Estimation with correlated censored survival data with missing covariates

STUART R. LIPSITZ
Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute,
44 Binney Street, Boston, MA 02115, USA

and

JOSEPH G. IBRAHIM*
Department of Biometry and Epidemiology, Medical University of South Carolina, Charleston, SC, USA

SUMMARY

Incomplete covariate data are a common occurrence in studies in which the outcome is survival time. Further, studies in the health sciences often give rise to correlated, possibly censored, survival data. With no missing covariate data, if the marginal distributions of the correlated survival times follow a given parametric model, then the estimates using the maximum likelihood estimating equations, naively treating the correlated survival times as independent, give consistent estimates of the relative risk parameters (Lipsitz et al., 1994). Now, suppose that some observations within a cluster have some missing covariates. We show in this paper that if one naively treats observations within a cluster as independent, that one can still use the maximum likelihood estimating equations to obtain consistent estimates of the relative risk parameters. This method requires the estimation of the parameters of the distribution of the covariates. We present results from a clinical trial (Lipsitz and Ibrahim, 1996b) with five covariates, four of which have some missing values. In the trial, the clusters are the hospitals in which the patients were treated.

Keywords: Clustered survival data; Score vector.

1. INTRODUCTION

Clustered survival data arise often in biomedical studies; observations within the cluster tend to be correlated. In this paper, the data consists of $N$ independent clusters; the response vector for each cluster is a vector of correlated, possibly censored, survival times. The example in this paper is from two Eastern Cooperative Oncology Group (ECOG) clinical trials to evaluate patients with primary liver cancer (Falkson et al., 1990, 1994). The primary interest here is how the outcome survival, time from entry on the study until death, differs with respect to five dichotomous baseline covariates. The five covariates are age, categorized as less than 60 or greater than or equal to 60; associated jaundice, yes or no; associated hepatitis, yes or no; and two biochemical markers, alpha fetoprotein and anti-hepatitis B antigen, each classified as normal or abnormal.

*To whom correspondence should be addressed

© Oxford University Press (2000)
In many clinical trials, including those undertaken by large cooperative cancer groups (CALGB, ECOG, SWOG, EORTC), patients enter the trial from an institution (i.e. hospital). Not infrequently, patients from the same institution have similar outcomes (i.e. outcomes which are correlated), due, possibly, to unmeasured variables such as the skill or training of the staff or the quality of the hospital equipment. Institutions can then be thought of as clusters, and patients as the units within these clusters. In the liver cancer study, the data include 190 observations from \( N = 31 \) institutions. Table 1 shows the data from eight of these clusters.

In Table 1, we see that there are missing covariate data. Table 2 shows the percent missing (over the 190 observations) on each covariate. Note that 52 (27\%) of the 190 observations are missing at least one variable. The biochemical markers require blood to be drawn and possibly complicated laboratory tests, and thus have the highest proportion missing. A previous analysis of this data (Lipsitz and Ibrahim, 1996b) assumed the failure time followed a Weibull distribution and ignored the cluster effect, but did account for the missing data using the EM algorithm assuming ignorable missingness. We assume, for a given member of a cluster, that the probability that data are missing only depends on the observed data for that member, and no observed data, from other members of the cluster, or observations in other clusters. In addition, we allow for right-censored observations and assume noninformative censoring for an observation within the cluster.

Even with no missing covariate data, with censored data, it is difficult to specify the correlation structure between the observed responses (minimum of censoring and survival times) of patients in the same institution, which we consider a nuisance. There is a large literature on fitting marginal regression models for censored survival data; some of these approaches naively treat observations in a cluster as independent to obtain consistent parameter estimates, and use a ‘sandwich’-type robust variance estimator to account for the within-cluster dependence (Wei et al., 1989; Lipsitz et al., 1994). Other methods for parameter estimation with clustered survival data have been proposed by Pepe (1991) and Cai and Prentice (1995). If there is no censoring and no missing covariate data, then the regression parameters can be estimated using generalized estimating equations (Liang and Zeger, 1986) with parametric marginal distributions for the survival times of patients within an institution and any desired correlation structure between the survival times. A generalized estimating equation has yet to be developed in the presence of censoring and missing covariate data.

There have also been some recent approaches to semi-parametric estimation in univariate censored survival data models with missing covariates, which includes methods by Zhou and Pepe (1995), Lipsitz and Ibrahim (1996b), Zhong et al. (1996) and Chen and Little (1999). All of these methods can be used with the Cox model for univariate survival data, and missing covariate data. Our future goal is to extend these methods to clustered censored survival data with missing covariates in which the marginal distributions follow a Cox model. However, because of the semi-parametric nature of these methods, an extension of them to the clustered data setting can be quite complicated. As a first step, in this paper, we have chosen to work with a fully parametric marginal model.

As discussed above, with censoring but no missing covariate data, if the marginal distributions of the correlated survival times follow a given parametric model, then the estimates using the maximum likelihood estimating equations, naively treating the correlated survival times as independent, give consistent estimates of the relative risk parameters (Lipsitz et al., 1994). Here, we extend this approach to missing covariate data. In particular, to consistently estimate the regression parameters, we naively treat the observations within the cluster as independent and use the maximum likelihood estimating equations. Because of the missing covariate data, for a given observation in a cluster, we specify a conditional distribution of the survival time given the covariates and a marginal distribution for the covariates. We then use the EM algorithm (Lipsitz and Ibrahim, 1996b; Chen and Little, 1999) in a regression framework to obtain the estimates. Although the regression parameter estimates are consistent when naively assuming the observations in a cluster as independent, because of the correlation between survival times in a cluster,
### Table 1. Data from liver cancer clinical trials

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Survival (weeks)</th>
<th>Age</th>
<th>Jaundice</th>
<th>Hepatitis</th>
<th>Abnormal Anti-Hep</th>
<th>Abnormal Alpha feto</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.000</td>
<td>≥60</td>
<td>no</td>
<td>—</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>1</td>
<td>6.000</td>
<td>≥60</td>
<td>yes</td>
<td>no</td>
<td>—</td>
<td>yes</td>
</tr>
<tr>
<td>1</td>
<td>9.143</td>
<td>≥60</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>1</td>
<td>14.571</td>
<td>≥60</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>1</td>
<td>187.857</td>
<td>≥60</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>3.857</td>
<td>≥60</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>11.000</td>
<td>&lt;60</td>
<td>yes</td>
<td>no</td>
<td>—</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>21.429</td>
<td>&lt;60</td>
<td>no</td>
<td>yes</td>
<td>—</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>60.857</td>
<td>≥60</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>—</td>
</tr>
<tr>
<td>3</td>
<td>5.571</td>
<td>&lt;60</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>9.429</td>
<td>&lt;60</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>12.000</td>
<td>&lt;60</td>
<td>no</td>
<td>no</td>
<td>—</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>14.571</td>
<td>≥60</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>28.286</td>
<td>≥60</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>10.143</td>
<td>≥60</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>13.714</td>
<td>≥60</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>15.000</td>
<td>&lt;60</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>2.857</td>
<td>&lt;60</td>
<td>yes</td>
<td>no</td>
<td>—</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>11.857</td>
<td>≥60</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>16.714</td>
<td>&lt;60</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>7.857</td>
<td>&lt;60</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>8.714</td>
<td>&lt;60</td>
<td>no</td>
<td>yes</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>5</td>
<td>9.000</td>
<td>≥60</td>
<td>no</td>
<td>no</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>7</td>
<td>25.143</td>
<td>&lt;60</td>
<td>no</td>
<td>no</td>
<td>—</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>25.143</td>
<td>&lt;60</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>13.714</td>
<td>≥60</td>
<td>yes</td>
<td>no</td>
<td>—</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>27.571</td>
<td>≥60</td>
<td>yes</td>
<td>—</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>30.714</td>
<td>—</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>—</td>
</tr>
<tr>
<td>8</td>
<td>99.429</td>
<td>&lt;60</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>124.714</td>
<td>&lt;60</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>8</td>
<td>3.857</td>
<td>≥60</td>
<td>no</td>
<td>no</td>
<td>—</td>
<td>no</td>
</tr>
<tr>
<td>8</td>
<td>449.143*</td>
<td>&lt;60</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

* Censored.
— Missing.
Table 2. Percent missing out of 190 on each variable

<table>
<thead>
<tr>
<th>Variable</th>
<th>Percent missing</th>
<th>Number missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survival</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>Anti-hep</td>
<td>21.14</td>
<td>40</td>
</tr>
<tr>
<td>Alpha feto</td>
<td>7.91</td>
<td>15</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>2.14</td>
<td>4</td>
</tr>
<tr>
<td>Age</td>
<td>1.12</td>
<td>2</td>
</tr>
<tr>
<td>Jaundice</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>Overall</td>
<td>27.45</td>
<td>52</td>
</tr>
</tbody>
</table>

the inverse of the information matrix may not be a consistent estimate of the asymptotic variance. To consistently estimate the variance of the estimates, one needs to use a robust variance estimate such as a ‘sandwich’ estimator (White, 1982).

Thus, the main contribution of this paper is to show that the likelihood equations for univariate censored survival data with missing covariates can be used as an estimating equation to obtain consistent regression parameter estimates from the marginal model of clustered censored survival data with missing covariates. Note that, even with no censoring, the method discussed here has not been previously proposed for consistently estimating marginal regression parameters for clustered outcome data with missing covariates. In fact, the method can be used when the marginal model is any generalized linear model and the outcome is not censored.

2. DISTRIBUTION

Let $T_{ik}$ be the failure time for the $k$th member of cluster $i$, $i = 1, \ldots, N$; $k = 1, \ldots, n_i$, and $x_{ik} = [x_{ik1}, \ldots, x_{ikP}]'$ be a $(P \times 1)$ vector of covariates. Let $(t_{ik}|x_{ik})$ have density $p(t_{ik}|x_{ik}, \beta)$ indexed by a parameter vector $\beta$. The survival time is often right-censored. Let $U_{ik}$ be the censoring time so that instead of $T_{ik}$ we observe $Y_{ik} = \min(T_{ik}, U_{ik})$ where the censoring indicator $\delta_{ik} = I[T_{ik} \leq U_{ik}]$ equals 1 if $T_{ik}$ is a failure time and 0 if it is right-censored. Conditional on the covariates and assuming non-informative censoring (see Lawless, 1982), the joint probability distribution for $(y_{ik}, \delta_{ik}|x_{ik})$ is proportional to

$$p(y_{ik}, \delta_{ik}|x_{ik}, \beta) = p(y_{ik}|x_{ik}, \beta)^{\delta_{ik}} S(y_{ik}|x_{ik}, \beta)^{1-\delta_{ik}},$$

where $S(t|x_{ik}, \beta) = pr(T_{ik} > t|x_{ik}, \beta)$ is the survivor function of $T_{ik}$.

Since some elements of $x_{ik}$ can be missing, we must also consider its distribution. Let $x_{ik}$ have density $p(x_{ik}|\alpha)$ indexed by $\alpha$, where $\alpha$ is distinct from $\beta$. With missing elements of $x_{ik}$, it is convenient to introduce a $(P \times 1)$ random vector for the $k$th member of cluster $i$, $R_{ik}$, whose $p$th component, $R_{ikp}$, equals 1 if the $p$th component of $x_{ik}$ is observed, and equals 0 if it is missing. The distribution of $R_{ik}$ given $(y_{ik}, \delta_{ik}, x_{ik})$ is called the ‘missing data mechanism’, and has a multinomial distribution with $2^P$ cell probabilities,

$$P(r_{ik}|y_{ik}, \delta_{ik}, x_{ik}, \omega) = pr[R_{ik1} = r_1, \ldots, R_{ikp} = r_p|y_{ik}, \delta_{ik}, x_{ik}, \omega],$$

parameterized by $\omega$. In this paper, we assume that (1) depends on $(y_{ik}, \delta_{ik})$, and the observed components of $x_{ik}$, say $x_{obs,ik}$. In general, we write $x_{ik} = (x_{mis,ik}, x_{obs,ik})$ where $x_{mis,ik}$ are the missing components.
of \( x_{ik} \). Thus, we assume

\[
p(r_{ik} | y_{ik}, \delta_{ik}, x_{ik}, \omega) = p(r_{ik} | y_{ik}, \delta_{ik}, \mathbf{x}_{\text{obs},ik}, \omega).
\]  

(2)

We assume that given \((y_{ik}, \delta_{ik}, \mathbf{x}_{\text{obs},ik})\), \( r_{ik} \) is independent of data for any other member of cluster \( i \), or data from any other cluster. If observations within a cluster are independent, (2) is a missing at random assumption (Little and Rubin, 1987). We let \( y' = (\beta', \alpha') \). Our main interest is in the estimation of \( \beta \), with \( \alpha \) being viewed as a nuisance parameter.

3. THE ESTIMATING EQUATIONS

First, we describe the maximum likelihood estimating equations when observations in a cluster are independent, and there are no missing data. The estimating equations are

\[
u(\gamma) = \begin{bmatrix} u_1(\beta) \\ u_2(\alpha) \end{bmatrix} = \sum_{i=1}^{N} \sum_{k=1}^{n_i} \begin{bmatrix} u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{ik}) \\ u_{2ik}(\alpha; x_{ik}) \end{bmatrix} = 0,
\]

(3)

where

\[
u_1(\beta) = \sum_{i=1}^{N} \sum_{k=1}^{n_i} u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{ik}) = \sum_{i=1}^{N} \sum_{k=1}^{n_i} \partial \log p(y_{ik}, \delta_{ik}, x_{ik}, \beta) / \partial \beta;
\]

(4)

and

\[
u_2(\alpha) = \sum_{i=1}^{N} \sum_{k=1}^{n_i} u_{2ik}(\alpha; x_{ik}) = \sum_{i=1}^{N} \sum_{k=1}^{n_i} \partial \log p(x_{ik} | \alpha) / \partial \alpha.
\]

(5)

Using method of moments theory (Casella and Berger, 1990, ch. 7), \( \gamma \) is asymptotically normal and consistent since the score vectors have mean 0 (Cox and Hinkley, 1974), i.e. \( E[u_1(\beta)] = 0 \) and \( E[u_2(\alpha)] = 0 \), and we are solving \( u(\gamma) = 0 \) for \( \gamma \).

With some of the covariates missing for some members of the clusters, assuming observations in the cluster are independent, the maximum likelihood estimate of \( \gamma \) can be obtained by setting the conditional expectation of the complete data score vector given the observed data, denoted \( u^*(\gamma) \), to 0, and solving for \( \gamma \). In particular, this conditional expectation is taken with respect to the conditional distribution of the missing data given the observed data for the \( k \)th member of cluster \( i \), i.e.

\[
u^*(\gamma) = \sum_{i=1}^{N} \sum_{k=1}^{n_i} \begin{bmatrix} u_{1ik}^*(\gamma) \\ u_{2ik}^*(\gamma) \end{bmatrix}
\]

\[
= \sum_{i=1}^{N} \sum_{k=1}^{n_i} E \left[ \begin{bmatrix} u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{ik}) \\ u_{2ik}(\alpha; x_{ik}) \end{bmatrix} \right| \text{observed data for } k \text{th member of cluster } i 
\]

\[
= \sum_{i=1}^{N} \sum_{k=1}^{n_i} E \left[ \begin{bmatrix} u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{ik}) \\ u_{2ik}(\alpha; x_{ik}) \end{bmatrix} \right| y_{ik}, \delta_{ik}, \mathbf{x}_{\text{obs},ik}, r_{ik}
\]

\[
= \sum_{i=1}^{N} \sum_{k=1}^{n_i} E \left[ \begin{bmatrix} u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{ik}) \\ u_{2ik}(\alpha; x_{ik}) \end{bmatrix} \right| y_{ik}, \delta_{ik}, \mathbf{x}_{\text{obs},ik}
\]

(6)

The conditional expectation at the bottom of (6) does not depend on \( r_{ik} \) because (2) holds, i.e. the
missing data mechanism is independent of \( x_{mis,ik} \), and

\[
p(x_{mis,ik} | y_{ik}, \delta_{ik}, x_{obs,ik}, r_{ik}, \omega, \beta, \alpha) = \frac{p(r_{ik} | y_{ik}, \delta_{ik}, x_{ik}, \omega) p(y_{ik}, \delta_{ik} | x_{ik}, \beta) p(x_{ik} | \alpha)}{p(r_{ik} | y_{ik}, \delta_{ik}, x_{obs,ik}, \omega) p(y_{ik}, \delta_{ik} | x_{ik}, \beta) p(x_{ik} | \alpha)}
\]

\[
= \frac{p(r_{ik} | y_{ik}, \delta_{ik}, x_{obs,ik}, \omega) p(y_{ik}, \delta_{ik} | x_{ik}, \beta) p(x_{ik} | \alpha)}{p(r_{ik} | y_{ik}, \delta_{ik}, x_{obs,ik}, \omega) p(y_{ik}, \delta_{ik} | x_{ik}, \beta) p(x_{ik} | \alpha)}
\]

(7)

To make the notation clear, even though \( u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{ik}) \) is just a function of \( \beta \),

\[
u_{1ik}^*(\gamma) = E[u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{ik}) | y_{ik}, \delta_{ik}, x_{obs,ik}]
\]

\[
= \int_{x_{mis,ik}} p(x_{mis,ik} | y_{ik}, \delta_{ik}, x_{obs,ik}, \gamma) u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{ik}), \quad (8)
\]

is a function of \( \gamma = (\alpha, \beta) \).

Under independence of observations within a cluster, if the distributions for \((y_{ik}, \delta_{ik}, x_{ik})\) and \(x_{ik}\) are correctly specified and (2) holds, using method of moment ideas, \((\hat{\alpha}, \hat{\beta})\) is consistent since \(E[u_{1ik}^*(\gamma)] = 0\) and \(E[u_{1ik}^*(\gamma)] = 0\), and we are again solving \(u^*(\hat{\gamma}) = 0\) for \(\hat{\gamma}\). Using a heuristic argument, \(E[u^*(\gamma)] = 0\) since

\[
E[u^*(\gamma)] = \sum_{i=1}^{N} \sum_{k=1}^{n_i} E \left[ \frac{u_{1ik}^*(\gamma)}{u_{2ik}^*(\gamma)} \right]
\]

\[
= \sum_{i=1}^{N} \sum_{k=1}^{n_i} E \left[ \frac{u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{ik})}{u_{2ik}(\alpha; x_{ik})} \right] \bigg| y_{ik}, \delta_{ik}, x_{obs,ik}, r_{ik}
\]

\[
= \sum_{i=1}^{N} \sum_{k=1}^{n_i} E \left[ \frac{u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{ik})}{u_{2ik}(\alpha; x_{ik})} \right] \bigg| y_{ik}, \delta_{ik}, x_{obs,ik}, r_{ik}
\]

\[
= 0.
\]

Now, suppose that survival times within the same cluster are correlated. When calculating the expectation in (9), the correlation between observations in the same cluster does not come into play. As long as our specification of

\[
E \left[ \frac{u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{ik})}{u_{2ik}(\alpha; x_{ik})} \bigg| y_{ik}, \delta_{ik}, x_{obs,ik}, r_{ik} \right]
\]

(10)

in (6) is the correct conditional expectation, then, when taking unconditional expectations in (9), \(E[u^*(\gamma)] = 0\), and, using method of moment ideas, \(\hat{\gamma}\) will be consistent and asymptotically normal. If our posed distributions for \((y_{ik}, \delta_{ik}, x_{ik})\) and \(x_{ik}\) are correct, and \(r_{ik} | y_{ik}, \delta_{ik}, x_{ik}\) does not depend on \(x_{mis,ik}\) so that (8) holds, then (10) will be the correct specification. Otherwise, if \(r_{ik} | y_{ik}, \delta_{ik}, x_{ik}\) does depend on \(x_{mis,ik}\), which is called non-ignorable missingness, then we will have to specify a model for \((r_{ik} | y_{ik}, \delta_{ik}, x_{ik})\), and use it to calculate (6). However, non-ignorable models have problems with non-identifiability (Ibrahim et al., 1999b), and are beyond the scope of this paper, in which we assume that (2) holds. We note here that non-ignorable missingness can be induced when the conditional distribution of \(r_{ik}\) given all data for cluster \(i\) depends only on observed data, but some of these observed data are from other members of the cluster. In this case, unfortunately, even though \(r_{ik}\) given all data in the dataset only depends on observed data, using conditional probability, one can show that \((r_{ik} | y_{ik}, \delta_{ik}, x_{ik})\) depends on \(x_{mis,ik}\). In the liver
cancer clinical trial, since the members of the clusters are different patients, we feel that it is unlikely that the probability of missingness in the covariates from one member of the cluster will depend on observed data from other members of the cluster.

Thus, although observations within clusters tend to be correlated, using method of moment ideas, the maximum likelihood estimating equations obtained by naively assuming the observations within a cluster are independent will produce consistent and asymptotically normal estimates of \( \beta \). Further, to obtain the solution to these maximum likelihood estimating equations, one can use the EM algorithm as proposed by Lipsitz and Ibrahim (1996b). For discrete \( x_{mis,ik} \), Ibrahim (1990) and Lipsitz and Ibrahim (1996a) have proposed an EM algorithm to solve \( u^*(\hat{\gamma}) = 0 \); for continuous \( x_{mis,ik} \), Ibrahim et al. (1999a) have proposed a Monte Carlo EM algorithm to solve \( u^*(\hat{\gamma}) = 0 \). We briefly describe the EM algorithm in the appendix.

Although the regression parameters estimates are consistent when naively assuming the observations in a cluster as independent, when the survival times within a cluster are correlated, the inverse of the information matrix under independence may be an inconsistent estimate of the asymptotic variance. The asymptotic variance of \( \hat{\gamma} \) can be consistently estimated using a robust ‘sandwich’ estimator (White, 1982),

\[
\text{Var}(\hat{\gamma}) = \left\{ \sum_{i=1}^{N} \sum_{k=1}^{n_i} \hat{u}_{ik}^*(\hat{\gamma}) \right\}^{-1} \sum_{i=1}^{N} \sum_{k=1}^{n_i} \left[ \hat{u}_{ik}^*(\hat{\gamma}) \right]^2 \left\{ \sum_{i=1}^{N} \sum_{k=1}^{n_i} \hat{u}_{ik}^*(\hat{\gamma}) \right\}^{-1} \tag{11}
\]

where

\[
\hat{u}_{ik}^*(\hat{\gamma}) = \left[ \frac{\partial u_{ik}(\gamma)}{\partial \gamma} \right]_{\gamma=\hat{\gamma}}.
\]

The upper \( P \times P \) block of this estimate is consistent for the asymptotic variance of \( \hat{\beta} \).

4. Example

Although there are some difficulties in determining the precise number of deaths, each year approximately 2500 persons in the United States die of liver cancer. The ECOG has undertaken a series of phase II clinical trials to evaluate new treatments in patients with primary liver cancer. To illustrate our proposed methods, we consider data from two such clinical trials, EST 2282 (Falkson et al., 1990) and EST 1286 (Falkson et al., 1994). A total of 190 eligible patients were accrued from 31 cooperating institutions. In similar studies, it has often occurred that responses within institutions tend to be correlated. Then, we can think of an institution as cluster \( i \), with \( N = 31 \); within the \( i \)th institution, we have \( n_i \) patients (with 190 patients, we have an average of 6.1 patients per institution). The observed response for the \( k \)th patient in institution \( i \) is the minimum of the censoring and survival time.

We are primarily interested in how survival, the time from entry on the study until death, differs with respect to five baseline characteristics. These five baseline characteristics are age (less than 60 or greater than or equal to 60); associated jaundice (yes, no); associated hepatitis (yes, no); and two biochemical markers alpha fetoprotein and anti-hepatitis B antigen, each classified as normal or abnormal. Older patients, patients with abnormal biochemical markers, jaundice, and/or hepatitis are all expected to have shorter survival. Each of the five covariates are coded 0 or 1, with 1 assigned to the category expected to lead to shorter survival. Unfortunately, as shown in Table 2, 27.4% of the patients have at least one covariate missing. The biochemical markers, which require blood to be drawn and possibly complicated laboratory tests, have the highest proportion missing.

Assuming a Weibull survival distribution with non-informative censoring, we have

\[
p(y_{ik}, \delta_{ik}|x_{ik}, \beta) = \left[ \lambda_{ik} \sigma^{-1}(\lambda_{ik} y_{ik})^{-1} \exp(-\sigma^{-1}(\lambda_{ik} y_{ik}))^{\delta_{ik}}[\exp(-\sigma^{-1}(\lambda_{ik} y_{ik}))^{\sigma^{-1}}]^{1-\delta_{ik}} \right]
\]
where \( \sigma \) is the scale parameter and
\[
\lambda_{ik} = \exp(\beta_0 + \beta_1 x_{ik})
\]
is the location parameter for subject \( i \). If \( \sigma = 1 \), then (12) reduces to the exponential distribution. Here, we are interested in a main effects model, where \( x_{ik} = (x_{ik1}, \ldots, x_{ik5})' \) contains the five dichotomous covariates. Since all covariates are discrete, we model \( p(x_{ik} | \alpha) \) as a saturated multinomial with \( 2^5 \) levels.

Before we fit the Weibull model, we would like to briefly see if the missing data mechanism for these data fits the assumptions given in (1) and (2). To test if (1) holds versus an alternative of non-ignorable missingness (i.e. missingness depends on \( x_{mix,ik} \)), we would need to fit a complicated joint likelihood of the survival time, missing data indicators, and covariate distributions; however, the main advantage of the method described here is to avoid having to specify and maximize such a likelihood. Further, such non-ignorable models are often non-identifiable, and lead to inestimable parameters (Baker and Laird, 1988). Thus, here we are interested in testing if the missing data are missing completely at random (i.e. missingness does not depend on \( y_{ik}, \delta_{ik}, \) or \( x_{obs,ik} \)) versus missing at random. To do this in the most rigorous sense, we would still need to fit a complicated joint likelihood of the survival times, the missing data indicators, and covariate distributions, which is again beyond the scope of this paper. Instead, to test if the covariates are missing completely at random, we define a simpler version of (2) which can be written as follows. Let the indicator \( R_{ik}^* = \prod_{p=1}^{P} R_{ikp} \) equal 1 when the \( k \)th member of cluster \( i \) has no missing covariate data and 0 otherwise. We then form a logistic regression model for \( R_{ik}^* \) as a function of the completely observed variables: jaundice, follow-up time (\( y_{ik} \)), and censoring time (any of the other covariates have at least one missing value). Further, since only one subject has a censored outcome (the last patient in Table 1), we could not fit a logistic regression for \( R_{ik}^* \) with the censoring indicator as a covariate. Thus, we model the probability of being a complete case \( (R_{ik}^* = 1) \) as a function of the completely observed data,
\[
\logit[p(R_{ik}^* = 1 | jaun_{ik}, s_{ik}, \alpha)] = \alpha_0 + \alpha_1 \log(y_{ik}) + \alpha_2 jaun_{ik}.
\]

We actually fit (13) using generalized estimating equations (Liang and Zeger, 1986), assuming \( R_{ik}^* \)'s in the same cluster had an exchangeable correlation, i.e.
\[
\rho = \text{Corr}(R_{ij}^*, R_{ik}^* | jaun_{ij}, y_{ij}, jaun_{ik}, y_{ik}).
\]
The estimate of \( \alpha \) and \( \rho \) are given in Table 3, using the robust estimate of variance of \( \hat{\alpha} \) and \( \hat{\rho} \) given in Prentice (1988).

Since jaundice is significant in Table 3, this suggests that the missing data are not missing completely at random. In particular, a patient with jaundice has approximately \( \exp(0.61) \approx 1.91 \) times the odds of having the other covariates observed, possibly due to the fact that doctors of patients with jaundice may want to see all of the lab tests and biochemical markers at baseline. However, follow-up time (\( \log(y_{ik}) \)), is not significant at the 5% level. Although not significant, the relationship (positive log–odds ratio) between \( R_{ik}^* \) and follow-up time suggests that the longer a patient is followed, the greater the probability

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard error</th>
<th>Z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.244</td>
<td>0.358</td>
<td>0.68</td>
<td>0.497</td>
</tr>
<tr>
<td>Jaundice</td>
<td>0.607</td>
<td>0.304</td>
<td>1.99</td>
<td>0.046</td>
</tr>
<tr>
<td>Log(follow-up)</td>
<td>0.123</td>
<td>0.090</td>
<td>1.36</td>
<td>0.174</td>
</tr>
<tr>
<td>( \rho )</td>
<td>0.031</td>
<td>0.037</td>
<td>0.83</td>
<td>0.409</td>
</tr>
</tbody>
</table>
of all covariates being measured. We suggest that this is possibly due to the fact that the data managers have more time to query the laboratories for the values of the baseline biochemical markers for a patient who lives longer; it may not be worth the trouble to update the file of a patient who has already passed away. Further, doctors may order follow-up lab tests for a patient who lives longer, and has a data manager query the lab for the baseline value in order to look at the change from baseline. Since the intra-cluster correlation in Table 3 is non-significant, this suggests that (2) and not (1), holds, i.e. that \( r_{ik} \) given all data for cluster \( i \) depends only on the observed data from patient \( k \) in cluster \( i \). Further, assuming the results in Table 3 are approximately correct, with missingness only depending on the variable jaundice, one can show that a complete case analysis, in which we only use the 138 patients with no missing covariates, gives consistent, although possibly inefficient, estimates.

Table 3 gives the estimates of \( \beta \) using our method (modified EM), and the complete case method, with the variance estimated using both the inverse of the information matrix under independence and the robust variance estimator in (11). The two approaches give similar estimates, as we thought they would from the above discussion of the missing data mechanism. The biggest difference between the two approaches is in the estimate of the jaundice effect, although the complete case estimate (−0.220) is well within the modified EM estimate plus or minus two robust standard errors (−0.321 ± 1.96 × 0.184). For this dataset, the big advantage of using our method over complete cases appears to be the gain in efficiency. The estimated relative efficiencies, defined as the EM estimated variance over the complete case, is 53% for the age effect, which also turns out to be the covariate which is significant at \( \alpha = 0.05 \) using the EM-type method, but not using complete cases. Similarly, the estimated relative efficiency for the complete case estimate of \( \sigma \) is 53%.

To check to see if the Weibull appears to be a satisfactory fit to the data, using our EM-type method,
we have fit the piecewise exponential model (which is an approximation to the Cox model) with varying numbers of intervals. The estimates of $\beta$ were very similar to the Weibull in each case. Thus, the Weibull fit appears satisfactory.

Looking the the EM-type estimates, using the usual variance estimate under independence, only alpha fetoprotein is significant at the 5% level. Using the robust variance estimate, we find that both alpha fetoprotein and age are significant at the 5% level. In fact, for age, the $p$-value drops from 0.176 using the usual variance estimate under independence to 0.047 using the robust variance estimate. Further, using the usual variance estimate under independence, $\sigma$ is significantly different from 1 at the 5% significance level, whereas using the robust variance estimate, it is not. For every parameter except for jaundice, there is at least a 50% relative difference between the usual variance estimate under independence and the robust variance estimate. As we see from the variance estimates, it is not always the case that the variance estimates under independence will be smaller than the robust variance estimate. Overall, this example shows that using the usual variance estimate under independence can give misleading results. This demonstrates how an analysis that does not use the robust variance estimate for clustering can give misleading and conflicting results.

We have tried to come up with a simple rule of thumb to tell when the variance under independence overestimates or underestimates the true variance, but have yet to find a simple rule. The bias in the variance estimate under independence for a given covariate depends on the intra-cluster correlation among the outcomes in the cluster, the intra-cluster correlation among the different values of the given covariate in the cluster, and the correlation between the given covariate and other covariates. Unfortunately, we have been not able to come up with a simple formula describing the interplay between these three correlations, and have found that the only way to see the possible bias is to calculate the robust variance estimate, as well as the estimate under independence.

To use the method here, one can use the usual EM algorithm for independent observations, as discussed in the Appendix. To find our proposed estimate $\hat{\beta}$, we iterate between the E and M steps until $\beta^{(m+1)} = \beta^{(m)} = \hat{\beta}$. As shown in the Appendix, the M step can be done in any survival data and logistic regression program that allows for weights. However, these weights, which are the conditional probabilities of the missing data given the observed data, can be quite involved to calculate, possibly requiring computation of all possible values of $x_{mis,ik}$ for all subjects. We are in the early stages of developing a general SAS macro to do this for any number of categorical covariates, and varying cluster sizes. We note here that the convergence criterion used for the EM-type algorithm was that the distance between the $m$th iteration and the $(m+1)$th iteration in each parameter was less than $10^{-6}$. The number of iterations required for convergence was 12, with the complete case estimates used as starting values. We programmed the algorithm in SAS, and our SAS program took 17 seconds in real time to calculate the EM-type estimate on a SPARC20 Workstation.

5. DISCUSSION

In this paper, we have shown that if one has clustered survival data with missing covariate, and naively treats observations within a cluster as independent, that one can still use the maximum likelihood estimating equations to obtain consistent estimates of the marginal relative risk parameters, but one must use a robust variance estimator. Although the estimates are consistent, they could be inefficient if there is high correlation between observations in a cluster. A useful extension would then be to develop estimating equations that somehow use the correlation between the clustered censored survival times. Unfortunately, there is not even a good estimating equations approach for jointly estimating the marginal relative risk parameters and intra-cluster correlation coefficient with clustered censored survival data with no missing covariates (Segal et al., 1997). Adding missing covariates on top of the censoring makes this a daunting
Our approach only requires correct specification of the marginal distribution. Thus, our approach is a relatively simple and useful method to use with clustered survival data with missing covariates.

**Acknowledgements**

We are grateful for the support provided by grants CA 55576, CA 55670, CA 70101, and CA 74015.

**Appendix: EM Algorithm**

**Discrete Xmis,ik**

For discrete $x_{mis,ik}$, one can pose a saturated multinomial, a log-linear model Agresti (1990) or a set of multinomial logistic regressions (Lipsitz and Ibrahim, 1996a). If $x_{mis,ik}$ is discrete, $u_{ik}^1(\gamma')$ and $u_{2ik}^2(\gamma')$ in the score vector in (6) become

$$E[u_{1ik}(\beta'; y_{ik}, \delta_{ik}, x_{ik})|y_{ik}, \delta_{ik}, x_{obs,ik}]$$

$$= \sum_{x_{mis,ik}} \Pr[X_{mis,ik} = x_{mis,ik}|y_{ik}, \delta_{ik}, x_{obs,ik}, \gamma] u_{1ik}(\beta'; y_{ik}, \delta_{ik}, x_{ik})$$

and

$$E[u_{2ik}(\alpha'; x_{ik})|y_{ik}, \delta_{ik}, x_{obs,ik}] = \sum_{x_{mis,ik}} \Pr[X_{mis,ik} = x_{mis,ik}|y_{ik}, \delta_{ik}, x_{obs,ik}, \gamma] u_{2ik}(\alpha'; x_{ik})$$

where the sum extends over all possible values of $x_{mis,ik}$. The score vector in (6) then becomes

$$u^*(\gamma) = \sum_{i=1}^{N} \sum_{k=1}^{n_i} \left[ \sum_{x_{mis,ik}} w_{ik,x_{mis,ik}} u_{1ik}(\beta'; y_{ik}, \delta_{ik}, x_{obs,ik}, x_{mis,ik}) \right]$$

where

$$w_{ik,x_{mis,ik}} = w_{ik,x_{mis,ik}}(\gamma) = \Pr(X_{mis,ik} = x_{mis,ik}|y_{ik}, \delta_{ik}, x_{obs,ik}, r_{ik}, \gamma)$$

$$= \frac{\Pr(y_{ik}, \delta_{ik}|x_{ik}, \beta) p(x_{ik}|x_{ik}, \alpha)}{\sum_{x_{mis,ik}} \Pr(y_{ik}, \delta_{ik}|x_{ik}, \beta) p(x_{ik}|x_{ik}, \alpha)}.$$  

(15)

We can think of (14) as a weighted score vector, with weights given by $w_{ik,x_{mis,ik}}$. The solution to $u^*(\hat{\gamma}) = 0$ can be obtained via the EM algorithm. Suppose we define the function

$$u^*(\gamma|\gamma^{(t)}) = \sum_{i=1}^{N} \sum_{k=1}^{n_i} \left[ \sum_{x_{mis,ik}} w_{ik,x_{mis,ik}}^{(t)} u_{1ik}(\beta'; y_{ik}, \delta_{ik}, x_{obs,ik}, x_{mis,ik}) \right]$$

where $w_{ik,x_{mis,ik}}^{(t)} = w_{ik,x_{mis,ik}}(\gamma^{(t)})$. The EM iterative algorithm to solve $u^*(\hat{\gamma}) = 0$ entails the following:

- Obtain an initial estimate $\gamma = \gamma^{(1)}$, say, by complete cases. At the $r$th step, we have $\gamma^{(t)}$.
- Using $\gamma^{(t)}$, calculate $w_{ik,x_{mis,ik}}^{(t)} = w_{ik,x_{mis,ik}}(\gamma^{(t)})$.
- Fixing $w_{ik,x_{mis,ik}}^{(t)} = w_{ik,x_{mis,ik}}(\gamma^{(t)})$, solve $u^*(\gamma^{(t+1)}|\gamma^{(t)}) = 0$ for $\gamma^{(t+1)}$, using any survival and multinomial logistic programs that allows for weights.
- We iterate until convergence, which gives the solution to $u^*(\hat{\gamma}) = 0$. 
Continuous $X_{\text{mix},ik}$

For continuous $x_{\text{mix},ik}$, one need not pose a multivariate normal distribution, but can pose a set of any continuous univariate distributions, as described in Ibrahim et al. (1999a). When the missing covariates are continuous, one can adapt the Monte Carlo EM algorithm of Wei and Tanner (1990) or the stochastic approximation Monte Carlo EM algorithm of Gu and Lin (1998). With $x_{\text{mix},ik}$ continuous, the score vector in (6) is

$$u^*(\gamma) = \sum_{i=1}^{N} \sum_{k=1}^{n_i} \left[ \int_{x_{\text{mix},ik}} w_{ik,x_{\text{mix},ik}} u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{\text{obs},ik}, x_{\text{mix},ik}) \int_{x_{\text{mix},ik}} w_{ik,x_{\text{mix},ik}} u_{2ik}(\alpha; x_{\text{obs},ik}, x_{\text{mix},ik}) \right],$$

where $w_{ik,x_{\text{mix},ik}}$ is the conditional density

$$w_{ik,x_{\text{mix},ik}} = w_{ik,x_{\text{mix},ik}}(\gamma) = p(x_{\text{mix},ik} | y_{ik}, \delta_{ik}, x_{\text{obs},ik}, r_{ik}, \gamma) = \frac{p(y_{ik}, \delta_{ik} | x_{ik}, \beta) p(x_{ik} | \alpha)}{\int_{x_{\text{mix},ik}} p(y_{ik}, \delta_{ik} | x_{ik}, \beta) p(x_{ik} | \alpha)}.$$

To obtain the solution $u^*(\gamma) = 0$, we draw $L$ values of $x_{\text{mix},ik}$ from

$$w_{ik,x_{\text{mix},ik}}(\gamma^{(l)}) = p(x_{\text{mix},ik} | y_{ik}, \delta_{ik}, x_{\text{obs},ik}, r_{ik}, \gamma^{(l)})$$

using the Gibbs Sampler (Gilks and Wild, 1992). We estimate $u^*(\gamma | \gamma^{(l)})$ with

$$u^{**}(\gamma | \gamma^{(l)}) = \sum_{i=1}^{N} \sum_{k=1}^{n_i} \left[ \sum_{l=1}^{L} \frac{1}{L} u_{1ik}(\beta; y_{ik}, \delta_{ik}, x_{\text{obs},ik}, x_{\text{mix},ik}^{(l)}) \sum_{l=1}^{L} \frac{1}{L} u_{2ik}(\alpha; y_{ik}, \delta_{ik}, x_{\text{obs},ik}, x_{\text{mix},ik}^{(l)}) \right],$$

where $x_{\text{mix},ik}^{(l)}$ is the $l$th draw from $w_{ik,x_{\text{mix},ik}}(\gamma^{(l)})$. In iterations of Monte Carlo EM algorithm, instead of solving $u^*(\gamma | \gamma^{(l)}) = 0$ for $\gamma^{(l+1)}$, we solve its approximation $u^{**}(\gamma | \gamma^{(l)}) = 0$. Then, the EM algorithm proceeds just as with discrete $x_{\text{mix},ik}$.

REFERENCES


Estimation with survival data with missing covariates


[Received September 7, 1999; revised January 27, 2000; accepted for publication February 22, 2000]