STATISTICS IN MEDICINE
*Statist. Med.* **18**, 3111–3121 (1999)

# LINEAR REGRESSION FOR BIVARIATE CENSORED DATA VIA MULTIPLE IMPUTATION

WEI PAN[1]* AND CHARLES KOOPERBERG[2]

[1] *Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, Box 303, 420 Delaware Street SE, Minneapolis, MN 55455-0378, U.S.A.*
[2] *Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue E/MP 1002, Seattle, WA 98109-1024, U.S.A.*

## SUMMARY

Bivariate survival data arise, for example, in twin studies and studies of both eyes or ears of the same individual. Often it is of interest to regress the survival times on a set of predictors. In this paper we extend Wei and Tanner's multiple imputation approach for linear regression with univariate censored data to bivariate censored data. We formulate a class of censored bivariate linear regression methods by iterating between the following two steps: 1, the data is augmented by imputing survival times for censored observations; 2, a linear model is fit to the imputed complete data. We consider three different methods to implement these two steps. In particular, the marginal (independence) approach ignores the possible correlation between two survival times when estimating the regression coefficient. To improve the efficiency, we propose two methods that account for the correlation between the survival times. First, we improve the efficiency by using generalized least squares regression in step 2. Second, instead of generating data from an estimate of the marginal distribution we generate data from a bivariate log-spline density estimate in step 1. Through simulation studies we find that the performance of the two methods that take the dependence into account is close and that they are both more efficient than the marginal approach. The methods are applied to a data set from an otitis media clinical trial. 

## 1. INTRODUCTION

Bivariate failure time data arise when study units are paired, such as the eyes, ears, lungs and kidneys from the same person. A well-known example involving bivariate failure times is the Diabetic Retinopathy Study,[1] which was conducted to investigate the effectiveness of laser photocoagulation in delaying the onset of blindness for diabetic retinopathy patients. One eye of each patient was randomly chosen to receive photocoagulation, whereas another eye served as control. The subjects were followed for several years and the conditions of their eyes were recorded. It seems reasonable to assume that the results for the two eyes of the same patient are correlated. This dependence, along with the presence of censoring, greatly complicates the

* Correspondence to: Wei Pan, Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, Box 303, 420 Delaware Street SE, Minneapolis, MN 55455-0378, U.S.A. E-mail: weip@biostat.umn.edu

Contract/grant sponsor: NCI
Contract/grant number: R29 CA 74841

CCC 0277–6715/99/223111–11$17.50
*Received June 1998*
*Accepted February 1999*

analysis of the data. Other frequently cited examples involving bivariate failure time data are the Australian Twins Study,[2] the Danish Twin Register,[3] and a ventilation tube study for otitis media patients.[4] We will use the otitis media data as an example in Section 5.

The Cox proportional hazards model and the accelerated failure time model (AFT) are the most popular regression methodologies for survival data. The Cox model assumes that the conditional hazard function for any set of covariates is proportional to an unknown baseline hazard function. In the AFT model there is a linear relation between the logarithm of the survival time and the covariates. (In generalizations of the AFT model, the logarithmic transformation of the survival times can be replaced by other monotone transformations from $(0, \infty)$ to the real line.) The regression analysis based on the AFT model is often called linear regression.[5,6]

For univariate failure time data, approaches for fitting AFT models include the Buckley–James method[7] and a method using linear rank statistics.[8,9] Ritov[10] has established their asymptotic equivalence. Recently Wei and Tanner[11] proposed a multiple imputation approach using two data augmentation schemes. Lee *et al.*[12] proposed a marginal approach to extend the Buckley–James and the linear rank statistics based methods to multivariate survival time data. The marginal approach only requires a correctly specified marginal model, and estimates the regression coefficients by ignoring the correlation of the failure times. It has the advantage of being simple to implement while being asymptotically valid. However, since it ignores the dependence structure in the data, the estimates are likely not efficient. In this paper we propose two new methods which account for the within-cluster correlation to different extents. These two methods, as well as a related marginal approach, are implemented via multiple imputation. We also show that our marginal approach implemented by multiple imputation is a Monte Carlo approximation to the marginal Buckley–James method.[12]

The paper is organized as follows. Section 2 describes a bivariate AFT model and explains the basic ideas of the two new methods as well as the marginal approach. In Section 3 we discuss the poor man's data augmentation technique in the context of linear regression for bivariate censored data. A simulation study that was conducted to assess the performance of the methods is described in Section 4. Section 5 contains an example of the methods to a ventilating tube duration data set from an otitis media clinical trial. We end the paper with a short discussion.

## 2. MODEL AND METHODS

### 2.1. A Bivariate Model

Let $T_{ij}$ be the logarithm of the failure time of the $j$th individual in cluster $i$, where $j = 1, 2$ and $i = 1, \ldots, n$. Because of censoring, $T_{ij}$ is not always observed. There is a censoring random variable $C_{ij}$ independent of $T_{ij}$. We observe $Y_{ij} = \min(T_{ij}, C_{ij})$ and $\delta_{ij} = I(T_{ij} \leqslant C_{ij})$, for each $i$ and $j$, where $I(.)$ is the usual indicator function. The linear (AFT) model is

$$T_{ij} = X_{ij}\beta + \varepsilon_{ij} \tag{1}$$

where $X_{ij}$ are covariates and $\beta$ is the unknown regression coefficient (vector) that is of primary interest. The mean-zero random vectors $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2})'$, $i = 1 \ldots, n$, are independent of each other, but the components $\varepsilon_{i1}$ and $\varepsilon_{i2}$ generally are not. Throughout we denote any vector with components $v_i$ by $v$, with the exception that $X$ will denote the design matrix with elements $X_{ij}$. In this paper we assume that the marginal distributions of $\varepsilon_{i1}$ and $\varepsilon_{i2}$ are equal, and we denote the marginal distribution as $F_0$ and the joint distribution as $G$. Let $V_0$ be the $2 \times 2$ covariance matrix

 

of $\varepsilon_i$, and let $V = \text{diag}(V_0, \ldots, V_0)$ be the $2n \times 2n$ covariance matrix of $(\varepsilon_1', \ldots, \varepsilon_n')'$. Note that we do not make a parametric assumption about $F_0$ or $G$. It is straightforward to extend the proposed methodologies to the situation with two unequal marginal distributions.

Now we briefly describe three methods of estimating $\beta$ for model (1). More details will be given in the Section 3. All three methods consist of two steps: 1, the data is augmented by imputing survival times $T_{ij}$ for censored observations; 2, a linear model is fit to the imputed complete data $T$.

### 2.2. The Marginal Approach

The marginal (independence) approach [12] estimates $\beta$ by ignoring the possible correlation between the two components of $\varepsilon_i$. Specifically, it has a working assumption that $V_0 = \text{diag}(1, 1)$. For complete data $T$, that is, $\delta_{ij} = 1$, for all $i$ and $j$, $\beta$ in model (1) can be estimated by ordinary least squares (OLS) regression. When some observations are censored, for any given estimate $\hat{\beta}$ the marginal distribution $F_0$ can be estimated by the Kaplan–Meier estimator $\hat{F}(\hat{\beta})$ from the residuals $\{e_{ij} = Y_{ij} - X_{ij}\hat{\beta}, \delta_{ij}\}$. Using $\hat{F}(\hat{\beta})$ we can impute values for those $T_{ij}$ that are censored. In particular, in the Buckley–James method a censored observation $T_{ij}$ is replaced by its conditional expectation $E(\varepsilon_{ij} | \varepsilon_{ij} \geqslant e_{ij}) + X_{ij}\hat{\beta}$, where the expectation is taken under the distribution $\hat{F}(\hat{\beta})$. In this paper we use multiple imputation: we draw a random sample from $\hat{F}(\hat{\beta})$, conditional on the observed $\{e_{ij}, \delta_{ij}\}$. Since the two steps of fitting a linear model and imputing censored observations are dependent on each other we use an iterative procedure.

Note that under the independence assumption the possibly correlated residuals $\{e_{i1}, \delta_{i1}\}$ and $\{e_{i2}, \delta_{i2}\}$ can be pooled together to obtain the Kaplan–Meier estimate $\hat{F}(\hat{\beta})$. Ying and Wei[13] have proved a consistency result of the Kaplan–Meier estimator when applied to such dependent data.

### 2.3. The Semi-Marginal Approach

In the semi-marginal approach we also estimate the marginal distribution $F_0$ by applying the Kaplan–Meier estimator to pooled residuals under the working assumption that they are independent and identically distributed, and we use this Kaplan–Meier estimate to impute censored data. However, when we fit the linear model (1) with imputed complete data we take account of the within-cluster correlation.

Given the model (1) it is well known that, with a known covariance matrix $V$, generalized least squares (GLS) yields more efficient estimates than OLS. In particular, the GLS estimate is $\hat{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}T$. Since $V = \text{diag}(V_0, \ldots, V_0)$ is not known, we estimate the covariance matrix $V_0$ of $\varepsilon_i$ from the residuals. Otherwise, the semi-marginal approach employs an iterative algorithm similar to the marginal approach. Note that the estimate of $V_0$ based on the imputed data may be a biased estimate of $V_0$, since the dependence of $\varepsilon_{i1}$ and $\varepsilon_{i2}$ is ignored in the imputation procedure; however, we expected that the bias is small when the censoring is not heavy.

### 2.4. The Bivariate Log-spline Density Estimation Approach

There are several reasons why we may improve over the marginal and semi-marginal approaches. First, although ignoring the within-cluster correlation in estimating the marginal distribution

 

using the Kaplan–Meier estimator is asymptotically valid, its finite sample performance is unclear. Second, we may lose some information when imputing, for example, $T_{i1}$ and ignoring the information about $T_{i2}$. The intuition is that if we can obtain an estimate $\hat{G}$ of the joint distribution $G$, then we can impute the two components in the same cluster as $(X_{i1}\hat{\beta}, X_{i2}\hat{\beta})' + (U_{i1}, U_{i2})'$, where $(U_{i1}, U_{i2})'$ is a sample from $\hat{G}(.,.|(e_{i1}, \delta_{i1}), (e_{i2}, \delta_{i2}))$. Intuitively, the within-cluster correlation is now better handled.

There is an extensive literature on the estimation of bivariate distribution functions from censored data. The reader is referred to van der Laan[14] and references therein. In our current context, most of the non-parametric bivariate distribution function estimators are unsuitable. For clusters of which only one of the two observations is censored, we want to generate a pair $(U_{i1}, U_{i2})$, where one of the two elements, say $U_{i1}$, is identical to the observed residual $e_{i1}$. However, because of the discreteness of many estimators of the bivariate distribution function, there are few points with a positive probability mass for fixed $e_{i1}$. Hence it is difficult to have good imputation for singly censored observations. Smooth bivariate estimators, such as the bivariate log-spline density estimate of Kooperberg,[15] may work better. In bivariate log-spline density estimation, first the data is transformed to the unit square. Then the bivariate log-density for the transformed data is estimated using linear splines and their tensor products, after which the estimate is transformed back to the original scale. It is adaptive in the sense that it chooses the number and location of spline knots automatically. The bivariate log-spline density estimation can deal with censored data, and seems to give reasonable estimates of the dependence structure even when some data is censored. For more details see Kooperberg.[15]

## 3. DATA AUGMENTATION FOR CENSORED DATA

The data augmentation algorithm which we use was originally proposed by Tanner and Wong[16] to compute the posterior density of a parameter based on incomplete data. Our implementation is based on the poor man's algorithm.

In the poor man's algorithm the augmented (complete) data $D_c$ is generated from $p(D_c|D_o, \theta_i)$, where $p(\,)$ represents a probability model of the complete data based on the observed data $D_o$ and the current parameter estimate $\theta_i$.

In our current context, the observed data are the censored failure times $Y$ and the censoring indicators $\delta$. The augmented data are the underlying uncensored failure times $T$ and the parameter is the regression coefficient $\beta$. Our algorithm now becomes:

1. Start with an initial estimate of the regression coefficient, $\hat{\beta}^{(0)}$; set $i = 0$.
2. Calculate the Kaplan–Meier estimate $\hat{F}^{(i)}$ or bivariate log-spline estimate $\hat{G}^{(i)}$ from $\{e^{(i)}, \delta\}$, where $e^{(i)} = Y - X\hat{\beta}^{(i)}$.
3. Do for $k = 1, \ldots, m$.
   (a) Sample a set of $n$ pairs of residuals $\tilde{e}_{(k)}^{(i+1)}$ from the distribution $\hat{F}^{(i)}$ or $\hat{G}^{(i)}$ conditional on the observed $\{e^{(i)}, \delta\}$; form the complete data $\tilde{T}_{(k)}^{(i+1)} = X\hat{\beta}^{(i)} + \tilde{e}_{(k)}^{(i+1)}$.
   (b) Fit a linear model via OLS or GLS to $\{\tilde{T}_{(k)}^{(i+1)}, X\}$ to obtain an estimate $\hat{\beta}_{(k)}^{(i+1)}$ and its estimated covariance matrix $\hat{\Sigma}_{(k)}^{(i+1)}$.
4. Set

$$\hat{\beta}^{(i+1)} = \frac{1}{m}\sum_{k=1}^{m} \hat{\beta}_{(k)}^{(i+1)}$$

 

and

$$\hat{\Sigma}^{(i+1)} = \frac{1}{m}\sum_{k=1}^{m}\hat{\Sigma}_{(k)}^{(i+1)} + \left(1+\frac{1}{m}\right)\sum_{k=1}^{m}\frac{(\hat{\beta}_{(k)}^{(i+1)} - \hat{\beta}^{(i+1)})^2}{m-1}. \quad (2)$$

5. Set $i = i + 1$. Go to step 2 until convergence.

The regression coefficient and its covariance matrix are estimated by $\hat{\beta}^{(i)}$ at convergence and $\hat{\Sigma}^{(i)}$ at convergence, respectively. The two terms in (2) correspond to the within-imputation and the between-imputation variances. The within-imputation variance is an average of the estimated covariance matrices for $\hat{\beta}_{(k)}^{(i+1)}$ obtained from the imputed complete data. The between-imputation variance involves the sample covariance matrix calculated from observations $\hat{\beta}_{(1)}^{(i+1)}, \ldots, \hat{\beta}_{(m)}^{(i+1)}$. An inflation factor of $(1 + 1/m)$ is used to account for the fact that we only carry out a finite number of imputations $m$.[17–20]

Note that:

(i) In step 2, the Kaplan–Meier estimate and the bivariate log-spline estimate are obtained under the assumption that the two marginal distributions are equal. If the distributions are not equal, we can obtain two separate Kaplan–Meier estimates, or a bivariate log-spline estimate with unequal marginals in step 2, and modify the data generation in step 3 accordingly. Currently there are no implementations of log-spline density estimation for multivariate data with more than two components.

(ii) It is possible to extend the marginal and semi-marginal methods to multivariate data. For instance, if we still assume that the marginals are equal, we do not need to modify the algorithm but we can directly implement the GLS.

(iii) The poor man's data augmentation implementation of the marginal approach is a Monte Carlo approximation to the marginal Buckley–James method.[12] To see this, note that from step 4 we have

$$\hat{\beta}^{(i+1)} = \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_{(k)}^{(i+1)} = \frac{1}{m}\sum_{k=1}^{m}(X'X)^{-1}X'\tilde{T}_{(k)}^{(i+1)} = (X'X)^{-1}X'\frac{1}{m}\sum_{k=1}^{m}\tilde{T}_{(k)}^{(i+1)}$$

and $\sum_{k=1}^{m}\tilde{T}_{(k)}^{i+1)}/m$ is a Monte Carlo realization of $E(T|e^{(i)}, \delta, X, \beta^{(i)})$ under the independence working assumption. This observation has been verified in our simulations (results not shown).

(iv) Since GLS estimates the covariance matrix $V$ of the survival times (see Section 2.3) it appears that we implicitly obtain an estimate of the correlation coefficient between two marginal survival times. Some limited computations (not reported) suggest that this is sometimes a downwards biased estimate of the variance. There is a large literature on estimating correlations for censored data, see Oakes,[21] Hsu and Prentice[22] and Shih and Louis.[23]

(v) For OLS and GLS the (ordinary) covariance estimates $\hat{\Sigma}_{(k)}^{(i+1)}$ are easy to calculate. Specifically, for OLS $\hat{\Sigma}_{(k)}^{(i+1)} = \hat{\sigma}^2(X'X)^{-1}$, where $\hat{\sigma}^2$ is the estimated variance of the survival time $\tilde{T}_{(k)}^{(i+1)}$. For GLS, suppose that $\hat{V}$ is the estimate of the covariance $V$ of $\tilde{T}_{(k)}^{(i+1)}$, then $\hat{\Sigma}_{(k)}^{(i+1)} = (X'\hat{V}^{-1}X)^{-1}$. These are our default estimates of the covariance matrix. Furthermore, a sandwich estimator based on the GLS estimating equations can be obtained as:[24]

$$\hat{\Sigma}_{(k)}^{(i+1)} = (X'\hat{V}^{-1}X)^{-1}[X'\hat{V}^{-1}(\tilde{T}_{(k)}^{(i+1)} - X\hat{\beta}_{(k)}^{(i+1)})(\tilde{T}_{(k)}^{(i+1)} - X\hat{\beta}_{(k)}^{(i+1)})'\hat{V}^{-1}X](X'\hat{V}^{-1}X)^{-1} \quad (3)$$

which in combination with (2) can be used to obtain a 'sandwich' covariance estimate of $\hat{\beta}^{(i+1)}$.

 

## 4. SIMULATIONS

We compare the performance of the marginal, semi-marginal and bivariate approach through a simulation study. The data were generated from the linear model

$$T_{ij} = X_{ij} + b_i + \varepsilon_{ij}$$

where $j = 1, 2$; $i = 1, \ldots, n$, so that the true value of $\beta$ is 1. The random effect $b_i$, which is the same within a cluster, introduces the within-cluster correlation. We used $n = 50$ and $n = 250$ pairs. For all simulations the covariates $X_{ij}$ are generated i.i.d. from a uniform distribution between 0 and 1. We used a variety of distributions for the random effects $b_i$ and the random errors $\varepsilon_{ij}$, to verify that our semi-parametric approaches are not much influenced by these distributions. In particular, the random errors were generated from:

N, a normal N $(0, \sigma^2)$ with mean 0 and variance $\sigma^2 = 1$ or 4;
CN, a contaminated normal distribution that is the mixture of two normal distributions: $0{\cdot}9 \times \mathrm{N}(0, 1) + 0{\cdot}1 \times \mathrm{N}(0, 9)$; and
D2, a distribution concentrated on two points: $p(-1) = p(1) = 0{\cdot}5$.

There were two levels of censoring; low (around 20 per cent of the marginal $T_{ij}$ in each component are censored) and high (around 60 per cent of the $T_{ij}$ are censored). The censoring variables $C_{ij}$ were generated i.i.d. from a normal distribution with mean and variance empirically determined to achieve approximately the right censoring percentage. By changing the variances of $b_i$ and $\varepsilon_{ij}$ we can adjust the within-cluster correlations. Table I lists the seven different set ups for the simulations study.

We stop the iterations if the difference between consecutive estimates of $\beta$ is less than 0·01, provided that we carried out at least four iterations. The maximum number of iterations was 10.

The results of the simulation study are summarized in Tables II ($n = 50$) and III ($n = 250$). It appears that all three approaches yield unbiased estimates of $\beta$ for all set-ups and both sample sizes, even for non-normal random effects or random errors. This is not surprising, since all three approaches are semi-parametric and do not depend on any specific parametric assumption. When there is no within-cluster correlation, as in set-up 1, or when the correlation is small, as in set-ups 2 and 4, the semi-marginal approach and the bivariate approach yield results that have almost the same efficiency as the marginal approach, which has the correct independence working assumption for set-up 1. However, when the within-cluster is 0·5, the semi-marginal and bivariate approach yield more efficient estimates than the marginal approach. The relative efficiency for the semi-marginal approach and the bivariate approach is slightly better for $n = 250$ than for $n = 50$. When the sample size increases, the performance of all three methods improves, suggesting desirable large-sample properties such as consistency.

Wei and Tanner[11] suggested that the standard error estimate from the poor man's data augmentation scheme may underestimate the true standard deviation. Indeed, the estimate of $\mathrm{SE}(\tilde{\beta})$ tend to be somewhat smaller than the standard deviations of $\tilde{\beta}$ in Tables II and III for most configurations for both sample sizes.

To investigate this downward bias in the standard errors further, we generated 200 uncensored data sets using the same configurations as in our simulation study. For each data set we computed $\hat{\beta}$ and $\mathrm{SE}(\hat{\beta})$ using OLS and using GLS with an estimated correlation matrix. In Table IV we report the mean value of $\mathrm{SE}(\hat{\beta})$. The OLS standard errors should be compared with

 

Table I. Simulation configurations

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Within-cluster correlation | 0 | 0·25 | 0·5 | 0·25 | 0·5 | 0·5 | 0·5 |
| Censoring level | low | low | low | high | high | low | low |
| Distribution of random effect $b$ | N | N | N | N | N | N | D2 |
| Distribution of error term $\varepsilon$ | N | N | N | N | N | CN | N |

Table II. Monte Carlo means of the estimates of the regression coefficient, their standard error estimates (SE) and the mean squared error (MSE) of the regression coefficient estimate (Monte Carlo standard deviations in parentheses), based on 200 independent replications with $n = 50$ pairs. The true regression coefficient is $\beta_0 = 1$. The relative efficiency (RE) of the semi-marginal or bivariate approach is the ratio the MSE of that approach over the MSE from the marginal approach

| Set-up | Marginal | | | Semi-marginal | | | | Bivariate | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\tilde{\beta}$ | SE($\tilde{\beta}$) | MSE | $\hat{\beta}$ | SE($\hat{\beta}$) | MSE | RE | $\hat{\beta}$ | SE($\hat{\beta}$) | MSE | RE |
| 1 | 0·96 (0·33) | 0·34 (0·03) | 0·112 (0·010) | 0·96 (0·34) | 0·34 (0·03) | 0·116 (0·011) | 1·03 | 0·95 (0·34) | 0·33 (0·03) | 0·116 (0·011) | 1·03 |
| 2 | 1·00 (0·47) | 0·39 (0·04) | 0·223 (0·021) | 0·99 (0·46) | 0·38 (0·04) | 0·213 (0·020) | 0·96 | 0·98 (0·46) | 0·37 (0·04) | 0·211 (0·020) | 0·95 |
| 3 | 1·02 (0·49) | 0·46 (0·04) | 0·238 (0·029) | 0·99 (0·46) | 0·41 (0·04) | 0·209 (0·025) | 0·88 | 0·99 (0·46) | 0·40 (0·04) | 0·208 (0·026) | 0·87 |
| 4 | 1·02 (0·50) | 0·41 (0·06) | 0·247 (0·023) | 1·02 (0·50) | 0·40 (0·06) | 0·250 (0·026) | 1·01 | 1·00 (0·51) | 0·38 (0·06) | 0·262 (0·023) | 1·06 |
| 5 | 0·98 (0·62) | 0·48 (0·07) | 0·385 (0·040) | 0·99 (0·58) | 0·46 (0·07) | 0·333 (0·035) | 0·87 | 0·98 (0·59) | 0·42 (0·07) | 0·352 (0·068) | 0·91 |
| 6 | 1·00 (0·70) | 0·58 (0·07) | 0·489 (0·053) | 1·03 (0·61) | 0·52 (0·07) | 0·370 (0·043) | 0·76 | 1·02 (0·61) | 0·50 (0·07) | 0·372 (0·042) | 0·76 |
| 7 | 1·01 (0·50) | 0·48 (0·05) | 0·250 (0·025) | 1·01 (0·44) | 0·42 (0·05) | 0·194 (0·019) | 0·78 | 1·02 (0·45) | 0·41 (0·05) | 0·205 (0·020) | 0·82 |

the standard errors of the marginal approach and the GLS standard errors should be compared with those for the semi-marginal and bivariate approaches. As can be seen, the estimated standard errors based on the poor man's data augmentation are very similar to those based on the uncensored data. Only for simulation set-up 6, where the error distribution is not normal, for both sample sizes are the standard error estimates based on the data augmentation scheme somewhat smaller than those for the uncensored data. For most set-ups and $n = 50$ the standard error estimates are very slightly smaller than those for the uncensored data. However, since in particular for the set-ups with higher censoring the information in the censored data is clearly smaller than in the uncensored data, this suggests that for larger amounts of censoring the standard error estimates may indeed be somewhat downward biased, as they may not adequately acknowledge the loss of information due to the censoring and the uncertainty in the Kaplan–Meier and the bivariate logspline estimate.

 

Table III. Monte Carlo means of the estimates of the regression coefficient, their standard error estimates (SE) and the mean squared error (MSE) of the regression coefficient estimate (Monte Carlo standard deviations in parentheses), based on 200 independent replications with $n = 250$ pairs. The true regression coefficient is $\beta_0 = 1$. The relative efficiency (RE) of the semi-marginal or bivariate approach is the ratio the MSE of that approach over the MSE from the marginal approach

| Set-up | Marginal | | | Semi-marginal | | | | Bivariate | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\tilde{\beta}$ | SE($\tilde{\beta}$) | MSE | $\hat{\beta}$ | SE($\hat{\beta}$) | MSE | RE | $\hat{\beta}$ | SE($\hat{\beta}$) | MSE | RE |
| 1 | 1·00 (0·17) | 0·16 (0·01) | 0·028 (0·003) | 1·00 (0·17) | 0·16 (0·01) | 0·028 (0·003) | 1·02 | 0·99 (0·17) | 0·15 (0·01) | 0·028 (0·003) | 1·01 |
| 2 | 1·00 (0·18) | 0·18 (0·01) | 0·034 (0·003) | 1·01 (0·18) | 0·17 (0·01) | 0·033 (0·003) | 0·99 | 1·00 (0·18) | 0·17 (0·01) | 0·033 (0·003) | 0·98 |
| 3 | 1·00 (0·21) | 0·21 (0·01) | 0·045 (0·004) | 0·99 (0·19) | 0·19 (0·01) | 0·037 (0·004) | 0·81 | 0·99 (0·19) | 0·19 (0·01) | 0·037 (0·004) | 0·81 |
| 4 | 1·02 (0·24) | 0·20 (0·02) | 0·059 (0·007) | 1·03 (0·24) | 0·20 (0·02) | 0·060 (0·007) | 1·01 | 1·03 (0·24) | 0·19 (0·02) | 0·059 (0·023) | 1·00 |
| 5 | 1·00 (0·27) | 0·23 (0·03) | 0·075 (0·007) | 1·00 (0·26) | 0·22 (0·02) | 0·066 (0·006) | 0·88 | 1·01 (0·26) | 0·20 (0·02) | 0·068 (0·007) | 0·91 |
| 6 | 1·00 (0·31) | 0·27 (0·02) | 0·095 (0·009) | 0·98 (0·27) | 0·24 (0·01) | 0·072 (0·007) | 0·76 | 0·98 (0·27) | 0·23 (0·01) | 0·073 (0·007) | 0·77 |
| 7 | 1·00 (0·24) | 0·22 (0·01) | 0·058 (0·005) | 0·99 (0·21) | 0·19 (0·01) | 0·044 (0·004) | 0·77 | 1·00 (0·21) | 0·19 (0·01) | 0·044 (0·004) | 0·77 |

Table IV. Monte Carlo means of the estimate of the standard error using OLS and GLS, with an estimated covariance matrix, based on uncensored data

| Set-up | $n = 50$ | | $n = 250$ | |
|---|---|---|---|---|
| | OLS | GLS | OLS | GLS |
| 1 | 0·35 | 0·35 | 0·15 | 0·15 |
| 2 | 0·41 | 0·39 | 0·18 | 0·17 |
| 3 | 0·49 | 0·43 | 0·22 | 0·19 |
| 4 | 0·41 | 0·41 | 0·18 | 0·17 |
| 5 | 0·49 | 0·49 | 0·22 | 0·19 |
| 6 | 0·66 | 0·67 | 0·29 | 0·26 |
| 7 | 0·50 | 0·43 | 0·22 | 0·19 |

We explored a variety of alternative methods to obtain standard errors of the regression coefficients, such as a sandwich estimator (see (3)), the asymptotic normal data augmentation,[11] the approximate Bayesian bootstrap data augmentation,[17] and the (regular) bootstrap.[25] Most of these methods could be equally well applied to the marginal, the semi-marginal and the bivariate log-spline approach. Since the downward bias seemed to be most severe for $n = 50$, while otherwise the three approaches had similar problems, we decided to focus on the semi-marginal approach with $n = 50$. The standard errors provided by the sandwich estimator were very similar

 

Table V. Monte Carlo mean of bootstrap standard error estimates of the estimated regression coefficients for the semi-marginal approach (with Monte Carlo standard deviations in parentheses), for the sample size $n = 50$ based on 50 bootstrap replications

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| SE($\hat{\beta}$) | 0·37 (0·06) | 0·41 (0·07) | 0·45 (0·08) | 0·50 (0·11) | 0·58 (0·12) | 0·59 (0·13) | 0·47 (0·10) |

to those in Tables II; the asymptotic normal data augmentation scheme and the approximate Bayesian bootstrap yielded somewhat improved standard error estimates, but only the bootstrap appeared to yield unbiased estimates. We believe that the reason of this may be that the correlation structure is already satisfactorily taken care of by the GLS procedure, but that the bootstrap standard errors acknowledge most the uncertainty in the Kaplan–Meier and the bivariate log-spline estimate, which we believe to be the main reason of the bias.

In our implementation of the bootstrap, for each bootstrap replication we resampled $n$ clusters from the original data and we computed the estimate of the regression coefficient. The bootstrap estimate of the standard error for the regression coefficient is now the standard deviation of the bootstrap estimates of the regression coefficient. Table V presents the results of the bootstrap standard error estimates for the semi-marginal approach and $n = 50$ based on 50 bootstrap replications. These standard error estimates are very similar to the standard deviations of $\tilde{\beta}$ for the semi-marginal approach in Table II, suggesting that any bias in these estimates is negligible.

It is somewhat surprising that the bivariate approach is not better than the semi-marginal approach. Thus, almost all of the efficiency gain comes from using GLS. There are several reasons why estimating the bivariate density may not improve the efficiency. First, it is hard to estimate a bivariate density based on a fairly small sample. Indeed, we do notice that for set-ups 3 and 7 the bivariate approach improves markedly with an increased sample size. Second, the bivariate approach would only improve the clusters with censored observations, which for most set-ups is only a smaller percentage, while for the set-ups with a high percentage of censoring (set-ups 5 and 6) only a smaller number of uncensored cases is available to estimate the bivariate density.

Note that the relative efficiency for the set-ups with heavy censoring (set-ups 4 and 5) is smaller than those for the same set-ups with less censoring (set-ups 2 and 3). In particular, when the censoring is heavy and the correlation is low (set-up 4), the efficiency gain is apparently completely offset by the extra variability that is introduced in the estimation procedures of the two more complicated approaches. This is consistent with Oakes' result in the context of the Cox regression.[26] It appears that we lose information about the within-cluster correlation from the heavy censoring.

## 5. EXAMPLE

As an illustration we apply the methods to the ventilation tubes duration data set given by Le and Lindgren.[4] One of the most common childhood diseases is the inflammation of the middle ear or otitis media (OM). One common surgical intervention is to install ventilating tubes inside two ears of a patient. This treatment reduces the incidence of OM episodes and improves hearing as

 

long as the tubes are still functioning. From February 1987 to January 1990, a clinical trial was conducted with children who would have therapeutic myringotomy for tympanostomy tube placement. Subjects were randomly chosen to receive two week trials of prednisone and sulphamethoprim treatment soon after surgery. The control group did not receive any additional treatment after surgery. The goal was to investigate whether the treatment prolongs the life of the tubes. Thus, the survival time is the lifetime of a functioning tube. There were a total of 78 subjects of which 40 were in the treatment group. Since each patient had one tube in each of his/her two ears, we expect that the survival times of the two tubes in the same subject are correlated. Only 12 out of 156 survival times were censored. We take the logarithm of the survival time as our response, and code the covariate $X_{i1} = X_{i2} = 1$ if the $i$th subject is in the treatment group, and 0 otherwise. The estimated regression coefficient from the marginal, semi-marginal and bivariate approaches are, respectively, 0·309, 0·304 and 0·306. Thus, the treatment seems to prolong the lifetime of the tubes. The standard errors of the regression coefficient estimate from the three approaches are 0·141, 0·161 and 0·162. For the semi-marginal approach we used the bootstrap (with 1000 bootstrap replications) to obtian a standard error estimate 0·158. The 95 per cent bootstrap percentile confidence interval is (0·019, 0·629). Hence the regression coefficient, the effect of prolonging the lifetime of tubes by the treatment, is different from 0 with statistical significance. (A similar conclusion was reached by Le and Lindgren[4]).

In addition to the regression coefficient, the bivariate log-spling procedure also provides us with an estimate $G$ of the joint baseline distribution. Such a distribution may be useful in that it gives a graphical interpretation of the correlation structure, which may shine more light on the dependence. See Kooperberg[15] for some bivariate log-spline density estimates and their applications.

## 6. DISCUSSION

To investigate whether we can improve the performance of the marginal approach in the bivariate linear regression, we proposed a semi-marginal approach, which takes account of the possible within-cluster correlation in fitting a linear regression model with imputed complete data, and a bivariate approach, which accounts for the within-cluster correlation both when fitting the linear model and when imputing the censored observations. We used the bivariate log-spline density method to estimate the bivariate joint distribution. Both new approaches, which perform similarly, improve over the marginal approach. Except when the censoring is too heavy or the within-cluster correlation is small, we recommend using the semi-marginal or bivariate approach.

We use Wei and Tanner's poor man's data augmentation algorithm to impute censored observations. It is conceptually simple, and easy to implement. After imputing, we can take advantage of existing techniques, such as GLS, to handle *complete* multivariate data. Though multiple imputation also has the potential to automatically take account of between-imputation variability, our simulation results suggest that the standard error estimates from the poor man's data augmentation may be slightly downward biased. If this bias is a problem, the bootstrap methods appear to be a promising approach for obtaining unbiased standard error estimates.

ACKNOWLEDGEMENTS

The first author is grateful to Drs. Tom Louis, John Connett and Chap Le for many helpful discussions. The first author was partially supported by a University of Minnesota Grant-in-aid. The second author was supported in part by NIH grand CA 74841.

 

## REFERENCES


1. Diabetic Retinopathy Study Research Group. 'Diabetic retionopathy study', *Investigative Opthalmology and Visual Science*, **21**, 149–226 (1981).
2. Duffy, D. L., Martin, N. G. and Matthews, J. D. 'Appendectomy in Australian twins', *American Journal of Human Genetics*, **47**, 590–592 (1995).
3. Haue, M., Harvald, B., Fischer, M., Jensen, K., Gotlieb, Juel-Nielsen, N., Reabild, J., Shapiro, R. and Videbech, T. 'The Danish twin register', *Acta Genetics Medicae et Gemillologiae*, **17**, 315–332 (1968).
4. Le, C. T. and Lindgren, B. R. 'Duration of ventilating tubes; a test for comparing two clustered samples of censored data', *Biometrics*, **52**, 328–334 (1996).
5. Wei, L. J. 'The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis (with discussion)', *Statistics in Medicine*, **11**, 1871–1879 (1992).
6. Cox, D. R. 'Some remarks on the analysis of survival data', in *proceedings of the First Seattle Symposium in Biostatistics*: *Survival Analysis*, Springer, 1997.
7. Buckley, J. and James, I. 'Linear regression with censored data', *Biometrika*, **66**, 429–436 (1979).
8. Louis, T. A. 'Non-parametric analysis of an accelerated failure time model', *Biometrika*, **68**, 381–390 (1981).
9. Tsiatis, A. A. 'Estimating regression parameters using linear rank test for censored data', *Annals of Statistics*, **18**, 354–372 (1990).
10. Ritov, Y. 'Estimation in a linear regression model with censored data', *Annals of Statistics*, **18**, 303–328 (1990).
11. Wei, G. C. G. and Tanner, M. A. 'Applications of multiple imputation to the analysis of censored regression data', *Biometrics*, **47**, 1297–1309 (1991).
12. Lee, C. W., Wei, L. J. and Ying, Z. 'Linear regression analysis for highly stratified failure time data', *Journal of the American Statistical Association*, **88**, 557–565 (1993).
13. Ying, Z. and Wei, L. J. 'The Kaplan–Meier estimate for dependent failure time observations', *Journal of Multivariate Analysis*, **50**, 17–29 (1994).
14. van der Laan, M. J. 'Efficient estimation in the bivariate censoring model and repairing NPMLE', *Annals of Statistics*, **24**, 596–627 (1996).
15. Kooperberg, C. 'Bivariate density estimation with an application to survival analysis', *Journal of Computational and Graphical Statistics*, **7**, 322–341 (1998).
16. Tanner, M. A. and Wong, W. H. 'The calculation of posterior distributions by data augmentation', *Journal of the American Statistical Association*, **82**, 528–549 (1987).
17. Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*, Wiley, 1987.
18. Schenker, N. and Welsh, A. H. 'Asymptotic results for multiple imputation', *Annals of Statistics*, **16**, 1550–1566 (1988).
19. Tanner, M. A. and Wong, W. H. 'An application of imputation to an estimation problem in grouped lifetime analysis', *Technometrics*, **29**, 23–32 (1987).
20. Rubin, D. B. and Schenker, N. 'Multiple imputation for interval estimation from simple random samples with ignorable nonresponse', *Journal of the American Statistical Association*, **81**, 366–374 (1986).
21. Oakes, D. 'Bivariate survival models induced by frailties', *Journal of the American Statistical Association*, **84**, 487–493 (1989).
22. Hsu, L. and Prentice, R. L. 'On assessing the strength of dependency between failure time variates', *Biometrika*, **83**, 491–506 (1996).
23. Shih, J. H. and Louis, T. A. 'Inferences on the association parameter in copula models for bivariate survival data', *Biometrics*, **51**, 1384–1399 (1995).
24. Liang, K. -Y. and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).
25. Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap* , Chapman & Hall, London, 1993.
26. Oakes, D. 'Frailty models for multiple event times', in *Survival Analysis*: *State of the Art*, Kluwer, 1992.


 