Missing Covariate Data in Parametric Survival Analysis - Modelling the Missing Data Mechanism.

KL Boyd and JL Hutton

March 2, 2006

Abstract

Aims and Motivation To examine the effect of level of disability on the survival of children with cerebral palsy using a cohort taken from Bristol. The data is subject to, possibly not missing at random (NMAR), unobserved covariate data.

Methods A joint survival model for the log-survival times and missing data mechanism is introduced. This approach enables us to model the missing data mechanism. This is then used to model the effect of level of ambulatory disability on survival in the cerebral palsy data. Extensions to the model are discussed to include continuous and multiple covariates.

Results Analysis suggests that the effect of severe ambulation on survival in individuals with cerebral palsy is underestimated if no account is taken of the missing data mechanism. Simulations show that this model, under various distribution assumptions, performs well in comparison to basic exclusion techniques.

Conclusions It is very important to consider the mechanism behind any missing data when studying survival. Slight deviances from the less restrictive assumptions can effect parameter estimates in survival models. In our data, we see an increased effect of severe ambulation on survival in those with cerebral palsy. A severe level of ambulatory disability causes a decrease in survival.

Key words: Survival analysis, missing data, NMAR, cerebral palsy.

1 Introduction

Survival analysis is an area of statistics in which primary observations correspond to the time from a well-defined time origin until the occurrence of some particular event or end-point. The aim of a survival analysis is often to investigate the association of recorded covariates with survival times.

Data are often not perfect and, in medical settings in particular, often have a proportion of covariate information missing. Previous work has looked at this issue but has focused on the less restrictive assumptions usually considered in a missing data framework.

Here we develop a method for modelling survival data in a parametric framework based on a model for the missing data mechanism that allows for the most restrictive of assumptions. Work is motivated by an example data set from Bristol concerning the survival of children with cerebral palsy in which there is concern regarding the form of the missing data mechanism.

After introducing the data in Section 2 we discuss the model in Section 3. Initially, we base ideas on a log-normal model but in Section 4 we extend this to other commonly used distributions in survival analysis.

An example is taken from the Bristol cerebral palsy data and results are displayed and discussed in Section 5. We then consider implications of the likelihood in Section 6.

We discuss initial results from an investigation in to the effect of factors on survival in our cerebral palsy data and in Section 7 provide results from a simulation study to investigate the reliability of these findings.

Finally, we again extend the model in Section 8 to allow for alternative covariate structures and truncation which feature in our motivating data and will be of importance in future analysis and conclude with discussion in Section 9.

2 The Bristol Data

The motivating data were ascertained from a part retrospective and part prospective 1940s and 1960s birth cohort based on a consultant paediatrician'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 Dr Woods were recorded on professionally designed punch cards. This later became the subject of her MD thesis (Woods 1957). The cohort is said to include all cerebral palsied children from Bristol and the surrounding area. The issue of left truncation arises as children are only included in the cohort if they survived until the study period and could be seen by Dr Woods.

The information held on the punch cards was subsequently compiled into an electronic database. Individuals were included if they met certain criteria and could be clearly diagnosed with CP. Only those with early impairment CP were included. Inevitably, some cases were excluded as there was not enough information to allow for diagnosis.

The data consist of information on birth weight, gestational age, mother's age at birth, and several disability covariates. These include levels of ambulation, manual dexterity, vision, and IQ. All can be grouped into severe and non-severe groups. Previous research (Hutton & Pharoah 2002) 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, September 2003. Full information on the available data can be found in Hemming, Hutton & Pharoah (2006).

We wish to consider the survival prospects of children diagnosed with cerebral palsy. However, there are complications with our data that mean that standard methods for coping with missing covariate data are not applicable. Firstly, we do not believe that the covariate data are missing at random (MAR) (Little & Rubin 2002). Most of the previous approaches require this assumption (e.g. Ibrahim, Chen & Lipsitz (1999) and Meng & Schenker (1999)). Also, the majority of methods are applied to the Cox proportional hazards model (e.g. Lipsitz & Ibrahim (1998) and Herring & Ibrahim (2001)). We would like to be able to allow for a parametric hazard (Collett 1999) as accelerated failure time models appear to be useful in modelling cerebral palsy survival. Finally, our data is subject to left truncation which is not considered in any previous research on missing data.

3 Modelling the Missing Data Mechanism

3.1 Notation

An individual i, $i = 1...n$, enters the target study population at time t_i^0 and reaches the end-point at time $t_i^0 + x_i$. However, this x_i may be right censored by time c_i, which is

independent of i. Therefore, for subject i, observed data on survival consists of (t_i, δ_i) where $t_i = \min(x_i, c_i)$ and $\delta_i = \mathrm{I}(x_i < c_i)$. Here, δ indicates whether an individual is uncensored. Left truncation occurs if an individual enters a study, not at the time that they enter the target population but, at some time w_i, where $t_i^0 < w_i < t_i$. This means that those individuals who experience an event before time w_i are not recorded in the data. Additionally, covariate data for individual i are denoted by z_i.

3.2 Missing Data Mechanisms

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 $Y \in \mathbb{Y}$ and note that we observe $Y = (y_{ij})$. This is, in reality, not entirely observed. With regards to survival analysis we can consider $Y = (T, \delta, Z)$ where T, δ, and Z are the observed survival 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 Y, where $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is observed. The missing data mechanism can then be characterized by the conditional distribution of M given Y,

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

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 Y and is called the *missing completely at random* (MCAR) assumption. This occurs if

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

A slightly less restrictive mechanism is in operation if the data is *missing at random* (MAR). Here, missingness is allowed to depend upon the observed values of Y but not on the unobserved values. Define the class of matrices $\mathbb{Y}^*$, in which each matrix shares observed entries with Y but has alternative values for those that we do not observe i.e.

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

We can then write the definition for MAR data as

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

Finally, if missingness is allowed to depend on both the observed and unobserved data (i.e. the full data, $Y = y$) then the data is said to be *not missing at random* (NMAR). This is the least restrictive of the three assumptions and to allow for it we must model the mechanism that causes the missing data.

3.3 Modelling the Mechanism

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. Therefore, we must build a model that allows us to model the missing data mechanism. The aim is to embed the MAR model within a range of plausible models that allow the NMAR assumption. We can follow

ideas already used in meta-analysis to deal with issues of publication bias (Copas & Shi 2001). The issue there is that there are likely to be a number of studies that are not published, and hence cannot be included in the meta-analysis. The reason for non-publication may depend on the outcome of the study. For example, studies that do not include a significant outcome may be less likely to be published.

There are two main approaches to formulating models for non-ignorable data. Assume that the observations to be modeled are independent. Selection models have the joint distribution of M and Y 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

$$f(m, y|\eta, \omega) = f(y|m, \eta)f(m|\omega)$$

where η and ω are again distinct parameter vectors.

3.4 The Joint Model for Survival and the Missing Data Mechanism

We can now go on to describe our joint model for survival and missing covariate data. Note that we are only considering the case when we have fully observed survival time and censoring information on n individuals. For simplicity, assume we observe just one binary covariate, $z = (z_1, ..., z_n)$, which has some missing data. Firstly, we construct a model for the survival times, T:

$$T_i' = \log(T_i) = \eta + \gamma z_i + \sigma\epsilon_i, \qquad \epsilon_i \sim N(0, 1), \qquad i = 1, ..., n. \tag{3.1}$$

We allow the survival of individual i to depend on the value of the covariate. The choice of a log-normal model for survival times is a sensible choice here as previous evidence suggests that survival in cerebral palsy follows a log-logistic distribution (Hutton, Cooke & Pharoah (1994)) and there are very strong similarities between the logit and probit forms (Cox (1966)). However, we can easily consider the log-logistic model itself as well as other parametric distributions (see Section 4.1). We present the log-normal model first to highlight the comparisons with the publication bias sensitivity analysis of Copas & Shi (2001). Here, η is the baseline log-survival (when $z_i = 0$), γ is the effect of the covariate on log-survival, and σ is the cohort variance of the log-survival times. Secondly, we construct a model for the missing data mechanism using a latent variable:

$$M_i = a + bz_i + cT_i' + \omega_i, \qquad \omega_i \sim N(0, 1). \tag{3.2}$$

We can, without loss of generality, state that an individual i has missing data if $m_i > 0$. Assume the residuals (ϵ, ω) are independent and jointly normal with $\text{corr}(\epsilon_i, \omega_i) = 0$.

We must also construct a model for the covariate. As we are using a simple binomial covariate here we can use the model $P(z = 0) = \theta_0 = 1 - \theta_1 = 1 - P(z = 1)$.

This model allows for all three missing data assumptions. For example, if we set $b = 0$ and $c = 0$ then we are assuming data are missing completely at random or, if all parameters, a, b, and c are non-zero then we are allowing the data to be not missing at random. We assume data are missing at random if b is zero. We can have prior beliefs about the values of b and c. 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, $b < 0$. Conversely, data are more likely to be observed if the individual has a longer lifetime which implies that $c > 0$. However, we can use

the likelihood to find estimates for all these parameters. This is further discussed in Section 6.

3.4.1 Comparison to the Meta-Analysis approach

As previously mentioned, this approach is adapted from the approach of Copas & Shi (2001) to the issue of publication bias. Publication bias occurs when a study's conclusions affect its probability of publication and hence, its inclusion in a meta-analysis. Here they assume that the *ith* study in the population 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.

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

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

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

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

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.

Note the similarities with our missing data model. However, we can combine these models in a slightly different way to avoid complex multivariate distributions.

3.5 Construction of the Model Likelihood

The likelihood can now be constructed. We can include right-censoring in our likelihood, but we do not as yet look at left truncation. This is possible, and likelihood contributions would be of a similar form to those in left truncated survival data with full information. We must start by looking at the joint distribution $P(M = m, T' = t', Z = z)$. Note that,

$$P(M = m, T' = t', Z = z) = P(M = m|T' = t, Z = z)P(T' = t'|Z = z)P(Z = z)$$

where

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

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

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

Therefore,

$$P(M=m,T'=t',Z=z)=\frac{1}{2\pi\sigma}\theta_z\exp\left[-\frac{1}{2}\left\{(m-a-bz-ct')^2+\frac{1}{\sigma^2}(t'-\eta-\gamma z)^2\right\}\right].$$

Now that we have the complete joint distribution we can think about the construction of the likelihood. We can divide the data set up into four subgroups based on their censoring indicator and their missing data indicator. The contribution to the likelihood for an individual in each group is then considered.

Group 1) Individual, i, with complete covariate data and failure time, total number of individuals$= n_1$

$$\begin{aligned}
L_1(\eta,\gamma,\sigma,a,b,c,\theta|t_i',z_i) =& P(M<0,T'=t_i',Z=z_i)\\
=&\int_{-\infty}^{0}P(M=m,T'=t_i',Z=z_i)\,dm\\
=&\int_{-\infty}^{0}\frac{1}{2\pi\sigma}\theta_{z_i}\exp\left[-\frac{1}{2}\left\{(m-a-bz_i-ct_i')^2+\frac{1}{\sigma^2}(t_i'-\eta-\gamma z_i)^2\right\}\right]\,\mathrm{dm}\\
=&\frac{1}{\sigma\sqrt{2\pi}}\theta_{z_i}\exp\left\{-\frac{1}{2\sigma^2}(t_i'-\eta-\gamma z_i)^2\right\}\\
&\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(m-a-bz_i-ct_i')^2\right\}\,\mathrm{dm}\\
=&\frac{1}{\sigma\sqrt{2\pi}}\theta_{z_i}\exp\left\{-\frac{1}{2\sigma^2}(t_i'-\eta-\gamma z_i)^2\right\}\Phi(-a-bz_i-ct_i').
\end{aligned}$$

This is the contribution from each individual in this group. Therefore, the group contribution is

$$\prod_{\binom{i:m_i<0}{\delta_i=0}}L_1(\eta,\gamma,\sigma,a,b,c,\theta|t_i',z_i)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n_1}\prod_{\binom{i:m_i<0}{\delta_i=0}}\theta_{z_i}\exp\left\{-\frac{1}{2\sigma^2}(t_i'-\eta-\gamma z_i)^2\right\}\Phi(-a-bz_i-ct_i').$$

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

$$\begin{aligned}
L_2(\eta,\gamma,\sigma,a,b,c,\theta|t_i',z_i) &= P(M<0,T'>t_i',Z=z_i)\\
&=\int_{t_i'}^{\infty}P(M<0,T'=u,Z=z_i)\,\mathrm{du}\\
&=\frac{1}{\sigma\sqrt{2\pi}}\theta_{z_i}\int_{t_i'}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}(u-\eta-\gamma z_i)^2\right\}\Phi(-a-bz_i-cu)\,\mathrm{du}.
\end{aligned}$$

This integration can be evaluated using numerical Gaussian quadrature methods. These will be further discussed in Section 6.1. 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,

$$\begin{aligned} L_3(\eta,\gamma,\sigma,a,b,c,\theta|t_i') &= P(M>0,T'=t_i') \\ &= \sum_{z=0}^{1} P(M>0,T'=t_i'|Z=z)P(Z=z) \\ &= \sum_{z=0}^{1}\left(\int_0^{\infty} P(M=m,T'=t_i'|Z=z)P(Z=z)\ \mathrm{dm}\right) \\ &= \sum_{z=0}^{1}\frac{1}{\sigma\sqrt{2\pi}}\theta_z \exp\left\{-\frac{1}{2\sigma^2}(t_i'-\eta-\gamma z)^2\right\}\Phi(a+bz+ct_i') \end{aligned}$$

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 log-likelihood contribution...

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

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. These are implemented using the *nlminb* function within S-Plus.

4 Alternative survival distributions

The previous section, Section 3, 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.

Changing the 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.

4.1 The log-logistic distribution

The log-logistic has proved to be useful when modelling 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,

$$T_i' = \log(T_i) = \eta + \gamma z_i + \sigma\epsilon_i, \qquad \epsilon_i \sim \log(0,1), \qquad i = 1,...,n.$$

Similarly, we construct the missing data mechanism model as

$$M_i = a + bz_i + ct_i' + \omega_i, \qquad \omega_i \sim \log(0,1).$$

Note that now the errors have logistic distributions. This means that the distribution of t_i, given z_i, is log-logistic with mean $\eta + \gamma z_i$ 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

$$P(M = m, T' = t', Z = z) = \frac{\exp\left\{-\left(m - a - bz - ct'\right)\right\} \exp\left\{-\left(t' - \eta - \gamma z\right)/\sigma\right\} \theta_z}{\sigma\left(1 + \exp\left\{-\left(m - a - bz - ct'\right)\right\}\right)^2 \left(1 + \exp\left\{-\left(t' - \eta - \gamma z\right)/\sigma\right\}\right)^2}.$$

The formulation of the likelihood can continue in a similar fashion to that shown in Section 3. 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.

$$\begin{aligned}
l(\eta, \gamma, \sigma, a, b, c, \theta | t_i', m_i, z_i) = & \sum_{\substack{(i:m_i<0) \\ \delta_i=1}} \log\left[\frac{\theta_{z_i} \exp\left\{-\left(t_i' - \eta - \gamma z_i/\sigma\right)\right\}}{\left[1 + \exp\left\{-\left(t_i' - \eta - \gamma z_i\right)/\sigma\right\}\right]^2} \left\{\frac{1}{1 + \exp\left(a + bz_i + ct_i'\right)}\right\}\right] \\
& + \sum_{\substack{(i:m_i<0) \\ \delta_i=0}} \log\left[\int_{t'}^{\infty} \frac{\theta_{z_i} \exp\left\{-\left(u_i - \eta - \gamma z_i/\sigma\right)\right\}}{\left[1 + \exp\left\{-\left(u_i - \eta - \gamma z_i/\sigma\right)\right\}\right]^2} \left\{\frac{1}{1 + \exp\left(a + bz_i + cu_i'\right)}\right\} \mathrm{du}\right] \\
& + \sum_{\substack{(i:m_i>0) \\ \delta_i=1}} \log\left[\sum_{z=0}^{1} \frac{\theta_z \exp\left\{-\left(t_i' - \eta - \gamma z/\sigma\right)\right\}}{\left[1 + \exp\left\{-\left(t_i' - \eta - \gamma z/\sigma\right)\right\}\right]^2} \left\{\frac{\exp\left\{a + b + ct_i'\right\}}{1 + \exp\left(a + b + ct_i'\right)}\right\}\right] \\
& + \sum_{\substack{(i:m_i>0) \\ \delta_i=0}} \log\left[\sum_{z=0}^{1} \int_{t'}^{\infty} \frac{\theta_z \exp\left\{-\left(u_i - \eta - \gamma z/\sigma\right)\right\}}{\left[1 + \exp\left\{-\left(u_i - \eta - \gamma z/\sigma\right)\right\}\right]^2} \left\{\frac{\exp\left(a + bz + cu_i\right)}{1 + \exp\left(a + bz + cu_i\right)}\right\} \mathrm{du}\right].
\end{aligned}$$

4.2 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

$$f(\epsilon) = \exp\left\{\epsilon - e^{\epsilon}\right\}.$$

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

$$P(M = m, T' = t', Z = z) = \frac{1}{\sigma} \exp\left\{m - a - bz - c_t' + \frac{t' - \eta - \gamma z}{\sigma} - e^{m-a-b-ct'} - e^{\frac{t'-\eta-\gamma z}{\sigma}}\right\}.$$

Using the same methodology as previously we can calculate the log-likelihood...

$$
\begin{aligned}
l(\eta,\gamma,\sigma,a,b,c,\theta|t_i',m_i,z_i) = & \sum_{\substack{(i:m_i<0)\\ \delta_i=1}} \log\left[\frac{\theta_{z_i}}{\sigma}\exp\left(\frac{t_i'-\eta-\gamma z_i}{\sigma} - e^{\frac{t_i'-\eta-\gamma z_i}{\sigma}}\right)\left\{1-\exp\left(-e^{-a-bz_i-ct_i'}\right)\right\}\right] \\
& + \sum_{\substack{(i:m_i<0)\\ \delta_i=0}} \log\left[\int_{t_i'}^{\infty}\frac{\theta_{z_i}}{\sigma}\exp\left(\frac{u_i-\eta-\gamma z_i}{\sigma} - e^{\frac{u_i-\eta-\gamma z_i}{\sigma}}\right)\left\{1-\exp\left(-e^{-a-bz_i-cu}\right)\right\}\,\mathrm{du}\right] \\
& + \sum_{\substack{(i:m_i>0)\\ \delta_i=1}} \log\left\{\sum_{z=0}^{1}\frac{\theta_z}{\sigma}\exp\left(\frac{t_i'-\eta-\gamma z}{\sigma} - e^{\frac{t_i'-\eta-\gamma z}{\sigma}} - e^{-a-bz-ct_i'}\right)\right\} \\
& + \sum_{\substack{(i:m_i>0)\\ \delta_i=0}} \log\left\{\sum_{z=0}^{1}\int_{t'}^{\infty}\frac{\theta_z}{\sigma}\exp\left(\frac{u_i-\eta-\gamma z}{\sigma} - e^{\frac{u_i-\eta-\gamma z}{\sigma}} - e^{-a-bz-cu_i}\right)\,\mathrm{du}\right\}.
\end{aligned}
$$

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. 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.

5 Example - Bristol Cerebral Palsy Data

As discussed in Section 2 one of our main motivations for this work is a large data set from Bristol, UK which looks at the survival of children diagnosed with cerebral palsy. This data set has a large amount of missing covariate data. Previous research suggests that the level of ambulatory disability has one of the most significant effects on survival time. Indeed, the largest effect seems to come from a distinction between not severe and severe level disability. Approximately, one sixth of the information on this binary variable is missing. A Kaplan-Meier estimate of survival by ambulatory group is given in Figure 1.

From Figure 1 we can see that severe ambulation has a large negative effect on survival except during approximately the first five years. This may be because those most at risk in the severe disability group do not survive until diagnosis. Those with a missing disability level have considerably lower survival in the first seven years. This is probably because those with short lifetimes failed or were censored before measurements could be taken. It is also probable that these were the most severely disabled because then it is harder to measure the level of disability.

Table 1 compares survival model estimates from our model using a missing data mechanism to those of a complete case analysis (when individuals with missing data are ignored in analysis). Complete case analysis assumes MCAR. We consider a range of possible survival distributions.

If we look at Table 1 we can see that regardless of the survival distribution used a complete case analysis noticeably underestimates the effect of a severe level of ambulatory. This obviously leads to consistently different fitted survival models.

We can also use the maximum likelihood to consider the most appropriate parametric model the log-logistic or Weibull distribution suggesting that there is little difference between the two models. Hutton & Monaghan (2002) discuss the choice of parametric models in survival analysis. They compare theoretical parameter values for fitted Weibull

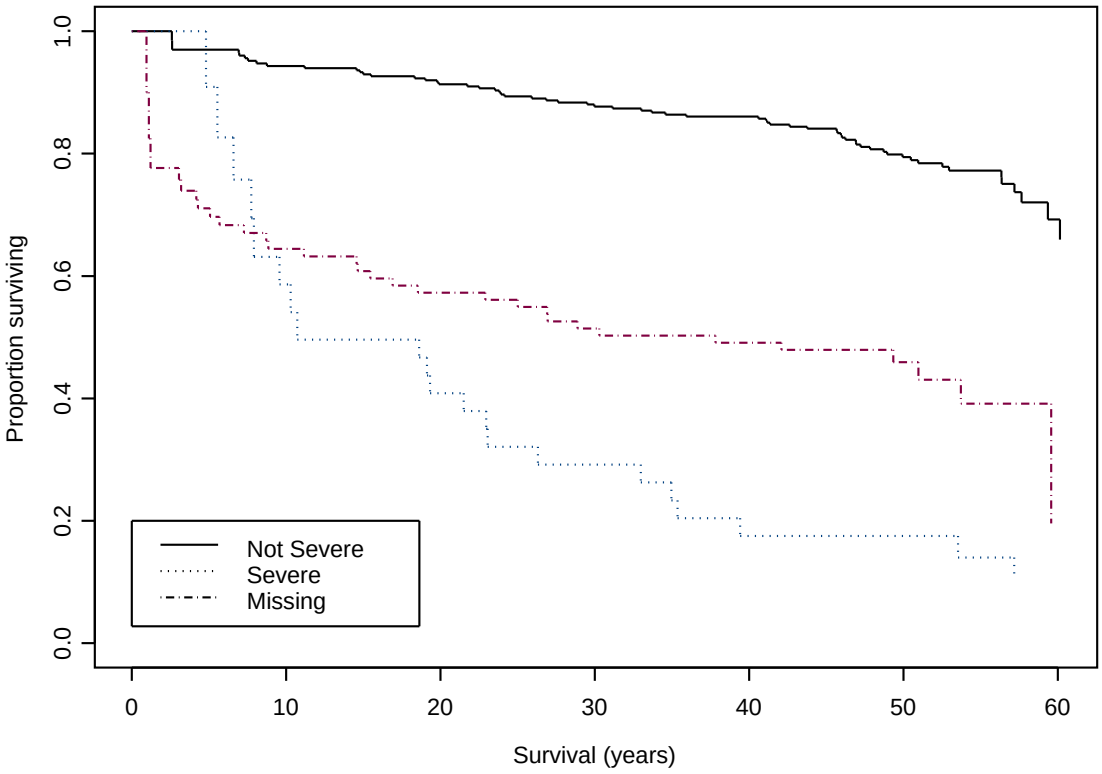


Figure 1: Non-parametric survival of subjects with cerebral palsy with relation to level of ambulation.

and log-logistic models and highlight that the shape of the log-logistic hazard, which switches from being monotonic to non-monotonic, can lead to model misspecification.

6 Considering the Likelihood

Before we continue with this work there are issues within the likelihood concerning its practical implementation that must be discussed. Consider here only the log-normal model. Discussions hold for the other distributions discussed.

6.1 Numerical integration and implementation

As discussed, the one of the main complication with implementing the model is the numerical integration required for the likelihood. Most methods for numeric integration, or quadrature, are based on the idea of adding up the value of the integrand at a sequence of abscissas within the range of the integration. One of the simplest of these is Simpson's rule. Romberg integration generalizes this to a higher order. In these the integral is approximated by the sum of its functional values at a set of equally spaced points, multiplied by suitably chosen weighting coefficients. Gaussian quadrature allows us the freedom of choosing the abscissas as well as the weighting coefficients. Thus we have twice the number of degrees of freedom at our disposal. One additional feature is that we can choose the weights and abscissas to make the integral exact for integrals of the form "some known function W multiplied by a polynomial". We implement 10-point Gaussian quadrature (Faires & Burden 2003) in S-Plus using tabulated weights and abscissas.

Model	η - baseline survival	γ - disability effect	σ - dispersion	θ - % severe	Likelihood
Log-normal					
Missing data	10.9	-2.3	1.3	11.9	-585.0
Complete case	9.9	-1.0	1.3	8.8	
Log-logistic					
Missing data	10.8	-2.1	0.7	10.7	-583.0
Complete case	9.8	-0.9	0.6	8.8	
Weibull					
Missing data	11.0	-1.7	0.9	7.4	-583.0
Complete case	10.2	-0.8	0.7	8.8	
Exponential					
Missing data	11.1	-1.8	1.0	7.0	-584.1
Complete case	10.4	-1.0	1.0	8.8	

Table 1: Comparison of missing data model and complete case analysis maximum likelihood estimates for cerebral palsy patients in Bristol.

6.2 Estimating the missing data mechanism

Our model is also able to find maximum likelihood estimates for the parameters (a, b, c) in the missing data mechanism. In order to consider whether we may rely on these estimates we can consider the profile likelihoods based upon the maximum likelihood estimates for the Bristol data analysis. These profile likelihoods are shown in Figure 2. We can see that all have an obvious peak so we should be able to identify the missing data mechanism parameters from our model.

We can also consider the missing data mechanism parameter maximum estimates. Referring back to Equation 3.2, the positive b shows that an increase in the covariate causes an increase in m i.e. those with a severe level of ambulatory disability have a higher probability of missing covariate data. The negative c shows that those with shorter observed lifetimes also have a higher probability of missing covariate data. This agrees with what we might expect.

These estimates also agree with the differences that we observe between the survival model estimates for our joint model and complete case analysis results. The estimated missing data mechanism suggests that those with the severest disability levels and shortest survival times are more likely to have missing data data. Hence, a complete case analysis is more likely to miss these people. This causes the decrease in the covariate effect in the complete case survival model that we observe in Table 1.

7 Simulation

Now that we have derived our model we must consider its robustness. We must develop a simulation study to consider the survival model estimates that the model arrives at under various situations. We can then compare these estimates to those from other missing data methods and, also, estimates with complete data. The main situations to consider are a change in the missing data mechanism and a change in the percentage of missing data. We try to consider data similar to our Bristol cerebral palsy data to test the validity of the results in Section 5.

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

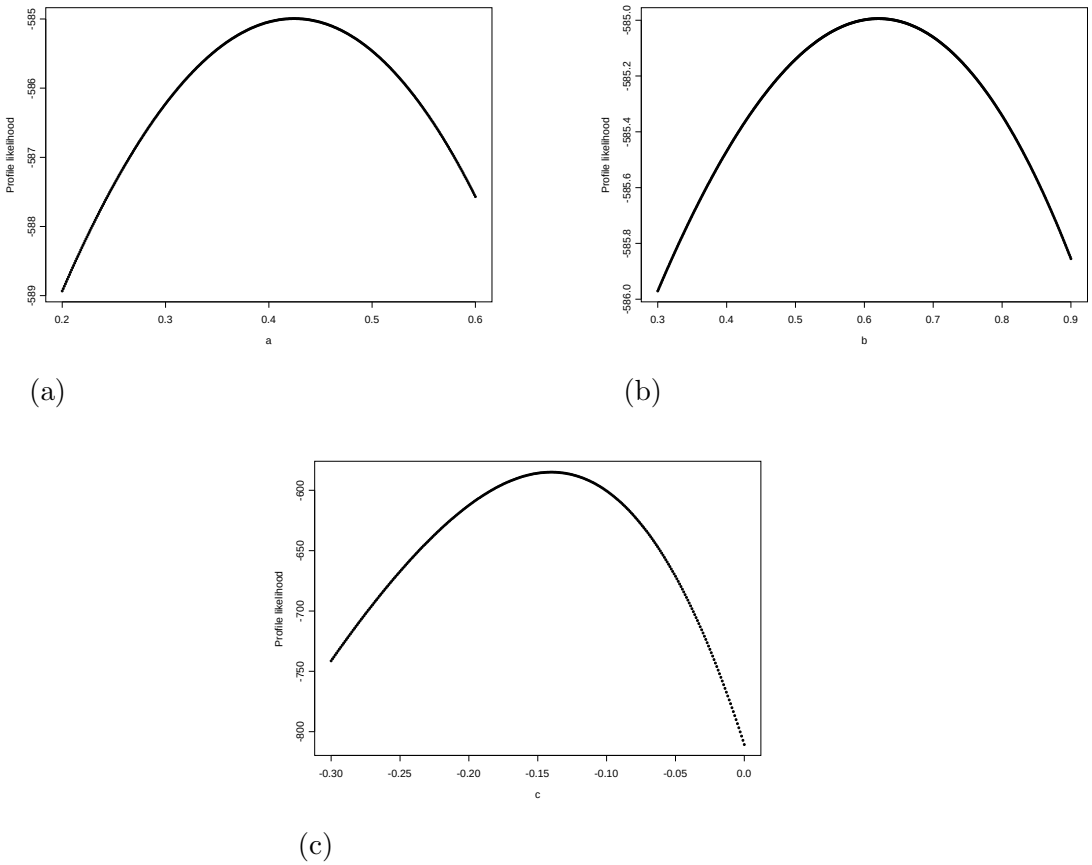


Figure 2: Profile likelihood for estimation of missing data mechanism parameters (a, b, and c respectively) within the log-normal model.

construction of the missing data mechanism.

Firstly, we create a vector of length n based on realizations from a Bernoulli distribution with probability θ. This gives us the true covariate values.

Secondly, we must construct the true survival times. This is done using the inverse probability transform.

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

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

$\Rightarrow$ *A real valued random variable X with distribution function $F(x) = P(X \leq x)$ 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, γ, and σ.

We impose a censoring distribution similar to that which we believe applies to the Bristol data i.e. independent uniform censoring on the interval [15 years, 65 years]. This is because censoring is mainly due to the end of the study and not due to lost follow-up. We have data up to 15 years following the last recruitment and the highest survival time in the data is approximately 65 years. We can then calculate the observed survival and censoring indicator. The enforced censoring mechanism led to approximately 80% censoring, similar to the level found in the Bristol data.

Again using the inverse probability transform we can construct the latent missing data mechanism variable. We can then force the covariate data to be missing according to the value of this.

This produces the full simulated data set and is programmed in S-plus.

7.1 Simulation 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 data set close in structure to the Bristol data as we wish to consider the reliability of the results that we obtained in Section 5. In order to do this we settled upon survival data with parameters $(\eta, \gamma, \sigma, \theta) = (11, -2.3, 1.3, 0.2)$ and subjected the simulations to four different missing data mechanisms. These mechanisms were defined by the values of (b, c) with a chosen to achieve approximately the right proportion of missing data. The four mechanisms presented here are

MCAR - b = 0, c = 0, 20% missing data
MAR - b = 0, c = -0.2, 20% missing data
NMAR - b = 0.2, c = -0.2, 20% missing data
NMAR - b = 0.2, c = -0.2, 60% missing data.

For each survival distribution previously discussed and each mechanism we then simulated 100 data sets (using methods discussed in Section 7) and compared parameter estimates from our model with complete case estimates and the true estimates (based on the true data) using the survReg function within S-Plus. We also considered log-normal survival with no censoring applied to the survival times. Each data set consisted of 400 individuals and there was one binary covariate. Results are presented in Appendix A-Appendix D.

7.2 Conclusion from the simulation study

Studying the boxplots in Appendix A-Appendix D we can discuss the reliability of our model. There are several things to note. Firstly, we consider the results when the data are MCAR. We can see that the distributions of parameter estimates are similar for our model compared to that of the complete case analysis and the 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 of missing data methods. We can look at the distributions of results as the missing data mechanism tends to NMAR and the proportion of missing data increases. We see that the complete case estimates tend to shift away from the true parameters while the model based estimates stay much more consistent. We know that complete case analysis can 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 the level of missing data rises to 60% the complete case estimates are considerably biased but our model remains more reliable. This is particularly obvious in the estimation of the dispersion parameter where complete case analyses vastly underestimate the magnitude. The bias within the complete case estimates is in the expected direction as we expect to miss the most severely disabled and hence the lowest survival times which would increase survival in general and decrease the effect of a severe level of ambulation.

Secondly, we can look at the range of the parameter estimates. While the model estimates are obviously more variable than the estimates assuming fully known data they are generally of a similar magnitude. This perhaps fails to be true when the level of missing data increases to 60% but this is unsurprising as we have much less information

upon which to base results. The range decreases when there is no censoring (note the change in the y-axes in Appendix D) but the problems at large quantities of missing data still arise.

It is also useful to note that our model works equally consistently over all the survival distributions. Results from the exponential distribution are not presented here but are also consistent.

These findings are encouraging and suggest that we may be reasonably trusting of the results regarding the Bristol data that we found in Section 5.

However, the simulation study highlighted the problems that occur with the implementation of the model. Mainly, this concerns the numerical integration required in the maximisation of the likelihood. A large number of points are needed in the Gaussian Quadrature and this significantly increases the time required to do a simulation study of this size.

8 Extensions to Categorical and Continuous Covariates and the Multivariate model including Left Truncation

Until now we have only considered the use of a single binary covariate. However, there is no reason why we must restrict ourselves to this. Within our Bristol data we have information on a large number of both categorical and continuous covariates which we may wish to investigate. We must consider how to include these different forms into our model.

We start by considering categorical covariates. This requires only a simple extension to our original likelihoods. In the case of a single binary covariate the information contributed to the likelihood was of the form $P(Z = 1) = \theta = 1 - P(Z = 0)$. For categorical covariate with n categories this can be extended to $P(Z = z) = \theta_z$ such that $\sum_{z=0}^{n} \theta_z = 1$. This poses no complication in the construction of the likelihood but will increase computation time.

In a multivariate setting we start by defining a multinomial distribution for the categorical variables denoted by $Z = (Z_1, ..., Z_p)'$. The p potentially observed covariates form a contingency table with $R = D_1 \times D_2 \times ... \times D_p$ cells where D_j is the number of levels of covariate Z_j. We can then number these cells. The probability assigned to cell r is denoted as θ_r such that $\sum_{r=1}^{R} \theta_r = 1$. If the covariate is observed we know exactly which cell the individual falls into, and the likelihood contribution can be calculated as before, but with missing data we only observe a set of possible cells based on the observed observation. Therefore, in calculating the contribution to the likelihood of an individual with missing information on some or all of the covariates we must sum over the possible cells in which they could lie. Let $W_i = (W_{i1}, ..., W_{iR})$ be a binary vector of length R indicating the cells in which an individual may be. If we know individual i falls in cell q then the qth element of W_i equals 1 and all other entries equal 0. Similarly, if the individual has missing data and we know only a range of possible cells then these cells are indicated in W_i by the value 1 and all others by 0. We can then calculate the contribution for the likelihood for an individual i with missing data as

$$L(\eta, \gamma, \sigma, a, b, c, \theta) = \sum_{r=1}^{R} W_{ir}\theta_r P(M > m | T' = t, Z = z_r) P(T' = t' | Z = z_r) \qquad (8.1)$$

if that individual fails and

$$L(\eta, \gamma, \sigma, a, b, c, \theta) = \sum_{r=1}^{R} W_{ir}\theta_r P(M > m | T' > t, Z = z_r) P(T' > t' | Z = z_r) \qquad (8.2)$$

if they are censored where z_r indicates the set of covariate values that map to cell r. We may believe that correlations exist between the covariates, in which case constraints may be placed upon θ.

To include continuous covariates we need to use their probability distribution function. This is obviously not known exactly but must be modeled. The parameter values will be estimated by the full joint model but a distribution form must be decided upon. This distribution (which may be multivariate) can then be placed over the discrete multinomial distribution so that we have a separate distribution in each multinomial cell. These distributions can be restricted to all be equal. This is useful when there are few individuals in many cells and will decrease computing time and difficulty.

8.1 Including Left Truncation

One of the main issues with our motivating data is the presence of left truncation. Left truncation arises when lifetimes of individuals who fail before a certain time are not recorded. This truncation occurs because patients were not recorded in the data until their first appointment with Dr Woods. Up until now we have ignored this complication but it can be shown that we can include this into our model.

Recall the earlier notation. Left truncation occurs if an individual enters a study at time $t_i^0 < w_i < t_i$. This means that those individuals who experience an event before time w_i are not recorded in the data.

In order to calculate the changed likelihood we again consider the full joint distribution. However, now our data is dependent upon survival until time w. Therefore,

$$P(M = m, T' = t, Z = z | T' > w) = \frac{P(M = m, T' = t, Z = z, T' > w}{T' > w} = \frac{P(M = m, T' = t, Z = z)}{P(T' > w)}.$$

Therefore, we condition on $T' > w$ by dividing by the survival function of T'. Note that...

$$P(T' > w) = \sum_z P(T' > w | Z = z) P(Z = z).$$

9 Discussion

We have presented here a method for including a model of the missing covariate data mechanism into a parametric survival analysis. Much work had been previously done in this field but little applied to the problem of data that was not missing at random. This was a potential concern in our motivating cerebral palsy example.

We have discussed the application of the theory to a wide variety of parametric survival models and extended initial models to allow for various types of data.

Simulation results are encouraging and suggest that our model is useful in approaching survival analysis problems with missing covariate data. They lead us to believe that estimates based upon complete case analysis for the effect of severe ambulation on the survival of individuals with cerebral palsy significantly underestimate the magnitude. This has a large impact on our conclusions from this data although there is still much more detailed analysis to be done.

There is still much left to investigate with this method. We need to do a much more in depth analysis of our example data, including allowing for left truncation, multiple covariates, and informative left truncation.

There are also more methodological issues. Firstly, how does our method compare with multiple imputation results if the MAR assumption holds. MI is used widely in the

literature and in practice and it is important that our model is not inferior to this idea. We can use our model to explicitly test for MAR or MCAR by checking the significance of the missing data mechanism model parameters.

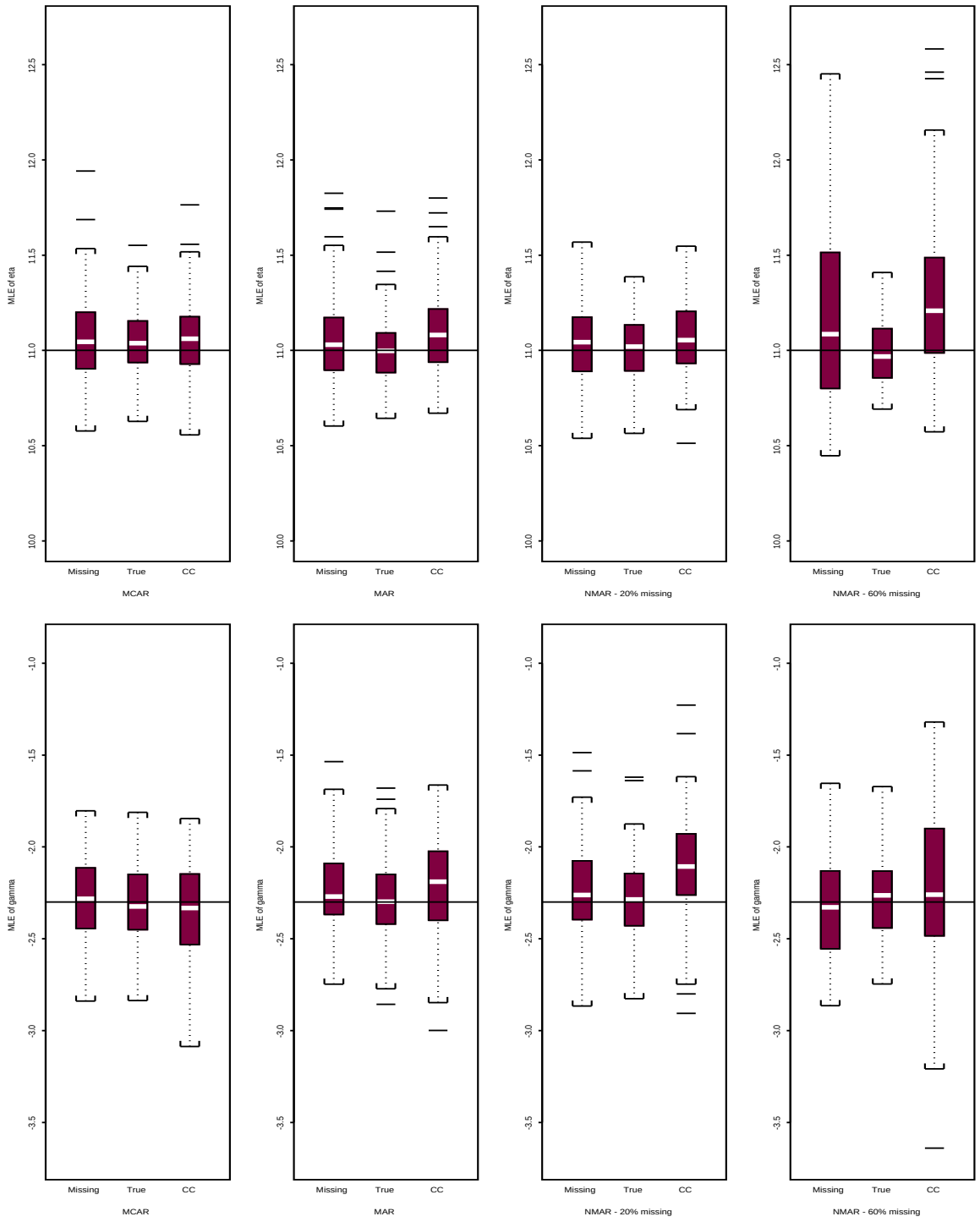
Secondly, how might we go about conducting a model selection. This is well established for complete data but choosing covariates to including in models for prediction and also the choice of survival distribution becomes less clear when there is a proportion of missing data, particularly if this proportion is large.

In Section 6.2 we discussed the profile likelihoods for the parameters in the missing mechanism in our example. However, from our simulations we have information on the estimated parameters and it would be interesting to compare these to the values used for simulation and see if we can still use the model to estimate them.

Finally, the Cox proportional hazards model (Cox 1972) is used extensively in survival analysis and it may be interesting to see if we can use our ideas in this semi-parametric framework.

10 Appendix

A Simulation with the log-normal distribution



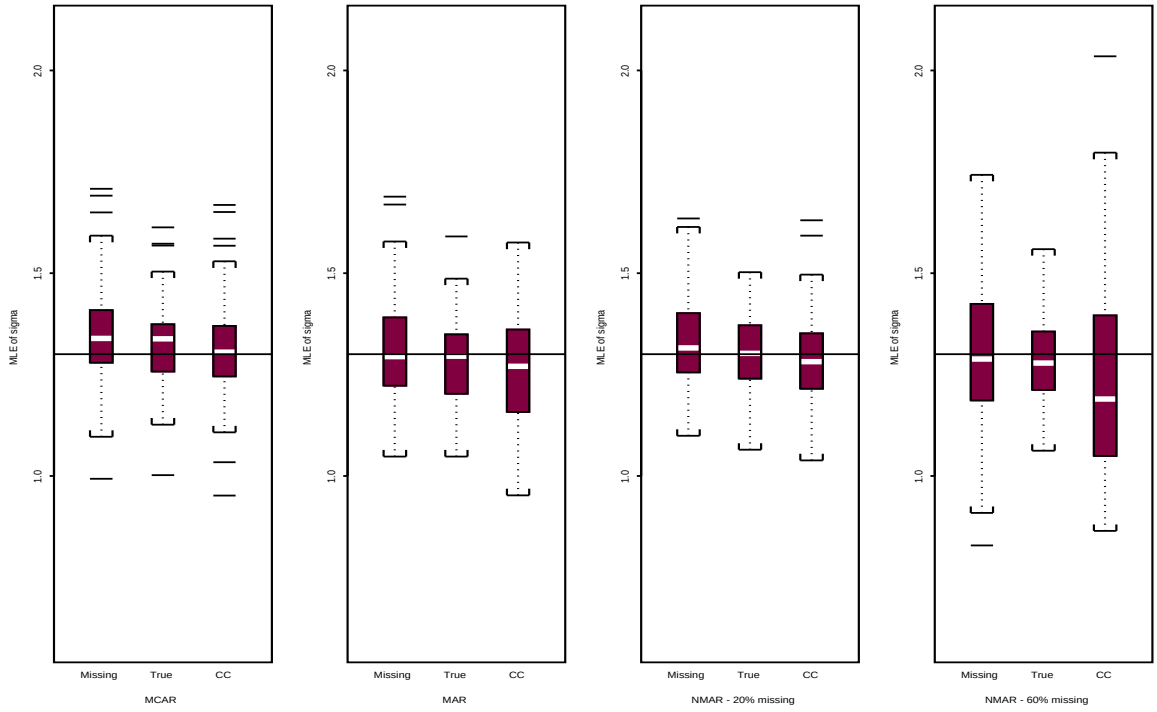
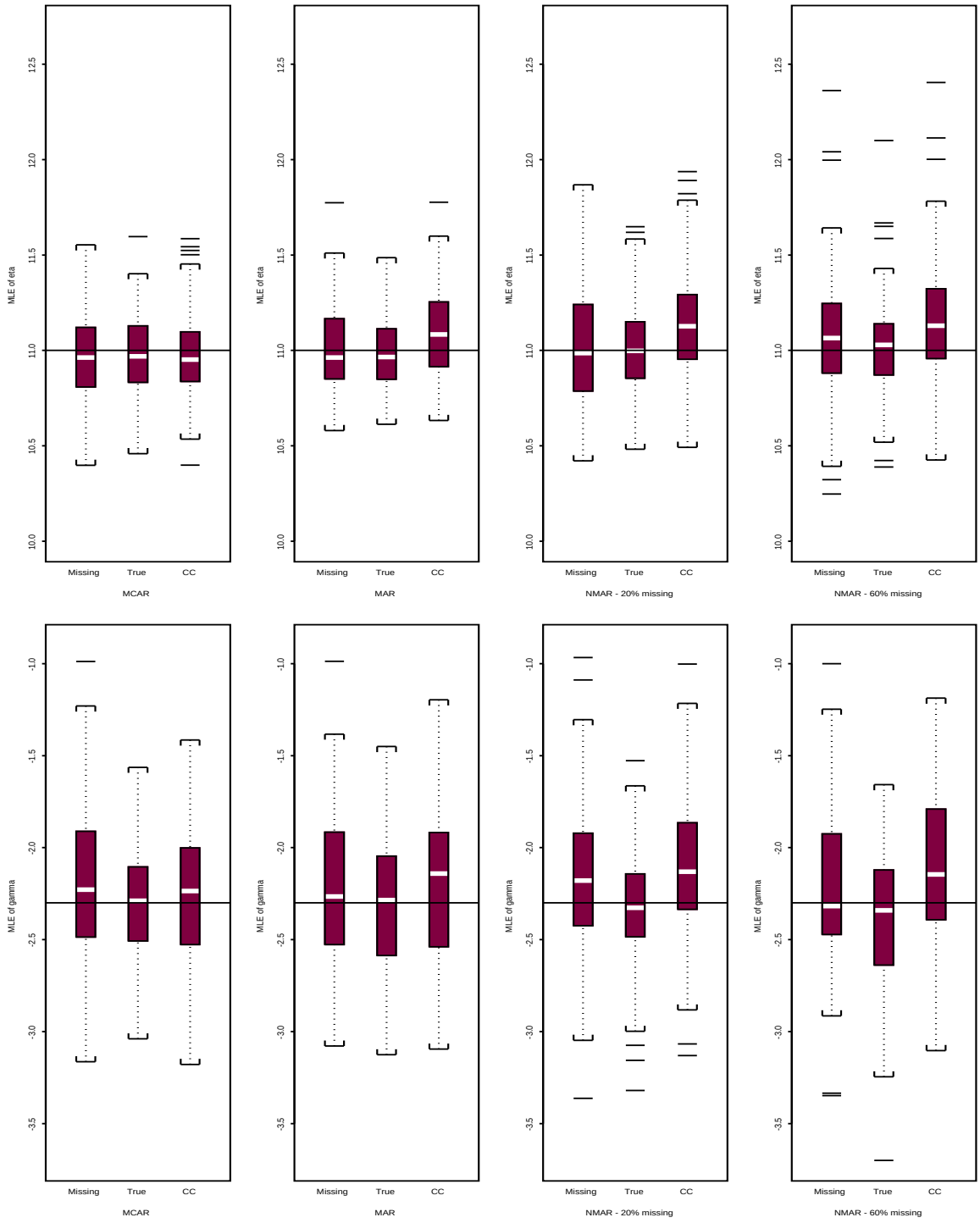


Figure 3: Parameter estimates for the log-normal survival model with four different missing data mechanisms based on 100 simulated data sets each (true parameters ——).

B Simulation with the log-logistic distribution



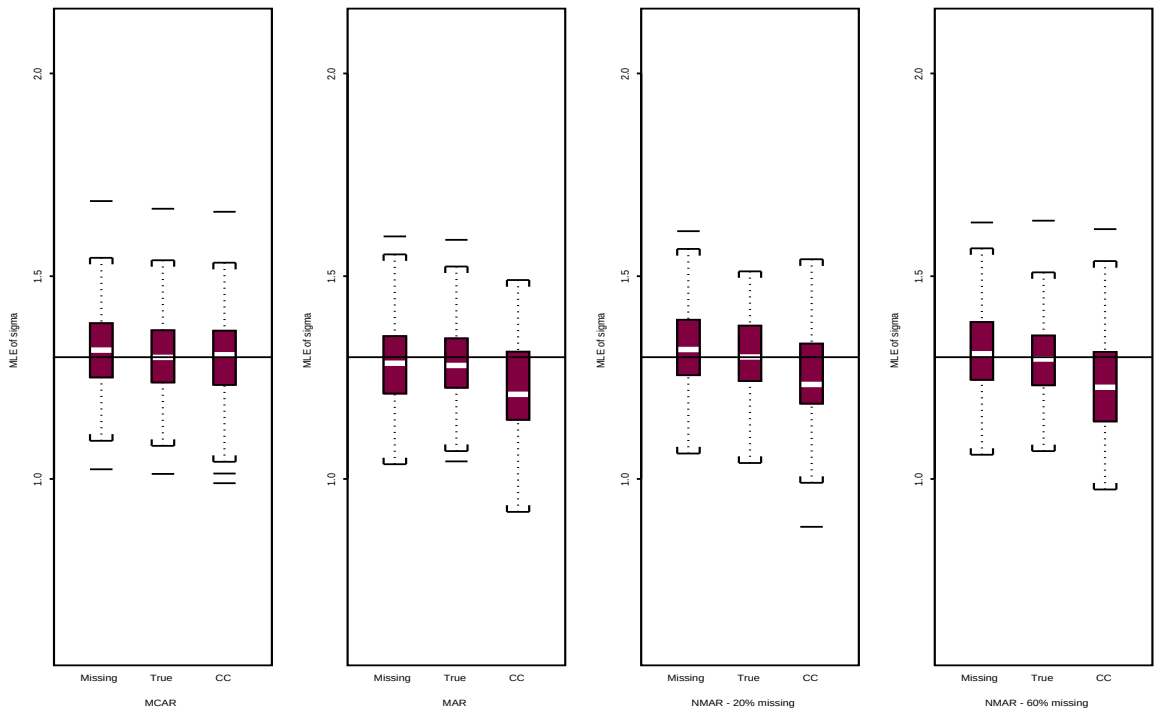
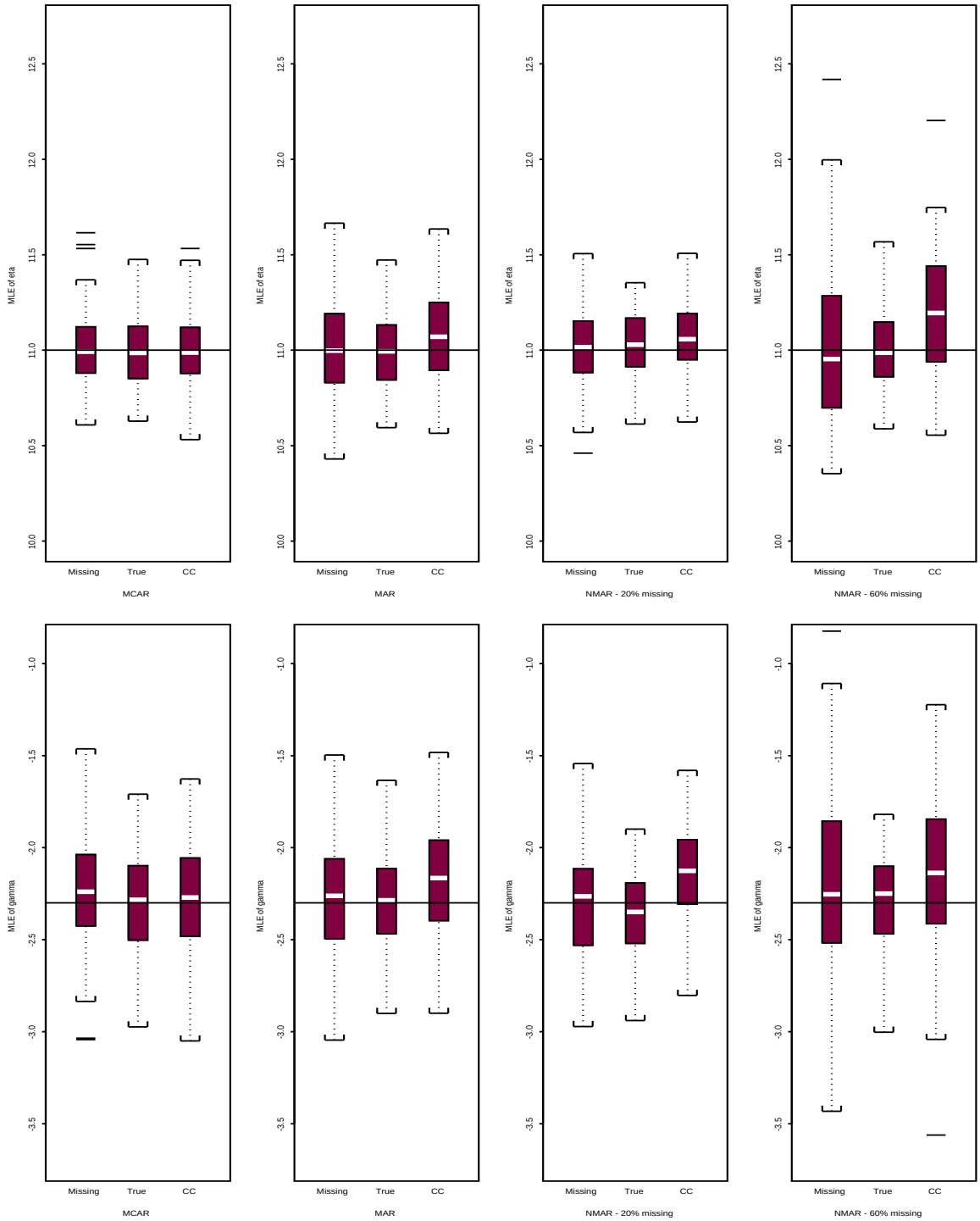


Figure 4: Parameter estimates for the log-logistic survival model with four different missing data mechanisms based on 100 simulated data sets each (true parameters ——).

C Simulation with the Weibull distribution



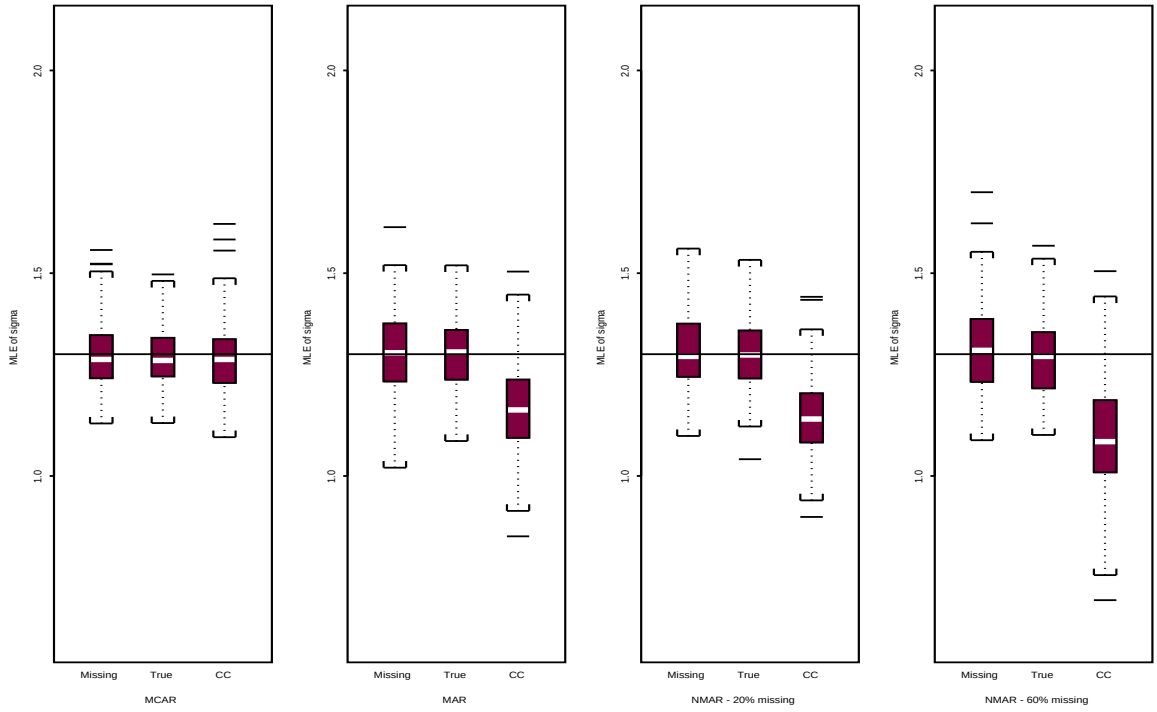
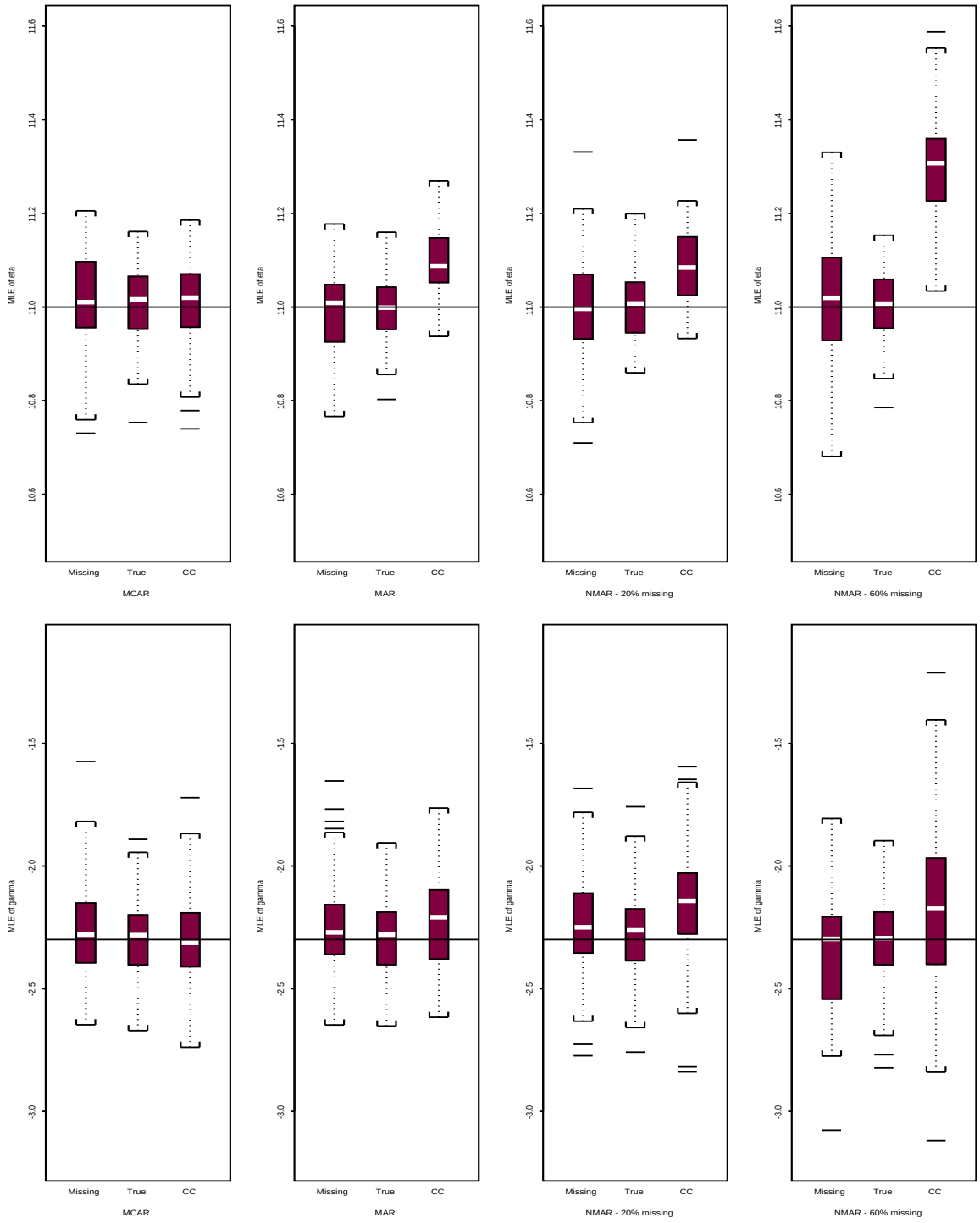


Figure 5: Parameter estimates for the Weibull survival model with four different missing data mechanisms based on 100 simulated data sets each (true parameters ——).

D Simulation with the log-normal model and no censoring



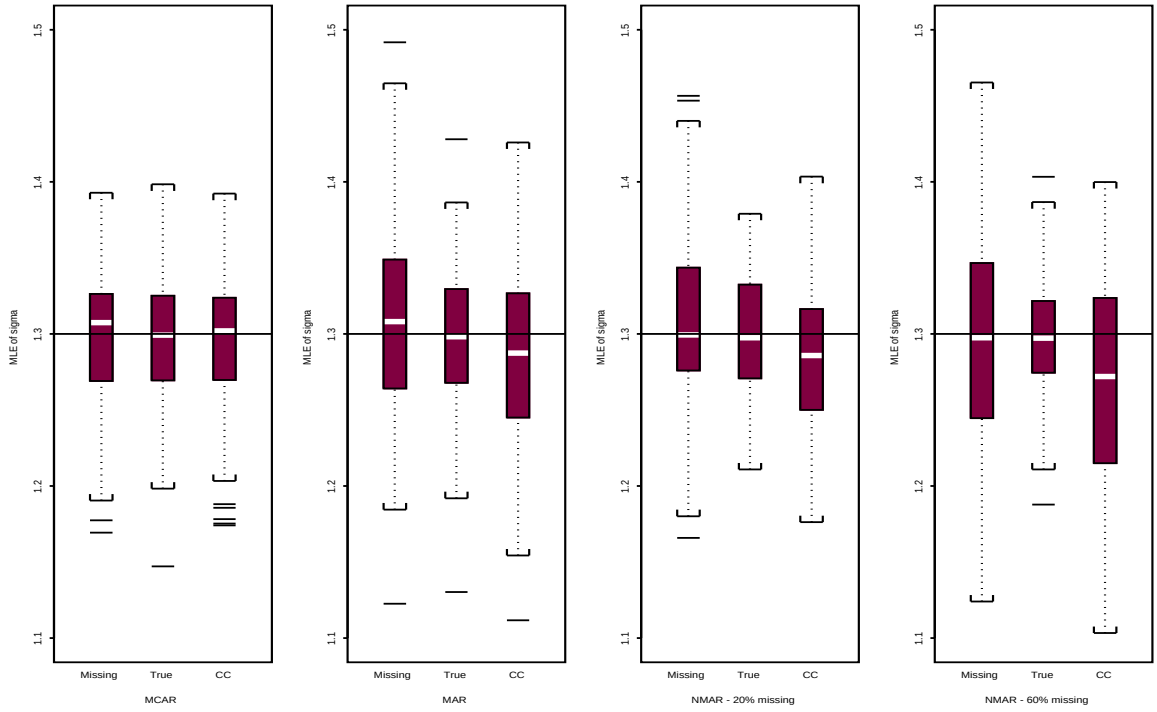


Figure 6: Parameter estimates for the log-normal survival model with four different missing data mechanisms and no censoring based on 100 simulated data sets each (true parameters ——).

References

Collett, D. (1999), *Modelling Survival Data in Medical Research*, Chapman & Hall / CRC, London.

Copas, J. & Shi, J. (2001), 'A sensitivity analysis for publicatin bias in systematic reviews', *Statistical Methods in Medical Research* **10**, 251–265.

Cox, D. (1966), *Festschrift for J. Neyman - Research Papers in Statistics*, Wiley, chapter 4, pp. 55–71.

Cox, D. (1972), 'Regression models and life tables', *Journal of the Royal Statistical Society, B* **34**, 187–220.

Faires, D. & Burden, R. (2003), *Numerical methods*, 3 edn, Brooks/Cole, chapter 4.

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.

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.

Ibrahim, J., Chen, M.-H. & Lipsitz, S. (1999), 'Monte carlo em for missing covariates in parametric regression models', *Biometrics* **591-596**, 55.

Lipsitz, S. & Ibrahim, J. (1998), 'Estimating equations with incomplete categorical covariates in the cox model', *Biometrics* **54**, 1002–13.

Little, R. & Rubin, D. (2002), *Statistical Analysis with Missing Data*, John Wiley and Sons, Inc., New York.

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.

Woods, G. (1957), Cerebral Palsy in childhood: The aetiology and clinica assessment with particular reference to findings in Bristol, PhD thesis, University of Bristol, Bristol, UK.