Abstract

In this paper, a cure fraction model for interval-censored data with a change point according to a covariate threshold is proposed. Maximum likelihood estimators of the model parameters are obtained using the Expectation Maximization (EM) algorithm. A critical challenge to this method was that the likelihood function is not differentiable with respect to the unknown change point parameter. Simulation studies were conducted to evaluate the performance of the proposed estimation method. The numerical results showed that the new model represents a valuable advancement of cure models.

Keywords: Change-Point Model, Cure Model, EM Algorithm, Interval-Censored Data, Smoothing

1. Introduction

A change-point problem is a problem in which changes at unknown points are identified and the locations of these changes are estimated. This type of problem occurs quite frequently in survival studies. For example, data obtained from a group of pre-school boys indicate that their weight/height ratios relate to their ages in one way before a certain age but that the observed functional relation changes afterwards. As another example, cancer incidence rates remain relatively stable in young people but change drastically after a certain age threshold.

A number of researchers have introduced change-point models into the field of survival analysis. Matthews and Farewell assumed that the hazard function is constant, with the exception of a jump. Muller and Wang proposed non-parametric method for estimation of the changes in the hazard rate. Sen and Pons considered a Cox model with change point in accordance with an unknown threshold of a covariate. However, these models do not incorporate the proportion of cure. A cure rate model represents a combination of cure fractions and survival models and can be applied to clinical studies of several types of cancer such as breast cancer, head and neck cancer, and prostate cancer. In such cases, a proportion of the population of patients may not be susceptible to the event of interest, namely, recurrence of cancer.

Survival cure models have been widely studied for decades. The interested readers can refer, for example, to the books of Maller and Zhou, as well as to the review article prepared by Tsodikov et al. or to the articles of Boag, Berkson et al., Chen et al., Lambert et al., and Yu et al. However, the literature on cure models with change-point problems is rather limited, and only few recent studies have addressed this type of model. For example, Zhao et al. proposed a mixture cure model with a change-point at an unknown threshold of failure time and Othus et al. developed a cure survival model that allowed for a change-point effect in covariates in order to investigate a potential change point in the age of diagnosis of prostate cancer. Both of these studies employed right-censored data. However, this paper presents an approach to analysis of the cure rate model in the presence of interval-censored data with a change-point at an unknown threshold of a covariate. Here, we study parametric maximum likelihood inference via the

*Author for correspondence
Expectation Maximization (EM) algorithm, which can accommodate a smooth transition as well as an abrupt change to determine the parameters’ values.

This article is organized as follows: in Section 2, the notations and model are described. In Section 3, we propose a model based on interval censoring and develop Maximum Likelihood Estimation (MLE) technique based on a covariate change point. Sections 4 and 5 present the estimation procedures, namely, the smoothed approach and the EM algorithm, while the simulation studies and their major findings are presented in Section 6. We conclude with discussion of the major research findings and conclusions in Section 7.

2. Model Formulation

In this section, we first present a brief review of the hazard function for survival data with a cure fraction. In this type of data, the survival time is assumed to take the form $T = vT^* + (1 - v)\infty$, where $T^* < \infty$ denotes the failure time of an uncured subject and $v$ is an indicator showing whether the sampled subject is cured ($v = 0$) or not ($v = 1$). Let $\eta = P_i(v = 0)$ express the proportion of cured subjects, $\eta \in (0,1]$. Then, the survival function for $T$ which is known as the cure model, is given by Zhou et al.\(^1\)

$$S(t) = P_i(T \geq t) = \eta + (1 - \eta)S^*(t)$$

where $S^*(t)$ is the survival function for those who are uncured. A parametric model can be specified for the failure time. In this work, we consider the lognormal distribution, which is one of the commonly-used distributions for modeling the failure time of the uncured subjects. It is positively skewed and hence can be applied to right-skewed data arising from medical, biological, environmental and reliability studies. Further details on this issue can be found in Boag\(^10\) and Frankel and Longmate\(^18\).

In medical studies, one often has to deal with censored survival times due to a variety of potential reasons, such as dropping-out of patients from the study or termination of the observation period. In general, data drawn from patients who are cured will always appear as censored. However, a censored patient is not necessarily a cured one. Let $C$ denote the censoring time and let $X$ be a covariate related to $T^*$ and the cure indicator $v$. Following the usual formulation, we define the observable survival time $\tilde{T}$ by $\tilde{T} = \min(T, C)$ and $\delta = I(T \leq C)$ where $I(\cdot)$ is the indicator function, which takes the value of 1 if $T$ is an exact failure time (uncensored) and the value of 0 if $T$ is censored. It is further assumed that $T$ is independent of $T^*$. Then, the observations consisting of $(\tilde{T}, \delta, X)$ for $i = 1,2,...,n$ are independent and identically-distributed consisting of $(\tilde{T}, \delta, X)$. Based on the cure model in \(^1\) and the lognormal distribution, the density function of $T$ at $t$ can be written

$$f(t) = \frac{1 - \eta(X)}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\ln \tilde{T} - \mu(X)}{\sigma}\right)^2\right)$$

as and the survival function can be written as,

$$S(t) = (1 - \eta(X))\left[1 - \Phi\left(\frac{\ln \tilde{T} - \mu(X)}{\sigma}\right)\right]$$

where $\Phi(.)$ is the distribution function of the standard normal. The complete likelihood function for the $n$ observed survival data is then\(^2\)

$$L = \prod_{i=1}^{n} \left[\frac{1 - \eta(X)}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\ln \tilde{T} - \mu(X)}{\sigma}\right)^2\right)\right]^{\delta_i} \times \left[\eta(X) + (1 - \eta(X))\left[1 - \Phi\left(\frac{\ln \tilde{T} - \mu(X)}{\sigma}\right)\right]\right]^{1-\delta_i}$$

3. The Model with Interval Censoring

Regarding interval-censored data, the exact failure time $\tilde{T}$ is unknown even for a subject who is not right censored. Instead, an interval $(t_{L}, t_{R})$ is observed, where $\tilde{T} \in (t_{L}, t_{R})$ and $t_{L} \leq t_{R}$. In this section, we describe the cure rate model for the interval-censored data and extend it to the cure rate model with a change-point effect in the covariate $X$. The model can be used for multiple covariates.

3.1 Data and Likelihood

Considering a setting in which the event time $\tilde{T}$ is known to have occurred within two points $(t_{L}, t_{R})$, where $t_{L}$ is the latest examination time before the event and $t_{R}$ is the earliest examination time after the event, and $t_{R} = \infty$ if a subject has not experienced the event of interest before
the last follow-up. Then, the observed data are \((t_{Li}, t_{Ri}, X_i)\), \(i = 1, 2, ..., n\) and model\(^4\) can be re-expressed as

\[
P_i(t_{Li} \leq t_i \leq t_{Ri}) = P_i(t_i \geq t_{Li}) - P_i(t_i \geq t_{Ri})
\]

\[
= (1 - \eta(X_i)) \left[ S^*(t_{Li} | X) - S^*(t_{Ri} | X) \right]
\]

The present study assumes that the change-point of the model depends on the covariate \(X\), and at this point, the probability of cure, or survival function, takes a sudden jump or fall. Suppose that the change-point is \(\tau\). If \(X \leq \tau\), then let \(\eta(X) = p_1\) and \(S(X) = S_1\). However, if \(X > \tau\), then \(\eta(X) = p_2\) and \(S(X) = S_2\). Now, reformat the censoring indicator \(d_i\) as follows: \(d_i = I(t_{Ri} < \infty)\) for \(t_{Li} < t_i \leq t_{Ri}\) and define the unknown parameters as \(\theta = (p_1, p_2, \mu_1, \mu_2, \sigma_1, \sigma_2, \tau)\). Then, the likelihood function for the \(n\) observed, interval-censored data based on the Mixture Cure Model (MCM) with a change-point effect in a covariate, referred to as \(L_n^*(\theta)\), is\(^3\)

\[
\prod_{i=1}^{n} \left\{ (1 - p_i) \left[ \Phi \left( \frac{\ln t_{Ri} - \mu_1}{\sigma_1} \right) - \Phi \left( \frac{\ln t_{Li} - \mu_1}{\sigma_1} \right) \right]^{d_i} 
\right\} \times\left\{ p_i + (1 - p_i) \left[ 1 - \Phi \left( \frac{\ln t_{Li} - \mu_1}{\sigma_1} \right) \right] \right\}^{1-d_i} I(X \leq \tau)
\]

\[
\times\left\{ (1 - p_2) \left[ \Phi \left( \frac{\ln t_{Ri} - \mu_2}{\sigma_2} \right) - \Phi \left( \frac{\ln t_{Li} - \mu_2}{\sigma_2} \right) \right]^{d_i} 
\right\} \times\left\{ p_2 + (1 - p_2) \left[ 1 - \Phi \left( \frac{\ln t_{Li} - \mu_2}{\sigma_2} \right) \right] \right\}^{1-d_i} I(X > \tau)
\]

Because the change point-parameter \(\tau\) is unknown and according to the classical likelihood approach, the likelihood function is not differentiable with respect to \(\tau\). In consequence, one of the regularity prerequisites for the normal asymptotic theory, that is, a specific degree of smoothness of the objective function concerning the parameters is violated. In order to address this critical problem, the smoothed likelihood function is suggested.

### 4. Smoothed Likelihood Approach

To circumvent the critical problem of non-smoothing, a smoothed likelihood approach is proposed. The idea of this approach is to use a continuous and differentiable function to approximate the indicator functions \(I(X \leq \tau)\) and \(I(X > \tau)\). Let \(K(u)\) be a continuous function such that \(K(u)\) satisfies \(\lim_{u \rightarrow -\infty} K(u) = 0\) and \(\lim_{u \rightarrow +\infty} K(u) = 1\).

Let \(K_n(u) = K(u / h_n)\) where \(h_n\) is a small positive constant that depends on the sample size. A special case of this class of functions is the logistic function where

\[
K_n(u) = \frac{\exp[u / h_n]}{1 + \exp[u / h_n]}. 
\]

It is recognized in the density estimation literature that the choice of the smoothing function does not affect the asymptotic properties of the density estimator. In view of this, we conducted the simulations with two different distribution functions for \(K(u)\): the logistic function and the distribution function of the standard normal distribution. Then, based on this function, the smoothed likelihood function for the observed data \((t_{Li}, t_{Ri}, d_i, X_i)\), which is referred to as \(L_n^*(\theta)\), is

\[
\prod_{i=1}^{n} \left\{ (1 - p_i) \left[ \Phi \left( \frac{\ln t_{Ri} - \mu_1}{\sigma_1} \right) - \Phi \left( \frac{\ln t_{Li} - \mu_1}{\sigma_1} \right) \right]^{d_i} 
\right\} \times\left\{ p_i + (1 - p_i) \left[ 1 - \Phi \left( \frac{\ln t_{Li} - \mu_1}{\sigma_1} \right) \right]^{1-d_i} K_n(\tau - X_i)
\right\}
\]

\[
\times\left\{ (1 - p_2) \left[ \Phi \left( \frac{\ln t_{Ri} - \mu_2}{\sigma_2} \right) - \Phi \left( \frac{\ln t_{Li} - \mu_2}{\sigma_2} \right) \right]^{d_i} 
\right\} \times\left\{ p_2 + (1 - p_2) \left[ 1 - \Phi \left( \frac{\ln t_{Li} - \mu_2}{\sigma_2} \right) \right] \right\}^{1-d_i} K_n(\tau - X_i) \}^{1-d_i} K_n(\tau - X_i) \}
\]

The likelihood function does not have closed-form solutions for the maximum likelihood estimators, and use of a numerical technique is thus necessary. Following Peng and Dear and Zhang and Peng, we simplify the computation by rewriting the likelihood function using partially-complete censored observations. Thus, the complete likelihood function can be converted into the following modified likelihood function:

\[
L_n = \prod_{i=1}^{n} \left\{ (1 - p_i) \left[ \Phi \left( \frac{\ln t_{Ri} - \mu_1}{\sigma_1} \right) - \Phi \left( \frac{\ln t_{Li} - \mu_1}{\sigma_1} \right) \right]^{d_i} 
\right\} \times\left\{ p_i + (1 - p_i) \left[ 1 - \Phi \left( \frac{\ln t_{Li} - \mu_1}{\sigma_1} \right) \right]^{1-d_i} K_n(\tau - X_i)
\right\}
\]

\[
\times\left\{ (1 - p_2) \left[ \Phi \left( \frac{\ln t_{Ri} - \mu_2}{\sigma_2} \right) - \Phi \left( \frac{\ln t_{Li} - \mu_2}{\sigma_2} \right) \right]^{d_i} 
\right\} \times\left\{ p_2 + (1 - p_2) \left[ 1 - \Phi \left( \frac{\ln t_{Li} - \mu_2}{\sigma_2} \right) \right] \right\}^{1-d_i} K_n(\tau - X_i) \}
\]
The simulation studies were conducted in the current work to investigate finite sample performance of the proposed estimator considering two smoothing functions $K(.)$; the logistic function and the distribution function of the standard normal distribution. For each studied function, two simulation scenarios were taken into consideration. In one scenario, a Uniform ($U(0,1)$) random variable with a change-point at 0.5 was used while in the other scenario a truncated Normal ($tN(1,1,0,2)$) random variable with a change-point at 1 was employed. The interval-censored survival times ($t_{Li}, t_{Ri}, \delta$) with the cure fraction were generated in a manner similar to that employed by Pan$^{23}$, and Kim and Jhun$^{19}$. First, we generated random numbers $u_i$ from uniform distribution based on (0, 1) to determine whether a subject is cured $u_i \leq \eta_i$ or not. Interval-censored data ($t_{Li}, t_{Ri}, \delta$) were then generated as follows:

- If the subject is cured, then sample $T_i = T_{Li}$ from the lognormal distribution with $(2,0.15)$ and $\delta_i = 0$.
- If the subject is not cured, then generate $T_i$ from a log-normal model with $(\mu, \sigma)$ and generate the censoring time $C_i$ from the lognormal distribution with $(2,0.15)$, which results in a censoring rate of approximately 35%. $\delta_i$ is equal to 1 if $T_i \leq C_i$ and to 0 otherwise.
- For $\delta_i = 1$, we generate lens from uniform distribution on $(0.2, 0.7)$ and $l_i$ from uniform distribution on $(0, 1)$. Then, from $(0, l_i], (l_i, l_i + l_{len}], ..., (l_i + k \cdot l_{len}, \infty)$, $k=1, 2, ...,$ we choose as that satisfying $T_{Li} < T_i \leq T_{Ri}$.
- For $\delta_i = 0$, let $T_{Li} < T_{Ri} = \infty$.

Summary statistics based on 500 replications with the sample sizes 200, 400 and 800 data points are presented in Table 1, which show the values of bias, Standard Error (SE), and Root Mean-Square Error (RMSE).

The results of simulation studies suggest that the proposed estimation method has small biases, thus implying that estimates of cure probability and change-point using the suggested method are quite accurate under all the investigated settings. The biases in the estimates for all of the examined parameters decreased with increasing sample size for both the normally- and uniformly-distributed covariates. Increasing the sample size ensures that the...
sample characteristics get closer to the properties defined by the data-generating model/process, hence reducing bias. This observation demonstrates that the estimator of the parameters is statistically consistent. Given the consistency of the estimator and the increased accuracy with increasing sample size, both the SE and RMSE decreased with increasing sample sizes for all considered parameters.

The aforementioned analysis was repeated with K(.) as the cumulative distribution function of the standard normal random variable, \( K(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt \), and the results are presented in Table 2. The results indicate that the estimated values are very similar for the two smoothing functions, and thus it is concluded that accuracies of the estimates are not sensitive to the type of the smoothing function employed.

### 7. Conclusion

Incorporation of the change point parameter into the cure rate model involves intriguing theoretical difficulties in both detection and estimation of such phenomena. In this article, we have proposed a change point mixture model.
cure model for interval-censored data. We obtained an estimate of the change-point under interval censoring using a modified objective function to eliminate the non-smoothness problem of the likelihood function. The estimation method is a combination of the maximum likelihood method and the EM algorithm. For various sample sizes, we implemented a simulation to generate samples with a cure fraction, and then under this setup we obtained the maximum likelihood estimators of the model. The values of the bias and MSE that were obtained from simulation studies show that the proposed estimation method performs well in the situations considered.

8. References