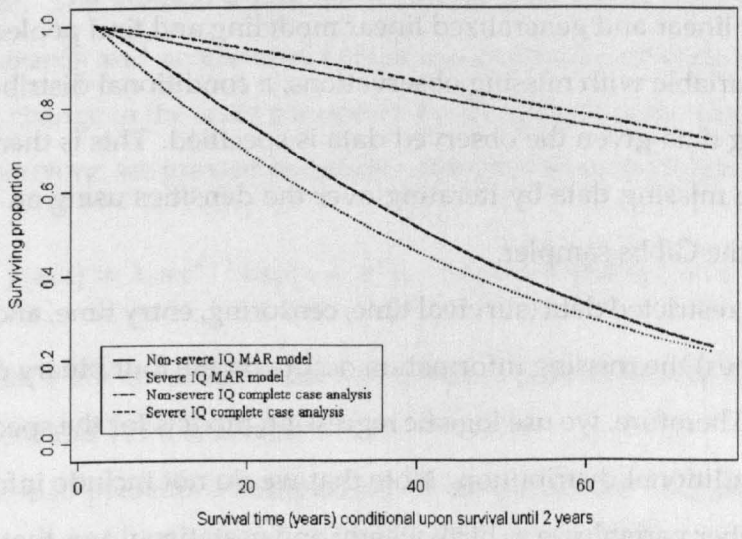


Figure 4.10: Survival by severity of IQ for the incident cohort under the MAR assumption

4.8 Multiple Imputation (MI)

Imputation is the method of "filling in" missing values to arrive at a complete data set. It is very flexible and has many attractive features as discussed in Section 3.1.2. We can compare survival estimates from analysis using available cases and the previously discussed likelihood based methods to the results from standard survival analyses conducted on imputed data sets. We expect similar estimates from the likelihood based analysis and from the imputed data as they are both conducted under the MAR assumption.

4.8.1 Calculating the Imputed Data - MICE

MICE (Van Buuren & Oudshoorn 1999) is a software library for S-Plus and R. There is also an implementation for STATA. MICE stands for "Multiple Imputation by Chained Equations". It can be used to impute data as well as conduct linear and generalized linear modeling and find pooled results. For each variable with missing observations, a conditional distribution for the missing data given the observed data is specified. This is then used to impute the missing data by iterating over the densities using an approximation to the Gibbs sampler.

On our restricted data (survival time, censoring, entry time, and disability covariates) the missing information occurs on the four binary disability variables. Therefore, we use logistic regression models for the specification of each conditional distribution. Note that we do not include information from the other variables, e.g. birth weight and gestational age, that we have as we have not included these in the likelihood based analysis. However, an advantage of MI is that we can easily use all available data to impute

the missing information. We use all the information in our restricted data set (including only the disability variables) for the imputation in order to minimise bias and maximise certainty.

MICE uses an approximate Gibbs sampler to draw from the conditional distributions. Convergence of the sampler must therefore be checked.

4.8.2 Comparing Survival Model Estimates

Separate univariate analyses are carried out on the adult and incident cohorts. Results are presented for both. We again consider the effect of each disability on survival separately. For each imputation analysis we combined the results from 5 imputations.

The Adult cohort

The effect of each covariate was best modeled by a Weibull distribution in each case. The Weibull distribution can be used under both the proportional hazards and accelerated failure assumptions and each assumption causes a change in the scale parameter λ only and not in the shape parameter κ. Therefore, we present parameter estimates from the following model:

$$f(x|z) = \lambda_z \kappa x^{\kappa-1} \exp(-\lambda_z x^{\kappa}), \qquad \lambda, \kappa > 0 \text{ and } 0 \leq x < \infty.$$

Recall that z is a covariate value and equals 0 if the disability of interest is non-severe and 1 if it is severe.

Table 4.10 presents a comparison of the parameter estimates across a available case, a multiple imputation, and a likelihood based analysis for the adult cohort. We can see that the available case analysis consistently underestimates the effect of a severe level disability in comparison to the

	$\hat{\lambda_0}$	$\hat{\lambda_1}$	$\hat{\kappa}$
Ambulation			
Available case	0.0023 (0.0012)	0.0050 (0.0032)	1.3109 (0.1487)
MI	0.0028 (0.0014)	0.0100 (0.0055)	1.2485 (0.1343)
Likelihood	0.0028 (0.0014)	0.0099 (0.0056)	1.2500 (0.1387)
Manual dexterity			
Available case	0.0028 (0.0015)	0.0078 (0.0054)	1.2618 (0.1457)
MI	0.0030 (0.0014)	0.0093 (0.0057)	1.2448 (0.1292)
Likelihood	0.0030 (0.0014)	0.0083 (0.0054)	1.2464 (0.1315)
Vision			
Available case	0.0027 (0.0014)	0.0047 (0.0036)	1.2779 (0.1431)
MI	0.0030 (0.0015)	0.0054 (0.0039)	1.2528 (0.1326)
Likelihood	0.0030 (0.0015)	0.0054 (0.0039)	1.2526 (0.1328)
IQ			
Available case	0.0012 (0.0007)	0.0031 (0.0020)	1.4258 (0.1720)
MI	0.0025 (0.0012)	0.0073 (0.0036)	1.2364 (0.1294)
Likelihood	0.0025 (0.0011)	0.0065 (0.0031)	1.2421 (0.1230)

Table 4.10: Comparison of parameters (s.e.) from available case, multiple imputation, and likelihood based analyses for univariate disabilities in the adult cohort

other methods i.e. λ_1 is smaller than for the MAR methods. However, the intercept λ_0 term is similarly estimated by each method. This suggests that, as suspected, we are missing more information on severe level disabilities. Conversely, the estimated shape parameter κ is greater for the available case analysis particularly when considering the effect of severe low IQ, implying that the hazard function increases less rapidly than estimated for the available case. We expected to be including more early deaths as we are incorporating those with more severe disability so this fits with this result. We can consider the estimated survival curves for each model.

Also in this table we present standard errors for the parameter estimates. These are found numerically using the *vcov.nlminb* function in the *MASS* library of S-Plus. This method uses a finite difference approximation

to the Hessian matrix. We are pleased to note the fact that the standard errors estimated using MI and the likelihood approach are almost identical. Our data is of quite a simple structure and the simple method for imputation that we used seems to have been accurate. Using the MAR assumption seems to increase our precision only on the shape parameter, κ, and not on the estimated scale. However, estimated errors are not too dissimilar over any of the three methods. The magnitude of the standard errors show that there is no significant difference between the parameters over the MCAR and MAR assumptions.

		90% survival		75% survival	
		Non-sev	Severe	Non-sev	Severe
Ambulation	AC	41.4	29.3	63.8	37.7
	MI	40.3	28.6	62.9	36.7
	Likelihood	40.2	28.6	62.7	36.8
Manual Dexterity	AC	39.7	29.9	61.3	39.4
	MI	39.4	29.0	61.1	37.8
	Likelihood	39.4	29.7	60.9	39.2
Vision	AC	39.6	33.4	60.6	47.0
	MI	39.1	32.7	60.2	45.9
	Likelihood	39.1	32.7	60.2	45.9
IQ	AC	45.1	33.9	68.7	46.0
	MI	42.6	30.7	68.4	41.5
	Likelihood	42.3	31.4	67.6	43.1

Table 4.11: Comparison of available case (AC), multiple imputation (MI), and likelihood based analyses for univariate disabilities in the adult cohort - 90% and 75% survival (in years from birth).

Table 4.11 gives estimated survival times for 90% and 75% of the individuals by level of disability. We can see that estimated survival times are generally similar across methods. In particular, survival is similar for non-severe levels of physical disability. Non-severe IQ differs as the 90% survival for the available case analysis is 2.5 years higher than for the MAR

methods. However, this gap has decreased by the 75% survival time to approximately 1 year. Severe level physical disability survival is also generally similarly estimated with 90% survival differing by a maximum of 0.9 years. However, differences in the parameter estimates for the effect of IQ cause a greater difference and also a difference between estimates from the likelihood and multiple imputation methods. This is very apparent when considering the severe level survival where the range of estimates for 75% survival times is 4.5 years.

Recall that each univariate model was best fitted using a Weibull hazard. With parameters similar to those estimated this implies a monotonically increasing hazard function. This is unsurprising as we are conditioning on survival until 22 years so we will not expect an initially larger hazard as may occur for the incident cohort.

The Incident Cohort

Recall that the incident cohort consists of individuals with survival greater than 2 years.

The Weibull was again the chosen model, for each disability. We are now studying earlier survival and would expect a larger number of earlier deaths particularly in those most disabled. When we conditioned on 22 year survival this initial period had passed and so the hazard was much more linear. Table 4.12 presents parameter estimates for the incident cohort. For shape parameters κ as estimated here the hazard is only slightly increasing over time.

As with the adult cohort the available case analysis is underestimating the scale parameters. Also the scale parameter is again larger in the

	$\hat{\lambda_0}$	$\hat{\lambda_1}$	$\hat{\kappa}$
Ambulation			
Available case	0.0026 (0.0018)	0.0136 (0.0091)	1.1788 (0.1598)
MI	0.0057 (0.0025)	0.0304 (0.0136)	1.0125 (0.1052)
Likelihood	0.0058 (0.0027)	0.0291 (0.0135)	1.0113 (0.1091)
Manual dexterity			
Available case	0.0026 (0.0016)	0.0126 (0.0082)	1.2050 (0.1481)
MI	0.0071 (0.0030)	0.0310 (0.0145)	0.9902 (0.1025)
Likelihood	0.0068 (0.0030)	0.0328 (0.0158)	0.9930 (0.1055)
Vision			
Available case	0.0030 (0.0017)	0.0163 (0.0100)	1.1798 (0.1381)
MI	0.0071 (0.0031)	0.0345 (0.0167)	0.9946 (0.1019)
Likelihood	0.0069 (0.0031)	0.0363 (0.0182)	1.0000 (0.1089)
IQ			
Available case	0.0018 (0.0005)	0.0082 (0.0021)	1.2066 (0.1491)
MI	0.0045 (0.0020)	0.0206 (0.0088)	1.0168 (0.1022)
Likelihood	0.0045 (0.0012)	0.0197 (0.0053)	1.0129 (0.1563)

Table 4.12: Comparison of parameters (s.e.) from available case, multiple imputation, and likelihood based analyses for univariate disabilities in the incident cohort

available case models which may lead to some "trade-off" when we consider the survival curves. Multiple imputation estimates and likelihood estimates are again very similar in magnitude. The shape parameter κ for each of these models is lower than for the same models in the adult cohort. This is caused by the lower ages and the higher number of deaths at these low ages increasing the early hazard.

One thing to note here is that we have presented models based on the distribution choice for the MAR likelihood model. Note that in each case the scale parameter κ is close to 1 which is the point at which a Weibull hazard switches from being monotonically increasing to decreasing. This means that the best model choice might be difficult to identify although survival estimates are unlikely to be too sensitive to the final choice.

We can again consider the estimated survival curves by looking at the 90th and 75th survival percentiles. These are presented in Table 4.13.

		90% survival		75% survival	
		Non-sev	Severe	Non-sev	Severe
Ambulation	AC	24.7	7.7	55.4	15.3
	MI	19.8	5.4	50.0	11.2
	Likelihood	19.6	5.6	49.6	11.7
Manual Dexterity	AC	23.7	7.8	51.9	15.4
	MI	18.2	5.0	45.9	10.1
	Likelihood	17.7	5.2	45.2	10.9
Vision	AC	22.6	6.9	50.3	13.4
	MI	17.2	5.0	43.6	10.1
	Likelihood	17.4	4.9	43.9	9.9
IQ	AC	31.1	10.3	68.9	21.0
	MI	24.4	7.2	62.3	16.1
	Likelihood	24.5	6.9	62.3	15.1

Table 4.13: Comparison of available case, multiple imputation, and likelihood based analyses for univariate disabilities in the incident cohort - 90% and 75% survival.

By considering Table 4.13 we can see that, unsurprisingly, survival estimates are lower under the MAR assumption than under the MCAR assumption. Unlike with the analysis of the adult cohort it seems that data is no longer missing completely at random. This means we get differences in estimated survival over the different approaches. Looking at the each disability model we see the same dramatic decrease in survival at severe levels in comparison to non-severe levels. For example, the estimated 75% survival time, under the MAR assumption, drops from 46 to 10 years when a child suffers from severe manual dexterity. Again, multiple imputation and model based survival estimates are quite similar.

4.9 Conclusions and Summary

In this chapter we have introduced our motivating data and summarized its main features. We have decided to consider survival for two different sub-cohorts conditional upon survival to 2 and 22 years, the adult cohort and the incident cohort.

Firstly, we introduced the available information. Little is known about how, or exactly when this data was collected. We know disabilities can not have been measured until after a child was referred to and seen by Dr Woods but we do not know if all the data was then immediately obtained. This has several implications. Issues can arise with longitudinal data if covariates change over time. However, we know that cerebral palsy is a non-degenerative condition so the level of disability will not change in the first years of life. Disability may, of course, increase at a much older age but we can be sure all our data were collected prior to this becoming an issue. What may be more of an issue with our data is that the methods used to assess disability do change overtime. The broad categories of the physical disabilities mean that it is unlikely that a child would have been differently assessed at different time points but we do not know how IQ was measured. Using a binary covariate her, instead of a fully continuous variable, means that we can minimise any effect that changing methods may have. While changes in time may result in small changes in an estimate of IQ the probability of being classified incorrectly into one of the two groups is low.

Changes in referral habits may have also changed over time meaning that the children we see from early on in the study period are actually from a slightly different population to those that we see later. In deciding which data to include in our analysis we looked at how the level of severe dis-

ability changed over time. We saw reasonably consistent levels of severity over time but this was complicated by changes proportions of missing data. This suggests that referral patterns stayed similar with regards to the type of children being referred over time.

We used standard survival analysis techniques, the life table and Kaplan-Meier survival estimates to look at non-parametric survival in each of the two cohorts. The level of failure was also looked at in each cohort. We observed a moderate proportion of missing data. We then compared survival using both the MCAR and MAR assumption via likelihood based methods and multiple imputation.

Severe disability is highly associated with a decrease in survival. However, an available case analysis seems to underestimate this association slightly. This difference does not appear to be significant in the adult cohort but is apparent in the incident cohort where we include a large number of early deaths. By decreasing the restriction on the missing data mechanism to be missing at random we find decreased levels of survival for, in particular, the severely disabled cases in the incident cohort. This suggests that data is not MAR and that missingness is particularly dependent on low survival times. However, as discussed in Section 4.5, it is possible that data is not missing at random so we will now consider a model for the missing data mechanism.

In this Chapter we compared estimates from a model allowing for the MAR assumption to estimates using standard survival analysis methods based upon imputed data. These models both assumed slightly different mechanisms. The MAR model allowed survival to depend upon the survival time while the imputation method did not. Therefore, it was slightly surprising that both methods led to very similar results. This was seen in

Tables 4.10 and 4.12. There is clearly a very close association between survival time and the covariates and it seems there may be little additional effect on missingness from the survival time once all the covariates have been accounted for.

We can use the analysis from this chapter to consider the shape of the hazard functions. This is discussed in the following chapter after we have also modelled the missing data mechanism.

Chapter 5

Modelling the Missing Data Mechanism

In Chapter 4 we were introduced to the motivating data set of a cohort of children from the Bristol area of the UK suffering from cerebral palsy. As discussed, out interest lies in looking at the effect of severity of disability upon survival, particularly at longer survival times. However, the covariates within the data are subject to a certain level of missingness which we believe may be not missing at random (NMAR). This was discussed in Section 4.5. Therefore, the techniques discussed in Chapter 3 for dealing with missing data in survival analysis may not appropriate as they generally require the more restrictive MAR assumption. We will have to model the missing data mechanism.

In Section 2.3.1 we introduced the notation of Rubin (1976), Y_{mis}, the missing data, Y_{obs}, the observed data, and M, the missing data mechanism indicator matrix. We discussed the confusion that arises from this particular notation and presented an alternative. However, for convenience we

will temporarily use the original notation to discuss the likelihood equations under the different missing data mechanisms. Firstly, note that the complete joint distribution of the data and the missing data mechanism can be written as

$$\begin{aligned} f(Y, M|\Theta, \Phi) &= f(Y|\Theta)f(M|Y, \Phi) \\ &= f(Y_{obs}, Y_{mis}|\Theta)f(M|Y_{obs}, Y_{mis}, \Phi) \end{aligned}$$

where Θ and Φ are parameter vectors. The actual data consist of (Y_{obs}, M). By integrating out Y_{mis} from the joint distribution we can obtain the distribution for the observed data

$$f(Y_{obs}, M|\Theta, \Phi) = \int f(Y_{obs}, Y_{mis}|\Theta)f(M|Y_{obs}, Y_{mis}, \Phi)\, dY_{mis}. \qquad (5.1)$$

Missing data is defined to be *ignorable* if the missing data is MAR and the parameters Θ and Φ are distinct i.e. the joint parameter space of (Θ, Φ) is the product of the two individual parameter spaces. This is because under these conditions we can ignore the missing data mechanism when constructing the likelihood as it will depend only on Y_{mis}. A likelihood function is a conditional probability function considered as a function of its second argument with its first argument held fixed. We can use a likelihood function to calculate maximum likelihood estimates of the model parameters, indeed, we previously used this method in the MAR model in Section 4.6. The likelihood ignoring the missing data mechanism can be defined as

$$L_{ign}(\Theta|Y_{obs}) \propto f(Y_{obs}|\Theta)$$

where $f(Y_{obs}|\Theta)$ is obtained by integrating Y_{mis} out of the density $f(Y|\Theta)$.

The full likelihood is defined as

$$L_{full}(\Theta, \Phi | Y_{obs}, M) \propto f(Y_{obs}, M | \Theta, \Phi) \tag{5.2}$$

where $f(Y_{obs}, M|\Theta, \Phi)$ is obtained by integrating Y_{mis} out of the density $f(Y, M|\Theta, \Phi)$ as in Equation 5.1. Maximum likelihood estimates can be found by maximising L_{full}, in Equation 5.2, with respect to Θ and Φ. Occasionally, the missing data mechanism is known but in general it is not and parameters Φ must be estimated. Examples with known missing data mechanisms can be found in Chapter 15 of Little & Rubin (2002). Grouped or rounded data are examples of known missing data mechanisms. Note that these are examples of coarsened data as discussed by Heitjan & Rubin (1991).

There are two main approaches to formulating models for non-ignorable data. Assume that the observations to be modelled are independent. Selection models have the joint distribution of M and Y, where M is the missing data mechanism and Y is the full data set (see Section 2.3.1) in the form

$$f(M, Y|\Theta, \Phi) = f(Y|\Theta) f(M|Y, \Phi)$$

where θ and Φ are distinct. Here, conditioning on any complete covariates is suppressed. The model that we go on to formulate is of this form. Alternatively, pattern mixture models have the form

$$f(M, Y|\Psi, \Omega) = f(Y|M, \Psi) f(M|\Omega)$$

where Ψ and Ω are again distinct parameter vectors.

5.1 Non-Ignorable Missing Data and Selection Bias

The issue of selection bias raises similar questions to that of NMAR missing covariate data. Selection bias occurs when a sample used for inference is not randomly selected and hence, calculated statistics are biased. If this bias is not taken into consideration then any conclusions drawn may be invalid. The use of a complete case sample in data that is not MCAR results in selection bias. Another example of selection bias is publication bias. This occurs in meta analyses when insignificant or contradictory results are not included due to non-publication, hence, magnifying the overall positive effect in a meta analysis. Another example occurs in economics when investigating wage levels and it is this area that sparked a development of models to deal with selection bias. Similar issues arise with drop-out in longitudinal studies. There has also been considerable research into non-ignorable missing data in categorical data particularly within survey data.

5.1.1 Normal Selection Models for Non-Ignorable Missing Data

Some of the most influential work in selection bias within the economics literature is the seminal research of Heckman (1974). He considered selection bias with particular reference to market wage studies. For example, the wages for migrants do not provide a reliable estimate of what non-migrants would have earned if they had migrated. To model selection bias he used a simple characterization involving two equations. Consider a random sample of n individuals. For individual i, $(i = 1, .., n)$ ...

$$\begin{aligned} Y_{1i} &= X_{1i}^T \beta_1 + U_{1i} \\ Y_{2i} &= X_{2i}^T \beta_2 + U_{2i} \end{aligned}$$

where X_{ji} is a vector of K_j regressors, β_j is a vector of K_j parameters and U_{1i} and U_{2i} are such that $E(U_{ji}) = 0$ and $E(U_{ji}U_{j'i''}) = \sigma_{jj'}$ for $i = i''$ (= 0 otherwise). Also assume that the regressor matrix is of full rank so that if all the data were available parameters could be estimated by least squares regression. Suppose that our variable of interest is Y_1 so we wish to estimate parameters β_1 but that some data on Y_1 are missing. The population regressor function can be written as

$$E(Y_{1i}|X_{1i}) = X_{1i}^T\beta_1, \quad i = 1, .., n.$$

However, the regressor function for the available data is

$$E(Y_{1i}|X_{1i}, \text{ sample selection rule}) = X_{1i}^T\beta_1 + E(U_{1i}|\text{ sample selection rule}) \tag{5.3}$$

for observed individuals only. If the conditional expectation of U_{1i} in Equation 5.3 is zero the regression functions for the available data and the full data are the same so ordinary least squares on the complete data may be used to estimate β_1 and the only cost is a loss of efficiency. In general, this is not the case. Assume that data is observed on Y_{1i} only if $Y_{2i} \geq 0$. If $Y_{2j} < 0$ then Y_{1i} is not observed and hence the individual is not included in the sub-sample. This implies that

$$\begin{aligned} E(U_{1i}|X_{1i}, \text{ sample selection rule}) &= E(U_{1i}|X_{1i}, Y_{2,i} \geq 0) \\ &= E(U_{1i}|X_{1i}, U_{2i} \geq -X_{2i}^T\beta_2). \end{aligned}$$

The selected sample regression function therefore depends on both X_{1i} and X_{2i}. Ignoring the condition expectation of U_{1i}, i.e. fitting the model to the

observed data only, results in bias arising from omitted variables. It should be noted that if the joint density of U_{1i} and U_{2i} is a singular normal density and $X_{1i} = X_{2i}$, $\beta_1 \equiv \beta_2$ then the Tobit model emerges (Tobin 1958).

Heckman type models have been used extensively in the economics literature. Heckman (1979) describes the construction of a model that uses a bivariate normal density for the model errors. Olsen (1980) extends this model by removing this bivariate normal assumption. He shows that Heckman's model does not in fact require bivariate normality of the errors but only normality of U_{2i} and of $U_{1i}|U_{2i}$. Bivariate normality is sufficient for this but not necessary.

Heckman introduces a procedure for fitting this model where it is necessary to estimate the Mill's ratio. Mill's ratio is defined as

$$\lambda_i = \frac{\phi(Z_i)}{\Phi(Z_i)}$$

where $\phi(.)$ is the standard normal probability density function, $\Phi(.)$ its cumulative distribution function and where $Z_i = -\frac{X_{2i}^T \beta_2}{\sqrt{Var(U_{2i})}}$. This is estimated via use of a probit model in the first step of a two step procedure. Olsen goes on to derive a model where U_{2i} is assumed to have a standard uniform distribution where it is necessary to estimate a linear selection model in place of Mill's ratio. If $U_{1i}|U_{2i}$ is normal this only really leads to obviously different results when the correlation between U_{1i} and U_{2i} is strong. Olsen also describes conditions for the identifiability of the two step fitting method used by Heckman. For the linear selection model in Olsen (1980), variables are required in X_2 that are not included in X_1. Whilst for the bivariate normal model described by Heckman (1979), the probit model is identifiable even if $X_1 = X_2$ provided X_1 contains terms other

than a constant. Although this does rely on the nonlinearity of the Mill's ratio. However, even though empirical experiments by Olsen suggest a degree of robustness of models of this type Little (1985) highlights a structural assumption needed for identifiability which may prove to be inappropriate. He also discusses specific covariate requirements for stability of these models. This instability can be overcome by using maximum likelihood instead of a two step procedure. However, such methods are sensitive to misspecification of the distribution of U_{1i}.

More recently, Puhani (2000) gives an overview of Monte Carlo studies of Heckman's two step method. He concludes that the procedure is often inefficient particularly when there is correlation between the covariates in the outcome and selection models. However, Heckman (1979) himself writes that the main purpose of his estimator is to provide good starting values for maximum likelihood estimation. Indeed, given the progress in computing power since Heckman introduced his procedure maximum likelihood methods are recommended.

Copas & Li (1997) looked at inference in non-random samples. They discussed Heckman's two step procedure within the statistical literature and present, in detail, its restrictions. They investigate a full likelihood approach to fitting the bivariate normal model. In particular, they look the sensitivity of model parameters close to the missing at random assumption. They conclude that the likelihood is often flat in shape suggesting that the data provide little information about sample selection. However, they are considering meta analysis in which there are typically only a few trials or studies with no individual patient data, we have a cohort study with considerably more available data. They also question whether any clear evidence may be the result of model misspecification. However, they

agree with previous work that using conventional methods assuming MAR when even a small level of selection bias is present can lead to conclusions that are grossly misleading. They suggest use of a sensitivity analysis.

Selection type models are attractive as they have an intuitive structure. They have the same factorization used in the definitions of MCAR, MAR, and NMAR (see Section 2.3.1). Also, MCAR and MAR models can be obtained as special cases of a NMAR model by setting certain parameters equal to zero. However, the parameter estimates rely heavily upon distributional assumptions which suggests that we should not use models of this type to test the MAR assumption (Kenward 1998).

This is an important point of discussion. If the survival estimates are so heavily reliant upon the choice of distribution then the bias caused by misspecification may cause a problem. However, this does not mean that models of this type are not of use. If our model for the missing data mechanism is suitably flexible we can use the results to guide our understanding of the unknown distribution. We must be careful not to rely to heavily on the exact estimates.

5.1.2 Normal Pattern-Mixture Models for Non-Ignorable Missing Data

Little (1994) discusses the use of pattern mixture models in modeling non-ignorable missing data. He extends earlier results looking at maximum likelihood estimates for ignorable data. He also considers a Bayesian approach to inference. The model applies to bivariate normal data. Little discusses his opinions on both selection models and pattern mixture models. He concludes that the efficiency of selection type models is better if

the distributional assumptions are correct as they allow direct estimation of certain parameters that pattern mixture models require to be fixed a priori. However, the information provided for these parameters can be weak and relies heavily on the distributional assumptions so is susceptible to misspecification. Little & Wang (1996) extend this model to a multivariate model with monotone missing data on one outcome variable. However, Tang et al. (2003) suggest that it cannot be extended to general multivariate settings.

5.1.3 Models for Publication Bias

As discussed in Section 5.1 a specific example of selection bias is publication bias. This is a major problem in meta analyses in medical statistics. We will discuss here the most recent work of Copas & Shi (2000, 2001).

They assume that the ith study in the population of interest has parameter estimate of interest y_i with

$$y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$$

and

$$\mu_i \sim \mathrm{N}(\mu, \tau^2).$$

This is the standard random effects population model. Random effects are used to describe the heterogeneity of the data. They also have a selection model where they assume that the probability of publication or selection depends upon the reported standard deviation s of y in such a way that

$$P(select|s) = \Phi\left(a + \frac{b}{s}\right).$$

Here Φ is the standard normal cumulative distribution function. This can

be written in an equivalent way using the model

$$z_i = a + \frac{b}{s_i} + \omega_i, \text{ where } \omega_i \sim \mathrm{N}(0,1).$$

Where, without loss of generality we can say that a model is selected if and only if $z > 0$. Noting that we can model y as

$$y_i = \mu_i + \sigma_i \epsilon_i, \text{ where } \epsilon_i \sim \mathrm{N}(0,1)$$

they combine the models by using the jointly normal errors (ϵ_i, ω_i) and defining that the $\mathrm{corr}(y_i, z_i) = \rho$. Therefore, the joint distribution of y and z is multivariate normal.

We can see that this model is the same as that of Heckman (1979) applied specifically to publication bias. It again uses a bivariate normal for the distribution of the two error terms.

5.1.4 Non-Ignorable Missing Categorical Data in Surveys

Non-ignorable missing data has also been considered extensively within the survey literature. For example, Baker & Laird (1988) consider categorical non-ignorable non-response. They propose a hierarchical log-linear model for the joint distribution of the categorical covariates and missing data indicator matrix. Another approach is that of Little (1982) and Nordheim (1984). They both introduce prior odds of response for the different categories. Little then uses the EM algorithm, as discussed in Section 3.1.3 while Nordheim uses closed form estimates. A nice summary is given by Molenberghs et al. (1998). They discuss a wide range of published literature.

5.1.5 Informative Dropout in Repeated Measures Data

When considering data with repeated measurements it is possible that we will have a level of dropout from the study or intermittently missing outcome data. Diggle & Kenward (1994) discuss the issue of dropout in longitudinal studies. The issue here is, of course, missing outcome. Such studies consider repeated measurements on a group of individuals over a period of time. The observed data consists of $\{(y_{ij}, t_{ij}) : i = 1, ..., m, j = 1, ..., n_i\}$ where y_{ij} is the jth measurement on individual i which is obtained at time t_{ij}. The typical objective is to consider the mean response as a function of time and other covariates. In these cases, it is the dropout process that is under consideration. This is when the series of measurements on a particular individual end prematurely. Dropout is considered *informative* or *non-ignorable* if the process depends upon the unobserved measurements i.e. those that would have been observed had the individual not dropped out. Diggle and Kenward allow dropout to depend on current and previous values of Y. This issue is also considered by Wu & Carroll (1988). Their interest lies in estimating the rates of change of a covariate over time between different groups. They use a selection model and allow dropout to depend on the slope of the data. This may be appropriate if people with a rapid decline in outcome dropout more frequently than those with a slow deterioration. Rotnitzky et al. (1998) present methods based upon augmented inverse probability of censoring weighted estimating equations. They propose that this method offers a degree of robustness to misspecification not provided by other likelihood based methods.

A good summary of parametric methods for incomplete longitudinal data is give by Kenward & Molenberghs (1999). They discuss in detail the

use of selection and pattern mixture models for informative dropout.

It should be noted that missing data is an example of coarse data. Where as in missing data literature we are concerned with the observation or complete non-observation of a data point we can consider the more general setting of coarse data. This is when we observe only a subset of the complete data sample space in which the true data lie. Examples of coarse data occur due to rounding, measurement error, censoring, data heaping (i.e. when data contains items reported with various levels of coarseness). Heitjan & Rubin (1991) present work on ignorability within the setting of coarse data. They present a general model for coarse data under a generalization of Rubin's (1976) MAR assumption.

5.2 Introducing the Joint Survival and Missing Data Mechanism Selection Model

As discussed in Chapter 3, the majority of likelihood based methods for dealing with missing data in parametric survival analysis require the MAR assumption. However, this does not seem to be a sensible assumption for our cerebral palsy data (see Section 4.5). Therefore, we must build a model that allows us to model the missing data mechanism. The aim is to embed the MCAR and MAR models within a range of plausible models that allow the NMAR assumption. We can follow ideas already discussed in Section 5.1.1 (Heckman 1979) and Section 5.1.3 (Copas & Shi 2001) and develop a selection type model.

There has been criticism against the use of selection models in miss-

ing data analysis (Section 5.1.2). The main issue is the heavy dependence of the models on distributional assumptions. However, Hutton & Monaghan (2002) discussed misspecification in accelerated failure models and note that they are reasonably robust. We do not intend to propose a prescriptive model but rather a basic model which a sensitivity analysis can be constructed around. In particular, we aim to remain quite flexible on any distributional assumptions.

Firstly, let us establish the notation we will be using. Our data consist of (T, δ, Z) for individuals $i = 1, , ., n$ where $T = t_i$ is the recorded (possibly censored) survival time, $\delta = \delta_i$ is the censoring indicator ($\delta_i = 1$ if t_i = death time), and $Z = z_i$ is the possibly missing covariate information. Note that we are only considering the case when we have fully observed survival time and censoring information on n individuals as this is the case in our motivating data. We will also assume that we have a missing data mechanism denoted as M. Initially, we ignore the issue of truncation.

We wish to construct a model to estimate the joint distribution $f(T = t_i, M = m_i, Z = z_i)$. We can factorize this distribution as follows:

$$f(T = t_i, M = m_i, Z = z_i) = f(M = m_i | T = t_i, Z = z_i) f(T = t_i | Z = z_i) f(Z = z_i).$$

We can see that this is a selection type model.

For simplicity, initially assume we have just one binary covariate, $z = (z_1, ..., z_n)$, which has some missing data. Firstly, we construct a model for the survival times, T, to describe $f(T = t_i | Z = z_i)$:

$$t_i' = \log(t_i) = \eta_0 + \eta_1 z_i + \sigma \epsilon_i, \qquad \epsilon_i \sim N(0, 1), \qquad i = 1, ..., n, \tag{5.4}$$

with $\eta = (\eta_0, \eta_1, \sigma)$. We allow the survival of individual i to depend on

the value of the covariate. Notice the log-linear structure of this model. In Collett (1999), Chapters 4 and 6, he discusses this form of the parametric proportional hazards model and accelerated failure time model respectively. We present the log-normal model first to highlight the comparisons with the selection bias model of Heckman (1979) and the publication bias sensitivity analysis of Copas & Shi (2001). We can also consider the log-logistic or Weibull models as well as other parametric distributions (see Section 5.3.1). Here, η_0 is the baseline log-survival (when $z_i = 0$), η_1 is the effect of the covariate on log-survival, and σ is the standard deviation of the log-survival times.

We choose to consider a fully parametric model not only because previous research suggests that log-logistic models may be useful but also because our interest lies in estimating survival and the main focus of Cox proportional hazard models is the investigation of relative risk.

Secondly, we construct a model for the missing data mechanism using a latent variable M:

$$m_i = \alpha_0 + \alpha_1 z_i + \alpha_2 t_i' + \omega_i, \qquad \omega_i \sim N(0,1) \tag{5.5}$$

where $\alpha = (\alpha_0, \alpha_1, \alpha_2)$. This time we use a linear regression model for a continuous variable M which allows the mechanism to depend upon the covariate and the log survival time. Of course, M is not exactly observed. However, we can, without loss of generality, state that an individual i has missing data on covariate Z if $m_i > 0$. As Z is either observed or missing we can construct an indicator vector for missingness and hence conclude on the sign of each m_i. Assume the residuals (ϵ, ω) are independent and jointly normal with $\text{corr}(\epsilon, \omega) = 0$.

Note that this independent errors assumption differs slightly from other selection models as we incorporate the dependence of the missing data mechanism on the missing covariate directly and not through correlation of the covariate and latent variable.

We must also construct a model for the covariate. As we are using a simple binomial covariate here we can use the model $P(z = 0) = \theta_0 = 1 - \theta_1 = 1 - P(z = 1)$. The structure of this model will have to be carefully considered when allowing for multiple or continuous covariates.

This model allows for all three missing data assumptions. The MAR and MCAR assumptions occur as special cases of the complete model. For example, if we set $\alpha_1 = 0$ and $\alpha_2 = 0$ then we are assuming data are missing completely at random or, if all parameters, α_0, α_1, and α_2 are non-zero then we are allowing the data to be not missing at random. We assume data are missing at random if α_1 is zero. In this case, we do not need to include the model for the missing data mechanism as it will have no bearing upon the maximum likelihood estimates for the survival model. We can have prior beliefs about the values of α_1 and α_2 although we do not include these in our model. If the covariate in question is a disability covariate then we might expect those with more severe forms of the disability to have a higher chance of missing data because children are more likely to die before their disability levels can be ascertained so, therefore, $\alpha_1 < 0$. Conversely, data are, perhaps,more likely to be observed if the individual has a longer lifetime which implies that $\alpha_2 > 0$. However, we can use the likelihood to find estimates for all these parameters. This identifiability is possible due to the linear constraint of the missing data mechanism model. We should consider this assumption and will conduct a sensitivity analysis. However, when adding additional terms to the model identifiability does become an

issue. We should also be cautious about using the parameter estimates as a test for the MCAR or MAR assumptions (Kenward 1998).

As discussed in Sections 5.1.1 and 5.1.2 there is some concern about the reliance of selection models upon the distributional assumptions made. We will have to consider the form of the likelihood to identify parameters and maybe consider a sensitivity analysis, particularly with regards to the model for the missing data mechanism.

5.2.1 Calculating the Likelihood Function

The log-likelihood can now be constructed. Let F be the set of individuals with recorded failure times and C those with censored times. Recall that we denote $t_i' = \log t_i$. With no missing data the likelihood for right censored survival data can be constructed as follows:

$$\begin{aligned} L(\eta, \sigma | t', z) &= \prod_F f(t_i'|z_i, \eta, \sigma) \prod_C S(t_i'|z_i, \eta, \sigma) \\ &= \prod_{i=1}^{n} h(t_i'|z_i, \eta, \sigma)^{\delta_i} S(t_i'|z_i, \eta, \sigma). \end{aligned}$$

As discussed in Section 2.1.2 $S(t|z)$ is the survival function and $h(t|z)$ is the associated hazard function. We do not as yet look at the issue of left truncation. This is possible, and contributions would be of a similar form to those in left-truncated survival data with full information.

We can split the individuals in the data set into four groups based upon their censoring information and missing data indicator. Within each group the subjects can then contribute the same form of information to the likeli-

hood function. The likelihood can be constructed as follows:

$$\begin{aligned} L(\eta,\sigma,\alpha,\theta|t,z,\delta) = & \prod_{F,O} f(T'=t_i', M<0, Z=z_i|\alpha,\eta,\sigma,\theta) \times \\ & \prod_{C,O} S(T'=t_i', M<0, Z=z_i|\alpha,\eta,\sigma,\theta) \times \\ & \prod_{F,M} f(T'=t_i', M>0|\alpha,\eta,\sigma,\theta) \times \\ & \prod_{O,M} S(T'=t_i', M<0|\alpha,\eta,\sigma,\theta). \end{aligned} \tag{5.6}$$

Here, F denotes the subset of individuals with a recorded failure time, C those with censored survival times, O those with an observed covariate, and M the subset with a missing covariate. Recall that the full joint density function can be calculated as the product of the conditional density functions which are given by

$$P(M=m|T'=t',Z=z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(m-\alpha_0-\alpha_1 z-\alpha_2 t')^2\right\},$$

$$P(T'=t'|Z=z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(t'-\eta_0-\eta_1 z)^2\right\}, \text{ and}$$

$$P(Z=z) = \theta_z \text{ such that } \sum_{z=0}^{1} \theta_z = 1.$$

Note also that we are still only considering discrete covariates as the variables in our data set are of this form. Let us now consider the contribution to the likelihood from an individual in each of the four groups:

Group 1) Individual, i, with complete covariate data and failure time, total

number of individuals $= n_1$

$$
\begin{aligned}
L_1(\eta,\sigma,\alpha,\theta|t_i',z_i,m_i) &= P(M<0,T'=t_i',Z=z_i)\\
&= \int_{-\infty}^{0} P(M=m,T'=t_i',Z=z_i)\,dm\\
&= \int_{-\infty}^{0} \frac{1}{2\pi\sigma}\theta_{z_i}\exp\left[-\frac{1}{2}\left\{(m-\alpha_0-\alpha_1 z_i-\alpha_2 t_i')^2+\frac{1}{\sigma^2}\left(t_i'-\eta_0-\eta_1 z_i\right)^2\right\}\right]\text{ dm}\\
&= \frac{1}{\sigma\sqrt{2\pi}}\theta_{z_i}\exp\left\{-\frac{1}{2\sigma^2}(t_i'-\eta_0-\eta_1 z_i)^2\right\}\times\\
&\qquad \int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(m-\alpha_0-\alpha_1 z_i-\alpha_2 t_i')^2\right\}\text{ dm}\\
&= \frac{1}{\sigma\sqrt{2\pi}}\theta_{z_i}\exp\left\{-\frac{1}{2\sigma^2}(t_i'-\eta_0-\eta_1 z_i)^2\right\}\Phi(-\alpha_0-\alpha_1 z_i-\alpha_2 t_i').
\end{aligned}
$$

This is the contribution from each individual in this group where Φ is the standard normal distribution function. Therefore, the group contribution, which is the product of the individual contributions, is

$$
L_1=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n_1}\prod_{\substack{(i:m_i<0)\\ \delta_i=0}}\theta_{z_i}\exp\left\{-\frac{1}{2\sigma^2}(t_i'-\eta_0-\eta_1 z_i)^2\right\}\Phi(-\alpha_0-\alpha_1 z_i-\alpha_2 t_i').
$$

Group 2) Individuals with complete covariate data but censored survival time, total number of individuals $= n_2$ i.e. $i:\ z_i$ obs, $\delta_i=0$.

$$
\begin{aligned}
L_2(\eta,\sigma,\alpha,\theta|t_i',z_i) &= P(M<0,T'>t_i',Z=z_i)\\
&= \int_{t_i'}^{\infty} P(M<0,T'=u,Z=z_i)\,\text{du}\\
&= \frac{1}{\sigma\sqrt{2\pi}}\theta_{z_i}\int_{t_i'}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}(u-\eta_0-\eta_1 z_i)^2\right\}\Phi(-\alpha_0-\alpha_1 z_i-\alpha_2 u)\,\text{du}.
\end{aligned}
$$

This integration can be evaluated using numerical Gaussian quadrature methods. This technique will be further discussed later in this section. Again the full contribution from this group is the product of the individual

contributions from all individuals within the group.

Group 3) Individuals with recorded failure time but missing covariate, total number of individuals = n_3 i.e. $i:\ z_i$ missing, $\delta_i = 1$.

Now we need to consider the distribution of survival times given that the covariate information is unknown. We must look at

$$P(M > 0, T' = t') = \sum_{z=0}^{1} P(M > 0, T' = t' | Z = z) P(Z = z).$$

Therefore,

$$\begin{aligned} L_3(\eta, \sigma, \alpha, \theta | t'_i, m_i) &= P(M > 0, T' = t'_i) \\ &= \sum_{z=0}^{1} P(M > 0, T' = t'_i | Z = z) P(Z = z) \\ &= \sum_{z=0}^{1} \left(\int_0^{\infty} P(M = m, T' = t'_i | Z = z) P(Z = z)\, \mathrm{dm} \right) \\ &= \sum_{z=0}^{1} \frac{1}{\sigma\sqrt{2\pi}} \theta_z \exp\left\{ -\frac{1}{2\sigma^2} (t'_i - \eta_0 - \eta_1 z)^2 \right\} \Phi(\alpha_0 + \alpha_1 z + \alpha_2 t'_i). \end{aligned}$$

Group 4) Individuals with incomplete data and censored failure time, total number of individuals = n_4 i.e. $i:\ z_i$ missing, $\delta_i = 0$.

Using our previous calculations we arrive at the following likelihood

contribution...

$$
\begin{aligned}
L_4(\eta, \sigma, \alpha, \theta | t_i', m_i) &= P(M > 0, T' > t_i') \\
&= \sum_{z=0}^{1} P(M > 0, T' > t_i' | Z = z) P(Z = z) \\
&= \sum_{z=0}^{1} \left(\int_{t_i'}^{\infty} P(M > 0, T' = u | Z = z) P(Z = z) \, \mathrm{du} \right) \\
&= \sum_{z=0}^{1} \left[\int_{t_i'}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \theta_z \exp\left\{ -\frac{1}{2\sigma^2} (u - \eta_1 - \eta_1 z)^2 \right\} \Phi(\alpha_0 + \alpha_1 z + \alpha_2 u) \, \mathrm{du} \right].
\end{aligned}
$$

Now that we have the full log-likelihood (which can be found from the sum of the natural logs of these group contributions) we can use this to fit the model described to our cerebral palsy data via Newton Raphson methods (see Section 3.1.3). These are implemented using the *nlminb* function within S-Plus.

Gaussian Quadrature

Gaussian quadrature is a method of numerical integration which seeks to find the optimal abscissas. It is a weighted sum of function values at specified points within the region of integration. The fundamental theorem of Gaussian quadrature states that the optimal abscissas of the n-point Gaussian quadrature formulas are precisely the roots of the orthogonal polynomial of degree n. The domain of integration for such a rule is conventionally taken as [-1, 1], so the rule is stated as

$$\int_{-1}^{1} f(x) \, \mathrm{dx} \approx \sum_{i=1}^{n} w_i f(x_i)$$

where w_i are the appropriate weights.

We use 10-point quadrature to try and estimate the integrals accurately using a function written in S-Plus and presented in Appendix B along with functions to implement the above log-normal model.

If we could assume that missingness was not dependent on time (i.e. $\alpha_2 = 0$) we could consider analytical integration methods.

5.3 Alternative survival distributions

The previous section, Section 5.2, gives details of our joint model based on a log-normal distribution. However, it is possible to use other survival distributions instead. These include the exponential, the Weibull, and the log-logistic. All of these distributions are commonly used in parametric survival analysis. The parametric forms of each of these distributions was discussed in Section 2.1.4.

Changing the error distribution used changes the likelihood function. Note that we use the same distribution for the survival and missing data mechanism errors. However, this is not necessary.

5.3.1 The log-logistic distribution

The log-logistic has proved to be useful when modeling the survival of cerebral palsy as the hazard initially reaches a peak and then declines. We start with the same model form but change the distribution of the error to change the survival distribution. Therefore,

$$t_i' = \log(t_i) = \eta_0 + \eta_1 z_i + \sigma\epsilon_i, \qquad \epsilon_i \sim \log(0,1), \qquad i = 1, ..., n.$$

Similarly, we construct the missing data mechanism model as

$$m_i = \alpha_0 + \alpha_1 z_i + \alpha_2 t_i' + \omega_i, \qquad \omega_i \sim \log(0,1).$$

Note that now the errors have independent logistic distributions. This means that the distribution of t_i, given z_i, is log-logistic with mean $\eta_0 + \eta_1 z_i$ and variance σ^2. The density function for the logistic distribution is

$$f(\epsilon) = \frac{\exp(-\epsilon)}{\{1+\exp(-\epsilon)\}^2}.$$

As before, we assume for now that we are working with one binary covariate. We can now construct the likelihood as before using the full joint distribution

$$P(M = m, T' = t', Z = z) = \frac{\exp\{-(m-\alpha_0-\alpha_1 z-\alpha_2 t')\}\exp\{-(t'-\eta_0-\eta_1 z)/\sigma\}\,\theta_z}{\sigma\left(1+\exp\{-(m-\alpha_0-\alpha_1 z-\alpha-2t')\}\right)^2\left(1+\exp\{-(t'-\eta_0-\eta_1 z)/\sigma\}\right)^2}.$$

The formulation of the likelihood can continue in a similar fashion to that shown in Section 5.2.1. We can split the data into four groups based on their censoring and missing data indicators and calculate their individual contributions to the likelihood within these groups. The full log-likelihood is then the sum of the natural logarithms of the individual contributions.

$$
\begin{aligned}
&l(\eta,\sigma,\alpha,\theta|t_i',m_i,z_i) = \\
&\sum_{\substack{(i:m_i<0)\\ \delta_i=1}} \log\left[\frac{\theta_{z_i}\exp\left\{-\left(t_i'-\eta_0-\eta_1 z_i\right)/\sigma\right\}}{\left[1+\exp\left\{-\left(t_i'-\eta_0-\eta_1 z_i\right)/\sigma\right\}\right]^2}\left\{\frac{1}{1+\exp\left(\alpha_0+\alpha_1 z_i+\alpha_2 t_i'\right)}\right\}\right] \\
&+\sum_{\substack{(i:m_i<0)\\ \delta_i=0}} \log\left[\int_{t'}^{\infty}\frac{\theta_{z_i}\exp\left\{-\left(u_i-\eta_0-\eta_1 z_i\right)/\sigma\right\}}{\left[1+\exp\left\{-\left(u_i-\eta_0-\eta_1 z_i\right)/\sigma\right\}\right]^2}\left\{\frac{1}{1+\exp\left(\alpha_0+\alpha_1 z_i+\alpha_2 u_i'\right)}\right\}\mathrm{du}\right] \\
&+\sum_{\substack{(i:m_i>0)\\ \delta_i=1}} \log\left[\sum_{z=0}^{1}\frac{\theta_z\exp\left\{-\left(t_i'-\eta_0-\eta_1 z\right)/\sigma\right\}}{\left[1+\exp\left\{-\left(t_i'-\eta_0-\eta_1 z\right)/\sigma\right\}\right]^2}\left\{\frac{\exp\left\{\alpha_0+\alpha_1+\alpha_2 t_i'\right\}}{1+\exp\left(\alpha_0+\alpha_1+\alpha_2 t_i'\right)}\right\}\right] \\
&+\sum_{\substack{(i:m_i>0)\\ \delta_i=0}} \log\left[\sum_{z=0}^{1}\int_{t'}^{\infty}\frac{\theta_z\exp\left\{-\left(u_i-\eta_0-\eta_1 z\right)/\sigma\right\}}{\left[1+\exp\left\{-\left(u_i-\eta_0-\eta_1 z\right)/\sigma\right\}\right]^2}\left\{\frac{\exp\left(\alpha_0+\alpha_1 z+\alpha_2 u_i\right)}{1+\exp\left(\alpha_0+\alpha_1 z+\alpha_2 u_i\right)}\right\}\mathrm{du}\right].
\end{aligned}
$$

5.3.2 The Weibull and exponential distributions

Another distribution used commonly in survival analysis is the Weibull distribution (and its restricted form, the exponential distribution). To use this distribution the survival model errors must follow a Gumbel distribution (See Collett (1999) for details) This is a type of extreme value distribution and has the density function $f(\epsilon) = \exp\{\epsilon - e^{\epsilon}\}$.

The joint distribution for the survival times and latent missing data variable is therefore...

$$
\begin{aligned}
&P(M=m,T'=t',Z=z) = \\
&\frac{1}{\sigma}\exp\left\{m-\alpha_0-\alpha_1 z-\alpha_2 t'+\frac{t'-\eta_0-\eta_1 z}{\sigma}-e^{m-\alpha_0-\alpha_1-\alpha_2 t'}-e^{\frac{t'-\eta_0-\eta_1 z}{\sigma}}\right\}.
\end{aligned}
$$

Using the same methodology as previously we can calculate the full log-

likelihood...

$$
\begin{aligned}
&l(\eta,\sigma,\alpha,\theta|t_i',m_i,z_i) = \\
&\sum_{\substack{(i:m_i<0)\\ \delta_i=1}} \log\left[\frac{\theta_{z_i}}{\sigma}\exp\left(\frac{t_i'-\eta_0-\eta_1 z_i}{\sigma}-e^{\frac{t_i'-\eta_0-\eta_1 z_i}{\sigma}}\right)\left\{1-\exp\left(-e^{-\alpha_0-\alpha_1 z_i-\alpha_2 t_i'}\right)\right\}\right] \\
&+\sum_{\substack{(i:m_i<0)\\ \delta_i=0}} \log\left[\int_{t_i'}^{\infty}\frac{\theta_{z_i}}{\sigma}\exp\left(\frac{u_i-\eta_0-\eta_1 z_i}{\sigma}-e^{\frac{u_i-\eta_0-\eta_1 z_i}{\sigma}}\right)\left\{1-\exp\left(-e^{-\alpha_0-\alpha_1 z_i-\alpha_2 u}\right)\right\}\,\mathrm{du}\right] \\
&+\sum_{\substack{(i:m_i>0)\\ \delta_i=1}} \log\left\{\sum_{z=0}^{1}\frac{\theta_z}{\sigma}\exp\left(\frac{t_i'-\eta_0-\eta_1 z}{\sigma}-e^{\frac{t_i'-\eta_0-\eta_1 z}{\sigma}}-e^{-\alpha_0-\alpha_1 z-\alpha_2 t_i'}\right)\right\} \\
&+\sum_{\substack{(i:m_i>0)\\ \delta_i=0}} \log\left\{\sum_{z=0}^{1}\int_{t'}^{\infty}\frac{\theta_z}{\sigma}\exp\left(\frac{u_i-\eta_0-\eta_1 z}{\sigma}-e^{\frac{u_i-\eta_0-\eta_1 z}{\sigma}}-e^{-\alpha_0-\alpha_1 z-\alpha_2 u_i}\right)\,\mathrm{du}\right\}.
\end{aligned}
$$

The exponential is a specific case of the Weibull distribution. It occurs when $\sigma = 1$. This means that its hazard function is constant and does not depend on time. The log-likelihood can be easily derived from the Weibull model log-likelihood.

We can, therefore, consider a variety of survival distributions and whilst details are given here for only three types of distribution we are not restricted to just these. However, problems arise in calculating the likelihood. The main problem occurs in the numerical integration as discussed previously. Perhaps allowing the distribution of the missing data mechanism latent variable to differ from the survival model would mean that we could find analytic forms of the integrals, although nothing became apparent during the model development.

5.4 Identifiability

We must check the identifiability of the model parameters. We present here the original log-normal model and the log-logistic and Weibull models can

be checked in the same manner. We could check identifiability by considering the log-likelihood and counting the number of sufficient statistics. We do not consider the effect of censoring on the likelihood as this does not change the number of sufficient statistics. Define W_i to be the indicator variable $\mathrm{I}_{(z\ \mathrm{obs})}$. In this case, the log-normal log-likelihood can be written as

$$
\begin{aligned}
l(\eta,\sigma,\alpha,\theta|t_i',z_i) = \sum_{i=1}^{n}\Bigg\{ W_i\Bigg[&-\log\sigma - \frac{1}{2\sigma^2}(t_i'-\eta_0-\eta_1 z_i)^2 + z_i\log\theta + \\
&(1-z_i)\log(1-\theta) + \log\left(\Phi(-\alpha_0-\alpha_1 z_i-\alpha_2 t_i')\right)\Bigg] + \\
&(1-W_i)\Bigg[\log\Big(\exp\{-\frac{1}{2\sigma^2}(t_i'-\alpha_0)^2\}\Phi(-\alpha_0-\alpha_2 t_i')(1-\theta) + \\
&\exp\{-\frac{1}{2\sigma^2}(t_i'-\alpha_0-\alpha_1)^2\}\Phi(-\alpha_0-\alpha_1-\alpha_2 t_i')\theta\Big)\Bigg]\Bigg\} \\
= &-\log\sigma\sum_{i=1}^{n}W_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}W_i t_i'^2 + \frac{1}{\sigma^2}\eta_0\sum_{i=1}^{n}W_i t_i' + \\
&\frac{1}{\sigma^2}\eta_1\sum_{i=1}^{n}W_i t_i' z_i - \frac{1}{2\sigma^2}\eta_0^2\sum_{i=1}^{n}W_i - \frac{1}{\sigma^2}\eta_0\eta_1\sum_{i=1}^{n}W_i z_i - \\
&\frac{1}{2\sigma^2}\eta_1^2\sum_{i=1}^{n}W_i z_i^2 + \log\theta\sum_{i=1}^{n}W_i z_i + \log(1-\theta)\sum_{i=1}^{n}W_i - \\
&\log(1-\theta)\sum_{i=1}^{n}W_i z_i + \sum_{i=1}^{n}W_i\log\Phi(-\alpha_0-\alpha_1 z_i-\alpha_2 t_i') + \\
&(1-W_i)\Bigg[\log\Big(\exp\{-\frac{1}{2\sigma^2}(t_i'-\alpha_0)^2\}\Phi(-\alpha_0-\alpha_2 t_i')(1-\theta) + \\
&\exp\{-\frac{1}{2\sigma^2}(t_i'-\alpha_0-\alpha_1)^2\}\Phi(-\alpha_0-\alpha_1-\alpha_2 t_i')\theta\Big)\Bigg]
\end{aligned}
\tag{5.7}
$$

However, we see that we cannot check the identifiability of the model in the log-normal setting as we cannot identify all the sufficient statistics here. We instead satisfy the identifiability issues by testing our model.

5.5 Application to the Cerebral Palsy Data

We can now apply this model to the adult cohort and compare results to the MAR and MCAR estimates of the previous Chapter. This was programmed in S-Plus as before and can be found in Appendix C. We again model the effect of each severe level disability upon survival. Multivariate extension will be discussed in Chapter 7.

5.5.1 The Adult Cohort

We now present the parameter estimates for the joint survival and missing data mechanism model based upon the adult cohort and compare this to the complete case and MAR likelihood methods as discussed in Chapter 4. These are presented in Table 5.1. Comparison of the maximised log-likelihoods again suggests that the Weibull distribution is most appropriate. Note that we now consider these using the log-linear parametrization but it is easy to switch between the two using the following equalities:

$$\begin{aligned} \gamma &= \frac{1}{\sigma}, \\ \lambda_0 &= \exp\left(-\frac{\eta_0}{\sigma}\right) \text{ and,} \\ \lambda_1 &= \exp\left(-\frac{\eta_0 + \eta_1}{\sigma}\right). \end{aligned}$$

Table 5.1 shows that parameter estimates for the Weibull models over each of the missing data mechanisms are quite similar. Although we do appear to see a reduction in the effect of the MAR assumption. The most apparent difference occurs with the effect of IQ. The estimated survival curves by IQ are shown in Figure 5.1. Recall that IQ had the highest level of missing data. We see that the survival curves are still very similar. It seems that

the naive MCAR analysis works well in this case although we must consider the appropriateness of the linearity of the missing data mechanism before concluding this.

		η_0	η_1	σ
Ambulation	Available case	4.6808	-0.9759	0.7658
		(0.1102)	(0.2601)	(0.0850)
	MAR Likelihood	4.7055	-1.0119	0.8000
		(0.0990)	(0.2010)	(0.0695)
	NMAR model	4.6875	-0.9758	0.7993
		(0.1386)	(0.2724)	(0.0747)
Manual Dexterity	Available case	4.6502	-0.8044	0.7925
		(0.1500)	(0.2500)	(0.0823)
	MAR Likelihood	4.6523	-0.8070	0.8023
		(0.0988)	(0.1300)	(0.0701)
	NMAR model	4.6509	-0.8005	0.8043
		(0.1359)	(0.3814)	(0.0757)
Vision	Available case	4.6402	-0.4446	0.7826
		(0.1473)	(0.1073)	(0.0646)
	MAR Likelihood	4.6256	-0.4529	0.7983
		(0.0912)	(0.1572)	(0.0589)
	NMAR model	4.6306	-0.4164	0.8032
		(0.1341)	(0.1750)	(0.0760)
IQ	Available case	4.7307	-0.6726	0.7013
		(0.1502)	(0.1745)	(0.0624)
	MAR Likelihood	4.8119	-0.7607	0.8051
		(0.1401)	(0.1593)	(0.0431)
	NMAR model	4.7003	-0.5835	0.7928
		(0.1406)	(0.2030)	(0.0717)

Table 5.1: Comparison of complete case, and MAR and NMAR likelihood based survival parameters (s.e.) for univariate disabilities in the adult cohort

In Table 5.2 we see the parameter estimates from the missing data mechanism and covariate models (see Equation 5.5). We see that severe IQ affects the mechanism slightly differently to the three physical disabilities and we

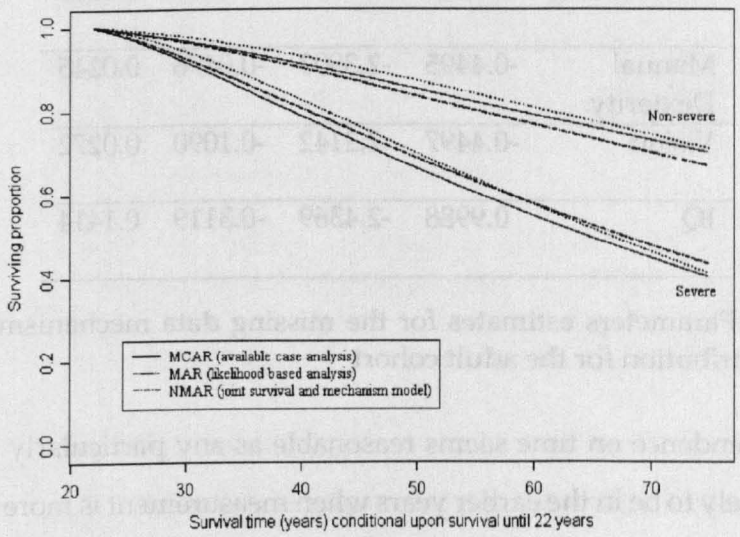


Figure 5.1: Survival by severity of IQ for the incident cohort under different missing data assumptions

also estimate a higher proportion of severe (as observed in the complete data). In particular, the survival time seems to have a greater effect on the missingness of IQ than the other covariates.

Sensitivity Analysis

So far we have considered a linear missing data mechanism. However, we need to consider the appropriateness of this model. With the adult cohort we are conditioning upon survival of 22 years so we need to think about possible mechanisms that may act upon the data after this time. We are working only with binary covariates at present so we focus upon the dependence on time. As we are conditioning on reaching adulthood, the individuals in our cohort will have reached physical maturity. Therefore, the

	α_0	α_1	α_2	θ
Ambulation	-0.2060	-2.6625	-0.1050	0.0489
Manual Dexterity	-0.4495	-2.3903	-0.0476	0.0245
Vision	-0.4497	-2.2142	-0.1090	0.0272
IQ	0.9988	-2.4369	-0.3119	0.1414

Table 5.2: Parameters estimates for the missing data mechanism and covariate distribution for the adult cohort

linear dependence on time seems reasonable as any particularly different effect is likely to be in the earlier years when measurement is more difficult.

Breaking away slightly from our model from it may be interesting to use year of entry or age at entry in the missing data mechanism model but we have not presented this here.

We can now consider the incident cohort and see if the same issues arise. We believe that we are more likely to have complex missing data mechanisms in this cohort as we are picking up individuals with shorter lifetimes. Very short times are likely to have a greater effect on missingness. Not only is there less time to collect the information but failure might occur before the children have fully developed so levels of disability may be impossible to ascertain.

5.5.2 The Incident Cohort

We repeat the analysis for the incident cohort except now we need to allow for the left-truncation of survival times. First we see that the Weibull is now no longer always the optimal model choice. Note that as each model estimates uses the same number of parameters we can compare the models

directly using the maximised log-likelihood.

	Model distribution		
Disability	Log-logistic	Weibull	Log-normal
Ambulation	-605.68	-605.47	-609.80
Manual dext.	-584.07	-584.93	-588.03
Vision	-520.08	-523.29	-523.30
IQ	-700.48	-699.30	-704.34

Table 5.3: Maximum log-likelihood values for univariate accelerated failure models over different distributions under the NMAR assumption

Maximised log-likelihoods for log-normal, log-logistic, and Weibull univariate NMAR models are shown in Table 5.3. We can see that the choice between the log-logistic and Weibull models is less clear. In the models for manual dexterity and vision, the optimal model chosen by finding the maximum log-likelihood is the log-logistic model. The Weibull distribution was chosen using this criteria in the MAR and MCAR analyses. We know that both distributions can be defined as special cases of the Burr XII distribution but as both models seem to fit well, and we had problems with optimizing the models with this distribution previously, we suspect that the Burr XII parameters would be unstable. We can consider the estimated survival model scale and shape parameters to investigate why there seems to be little to distinguish between the two models.

The hazard functions for the Weibull and log-logistic models respectively are

$$h_W(x|\lambda,\gamma) = \lambda\gamma x^{\gamma-1} \quad \text{and} \quad h_L(x|\theta,\xi) = \frac{\theta\xi x^{\xi-1}}{1+\theta x^{\xi}}.$$

When the scale is < 1 the hazards have similar forms: they are both mono-

tonically decreasing. We might expect to see this now that we are considering survival from diagnosis age. For scale > 1 the Weibull has a monotonically increasing hazard (constant when the scale is unity) and the log-logistic has a single early peak followed by a slower decline. Recall that for the adult cohort the scale parameter was consistently greater than 1 implying an increasing hazard under the Weibull model (see Table 5.1).

Table 5.4 presents 90 and 75% survival (in years) conditional upon survival until 2 years for the ambulation and IQ models using the optimum models previously identified in Table 5.3 (i.e. Weibull models for the effect of severe ambulation and IQ and log-logistic models for the effect of manual dexterity and vision). We see that the linear NMAR model leads to a consistent decrease in estimated survival for those with non-severe level disabilities. We are now estimating 75% survival to 46.8 and 51.8 years for individuals with non-severe ambulation and IQ respectively.

		90% survival		75% survival	
		Non-sev	Severe	Non-sev	Severe
Ambulation	AC	24.7	7.7	55.4	15.3
	MAR	19.8	5.4	50.0	11.2
	NMAR	18.6	5.9	46.8	12.5
IQ	AC	31.1	10.3	68.9	21.0
	MAR	24.4	7.2	62.3	16.1
	NMAR	20.6	7.8	51.8	17.5

Table 5.4: Comparison of available case, MAR likelihood, and NMAR based analyses for ambulation and IQ in the incident cohort - 90% and 75% survival (age in years).

Survival in the manual dexterity and vision models is discussed in the following section where we also look the sensitivity to the linear constraint on the missing data mechanism model.

Sensitivity Analysis

We can look at the actual estimates for the missing data mechanism model in Table 5.5. As discussed we should not place too much importance on these values but we can use them to consider the type of mechanisms that might be working. We can see that each model estimates a decrease in the probability of missingness with an increase in survival time as we previously suspected. However, the structure of the mechanism then differs over the different covariates. This may suggest that the mechanisms are not the same over the different disabilities but we have only used a simple model and so cannot place a great deal of weight on any conclusion. If we compare the estimates to those in Table 5.2 we see a much greater dependence on the true covariate value suggesting that the incident cohort is further from the MAR assumption than the adult cohort.

	$\hat{\alpha_0}$	$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\theta}$
Ambulation	0.58	-1.06	-0.27	0.08
Manual dexterity	-1.44	2.04	-0.12	0.11
Vision	-1.69	2.46	-0.24	0.09
IQ	1.56	-1.79	-0.40	0.18

Table 5.5: Parameter estimates from the missing data mechanism and covariate model for the incident cohort

We need to consider the sensitivity of our conclusions to diversions from this linear assumption. Unlike with the adult cohort we have reasons to suspect this assumption. We are now considering children who may not reached physical maturity. Therefore, assessing the level of disability will be harder. This will lead to a higher probability of missingness before approximately 10 years. We are working only with binary covariates at present so we focus upon the dependence on time. Therefore, we

consider a missing data mechanism of the form:

$$m_i = \alpha_0 + \alpha_1 z_i + \alpha_2 t_i' + \alpha_3 exp(-t_i') + \omega_i, \qquad \omega_i \sim N(0,1). \tag{5.8}$$

Note the inclusion of the exponential term with parameter α_3. We start by setting $\alpha_2 = 0$ i.e. having an exponential effect of t_i' on missingness. This model is again identifiable and we can compare survival model estimates over mechanisms.

		η_0	η_1	σ
Ambulation	Available case	5.034	-1.388	0.848
	MAR Likelihood	5.098	-1.596	0.991
	NMAR linear model	5.030	-1.448	0.986
	NMAR exponential model	5.184	-1.633	0.984
Manual Dexterity	Available case	4.655	-1.323	0.604
	MAR Likelihood	4.683	-1.546	0.699
	NMAR linear model	4.811	-2.329	0.805
	NMAR exponential model	4.834	-2.297	0.813
Vision	Available case	4.648	-1.447	0.623
	MAR Likelihood	4.657	-1.616	0.699
	NMAR linear model	4.765	-2.606	0.815
	NMAR exponential model	4.795	-2.522	0.829
IQ	Available case	5.235	-1.257	0.829
	MAR Likelihood	5.330	-1.450	0.987
	NMAR linear model	5.127	-1.168	0.978
	NMAR exponential model	5.419	-1.479	0.987

Table 5.6: Comparison of complete case, and MAR and NMAR (linear and exponential) likelihood based survival analyses for univariate disabilities in the incident cohort

Table 5.6 presents the comparison of survival model parameters over MCAR, MAR, and two NMAR mechanisms. Let us consider the comparison for each univariate model. Note, that we are comparing Weibull models for the effect of severe ambulation and IQ and log-logistic models for the

		90% survival		75% survival	
		Non-sev	Severe	Non-sev	Severe
Manual dexterity	AC	29.8	9.4	56.1	16.4
	MAR	25.3	7.0	52.2	12.7
	NMAR (linear)	23.0	4.0	52.7	6.9
	NMAR (exponential)	23.1	4.1	53.5	7.2
Vision	AC	28.6	8.2	54.6	14.4
	MAR	24.7	6.5	50.9	11.7
	NMAR (linear)	21.6	3.4	49.9	5.5
	NMAR (exponential)	21.6	3.6	50.6	5.9

Table 5.7: Comparison of available case, MAR likelihood, and NMAR based analyses for manual dexterity and vision in the incident cohort - 90% and 75% survival.

effect of severe manual dexterity and vision. There seems to be a difference according to this model choice, with a more obvious difference in η_0 and η_1 between the MCAR and MAR mechanisms and the two NMAR mechanisms in the log-logistic models. This may be because under the more restrictive mechanisms the favoured model was the Weibull distribution so we are actually comparing different models. However, the estimated parameters are also closer over the two NMAR mechanisms (the linear and exponential models, Equation 5.8 with $\alpha_3 = 0$ and $\alpha_2 = 0$ respectively) suggesting that the main effect on the probability of missing data comes form the true severity level. This leads to an increase in the magnitude of the severe level effect in the survival model and a smaller increase in the baseline survival. These estimates were presented in Table 5.6.

It is also interesting to note that for the ambulation and IQ models, where the optimal distribution is the Weibull, we estimate near exponential models ($\sigma \approx 1$).

Table 5.7 presents estimated survival for the manual dexterity and vi-

sion models. The estimated survival curve is considerably lower for those with a severe impairment under the NMAR mechanisms. In particular the 75% survival rates drop from approximately 13 to 7 years for those with severe manual dexterity and from 12 to 6 years for those with severe vision impairment. From this we can see that a naive MCAR or MAR model vastly overestimates survival for those at severe levels. The estimated survival curve remains similar for those with non-severe levels of disability in the sense that the absolute change is approximately the same but the proportional change is less. If we compare estimates under NMAR to those under the optimal MCAR and MAR Weibull models we see similar patterns although the magnitude of differences changes.

Conclusions are not quite as clear for the ambulation and IQ models. We see that there is no clear trend in parameter estimates over the different mechanisms and there are differences between the estimates using the linear and exponential missing data mechanisms. Therefore, we consider including both terms in our mechanism. We cannot identify each parameter in this model (see Section 5.4) therefore we must consider a range of suitable values for α_3.

Figure 5.2 shows the estimated probability of missing data for the ambulation covariate for the linear missing data mechanism already fitted in Section 5.5.2 and the missing data mechanism using the same maximum likelihood estimates for α_0, α_1, and α_2 but with the addition of a $10\exp(-t_i')$ term. This results in a higher probability of missing data at low survival times. This is quite an extreme mechanism given our understanding of the data, which we discussed at the start of this section, so we look at values $0 < \alpha_3 \leq 10$. We can now fit this model, which is still identifiable, and compare the survival estimates to those of the previous no interaction model.

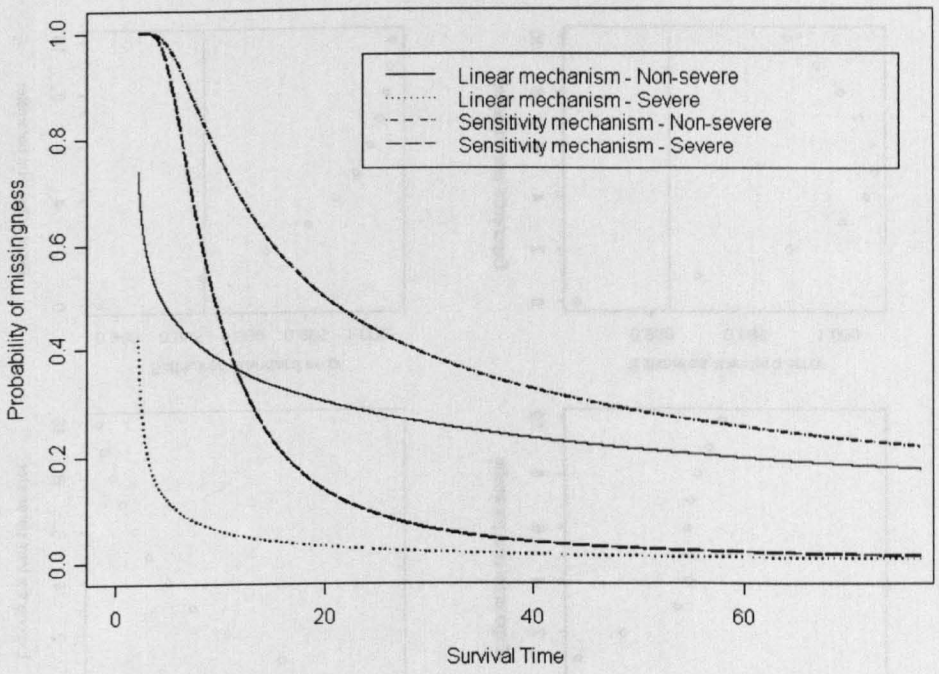


Figure 5.2: Probability of missing data for linear and exponential mechanisms for the effect of severe ambulation with a Weibull model by survival (age in years).

Figure 5.3 shows results of the sensitivity analysis. We see that for the chosen range of α_3 the survival model estimates are similar in size to the estimates under the linear model (i.e. $\alpha_3 - 0$). Given that we believe we have considered a range of missing data mechanisms that contains a large proportion of mechanisms that fit our prior beliefs we might be satisfied that a linear missing data model is adequate here also. We can again look at estimated survival times. The 75% survival times for those with non-severe ambulation range from 46.8 to 47.2 years over the sensitivity analysis

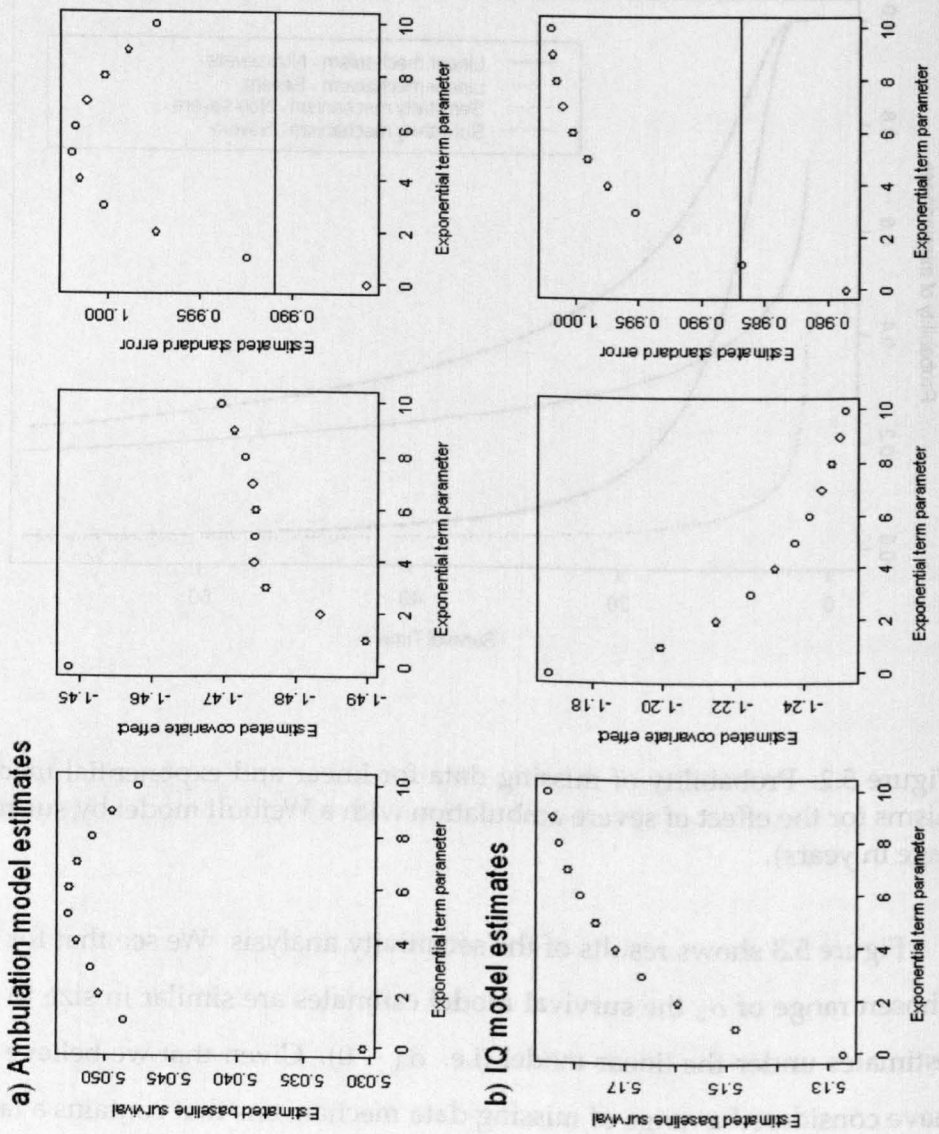


Figure 5.3: Survival model estimates for the effect of a) ambulation and b) IQ by fixed exponential parameter in the mechanism sensitivity analysis (- = MAR estimate)

and from 12.1 to 14.2 for those with severe ambulation. The corresponding ranges for those without and with severe IQ levels are 51.8 to 59.5 and 16.6 to 19.9 years. Recall from Table 5.4 that the linear NMAR estimates were 46.8 and 12.5 years for the ambulation model and 51.8 and 17.5 years for the IQ model. The MAR estimates were 50.0 and 11.2 years for ambulation and 62.3 and 16.1 years for IQ. Therefore, survival for those with non-severe disabilities seems slightly decreased under the NMAR model over plausible missing data models. Survival with severe disabilities seems possibly higher than thought under the MAR assumption, yet less than that from a MCAR analysis. However, the key to analysis of this kind is to realise that we are not looking for point estimates but rather wishing to estimate a range of values in which they may lie based upon our understanding of the data and collection method.

5.6 Discussion and Conclusions

We have now attempted to model the missing data mechanism. We introduced a univariate joint survival and missing data model following the selection model ideas of Heckman (1974). We showed how to construct the likelihood function under log-normal, log-logistic and Weibull distributional assumptions and also with left-truncated survival times. This model could then be used to model our cerebral palsy data described in Chapter 4 under the NMAR assumption.

Firstly, we considered the adult cohort consisting of individuals who survived at least 22 years. This removed the need to consider left-truncation as all participants are first observed before this time. We saw that the linear NMAR model resulted in similar estimated survival for physical disabili-

ties suggesting MCAR or MAR missingness is appropriate. We discussed the appropriateness of the linear missing data mechanism and decided that we had no obvious reasons to consider other models as this fitted with our understanding of the data collection method.

We then considered the incident cohort. This includes all individuals from the large data set with recorded entry time and survival greater than 2 years. We now saw a change in estimated survival over the different missing data mechanisms. Under the linear NMAR model we estimated a decrease in estimated survival time particularly for those with a severe level of disability. However, we considered a sensitivity analysis for the missing data mechanism based on the inclusion of an exponential of time term. As we were only considering a binary covariate this was the obvious change to make. This would be complicated for factors with more levels or continuous covariates. This sensitivity analysis was designed based upon our knowledge of the data collection method and our beliefs concerning the missing data mechanism and should not be universally applied in other analyses where these factors differ. This sensitivity analysis showed a greater dependence of the missing data mechanism on survival time. The main conclusion drawn here was the obvious bias in the naive available case analysis which overestimated survival for all individuals regardless of level of impairment. It was more complicated to draw conclusions from the NMAR analysis concerning estimated survival but this can not be the aim of an analysis such as this. Our interest must lie in finding a plausible range of estimates given our beliefs and understandings about the data. Given that we are modeling using untestable assumptions we need to acknowledge this.

We have had to make certain assumptions about the data itself and its

collection method to use this model. These focus around the disability covariates. Firstly, we are assuming that there was adequate attempt to collect data over the study period and that if a child was simply too young at referral to obtain the information they were reassessed at later visits. This is important as we are not conditioning the missing data mechanism on entry time. However, this also requires that the level of disability does not change. CP is a non-degenerative condition so disability should not get worse until much older ages. Changes might arise in situations such as these due to a change in the testing procedures. Fortunately, the disability covariates are recorded on a very simple clear scale meaning they should be reasonably consistently estimated. We also assume that patterns of diagnosis and referral remain the same over the course of the study period. This is a more difficult assumption to make. If we refer back to Table 4.2 we see that the levels of disability remained reasonably constant over time although this is complicated by the missing data. This gives us some belief that children entering the study later are essentially the same as the children entering earlier.

There are obvious possible extensions to this model. Clearly we can extend to discrete covariates. Continuous covariates could also be included but their probability distribution may be harder to model. In Chapter 7 we will consider a multivariate model for our data by considering a model for all four binary disability covariates. However, first we present results from a simulation study investigating the accuracy of the model.

Chapter 6

Simulation Study

In Chapter 5 we presented a flexible joint univariate model for the survival time and missing data mechanism. We must consider its reliability. In this chapter we describe a simulation study to look at the ability of the model to accurately estimate survival parameters assuming it is the true model. We consider the log-normal, log-logistic, and Weibull distributions. Estimates will be compared to corresponding results from available case and MAR likelihood analyses using the methods discussed in Chapter 4 as well as estimates based upon the complete data. Simulations are conducted using different missing data mechanisms and also a change in the proportion of missing data. Otherwise, we consider data similar to the cerebral palsy data set that is our motivation i.e. we use a similar level of censoring and the joint model maximum likelihood estimates are used as the model for the simulated data.

Note that we are conducting this simulation study in order to investigate the reliability of our model to accurately estimate parameters assuming the model is correct. We are not looking at the robustness of our model

to misspecification. This would require an additional study that is beyond the scope of this thesis as this is a more complicated, although also interesting and important, issue.

6.1 Joint Model Simulation Study

We will now discuss the methods by which we simulate the missing data, the design of the study, and present the results

6.1.1 Generating Data

In order to simulate data, of size n, we have to consider how to draw from the various survival distributions, how to apply censoring, the distribution of the covariate, and the construction of the missing data mechanism.

Firstly, we create a vector of length n based on realisations from a Bernoulli distribution with probability Θ. This gives the true covariate values which we then subject to our required missingness mechanism.

We must then construct the true survival times based upon these simulated covariate values and the maximum likelihood survival estimates from the joint model for the adult cohort. Note that we are going to consider this cohort as it does not include the issue of left-truncation hence simplifying the study. Like censoring, truncation leads to a decrease in accuracy of model estimates. If the truncation is independent it can be incorporated in the likelihood as discussed and does not lead to bias. We can use established techniques to generate random numbers from a standard uniform distribution via the *runif* function in S-Plus. Effective random number generation has a vast literature but we are using it simply so that is not our concern here. To transform these to random numbers from the normal,

logistic, and Gumbel distributions we require the inverse probability transform.

Theorem *If* $F : \mathbb{R} \rightarrow [0,1]$*is increasing and left-continuous then we define its inverse as follows*

$$F^{-1}(u) = \inf(t : F(t) > u)$$

$\Rightarrow$ *A real valued random variable X with distribution function* $F(x) = P(X <= x)$ *can be represented using the inverse probability transform* $X = F^{-1}(U)$ *for U a uniform [0,1] random variable.*

We can therefore construct survival times based on errors drawn from the relevant distribution, the corresponding covariate value, and the required η and σ, the survival model parameters as presented in the previous chapter.

We impose a censoring distribution similar to that we believe applies to the Bristol CP data i.e. independent uniform censoring on the interval [23 years, 53 years]. This is because censoring is mainly due to the end of the study period and not due to individuals being lost during the study. Recall that as we are trying to simulate data similar to the adult cohort we are conditioning on survival until 22 years. The last entry into the study is approximately 45 prior to the final censoring date of 2005 and the highest survival time in the data is approximately 75 years. Therefore, observed additional survival must be less than 53 years. We can then calculate the observed survival and censoring indicator. The enforced censoring mechanism leads to approximately 80% censoring, similar to that found in the Bristol cerebral palsy data.

Again using the inverse probability transform we can construct the la-

tent missing data mechanism variable, m. We can then force the covariate data to be missing according to the value of this. This method produces the full simulated data set and was programmed in S-Plus.

6.1.2 Study design

The aim of our simulation study is to investigate the success of the model at correctly estimating the survival model for different data sets with a variety of missing data mechanisms. We look at look at data sets similar in structure to the adult sub-cohorts discussed in Section 4.2.3 as we wish to investigate the reliability of the estimates obtained in Chapter 5. In order to do this we simulate survival data with parameters $(\eta_0, \eta_1, \sigma, \theta) = (4.7, -0.7, 0.8, 0.05)$. Recall that the survival model is defined as

$$\log t_i = \eta_0 + \eta_1 z_i + \sigma \epsilon_i$$

and that θ defines the covariate model. Refer back to Table 5.1 to see that these are approximate averages of the estimated parameters for the survival in the adult cohort. We subject the simulations to four different missing data mechanisms. These mechanisms are defined by the value of (α_1, α_2) with α_0 chosen to result in approximately the right proportion of missing data. Recall that we modeled the missing data mechanism using the linear model,

$$m_i = \alpha_0 + \alpha_1 z_i + \alpha_2 \log t_i + \omega_i.$$

The four mechanism we consider are

a) MCAR $\alpha_1 = \alpha_2 = 0$, 20% missing data,

b) MAR $\alpha_1 = 0, \alpha_2 = -0.2$, 20% missing data,

c) NMAR $\alpha_1 = 1.5, \alpha_2 = -0.2$, 20% missing data,

d) NMAR $\alpha_1 = 1, \alpha_2 = 0$, 50% missing data.

We are forced to drop the dependence of the missing data mechanism on survival time in the model with 50% missing data as otherwise we are left with no severe level cases and we wish to maintain the high dependence on the true covariate value. For each survival distribution previously discussed and each mechanisms we then simulate 100 data sets using the method discussed in the previous section and compare survival model estimates from our model with those from available case and MAR likelihood estimates and also the true estimates (based upon the true data). Each data set consists of 400 individuals with one binary covariate. Full results are presented as box plots.

6.2 Simulation Study Results

Results are presented for the simulations based upon the Weibull distribution as this is the model chosen as fitting the data best most frequently. Estimates for log-logistic and log-normal models display similar distributions and summaries can be obtained from the author.

Figures 6.1 to 6.3 present simulated results for the estimates of the survival model intercept (η_0), the covariate effect (η_1), and the dispersion (σ) over the three mechanisms with 20% missing data. They compare the "True" estimates (i.e. estimates based on the complete data) with estimates from available case (AC), MAR likelihood based (MAR), and NMAR joint model (NMAR) analyses. We can also consider the effect of an increased proportion of missing data on the different approach estimates. Results for

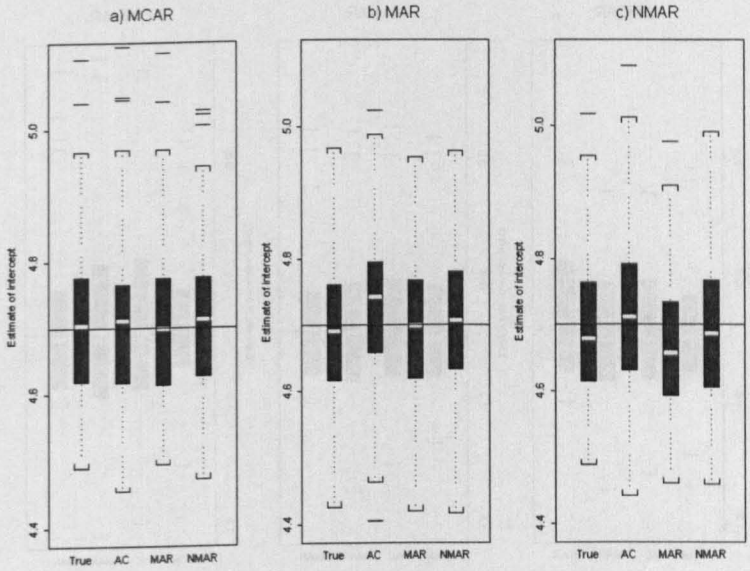


Figure 6.1: Simulation model estimates for survival model intercept under a) MCAR b) MAR c) 20 percent NMAR mechanisms

the missing data mechanism leading to approximately 50% missing data can be found in Figure 6.4.

6.3 Discussion of the Results

Studying Figures 6.1-6.3 we can discuss the reliability of our model over increasingly less restrictive missing data mechanisms. There are several things to note. Firstly, we consider our results when the data are MCAR. We can see that the distributions of all parameter estimates are similar for our joint model compared to that of the alternative methods and estimates assuming known data. This is encouraging as it suggests that modelling the missing data mechanism does not lead to less reliable results compared to the most simple missing data methods when it is actually unnecessary. We

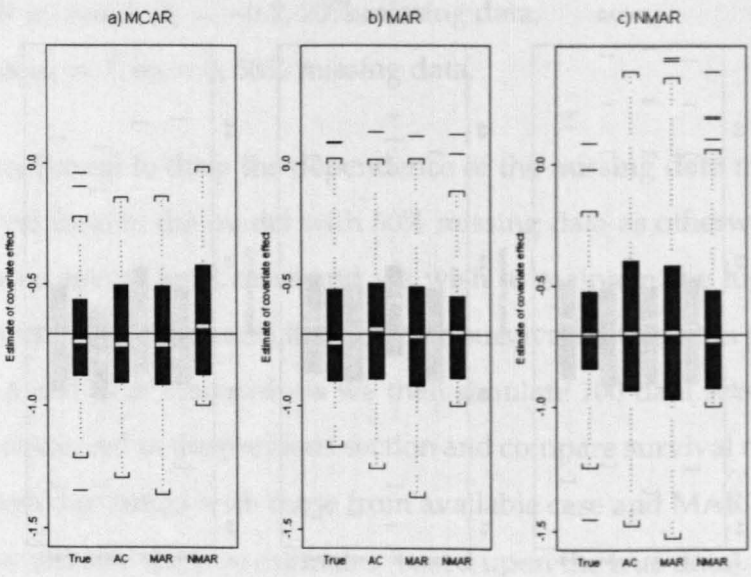


Figure 6.2: Simulation model estimates for survival model covariate effect under a) MCAR b) MAR c) 20 percent NMAR mechanisms

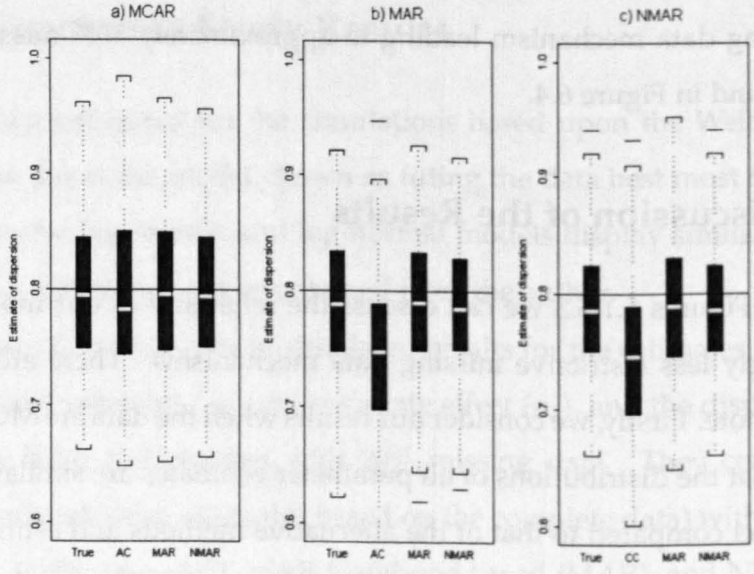


Figure 6.3: Simulation model estimates for survival model dispersion under a) MCAR b) MAR c) 20 percent NMAR mechanisms

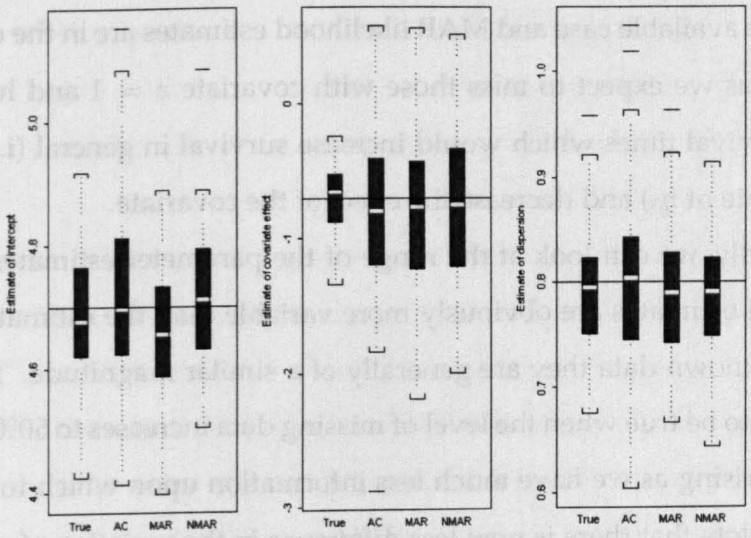


Figure 6.4: Simulation model estimates for survival model with 50 percent missing data

can then look at the distributions of results as the missing data mechanism tend to NMAR and the proportion of missing data increases. We can see that the available case estimates shift away from the true parameters under MAR conditions (in particular the σ and η_0 terms) and the likelihood based estimates shifts slightly under the NMAR mechanism (observe the bias in the estimate of η_0). However, the median of the joint model estimates remain consistently close to the "true" estimates. We know that available case analysis lead to bias in parameter estimates when data are not MCAR. This is why it is so important to be able to relax this assumption. In particular, when we have an NMAR mechanism the available case estimates are considerably biased but our model remains more reliable. This is particularly obvious in the estimation of the dispersion parameter, σ, where

available case analyses vastly underestimate the magnitude. The biases within the available case and MAR likelihood estimates are in the expected direction as we expect to miss those with covariate $z = 1$ and hence the lower survival times which would increase survival in general (i.e. lower the estimate of η_0) and decrease the effect of the covariate.

Secondly, we can look at the range of the parameter estimates. While the model estimates are obviously more variable than the estimates using the fully known data they are generally of a similar magnitude. This perhaps fails to be true when the level of missing data increases to 50% but this is unsurprising as we have much less information upon which to base estimates. Note that there is now less difference in the variation of estimates form the MAR and NMAR models. This is most probably caused by the fact that the missing data mechanism used to generate this data is closer to a MAR mechanism. In situations such as this it does not seem to be a sensible idea to try any analysis as we have so little data. The range may decrease when there is no or less censoring, recall we have 80% censoring, as censoring does lead to a decrease in precision of parameter estimates even with standard survival analysis models in data with no missing values. Changes would also be likely under varying values of θ, the probability of a "severe" covariate level. The low values that we are simulating with lead to small number of severe individuals, as observed in the cerebral palsy data.

Thirdly, if we consider Figure 6.4, results for the NMAR mechanism resulting in 50% missing data, we see considerably larger variance in estimates, particularly the covariate effect parameter η_1. It should be noted that Figure 6.4 differs in content to the preceding three figures as we are now looking at each of the survival model parameters at once. They are

presented separately as we are now considering a different proportion of missing data. This larger variance may potentially be masking biases in simpler models, the result of dropping the dependence of missingness on survival time, or the fact that there is simply not enough data for the joint model to be able to extract any more information than the simplest available case and MAR models.

It is also useful to note that our model works equally over all the survival distributions, although full results are not displayed here. These results suggest that we may be reasonably confident in the precision of our model. However, we have already discussed the dependence of the model on the assumption of the linear missing data mechanism. In the analysis of the two cerebral palsy sub-cohorts we conducted a sensitivity analysis to the linearity assumption.

It would also be interesting to extend the study to consider the accuracy of standard errors and confidence intervals. These are as important in analysis as the actual point estimate. However, time constraints meant that this was not possible but might be considered as future work for investigation.

Chapter 7

Multivariate Analysis

In Chapter 5 we have considered a joint univariate selection model for the survival time and the missing data mechanism. We might use this to study the possible effect of each of the four disabilities upon survival. We have seen that severe disability causes a significant decrease in the estimated survival time. We can now consider multivariate extensions to this idea. These are of particular interest for our incident cohort as our previous analysis suggested that data were not MAR so standard multivariate survival models are not appropriate. We wish to investigate the multivariate model for the combined effect of severe ambulation, manual dexterity, vision, and IQ upon survival. Models of this type help us to investigate the relative impact of covariates as predictors of survival and also the dependence structure of the dependent and independent covariates. There is a definite correlation between the severity of the disabilities and a multivariate model helps us to identify this structure.

Here we discuss the possible structures for this multivariate model. As with the univariate case we must consider the form of the covariate model,

the survival model, and the missing data mechanism.

7.1 The Multivariate Model

Recall that in the univariate case, as stated in Section 5.2.1, the joint likelihood for our latent variable model can be constructed as

$$\begin{aligned} L(\eta,\sigma,\alpha,\theta|t,z,\delta) = &\prod_{F,O} f(T'=t_i', M<0, Z=\underline{z}_i|\alpha,\eta,\sigma,\theta) \times \\ &\prod_{C,O} S(T'=t_i', M<0, Z=\underline{z}_i|\alpha,\eta,\sigma,\theta) \times \\ &\prod_{F,M} f(T'=t_i', M>0|\alpha,\eta,\sigma,\theta) \times \\ &\prod_{O,M} S(T'=t_i', M>0|\alpha,\eta,\sigma,\theta) \end{aligned}$$

where T' is the log of the observed survival time, M is a latent variable controlling the missing data mechanism, and Z is the vector of covariates.

In turn we can express the joint density function in terms of the product of conditional densities:

$$\begin{aligned} &f(T'=t_i', M<0, Z=\underline{z}_i) = \\ &\qquad f(M=m|T'=t',Z=\underline{z})f(T'=t'|Z=\underline{z})f(Z=\underline{z}). \end{aligned}$$

However, now we have the issue that some data for an individual may be observed and some missing. Therefore the likelihood must be further divided to allow for the different patterns of missingness. Similarly, we will have to further separate the distribution of the missing data mechanism in the joint density function.

7.1.1 The Covariate Model

In the early model, where we considered one binary covariate, a simple model was adequate for Z. However, as Z is now a vector we will have to describe a model for the distribution of the cell probabilities for the contingency table constructed via the components of Z. Using the same notation as in the missing at random model of Section 4.6, assume that Z is a vector of p factor variables. We can then construct a contingency table, based upon the covariates, of dimension $I_1 \times ... \times I_p$ where I_j is the number of levels of the jth covariate. We then place a model on the probability distribution of the table cells. We can consider a fully saturated model or a restricted model. Investigation of the observed data might suggest possible simplified models.

Our main interest does not lie with the distribution of the covariates and therefore the estimation of these model parameters is a nuisance. Using a restricted model decreases the number of degrees of freedom required to fit it although the effect of using an unsaturated model should be considered. We have not looked at the sensitivity of our model to the form of the covariate distribution. While this was not a particular issue in the single binary variable case it becomes more important in a multivariate or continuous setting.

Note that in a multivariate setting we may have some observed information for an individual but not all. We will need to model the unobserved conditional upon the observed data using Bayes theorem.

7.1.2 The Survival Model

We also need to define the model used for the survival time. This is of the same form as before, except we now have a vector of covariates and hence a vector of covariate effect parameters. Our interest in the cerebral palsy data lies with the effect of the four binary disability covariates; ambulation, manual dexterity, vision, and IQ. Previous work (Hutton & Pharoah 2002) has considered the effect of the number of severe level disabilities upon survival. As was shown in Table 4.3 data are exceptionally sparse if we considered this parametrization. Our univariate (binary) simulations showed that with high levels of missing data estimates became less reliable and this would not be helped by the increased number of levels if we looked at the number of disabilities. We could consider constructing a model here where we considered the variable to be coarsened as opposed to entirely missing i.e. if we observed severe ambulation but non-severe manual dexterity and IQ then if the data for vision was missing we would know that the number of severe level disabilities could only equal one or two dependent upon the true level of sight. However, this is beyond the scope of this thesis.

We again use a log-linear construction for the survival model:

$$\log t_i = t_i' = \eta_0 + \eta_1^T \underline{z}_i + \sigma\epsilon_i, \qquad \epsilon_i \sim N(0,1), \qquad i = 1, ..., n.$$

Note that η_1 is now a vector of parameters of length equal to the covariate vector $\underline{z}$. We may now have any number of data values missing for each individual.

Our main purpose in this chapter is to show how the univariate joint model may be extended to multivariate settings and to discuss the complications in doing so. Therefore, the analysis here should be considered as

an example of what can be done and not the only choice. We could also consider interaction terms or, if using continuous covariates, exponential or polynomial terms.

7.1.3 The Missing Data Mechanism

As we are now in a multivariate situation our missing data mechanism must be able to allow for any pattern of missingness. Therefore, we employ a vector of length equal to the number of survival model covariates consisting of latent variables each of which works in the same fashion as in the univariate case. Again, we need to consider the data collection method in order to decide upon a sensible model for the missing data mechanism. It does not seem necessary, in our data, to allow the probability of missing a disability observation to depend upon the severity of the other disabilities. Therefore, as before, each missing data latent variable will be modelled using the corresponding true individual covariate value and the log survival time. This means that

$$m_j = \alpha_{j,0} + \alpha_{j,1} z_j + \alpha_{j,2} \log t + \omega_j \tag{7.1}$$

where $j = 1, ...p$. As before, if $m_j > 0$ this implies that the covariate value z_j is unobserved. We assume independence between each ω_j. Therefore, we can separate the density function of the mechanism into the product of the individual densities. This independence may not be appropriate in other situations. In this case we would need to consider how we would allow for the dependence structure and then calculate the likelihood function. This would involve multivariate numerical integration and there may be issues with identifiability.

7.1.4 The Likelihood Function

Let us look specifically at the likelihood for a bivariate survival model. The bivariate survival model is of the form

$$\log T_i = t'_i = \eta_0 + \eta_1 z_{1,i} + \eta_2 z_{2,i} + \sigma\epsilon_i$$

where $z_i = (z_{1,i}, z_{2,i})^T$ is the column vector of two covariates. Therefore, the missing data model used is of the form

$$\begin{pmatrix} m_{1,i} \\ m_{2,i} \end{pmatrix} = \begin{pmatrix} \alpha_{1,0} \\ \alpha_{2,0} \end{pmatrix} + \begin{pmatrix} \alpha_{1,1} \\ 0 \end{pmatrix} z_{1,i} + \begin{pmatrix} 0 \\ \alpha_{2,1} \end{pmatrix} z_{2,i} + \begin{pmatrix} \alpha_{1,2} \\ \alpha_{2,2} \end{pmatrix} t'_i + \begin{pmatrix} \omega_{1,i} \\ \omega_{2,i} \end{pmatrix},$$

where ω_1 and ω_2 are independent error terms. We will start by using a fully saturated model for the covariate model.

We can now construct the likelihood as follows. As before, we split the likelihood into different components based upon the missing data pattern, initially assuming that we have a recorded failure.

Group 1) Individuals with observed data on both covariates

$$\begin{aligned} L_1(\eta, \sigma, \alpha, \theta | t'_i, z_i) &= P(M_1 < 0, M_2 < 0, T' = t'_i, Z = (z_{1,i}, z_{2,i})) \\ &= P(M_1 < 0 | T' = t', Z_1) P(M_2 < 0 | T' = t', Z_2) P(T' = t' | Z_1, Z_2) P(Z_1, Z_2). \end{aligned}$$

The exact parametrization is obviously determined by the choice of distribution.

Group 2) Individuals with missing data on both covariates.

$$\begin{aligned} L_1(\eta, \sigma, \alpha, \theta | t'_i, z_i) &= P(M_1 > 0, M_2 > 0, T' = t'_i) \\ &= P(M_1 > 0 | T' = t') P(M_2 > 0 | T' = t') P(T' = t') \\ &= \sum_{Z_1, Z_2} P(M_1 > 0 | T' = t', Z_1) P(M_2 > 0 | T' = t', Z_2) P(T' = t' | Z_1, Z_2) P(Z_1, Z_2). \end{aligned}$$

Group 3) Individuals with missing data on covariate Z_1 but observed data on Z_2.

$$\begin{aligned} L_1(\eta,\sigma,\alpha,\theta|t_i',z_i) &= P(M_1>0, M_2<0, T'=t_i', Z_2) \\ &= P(M_1>0|T'=t')P(M_2<0|T'=t', Z_2)P(T'=t'|Z_2)P(Z_2) \\ &= \sum_{Z_1} P(M_1>0|T'=t', Z_1)P(M_2<0|T'=t', Z_2)P(T'=t'|Z_1, Z_2)P(Z_1|Z_2)P(Z_2). \end{aligned}$$

Group 4) Individuals with missing data on covariate Z_2 but observed date on Z_1.

This is the same as for Group 3) above but with the covariates inverted.

We then construct the full likelihood as the sum of the log of each component, remembering to integrate to find the survival function if the time is censored.

As with the univariate model we need to consider the identifiability of this multivariate case. This can again be done using significant statistics. Details are not presented here but follow from the calculations of Section 5.4. Identifiability would become a point of concern if we did not force the missing data mechanisms to be independent conditional upon time as we have done.

7.2 Multivariate Analysis of Cerebral Palsy Data

We now go on to present the results and discussion of a multivariate analysis of the adult sub-cohort of the cerebral palsy data. We focus on the anal-

ysis from the adult cohort because our analysis so far has suggested that the data is MAR (or, at least, our NMAR model provides little additional information). This means we can compare the reliability of the multivariate model in comparison to available case estimates. We start by looking at bivariate models. We use a fully saturated multinomial model for the covariates.

7.2.1 Fitting Bivariate Models to the Adult Cohort

In attempting to fit the bivariate models we run in to serious issues with convergence and a lack of data. This is because the small proportions of severe disabilities mean that when considering two covariates we see very few individuals in some of the contingency table cells. Looking at available case models the optimal choice uses the ambulation and IQ covariates. Table 7.1 shows the breakdown of the data for the two covariates.

	IQ							
	Not severe		Severe		Missing		Total	
Ambulation								
Not severe	218	(59.2)	33	(9.0)	47	(12.8)	298	(81.0)
Severe	3	(0.8)	11	(3.0)	4	(1.1)	18	(4.9)
Missing	15	(4.1)	8	(2.2)	29	(7.9)	52	(14.1)
Total	236	(64.0)	52	(14.0)	80	(22.0)	368	(-)

Table 7.1: Number (percent) of severe ambulation and IQ in the adult cohort

The proportion with non-severe IQ but severe ambulation is very low. This means that the missing data mechanism is going to be very hard to identify.

In Figure 7.1 we show the estimated survival curves from the available

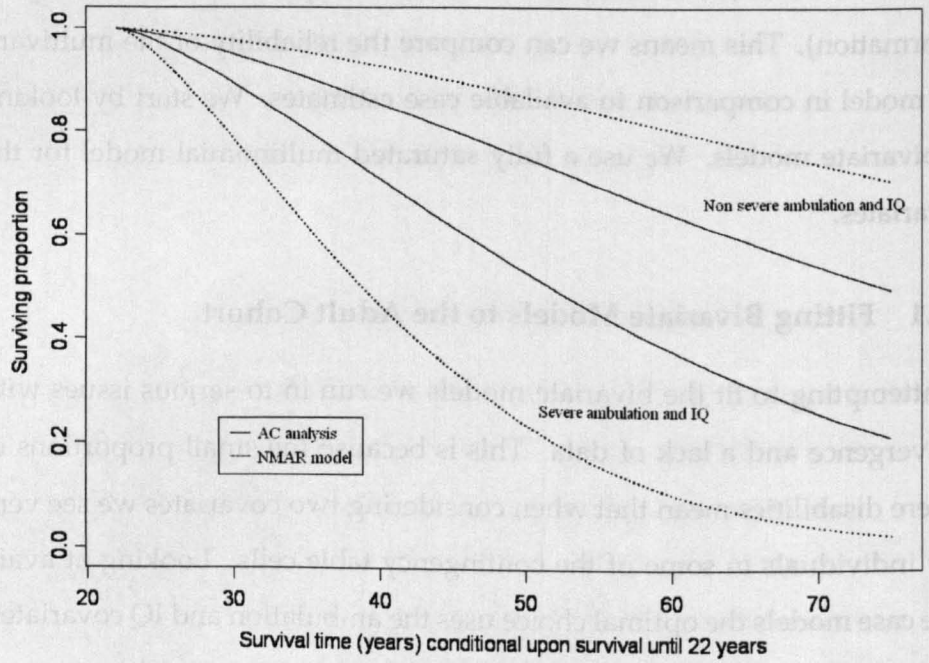


Figure 7.1: Survival for those with non-severe ambulation and IQ or severe ambulation and IQ (age in years).

case and NMAR model for ambulation and IQ but only for those with two non-severe level disabilities and those with two severe level disabilities. The survival curves for those with one severe disability are similar over the available case and NMAR model and are also quite similar for each of the two covariates. This suggests that it may be the number of severe disabilities that is the important factor when considering survival. We can again see how there is considerable survival into older age in this cohort. However, we see that with the NMAR model we arrive at vastly different survival estimates for those with no severe disabilities and for those with

two severe levels. When modelling the missing data mechanism survival considerably improves in those with neither severe ambulatory disability with 90% survival increasing from 37.4 to 44.5 years and gets much worse in those with both severe impairments with 90% survival decreasing from 30.7 to 26.6 years.

The model estimated here is used as an example of the implementation of a multivariate model. We should look at a simulation study to investigate the efficiency and precision of the model in multiple dimensions. We have not used the bivariate model to investigate survival in the incident cohort. The issue of left-truncation in the incident cohort complicates the likelihood and makes convergence of the maximization of the likelihood difficult to achieve particularly given the small number of events in the available "exposed" time. The higher levels of missing data also mean that we are seeing very few individuals with fully observed data at severe levels on two covariates.

7.2.2 Further Multivariate Models

It would also be interesting to look at a model for survival involving all of the four disability covariates. Given the lack of data at severe levels it would be probably impossible to fit but can be discussed in theory. However, the increase in covariates obviously adds a considerable number of parameters to the model which will lead to an increase in work in the optimization. Therefore, we look at reducing the number of parameters by restricting the covariate model.

If we look at the cerebral palsy data and fit a series of different models to the observed data we see that the best model uses a factor for the num-

ber of severe disabilities (rather than the separate disability covariates) and the severity of IQ. This means using a variable with levels 0,1,2,3 and, 4 instead of four binary covariates. It suggests that the three physical disabilities have a similar distribution and action but that IQ behaves slightly differently and independently of the others. This is quite interesting. We would, perhaps, expect IQ to behave differently as the others are physical disabilities.

However, even after reducing the model our program we are still unable to fit the model due to a lack of data. Note also that we are trying to simultaneously find 23 parameters and the likelihood involves considerable numerical integration which would complicate the optimization even if we had more data. Despite this it is interesting to look at a possible alternative model to avoid this issue.

Model using the Number of Severe Level Disabilities

Another alternative would be to consider the effect of the total number of severe level disabilities through a four level factor variable, the levels corresponding to 1, 2, 3, or 4 severe disabilities. Hutton & Pharoah (2002) show how survival decreases rapidly with an increase in the number of severe level disabilities. This approach may not be appropriate in the analysis of the adult cohort as we observe very few cases with 3 or 4 severe disabilities (and, hence, very few deaths) but may be of more interest in the incident cohort. However, particular issues with such a model require consideration.

The problem here is that we have partial information on the factor as it is the sum of four partially observed covariates. For example, we may ob-

serve, for a specific individual, non-severe levels of ambulation and manual dexterity but have missing values of vision and IQ. This means that the number of severe level disabilities can be at most 2. This pattern can vary for each individual causing complications. If we do not take this partial information into account we will be missing important information and hence losing precision. However, in order to do this we can cause complications with the construction of the missing data mechanism model. How will this model work now that our data is not simply missing or observed. Perhaps we can use four separate missing data models as before, although of course this will not reduce the number of parameters which was part of our aim. Alternatively, we might develop a model based upon a different latent variable which allows for the different patterns of missing data. Methods for this have not been considered and are beyond the scope of this thesis.

7.3 Discussion and Conclusions

In this chapter we have tried to consider multivariate models to look at the survival of children with cerebral palsy. We have shown that survival decreases with the number of severe disabilities but that there is still considerable survival into later years. This implies that funding for resources specific to older sufferers is important and exact levels should be thought about.

However, we have been hindered by the complexity of the model and the simplicity of our program despite attempts to simplify as much as possible. Identifiability would be a definite issue in more complex models. This is definitely an area for further development. In particular, multivari-

ate models are difficult to fit when the proportion of cases at some factor levels are low. We should consider looking at sensitivity type model of the type used by (Copas & Shi 2001) in the case of selection bias. The lack of data means that we are in a similar situation and the missing data mechanism cannot be well identified using the data alone.

7.4 Further Extensions

There are several extensions to this model that we have not considered, mainly because they were not applicable to our motivating data but partly because the computation can become very difficult. We will briefly discuss some of these here.

7.4.1 Incorporating Continuous Covariates

Our interest lay in looking at the association of discrete (binary) covariates with survival time. However, it is worth noting that, theoretically, we can incorporate continuous covariates into our model. One way of doing this would be to use a general location model as used by Cho & Schenker (1999) and Lipsitz & Ibrahim (1998) for example. This type of model splits the covariates into discrete and continuous; uses a multinomial model (possibly restricted) for the discrete covariates and then places a continuous distribution over each cell to model the remaining continuous covariates.

There are several issues with this model. Firstly, while deciding upon an appropriate multinomial model is relatively straightforward (we can use a fully saturated model) the choice of continuous distribution is less obvious. A Gaussian distribution is the normal choice but an investigation into the effects of misspecification on the accuracy of parameter estimates would

be important. Secondly, if we are trying to include discrete and continuous covariates the number of parameters to be estimated increases rapidly and approaches to restricting the model become increasingly unclear. Another complication occurs when using multiple continuous covariates. The possibility of different distributions describing the different covariates means that multivariate distributions, and hence the covariance structures, may be very complicated.

The inclusion of continuous covariates highlights the issues that arose when we considered the multivariate model. This joint model becomes complicated with more complex covariate structures and the possibilities of misspecification increase.

7.4.2 Allowing for Informative Truncation

In Section 2.2 we discussed the issue of left truncation and discussed its comparison to censoring. Throughout we have assumed independent censoring. This seems a valid assumption as censoring is almost totally forced by the current censoring date, in this case December 2005. This means that we can ignore the censoring mechanism.

We have also explained why our survival times are subject to left truncation. However, we have not focussed upon the mechanism behind this truncation. We have assumed independent truncation so that, as with the independent censoring, we can ignore the distribution of the truncation times and easily incorporate the conditioning upon survival until entry into the likelihood function.

As with any assumption it is important that we consider its validity. Children become known to the study only when they were referred to Dr

Woods, the paediatrician who collected the data. The question is, how quickly did she see the children and was this associated with their survival. As we are unsure about the exact process of referral this is difficult to consider but its seems reasonable that entry time may be associated with severity but once adjusting for this there is no further association with survival time. If this assumption was not valid we would have to model the entry time distribution.

Chapter 8

Conclusions and Discussion

We have now completed the main body of research for this thesis. We will briefly summarise the content of the preceding chapters, drawing together the results, before presenting the conclusions that can be drawn. The main motivation was to look at the long term survival of a cohort of children diagnosed with cerebral palsy. The data also posed interesting theoretical statistics questions.

Firstly, we amassed and presented the established theoretical background that we considered would be required for the later analysis and methodological work. Established methods include models for the analysis of survival data and the framework used for handling missing data. Survival analysis is concerned with the analysis of time to event data. Our focus lay with parametric models for survival although we also considered semi-parametric methods as these are exceptionally popular with applied statisticians and epidemiologists. In particular we presented methods adapted to deal with left truncated survival times. Within this chapter we introduced the commonly used taxonomy of missing data assumptions derived

by Rubin (1976). We commented on the application of these assumptions and developed a correction to the MAR assumption to avoid the possible confusion within it. Having presented these established methods we could focus on more recent work discussing issues similar to those posed by our motivating data.

Chapter 3 contained a full literature review of the handling of missing data in survival analysis. We opened with the standard methods used commonly in the analysis of data sets with missing observations and compared the advantages and disadvantages of each. These methods included case deletion and imputation techniques. We then described methods developed in the literature to specifically incorporate missing data into survival analysis models. These focused on the Cox proportional hazards model. This is partly because it is a popular model useful for investigating the relative risks of failure within different groups and partly because the semi-parametric nature means that its implementation requires a profile likelihood.

This review left several open questions. We had seen a focus on the Cox model. While this is a useful flexible model the assumption of proportional hazards is not necessarily appropriate for cerebral palsy data. Even if the assumption is sensible if we can correctly fit a parametric form to the hazard we achieve higher power in our model.

Research has also focussed upon Rubin's MAR (missing at random) assumption. This assumption is generally of use in missing data problems and several situations were discussed by various authors. However, this is an untestable assumption so as with any analysis we should look at the sensitivity of models to it. Choosing a more flexible model that allows us to model the missing data mechanism means we can consider less restric-

tive assumptions. Possible approaches for doing this are discussed later in Chapter 5. However, first we start with a more basic statistical analysis of our data. This is to further our understanding of the demographics and association within our cohort.

In Chapter 4 we fully summarised the motivating data. The data came from an early study into children with cerebral palsy. It contains full information on survival time and complete censoring information. However, some of the covariate data is missing.

Our interest in the data lies with its possible use in looking at long term survival rates by level of disability. Specifically, we construct two sub-cohorts of the whole data set to answer two slightly different questions. Firstly, what is the survival from diagnosis and how is this associated with the level of disability and, secondly, given a child has survived into adulthood what is their future expected survival, does this still depend on the baseline level of disability, and how does survival differ to that taken from diagnosis?

The first set of questions are looked at using all data available, conditional upon survival until 2 years of age. This is an approximate average age of diagnosis. However, not all children have entered the study by this age so we have the issue of left truncation, as discussed in Chapter 2. The issue of left truncation is avoided in the second sub-cohort as we now look at survival conditional upon age 22 years, an age older than any of the individuals at the time of their first assessment i the study. This sub-cohort is used to look at the second set of questions as discussed above.

We see that, in both sub-cohorts, levels of severe physical disability are low with a severe impairment being strongly associated with a decrease in survival time. This supports alternative work on both this, and other,

data. There are higher levels of severely low IQ, the measure of intellectual capacity used, and this is again associated with a significant decrease in survival. The association with survival time was looked at via complete case non-parametric survival methods. However, it is highly unlikely that data are missing completely at random and hence estimates will be biased. Therefore, we must look at the pattern of missing data and think about possible mechanisms that may be underlying it.

After concluding that the MCAR assumption was almost certainly not valid we used a likelihood based analysis to allow for the MAR assumption. This method was an extension of earlier work by Schluchter & Jackson (1989). It allows us to calculate survival estimates based upon parametric hazards via maximum likelihood techniques. The MAR assumption means that the mechanism is *ignorable* and so we do not have to directly model it. We compared the estimates from this likelihood analysis to those from multiple imputation techniques concluding that they led to similar results. Multiple imputation is a valuable technique. We can develop strategies for filling-in the missing data and then continue analysis as we would on complete data. Discussion of our likelihood based MAR model and the imputation techniques we used highlights the complex associations in our data. There is clearly a strong relationship between the disabilities and the survival time. It seems that the information held in the disability variables with regard to the missing data mechanism is very similar to that held by the survival time. This information, along with the need to look at the sensitivity of our model to the MAR assumption, leads us to consider the development of a more complex joint model.

As previously mentioned, our review of possible methods continued in Chapter 5 when we discussed selection and pattern mixture models for

non-ignorable missing data. These methods attempt to allow for less restrictive missing data mechanisms.

We compared the issue of missing covariate data to that of selection bias and then focussed on a selection type model to jointly estimate the survival time and the missing data mechanism. The construction of this model was based upon ideas that have previously arisen in the selection bias literature. However, we are in a slightly different situation as we are assuming that we are seeing all children affected by cerebral palsy so our survival, and censoring, data are complete, we are only missing some covariate information. In selection bias issues we have the situation when it is possible that the cohort is not complete or is not representative of the population.

We discussed in detail the formulation of this model and the calculation of the likelihood function. The structure of the model followed from the results of our MAR analysis. We also discussed the practical evaluation of the likelihood function. This is a complicated model and the maximisation of the likelihood requires both numerical integration and optimisation techniques. We showed how this parametric model is flexible enough to take a variety of survival distributions and presented the likelihood for each. We also extended this model to allow for left-truncation.

The chapter concluded with an application of this joint model to the cerebral palsy data and a comparison of survival estimates across the range of missing data mechanisms. Estimates in the cohort looking at survival from diagnosis have changed considerably over the different mechanisms. We have previously concluded that missingness is dependent upon survival time, the shorter the survival the higher chance of missing data as there is less time in which to collect it. Indeed, this relationship forms part of our NMAR model. Therefore, we would expect estimates to change in

this sub-cohort as we are including more children with short survival times. We also believe that shorter survival is associated with severe level impairments, implying that we are now seeing more children with greater disability. If our belief that missingness is also associated with the true level of severity is true this would also cause bias in our survival model estimates if not accounted for. One of the key sections in this chapter discussed the importance of sensitivity analysis to the linear structure of the missing data model.

The performance of this univariate joint model was investigated in the simulation study, results of which were shown in Chapter 6. This performance was compared to case deletion and MAR methods, and also to the "true" data estimates, and proved to be effective at moderate levels of missing data. In any statistical analysis it is important to test the sensitivity of the model to any untestable assumptions. We want our model to be accurate under the NMAR mechanism but it must still produce reliable estimates under the more restrictive assumptions. We also want it to be efficient. We do not look at the robustness of our model to deviations from the assumed distribution and structure. This is also important as we can not say conclusively if we have specified the correct model. However, this was beyond the scope and time constraints of this thesis.

Finally, in Chapter 7 we discussed the model within a multivariate setting. Until now we have only looked at univariate analyses but it is important to look at multivariate models for this data and, also, for generalisation of the techniques. Examples were again taken from the motivating cerebral palsy data. We showed that while multivariate models could be constructed they became difficult to fit within our data. This would likely be a problem even in larger data sets with more even risk sets. We also

suggested extension to the model to allow for continuous covariates and informative censoring or truncation. There are limitations to our model with regard to these issues. We are already making many assumptions that we need to investigate and these extensions require more.

8.1 Long-term Survival in Cerebral Palsy

Our main focus for this thesis was the analysis of long term survival for children diagnosed with cerebral palsy. We noted at the start that this was vitally important for the allocation of funding and resources in an ageing cohort. Other UK databases can look at 40 year survival but we have seen that moderately severely disabled people can easily live until 70 years old. Univariate models show that severe physical and cognitive disabilities have a large negative impact upon survival, reducing 75% survival by approximately 25 years. Severe IQ (i.e. an IQ < 50), in particular, has a major effect. Interestingly we saw that IQ behaves differently to the physical covariates when looking at the joint covariate distribution in the multivariate models.

We can consider both of our chosen sub-cohorts in turn. Firstly, the adult cohort. Here we conditioned upon survival until 22 years to avoid the need to allow for left truncation. We could use this cohort to investigate survival in those individuals who have managed to survive into adulthood, possibly those who are naturally "better" survivors. In this cohort, we, unsurprisingly, see lower levels of severe disability. It transpires that the data is well modelled using Weibull models. Once a child has survived two decades they have made it through the periods of greatest risk so it is reasonable to believe that their hazard functions will be similar to that of the

complete adult population. We also saw similar survival model estimates for the adult cohort over increasingly flexible missing data mechanisms. This suggests data is missing with reasonable randomness perhaps contrasting to our initial belief that missingness depends on both severity and survival time.

In the incident cohort we included children from the age of 2 years meaning that we had to handle the issue of left truncation. We assumed non-informative truncation. However, we also discussed how we might go about modelling the late entry. In this cohort we saw lower survival rates and higher levels of disability. In particular, we captured more of those with three or four severe level disabilities. Failure rates for the most severely disabled were very high.

Survival in the incident cohort was modelled slightly differently to that in the adult cohort. The choice of optimal distribution becomes less clear. In particular, it was slightly difficult to distinguish between the Weibull and Log-logistic distributions. This is possibly because the estimated log-logistic parameters are close to the point where the hazard function switches from a monotonic decrease and a single early peaked function.

8.2 Modelling the Missing Data Mechanism

Whenever we are analysing data with missing information it is important that we think about the possible mechanisms that might be underlying the data. Without the right approach analysis can result in biased estimates. This was the second focus for this thesis. We developed a joint model for the survival time and the missing data mechanism in order to allow for NMAR patterns. This selection model came from work in selection and

publication bias. We were able to directly estimate the complete model as our data was only partially unobserved i.e. we could assume we had the whole population of individuals and that it was only some of their baseline covariate data that we were missing.

There are several important points raised by such analyses. Firstly, the main advantage of this model was that we could use it to investigate the missing data mechanism and see how allowing for simple NMAR mechanisms changed survival estimates. We could use its results to increase our understanding of the data structure. However, due to the nature of the model we had to keep the model for the missing data mechanism quite simple in order to be able to identify it. Therefore, we could not place too great a reliance on the exact point estimates. Instead, we had to consider the sensitivity of our results to the model and possible consider a range of sensible mechanisms and, hence, survival models. With our univariate models the obvious sensitivity analysis was to change the dependence of missingness upon survival time but this would become much more difficult in more complicated setting. For example, how might we adapt the model to conduct a sensitivity analysis if we had multiple or continuous covariates or the data was collected via unusual techniques.

Secondly, our work highlighted the issue of how important it is to consider the missing data mechanism. The commonly used complete or available case methods are probably rarely appropriate and, as displayed by our review, there is a large literature available to implement methods that assume only the MAR method. Multiple imputation is particularly useful as it can exist separately from the analysis model so can be used in many situations. The availability of computer software for simulating imputations and combining results is increasing and should be recommended.

The main disadvantage of the full model is the computational difficulties in implementing it. As each covariate added to the survival model results in the need to estimate at least three more parameters for the missing data mechanism as well as the contribution to the covariate model we quickly see a vast increase in computation time. This is before we even allow for more complex dependence structures in the missing data mechanism. Therefore, it not possible to use a large saturated model as a starting place for investigation.

This model seems to be of use in this analysis only in the univariate models. A lack of data means that multivariate models are hard to identify. This may not be the case if the proportions observed at the different levels were of a similar magnitude. Sensitivity analyses such as seen in the selection bias literature (Copas & Shi 2001) are necessary here. They are also needed if we wish to fit more complex missing data mechanisms, thus making the model unidentifiable.

8.3 Discussion, Criticism and, Further Work

This thesis is, of course, limited in content. Time constraints mean that we are unable to consider all possible methods for analysis or all the questions we may wish to ask.

This was a challenging data set and we have only dealt with some of the issues it raises. From a epidemiological point of view it can provide valuable information that other cohorts can not due to its length and completeness. The missing data and truncation are the obvious problems. Follow up for a period as long as that in this study is always difficult. This raises problems that can also be seen in the data. As data were collected

over 50 years ago it is difficult to go back and learn more about the collection process and methods. Also many factors affecting the distribution of the data are likely to change over both the study period and the complete follow up. It is possible that during the study period the tests used to measure disability or the way the results were interpreted changed. It is also possible that the behaviour of the paediatrician changed with regards to diagnosis or treatment. We are unfortunately unable to go back and find answers to these questions.

Another issue caused by the long follow up is the almost definite improvement in medical care and expertise over the period. This suggests a possible extension in this work, to include calender time in the model. This could allow survival patterns to change over time. This of course raises the issue of how exactly this could be done and complicates the model further.

We have presented here an interesting and flexible class of models. Sensitivity analysis suggested that making the simple linear assumption in the model for the missing data mechanism was reasonably adequate and this enabled direct estimation of a single survival model. However, as discussed in Chapter 7 there are obvious extensions to the model that would be useful for our cerebral palsy research that are difficult with our model. These include the multivariate models we discussed in detail as well as the inclusion of continuous covariates and informative truncation or censoring. These are all examples of methodological research that would be of further interest.

As discussed this is only one approach to analysing data such as these. Alternatives might be NMAR imputation or a fully Bayesian approach. There are both advantages and disadvantages to these techniques. NMAR

imputation usually uses pattern mixture models. Where as in normal MI we assume that those with observed and missing data are similar now we assume they are different but we don't know how. We have to consider imputations under a variety of different mechanisms and look at the sensitivity of the model estimates to the choice of imputation model. We do not model the mechanism as we do in our analysis but specify it. Techniques such as this require a really good understanding of exactly what data we would expect if it were complete and, using this, what the possible mechanism might be. However, once this has been considered and we have the complete data analysis is much more straightforward although we arrive at a set of possible models. We were able to estimate the mechanism after making some distributional assumptions about it and then we only had to look at sensitivity to the distributional assumptions.

An alternative method would be to take a Bayesian approach and specify priors for the mechanism model parameters. This would lead to greater stability in the model estimates, particularly in the event of sparse data as in the multivariate analysis, and also mean we could arrive at a single model estimate. Maximum likelihood based methods can often be well approximated using Bayesian machinery. There is a growing literature in Bayesian approaches to this problem (e.g. Scharfstein et al. (2003)) particularly with reference to non-ignorable dropout (e.g Rotnitzky et al. (1998)). As with any Bayesian analysis priors have to be first elicited and in the case of parameters in a missing data mechanism this is extremely complicated.

The use of a Bayesian approach might mean that numerical convergence is easier to obtain. We encountered some difficulties with obtaining convergence in the maximisation of the likelihood although refinement of the numerical methods used may have helped with this. In particular, there

are possible modifications to the starting values and Newton-Raphson iteration that can help reach convergence.

8.4 Final Remarks

This thesis looked at both the applied and the theoretical issues involved in the analysis of a data set. It attempted to collate the available literature looking at missing data in survival analysis and apply these, and a new model, to estimate long term survival for sufferers of cerebral palsy. It showed how important it is to fully consider the missing data mechanism and highlighted the sometimes forgotten issue of how important it is to understand the data and the collection methods before embarking on an analysis. What should not be forgotten is that any attempt to model data with potentially NMAR observations is extremely complicated and heavily dependent upon assumptions meaning that we should make every effort when collecting data to render at least the MAR assumption plausible.

Appendix A

MAR model extension programs

Here, we present the S-Plus functions to calculate the likelihoods for the Weibull and log-normal extensions to the stepwise method of Schluchter & Jackson (1989). Density function for the two distributions can be found in Section2.1.4. We are fitting under the accelerated failure assumption and are assuming for the moment independent left-truncation and right-censoring. These S-Plus functions can be used to maximise the likelihoods by Newton-Raphson methods via the S-Plus in-built *nlminb* function. See Venables & Ripley (2002) for details. Function *nlminb* only calculates the Hessian matrix at the solution if a means to calculate it is provided. This becomes particularly complicated for the Burr distribution so we can use the function *vcov.nlminb* in the `MASS` library of uses a finite difference approximation to the Hessian.

Function variables

para Vector of model parameters. If we assume that the number of cells in the contingency table is M then for the Weibull distribution `para` $= (\theta_1, ..\theta_{M-1}, \lambda_1, .., \lambda_M, \gamma)$ and for the log-normal distribution `para` $= (\theta_1, ..\theta_{M-1}, \mu_1, .., \mu_M, \sigma)$, where θ_m is the probability of being in cell m and (λ_m, γ) are the parameters for the Weibull distribution in cell m and (μ_m, σ) are the parameters for the log-normal distribution in the mth cell.

surv The vector of survival times t_i $(i = 1, .., n)$, possibly subject to independent left-truncation and right-censoring.

enter The vector of entry times (truncation times)

censor The vector of censoring indicators

w Matrix of dimension $(n \times M)$ where

$$\begin{aligned} w[i,m] &= 1 \quad \text{if individual } i \text{ can lie in cell m,} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

This can be constructed from the vector, or matrix, of covariate values.

Weibull distribution

```
# Hazard function for the Weibull
haz.comp<-function(lamda,gamma,time)
{ lambda*gamma*(time^(gamma-1)) }

# Survival function for the Weibull
surv.comp<-function(lamda,gamma,time)
{ exp(-(lamda*(time^gamma))) }
```

```
# Constructs likelihood by sum of contributions over \(i\)
and \(m\)
weibull.likelihood.func<-function(para,surv,enter,censor,w)
{
like.temp<-matrix(0,length(surv),ncol(w))
like<-rep(0,length(surv)) for(i in 1:length(surv)) {
  for(m in 1:(ncol(w)-1))
  {
    like.temp[i,m]<-((w[i,m]*para[m])*
    (haz.comp(para[ncol(w)-1+m],para[2*ncol(w)],surv[i])
    ^(censor[i]))*
    (surv.comp(para[ncol(w)-1+m],para[2*ncol(w)],surv[i])))/
    (surv.comp(para[ncol(w)-1+m],para[2*ncol(w)],enter[i]))
  }
    like.temp[i,ncol(w)]<-((w[i,m]*
    (1-sum(para[1:(ncol(poss)-1)])))*
    (haz.comp(para[(2*ncol(w))-1],para[2*ncol(w)],surv[i])
    ^(censor[i]))*
    (surv.comp(para[(2*ncol(w))-1],para[2*ncol(w)],surv[i])))/
    (surv.comp(para[(2*ncol(w))-1],para[2*ncol(w)],enter[i]))
  for(i in 1:length(surv))
  {
    like[i]<-sum(like.temp[i,])
  }
  return(-sum(log(like)))
}
```

Log-Normal distribution

```
# Density function for the log-normal
```

```
frac.comp<-function(sigma,time)
{ 1/(time*sigma*sqrt(2*pi)) }

exp.comp<-function(mu,sigma,time)
{ exp(-((log(time)-mu)^2)/(2*(sigma^2))) }

# Survival function for the log-normal surv.comp<-function(mu,sigma,time)
{ 1-pnorm((log(time)-mu)/sigma) }

# Constructs likelihood by sum of contributions over \(i\)
and \(m\)
lognorm.likelihood.func<-function(para,surv,enter,censor,w)
{
like.temp<-matrix(0,length(surv),ncol(w))
like<-rep(0,length(surv))
for(i in 1:length(surv))
{
  for(m in 1:(ncol(w)-1))
  {
    like.temp[i,m]<-((w[i,m]*para[m])*
    ((frac.comp(para[2*ncol(w)],surv[i])*
    exp.comp(para[ncol(w)-1+m],para[2*ncol(w)],surv[i]))
    ^(censor[i]))*
    (surv.comp(para[ncol(w)-1+m],para[2*ncol(w)],surv[i])
    ^(1-censor[i])))/
    (surv.comp(para[ncol(w)-1+m],para[2*ncol(w)],enter[i]))
  }
    like.temp[i,ncol(w)]<-((w[i,ncol(w)]*
    (1-sum(para[1:(ncol(w)-1)])))*
    ((frac.comp(para[2*ncol(w)],surv[i])*
```

```
    exp.comp(para[(2*ncol(w))-1],para[2*ncol(w)],surv[i]))
    ^(censor[i]))*
    (surv.comp(para[(2*ncol(w))-1],para[2*ncol(w)],surv[i])
    ^(1-censor[i])))/
    (surv.comp(para[(2*ncol(w))-1],para[2*ncol(w)],enter[i]))
  for(i in 1:length(surv))
  {
    like[i]<-sum(like.temp[i,])
  }
  return(-sum(log(like)))
}
```

Appendix B

Gaussian Quadrature

This is the program used to implement 10 point Gaussian quadrature used in the implementation of the joint survival time and missing data model.

```
gauss.int.10<-function(f, umin,umax, n, entry, parameters) {
    w <- c(0.0666713, 0.1494513, 0.2190864, 0.2692667, 0.2955242,
        0.2955242, 0.2692667, 0.2190864, 0.1494513, 0.0666713)
    p <- c(-0.9739065, -0.8650634, -0.6794096, -0.4333954, -0.1488743,
        0.1488743, 0.4333954, 0.6794096, 0.8650634, 0.9739065)
    d <- (umax - umin)/n
    ans <- 0
    for(j in 1:n) {
        uj <- umin + d * (j - 1)
        sumj <- 0
        for(k in 1:10)
            sumj <- sumj + w[k] * f(d/2 * p[k] + uj + d/2, entry,
                parameters)
        ans <- ans + sumj
    }
```

```
ans <- (ans * d)/2
ans}
```

Appendix C

The NMAR joint model with left truncation

This is the S-Plus code for the univariate joint model of the survival time and the missing data mechanism. This is for the Weibull distribution model with left truncation.

```
like.one.zero<-function(surv, parameters)
{
(1/(parameters[3]*sqrt(2*pi)))*
exp(-((surv-parameters[1])^2)/(2*(parameters[3]^2)))*
(1-parameters[7])*
(1-pnorm(parameters[4]+(parameters[6]*surv))) }

like.one.zero.left<-function(surv, entry, parameters)
{
like.one.zero(surv, parameters)/
(((pnorm((-entry+parameters[1])/parameters[3]))*(1-parameters[7]))+
((pnorm((-entry+parameters[1]+parameters[2])/parameters[3]))*parameters[7]))
```

```
}

like.one.one<-function(surv, parameters)
{
(1/(parameters[3]*sqrt(2*pi)))*
exp(-((surv-parameters[1]-parameters[2])^2)/(2*(parameters[3]^2)))*
parameters[7]*
(1-pnorm(parameters[4]+parameters[5]+(parameters[6]*surv)))
}

like.one.one.left<-function(surv, entry, parameters)
{
like.one.one(surv, parameters)/
(((pnorm((-entry+parameters[1])/parameters[3]))*(1-parameters[7]))+
((pnorm((-entry+parameters[1]+parameters[2])/parameters[3]))*parameters[7]))
}


like.three.zero<-function(surv, parameters)
{
(1/(parameters[3]*sqrt(2*pi)))*
exp(-((surv-parameters[1])^2)/(2*(parameters[3]^2)))*
(1-parameters[7])*
pnorm(parameters[4]+(parameters[6]*surv)) }

like.three.zero.left<-function(surv, entry, parameters)
{
like.three.zero(surv, parameters)/
(pnorm((-entry+parameters[1])/parameters[3]))
}

like.three.one<-function(surv, parameters)
{
```

```
(1/(parameters[3]*sqrt(2*pi)))*
exp(-((surv-parameters[1]-parameters[2])^2)/(2*(parameters[3]^2)))*
parameters[7]*
pnorm(parameters[4]+parameters[5]+(parameters[6]*surv))
}

like.three.one.left<-function(surv, entry, parameters)
{
like.three.one(surv, parameters)/
(pnorm((-entry+parameters[1]+parameters[2])/parameters[3]))
}

like.four.left<-function(surv, entry, parameters)
{
like.three.zero.left(surv, entry, parameters)+
like.three.one.left(surv, entry, parameters)
}

log.like.func.left(parameters.start, incident$SurvTimeYrs-2,
entry.adjusted, incident$Censoring, incident$SevAmb)
log.like.func.left<-function(parameters, surv, entry, delta, cov)
{
ifelse(!is.na(cov),
ifelse(delta==1,
ifelse(cov==0,
log.like<-log(like.one.zero.left(surv, entry, parameters)),
log.like<-log(like.one.one.left(surv, entry, parameters))),
ifelse(cov==0,
log.like<-log(gauss.int.10(like.one.zero.left, surv,
100, n=20, entry, parameters)),
log.like<-log(gauss.int.10(like.one.one.left, surv, 100, n=20,
entry, parameters)))),
ifelse(delta==1,
```

```
log.like<-log(like.four.left(surv, entry, parameters)),
log.like<-log(gauss.int.10(like.four.left, surv, 100, n=20, entry,
parameters))))
}

max.log.like.left<-function(parameters, surv, entry, delta, cov)
{
-sum(log.like.func.left(parameters, surv, entry, delta, cov))
}
```

Bibliography

Anderson, P., Borgan, O., Gill, R. & Kielding, N. (1989), *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

Baker, S. (1994), 'Regression analysis of grouped survival data with incomplete covariates: nonignorable missing data and censoring mechanisms', *Biometrics* **50**, 821–826.

Baker, S. & Laird, N. (1988), 'Regression analysis for categorical survey variables with outcome subject to nonignorable nonresponse', *Journal of the American Statistical Association* **83**, 62–69.

Bishop, Y., Fienberg, S. & Holland, P. (1975), *Discrete Multivariate Analysis Theory and Practice*, MIT Press, Cambridge, MA.

Blair, E., Watson, L., Badawi, N. & Stanley, F. (2001), 'Life expectancy among people with cerebral palsy in western australia', *Developmental Medicine and Child Neurology* **43**, 508–15.

Burr, I. (1942), 'Cumulative frequency functions', *Annals of Mathematical Statistics* **13**, 215–232.

Cerebral Palsy - Ask the Doctor (2006), http://www.about-cerebral-palsy.org.

Chen, H. & Little, R. (1999), 'Proportional hazards regression with missing covariates', *Journal of the American Statistical Association* **94**, 896–908.

Cho, M. & Schenker, N. (1999), 'Fitting the log-f accelerated failure time model with incomplete covariate data', *Biometrics* **55**, 826–833.

Collett, D. (1999), *Modelling Survival Data in Medical Research*, Chapman & Hall / CRC, London.

Copas, J. & Li, H. (1997), 'Inference for non-random samples', *Journal of the Royal Statistical Society, Series B* **59**, 55–95.

Copas, J. & Shi, J. (2000), 'Meta analysis, funnel plots and sensitivity analysis', *Biostatistics* **1**, 247–262.

Copas, J. & Shi, J. (2001), 'A sensitivity analysis for publication bias in systematic reviews', *Statistical Methods in Medical Research* **10**, 251–265.

Cox, D. (1972), 'Regression models and life tables', *Journal of the Royal Statistical Society, B* **34**, 187–220.

Cox, D. & Oakes, D. (1984), *Analysis of Survival Data*, Chapman & Hall / CRC, London.

Crowder, M., Kimber, A., Smith, R. & Sweeting, T. (1991), *Statistical Analsysi of Reliability Data*, Chapman & Hall, London.

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society, B* **39**, 1–38.

Diggle, P. & Kenward, M. (1994), 'Informative drop-out in longitudinal data analysis (with discussion)', *Applied Statistics* **43**, 49–93.

Dinse, G. (1982), 'Nonparametric estimation for partially-complete time and type of failure data', *Biometrics* **38**, 417–431.

Evans, P., Evans, S. & Alberman, E. (1990), 'Cerebral palsy: Why we must plan for survival', *Archives of Disease in Childhood* pp. 1329–33.

Fleming, T. & Harrington, D. (1991), *Counting Processes and Survival Analysis*, Wiley, New York.

Freireich, E. e. a. (1963), 'The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: a model for evaluation of other potentially useful therapy', *Blood* **21**, 699–716.

Gamerman, D. (2002), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman & Hall / CRC, Florida.

Goetghebeur, E. & Ryan, L. (1995), 'Analysis of competing risks survival data when some failure types are missing', *Biometrika* **82**, 821–833.

Hall, W. & Welner, J. (1980), 'Confidence bands for a survival curve from censored data', *Biometrika* **67**, 133–43.

Heckman, J. (1974), 'Shadow prices, market wages, and labor supply', *Econometrica* **42**, 679–694.

Heckman, J. (1979), 'Sample selection bias as a specification error', *Econometrica* **47**, 153–161.

Heitjan, D. & Rubin, D. (1991), 'Ignorability and coarse data', *The Annals of Statistics* **19**, 2244–2253.

Hemming, K., Hutton, J., Colver, A. & Platt, M.-J. (2005), 'Regional variation in survival of people with cerebral palsy in the united kingdom', *Pediatrics* **116**, 1383–90.

Hemming, K., Hutton, J. & Pharoah, P. (2006), 'Long-term survival for a cohort of adults with cerebral palsy', *Developmental Medicine and Child Neurology* **48**, 90–5.

Herring, A. & Ibrahim, J. (2001), 'Likelihood-based methods for missing covariates in the cox proportional hazards model', *Journal of the American Statistical Association* **96**, 292–302.

Herring, A., Ibrahim, J. & Lipsitz, S. (2004), 'Non-ignorable missing covariate data in survival analysis: a case study of an international breast cancer study group trial', *Journal of the Royal Statistical Society, C* **53**, 293–310.

Hutton, J., Colver, A. & Mackie, P. (2000), 'Effect of severity of disability on survival in north east england cerebral palsy cohort', *Archives of Disease in Childhood* **83**, 468–73.

Hutton, J., Cooke, T. & Pharoah, P. (1994), 'Life expectancy in children with cerebral palsy', *British Medical Journal* **309**, 431–5.

Hutton, J. & Monaghan, P. (2002), 'Choice of parametric accelerated life and proportional hazards models for survival data: asymptotic results', *Lifetime Data Analysis* **8**, 375–393.

Hutton, J. & Pharoah, P. (2002), 'Effects of cognitive, motor, and sensory disabilities on survival in cerebral palsy', *Archives of Disease in Childhood* **86**, 84–9.

Hyde, J. (1980), *Biostatisics Casebook*, John Wiley and Sons, New York, chapter Survival analysis with incomplete observations, pp. 31–46.

Ibrahim, J. (1990), 'Incomplete data in generalized linear models', *Journal of the American Statistical Association* **85**, 765–769.

Ibrahim, J., Chen, M.-H. & Lipsitz, S. (1999), 'Monte carlo em for missing covariates in parametric regression models', *Biometrics* **591-596**, 55.

Insightful Corporation, Seattle, W. (2002), 'S-plus version 6.1 release 1', http://www.insightful.com.

Kaplan, E. & Meier, P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association* **53**, 457–81.

Katz, R. (2003), 'Life expectancy for children with cerebral palsy and mental retardartion: Implications for life care planning', *NeuroRehabilitation* **18**, 261–270.

Kenward, M. (1998), 'Selection models for repeated measurements for nonrandom dropout: An illustration of sensitivity', *Statistical Science* **17**, 2723–2732.

Kenward, M. & Molenberghs, G. (1999), 'Parametric models for incomplete continuous and categorical longitudinal data', *Statistical Methods in Medical Research* **8**, 51–83.

Klein, J. & Moeschberger, M. (1997), *Survival Analysis - Techniques for Censored and Truncated Data*, Springer-Verlag, New York.

Lawless, J. (2003), *Statistical Models and Methods for Lifetime Data*, Wiley, New Jersey.

Leathem, A. & Brooks, S. (1987), 'Predictive value of lectin binding on breast cancer recurrence and survival', *The Lancet* **1**, 1054–1056.

Leong, T., Lipsitz, S. & Ibrahim, J. (2001), 'Incomplete covariates in the cox model with applications to biological marker data', *Journal of the Royal Statistical Society, C* **50**, 467–484.

Lin, D. & Ying, Z. (1993), 'Cox regression with incomplete covariate measurements', *Journal of the American Statistical Association* **88**, 1341–1349.

Lipsitz, S. & Ibrahim, J. (1996), 'A conditional model for incomplete covariates in parametric regression models', *Biometrika* **83**, 916–922.

Lipsitz, S. & Ibrahim, J. (1998), 'Estimating equations with incomplete categorical covariates in the cox model', *Biometrics* **54**, 1002–13.

Little, R. (1982), 'Models for nonresponse in sample surveys', *Journal of the American Statistical Association* **77**, 237–250.

Little, R. (1985), 'A note about models for selectivity bias', *Econometrica* **53**, 1468–1474.

Little, R. (1992), 'Regression with missing x's: a review', *Journal of the American Statistical Association* **87**, 1227–1237.

Little, R. (1994), 'A class of pattern-mixture models for normal incomplete data', *Biometrika* **81**, 471–483.

Little, R. & Rubin, D. (2002), *Statistical Analysis with Missing Data*, John Wiley and Sons, Inc., New York.

Little, R. & Wang, Y. (1996), 'Pattern-mixture models for multivariate incomplete data with covariates', *Biometrics* **52**, 98–111.

Lui, K., Lawrence, D., Morgan, W., Peterman, T., Haverkos, H. & Bragman, D. (1986), 'A model-based aproach for estimating the mean incubation period of transfusion associated acquired immunodeficiency syndrome', *Proceedings of the National Academy of Science, USA* **83**, 3051–3055.

Lynden-Bell, D. (1971), 'A method for allowing for known observational selection in small samples applied to 3cr quasars', *Monthly Notices of the Royal Astronomical Society* **155**, 95–118.

Martinussen, T. (1999), 'Cox regression with incomplete covariate measurements using the em-algorithm', *Scandinavian Journal of Statistics* **26**, 479–491.

McKendrick, A. (1928), 'Applications of mathematics to medical problems', *Proceedings of the Edinburgh Mathematical Society* **44**, 98–130.

Medley, G., Anderson, R., Cox, D. & Billard, L. (1987), 'Incubation period of aids in patients infected via blood transfusion', *Nature* **328**, 719–721.

Meng, X. & Schenker, N. (1999), 'Maximum likelihood estimation for linear regression models with right censored outcomes and missing predictors', *Computational Statistics and Data Analysis* **29**, 471–483.

Molenberghs, G., Goetghebeur, E., Lipsitz, S. & Kenward, M. (1998), 'Non-random missingness in categorical data: Strengths and limitations', *American Statistician* **53**, 110–118.

Nahman, N. e. a. (1992), 'Modification of the percutaneous approach to peritonal dialysis catheter placement under peritoneoscopic visual-

ization: clinical results in 78 patients', *Journal of the American Society of Nephrology* **3**, 103–107.

Nelson, W. (1970), 'Hazard plotting methods for analysis of life data with different failure modes', *Journal of Quality Technology* **2**, 126–149.

Nordheim, E. (1984), 'Inference from nonrandomly missing data: An example from a genetic study on turner's syndrome', *Journal of the American Statistical Society* **79**, 772–780.

Oakes, D. (1982), 'Survival analysis', *European Journal of Operational Research* **12**, 3–14.

Olsen, R. (1980), 'A least squares correction for sensitivity bias', *Econometrica* **48**, 1815–1820.

Orchard, T. & Woodbury, M. (1972), 'A missing information principle: theory and applications', *Proceedings of the 6th Berkley Symposium on Mathematical Statistics and Probability* **1**, 697–715.

Paik, M. (1997), 'Multiple imputation for the cox proportional hazards model with missing covariates', *Lifetime Data Analysis* **3**, 289–298.

Paik, M. & Tsai, W.-Y. (1997), 'On using the cox proportional hazards model with missing covariates', *Biometrika* **84**, 579–593.

Pan, W. & Chappell, R. (1998), 'A nonparametric estimator of survival functions for arbitrarily truncated and censored data', *Lifetime Data Analysis* **4**, 187–202.

Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J. & Smith, P. (1977), 'Design and analysis

of randomized clinical trials requiring prolonged observation of each patient II. Analysis and examples.', *British Journal of Cancer* **35**, 1–39.

Pugh, M., Robins, J., Lipsitz, S. & Harrington, D. (1993), Inference in the cox proportional hazards model with missing covariate data, Technical report, Dept. Biostatistics, Harvard University.

Puhani, P. (2000), 'The heckman correction for sample selection and its critique', *Journal of Economic Surveys* **14**, 53–68.

Rotnitzky, A., Robins, J. & Scharfstein, D. (1998), 'Semiparametric regression for repeated outcomes with nonignorable nonresponse', *Journal of the American Statistical Association* **93**, 1321–1339.

Rubin, D. (1976), 'Inference and missing data (with discussion)', *Biometrika* **63**, 581–592.

Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons, Inc., New York.

Rubin, D. (1996), 'Multiple imputation after 18+ years', *Journal of the American Statistical Association* **91**, 473–489.

Schafer, J. (1999), 'Multiple imputation: a primer', *Statistical Methods in Medical Research* **8**, 3–15.

Schafer, J. & Graham, J. (2002), 'Missing data: Our view of the state of the art', *Psychological Methods* **7**, 147–177.

Scharfstein, D., Daniels, M. & Robins, J. (2003), 'Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes', *Biostatistics* **4**, 495–512.

Schluchter, M. & Jackson, K. (1989), 'Log-linear analysis of censored survival data with partially observed covariates', *Journal of the American Statistical Association* **84**, 42–52.

Scope: About Cerebral Palsy (2006), http://www.scope.org.uk.

Shao, Q. & Zhou, X. (2004), 'A new parametric model for survival data with long-term survivors', *Statistics in Medicine* **23**, 3525–3543.

Strauss, D., Ojdana, K., Shavelle, R. & Rosenbloom, L. (2004), 'Decline in function and life expectancy of older persons with cerebral palsy', *NeuroRehabilitation* **19**, 69–78.

Strauss, D. & Shavelle, R. (1998), 'Life expectancy in adults with cerebral palsy', *Developmental Medicine and Child Neurology* **40**, 369–75.

Strauss, D., Shavelle, R. & Anderson, T. (1998), 'Life expectancy of children with cerebral palsy', *Pediatric Neurology* **18**, 143–9.

Struthers, C. & Farewell, V. (1989), 'A mixture model for time to aids data with left truncation and an uncertain origin', *Biometrika* **76**, 814–817.

Tang, G., Little, R. & Raghunathan, T. (2003), 'Analysis of multivariate missing data with nonignorable nonresponse', *Biometrika* **90**, 747–764.

Taylor, J., Murray, S. & Hsu, C.-H. (2002), 'Survival estimation and testing via multiple imputation', *Statistics and Probability Letters* **58**, 221–232.

Tobin, J. (1958), 'Estimation of relationships for limited dependent variables', *Econometrica* **26**, 24–36.

Tsai, W.-Y., Jewell, N. & Wang, M.-C. (1987), 'A note on the product-limit estimator under right censoring and left truncation', *Biometrika* **74**, 883–886.

Van Buuren, S. & Oudshoorn, C. (1999), Flexible multivariate imputation by mice. Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054. For associated software see http://www.multiple-imputation.com.

Van Der Laan, M. & McKeague, I. (1998), 'Efficient estimation from right-censored data when failure indicators are missing at random', *Annals of Statistics* **26**, 164–182.

Venables, W. & Ripley, B. (2002), *Modern Applied Statistics with S*, Springer, chapter 16.

Vu, H. & Zhou, X. (1997), 'Generalization of likelihood ratio test under non-standard conditions', *Annals of Statistics* **25**, 897–916.

Wei, G. & Tanner, M. (1990), 'A monte carlo em algorithm and the poor man's data augmentation algorithms', *Journal of the American Statistical Association* **85**, 699–704.

Woodroofe, M. (1985), 'Estimating a distribution function with truncated data', *The Annals of Statistics* **13**, 163–177.

Woods, G. (1957), *Cerebral Palsy in Childhood: The Aetiology and Clinical Assessment with Particular Reference to the Findings in Bristol*, John Wright, Bristol, pp. 7–13.

Wu, C. (1983), 'On the convergence properties of the em algorithm', *Annals of Statistics* **11**, 95–103.

Wu, M. & Carroll, R. (1988), 'Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process', *Biometrics* **44**, 175–188.

Zhou, H. & Pepe, M. (1995), 'Auxiliary covariate data in failure time regression', *Biometrika* **82**, 139–149.