Australian Journal of Basic and Applied Sciences, 5(3): 546-554, 2011
ISSN 1991-8178

An Analytical Approach on Parametric Estimation of Cure Fraction Based on Weibull Distribution Using Interval Censored Data

[1]Bader Ahmad I. Aljawadi, [2]Mohd Rizam A. Bakar, and [3]Noor Akma Ibrahim

[1]Institute for Mathematical Research, Universiti Putra Malaysia, Malaysia
[2]Department of Mathematics, Universiti Putra Malaysia, Malaysia
[3]Institute for Mathematical Research, Universiti Putra Malaysia, Malaysia

Abstract: In this article, we consider the Bounded Cumulative Hazard (BCH) model that is more appropriate than mixture cure model in case of cancer clinical trials when the population of interest contains long-term survivors or *cured*. We propose this cure rate model based on the Weibull distribution with interval censored data. Maximum likelihood estimation (MLE) method is proposed to estimate the parameters within the framework of expectation-maximization (EM) algorithm, Newton Raphson method also employed. The analysis showed that the cure fraction cannot be obtained analytically, but may be obtained from the numerical solution of the estimated equations. A simulation study is also provided for assessing the efficiency of the proposed estimation procedure.

Key words: BCH model, Cure fraction, Interval censoring, MLE method, EM algorithm.

INTRODUCTION

Survival models that incorporate a cure fraction are called cure rate models. These models are being widely used in analyzing survival data from clinical trials; especially cancer clinical trials where a significant proportion of patients are cured.

The simplest cure rate model was constructed by Boag in 1949 and later developed by Berkson and Gage in 1952 which is called *mixture cure rate model*. In this model, the survival at any given time is equal to the proportion of those that are cured (π) plus those that are not cured (1-π), but who have not yet died or, in the case of diseases that feature asymptomatic remissions, have not yet re-developed signs and symptoms of the disease.

The mixture cure model equation is: $S_p(t) = \pi + (1-\pi)S_u(t)$.

Where $S_p(t)$ and $S_u(t)$ are survival functions for the entire population and the uncured patients respectively.

In this model, the survival function of the uncured patients $S_u(t)$ can be estimated parametrically or non-parametrically, which leads to parametric or semi-parametric survival models respectively. In parametric cure models we assume a particular distribution for the failure time distribution of uncured patients such as Exponential, Weibull, Gompertz, Generalized F and Log normal distributions. Because of its historical significance and important properties, the Weibull distribution is one of the most common parametric models which has a broad range of applications in survival analysis, Ahmed and Noor, (2010) and Cantor, (2003), therefore we will consider this distribution for the subsequent analysis.

The Weibull distribution has a survival function $S_p(t) = e^{-\lambda t\gamma}$ and the probability density function, $f(t) = \lambda\gamma(t)^{\gamma-1}e^{-\lambda t\gamma}$ for $t \geq 0$. Here $\lambda > 0$ is the scale parameter and $\gamma > 0$ is a shape parameter.

With rapid developments in medical and health sciences, an alternative model to mixture cure model was proposed to account the drawbacks of mixture model that have been discussed by Chen, *et al.,* in 1999, this alternative model developed by Yakovlev *et al.,* in 1993 and known as the *bounded cumulative hazard* (BCH) model which has the following particular features:

Corresponding Author: Bader Ahmad I. Aljawadi, Institute for Mathematical Research, Universiti Putra Malaysia, Malaysia

i) It is derived from a natural biological motivation.
ii) It has proportional hazard structure through the cure rate parameter.
iii) It is extremely computationally attractive.
iv) It has a mathematical relationship with the mixture cure rate model.

BCH Model:

In the BCH model which is also known as the promotion time cure model, it is assumed that for a cancer patient in the treated group after the initial treatment the number of cancer cells (often called clonogens) for that patient left active, these clonogens grow rapidly to produce a detectable cancer mass later on which may replace the normal tissue (cancer relapse), the number of clonogens N follows a Poisson distribution with mean θ, (Chen, M.H., 1999). Such that:

N~Pois (θ), where $f(N;\theta)=\dfrac{\theta^{N}e^{-\theta}}{N_{!}}$.

Let the variables Z_i ,i=1,2,..N, denote the i^{th} clonogen time to produce a detectable cancer mass, Z_i are assumed to be independent of N and have a cumulative distribution function (c.d.f.) $F(t) = P(Z\leq t)$, giving the probability of development a detectable tumor cancer mass by duration, Chen *et al.,* (1999). It will often be convenient to work with the complement of the c.d.f, the survival function S_p $(t) = P(Z>t) = 1\text{-}F(t)$ which gives the probability of being alive at duration t or more generally the probability that the event of interest has not occurred by duration, t Cooner, *et al.,* (2007). Note that $F(0) = 0$ and, $F(\infty) = 1$so that S_p $(0) = 1$ and S_p (∞) $= \pi$ the plateau value. The hazard function corresponding to this model is

$$h(t)=\frac{f(t)}{S_p(t)},$$

where $f(t)$ is the density function corresponding to $F(t)$.

There is biological evidence that the majority of recurrent tumors are arising from a single cancer cell, Tsodikov *et al.,* (2007). Thus, the time it takes the cancer to relapse can be defined by the random variable T = min $\{Z_i,\ 0\leq\ \text{i}\ \leq N\}$, where $P(Z_0 = \infty) = 1$ and Z_i is independent of N. Thus, the survival function for the entire population can be defined as follows:

$$\begin{aligned}
S_p\ (t) &= P\ [\text{There is no detectable cancer mass by the time } t]\\
&= P(N=0) + P(Z_1 > t,\ Z_2 > t\ldots,\ Z_N > t,\ N\geq 1)\\
&= \exp\ (\text{-}\theta) + [P(Z_1) > t)P\ (N = 1)] + [P\ (Z_1 > t)P\ (Z_2 > t)P\ (N = 2)]+\\
&\qquad [P(Z_1) > t)P\ (Z_2 > 1)P\ (Z_3 > t)P\ (N=3)] +\ldots+\\
&\qquad [P(Z_1) > t)P\ (Z_2 > t)\ \ldots\ P(Z_{\text{n}} > t)P\ (N = n)]\\
&= \exp\ (\text{-}\theta) + [S(t_1)\ P\ (N = 1)] + [S\ (t)^2\ P\ (N{=}2)]+[S(t)\ (N = 3)]+\ \ldots\\
&\qquad + [S(t)^n P(N{=}n)]
\end{aligned}$$

$$=\exp(-\theta)+\sum\nolimits_{n=1}^{N}[S(t)^{n}P(N=n)]$$

Since N ~ Pois (θ) then

$$=\exp(-\theta)+\sum\nolimits_{n=1}^{N}\frac{(S(t))^{n}\exp(-\theta)(\theta^{n})}{n!}$$

$$=\exp(-\theta)+\sum\nolimits_{n=1}^{N}\frac{(S(t))^{n}\exp(-\theta)}{n!}$$

$$=\exp(-\theta)+\exp(-\theta)\sum\nolimits_{n=1}^{N}\frac{(S(t)\theta)^{n}}{n!}$$

$$=\exp(-\theta)\left[1+\sum\nolimits_{n=1}^{N}\frac{(S(t)\theta)^{n}}{n!}\right]$$

$$
\begin{aligned}
&= \exp(-\theta)\left[\sum\nolimits_{n=1}^{N} \frac{(S(t)\theta)^{n}}{n!}\right] \\
&= \exp(-\theta)\exp(\theta S(t)) \\
&= \exp(-\theta F(t))
\end{aligned}
\tag{1}
$$

(*See* Aljawadi *et al.,* 2011)

Since $F(\infty) = 1$ and $S(\infty) = \exp(-\theta)$ model (1) is an improper survival function. However, the cure fraction based on BCH model can be defined as follows:

$$
\begin{aligned}
\pi &= \lim_{t\to\infty} S_p(t) = P(N=0) \\
&= \lim\nolimits_{t\to\infty} \exp(-\theta F(t)) \\
&= \exp(-\theta).
\end{aligned}
\tag{2}
$$

Note that when $\theta \to \infty$ then $\pi \to 0$, whereas as $\theta \to 0$ then $\pi \to 1$.

Parametric Estimation under Interval Censored Data.:

Interval censoring occurs when the event of interest is known only to occur within a given period of time, where the time t_i, $i = 1,...,n$, until the occurrence of the interested event for each individual is only known (whether it occurs) that it belongs to the interval between pre assigned visits points, *i.e.* between the visit at time L_i and the visit at time R_i, the survival time t is not exactly known, it is only known that the interested event has occurred within the interval $(L_i, R_i$. Both left and right censored data are special cases of interval censored data, where in left censored data the lower endpoint L is 0 and in right censored data the upper endpoint R is ∞, Davison, (2006).

In case of interval censored data $S(t_i) = S(L_i) - S(R_i)$ (Klein, J.P., 2003). Thus, for left censoring $S(t_i) = S(0)-S(R_i) = 1 - S(R_i)$, since $S(0) = 1$, and similarly for right censoring $S(t_i) = S(L_i) - S(\infty) = S(L_i)$ since $S(\infty) = 0$. Regarding uncensored individuals t_i is not observed, therefore, we will use the Mid-point estimation for t_i.

In parametric maximum likelihood estimation method the survival function S_p (.) and the probability density function f(.) for the entire population are known given that θ is unknown, let α_i be an indicator of censoring with zero if t_i is censored time and one otherwise, and c_i is an indicator of cure status of the i^{th} patient, namely c_i is zero if the patient is cured and one otherwise, i= 1,2,...,n . Obviously, if $\alpha_i = 1$, then $c_i = 1$, but if $\alpha_i = 0$ then is not observed and it can be one or zero. We assume throughout this analysis that the censoring is independent of failure times.

Given α_i and c_i, i.e. the complete data are available, then the log likelihood function can be written as follows:

$$
l_c = \log \Pi_{i=1}^{n} [\{f(t_i).(1-\pi)\}^{c_i}]^{\alpha_i}.[\{\pi\}^{1-c_i}.\{(1-\pi).S_u(t_i)\}^{c_i}]^{1-\alpha_i}
$$

Based on the survival function and the probability density function corresponding to Weibull distribution, the log-likelihood function can be rewritten as follows:

$$
l_c = \log \Pi_{i=1}^{n} [\{(\lambda\gamma(t_i)^{\gamma-1} e^{-\lambda t_i^{\gamma}}).(1-e^{-\theta})\}^{c_i}]^{\alpha_i} \times [\{e^{-\theta}\}^{1-c_i}.\{(1-e^{-\theta}).(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}})\}^{c_i}]^{1-\alpha_i}
$$

This can be simplified and expressed by:

$$
\begin{aligned}
l_c &= \log(\lambda\gamma)\sum\nolimits_{i=1}^{n} \alpha_i c_i + (\gamma-1)\sum\nolimits_{i=1}^{n} \alpha_i c_i + \log(t_i) - \sum\nolimits_{i=1}^{n} (t_i^{\gamma})\alpha_i c_i - \\
&\theta\sum\nolimits_{i=1}^{n} (1-\alpha_i)(1-c_i) + \log(1-e^{-\theta})\sum\nolimits_{i=1}^{n} c_i + \sum\nolimits_{i=1}^{n} (1-\alpha_i)c_i \log(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}}).
\end{aligned}
\tag{3}
$$

The solutions of $\frac{\partial l_c}{\partial \theta} = 0, \frac{\partial l_c}{\partial \lambda} = 0$ and $\frac{\partial l_c}{\partial \gamma} = 0$ are the desired estimates of θ, λ and γ. Where

$$\frac{\partial l_c}{\partial \theta} = \frac{(e^{-\theta})(\sum_{i=1}^{n} c_i)}{1-e^{-\theta}} - \sum_{i=1}^{n}(1-\alpha_i) = 0,$$

or

$$\theta = \log\left[\frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n}(1-\alpha_i)(1-c_i)} + 1\right]. \tag{4}$$

$$\begin{aligned}\frac{\partial l_c}{\partial \lambda} &= \frac{1}{\lambda}\sum_{i=1}^{n}(t_i^{\gamma})\alpha_i c_i - \sum_{i=1}^{n}(t_i^{\gamma})\alpha_i c_i + \sum_{i=1}^{n} c_i(1-\alpha_i)\left(\frac{R_t^{\gamma} e^{-\lambda R_i^{\gamma}} - L_t^{\gamma} e^{-\lambda L_i^{\gamma}}}{e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}}}\right) = 0\\ \frac{\partial l_c}{\partial \gamma} &= \frac{1}{\gamma}\sum_{i=1}^{n}\alpha_i c_i + \sum_{i=1}^{n}\alpha_i c_i \log(t_i) - \lambda\gamma\sum_{i=1}^{n}(t_i^{\gamma-1})\alpha_i c_i + \sum_{i=1}^{n} c_i(1-\alpha_i)\left(\frac{(\lambda\gamma R_t^{\gamma-1})e^{-\lambda R_i^{\gamma}} - (\lambda\gamma L_t^{\gamma-1} e^{-\lambda L_i^{\gamma}}}{e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}}}\right) = 0\end{aligned} \tag{5}$$

$$= \frac{1}{\gamma}\sum_{i=1}^{n}\alpha_i c_i + \sum_{i=1}^{n}\alpha_i c_i \log(t_i) - \lambda\gamma\left[\sum_{i=1}^{n}(t_i^{\gamma-1})\alpha_i c_i - \sum_{i=1}^{n} c_i(1-\alpha_i)\left(\frac{(R_i^{\gamma-1})e^{-\lambda R_i^{\gamma}} - (L_i^{\gamma-1} e^{-\lambda L_i^{\gamma}}}{e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}}}\right)\right] = 0. \tag{6}$$

Equations (5) and (6) need to be solved numerically to find λ and γ.

However, since the cure status is not fully observed then we need to implement the expectation maximization (**EM)** algorithm to estimate the desired parameters.

Before the implementation of the EM algorithm let us define g_i as the expected value of c_i for the i^{th} patient to be uncured conditional in the current estimates of α_i and the survival function of uncured patients $S_u(t_i)$, this definition proposed by [10] where:

$$g_i = \alpha_i + (1-\alpha_i).\left[\frac{[1-e^{-\theta}].S_u(t_i)}{[e^{-\theta}]+[1-e^{-\theta}].S_u(t_i)}\right]. \tag{7}$$

For censored individuals, *i.e.* $\alpha_i = 0$, then

$$g_i = \left[\frac{[1-e^{-\theta}].S_u(t_i)}{[e^{-\theta}]+[1-e^{-\theta}].S_u(t_i)}\right] = \left[\frac{[1-e^{-\theta}]\left(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}}\right)}{(e^{-\theta}) + (1-e^{-\theta})\left(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}}\right)}\right].$$

For simplicity we can define p_i as the probability of cured, such that

$$\begin{aligned} p_i &= 1 - \left[\frac{[1-e^{-\theta}]\left(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}}\right)}{(e^{-\theta}) + (1-e^{-\theta})\left(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}}\right)}\right]\\ &= \left[\frac{e^{-\theta}}{(e^{-\theta}) + (1-e^{-\theta})\left(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}}\right)}\right] = \left[\frac{1}{1 + (e^{-\theta}-1)\left(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}}\right)}\right]\end{aligned} \tag{8}$$

EM Algorithm:

The EM algorithm composed of two steps; the expectation step (E-Step) followed by the maximization step (M-step), where the E-step calculates the expectation of the log likelihood function defined in equation

(3) for the given values of α_i , c_i and $[L_i , R_i]$. Suppose that we have individuals where for i=1,...,m then α_i and c_i are observed and both are equal to 1, while for i=m+1,...n, α_i is observed and equal to 0 but c_i is not observed and need to be estimated, so the expected value of the log likelihood function can be written as follows:

$$E(l_c/\alpha, c_i, t_i) = E_1(l_c/\alpha_i = 1, c = 1, [L_i, R_i], 1 \le i \le m) + E_2(l_c/\alpha = 0, [L_i, R_i], m+1 \le i \le n),$$

where

$$E_1(l_c) = m\log(\lambda\gamma) + (\gamma-1)\sum\nolimits_{i=1}^{m}\log(t_i) + m\log(1-e^{-\theta}) - \lambda\sum\nolimits_{i=1}^{m} t_i^{\gamma}$$

and

$$E_2(l_c) = m\log(\lambda\gamma) + (\gamma-1)\sum\nolimits_{i=1}^{m}\log(t_i) + m\log(1-e^{-\theta}) - \lambda\sum\nolimits_{i=1}^{m} t_i^{\gamma}$$

It is clear that the expected value of the log likelihood function cannot be calculated unless the expressions $\sum\nolimits_{i=m+1}^{n}(1-c_i), \sum\nolimits_{i=m+1}^{n} c_i$ and $\sum\nolimits_{i=m+1}^{n} c_i \log(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}})$ could be evaluated, since the cure status c_i for the (n-(m+1)) censored individuals is not provided. Thus, the three expressions are called the sufficient statistics, where it is necessary to find the expected value for these sufficient statistics.

It follows that the log-likelihood function is linear in the complete data sufficient statistics, and then the E-step requires the computation of, $E\left(\sum\nolimits_{i=m+1}^{n}(1-c_i),\right)$, $E\left(\sum\nolimits_{i=m+1}^{n} c_i\right)$ and $E\left(\sum\nolimits_{i=m+1}^{n} c_i \log(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}})\right)$.

Let

$$S_1 = E\left(\sum\nolimits_{i=m+1}^{n}(1-c_i),\right) = \sum\nolimits_{i=m+1}^{n}\left[\frac{1}{1+(e^{\theta}-1)(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}})}\right],$$

$$S_2 = E\left(\sum\nolimits_{i=m+1}^{n} c_i\right) = \sum\nolimits_{i=m+1}^{n}\left(1-\left[\frac{1}{1+(e^{\theta}-1)(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}})}\right]\right),$$

and

$$S_3 = E\left(\sum\nolimits_{i=m+1}^{n} c_i \log(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}})\right).$$

$$= \sum\nolimits_{i=m+1}^{n}\left(1-\left[\frac{1}{1+(e^{\theta}-1)(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}})}\right]\right)\log(e^{\theta}-1)(e^{-\lambda L_i^{\gamma}} - e^{-\lambda R_i^{\gamma}}).$$

For the M-Step we can use the complete data maximum likelihood estimates given by equations (4), (5) and (6). Such that the maximum likelihood estimate of θ can be obtained by

$$\begin{aligned} \theta^{t+1} &= \log\left[\frac{\sum\nolimits_{i=1}^{n} c_i}{\sum\nolimits_{i=1}^{n}(1-\alpha_i)(1-c_i)}\right] = \log\left[\frac{\sum\nolimits_{i=1}^{n} c_i + \sum\nolimits_{i=m+1}^{n} c_i}{\sum\nolimits_{i=1}^{m}(1-\alpha_i)(1-c_i) + \sum\nolimits_{i=m+1}^{m}(1-\alpha_i)(1-c_i)} + 1\right] \\ \theta^{t+1} &= \log\left[\frac{m+S_2}{S_1} + 1\right]. \end{aligned} \tag{9}$$

While for equations (5) and (6) we can't find explicit solutions with respect to λ and γ. Therefore, the maximum likelihood estimates of these parameters λ^{t+1} and γ^{t+1} could be solved using any appropriate numerical method such as Newton Raphson Method. Where these equation could be simplified as follows:

$$\begin{aligned}\frac{\partial l_c}{\partial\lambda} &= \frac{1}{\lambda}\left[\sum\nolimits_{i=1}^{m}\alpha_i c_i+\sum\nolimits_{i=m+1}^{n}\alpha_i c_i\right]-\left[\sum\nolimits_{i=1}^{m}(t_i^{\gamma})\alpha_i c_i+\sum\nolimits_{i=m+1}^{n}(t_i^{\gamma})\alpha_i c_i\right]+\\ &\left[\sum\nolimits_{i=1}^{m}c_i(1-\alpha_i)\left(\frac{R_i^{\gamma}e^{-\lambda R_i^{\gamma}}-L_i^{\gamma}e^{-\lambda L_i^{\gamma}}}{e^{-\lambda L_i^{\gamma}}-e^{-\lambda R_i^{\gamma}}}\right)+\sum\nolimits_{i=m+1}^{n}c_i(1-\alpha_i)\left(\frac{R_i^{\gamma}e^{-\lambda R_i^{\gamma}}-L_i^{\gamma}e^{-\lambda L_i^{\gamma}}}{e^{-\lambda L_i^{\gamma}}-e^{-\lambda R_i^{\gamma}}}\right)\right]=0\\ &=\frac{m}{\lambda}-\left[\sum\nolimits_{i=1}^{m}(t_i^{\gamma})\sum\nolimits_{i=m+1}^{n}(t_i^{\gamma})\alpha_i c_i\right]+\left[\sum\nolimits_{i=m+1}^{n}c_i\left(\frac{R_i^{\gamma}e^{-\lambda R_i^{\gamma}}-L_i^{\gamma}e^{-\lambda L_i^{\gamma}}}{e^{-\lambda L_i^{\gamma}}-e^{-\lambda R_i^{\gamma}}}\right)\right]=0\end{aligned}\tag{10}$$

$$\begin{aligned}\frac{\partial l_c}{\partial\lambda} &= \frac{1}{\lambda}\left[\sum\nolimits_{i=1}^{m}\alpha_i c_i+\sum\nolimits_{i=m+1}^{n}\alpha_i c_i\right]+\left[\sum\nolimits_{i=1}^{m}\alpha_i c_i\log(t_i^{\gamma})+\sum\nolimits_{i=m+1}^{n}\alpha_i c_i\log(t_i^{\gamma})\right]-\\ &\lambda\gamma\left[[\sum\nolimits_{i=1}^{m}(t_i^{\gamma-1})\alpha_i c_i+\sum\nolimits_{i=m+1}^{n}(t_i^{\gamma-1})\alpha_i c_i]-\left[\sum\nolimits_{i=1}^{m}c_i\left(\frac{(R_i^{\gamma-1})(L_i^{\gamma-1})e^{-\lambda(R_i^{\gamma}-L_i^{\gamma})}}{e^{-\lambda(R_i^{\gamma}-L_i^{\gamma})}-1}\right)+\right.\right.\\ &\left.\left.\sum\nolimits_{i=m+1}^{n}c_i(1-\alpha_i)\left(\frac{(R_i^{\gamma-1})(L_i^{\gamma-1})e^{-\lambda(R_i^{\gamma}-L_i^{\gamma})}}{e^{-\lambda(R_i^{\gamma}-L_i^{\gamma})}-1}\right)\right]\right]=0\\ &=\frac{m}{\gamma}+\sum\nolimits_{i=1}^{m}\log(t_i)-\lambda\gamma\left[\sum\nolimits_{i=1}^{m}(t_i^{\gamma-1})-\sum\nolimits_{i=m+1}^{n}c_i\left(\frac{(R_i^{\gamma-1})(L_i^{\gamma-1})e^{-\lambda(R_i^{\gamma}-L_i^{\gamma})}}{e^{-\lambda(R_i^{\gamma}-L_i^{\gamma})}-1}\right)\right]=0.\end{aligned}\tag{11}$$

As a result, the E-step composed of the evaluation of the sufficient statistics defined above starting with some appropriate pre-assigned initial values ($\theta^{\circ},\lambda^{\circ},\gamma^{\circ}$) then, the M-step composed of the substitution of the sufficient statistics in equation (9) to get a new value for the parameter θ and based on the same initial values we can solve equations (10) and (11) using Newton Raphson method to get the new values of the parameters λ and γ. Using the new values of these parameters and repeat until a stopping condition such as:

$\theta^{t+1}-\theta^{t}\le\varepsilon,\varepsilon$, is a small positive value such as ε=0.0001.

Then $\theta^{t+1},\lambda^{t+1}$ and γ^{t+1} are the maximum likelihood estimates of the interested parameters.

Simulation and Results:

In this simulation study, the Weibull distribution with various values of the scale and shape parameters are considered for the data generation, where varying the parameters will implies various censoring rates (*P*). Each data set contained 100 interval censored observations of which different censoring rate depends on the value of λ. Here we ignored the left censoring case. To control the generation process we assumed that the true survival time *t* follows Weibull distribution; the steps used for data generation are as follows:

a) Generate from Weibull distribution with different scale parameter values to control the censoring rate.
b) Generate a vector *V* for the clinic visits, assuming that there are 20 clinic visits, in case of Weibull distribution the first visit v_1 was generated from *U*(0,0.1). Then the next visit v_2 was generated from *U*(v_1,v_1+0.1). The other visit times were generated in the same manner.
c) Generate a 100×2 empty matrix named "bound" for each data set. The entries of bound matrix are the intervals endpoints for each individual after comparing the true survival time with the 20 visit times. In case of right censoring the right end point can be assigned to be a large number beyond the last visit time. The formula used for end points determination is:

For i=1,...,100, j=1,...,20

$$bound[i,1] = \begin{cases} 0 & : if \;\; t[i] < V[1] \\ V[j] & : if \;\; V[j] < t[i] < V[j+1] \\ V[20] & : if \;\; t[i] > V[20] \end{cases}$$

$$bound[i,2] = \begin{cases} V[1] & : if \;\; t[i] < V[1] \\ V[j+1] & : if \;\; V[j] < t[i] < V[j+1] \\ Inf & : if \;\; t[i] > V[20] \end{cases}$$

d) Generate a 100×2 empty matrix named “status”. Based on the bound matrix let:

Status [i,1] ≡ censoring indicator $\alpha_i = \begin{cases} 0 : if \;\; bound[i,2] = 100 \\ 1 : \;\; otherwise \end{cases}$

Status [i,1] ≡ cured indicator $c_i = \begin{cases} 0 : if \; \alpha_i = 0 \\ 1 : \;\; otherwise \end{cases}$

In this simulation we are interested in the *bias* and *mean square error* (*MSE*), where bias can be defined as follows:

$$bias = \pi - E(\hat{\pi}), \tag{17}$$

where $\hat{\pi}$ is the maximum likelihood estimate for π.

A Smaller bias indicates that the parameter is closer to the true value on average and hence more accurate. Although bias can be used to measure the accuracy of an estimator, the mean square error (MSE) provides a better assessment of the quality of parameters. This is evident in a simulation study where the true parameter values are assumed known at the outset. The MSE of an estimator is known as the expected squared deviation of the estimated parameter value from the true parameter value, and by using a standard notation for a scalar parameter it can be decomposed into the following form:

$$MSE_{\pi} = variance(\hat{\pi}) + bias^2 \; . \tag{18}$$

The simulation was carried out with the built-in random generators in “R”. The Results are shown in tables 4.1, 4.2 and 4.3.

Table 4.1: Censoring Rates, Real Cure Rates, Expected Cure Rates, Bias and the *MSE* for *P*∈[20%, 30%].

Run	*Censoring Rate (%)*	*Real Cure (%)*	*Expected Cure (%)*	*Bias (%)*	*MSE × 1000*
1	21	21	19	2	0.615
2	23	21	19	2	0.615
3	23	21	18	3	1.115
4	24	21	19	2	0.615
5	24	22	20	2	0.615
6	25	22	19	3	1.115
7	25	23	21	2	0.615
8	26	23	20	3	1.115
9	26	23	21	2	0.615
10	26	23	20	3	1.115
11	27	24	21	3	1.115
12	27	24	20	4	1.815
13	27	24	22	2	0.615
14	28	24	20	4	1.815
15	28	25	22	3	1.115
16	29	25	21	4	1.815
17	29	25	22	3	1.115
18	30	26	22	4	1.815
19	30	26	22	4	1.815
20	30	27	24	3	1.115

Table 4.2: Censoring Rates, Real Cure Rates, Expected Cure Rates, Bias and the *MSE* for $P\in[30\%, 40\%]$.

Run	*Censoring Rate (%)*	*Real Cure (%)*	*Expected Cure (%)*	*Bias (%)*	*MSE × 1000*
1	31	30	26	4	1.857
2	31	31	26	5	2.757
3	32	31	26	5	2.757
4	32	31	27	4	1.857
5	32	32	27	5	2.757
6	33	32	27	5	2.757
7	33	33	29	4	1.857
8	34	33	28	5	2.757
9	34	33	29	4	1.857
10	35	34	29	5	2.757
11	35	34	29	5	2.757
12	36	34	29	5	2.757
13	38	35	29	6	3.857
14	38	35	30	5	2.757
15	38	35	30	5	2.757
16	39	36	30	6	3.857
17	39	36	30	6	3.857
18	40	36	29	7	5.157
19	40	37	31	6	3.857
20	40	37	31	6	3.857

Table 4.3: Censoring Rates, Real Cure Rates, Expected Cure Rates, Bias and the *MSE* for $P\in[40\%, 50\%]$.

Run	*Censoring Rate (%)*	*Real Cure (%)*	*Expected Cure (%)*	*Bias (%)*	*MSE × 1000*
1	41	40	34	6	3.738
2	41	40	33	7	5.038
3	42	40	33	7	5.038
4	42	41	34	7	5.038
5	44	41	33	8	6.538
6	44	42	35	7	5.038
7	44	42	35	7	5.038
8	45	42	34	8	6.538
9	45	43	35	8	6.538
10	46	43	34	9	8.238
11	46	43	35	8	6.538
12	46	44	35	9	8.238
13	47	44	34	10	10.138
14	47	44	35	9	8.238
15	48	45	35	10	10.138
16	48	45	36	9	8.238
17	48	46	35	11	12.238
18	49	46	35	11	12.238
19	49	47	36	11	12.238
20	50	50	38	12	14.538

The bias and mean square error values for the corresponding censoring rates indicate that the proposed method of cure rate estimation is more efficient when censoring rate decreases, and the estimation start to diverge in case of heavy censoring occurs. This result can be detected explicitly from figure 4.1 below which is constructed by allocating the censoring rates versus bias values from the results shown in the above tables ignoring the repeated censoring rates and considering the average of the corresponding bias values.

Conclusion:

We have investigated the parametric maximum likelihood estimation approach to estimate the cure fraction based on bounded cumulative hazard model under interval censored data. We have considered also the Weibull distribution to represent the distributional function of the uncured patients. The estimation method is a combination of the straight forward maximum likelihood estimation via the EM algorithm. As a result, the estimating equations are solved numerically since no explicit solutions could be found.

However, based on results obtained from the simulation study, we were able to conclude from observing bias and mean square error values that the parametric estimation of cure fraction under the assigned specifications and methods provides inefficient estimates of this fraction when the censoring rate is rapidly increasing, and this parametric estimation seems to provide consistent and better estimates of the cure fraction when the censoring rate is a bit low. Thus, when much censored observations are founded in the data set, the proposed estimation procedure is often producing biased estimators. This result was to be expected because the parametric assumption in case of high censoring rate is not valid, since the censored observations 'shrink' the parametric assumption toward the survival data set, and it is wrongfully to consider the parametric method when the censoring rate is high.

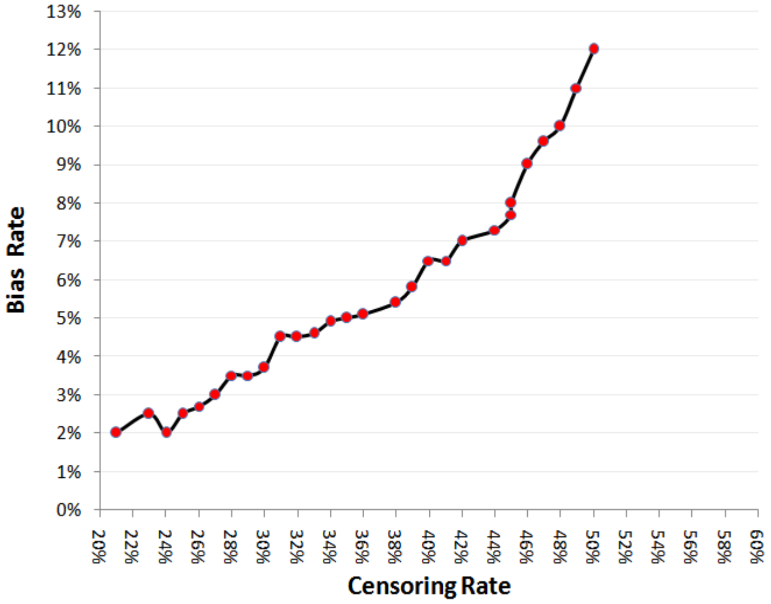

Fig. 4.1: Censoring rate versus bias for some generated samples based on Weibull Distributions.

REFERENCES

Ahmed, A.O.M. and A.I. Noor, 2010. *Bayesian survival estimator for Weibull distribution with censored data*. J. Applied Sci., 11: 393-396.

Aljawadi, B.A.I., M.R. Abu Bakar and N.A. Ibrahim, 2011. *Non-parametric Maximum Likelihood Estimation of Cure Fraction for Interval Survival Data*. International Journal of Applied Mathematics and Statistics, in press.

Berkson, J. and R.P. Gage, 1952. *Survival curves for cancer patients following treatment*. Journal of the American Statistical Association, 47: 501-515.

Boag, J.W., 1949. *Maximum likelihood estimates of the proportion of patients cured by cancer therapy*. Journal of the Royal Statistical Society, Series B, 11: 15-44.

Cantor, A.B., 2003. *SAS Survival Analysis Techniques for Medical Research*. Cary, NC: SAS Publishing.

Chen, M.H., J.G. Ibrahim and D. Sinha, 1999. *A new Bayesian model for survival data with a surviving fraction*. Journal of the American Statistical Association, 94: 909-919.

Cooner, F., S. Banerjee, B.P. Carlin and D. Sinha, 2007. *Flexible cure rate modeling under latent activation schemes*. Journal of the American Statistical Association, 102: 560-572.

Davison, A.C., 2006. *Survival and Censored Data*: Ecole Polytechnique Federal De Lausanne. Semester Project, pp: 1-44.

Klein, J.P. and M.L. Moeschberger, 2003. Survival Analysis Techniques for Censored and Truncated Data, (ed.). New York, USA: Springer.

Peng, Y., 2003. Fitting semi-parametric cure models. Computational Statistics and Data Analysis, 41: 481-490.

Tsodikov, A.D., J.G. Ibrahim and A.Y. Yakovlev, 2003. *Estimating cure rates from survival data*. Journal of the American Statistical Association, 98: 1063-1078.

Yakovlev, A.Y., B. Asselain, V.J. Bardou, A. Fourquet, T. Hoang, A. Rochefediere and A.D. Tsodikov, 1993. *A Simple Stochastic Model of Tumor Recurrence and Its Applications to Data on pre-menopausal Breast Cancer*. In Biometrie et Analyse de Dormees Spatio – Temporelles 12 (Eds. B. Asselain, M. Boniface, C. Duby, C. Lopez, J.P.Masson, and J.Tranchefort). Société Francaise de Biométrie, ENSA Renned, France, pp: 66-82.