Statistics and Probability Letters 131 (2017) 51–55

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On the likelihood of mixture cure models

Antai Wang [a,*], Yilong Zhang [b], Yongzhao Shao [b]

[a] Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ, USA
[b] Division of Biostatistics, New York University School of Medicine, New York, NY, USA

ARTICLE INFO

Article history:
Received 23 February 2016
Received in revised form 9 August 2017
Accepted 10 August 2017
Available online 24 August 2017

Keywords:
Complete-data likelihood
EM algorithm
Mixture cure model
Survival transformation models

ABSTRACT

The EM algorithm has been used for inference of the mixture cure models. However, the complete-data and incomplete-data specifications have never been postulated appropriately in literature. The goal of this paper is to fill in this gap by deriving proper specifications.

1. Introduction

Survival data with a latent cure fraction can be naturally modelled using mixture cure models (Berkson and Gage, 1952; Farewell, 1982; Kuk and Chen, 1992; Peng and Dear, 2000; Sy and Taylor, 2000; Lu and Ying, 2004; Othus et al., 2012). We first give notation. Let T^* denote the survival time and Y the indicator of the latent cure status. Let x be an observed baseline covariate vector with effects on the survival of uncured subjects ($Y = 1$) and z a covariate vector with effects on cure probability and may share common components with x. More precisely, given covariates (z, x), without loss of generality, we assume $p(Y = 1|z, x)$ only depends on z and denoted as $\pi(z)$; and $p(T^* > t|Y = 1, z, x)$ only depends on x and denoted as $S(t|Y = 1, x)$. Let $h(t|Y = 1, x) = -\frac{\partial}{\partial t} \log S(t|Y = 1, x)$ be the hazard rate. Then the marginal survival function $S(t|z, x) = P(T^* > t|z, x)$ of an individual is

$$S(t|z, x) = 1 - \pi(z) + S(t|Y = 1, x)\pi(z). \tag{1}$$

In the literature, $\pi(z) = p(Y = 1|z)$ is often modelled using the logistic model $\pi(z) = \pi_\gamma(z) = [1 + \exp(-z'\gamma)]^{-1}$ with γ being a vector of coefficients of z (Farewell, 1982). The probit model and other binary generalized linear models (GLM) can also be used. The effects of x on the survival of the uncured group $S(t|Y = 1, x)$ have also been modelled in many ways. For examples, Farewell (1982) assumed $S(t|Y = 1, x) = S_\beta(t|x)$ to be the parametric Weibull survival function with a parameter vector β. Kuk and Chen (1992) generalized the Weibull model $S_\beta(t|x)$ to the semiparametric Cox PH model (Cox, 1972)

$$S_\beta(t|x) = [S_0(t)]^{\exp(x'\beta)}, \quad h_\beta(t|x) = h_0(t)\exp(x'\beta),$$

where $S_0(t)$ and $h_0(t)$ are the baseline survival and hazard functions, respectively. In addition, the accelerated failure time (AFT) model (Buckley and James, 1979; Jin et al., 2006), the accelerated hazard models (Chen and Wang, 2000), and various transformation models including the proportional odds model (Lu and Ying, 2004) have been used to model the failure times of the uncured subjects. Thus, the mixture cure models include a large class of useful models.

[*] Corresponding author.
E-mail address: aw224@njit.edu (A. Wang).

http://dx.doi.org/10.1016/j.spl.2017.08.006
0167-7152/© 2017 Elsevier B.V. All rights reserved.

The EM algorithm has played a significant role in effective inference of mixture cure models. In particular, Sy and Taylor (2000) and Peng and Dear (2000) have successfully developed EM algorithms for the semiparametric Cox PH cure models. EM algorithms for the AFT cure models and other mixture cure models have been developed by Zhang and Peng (2007) and others.

As explained in Dempster et al. (1977), proper complete-data specifications are crucial in the formal development of the EM algorithm to maximize the incomplete-data specification. However, neither the complete-data specification nor the incomplete-data specification has been rigorously derived for mixture cure models. To fill in this gap, we derive both complete-data and incomplete-data specifications for mixture cure models. The incorrect likelihood function in the literature does lead to correct estimates because the missing term $(1-\delta)^{1-y}$ (see Section 2) is not dependent upon the parameters, but using our correct derivation obviates the needs for any corrections to the incorrect likelihood function as have been previously proposed in the literature.

Our paper is organized in the following way, we derive the correct complete likelihood in Section 2 and then we present the incorrect complete likelihood function widely used in the literature and point out the differences between two likelihood functions in Section 3. We end our paper with some discussion in Section 4.

2. Proper specifications of mixture cure models

Let (T_i^*, C_i, Y_i) be the survival time, censoring time, and uncure indicator of the ith subject with $i = 1, \ldots, n$. Let z_i, x_i be the observed covariate values for the ith subject. We use (t_i, δ_i) to denote the observed survival time (possibly censored) and censoring indicator (T_i, Δ_i), i.e.

$$T_i = \min(T_i^*, C_i) \text{ and } \Delta_i = I(T_i^* < C_i).$$

The subscript i is generally omitted in deriving the complete-data specification for a generic subject. Denote $\Theta = (\gamma, \beta, S_0(\cdot))$ where $S_0(\cdot)$ is the baseline survival function for uncured subjects. We assume $\pi_\gamma(z) = p(Y = 1|z, x, \Theta)$ is modelled using a binary GLM. Let $h_\beta(t|x)$ be the hazard rate for $S_\beta(t|x) = S(t|Y = 1, z, x, \Theta)$ with a baseline survival function $S_0(\cdot)$. Similarly, for the random censoring time C, we will use $S_c(t|z, x)$ and $h_c(t|z, x)$ to denote its unknown survival and hazard functions that do not involve the parameter Θ. We assume that, conditional on covariates, C is independent of T^* among uncured subjects. Then the conditional distribution of (t, δ) given $Y = 1$ can be written as

$$p(t, \delta|Y = 1, z, x, \Theta) = [h_\beta^\delta(t|x)S_\beta(t|x)][h_c^{1-\delta}(t|z, x)S_c(t|z, x)]. \tag{2}$$

By the fact that $p(t, \delta, Y = 1|z, x; \Theta) = p(t, \delta|Y = 1, z, x, \Theta)p(Y = 1|z, x, \Theta)$, we have

$$p(t, \delta, Y = 1|z, x, \Theta) = \pi_\gamma(z)[h_\beta(t|x)]^\delta S_\beta(t|x)[h_c(t|z, x)]^{1-\delta} S_c(t|z, x). \tag{3}$$

It is obvious that $Y = 0$ implies $\delta = 0$ so that

$$p(t, 0|Y = 0, z, x, \Theta) = h_c(t|z, x)S_c(t|z, x)$$

and

$$p(t, 1|Y = 0, z, x, \Theta) = 0.$$

Thus

$$p(t, \delta|Y = 0, z, x, \Theta) = (1 - \delta)[h_c(t|z, x)S_c(t|z, x)]^{1-\delta}.$$

Note that $p(Y = 0|z, x, \Theta) = 1 - \pi_\gamma(z)$. Similar to (3), we have

$$p(t, \delta, Y = 0|z, x, \Theta) = (1 - \pi_\gamma(z))(1 - \delta)[h_c(t|z, x)S_c(t|z, x)]^{1-\delta}. \tag{4}$$

From (3) and (4), and note that $\delta y = \delta$, given the observed data (t, δ, z, x) and assuming knowing values of Y, the complete-data specification $p(t, \delta, y|z, x, \Theta)$, i.e., the conditional density of (T_i, Δ_i, Y_i) given covariates, can be given as

$$p(t, \delta, y|z, x, \Theta) = f(t, \delta, y|z, x, \Theta)[h_c^{1-\delta}(t|z, x)S_c(t|z, x)], \tag{5}$$

where $f(t, \delta, y|z, x, \Theta)$ is

$$f(t, \delta, y|z, x, \Theta) = (1 - \delta)(1 - \pi_\gamma(z))(1 - y) + h_\beta^\delta(t|x)S_\beta(t|x)\pi_\gamma(z)y. \tag{6}$$

From equation (1.1) of Dempster et al. (1977), the incomplete-data specification is

$$p(t, \delta|z, x, \Theta) = p(t, \delta, y = 0|z, x, \Theta) + p(t, \delta, y = 1|z, x, \Theta).$$

From (3) and (4), the incomplete-data specification is

$$p(t, \delta|z, x, \Theta) = [(1 - \pi_\gamma(z))(1 - \delta) + \pi_\gamma(z)h_\beta^\delta(t|x)S_\beta(t|x)]h_c^{1-\delta}(t|z, x)S_c(t|z, x). \tag{7}$$