On Dealing with Censored Largest Observations under Weighted Least Squares

By Md Hasinur Rahaman Khan
ISRT, University of Dhaka, Bangladesh
Email: hasinur@isrt.ac.bd

J. Ewart H. Shaw
Department of Statistics, University of Warwick, UK
Email: Ewart.Shaw@warwick.ac.uk

Summary

When observations are subject to right censoring, weighted least squares with appropriate weights (to adjust for censoring) is sometimes used for parameter estimation. With Stute's weighted least squares method (Stute, 1993, 1994), when the largest observation ($Y^+_{(n)}$) is censored, it is natural to apply the redistribution to the right algorithm of Efron (1967). However, Efron's redistribution algorithm can lead to bias and inefficiency in estimation. This study explains the issues and proposes alternative ways of treating $Y^+_{(n)}$. The new schemes use penalized weighted least squares that is optimized by a quadratic programming approach, applied to the log-normal accelerated failure time model. The proposed approaches generally outperform Efron's redistribution approach and lead to considerably smaller mean squared error and bias estimates.

Some key words: Accelerated failure time (AFT) model and Efron's tail correction and Imputation and Right censoring and Weighted least squares

1. Introduction

The accelerated failure time (AFT) model is a linear regression model where the response variable is usually the logarithm of the failure time (Kalbfleisch & Prentice, 2002). Let $Y_{(1)}, \cdots, Y_{(n)}$ be the ordered logarithm of survival times, and $\delta_{(1)}, \cdots, \delta_{(n)}$ the corresponding censoring indicators. Then the AFT model is defined by

$$Y_i = \alpha + X_i^T \beta + \varepsilon_i, \quad i = 1, \cdots, n, \tag{1}$$

where $Y_i = \log(T_i)$, X_i is the covariate vector, α is the intercept term, β is the unknown $p \times 1$ vector of true regression coefficients and the ε_i's are independent and identically distributed (iid) random variables whose common distribution may take a parametric form, or may be unspecified, with zero mean and bounded variance. For example, a log-normal AFT model is obtained if the error term ε is normally distributed. As a result, we have a log linear type model that appears to be similar to the standard linear model that is typically estimated using ordinary least squares (OLS). But this AFT model can not be solved using OLS because it can not handle censored data. The trick to handle censored data turns out to introduce weighted least squares method, where weights are used to account for censoring.

There are many studies where weighted least squares is used for AFT models (Huang et al., 2006; Hu & Rao, 2010; Khan & Shaw, 2012). The AFT model (1) is solved using a penalized version of Stute's weighted least squares method (SWLS) (Stute, 1993, 1996). The SWLS estimate $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ of $\theta = (\alpha, \beta)$ is defined by

$$\hat{\theta} = \underset{\theta}{\arg\min} \left[\frac{1}{2} \sum_{i=1}^{n} w_i \left(Y_{(i)} - \alpha - X_{(i)}^T \beta \right)^2 \right], \tag{2}$$

where $\frac{1}{2}$ is a normalizing constraint for convenience, and the w_i's are the weights which are typically determined by two methods in the literature. One is called inverse probability of censoring weighting (IPCW) and the other is called Kaplan−Meier weight which is based on the jumps of the K−M estimator. The IPCW approach is used in many studies in survival analysis (e.g. Robins & Finkelstein, 2000; Satten & Datta, 2001). The K−M weighting approach is also widely used in many studies such as Stute (1993, 1994, 1996), Stute & Wang (1994), Hu & Rao (2010), Khan & Shaw (2012). The SWLS method in Equation (2) uses the K−M weights to account for censoring.

The data consist of (T_i^*, δ_i, X_i), $(i = 1, \cdots, n)$, where $t_i^* = \min(t_i, c_i)$ and $\delta_i = I(t_i \leq c_i)$ where t_i and c_i represent the realization of the random variables T_i and C_i respectively. Let the ordered failure and censoring times be t_j^* $(j = 1, \cdots, n)$, d_j be the number of individuals who fail at time t_j and e_j be the number of individuals censored at time t_j. Then the K−M estimate of $S(t) = \mathrm{P}(T_i > t)$ is defined as

$$\widehat{S}(t) = \prod_{\{j: t_j \leq t\}} \left(1 - \frac{d_j}{r_j} \right), \tag{3}$$

where $r_j = \sum_{i=1}^{n} I(t_j \geq t)$ is the number of individuals at risk at time t. In Stute (1993, 1996) the K−M weights are defined as follows

$$w_1 = \frac{\delta_{(1)}}{n}, \; w_i = \frac{\delta_{(i)}}{n - i + 1} \prod_{j=1}^{i-1} \left(\frac{n - j}{n - j + 1} \right)^{\delta_{(j)}}, \quad j = 2, \cdots, n. \tag{4}$$

Note that this assigns $\frac{1}{n}$ weight to each observation if all observations in the dataset are uncensored.

As can be observed, the K–M weighting method (Equation 4) gives zero weight to the censored observations $Y_{(.)}^+$. The method also gives zero weight to the largest observation if it is censored $\delta_{(n)} = 0$. Furthermore we know from the definition of the K−M estimator (Equation (3)) that the K−M estimator $\widehat{S}(t)$ is not defined for $Y > Y_{(n)}^+$ i.e.

$$\widehat{S}(t) = \begin{cases} \prod_{\{j | t_j \leq t\}} \left(1 - \frac{d_j}{r_j} \right), & \text{for} \quad t \leq T_{(n)} \\ \begin{cases} 0, & \text{if } \delta_{(n)} = 1 \\ \text{undefined}, & \text{if } \delta_{(n)} = 0 \end{cases}, & \text{for} \quad t > T_{(n)}. \end{cases} \tag{5}$$

This problem is usually solved by making a tail correction to the K−M estimator, known as the redistribution to the right algorithm proposed by Efron (1967). Under this approach $\delta_{(n)} = 0$ is reclassified as $\delta_{(n)} = 1$ so that the K−M estimator drops to zero at $Y_{(n)}^+$ and beyond, leading to obtaining proper (weights adding to one) weighting scheme.

Several published studies give zero weight to the observation $Y^+_{(n)}$ (e.g. Huang et al., 2006; Datta et al., 2007). This has adverse consequences, as shown below.

1·1. *An Illustration*

In this study we consider only the K−M weights. Table 1 presents the hypothetical

Table 1. *Survival times for 10 rats and their corresponding K–M weights with tail correction (w_1) and without tail correction (w_0)*

Rat	1	2	3	4	5	6	7	8	9	10
T_i	9	13	13+	18	23	28+	31	34+	45	48+
w_0	0.100	0.100	0.000	0.114	0.114	0.000	0.143	0.000	0.214	0.000
w_1	0.100	0.100	0.000	0.114	0.114	0.000	0.143	0.000	0.214	0.214

survival times for ten rats, subject to right censoring. The table also presents the weight calculations with (w_1) and without (w_0) tail correction.

The Table 1 reveals that weighting without tail correction causes improper weighting scheme. For the AFT model analyzed by weighted least squares as defined by Equation (2), the improper weights will not contribute to the term $\frac{1}{2}\sum_{i=1}^{n} w_i\,(Y_{(i)} - \alpha - X^T_{(i)}\beta)^2$ for the observation $Y^+_{(n)}$. Since the term $w_i\,(Y_{(i)} - \alpha - X^T_{(i)}\beta)^2$ is non-negative this leads to a smaller value of weighted residual squares compared to its actual value, resulting in a biased estimate for β. As the censoring percentage $P_{\%}$ increases, the chance of getting the censored observation $Y^+_{(n)}$ also increases.

Therefore, both approaches with and without the tail correction affect the underlying parameter estimation process, giving biased and inefficient estimates in practice. In the following section we introduce some alternative options of dealing with the a censored largest observation. In the study we only consider datasets whose the largest observation is censored.

2. Penalized SWLS

Here we introduce a ℓ_2 penalized WLS method to solve the AFT model (1). We first adjust $\mathbf{X}_{(i)}$ and $Y_{(i)}$ by centering them by their weighted means

$$\bar{X}_w = \frac{\sum_{i=1}^{n} w_i\, X_{(i)}}{\sum_{i=1}^{n} w_i}, \qquad \bar{Y}_w = \frac{\sum_{i=1}^{n} w_i\, Y_{(i)}}{\sum_{i=1}^{n} w_i}.$$

The weighted covariates and responses become $X^w_{(i)} = (w_i)^{1/2}(X_{(i)} - \bar{X}_w)$ and $Y^w_{(i)} = (w_i)^{1/2}(Y_{(i)} - \bar{Y}_w)$ respectively, giving the weighted data $(Y^w_{(i)}, \delta_{(i)}, X^w_{(i)})$. By replacing the original data $(Y_{(i)}, \delta_{(i)}, X_{(i)})$ with the weighted data, the objective function of the SWLS (2) becomes

$$\ell(\beta) = \frac{1}{2}\sum_{i=1}^{n}(Y^w_{(i)} - X^{w\,T}_{(i)}\beta)^2.$$

Then, the ridge penalized estimator, $\hat{\beta}$, is the solution that minimizes

$$\ell(\beta) = \frac{1}{2}\sum_{i=1}^{n}(Y_{(i)}^{w} - X_{(i)}^{w\,T}\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2, \tag{6}$$

where λ_2 is the ridge penalty parameter. The reason for choosing ℓ_2 penalized estimation is to deal with any collinearity among the covariates. We use $\lambda_2 = 0.01\sqrt{2\log p}$ for the log-normal AFT model because the $\sigma\sqrt{2\log p}$ term is a natural adjustment for the number of variables (p) for model with Gaussian noise (Candes & Tao, 2007). The scaling factor (0.01) was found empirically to give good results in our simulations.

We further develop this penalized WLS (6) in the spirit of the study by Hu & Rao (2010). The objective function of the modified penalized WLS is defined in matrix form below

$$\ell(\beta) = \frac{1}{2}(Y_u^w - X_u^w\beta)^T(Y_u^w - X_u^w\beta) + \frac{1}{2}\lambda_2\beta^T\beta \;\text{ subject to } Y_{\bar{u}}^w \le X_{\bar{u}}^w\beta, \tag{7}$$

where Y_u^w and X_u^w are the response variables and the covariates respectively both corresponding to the uncensored data. For censored data they are denoted by $Y_{\bar{u}}^w$ and $X_{\bar{u}}^w$ respectively. The censoring constraints $Y_{\bar{u}}^w \le X_{\bar{u}}^w\,\beta$ arise from the right censoring assumption (Hu & Rao, 2010). The optimization of Equation (7) is then carried out using a standard quadratic programming that has the form

$$\underset{b}{\arg\min}\;\left[-d^Tb + \frac{1}{2}b^TDb\right] \qquad \text{subject to } Ab \ge b_0, \tag{8}$$

where

$$b = \left[\beta\right]_{p\times 1}, \quad d = \left[X_u^w\right]_{1\times p}, \quad D = \left[X_u^{wT}X_u^w + \lambda_2 I_{p\times p}\right]_{p\times p},$$

$$A = \left[X_{\bar{u}}^w\right]_{n_{\bar{u}}\times p}, \quad b_0 = \left[Y_{\bar{u}}^w\right]_{n_{\bar{u}}\times 1}.$$

Here $n_{\bar{u}}$ indicates the number of censored observations.

3. Proposed Approaches for Imputing $Y_{(n)}^{+}$

Let $T_{\bar{u}}$ be the true log failure times corresponding to the unobserved so that $T_{\bar{u}i} > Y_{\bar{u}i}$ for the i-th censored observation. We propose the following five approaches for imputing the largest observation $Y_{\bar{u}(n)}$. The first four approaches are based on the well known Buckley−James (1979) method of imputation for censored observations incorporating Efron's tail correction. The last approach is indirectly based on the mean imputation technique as discussed in Datta (2005). Therefore, under the first four approaches, the lifetimes are assumed to be modeled using the associated covariates whereas under the last approach there is no such assumption.

3·1. *Adding the Conditional Mean or Median*

The key idea of the Buckley−James method for censored data is to replace the censored observations (i.e. $Y_{\bar{u}i}$) by their conditional expectations given the corresponding censoring times and the covariates, i.e. $E(T_{\bar{u}i}|T_{\bar{u}i} > Y_{\bar{u}i}, X_i)$. Let ξ_i is the error term associated with the data (T_i^*, δ_i, X_i) i.e. $\xi_i = Y_i - X_i^T\beta$ where $Y_i = \log(T_i^*)$ such that solving the equation $\sum_{i=1}^{n} X_i^T(Y_i - X_i^T\beta) = \sum_{i=1}^{n} X_i^T\xi_i = 0$ yields the least squares estimates for β.

According to the Buckley−James method the quantity $E(T_{\bar{u}i}|T_{\bar{u}i} > Y_{\bar{u}i}, X_i)$ for the i-th censored observation is calculated as

$$E(T_{\bar{u}i}|T_{\bar{u}i} > Y_{\bar{u}i}, X_i) = X_i^T\beta + E(\varepsilon_i|\varepsilon_i > \xi_i, X_i). \tag{9}$$

We do not impute the largest observation, $Y_{\bar{u}(n)}$ using Equation (9), rather we add the conditional mean of $(\varepsilon_i|\varepsilon_i > \xi_i, X_i)$ i.e. $E(\varepsilon_i|\varepsilon_i > \xi_i) = \tau_m$ (say) or the conditional median of $(\varepsilon_i|\varepsilon_i > \xi_i, X_i)$ i.e. Median$(\varepsilon_i|\varepsilon_i > \xi_i) = \tau_{md}$ (say) to $Y_{\bar{u}(n)}$. Here the quantity $(\varepsilon_i|\varepsilon_i > \xi_i, X_i)$ is equivalent to $(\varepsilon_i|\varepsilon_i > \xi_i)$ since $\varepsilon_i \perp X_i$ in linear regression (1). The quantity $Y_{\bar{u}(n)} + \tau_m$ or $Y_{\bar{u}(n)} + \tau_{md}$ is therefore a reasonable estimate of the true log failure time $T_{\bar{u}(n)}$ for the largest observation.

The quantity τ_m can be calculated by

$$\tau_m = E(\varepsilon_i|\varepsilon_i > \xi_i) = \int_{\xi_i}^{\infty} \varepsilon_i \frac{\mathrm{d}F(\varepsilon_i)}{1 - F(\xi_i)}, \tag{10}$$

where $F(\cdot)$ is the distribution function. distribution. Buckley & James (1979) show that the above $F(\cdot)$ can be replaced by its Kaplan-Meier estimator $\hat{F}(\cdot)$. Using this idea Equation (10) can now be written as below

$$\tau_m = \sum_{j:\xi_j > \xi_i} \xi_j \frac{\Delta\hat{F}(\xi_j)}{1 - \hat{F}(\xi_i)}, \tag{11}$$

where $\hat{F}$ is the Kaplan−Meier estimator of F based on $[(\xi_i,\, \delta_i),\, i = 1, \cdots, n]$ i.e.,

$$\widehat{F}(\xi_i) = 1 - \prod_{j:\xi_j > \xi_i} \left(1 - \frac{\delta_i}{\sum_{j=1}^{n} 1_{\{\xi_j \geq \xi_i\}}}\right). \tag{12}$$

The conditional median τ_{md} can be calculated from the following expression.

$$\int_{\xi_i}^{\tau_{md}} \frac{\mathrm{d}F(\varepsilon_i)}{1 - F(\xi_i)} = 0.5. \tag{13}$$

After replacing $F(\cdot)$ by its K−M estimator $\hat{F}(\cdot)$ (12) Equation (13) can be written as below

$$\sum_{j:\xi_j > \xi_i} \frac{\Delta\hat{F}(\xi_j)}{1 - \hat{F}(\xi_i)} = 0.5. \tag{14}$$

3·2. *Adding the Resampling-based Conditional Mean and Median*

The approaches are similar to adding the conditional mean and median as discussed in section (3·1) except that τ_m and τ_{md} are calculated using a modified version of an iterative solution to the Buckley−James estimating method (Jin et al., 2006) rather than the original Buckley−James (1979) method. We have followed the iterative Buckley−James estimating method (Jin et al., 2006) along with the associated imputation technique because it provides a class of consistent and asymptotically normal estimators. We have modified this iterative procedure by introducing a quadratic programming based weighted least square estimator (Hu & Rao, 2010) as the initial estimator. Under this scheme we replace the unobserved $Y_{\bar{u}(n)}$ by $Y_{\bar{u}(n)} + \tau_m^*$ or $Y_{\bar{u}(n)} + \tau_{md}^*$ where τ_m^* and τ_{md}^* are the

resampling based conditional mean and median calculated by

$$\tau_m^* = \sum_{j:\xi_j > \xi_i} \xi_j \frac{\Delta \hat{F}^*(\xi_j)}{1 - \hat{F}^*(\xi_i)}, \tag{15}$$

and

$$\sum_{j:\xi_j > \xi_i} \frac{\Delta \hat{F}^*(\xi_j)}{1 - \hat{F}^*(\xi_i)} = 0.5 \tag{16}$$

respectively. Here $\widehat{F}^*(\xi_i)$ is calculated using Equation (12) based on the modified iterative Buckley−James estimating method. The procedure is described below.

Buckley & James (1979) replaced the i-th censored $Y_{\bar{u}i}$ by $E(T_{\bar{u}i}|T_{\bar{u}i} > Y_{\bar{u}i}, X_i)$, yielding

$$\widehat{Y}_i(\beta) = \delta_i Y_i + (1 - \delta_i) \left[\int_{\xi_i}^{\infty} \varepsilon_i \frac{\mathrm{d}\hat{F}(\varepsilon_i)}{1 - \hat{F}(\xi_i)} + X_i^T \beta \right],$$

where $\hat{F}$ is the K−M estimator of F based on the transformed data (ξ_i, δ_i) and is defined by the Equation (12). The associated Buckley−James estimating function $U(\beta, b) = \sum_{i=1}^n (X_i - \bar{X})\{\hat{Y}_i(b) - X_i^T\beta\}$ is then defined by $U(\beta, b) = \sum_{i=1}^n (X_i - \bar{X})\{\hat{Y}_i(b) - \bar{Y}(b) - (X_i - \bar{X}^T\beta)\}$ for $\bar{Y}(b) = n^{-1}\sum_{i=1}^n \hat{Y}_i(b)$. The Buckley−James estimator $\hat{\beta}_{bj}$ is the root of $U(\beta, \beta) = 0$. This gives the following solution:

$$\beta = \mathrm{L}(b) = \left\{ \sum_{i=1}^n (X_i - \bar{X})^{\bigotimes 2} \right\}^{-1} \left[\sum_{i=1}^n (X_i - \bar{X})\{\hat{Y}_i(b) - \bar{Y}(b)\} \right], \tag{17}$$

where $a^{\bigotimes 2}$ means aa^T for a vector. The expression (17) leads to following iterative algorithm.

$$\hat{\beta}_{(m)} = \mathrm{L}(\hat{\beta}_{(m-1)}), \qquad m \geq 1. \tag{18}$$

In Equation (18) we set the initial estimator $\hat{\beta}_{(0)}$ to be the penalized weighted least square estimator $\hat{\beta}_{qp}$ that is obtained by optimizing the objective function specified by the Equation (7). The initial estimator $\hat{\beta}_{qp}$ is a consistent and asymptotically normal estimator such as the Gehan-type rank estimator that was used as the initial estimator in Jin et al. (2006). Therefore using $\hat{\beta}_{qp}$ as the initial estimator will satisfy the following corollary that immediately follows from Jin et al. (2006).

COROLLARY 1. *The penalized weighted least squares estimator $\hat{\beta}_{qp}$ leads to a consistent and asymptotically normal $\hat{\beta}_{(m)}$ for each fixed m. In addition, $\hat{\beta}_{(m)}$ is a linear combination of $\hat{\beta}_{qp}$ and the Buckley−James estimator $\hat{\beta}_{bj}$ in that*

$$\widehat{\beta}_{(m)} = (I - D^{-1}A)^m \hat{\beta}_{qp} + \{I - (I - D^{-1}A)^m\}\hat{\beta}_{bj} + o_p(n^{-\frac{1}{2}})$$

where I is the identity matrix, $D := \lim_{n\to\infty} n^{-1} \sum_{i=1}^n (X_i - \bar{X})^{\bigotimes 2}$ is the usual slope matrix of the least-squares estimating function for the uncensored data, and A is a slope matrix of the Buckley−James estimation function as defined in Jin et al. (2006).

Jin et al. (2006) also developed a resampling procedure to approximate the distribution of $\hat{\beta}_{(m)}$. Under this procedure n iid positive random variables $Z_i (i = 1, \cdots, n)$ with $E(Z_i) = \text{var}(Z_i) = 1$ are generated. Then define

$$\widehat{F}^*(\xi_i) = 1 - \prod_{j:\xi_j > \xi_i} \left(1 - \frac{Z_i \delta_i}{\sum_{j=1}^n Z_i 1_{\{\xi_j \geq \xi_i\}}}\right), \tag{19}$$

and

$$\widehat{Y}_i^*(\beta) = \delta_i Y_i + (1 - \delta_i) \left[\int_{\xi_i}^{\infty} \varepsilon_i \frac{\mathrm{d}\hat{F}^*(\varepsilon_i)}{1 - \hat{F}^*(\xi_i)} + X_i^T \beta\right],$$

and

$$\mathrm{L}^*(b) = \left\{\sum_{i=1}^n Z_i (X_i - \bar{X})^{\bigotimes 2}\right\}^{-1} \left[\sum_{i=1}^n Z_i (X_i - \bar{X})\{\hat{Y}_i^*(b) - \bar{Y}^*(b)\}\right]. \tag{20}$$

Equation (20) then leads to an iterative process $\hat{\beta}^*_{(m)} = \mathrm{L}^*(\hat{\beta}^*_{(m-1)})$, $m \geq 1$. The initial value $\hat{\beta}^*_{(0)}$ of the iteration process becomes $\hat{\beta}^*_{qp}$ which is the optimized value of

$$\frac{1}{2} Z (Y_u^w - X_u^w \beta)^T (Y_u^w - X_u^w \beta) + \frac{1}{2} \lambda_2 \beta^T \beta, \qquad \text{subject to } Z Y_{\bar{u}}^w \leq Z X_{\bar{u}}^w \beta. \tag{21}$$

This objective function (21) is obtained from the function specified in Equation (7). For a given sample $(Z_1, \cdots, Z_n)$, the iteration procedure $\widehat{\beta}^*_{(k)} = \mathrm{L}^*(\widehat{\beta}^*_{(k-1)})$ yields a $\widehat{\beta}^*_{(k)}$ $(1 \leq k \leq m)$. The empirical distribution of $\widehat{\beta}^*_{(m)}$ is based on a large number of realizations that are computed by repeatedly generating the random sample $(Z_1, \cdots, Z_n)$ a reasonable number of times. Then the empirical distribution is used to approximate the distribution of $\widehat{\beta}_{(m)}$. See Jin et al. (2006) for more details.

3·3. *Adding the Predicted Difference Quantity*

Suppose $Y^*_{\bar{u}i}$ is the modified failure time (imputed value) obtained using mean imputation technique to the censored $Y_{\bar{u}i}$. Under mean imputation technique the censored observation, $Y_{\bar{u}i}$ is replaced with the conditional expectation of T_i given $T_i > C_i$. The formula for estimating $Y^*_{\bar{u}i}$ is given by

$$Y^*_{\bar{u}i} = \{\hat{S}(C_i)\}^{-1} \sum_{t_{(r)} > C_i} \log(t_{(r)}) \triangle \widehat{S}(t_{(r)}), \tag{22}$$

where $\hat{S}$ is the K−M estimator of the survival function of T as defined by Equation (3), $\triangle \widehat{S}(t_{(r)})$ is the jump size of $\hat{S}$ at time $t_{(r)}$. This mean imputation approach is used in many other studies (For example, Datta (2005), Datta et al. (2007), Engler & Li (2009) etc.)

We note that using mean imputation technique the modified failure time $Y^*_{\bar{u}(n)}$ to the largest observation $Y_{\bar{u}(n)}$ can not be computed since the K−M estimator $\widehat{S}$ is undefined for $\delta_{(n)} = 0$. In particular, the quantity $\triangle \widehat{S}(t_{(r)})$ in Equation (22) can not be calculated for the n-th observation $\delta_{(n)} = 0$. This issue is clearly stated in Equation (5). Here we present a different strategy to impute $Y_{\bar{u}(n)}$. We assume that the mean imputation technique is used for imputing all other censored observations except the last largest censored

observation. Let us assume that ν be a non-negative quantity such that $Y^*_{\bar{u}(n)} - Y_{\bar{u}(n)} = \nu$. One can now estimate ν using many possible ways. Here we choose a very simple way that uses the imputed values obtained by the above mean imputation approach. We estimate ν as a predicted value based on the differences between the imputed values and the censoring times for all censored observations except the largest observations.

Suppose D_i for $i = 1, \cdots, (n_{\bar{u}i} - 1)$ represents the difference between the imputed value and the unobserved value for the i-th censored observation. So, the quantity ν can be treated as a possible component of the D family. We examine the relationship between D_i and $Y_{\bar{u}i}$ by conducting various numerical studies. Figures 1 and 2 show the most approximate relationships between D_i and $Y_{\bar{u}i}$ from two real datasets. Figure 1 is based on the Larynx dataset (Kardaun, 1983) (Details are given in the real data analysis section). Figure 2 is based on the Channing House dataset (Hyde, 1980) that also

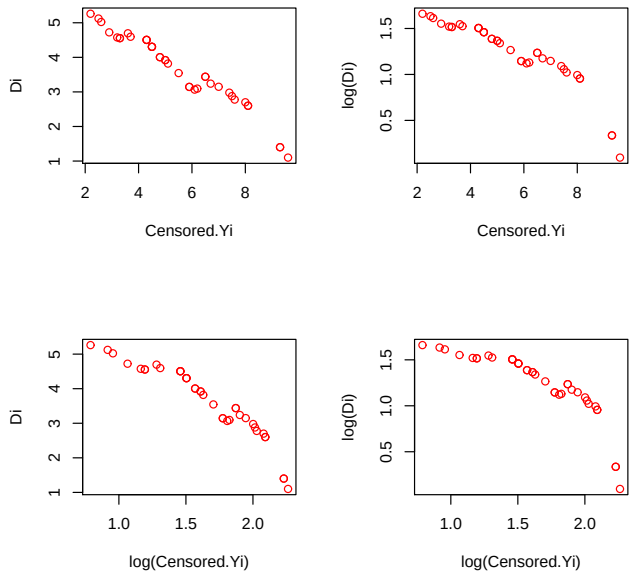

Fig. 1. Relationship between D_i and $Y_{\bar{u}i}$ for $i = 1, \cdots, n_{\bar{u}i} - 1$ based on the Larynx data

discussed in the section of real data analysis. Both male and female data have heavy censoring toward the right.

Both Figures 1 and 2 clearly suggest a negative linear relationship between D_i and $Y_{\bar{u}i}$. The trend based on other transformations (logarithmic of either D_i or $Y_{\bar{u}i}$ or both) appears to be nonlinear. Hence we set up a linear regression for D_i on $Y_{\bar{u}i}$ which is given by

$$D_i = \tilde{\alpha} + Y_{\bar{u}i}\,\tilde{\beta} + \tilde{\varepsilon}_i, \quad i = 1, \cdots, (n_{\bar{u}i} - 1), \tag{23}$$

where $\tilde{\alpha}$ is the intercept term, $\tilde{\beta}$ is the coefficient for the unobserved censored time, and $\tilde{\varepsilon}_i$ is the error term. We fit the model (23) with a WLS method that gives the objective

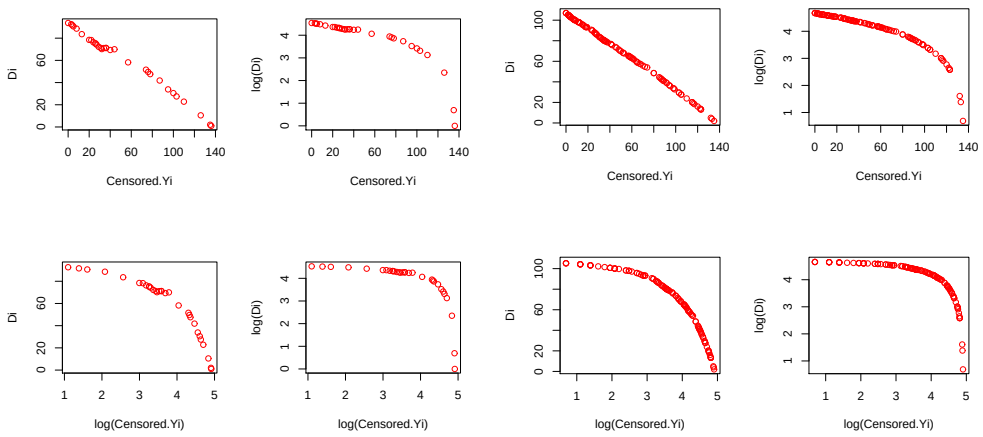

Fig. 2. Relationship between D_i and $Y_{\bar{u}i}$ for $i = 1, \cdots, n_{\bar{u}i} - 1$ based on Channing House data. Left two column plots based on male data and the right two column plots are based on female data

function.

$$\sum_{i=1}^{n_{\bar{u}i}-1} \tilde{w}_i \, (D_i - \tilde{\alpha} - Y_{\bar{u}i} \, \tilde{\beta})^2, \tag{24}$$

where $\tilde{w}_i$ are the data dependent weights. The weight for the i-th observation in Equation (24) is chosen by $\{Y_{\bar{u}(n)} - Y_{\bar{u}i}\}^{-1}$. We choose WLS method for fitting model (23) because it is observed from Figure 1 that the D_i occurs more frequently for the lower and middle censoring times than that for the higher censoring time. Finally the quantity ν is obtained by

$$\hat{\nu} = \begin{cases} \hat{D}(Y_{\bar{u}(n)}), & \text{if } \hat{D}(Y_{\bar{u}(n)}) > 0 \\ 0, & \text{if } \hat{D}(Y_{\bar{u}(n)}) \leq 0. \end{cases} \tag{25}$$

4. Estimation Procedures

The performance of the proposed imputation approaches along with Efron's (1967) redistribution technique is investigated with the AFT model evaluated by the quadratic program based Stute's weighted least squares method as discussed in Section (3·2). For convenience, let W_0 represents the estimation process when no imputation is done for $Y_{\bar{u}(n)}$, i.e. only Efron's (1967) redistribution algorithm is applied. Let W_{τ_m}, $W_{\tau_{md}}$, $W_{\tau_m^*}$, $W_{\tau_{md}^*}$, and W_ν represent estimation where $Y_{\bar{u}(n)}$ is imputed by adding the conditional mean, the conditional median, the resampling based conditional mean, the resampling based conditional median and the predicted difference quantity to the $Y_{\bar{u}(n)}$ respectively.

4·1. *W_0: Efron's Approach*

1. Set $\delta_{(n)} = 1$.
2. Solve Equation (7) using the QP approach (8) to estimate β.

4·2. W_{τ_m}: *Conditional Mean Approach*

1. Set $\delta_{(n)} = 1$ and solve Equation (7) using the QP approach (8) to estimate β.
2. Compute weighted least squares errors based on the estimated β from Step 1 and then calculate the K-M estimator of the errors using the Equation (12) and τ_m using Equation (11).
3. Add the quantity τ_m found in Step 2 to $Y_{\bar{u}(n)}$.
4. Get a new and improved estimate of β based on modified $Y_{\bar{u}(n)}$ found in Step 3 by solving Equation (7) using the QP approach (8).

4·3. $W_{\tau_{md}}$: *Conditional Median Approach*

The process of $W_{\tau_{md}}$ is similar to W_{τ_m} except that it uses τ_{md} instead of τ_m in Step 2 and Step 3.

4·4. $W_{\tau_m^*}$: *Resampling based Conditional Mean Approach*

1. Set $\delta_{(n)} = 1$ and solve Equation (20) to estimate β.
2. Compute weighted least squares errors based on the estimated β from Step 1 and then calculate the K-M estimator of the errors using the Equation (12) and τ_m^* using Equation (15).
3. Add the quantity τ_m^* found in Step 2 to $Y_{\bar{u}(n)}$.
4. Get a new and improved estimate of β based on modified $Y_{\bar{u}(n)}$ found in Step 3 by solving Equation (7) using the QP approach (8).

4·5. $W_{\tau_{md}^*}$: *Resampling based Conditional Median Approach*

The process of $W_{\tau_{md}^*}$ is similar to $W_{\tau_m^*}$ except that it uses τ_{md}^* instead of $W_{\tau_m^*}$ in Step 2 and Step 3.

4·6. W_ν: *Predicted Difference Quantity Approach*

1. Set $\delta_{(n)} = 1$ and compute the modified failure time using Equation (22).
2. Compute ν using Equation (25).
3. Add the quantity ν found in Step (ii) to $Y_{\bar{u}(n)}$.
4. Get an estimate of β based on modified $Y_{\bar{u}(n)}$ found in Step 3 by solving Equation (7) using the QP approach (8).

5. SIMULATION STUDIES

Here we investigates the performance of the imputation approaches using a couple of simulation examples. The datasets are simulated from the following log-normal AFT model, conditional on the largest observation being censored (i.e. $\delta_{(n)} = 0$):

$$Y_i = \alpha + X_i^T\beta + \sigma\varepsilon_i, \qquad \varepsilon_i \sim N(0,1) \text{ for } i = 1, \cdots, n. \tag{26}$$

The pairwise correlation (r_{ij}) between the i-th and j-th components of X is set to be $0.5^{|i-j|}$. The censoring times are generated using $U(a, 2a)$, where a is chosen such that pre-specified censoring rates $P_{\%}$ are approximated (see appendix for analytical detail). The true covariate $\mathbf{X_i} = (X_{i1}, X_{i2}, \cdots, X_{ip})^T$ are generated from $U(0, 1)$.

5·1. *First Example*

We choose $n = 100$, $p = 5$, and $\sigma = 1$, and $r_{ij} = 0$ and 0.5, three censoring rates 30%, 50%, and 70%. We choose $\beta_j = j + 1$ for $j = 1, \cdots, p$. The bias, variance, and mean squared error (MSE) for β are estimated by averaging the results from 1,000 runs.

Table 2. *Summary statistics for first simulation example $r_{ij} = 0$. Comparison between the imputation approaches W_0: Efron's redistribution, W_{τ_m}: conditional mean, $W_{\tau_{md}}$: conditional median, $W_{\tau_m^*}$: resampling based conditional mean, $W_{\tau_{md}^*}$: resampling based conditional median, and W_ν: predicted difference quantity.*

	$P_\% = 30$						$P_\% = 50$						$P_\% = 70$					
	W_0	W_{τ_m}	$W_{\tau_{md}}$	$W_{\tau_m^*}$	$W_{\tau_{md}^*}$	W_ν	W_0	W_{τ_m}	$W_{\tau_{md}}$	$W_{\tau_m^*}$	$W_{\tau_{md}^*}$	W_ν	W_0	W_{τ_m}	$W_{\tau_{md}}$	$W_{\tau_m^*}$	$W_{\tau_{md}^*}$	W_ν
Bias																		
β_1	0.391	0.363	0.380	0.338	0.367	0.408	0.467	0.402	0.450	0.403	0.430	0.472	0.729	0.595	0.665	0.488	0.588	0.709
β_2	0.655	0.616	0.639	0.605	0.622	0.678	0.629	0.535	0.596	0.498	0.562	0.636	0.958	0.733	0.841	0.678	0.716	0.942
β_3	0.839	0.789	0.820	0.777	0.799	0.867	0.877	0.761	0.841	0.748	0.793	0.885	1.299	1.030	1.143	0.921	0.998	1.270
β_4	0.988	0.935	0.963	0.911	0.941	1.018	0.981	0.835	0.940	0.806	0.886	0.986	1.539	1.176	1.361	1.059	1.196	1.494
β_5	1.226	1.160	1.201	1.142	1.177	1.258	1.271	1.087	1.225	1.065	1.162	1.282	2.193	1.812	1.992	1.642	1.798	2.151
Variance																		
β_1	0.154	0.156	0.155	0.162	0.159	0.152	0.283	0.348	0.290	0.328	0.296	0.291	0.483	0.632	0.523	0.890	0.594	0.520
β_2	0.174	0.178	0.178	0.187	0.181	0.174	0.282	0.324	0.285	0.331	0.294	0.290	0.631	0.742	0.645	0.828	0.664	0.657
β_3	0.142	0.146	0.143	0.146	0.143	0.144	0.279	0.319	0.273	0.327	0.268	0.291	0.536	0.666	0.561	0.774	0.581	0.581
β_4	0.182	0.181	0.181	0.190	0.182	0.180	0.276	0.316	0.268	0.371	0.273	0.301	0.618	0.628	0.543	0.665	0.550	0.616
β_5	0.160	0.164	0.161	0.164	0.158	0.166	0.246	0.281	0.239	0.274	0.236	0.277	0.620	0.690	0.561	0.784	0.605	0.659
MSE																		
β_1	0.305	0.286	0.298	0.275	0.292	0.317	0.499	0.506	0.489	0.487	0.478	0.511	1.009	0.980	0.959	1.119	0.934	1.017
β_2	0.601	0.556	0.585	0.550	0.567	0.632	0.675	0.607	0.638	0.576	0.607	0.692	1.543	1.273	1.347	1.280	1.171	1.539
β_3	0.844	0.786	0.814	0.748	0.781	0.894	1.046	0.895	0.978	0.883	0.895	1.072	2.219	1.721	1.861	1.614	1.571	2.189
β_4	1.157	1.054	1.107	1.019	1.067	1.213	1.237	1.010	1.148	1.016	1.055	1.270	2.981	2.004	2.389	1.779	1.975	2.842
β_5	1.663	1.508	1.601	1.465	1.542	1.746	1.860	1.460	1.737	1.406	1.584	1.918	5.422	3.969	4.525	3.472	3.832	5.277

Table 3. *Summary statistics for the first simulation example $r_{ij} = 0.5$. Comparison between the imputation approaches W_0: Efron's redistribution, W_{τ_m}: conditional mean, $W_{\tau_{md}}$: conditional median, $W_{\tau_m^*}$: resampling based conditional mean, $W_{\tau_{md}^*}$: resampling based conditional median, and W_ν: predicted difference quantity.*

	$P_\% = 30$						$P_\% = 50$						$P_\% = 70$					
	W_0	W_{τ_m}	$W_{\tau_{md}}$	$W_{\tau_m^*}$	$W_{\tau_{md}^*}$	W_ν	W_0	W_{τ_m}	$W_{\tau_{md}}$	$W_{\tau_m^*}$	$W_{\tau_{md}^*}$	W_ν	W_0	W_{τ_m}	$W_{\tau_{md}}$	$W_{\tau_m^*}$	$W_{\tau_{md}^*}$	W_ν
Bias																		
β_1	-0.295	-0.306	-0.301	-0.307	-0.303	-0.291	-0.329	-0.362	-0.345	-0.362	-0.349	-0.329	-0.387	-0.552	-0.542	-0.561	-0.524	-0.441
β_2	-0.042	-0.071	-0.065	-0.077	-0.070	-0.017	0.113	0.041	0.060	0.016	0.038	0.126	0.251	0.290	0.252	0.104	0.253	0.466
β_3	0.372	0.318	0.334	0.306	0.319	0.290	0.413	0.364	0.411	0.345	0.376	0.481	0.580	0.248	0.311	0.145	0.280	0.533
β_4	0.606	0.555	0.568	0.540	0.552	0.527	0.652	0.557	0.608	0.539	0.563	0.689	1.235	0.814	0.953	0.743	0.905	1.253
β_5	1.015	0.963	0.977	0.943	0.958	0.920	1.061	0.944	0.997	0.903	0.945	1.093	2.256	1.772	1.924	1.809	1.862	1.950
Variance																		
β_1	0.186	0.184	0.182	0.181	0.181	0.191	0.254	0.260	0.259	0.268	0.264	0.243	1.539	0.896	0.840	0.836	0.787	1.123
β_2	0.165	0.165	0.163	0.164	0.164	0.165	0.286	0.266	0.247	0.261	0.238	0.278	0.382	0.301	0.266	0.536	0.273	0.422
β_3	0.169	0.168	0.166	0.166	0.166	0.185	0.259	0.245	0.245	0.260	0.252	0.291	0.551	0.396	0.357	0.403	0.370	0.610
β_4	0.155	0.157	0.151	0.157	0.152	0.168	0.312	0.293	0.248	0.258	0.240	0.338	0.909	0.519	0.384	0.493	0.391	0.743
β_5	0.169	0.170	0.167	0.171	0.167	0.173	0.290	0.264	0.241	0.239	0.237	0.319	0.714	0.579	0.267	0.556	0.397	0.938
MSE																		
β_1	0.272	0.276	0.270	0.274	0.271	0.273	0.360	0.388	0.375	0.396	0.384	0.350	1.535	1.111	1.050	1.068	0.983	1.183
β_2	0.165	0.168	0.166	0.168	0.167	0.164	0.296	0.265	0.248	0.259	0.237	0.291	0.407	0.355	0.303	0.493	0.310	0.380
β_3	0.306	0.267	0.276	0.258	0.266	0.353	0.492	0.375	0.412	0.377	0.390	0.519	0.832	0.418	0.418	0.384	0.411	0.659
β_4	0.520	0.463	0.472	0.447	0.455	0.590	0.795	0.601	0.615	0.546	0.554	0.810	2.343	1.130	1.253	0.995	1.171	1.683
β_5	1.198	1.096	1.120	1.058	1.082	1.297	1.473	1.152	1.232	1.052	1.127	1.511	5.732	3.662	3.942	3.773	3.826	4.486

The results for uncorrelated and correlated datasets are reported in Table 2 and 3 respectively. The results generally suggest that the resampling based conditional mean adding $W_{\tau_m^*}$ and the resampling based conditional median adding $W_{\tau_{md}^*}$ provide the smallest MSE in particular, smaller than Efron's approach W_0 at all censoring levels. They seem to provide generally lower bias except for β_1 and β_2 in correlated case.

We also find that at lower and medium censoring levels both Efron's approach and the predicted difference quantity approach W_ν perform similarly to each other in terms of all three indicators. The predicted difference quantity approach performs less well but still better than Efron's approach for $P_\% = 70$. The MSE of β is usually decomposed by bias and variance i.e. $\text{MSE}(\hat{\beta}) = \text{Var}(\hat{\beta}) + [\text{Bias}(\hat{\beta})]^2$. If the bias is large, the bias then dominates the MSE. This may explain why the predicted difference quantity approach attains the highest MSE for β in two lower censoring levels.

The following simulation example is conducted particularly to understand how the effects of the imputation approaches change over the censoring levels and different correlation structures.

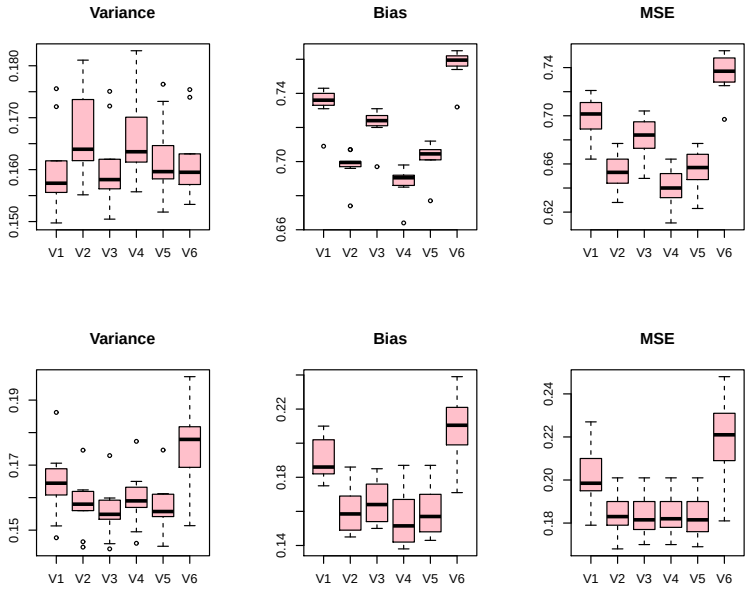

Fig. 3. Simulation study for second example under $P_{\%} = 30\%$. Box-plots for variance, bias, mean squared error of β. Here V1, V2, V3, V4, V5 and V6 represent the estimation approaches W_0: Efron's redistribution, W_{τ_m}: conditional mean, $W_{\tau_{md}}$: conditional median, $W_{\tau_m^*}$: resampling based conditional mean, $W_{\tau_{md}^*}$: resampling based conditional median, and W_ν: predicted difference quantity respectively. Graphs in first row show the results when r_{ij} is 0; graphs in second row show the results when r_{ij} is 0.5.

5·2. *Second Example*

We keep everything similar to the previous example except that $p = 10$ and $\beta_j = 3$ for $j = 1, \cdots, p$. The results are presented as summary box plots. Figures 3 to 5 represent the results corresponding to three censoring levels $P_{\%}$: 30, 50, and 70. The results for this example are similar to the results of the first example. The results for the uncorrelated datasets shown in Figure 4 suggest that adding the resampling based conditional mean gives the lowest bias and the lowest MSE, but yields the highest variance. For correlated datasets the other four approaches give lower variance, bias and MSE than either Efron's approach or the predicted difference quantity approaches. It is also noticed from the correlated data analysis that the median based approaches i.e. adding the conditional median or adding the resampling based conditional median perform slightly better than the mean based approaches i.e. adding the conditional mean or adding the resampling based conditional mean.

6. Real Data Examples

We present two well-known real data examples. The analysis for the first example is done with the larynx cancer data and for the second example, the analysis is done using

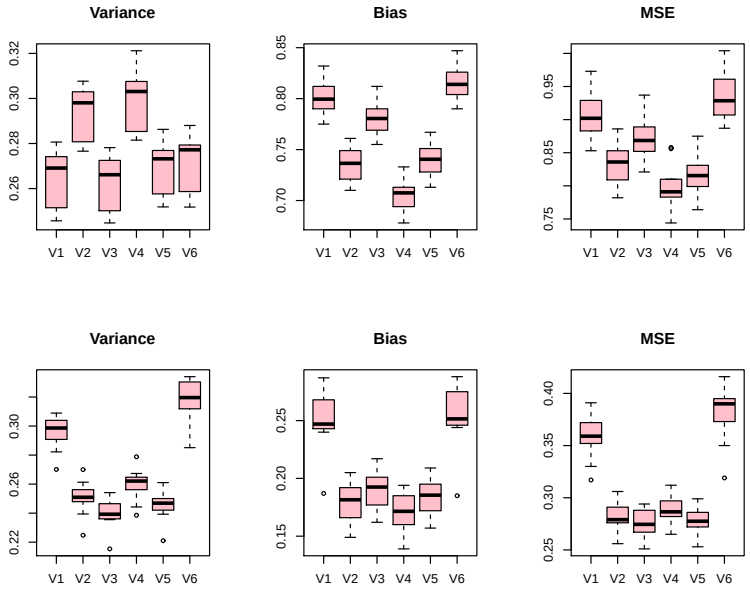

Fig. 4. Simulation study for second example under $P_{\%} = 50\%$. Box-plots for variance, bias, mean squared error of β. Here V1, V2, V3, V4, V5 and V6 represent the estimation approaches W_0: Efron's redistribution, W_{τ_m}: conditional mean, $W_{\tau_{md}}$: conditional median, $W_{\tau_m^*}$: resampling based conditional mean, $W_{\tau_{md}^*}$: resampling based conditional median, and W_ν: predicted difference quantity respectively. Graphs in first row show the results when r_{ij} is 0; graphs in second row show the results when r_{ij} is 0.5.

the Channing House data. The Channing House data is different from the larynx data since the data has heavy censoring and also has many largest censored observations.

6·1. *Larynx Data*

This example uses hospital data where 90 male patients were diagnosed with cancer of the larynx, treated in the period 1970−1978 at a Dutch hospital (Kardaun, 1983). An appropriate lower bound either on the survival time (in years) or on the censored time (whether the patient was still alive at the end of the study) was recorded. Other covariates such as patient's age at the time of diagnosis, the year of diagnosis, and the stage of the patient's cancer were also recorded. Stage of cancer is a factor that has four levels, ordered from least serious (I) to most serious (IV). Both the stage of the cancer and the age of the patient were a priori selected as important variables possibly influencing the survival function. We have therefore, $n = 90$, $p = 4$ (X_1: patient's age at diagnosis; X_2: 1 if stage II cancer, 0 otherwise; X_3: 1 if stage III cancer, 0 otherwise; X_4: 1 if stage IV cancer, 0 otherwise). The censoring percentage $P_{\%}$ is 44 and the largest observation is censored (i.e. $Y^+_{(n)} = 10.7+$). The dataset is also analysed using various approaches, including log-normal AFT modeling, in Klein & Moeschberger (1997).

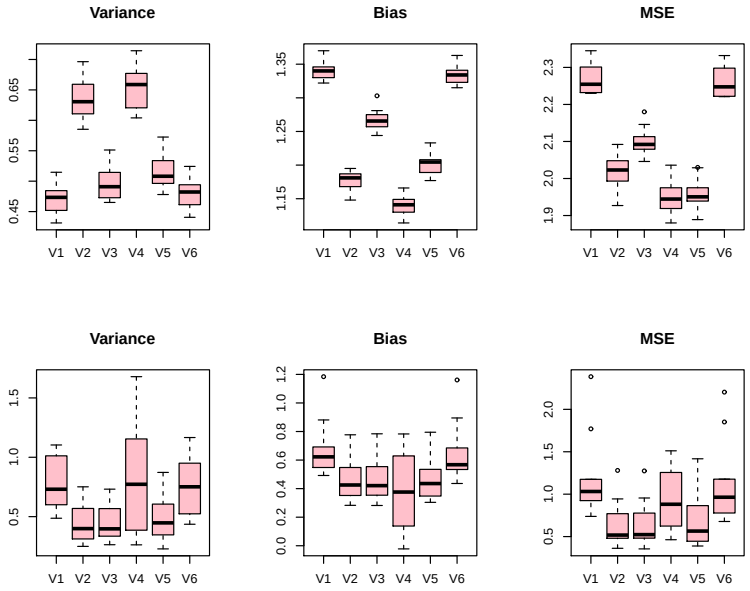

Fig. 5. Simulation study for second example under $P_{\%} = 70\%$. Box-plots for variance, bias, mean squared error of β. Here V1, V2, V3, V4, V5 and V6 represent the estimation approaches W_0: Efron's redistribution, W_{τ_m}: conditional mean, $W_{\tau_{md}}$: conditional median, $W_{\tau_m^*}$: resampling based conditional mean, $W_{\tau_{md}^*}$: resampling based conditional median, and W_ν: predicted difference quantity respectively. Graphs in first row show the results when r_{ij} is 0; graphs in second row show the results when r_{ij} is 0.5.

Table 4. *Parameter estimation under the approaches W_0: Efron's redistribution, W_{τ_m}: conditional mean, $W_{\tau_{md}}$: conditional median, $W_{\tau_m^*}$: resampling based conditional mean, $W_{\tau_{md}^*}$: resampling based conditional median, and W_ν: predicted difference quantity approach to the Laryngeal cancer data. Estimates under LN-AFT are based on log-normal AFT model solved by least squares method without tail correction (see Klein and Moeschberger, 1997).*

	Parameter Estimate						
Variable	W_0	W_{τ_m}	$W_{\tau_{md}}$	$W_{\tau_m^*}$	$W_{\tau_{md}^*}$	W_ν	LN-AFT
X_1 : Age	0.008 (0.020)	0.009 (0.022)	0.009 (0.022)	0.009 (0.024)	0.009 (0.021)	0.008 (0.019)	-0.018 (0.014)
X_2 : Stage II	-0.628 (0.420)	-0.846 (0.539)	-0.840 (0.514)	-1.052 (0.535)	-0.966 (0.500)	-0.649 (0.468)	-0.199 (0.442)
X_3 : Stage III	-0.945 (0.381)	-1.176 (0.443)	-1.169 (0.419)	-1.395* (0.451)	-1.304* (0.458)	-0.967 (0.390)	-0.900 (0.363)
X_4 : Stage IV	-1.627** (0.461)	-1.848** (0.444)	-1.841** (0.495)	-2.056** (0.506)	-1.969** (0.581)	-1.648** (0.478)	-1.857** (0.443)

$(\cdot)^{**}$ and $(\cdot)^{*}$ indicate significant at $p < 0.001$ and $p < 0.01$ respectively

We apply the proposed imputation approaches to the log-normal AFT model (26) using regularized WSL (7). We use two main effects, age and stage. The estimates of the parameters under different imputation techniques are reported in Table 4. These give broadly similar results, but differ from those found by Klein & Moeschberger (1997) where Efron's tail correction was not applied, shown in the last column of the table (LN-AFT). Klein & Moeschberger (1997) found Stage IV to be the only significant factor

influencing the survival times. All our imputation methods found Stage IV as highly significant factor. In addition, Stage III factor is found as significant at $p < 0.01$ by adding the resampling based mean and median methods.

6·2. *Channing House Data*

Channing House is a retirement centre in Palo Alto, California. The data were collected between the opening of the house in 1964 and July 1, 1975. In that time 97 men and 365 women passed through the centre. For each of these, their age on entry and also on leaving or death was recorded. A large number of the observations were censored mainly due to the residents being alive on July 1, 1975, the end of the study. The left truncation variable here is the entry age into Channing House. It is clear that only subjects with entry age smaller than or equal to age on leaving or death can become part of the sample. Over the time of the study 130 women and 46 men died at Channing House. Differences between the survival of the sexes was one of the primary concerns of that study.

Of the 97 male lifetimes, 51 observations were censored and the remaining 46 were observed exactly. Of the 51 censored lifetimes, there are 19 observations each of which has lifetime 137 months (which is the largest observed lifetime). Similarly, of 365 female lifetimes, 235 observations were censored and the remaining 130 were observed exactly. Of the 235 censored lifetimes, 106 take the maximum observed value of 137 months. Therefore, the imputation approaches impute the lifetime of 19 observations for the male dataset and 106 observations for the female dataset.

The K–M survival curve for male and female data, (Figure 6) shows that survival chances clearly differ between the sexes. We now investigate whether the imputed value

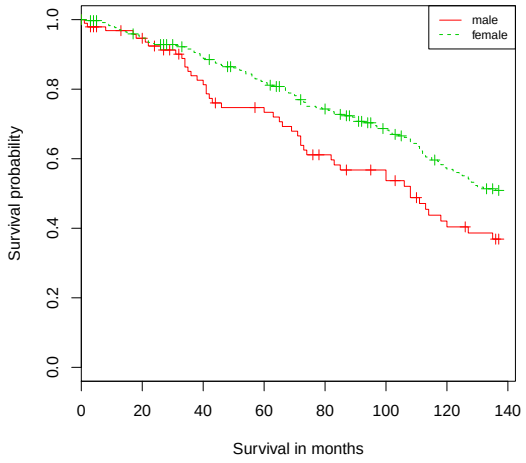

Fig. 6. K–M plots for Channing House data

and the estimate from the log-normal AFT model (26) of lifetimes on the calender ages (the only covariate) fitted by the WLS method (2) differ between male and female. Of interest we implement the imputing approaches except the resampling based conditional mean and conditional median approaches for male and female Channing House data separately. The two resampling based approaches can not be implemented for AFT models with one single covariate. They need at least two covariates to be executed. The results

are shown in Table 5. The results clearly depict that the estimates for age by the methods differ significantly between male and female. So does happen also for the imputed values obtained by the methods. For both datasets the conditional mean approach imputes with much higher value.

Table 5. *Parameter estimation for calender age from the log-normal AFT model and imputed value estimation for the largest observations by the approaches W_0: Efron's redistribution, W_{τ_m}: conditional mean, $W_{\tau_{md}}$: conditional median, and W_ν: predicted difference quantity approach for the Channing House data.*

	W_0	W_{τ_m}	$W_{\tau_{md}}$	W_ν
Age (male)	-0.153	-0.201	-0.154	-0.154
Age (female)	-0.180	-0.198	-0.182	-0.186
Imputed value (male)	137*	176.5	138.1	137.9
Imputed value (female)	137*	143.1	137.6	138.8

Note: The value with * is not imputed rather than obtained using Efron's redistribution algorithm.

Hence, in the present of heavy censoring the imputed methods are able to impute all the largest observations with only a single value. In this case one might be interested in imputing the observations with different lifetimes as if they would have been observed rather than impute them with a single value. In order to acknowledge this issue we propose two alternative approaches for imputing the heavy tailed censoring observations. The approaches do not require to take the underlying covariates into account. The practical implications of such imputation techniques might be found in many fields such as economics, industry, life sciences etc. One approach is based on the technique of the predicted difference quantity. The other approach is based on the trend of the survival probability in the K–M curve (i.e. Figure 6).

6·3. *Proposed Additional Approaches for Imputing Tail Ties*

Iterative Procedure:

Let us assume that there are m tied largest observations which are denoted, without loss of generality, by $Y_{\bar{u}(n_k)}$ for $k = 1, \cdots, m$. The first technique is an m iterative procedure where the k-th observation is imputed using the predicted difference method after assuming that the k-th observation is the unique largest censored observation in the dataset. The computational procedure is summarized briefly as below

1. Compute the modified failure time using Equation (22).
2. Set $\delta_{(n_k)} = 1$ for any $1 \leq k \leq m$.
3. Compute ν using Equation (25).
4. Add the quantity ν found in Step 3 to $Y_{\bar{u}(n_k)}$.
5. Repeat Step 2 to 4 for $m-1$ times for imputing $m-1$ observations. ν in Step 3 under each imputation is based on all modified failure times including the imputed values found in Step 4.

Extrapolation Procedure:

Under this approach we first follow the trend of the K–M survival probabilities $\hat{S}(t)$ versus the lifetimes for the subjects. If the trend for original K–M plot is not linear then we may first apply a transformation of the survival probability (e.g. $[\hat{S}(t)]^\psi$ for suitable

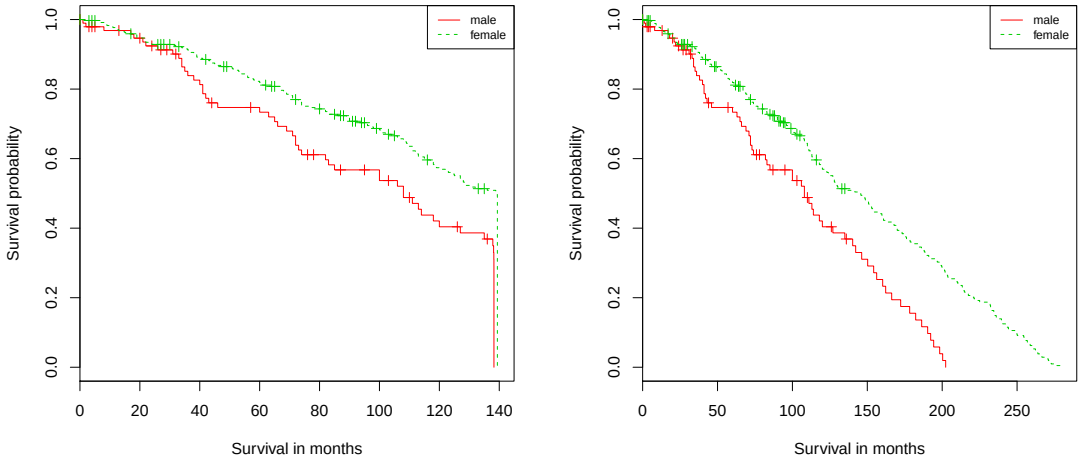

Fig. 7. Survival comparison between the male and female for Channing House data using the iterative (left panel) and extrapolation (lower panel) additional imputing methods to the tail tied observations

ψ). When linear trend is established we fit a linear regression of lifetimes on $[\hat{S}(t)]^{\psi}$. Now the lifetimes against the expected survival probabilities can easily be obtained using the fitted model.

Table 6. *Results by the additional imputation methods for the Channing House male data.*

Method	Imputed lifetimes for 19 tail tied observations
First method	137.85, 138.11, 138.19, 138.22, 138.22, 138.23, 138.23, 138.23, 138.23, 138.23, 138.23, 138.23, 138.23, 138.23, 138.23, 138.23, 138.23, 138.23, 138.23
Second method	134.23, 136.23, 150.25, 152.25, 156.26, 158.26, 160.26, 162.27, 166.27, 176.29, 178.29, 180.29, 184.30, 186.30, 188.30, 194.31, 196.32, 198.32, 200.32

Both techniques are implemented to male and female Channing House data separately. We report here results as given by the Table 6 only for male data. We follow the above computational procedure for applying the second method to the Channing House data, where we find a linear trend between the K–M survival probabilities $\hat{S}(t)$ and the lifetimes in first attempt. As part of the remaining procedure we first fit a linear regression of lifetimes on $\hat{S}(t)$ and then compute the predicted lifetime against each censored lifetime. For clarity the imputed values obtained using the second method are put in ascending order in Table 6. Results show that the second method tends to impute the largest observations with huge variations among the imputed values. The method also doesn't produce any tied imputed values. The method produces the imputed values in a way as if the largest censored (also tied) observations would have been observed. On the contrary, the first method produces imputed values with many ties and all imputed values close to the imputed value 137.9 obtained by the predicted difference approach.

The K–M plot with the 19 imputed lifetimes for male and 106 imputed lifetimes for female data under two imputing approaches is given by Figure 7. The K–M plots show that the second method outperforms the first by a huge margin. This also leads to major

changes in the coefficient value for the age covariate from fitting the AFT model. The estimated coefficient for male data using the two methods are -0.154 and -0.218 and those for female data are -0.187 and -0.385. Hence it might suggest that when there are many largest censored observations the second method prefers to the first for imputing them under the AFT models fitted by the WLS method. The first method might be useful when there are very few largest censored observations.

7. Discussion

We propose five imputation techniques for when the largest observation in a dataset is a censored observation. Each technique satisfies the basic right censoring assumption that the unobserved lifetime is greater than the observed censored time.

We examine the performance of these approaches taking into account different censoring levels and different correlation structures among the covariates in the dataset. The analysis from the simulation examples suggests all five imputation techniques except the predicted difference quantity approach can perform much better than Efron's redistribution technique for both type of datasets—correlated and uncorrelated. At higher censoring the predicted difference quantity approach outperforms the Efron's technique while at both lower and medium censoring they perform almost similarly of each other. For both type of datasets and for each censoring level, the conditional mean adding and the resampling based conditional mean adding provide the least bias and the least mean squared errors for the estimates. However, more investigation is required, particularly for different correlation structures. For implementing all proposed imputation approaches we have provided a publicly available package *imputeYn* (Khan & Shaw, 2012) implemented in the R programming system.

References

Buckley J, James I (1979) Linear regression with censored data. Biometrika 66:429–436

Candes E, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n. The Annals of Statistics 35(6):2313–2351

Datta S (2005) Estimating the mean life time using right censored data. Statistical Methodology 2:65–69

Datta S, Le-Rademacher J, Datta S (2007) Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. Biometrics 63:259–271

Efron B (1967) The two sample problem with censored data. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol 4, New York: Prentice Hall, pp 831–853

Engler D, Li Y (2009) Survival analysis with high-dimensional covariates: An application in microarray studies. Statistical Applications in Genetics and Molecular Biology 8(1):Article 14

Hu S, Rao JS (2010) Sparse penalization with censoring constraints for estimating high dimensional AFT models with applications to microarray data analysis. Technical Reports, University of Miami

Huang J, Ma S, Xie H (2006) Regularized estimation in the accelerated failure time model with high-dimensional covariates. Biometrics 62:813–820

Hyde J (1980) Testing Survival With Incomplete Observations. John Wiley, Biostatistics Casebook, New York

Jin Z, Lin DY, Ying Z (2006) On least-squares regression with censored data. Biometrika 93(1):147–161

Kalbfleisch J, Prentice RL (2002) The Statistical Analysis of Failure Time Data, 2nd edn. John Wiley and Sons, New Jersey

Kardaun O (1983) Statistical survival analysis of male larynx-cancer patients— a case study. Statistica Neerlandica 37(3):103–125

Khan MHR, Shaw JEH (2012a) imputeYn: Imputing the last largest censored observation/observations under weighted least squares. R package version 1.1

Khan MHR, Shaw JEH (2012b) Variable selection using the elastic net type regularized technique for high–dimensional survival data. Preprint

Robins JM, Finkelstein DM (2000) Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics 56:779–788

Satten GA, Datta S (2001) The Kaplan-Meier estimator as an Inverse-Probability-of-Censoring weighted average. The American Statistician 55(3):207–210

Stute W (1993) Consistent estimation under random censorship when covariables are available. Journal of Multivariate Analysis 45:89–103

Stute W (1996a) Distributional convergence under random censorship when covariables are present. Scandinavian Journal of Statistics 23:461–471

Stute W (1996b) The jackknife estimate of variance of a Kaplan-Meier integral. The Annals of Statistics 24(6):2679–2704

Stute W, Wang J (1994) The jackknife estimate of a Kaplan-Meier integral. Biometrika 81(3):602–606