STATISTICA, anno LXXVI, n. 1, 2016

A SEMI-PARAMETRIC REGRESSION MODEL FOR ANALYSIS OF MIDDLE CENSORED LIFETIME DATA

S. Rao Jammalamadaka
Department of Statistics and Applied Probability, University of California, Santa Barbara, USA.
S. Prasad
Department of Statistics, Cochin University of Science and Technology, Kerala, India.
P. G. Sankaran [1]
Department of Statistics, Cochin University of Science and Technology, Kerala, India

1. Introduction

Middle censoring introduced by Jammalamadaka and Mangalam (2003) occurs in situations where a data point becomes unobservable if it falls inside a random censoring interval. In such situations, the exact values are available for some observations and for some others, random censoring intervals are observed. We may find several such situations in survival studies and reliability applications. For example in biomedical studies, the patients under observation may be withdrawn from the study for a short period of time and the exact lifetimes of those patients may not be available if the event happens during this period. In reliability applications, the failure of equipment could occur during a period of time, which is not possible to observe. In such contexts we only observe a censorship indicator and the interval of censorship.

As was pointed out by Jammalamadaka and Mangalam (2003), left censored data and right censored data can be considered as special cases of this more general middle censoring, by suitable choices of the interval. Also such a censoring scheme is not complementary to the usual double censoring discussed in Klein and Moeschberger (2005) and Sun (2006). Jammalamadaka and Mangalam (2003) pointed out various applications of middle censoring and developed a nonparametric maximum likelihood estimator(NPMLE), which is the maximum likelihood estimator of the distribution function, where no specific parametric assumptions are made on the parent population. They have proved that an NPMLE is always a Self Consistent Estimator (SCE) (see Tarpey and Flury (1996)). Jammalamadaka and Iyer (2004) proposed a variant of this self consistent estimator for which the weak convergence was established. In the parametric context, Iyer *et al.* (2008)

[1] Corresponding Author e-mail: sankaran.p.g@gmail.com

studied middle censoring for the exponential distribution. Mangalam *et al.* (2008) developed a necessary and sufficient condition for the equivalence of self-consistent estimators and NPMLEs. Jammalamadaka and Mangalam (2009) discussed it for the von Mises model in the context of directional data. Shen (2010) proposed an inverse-probability-weighted estimator for the distribution function for data arising from such a censoring scheme, while Davarzani and Parsian (2011) discussed it in the discrete case for the geometric distribution. Shen (2011) showed that the nonparametric maximum likelihood estimator (NPMLE) of distribution function can be obtained by using Turnbull's EM algorithm (Turnbull, 1976) or self-consistent estimating equation (Jammalamadaka and Mangalam, 2003) with an initial estimator which puts mass only on the innermost intervals. Sankaran and Prasad (2014) disussed a Weibull regression model for a middle censored lifetime data. Jammalamadaka and Leong (2015) analysed discrete middle censored data in the presence of covariates while Abuzaid *et al.* (2015) discussed robustness middle censoring scheme in parametric survival models.

In survival studies, covariates or explanatory variable are usually used to represent heterogeneity in a population. The main objective in such situations is to understand and exploit the relationship between the lifetime and covariates. Regression models are commonly employed to study such relationship. The most widely used semi-parametric regression model is the well known proportional hazards model by David (1972). For a comprehensive review on properties and inference procedures of the proportional hazards model, one may refer to Kalbfleisch and Prentice (2011) and Lawless (2011). The analysis of middle censored data in the presence of covariates has not yet been developed, which is the goal of the present work. Accordingly we study the regression problem for middle censored lifetime data in which the hazard rate function may depend on some covariates.

In Section 2 we present the model and state the inference procedure for the problem. Section 3 provides a simulation study to assess the finite sample properties of the estimators, while Section 4 describes an application of the proposed model to a real life problem. Section 5 concludes the paper with a summary.

2. Model and Inference Procedure

Let T be a non-negative random variable representing lifetime of a study subject with an unknown cumulative distribution function $F_0(\cdot)$. Let (U, V) be a random vector which represents the censoring interval having bivariate cumulative distribution function $G(\cdot, \cdot)$. Assume that (U, V) is independent of T, with $P(U < V) = 1$. Let Z be a $p \times 1$ vector of covariates. The covariates may be continuous or they may be indicator variables. Assume that lifetime T is middle censored by the random interval (U, V). Thus one can observe the vector (X, δ, Z), where

$$X = \begin{cases} T & \text{if } \delta = I(X = T) = 1 \text{ (uncensored case)}, \\ (U, V) & \text{if } \delta = I(X = T) = 0 \text{ (censored case)}. \end{cases}$$

Let us assume that there are n individuals under study and for the i'th individual we observe $\big(X_i, \delta_i, Z_i\big)$, for $i = 1, 2, ..., n$.

When we have incomplete data due to censoring, the idea of self-consistency

plays a pivotal role in the estimation of the unknown population distribution function F_0. If we estimate F_0 via the Expectation-Maximisation algorithm (Dempster *et al.*, 1977), as described by Tsai and Crowley (1985), the resulting estimating equation takes the form $\hat{F}_0(t) = E_{\hat{F}_0}[E_n(t)|\text{observed data}]$, where $\hat{F}(\cdot)$ is the required estimate and $E_n(\cdot)$ is the empirical distribution function. This equation is known as Self-Consistency Equation (SCE) and was first introduced by Efron (1967) where he used this to describe a class of estimators of F_0 for the case of right censored data. Note that this equation is an implicit equation and hence the unknown quantity to be estimated appears on both sides of the estimating equation. Jammalamadaka and Mangalam (2003) have shown that the NPMLE of F_0 is always a Self Consistent Estimator(SCE) which takes the form:

$$\hat{F}_0(t) = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{\delta_i I(X_i \le t) + (1-\delta_i)I(V_i \le t) + (1-\delta_i)I(t \in (U_i, V_i))\frac{\hat{F}_0(t) - \hat{F}_0(U_i)}{\hat{F}_0(V_i) - \hat{F}_0(U_i)}\Bigg\}. \tag{1}$$

With Cox proportional hazards assumption, the survival function of T at t conditional on $Z = z$ is given by

$$S(t|z) = (S_0(t))^{\exp(\theta' z)}, \tag{2}$$

where $S_0(t)$ is baseline survival function and $\theta = (\theta_1, \theta_2, ..., \theta_p)'$ is a $p \times 1$ vector of regression coefficients. Differentiating (2) with respect to t we get the density function of T given $Z = z$ as

$$f(t|z) = f_0(t)\exp(\theta' z)\,(S_0(t))^{\exp(\theta' z)-1},$$

where $f_0(t)$ is the baseline density of T. Our objective is to estimate θ and $S_0(t)$ under middle censored observation scheme.
The likelihood corresponding to the observed data is given by

$$L(\theta) \propto \prod_{i=1}^{n} f(t_i|z_i)^{\delta_i}\left[(S_0(u_i))^{\exp(\theta' z_i)} - (S_0(v_i))^{\exp(\theta' z_i)}\right]^{1-\delta_i}.$$

Without loss of generality, assume that the first n_1 observations are exact lifetimes, and the remaining n_2 are censored intervals, with $n_1 + n_2 = n$.
Now the likelihood, excluding the normalizing constant is:

$$L(\theta) = \prod_{i=1}^{n_1} f(t_i|z_i) \cdot \prod_{i=n_1+1}^{n_1+n_2}\left((S_0(u_i))^{\exp(\theta' z_i)} - (S_0(v_i))^{\exp(\theta' z_i)}\right), \tag{3}$$

and the log-likelihood is given by

$$l(\theta) = \sum_{i=1}^{n_1}[\log f_0(t_i) + \theta' z_i + \exp(\theta' z_i)\log S_0(t_i)] + \sum_{i=n_1+1}^{n_1+n_2} \log\left((S_0(u_i))^{\exp(\theta' z_i)} - (S_0(v_i))^{\exp(\theta' z_i)}\right). \quad (4)$$

The first order partial derivative with respect to θ_r, for $r = 1, 2, .., p$, is given by

$$\frac{\partial}{\partial\theta_r} l(\theta) = \sum_{i=1}^{n_1}(z_{ir}(1 + \exp(\theta' z_i)\log S_0(t_i))) + \sum_{i=n_1+1}^{n_1+n_2}\Big\{z_{ir}\exp(\theta' z_i)\left((S_0(u_i))^{\exp(\theta' z_i)} - (S_0(v_i))^{\exp(\theta' z_i)}\right)^{-1} \left((S_0(u_i))^{\exp(\theta' z_i)}\log S_0(u_i) - (S_0(v_i))^{\exp(\theta' z_i)}\log S_0(v_i)\right)\Big\}, \quad (5)$$

where z_{ir} is the r'th component in the covariate vector corresponding to i'th individual. We observe that (5) does not involve the baseline density $f_0(t)$. We now give an algorithm for estimating the parameters θ and $S_0(t)$):

Step 1. Set the vector $\theta = 0$.

Step 2. At the first iteration, find the SCE $S_0^{(1)}(t)$ of $S_0(t)$ using (1) and substitute this in (5) and solve $\partial l(\theta)/\partial\theta_r = 0, r = 1, 2, ..., p$ to get the estimator $\theta^{(1)}$ of θ.

Step 3. Find $\tilde{t_i}^{(1)} = {S_0^{(1)}}^{-1}\left[S_0^{(1)}(t_i)^{\exp({\theta^{(1)}}' z_i)}\right]$ and similarly find $\tilde{u}_i^{(1)}$ and $\tilde{v}_i^{(1)}$ as our updated observations at first iteration.

Step 4. At the j'th iteration ($j > 1$), use $\tilde{t_i}^{(j-1)}, i = 1, 2, ..., n_1$ and $(\tilde{u}_i^{(j-1)}, \tilde{v}_i^{(j-1)})$, $i = n_1 + 1, ..., n$ as our data points in (1) and obtain $S_0^{(j)}(t)$. Substitute $S_0^{(j)}(t)$ in (5) and solve $\partial l(\theta)/\partial\theta_r = 0, r = 1, 2, ..., p$ to obatain the j'th iterated update $\theta^{(j)}$ of θ.

Step 5. Repeat Step 4 until convergence is met, say when $\|\theta^{(k)} - \theta^{(k+1)}\| < 0.0001$ and $\sup_t\left\{|S_0^{(k)}(t) - S_0^{(k+1)}(t)|\right\} < 0.001$, for some k.

Note that Step 3. in the algorithm is justified because if $a_i = (S_0^{(1)}(t_i))^{\exp({\theta^{(1)}}' z_i)}$, then the a_i 's have a uniform distribution over $[0, 1]$. Therefore to scale these back to baseline distribution we need to find $\tilde{t_i} = \inf\{t : S_0^{(1)}(t) \le a_i\}$. Thus the correct choice is $\tilde{t_i} = {S_0^{(1)}}^{-1}(a_i) = {S_0^{(1)}}^{-1}\left(S_0^{(1)}(t_i)^{exp({\theta^{(1)}}' z_i)}\right)$.

Now consider a situation where the support of U and V is contained in the support of T. Then clearly $S_0(\cdot)$ will stay away from 1 and 0 on the support of U and V. But if the least observation happens to be an observation on U say u_k, to maintain the monotonicity property, $\hat{S}_{0n}(u_k) = 1$. This leads to the inconsistency of the estimator at that end point. The same is true with the other end also. This motivates an additional restriction resulting in a bounded MLE (BMLE) of the parameter θ and $S_0(t)$:

A1: There exist $0 < \alpha_0 < \alpha_1 < \infty$ and $0 < m_0 < M_0 < 1$ such that $P(\alpha_0 \le U <$

$V \leq \alpha_1) = 1$ and $m_0 < S_0(\alpha_1) < S_0(\alpha_0) < M_0$.
We will incorporate this restriction with our estimator and define our parameter space be (Θ, Φ), where $\Theta \subseteq \mathbb{R}_p$ be a parameter space of θ and Φ be defined by

$$\Phi = \{S_0 : [\alpha_0, \alpha_1] \to [m_0, M_0] \text{ and } S_0 \text{ is decreasing } \}.$$

Let us name the estimator thus obtained for θ as $\hat{\theta}_n$ and that for $S_0(t)$ as $\hat{S}_{0n}(t)$. The following conditions are necessary to establish the consistency property.
A2: Conditional on Z, T is independent of (U, V).
A3: The joint distribution of (U, V, Z) does not depend on the true parameter $(\theta, S_0(t))$.
A4: Z is bounded. That is there exist some finite $M > 0$ such that $P\{\|Z\| \leq M\} = 1$, where $\|\cdot\|$ is the usual metric on $\mathbb{R}_p$.
A5: Distribution of Z is not concentrated on any proper affine subspace of $\mathbb{R}_p$.

THEOREM 1. *Suppose that $\Theta \in \mathbb{R}_p$ is bounded and assumptions (A1) to (A5) hold. Then estimator $(\hat{\theta}_n, \hat{S}_{0n})$ is consistent for the true parameter (θ_0, S_{00}) in the sense that if we define a metric $d : \Theta \times \Phi \to \mathbb{R}$, by*

$$d\left((\theta_1, S_{01}), (\theta_2, S_{02})\right) = \|\theta_1 - \theta_2\| + \int |S_{01}(t) - S_{02}(t)| dF_0(t) + \left[\int \left((S_{01}(u) - S_{02}(u))^2 + (S_{01}(v) - S_{02}(v))^2\right) dG(u, v)\right]^{\frac{1}{2}} \tag{6}$$

then $d\left((\hat{\theta}_n, \hat{S}_{0n}), (\theta_0, S_{00})\right) \to 0$ almost surely (a.s.).

PROOF. In the following discussion we denote $Y_i = (X_i, \delta_i)$. Let the probability function of $Y = (X, \delta)$ be given by

$$p(y; \theta, S_0) = \prod_{i=1}^{n} f(t_i|z_i)^{\delta_i} [(S_0(u_i))^{\exp(\theta' z_i)} - (S_0(v_i))^{\exp(\theta' z_i)}]^{1-\delta_i} g(u_i, v_i|z_i) q(z_i), \tag{7}$$

where g is the joint density of (U, V), conditional on Z and q is the density of Z. Using $(A2)$ and $(A3)$, the log-likelihood function scaled by $1/n$ for the sample $(y_i, z_i), i = 1, 2, ..., n$ up to terms not depending on (θ_0, S_{00}) is

$$l(\theta, S_0) = \frac{1}{n} \sum_{i=1}^{n} \left\{\delta_i \log f_0(t_i|z_i) + (1 - \delta_i) \log \left[(S_0(u_i))^{\exp(\theta' z_i)} - (S_0(v_i))^{\exp(\theta' z_i)}\right]\right\}. \tag{8}$$

We write $p_n(y) = p(y; \hat{\theta}_n, \hat{S}_{0n})$ and $p_0(y) = p(y; \theta_0, S_{00})$ where $(\hat{\theta}_n, \hat{S}_{0n})$ is the MLE that maximizes the likelihood function over $\Theta \times \Phi$ and $(\theta_0, S_{00}) \in \Theta \times \Phi$. Therefore

$$\sum_{i=1}^{n} \log p_n(Y_i) \geq \sum_{i=1}^{n} \log p_0(Y_i)$$

and hence

$$\sum_{i=1}^{n} \log \frac{p_n(Y_i)}{p_0(Y_i)} \geq 0.$$

By the concavity of the function $x \mapsto \log x$, for any $0 < \alpha < 1$,

$$\frac{1}{n} \sum_{i=1}^{n} \log \left((1-\alpha) + \alpha \frac{p_n(Y_i)}{p_0(Y_i)} \right) \geq 0. \tag{9}$$

The left hand side can be written as

$$\int \log \left((1-\alpha) + \alpha \frac{p_n(Y_i)}{p_0(Y_i)} \right) \mathrm{d}(\mathbb{P}_n - \mathbb{P})(Y) + \int \log \left((1-\alpha) + \alpha \frac{p_n(Y_i)}{p_0(Y_i)} \right) \mathrm{d}\mathbb{P}(Y), \tag{10}$$

where $\mathbb{P}_n$ is the empirical measure of Y and $\mathbb{P}$ is the joint probability measure of Y.

Let us assume that the sample space Ω consists of all infinite sequences $Y_1, Y_2, ...$, along with the usual sigma field generated by the product topology on $\prod_1^\infty (\mathbb{R}^3 \times \{0,1\})$ and the product measure $\mathbf{P}$. For p defined in (7) let us define a class of functions $\mathcal{P} = \Big\{ p(y, \theta, S_0), (\theta, S_0) \in (\Theta \times \Psi) \Big\}$ and a class of functions $\mathcal{H} = \Big\{ \log(1 - \alpha + \alpha p/p_0) : p \in \mathcal{P} \Big\}$, where $p_0 = p(y, \theta_0, S_{00})$. Then it follows from Huang and Wellner (1995) that $\mathcal{H}$ is a Donsker class. With this and Glivenko-Cantelli theorem there exists a set $\Omega_0 \in \Omega$ with $\mathbf{P}(\Omega_0) = 1$ such that for every $\omega \in \Omega_0$, the first term of (10) converges to zero. Now fix a point $\omega \in \Omega_0$ and write $\hat{\theta}_n = \hat{\theta}_n(\omega)$ and $\hat{S}_{0n}(\cdot) = \hat{S}_{0n}(\cdot, \omega)$. By our assumption Θ is bounded, and hence for any subsequence of $\hat{\theta}_n$, we can find a subsequence converging to $\theta_* \in \Theta^{'}$, the closure of Θ. Also by Helly's selection theorem, for any subsequence of $\hat{S}_{0n}$, we can find a further subsequence converging to some decreasing function S_{0*}. Choose the convergent subsequence of $\hat{\theta}_n$ and the convergent subsequence of $\hat{S}_{0n}$ so that they have the same indices, and without loss of generality, assume that $\hat{\theta}_n$ converges to θ_* and that $\hat{S}_{0n}$ converges to $S_{0*}(\cdot)$.

Let $p_*(y) = p(y, \theta_*, S_{0*})$. By the bounded convergence theorem, the second term of (10) converges to

$$\int \log \left((1-\alpha) + \alpha \frac{p_*(y)}{p_0(y)} \right) d\mathbb{P}(y)$$

and by (9) this is nonnegative. But by Jensen's inequality, it must be non-positive. Therefore it must be zero and it follows that

$$p_*(y) = p_0(y) \quad \mathbb{P} - \text{almost surely.}$$

This implies

$$S_{0*}(t) = S_{00}(t) \quad F_0 - \text{almost surely.}$$

Therefore by bounded convergence theorem,

$$\int |\hat{S}_{0n}(t) - S_{00}(t)| dF_0(t) \to 0 \tag{11}$$

and also

$$(S_{0*}(u))^{\exp(\theta_*' z)} = (S_{00}(u))^{\exp(\theta_0' z)} \text{and} (S_{0*}(v))^{\exp(\theta_*' z)} = (S_{00}(v))^{\exp(\theta_0' z)} \quad \mathbb{P}-\text{almost surely.}$$

This together with (A5) imply that there exist $z_1 \neq z_2$ such that for some $c \in [\alpha_0, \alpha_1]$,

$$(S_{0*}(c))^{\exp(\theta_*' z_1)} = (S_{00}(c))^{\exp(\theta_0' z_1)} \text{and} (S_{0*}(c))^{\exp(\theta_*' z_2)} = (S_{00}(c))^{\exp(\theta_0' z_2)}.$$

Since $S_{0*}(c) \geq m_0 > 0$ and $S_{00}(c) \geq m_0 > 0$, this implies

$$(\theta_* - \theta_0)'(z_1 - z_2) = 0.$$

Again by (A5), the collection of such z_1 and z_2 has positive probability and there exist at least p such pairs that constitute a full rank $p \times p$ matrix, it follows that $\theta_* = \theta_0$, This in turn implies that

$$S_{0*}(u) = S_{00}(u) \quad \text{and} \quad S_{0*}(v) = S_{00}(v) \quad G - \text{almost surely.}$$

Therefore by bounded convergence theorem,

$$\int \Big((\hat{S}_{0n}(u) - S_{00}(u))^2 + (\hat{S}_{0n}(v) - S_{00}(v))^2 \Big) \, dG(u, v) \to 0. \tag{12}$$

Equations (11) and (12) together with $\theta_* = \theta_0$ hold for all $\omega \in \Omega_0$ with $\mathbf{P}(\Omega_0) = 1$. This completes the proof. □

Remark 2. *A likelihood ratio test can be carried out to test the significance of regression coefficients. The null hypothesis $H_0 : \theta = 0$ can be tested against $H_1 : \theta \neq 0$, where 0 is the null vector of same order, with the test statistic $-2\, log\, \frac{L(0)}{L(\hat{\theta})}$ which is a $\chi^2_{(p-1)}$ variate. The test results in rejecting the null hypothesis for small P-values.*

3. Simulation Studies

A simulation study is carried out to assess the finite sample properties of the estimators. We consider the exponential distribution with mean λ^{-1} as the distribution of lifetime variable T. Also we choose independent exponential distributions with fixed means λ_1^{-1} and λ_2^{-1}, which themselves are independent of T, as the distributions for the censoring random variates U and $V - U$ respectively, so that the distribution of V is gamma. We only consider a single covariate z in the present study which is generated from uniform distribution over $[0, 10]$ and let θ be the corresponding regression coefficient. Under the proportional hazards assumption, the survival function of T given z is given by

$$S(t|z) = \exp\big(- \lambda \exp(\theta z) t \big). \tag{13}$$

A large number of observations are generated on T for fixed values of λ and θ. Now corresponding to each and every observation on T, a random censoring interval is

TABLE 1
Empirical bias and MSE of the estimator of θ.

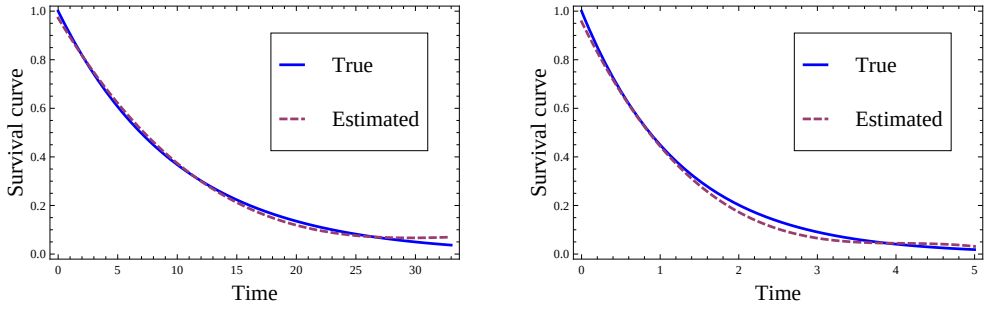
		$n = 500$			$n = 750$		
λ	θ	Bias	MSE	ECP	Bias	MSE	ECP
0.1	0.25	1.73e-2	9.62e-4	0.961	3.88e-3	7.52e-4	0.969
0.8	0.01	1.60e-3	7.16e-3	0.948	7.80e-4	1.09e-3	0.964
1.0	0.50	-1.26e-2	9.12e-4	0.935	-2.00e-4	1.01e-4	0.951
1.0	-0.50	3.56e-2	1.89e-3	0.969	2.73e-2	8.14e-4	0.977
1.5	0.80	-8.52e-3	4.11e-5	0.970	-6.94e-3	1.28e-5	0.982
3.0	-0.90	1.81e-2	5.19e-4	0.968	-3.06e-3	4.36e-5	0.979
5.0	-0.01	6.98e-3	9.71e-5	0.944	4.17e-3	6.38e-5	0.962

generated with (U, V) and if we find $T \notin (U, V)$ then T is selected in the sample, otherwise we choose the interval as the observation. As we generate large number of observations we can now choose a sample of required size n such that it contains about 25% censoring intervals. This process, now, can be repeated with various choices of λ and θ. The estimation procedure given in Section 2 can be employed to obtain the estimates of $S_0(t)$ and θ and using 1000 iterations. The empirical bias and mean squared error(MSE) are computed and given in Table 1. The Wald 95% confidence intervals of regression parameter are computed by using the empirical percentiles of the estimated regression coefficients. The proportion of times the true parameter value lies in such intervals is called empirical coverage probabilities (ECP) which is found out and is reported in Table 1.

Clearly both bias and MSE are small and as the sample size n increases they decrease, as one would expect and the empirical coverage probabilities are found fairly large, closer to one. Also for each set of parameter values and with sample size 750 we shall find out a cubic polynomial estimate of the form $S_0(t) = c_0 + c_1 t + c_2 t^2 + c_3 t^3$ with each of its coefficients being the average of corresponding coefficients obtained for all the iterations, for the baseline survival function and are compared in Figure 1, where continuous curve represents the true baseline survival function and dotted curve represents corresponding estimate. We see that both the estimated curve and actual curve are very close to each case.

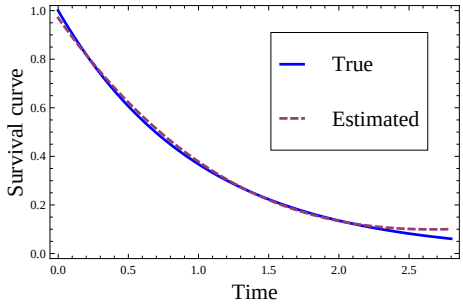
4. Data Analysis

In this section we apply our model to a real life data set. We consider the data on survival times (in years) for 149 diabetic patients followed for 17 years studied by Lee *et al.* (1988) and the data is given in Lee and Wang (2003), pages 58-63. The original data consists of 8 potential prognostic variables, but for illustrative purpose we take only two among them namely age denoted by z_1 and coronary heart disease(CHD) denoted by z_2 as the covariates with respective regression coefficients θ_1 and θ_2. Censoring is made by the method follows. A random censoring interval (U, V), where U and $V - U$ are independent exponential variates with means $\lambda_1^{-1} = 20$ and $\lambda_2^{-1} = 12.5$ is generated first. Then an individual from among 149 patients is selected at random and if lifetime of the patient happens to fall in the generated censoring interval, that lifetime is assumed to have middle censored and that interval is considered as the corresponding observation. Otherwise the lifetime is maintained. This process is repeated until around 25% of the observations are censored. The data resulted consists of 36 censored observations. We apply the model given in Section 2 and find that the estimate of the baseline survival function is $\hat{S}_0(t) = 7.886 \times 10^{-7} t^3 - 0.002272\, t^2 - 0.22181\, t + 0.880574$ and the estimates of the regression coefficients are obtained to be $\hat{\theta}_1 = 0.004418$ and $\hat{\theta}_2 = 0.1096$. *i.*e., the covariates have a positive effect on the lifetime. To test the significance of the covariate effect we consider the null hypothesis $H_0 : \theta = 0$, where $\theta = (\theta_1, \theta_2)$ and 0 is null vector of same order, and we use the likelihood ratio test described in Remark 2. The P-value of 0.0382 indicates that the covariate effects are significant. To check the validity of proportional hazards assumption, we shall plot $-\log(-\log(S(t|z)))$ against $\log t$ for the two covariates separately and the plots are given in Figure 2. We find that the two step functions are parallel and thus the proportionality assumption is justified.

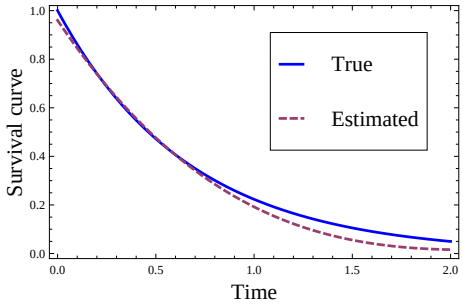


(a) $\lambda = 0.1$

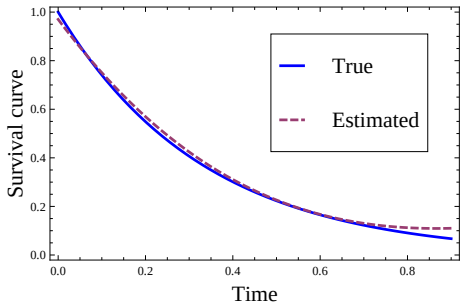
(b) $\lambda = 0.8$



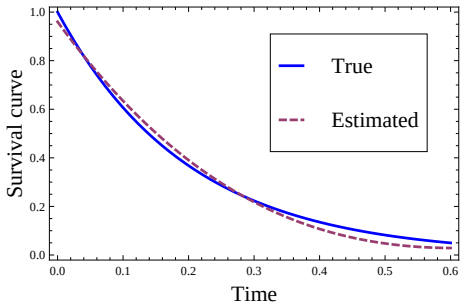
(c) $\lambda = 1$



(d) $\lambda = 1.5$



(e) $\lambda = 3$



(f) $\lambda = 5$

Figure 1 – Plots of baseline survival curve and its estimate.

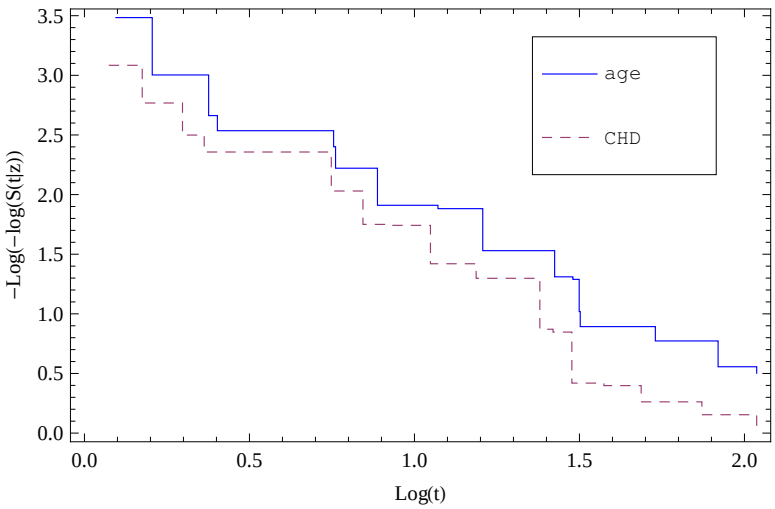


Figure 2 – Plots of $-\log(-\log(S(t|z)))$ against $\log t$

5. Conclusion

The present study discussed the semi-parametric regression problem for the analysis of middle censored data. A maximization procedure for finding the NPMLE is developed and its consistency established. The model is applied to a real data set. Simulation studies show that the inference procedure is efficient. Although we have considered a cubic curve for approximating $S_0(t)$ in our example, the degree and nature of the curve, of course, depends on the data set and the censoring distribution. More studies in this direction will be carried out in a separate work. Asymptotic normality of $\hat{\theta}$ and weak convergence of $\hat{S}_0(t)$ do not appear to be easy to establish, although one can perhaps extend the ideas used in Huang and Wellner (1995).

Acknowledgements

We thank the editor and referee for their constructive comments on the manuscript.

References

A. Abuzaid, M. Abu El-Qumsan, A. El-Habil (2015). *On the robustness of right and middle censoring schemes in parametric survival models.* Communications in Statistics-Simulation and Computation, , no. To appear.

N. Davarzani, A. Parsian (2011). *Statistical inference for discrete middle-censored data.* Journal of Statistical Planning and Inference, 141, no. 4, pp. 1455–1462.

C. R. David (1972). *Regression models and life tables (with discussion).* Journal of the Royal Statistical Society, 34, pp. 187–220.

A. P. Dempster, N. M. Laird, D. B. Rubin (1977). *Maximum likelihood from incomplete data via the em algorithm.* Journal of the royal statistical society. Series B (methodological), pp. 1–38.

B. Efron (1967). *The two sample problem with censored data.* In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.* vol. 4, pp. 831–853.

J. Huang, J. A. Wellner (1995). *Efficient estimation for the proportional hazards model with" case 2" interval censoring.* Technical Report No. 290, Department of Statistics, University of Washington, Seattle, USA.

S. K. Iyer, S. R. Jammalamadaka, D. Kundu (2008). *Analysis of middle-censored data with exponential lifetime distributions.* Journal of Statistical Planning and Inference, 138, no. 11, pp. 3550–3560.

S. R. Jammalamadaka, S. K. Iyer (2004). *Approximate self consistency for middle-censored data.* Journal of statistical planning and inference, 124, no. 1, pp. 75–86.

S. R. Jammalamadaka, E. Leong (2015). *Analysis of discrete lifetime data under middle-censoring and in the presence of covariates.* Journal of Applied Statistics, 42, no. 4, pp. 905–913.

S. R. Jammalamadaka, V. Mangalam (2003). *Nonparametric estimation for middle-censored data.* Journal of nonparametric statistics, 15, no. 2, pp. 253–265.

S. R. Jammalamadaka, V. Mangalam (2009). *A general censoring scheme for circular data.* Statistical Methodology, 6, no. 3, pp. 280–289.

J. D. Kalbfleisch, R. L. Prentice (2011). *The statistical analysis of failure time data*, vol. 360. John Wiley & Sons.

J. P. Klein, M. L. Moeschberger (2005). *Survival analysis: techniques for censored and truncated data.* Springer Science & Business Media.

J. F. Lawless (2011). *Statistical models and methods for lifetime data*, vol. 362. John Wiley & Sons.

E. T. Lee, J. Wang (2003). *Statistical methods for survival data analysis*, vol. 476. John Wiley & Sons.

E. T. Lee, M. A. Yeh, J. L. Cleves, D. Shafer (1988). *Vascular complications in noninsulin dependent diabetic oklahoma indians.* Diabetes, 37,(Suppl. 1).

V. Mangalam, G. M. Nair, Y. Zhao (2008). *On computation of npmle for middle-censored data.* Statistics & Probability Letters, 78, no. 12, pp. 1452–1458.

P. G. Sankaran, S. Prasad (2014). *Weibull regression model for analysis of middle-censored lifetime data.* Journal of Statistics and Management Systems, 17, no. 5-6, pp. 433–443.

P.-s. Shen (2010). *An inverse-probability-weighted approach to the estimation of distribution function with middle-censored data.* Journal of Statistical Planning and Inference, 140, no. 7, pp. 1844–1851.

P.-s. Shen (2011). *The nonparametric maximum likelihood estimator for middle-censored data.* Journal of Statistical Planning and Inference, 141, no. 7, pp. 2494–2499.

J. Sun (2006). *The statistical analysis of interval-censored failure time data.* Springer Science & Business Media.

T. Tarpey, B. Flury (1996). *Self-consistency: a fundamental concept in statistics.* Statistical Science, pp. 229–243.

W.-Y. Tsai, J. Crowley (1985). *A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency.* The Annals of Statistics, pp. 1317–1334.

B. W. Turnbull (1976). *The empirical distribution function with arbitrarily grouped, censored and truncated data.* Journal of the Royal Statistical Society. Series B (Methodological), pp. 290–295.

Summary

Middle censoring introduced by Jammalamadaka and Mangalam (2003), refers to data arising in situations where the exact lifetime becomes unobservable if it falls within a random censoring interval, otherwise it is observable. In the present paper we propose a semi-parametric regression model for such lifetime data, arising from an unknown population and subject to middle censoring. We provide an algorithm to find the non-parametric maximum likelihood estimator (NPMLE) for regression parameters and the survival function. The consistency of the estimators are established. We report simulation studies to assess the finite sample properties of the estimators. We then analyze a real life data on survival times for diabetic patients studied by Lee *et al.* (1988).

Keywords: Middle censoring; Proportional Hazards model; Self consistent estimator.