Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/100101

A Semiparametric Mixture Cure Model for Partly Interval Censored Failure Time Data

Yeqian Liu[1,*] *and Shuwei Li*[2]

[1]Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, USA
[2]School of Economics and Statistics, Guangzhou University, Guangzhou, China

Received: 23 Mar. 2020, Revised: 20 Apr. 2020, Accepted: 22 May 2020
Published online: 1 Mar. 2021

Abstract: This paper addresses regression analysis of partly interval censored data. Partly interval censored failure time data consist of both exact observed and interval censored observations on the survival time of interest. Furthermore, there may exist a cured subgroup, indicating that a proportion of study subjects are not susceptible to the failure event of interest. For the problem, we assume a logistic model for the cure probability and that the failure times of the uncured group come from a wide class of transformation models, which includes proportional hazards and proportional odds models as special cases. For the determination of the proposed estimators, an EM algorithm based on some subject-specific independent Poisson variables is developed to calculate the maximum likelihood estimators. Extensive simulation studies are conducted and indicate that the proposed method works well for practical situations. A motivating application from NASA's Hypobaric decompression sickness experiment is also provided.

Keywords: EM algorithm, Maximum likelihood estimation, Mixture cure models, Partly interval censored data, Semiparametric efficiency

1 Introduction

Partly interval censored failure time data arise when the failure times can be observed exactly for some subjects and for the remaining subjects. The event time of interest is only known to belong to an finite or half-open time interval. Such data include interval censored failure time data as special case and often occur in many fields especially in medical follow-up studies (Odell et al. [1]). Many authors discussed the analysis of such data under various situations including estimating the survival function, regression analysis and log-rank test (Turnbull [2], Kim [3] and Zhao et al. [3]).

A typical underlying assumption in failure time data analysis is that all subjects can eventually experience the event of interest when the follow-up time is long enough and it is well-known that sometimes this assumption may be untrue since some subjects may be cured or immune to the event. For the latter situation, we mean that there exists a cured subgroup in the whole population and mixture cure model proposed by Farewell [4], which treats the whole population as a mixture of cured subgroup and non-cured subgroup and assumes a cure rate model for cure probability. A survival model of the non-cure subjects is usually used to take this into account.

The mixture cure model has been extensively investigated in the literature. Among others, studies of mixture cure model under right censored data include Ku and Chen [5], Sy and Taylor [6], Lu and Ying [7] and Fang et al. [8]. For current status data with a cured subgroup, Lam and Xue [9] and Ma [10] investigated the fitting of accelerate failure time model and cox model to such data, respectively. Furthermore, previous works on the analysis of interval censored data with cure models include Kim and Jhun [11] and Ma [12], who assume cox model for the failure times in the non-cured subgroup. The book by Maller and Zhou [13] presents a detailed discussion on the statistical inference method for mixture cure model. In this paper, we address regression analysis of partly interval censored failure time data with a cured subgroup through a mixture of the general linear model for the cure probability and a general class of transformation models for the failure times of non-cured subjects. For inference, an EM algorithm using Poisson variables will be developed to calculate the maximum likelihood estimators, which have the advantages of estimating the parameters in the cure model and survival model separately, and prove close-form estimator for the baseline cumulative hazards function of failure time.

* Corresponding author e-mail: yeqian.liu@mtsu.edu

The rest of this paper is organized, as follows: In Section 2, we introduce some notations, the model and some assumptions to be used throughout the paper and the corresponding likelihood function. The nonparametric maximum likelihood estimation procedure is developed in Section 3 and for the implementation of the procedure, an EM algorithm is proposed with the use of the subject-specified independent Poisson variables. In Section 4, we develop asymptotic properties of the estimators, including the consistency and asymptotic normality. Simulation results shown in Section 5 indicate that the proposed approach works well in finite sample. In Section 6, we apply the approach to NASA's Hypobaric decompression sickness data. Section 7 is dedicated to some discussion and concluding remarks.

2 Assumptions, Models and the Likelihood Function

Consider a failure time study in which there may exist a cured or non-susceptible subgroup. Let T denote the failure time of interest and D be cure indicator, which indicates, by the value 1 or 0, whether the subject is susceptible or not. Under the mixture cure modeling approach, the failure time of interest can be decomposed as

$$T = DT^* + (1-D)\infty,$$

where $T^* < \infty$ denotes the failure time of the susceptible subject. To describe the covariate effects on T and D, we will assume that given X, the cumulative hazard function of T^* takes the form

$$G\{\Lambda(t)\exp(X^T\beta)\}, \tag{1}$$

where G is a prespecified increasing transformation function, $\Lambda(\cdot)$ is an unknown baseline cumulative hazard function that is also increasing and β is a p-dimensional vector of regression parameters. Model (1) gives many commonly used models as special cases. For example, the choice of $G(x) = x$ yields proportional hazards model, and it gives proportional odds model when $G(x) = \log(1+x)$. It is apparent that covariates may have some effects on D. For this, we will assume that D follows the logistic model,

$$P(D=1) = p = \frac{\exp(\tilde{X}^T\gamma)}{1+\exp(\tilde{X}^T\gamma)}, \tag{2}$$

where the first component of $\tilde{X}$ is 1 and $\tilde{X}$ may share the same components as X, and γ is a q-dimensional vector of regression parameters. Therefore, the survival function of T under the mixture cure modeling approach is given by

$$1 - p + p\exp\left\{-G\left(\Lambda(t)\exp(X^T\beta)\right)\right\}.$$

Suppose that the study consists of n independent subjects and yields partly interval censored failure time data. For subject i, define $\Delta_i = 1$ if the failure time T_i can be observed exactly. For censored subject, let $(L_i, R_i]$ be the smallest interval that contains T_i. We define $\delta_{1i} = 1$ if ith subject is left censored by R_i with $L_i = 0$ in this situation, $\delta_{2i} = 1$ if the failure time of ith subject falls in a finite time interval $(L_i, R_i]$, and $\delta_{3i} = 1$ if the failure time of ith subject has not yet occurred at the last observation time L_i and set $R_i = \infty$ with the constraint that $\Delta_i + \delta_{1i} + \delta_{2i} + \delta_{3i} = 1$. Under the independent assumption between the failure times and the observation times, the likelihood function is given by

$$\begin{aligned}
&\prod_{i=1}^{n}\left\{p_i\lambda(t_i)e^{X_i^T\beta}G'\left(\Lambda(t_i)\exp(X_i^T\beta)\right)\exp\left[-G\left(\Lambda(t_i)\exp(X_i^T\beta)\right)\right]\right\}^{\Delta_i} \\
&\times\left\{p_i - p_i\exp\left[-G\left(\Lambda(R_i)\exp(X_i^T\beta)\right)\right]\right\}^{\delta_{1i}} \\
&\times p_i^{\delta_{2i}}\left\{\exp\left[-G\left(\Lambda(L_i)\exp(X_i^T\beta)\right)\right] - \exp\left[-G\left(\Lambda(R_i)\exp(X_i^T\beta)\right)\right]\right\}^{\delta_{2i}} \\
&\times\left\{1 - p_i + p_i\exp\left[-G\left(\Lambda(L_i)\exp(X_i^T\beta)\right)\right]\right\}^{\delta_{3i}}.
\end{aligned} \tag{3}$$

The transformation function G can be derived by Laplace transformation of frailty variable with support $[0,\infty)$ as the following form,

$$\exp\{-G(x)\} = \int_0^\infty \exp(-xt)\,\phi(t|r)\,\mathrm{d}t.$$

When $\phi(t|r)$ is the density function of a gamma variable with mean 1 and variance r, we can obtain $G(x) = \log(1+rx)/r$, the logarithmic transformation function. One can find more detailed discussion on the frailty based transformations in

Kosorok et al. [14]. Therefore, one can convert the transformation model (1) into proportional hazards frailty model and the likelihood function can be re-expressed, as follows:

$$\prod_{i=1}^{n}\int_{\mu_i}\left\{p_i\lambda(t_i)e^{X_i^T\beta}\mu_i\exp\left(-\Lambda(t_i)\exp(X_i^T\beta)\mu_i\right)\right\}^{\Delta_i}$$
$$\times\left\{p_i-p_i\exp\left(-\Lambda(R_i)\exp(X_i^T\beta)\mu_i\right)\right\}^{\delta_{1i}}$$
$$\times p_i^{\delta_{2i}}\left\{\exp\left(-\Lambda(L_i)\exp(X_i^T\beta)\mu_i\right)-\exp\left(-\Lambda(R_i)\exp(X_i^T\beta)\mu_i\right)\right\}^{\delta_{2i}}$$
$$\times\left\{1-p_i+p_i\exp\left(-\Lambda(L_i)\exp(X_i^T\beta)\mu_i\right)\right\}^{\delta_{3i}}\phi(\mu_i|r)\,\mathrm{d}\mu_i. \tag{4}$$

3 Maximum Likelihood Estimation

To estimate β, γ and $\Lambda(\cdot)$, we consider nonparametric maximum likelihood estimating approach, which leads to assume that $\Lambda(\cdot)$ is a step function with non-negative jumps at the distinct uncensored failure times, the observation times for the left-censored subjects and the endpoints of the smallest time intervals that contain the failure times of interest for the interval-censored subjects. Let $c_1<\ldots<c_{K_n}$ be the ordered distinct time points above. Denote λ_k for the non-negative jump size at c_k for $k=1,\ldots,K_n$. Then, we can rewrite the likelihood function, as follows:

$$\prod_{i=1}^{n}\int_{\mu_i}\left\{p_i\left(\prod_{k=1}^{K_n}\lambda_k^{I(c_k=t_i)}\right)\exp(X_i^T\beta)\mu_i\exp\left(-\sum_{c_k\le t_i}\lambda_k\exp(X_i^T\beta)\mu_i\right)\right\}^{\Delta_i}$$
$$\times\left\{p_i-p_i\exp\left(-\sum_{c_k\le R_i}\lambda_k\exp(X_i^T\beta)\mu_i\right)\right\}^{\delta_{1i}}$$
$$\times p_i^{\delta_{2i}}\left\{\exp\left(-\sum_{c_k\le L_i}\lambda_k\exp(X_i^T\beta)\mu_i\right)-\exp\left(-\sum_{c_k\le R_i}\lambda_k\exp(X_i^T\beta)\mu_i\right)\right\}^{\delta_{2i}}$$
$$\times\left\{1-p_i+p_i\exp\left(-\sum_{c_k\le L_i}\lambda_k\exp(X_i^T\beta)\mu_i\right)\right\}^{\delta_{3i}}\phi(\mu_i|r)\,\mathrm{d}\mu_i. \tag{5}$$

In the following, we will describe the derivation of our proposed EM algorithm, which relies on two-stage data augmentation involving subject-specific independent Poisson variables. In the first stage, it is natural to treat the cure indicator D_i and latent variable μ_i as missing values. Let θ be the parameters to be estimated, then the likelihood function would have the form

$$L_1(\theta)=\prod_{i=1}^{n}p_i^{D_i}(1-p_i)^{1-D_i}\left\{\left(\prod_{k=1}^{K_n}\lambda_k^{I(c_k=t_i)}\right)\exp(X_i^T\beta)\mu_i\exp\left(-\sum_{c_k\le t_i}\lambda_k\exp(X_i^T\beta)\mu_i\right)\right\}^{\Delta_i}$$
$$\times\left\{1-\exp\left(-\sum_{c_k\le R_i}\lambda_k\exp(X_i^T\beta)\mu_i\right)\right\}^{\delta_{1i}}$$
$$\times\left\{\exp\left(-\sum_{c_k\le L_i}\lambda_k\exp(X_i^T\beta)\mu_i\right)-\exp\left(-\sum_{c_k\le R_i}\lambda_k\exp(X_i^T\beta)\mu_i\right)\right\}^{\delta_{2i}}$$
$$\times\exp\left(-\sum_{c_k\le L_i}\lambda_k\exp(X_i^T\beta)\mu_i\right)^{\delta_{3i}D_i}\phi(\mu_i|r). \tag{6}$$

For the subject whose failure time can be observed exactly and the left censored or interval censored subject, it is apparent that $P(D_i=1)=1$. Then the likelihood function above is equivalent to the following form, which motivates us to further extend the pseudo observed data in the following second stage data augmentation.

$$\prod_{i=1}^{n} p_i^{D_i}(1-p_i)^{1-D_i}\left\{\left(\prod_{k=1}^{K_n}\lambda_k^{I(c_k=t_i)}\right)\exp(X_i^T\beta)\mu_i D_i\exp\left(-\sum_{c_k\le t_i}\lambda_k\exp(X_i^T\beta)\mu_i D_i\right)\right\}^{\Delta_i}$$
$$\times\left\{1-\exp\left(-\sum_{c_k\le R_i}\lambda_k\exp(X_i^T\beta)\mu_i D_i\right)\right\}^{\delta_{1i}}$$
$$\times\left\{\exp\left(-\sum_{c_k\le L_i}\lambda_k\exp(X_i^T\beta)\mu_i D_i\right)-\exp\left(-\sum_{c_k\le R_i}\lambda_k\exp(X_i^T\beta)\mu_i D_i\right)\right\}^{\delta_{2i}}$$
$$\times\exp\left(-\sum_{c_k\le L_i}\lambda_k\exp(X_i^T\beta)\mu_i D_i\right)^{\delta_{3i}}\phi(\mu_i|r). \tag{7}$$

In the second stage, for the ith subject, we introduce a set of new latent variables $\{Z_{ik}, k=1,\ldots,K_n\}$, where Z_{ik} is a Poisson random variable with the parameter $\lambda_k\exp(X_i^T\beta)\mu_i D_i$. By treating the latent variables μ_i's, D_i's and Z_{ik}'s as missing values, we would have the following pseudo complete data likelihood function,

$$L_c(\theta)=\prod_{i=1}^{n} p_i^{D_i}(1-p_i)^{1-D_i}\prod_{k=1}^{K_n}p(Z_{ik})\,\phi(\mu_i|r),$$

where $\sum_{c_k<t_i}Z_{ik}=0$ and $Z_{ik}|_{c_k=t_i}=1$ if $\Delta_i=1$, $\sum_{t_k\le R_i}Z_{ik}>0$ if $\delta_{1i}=1$, $\sum_{t_k\le L_i}Z_{ik}=0$ and $\sum_{L_i<t_k\le R_i}Z_{ik}>0$ if $\delta_{2i}=1$ and $\sum_{t_k\le L_i}Z_{ik}=0$ if $\delta_{3i}=1$. Note that similar Poisson variables were adopted by McMahan et al.[15], Wang et al.[16] and Zeng et al.[17] for the analysis of current status data or interval censored failure time data without considering the cured subgroup.

In the E-step, we take conditional expectations with respect to all latent variables in the log-likelihood function $l_c(\theta)=\log L_c(\theta)$. This yields $Q(\theta,\theta^{(m)})=Q_1(\gamma,\theta^{(m)})+Q_2(\theta_1,\theta^{(m)})$, where

$$Q_1(\gamma,\theta^{(m)})=\sum_{i=1}^{n}\tilde{X}_i^T\gamma\,\mathrm{E}(D_i)-\log\left\{1+\exp(\tilde{X}_i^T\gamma)\right\},$$

and

$$Q_2(\theta_1,\theta^{(m)})=\sum_{i=1}^{n}\sum_{k=1}^{K_n}X_i^T\beta\,\mathrm{E}(Z_{ik})+\log(\lambda_k)\mathrm{E}(Z_{ik})-\lambda_k\exp(X_i^T\beta)\,\mathrm{E}(\mu_i D_i).$$

To explain, we need to calculate the conditional expectations $\mathrm{E}(Z_{ik})$ and $\mathrm{E}(\mu_i D_i)$ and $\mathrm{E}(D_i)$ given in the following forms,

$$\mathrm{E}(Z_{ik})=\Delta_i\left\{I(c_k=t_i)+\lambda_k\exp(X_i^T\beta)\,\mathrm{E}(\mu_i D_i)\,I(c_k>t_i)\right\}$$
$$+\delta_{1i}\left\{\frac{\lambda_k\exp(X_i^T\beta)}{1-\exp(-G(W_i))}I(c_k\le R_i)+\lambda_k\exp(X_i^T\beta)\,\mathrm{E}(\mu_i D_i)\,I(c_k>R_i)\right\}$$
$$+\delta_{2i}\lambda_k\exp(X_i^T\beta)\frac{\int_{\mu_i}\mu_i(\exp(-\mu_i V_i)-\exp(-\mu_i W_i))\{1-\exp(-\mu_i(W_i-V_i))\}^{-1}\phi(\mu_i|r)\,\mathrm{d}\mu_i}{\exp(-G(V_i))-\exp(-G(W_i))}I(L_i<c_k\le R_i)$$
$$+\delta_{2i}\lambda_k\exp(X_i^T\beta)\,\mathrm{E}(\mu_i D_i)\,I(c_k>R_i)+\delta_{3i}\lambda_k\exp(X_i^T\beta)\,\mathrm{E}(\mu_i D_i)\,I(c_k>L_i),$$

$$\mathrm{E}(\mu_i D_i)=\Delta_i\frac{\int_{\mu_i}\mu_i^2\exp(-\mu_i M_i)\,\phi(\mu_i|r)\,\mathrm{d}\mu_i}{\exp(-G(M_i))\,G'(M_i)}+\delta_{1i}\frac{1-\exp(-G(W_i))\,G'(W_i)}{1-\exp(-G(W_i))}$$
$$+\delta_{2i}\frac{\exp(-G(V_i))\,G'(V_i)-\exp(-G(W_i))\,G'(W_i)}{\exp(-G(V_i))-\exp(-G(W_i))}+\delta_{3i}\frac{p_i\exp(-G(V_i))G'(V_i)}{1-p_i+p_i\exp(-G(V_i))},$$

and

$$\mathrm{E}(D_i)=\Delta_i+\delta_{1i}+\delta_{2i}+\delta_{3i}\frac{p_i\exp(-G(V_i))}{1-p_i+p_i\exp(-G(V_i))},$$

where $V_i = \sum_{c_k \le L_i} \lambda_k \exp(X_i^T \beta)$, $W_i = \sum_{c_k \le R_i} \lambda_k \exp(X_i^T \beta)$, $M_i = \sum_{c_k \le t_i} \lambda_k \exp(X_i^T \beta)$, and

$$G'(x) = \frac{\int_{\mu_i} \mu_i \exp(-x\mu_i)\,\phi(\mu_i|r)\,\mathrm{d}\mu_i}{\exp(-G(x))} = \frac{(rx+1)^{-r^{-1}-1}}{\exp(-G(x))},$$

and

$$\int_{\mu_i} \mu_i^2 \exp(-\mu_i x)\,\phi(\mu_i|r)\,\mathrm{d}\mu_i = (1+r)(rx+1)^{-r^{-1}-2}.$$

when $\phi(\mu_i|r)$ is the gamma density function with known parameter r. Furthermore, we propose to employ Gauss-Laguerre quadrature technique to calculate the following conditional expectation that has no closed-form,

$$\int_{\mu_i} \mu_i(\exp(-\mu_i V_i) - \exp(-\mu_i W_i))\,\{1-\exp(-\mu_i(W_i - V_i))\}^{-1}\,\phi(\mu_i|r)\,\mathrm{d}\mu_i.$$

In the M-step, we can update γ by solving the following score function,

$$\sum_{i=1}^{n} \tilde{X}_i \left\{ \mathrm{E}(D_i) - \frac{\exp(\tilde{X}_i^T \gamma)}{1+\exp(\tilde{X}_i^T \gamma)} \right\}. \tag{8}$$

Setting $\frac{\partial Q_2(\theta_1, \theta^{(m)})}{\partial \lambda_k} = 0$, we can update λ_k with the following closed-form expression,

$$\lambda_k = \frac{\sum_{i=1}^{n} \mathrm{E}(Z_{ik})}{\sum_{i=1}^{n} \mathrm{E}(\mu_i D_i)\exp(X_i^T \beta)}, k = 1, \ldots, K_n. \tag{9}$$

Plugging the estimator above into equation $Q_2(\theta_1, \theta^{(m)})$, we can get the estimating equation for the regression parameter β,

$$\sum_{i=1}^{n} \sum_{k=1}^{K_n} \mathrm{E}(Z_{ik}) \left\{ X_i - \frac{\sum_{i=1}^{n} \mathrm{E}(\mu_i D_i)\exp(X_i^T \beta) X_i}{\sum_{i=1}^{n} \mathrm{E}(\mu_i D_i)\exp(X_i^T \beta)} \right\} = 0. \tag{10}$$

By combining all the above-mentioned discussions, the proposed EM algorithm can be summarized, as follows:
Step 0. Choose an initial value $\theta^{(0)}$.
Step 1. At the $(m+1)$th iteration, first calculate the conditional expectations $\mathrm{E}(Z_{ik})$, $\mathrm{E}(\mu_i D_i)$ and $\mathrm{E}(D_i)$ at $\theta = \theta^{(m)}$.
Step 2. Determine the updated estimators $\gamma^{(m+1)}$ from the estimating function (8) using one step Newton-Raphson method.
Step 3. Update $\beta^{(m+1)}$ by solving the score equation (10) with one step Newton-Raphson method.
Step 4. Update $\lambda_k^{(m+1)}$ by (9) by replacing β with $\beta^{(m+1)}$.
Step 5. Repeat Steps 1 - 4 until the convergence is satisfied.

For the convergence, various criteria can be applied as well as a simple and commonly used one is to check $\|\hat{\theta}^{(m+1)} - \hat{\theta}^{(m)}\| \le \varepsilon$, $\|\hat{\beta}^{(m+1)} - \hat{\beta}^{(m)}\| \le \varepsilon$, $\|\hat{\gamma}^{(m+1)} - \hat{\gamma}^{(m)}\| \le \varepsilon$ for a given positive number $\varepsilon = 10^{-6}$. We also set an addition termination criterion that is to stop the EM algorithm when the standard errors of the $\alpha^{(k)}$'s, $\beta^{(k)}$'s, $\gamma^{(k)}$'s from 30 consecutive iterations are all smaller than $\varepsilon_1 = 10^{-3}$, also a given positive number.

For inference about the parameters of interest, $\eta = (\beta^T, \gamma^T)^T$, we propose to employ nonparametric bootstrap method to estimate the the asymptotic covariance matrix of η. To be specific, we first draw new data sets, say $O^{(q)}$'s, of sample size n with replacement from the original observed data Q times and let $\tilde{\eta}$ contain all the resulting estimators of η based on $O^{(q)}$ for $q = 1, \ldots, Q$. Then one can use the sample covariance matrix of $\tilde{\eta}$ as an estimated covariance matrix of η. The theoretical justifications of the bootstrap method under semiparametric models were given by Cheng and Huang [18], and Cheng [19], with focusing on the distribution consistency and moment consistency, respectively.

4 A Simulation Study

In this section, we present some results obtained from a simulation study conducted to assess the finite sample performance of the inference procedure proposed in the previous sections. The cure indicator and failure time of interest (if not cured) were generated from model (1) and model (2) with $G(x) = \log(1+rx)/r (r \ge 0)$ with different values for r and set $\Lambda(t) = 0.2t$. Note that as mentioned above, it gives the proportional hazards model and the proportional odds model with $r = 0$ and 1, respectively. For each subject i, we first generated three random variables, U_{i1}, U_{i2} and U_{i3}, which followed

the uniform distribution over $(0, 1)$, $(0.2+U_{i1}, \tau/2)$, and $(0.5+U_{i2}, \tau)$, respectively. If T_i belongs to $(U_{i1}, U_{i2}]$, then we assumed that T_i can be observed exactly, T_i was left-censored by U_{i1} if $T_i \leq U_{i1}$, and if $T_i > U_{i3}$, then T_i was right-censored or cured. Otherwise, T_i was interval-censored by U_{i2} and U_{i3}. Here, we considered the situations that $\tilde{X}_i = (1, X_i^T)^T$ and $\tilde{X}_i \neq (1, X_i^T)^T$. The results given below are based on 1000 replications, $\tau = 3$, $Q = 100$, and $n = 200$ or 400.

Tables 1 and 2 present the results on the estimation of regression parameters β and γ with $\tilde{X}_i = (1, X_i^T)^T$. The former investigated the one covariate situation with the X_i's following the Bernoulli distribution with the success probability of 0.5 and we set $\beta = 0$ and $\gamma = (1, -0.5)^T$, which yields about 39% and 32% right censored and cure rates, respectively. The latter considered the two covariate situation, where $X_i = (X_{1i}, X_{2i})^T$, with the X_{1i}'s being generated from the Bernoulli distribution with the success probability of 0.5 and the X_{2i}'s from the uniform distribution over $(-1, 1)$. In this situation, we set $\beta = (0.5, -0.5)^T$ and $\gamma = (1, -0.5, -0.5)^T$ corresponding to about 39% and 32% right censored and cure rates, respectively. In both tables, the results include the estimated bias (Bias) given by the average of the estimates minus the true value, the sample standard error (SSE) of the obtained estimates, the average of the standard error estimates (SEE), and the 95% empirical coverage probability (CP). One can see that they suggest that the proposed estimator seems to be unbiased and the variance estimation based on the bootstrap procedure seems to be reasonable. In addition, all empirical coverage probabilities are in well accordance with the nominal value, indicating that the normal approximation to the distribution of the proposed estimator appears to be appropriate, and as expected, the results become better when the sample size increases.

Table 3 present the results on the estimation of regression parameters β and γ with $\tilde{X}_i \neq (1, X_i^T)^T$, where $\tilde{X}_i = (1, \tilde{X}_{1i}, \tilde{X}_{2i})^T$ and $X_i = (X_{1i}, X_{2i})^T$. Specially, we generated X_{1i} and $\tilde{X}_{1i}$ from two independent Bernoulli distribution with the success probability of 0.5, X_{2i} and $\tilde{X}_{2i}$ from two independent uniform distribution over $(-1, 1)$. We set the true values of regression parameters the same as those in Table 2 and so obtain the similar right censored and cure rates as above. It is apparent that the results given in Table 3 yielded similar conclusions as above and indicate that the proposed inference procedure works well for the practical situation considered here.

5 An Application

Now we apply the methodology proposed in the previous sections to the NASA's Hypobaric decompression sickness database (HDSD). The database consists of 177 male (SEX=1) and 61 female (SEX=0) volunteers, aged between 20 to 54 years, who underwent dehydrogenation test procedures and were then exposed to a hypobaric environment. The variable of interest is the time to onset of grade IV venous gas emboli (VGE). The study aims to investigate the mechanism of the onset of grade IV VGE and its association with four explanatory variables, age, gender and two experimental variables TR360 and NOADYN. The variable TR360 is the tissue ratio at 360 degrees, which is a measure of decompression stress. The higher the TR360, the more quickly a high grade bubble is expected to occur. It is a continuous variable ranging from 1.04 to 1.89. The variable NOADYN is also an experimentally manipulated variable, indicating whether the test subject was ambulatory (NOADYN = 1; 195 subjects) or lower body adynamic (NOADYN = 0; 43 subjects) during the test session.

The time to onset of grade IV VGE is either exactly observed or known to lie between certain examination time points, leading to interval-censored data. Furthermore, it has been suggested that not every subject will develop Grade IV VGE, so there is an obvious need for a cure model. Figure 1 presents the Turnbull nonparametric estimate of the survival function and one can see that the estimated survival curve has a clear non-zero plateau at the tail, indicating the possible existence of a cured subgroup. In other words, it seems to be reasonable and more appropriate to incorporate the cure model for analyzing the data here.

Previous authors (Conkin and Powell [20]) suggested that only subject-specific covariates can have an influence on the susceptibility to experience the event. Therefore, only age and gender are considered in the logistic component for the mixture cure rate model. Table 4 presents the estimation results obtained by the application of the proposed method. On the effect of the TR360, the estimates indicate that the subjects with higher TR360 had longer time-to-onset of Grade IV VGE. Furthermore, the estimated coefficients suggest that both age and NOADYN are significant, indicating that older people may develop Grade IV VGE less rapidly and ambulatory subjects may develop the event more rapidly than lower body adynamic subjects. The estimate also indicates that male subjects may have shorter time to the event than female subjects, but this is not significant at the 5% level. For the cure probability, both age and sex are significant, indicating that older male subjects were more susceptible to Grade IV VGE.

6 Discussion and Concluding Remarks

The classical hazard-based regression model is inappropriate for analyzing failure time data when there is a non-ignorable proportion of cured subjects in the whole population who would never experience the event of interest. In this paper,

mixture cure model was proposed to model partly interval censored data in the presence of a cured subgroup. Specially, we assumed a logistic model for the cured probability and a wide class of transformation models for the failure times of the uncured group. For inference, we developed an EM algorithm with the use of subject-specific independent Poisson variables and by treating the cured indicators as missing values to obtain the maximum likelihood estimators. The proposed EM algorithm has some very attractive features; for example the unknown high-dimensional baseline cumulative hazard function denoted by the λ_k's can be calculated explicitly and the low-dimensional regression parameters can be easily updated with one-step Newton-Raphson method. In addition, numerical results indicated that the proposed estimating method is reliable and appropriate for practical situations.

Model-checking is often of interest and challenging in mixture cure model. In practice, we selected the transformation function by maximizing the log-likelihood function. It would be helpful to develop a formal model selection or checking method to assess the adequacy of the transformation models for the failure times of uncured subgroup and the logistic regression model for the cure probability. Furthermore, the additive hazards model attracts great attention in survival analysis, which assumes a different type of covariate effect on the hazards function of the failure time from the transformation models discussed above. It is worthwhile to develop the estimating method to analyze partly interval censored data with a cured subgroup under additive hazards model.

Table 1: Simulation results for the regression parameter with one common covariate.

		$n = 200$				$n = 400$			
r		Est	SSE	SEE	CP	Est	SSE	SEE	CP
0	γ_0	0.003	0.230	0.232	95.2	-0.006	0.157	0.161	95.3
	γ_1	0.012	0.311	0.315	95.7	0.002	0.212	0.219	96.4
	β	-0.008	0.296	0.300	95.3	0.008	0.206	0.206	95.5
0.5	γ_0	0.003	0.229	0.233	95.0	0.002	0.156	0.161	95.8
	γ_1	0.012	0.309	0.315	94.8	-0.005	0.219	0.218	95.0
	β	0.008	0.339	0.344	96.2	0.001	0.235	0.236	94.3
1	γ_0	0.007	0.228	0.233	95.4	0.007	0.155	0.162	95.8
	γ_1	-0.005	0.317	0.314	94.7	-0.007	0.210	0.219	95.5
	β	0.017	0.376	0.385	95.7	-0.009	0.256	0.265	96.4

Table 2: Simulation results for the regression parameters with two common covariates.

		$n=200$				$n=400$			
r		Est	SSE	SEE	CP	Est	SSE	SEE	CP
0	γ_0	0.024	0.240	0.239	95.3	0.013	0.163	0.164	94.9
	γ_1	-0.020	0.319	0.322	95.0	-0.009	0.229	0.222	95.2
	γ_2	-0.015	0.273	0.282	95.4	0.003	0.190	0.193	94.7
	β_1	-0.016	0.276	0.275	94.1	-0.008	0.182	0.186	95.3
	β_2	0.012	0.241	0.246	95.4	0.002	0.172	0.167	93.8
0.5	γ_0	0.027	0.223	0.238	96.2	0.017	0.163	0.164	95.3
	γ_1	-0.025	0.314	0.322	94.7	-0.018	0.221	0.221	94.9
	γ_2	-0.009	0.280	0.281	94.6	-0.009	0.194	0.195	94.5
	β_1	-0.019	0.322	0.326	95.2	-0.002	0.221	0.224	95.5
	β_2	-0.004	0.280	0.290	96.4	0.002	0.203	0.200	94.8
1	γ_0	0.026	0.241	0.240	95.1	0.012	0.161	0.164	94.5
	γ_1	-0.018	0.324	0.321	95.1	-0.009	0.219	0.221	94.9
	γ_2	-0.010	0.279	0.282	95.5	-0.009	0.202	0.196	94.5
	β_1	-0.026	0.367	0.383	96.3	0.002	0.237	0.255	96.9
	β_2	-0.012	0.338	0.337	94.9	0.004	0.220	0.228	96.0

Table 3: Simulation results for the regression parameters with two different covariates.

		$n=200$				$n=400$			
r		Est	SSE	SEE	CP	Est	SSE	SEE	CP
0	γ_0	0.010	0.232	0.238	96.1	0.008	0.154	0.165	96.6
	γ_1	0.001	0.318	0.322	95.7	0.000	0.209	0.220	95.8
	γ_2	-0.019	0.277	0.281	95.7	-0.007	0.197	0.194	94.5
	β_1	-0.014	0.267	0.278	95.2	-0.010	0.186	0.189	95.7
	β_2	0.019	0.244	0.241	95.0	0.016	0.167	0.166	95.0
0.5	γ_0	0.013	0.221	0.238	96.0	0.008	0.157	0.163	95.9
	γ_1	0.006	0.304	0.320	95.8	-0.001	0.208	0.220	96.1
	γ_2	-0.014	0.285	0.282	95.7	-0.015	0.187	0.194	96.1
	β_1	-0.030	0.336	0.329	94.6	-0.009	0.217	0.225	95.5
	β_2	-0.008	0.295	0.290	95.0	0.011	0.192	0.196	95.1
1	γ_0	0.026	0.241	0.239	95.3	0.012	0.162	0.164	95.6
	γ_1	-0.007	0.322	0.321	94.4	-0.010	0.218	0.222	95.1
	γ_2	-0.015	0.285	0.282	95.1	-0.002	0.193	0.194	94.6
	β_1	-0.018	0.363	0.380	96.0	-0.008	0.258	0.257	94.7
	β_2	0.007	0.312	0.331	96.6	0.004	0.224	0.224	95.4

Table 4: Analysis results of NASA's Hypobaric Decompression Sickness Data

			EST	SEE	p-value
$m=3$	Logistic model	Intercept	-3.063	0.581	<0.001
		Age	1.179	0.363	0.001
		Gender	1.589	0.579	0.006
	PH model	Age	-0.318	0.071	<0.001
		Gender	-0.163	0.186	0.380
		TR360	-1.196	0.536	0.026
		NOADYN	0.865	0.321	0.007
$m=4$	Logistic model	Intercept	-2.944	0.562	<0.001
		Age	1.106	0.355	0.002
		Gender	1.504	0.562	0.007
	PH model	Age	-0.294	0.068	<0.001
		Gender	-0.152	0.173	0.383
		TR360	-1.092	0.508	0.031
		NOADYN	0.809	0.306	0.008
$m=5$	Logistic model	Intercept	-2.815	0.546	<0.001
		Age	1.028	0.339	0.002
		Gender	1.435	0.540	0.007
	PH model	Age	-0.274	0.062	<0.001
		Gender	-0.135	0.159	0.395
		TR360	-1.028	0.487	0.034
		NOADYN	0.785	0.287	0.006

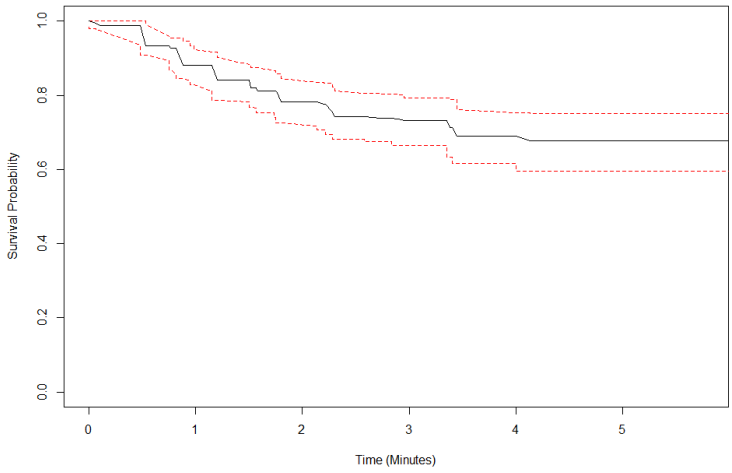

Fig. 1: Turnbull non-parametric estimated survival probability for the time to Grade IV VGE

Acknowledgement

The authors are grateful to the anonymous referee for the careful checking of the details and the constructive comments that improved this paper.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article

References

[1] P. M. Odell, K. M. Anderson and R.B. D'Agostino. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, **48**, 951-959 (1992).
[2] B. W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Biometrika*, **38**, 290-295 (1976).
[3] J. S. Kim. Maximum likelihood estimation for the proportional hazards model with partly interval- censored data. *Journal of the Royal Statistical Society: Series B*, **65**, 489-502 (2003).
[4] X. Zhao, Q. Zhao, J. Sun, J. S. Kim. Generalized log-rank tests for partly interval-censored failure time data. *Biometrical Journal*, **3**, 375-385 (2008).
[5] V. T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *The Annals of Statistics*, **38**, 1041-1046 (1987).
[6] J. P. Sy and J.M.G. Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, **56**, 227-236 (2000).
[7] W. Lu, and Z. Ying. On semiparametric transformation cure models. *Biometrika*, **91**, 331-343 (2004).
[8] H. Fang, G. Li and J. Sun. Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics*, **32**, 59-75 (2005).
[9] K. F. Lam and H. Xue. A Semiparametric Regression Cure Model with Current Status Data. *Biometrika*, **92**, 573-586 (2005).
[10] S. Ma. Cure model with current status data. *Statistica Sinica*, **19**, 233-249 (2009).
[11] Y. J. Kim and M. Jhun. Cure rate model with interval censored data. *Statistics in medicine*, **27**, 3-14 (2008).
[12] S. Ma. Mixed case interval censored data with a cured subgroup. *Statistica Sinica*, **20**, 1165-1181 (2010).
[13] R. Maller and X. Zhou. *Survival analysis with long-term survivors*, Springer, Wiley, Hoboken, New Jersey USA, 112-117 (1996).
[14] M. R. Kosorok, B.L. Lee and J. P. Fine. Robust inference for univariate proportional hazards frailty regression models. *The Annals of Statistics*, **32**, 1448-1491 (2004).
[15] C. S. McMahan, L. Wang and J. M. Tebbs. Regression analysis for current status data using the EM algorithm. *Statistics in Medicine*, **32**, 4452-4466 (2013).
[16] L. Wang, C. S. McMahan, M. G. Hudgens and Z. P. Qureshi. A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics*, **72**, 222-231 (2016).
[17] D. Zeng, L. Mao and D. Y. Lin. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, **103**, 253-271 (2016).
[18] G. Cheng and J. Huang. Bootstrap consistency for general semiparametric M-estimation. *The Annals of Statistics*, **38**, 2884-2915 (2010).
[19] G. Cheng. Moment consistency of the exchangeably weighted bootstrap for semiparametric M-estimation. *Scandinavian Journal of Statistics*, **42**, 665-684 (2015).
[20] J. Conkin and M. Powell. Lower body adynamia as a factor to reduce the risk of hypobaric decompression sickness. *Aviation, Space and Environmental Medecine*, **72**, 202–214 (2001).