Computational Statistics and Data Analysis 54 (2010) 2103–2112

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Extending the long-term survivor mixture model with random effects for clustered survival data

Xin Lai [a,b], Kelvin K.W. Yau [a,*]

[a] Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

[b] Department of Statistics and Finance, University of Science and Technology of China, Hefei, China

ARTICLE INFO

Article history:
Received 15 October 2008
Received in revised form 3 February 2010
Accepted 23 March 2010
Available online 1 April 2010

Keywords:
Cured patients
EM algorithm
GLMM
Long-term survivors
Power transformation
Random effects
REML

ABSTRACT

To provide a class of hazard functions in analyzing survival data, the power family of transformations has been proposed in the literature. Our work in this paper considers the existence of cured patients and random effects due to clustering of survival data in a long-term survivor model setting. A power family of transformations is assumed for the relative risk in the hazard function component. Such an extension allows us to flexibly base the inferences on various hazard function assumptions, particularly taking exponential and linear relative risk as two special cases. The parameter governing the power transformation could be determined by means of a modified Akaike information criterion (AIC). Applications to two sets of survival data illustrate the use of the proposed long-term survivor mixture model. A simulation study is carried out to examine the performance of the estimators under the proposed numerical estimation scheme.

1. Introduction

Cured patients (or the so-called long-term survivors) have been increasingly observed in clinical trial studies with the advancement in medical treatment methods and operative techniques in recent years; see for example Maller and Zhou (1996). The traditional assumption that all individuals will eventually experience a failure event does not seem to be appropriate in these cases where the individuals are cured and thus free of the disease or risk. The mixture cure model (or long-term survivor model), combining the cured fraction and at-risk group, has been proposed by Farewell (1982) to analyze the survival data with cured patients. Particularly, a logistic transform is used to model the cured proportion, and the baseline hazard function for the at-risk group is assumed to follow the Weibull distribution. Kuk and Chen (1992) considered a semi-parametric approach to construct the likelihood function and estimate the parameters by a Monte Carlo approximation. Peng and Dear (2000) and Sy and Taylor (2000) modified Kuk and Chen's approach and the estimation of parameters was achieved by implementing an EM algorithm.

In some cases, failure time data obtained from clinical trials are conducted in several clusters. The individuals in the same hospital (or cluster) may share some common factors which originate from the unobservable characteristics of the hospital's operations or its environment. Ignoring this potential heterogeneity due to the presence of random hospital effects may result in biased estimates of the fixed effect parameters. In previous studies, the random effects were incorporated into the underlying model for accommodating the dependence structure within clusters. Yau and Ng (2001) adapted Farewell's model formulation and take the random hospital effects into account. The generalized linear mixed models (GLMM)

* Corresponding author. Tel.: +852 3442 8585; fax: +852 3442 0189.
E-mail address: mskyau@cityu.edu.hk (K.K.W. Yau).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.03.017

(McGilchrist, 1994) approach was adopted to obtain the approximate residual maximum likelihood (REML) estimates for the fixed effect and variance component parameters.

For proportional hazards regression model, the exponential relative risk function is commonly adopted. However, for some kinds of survival data in practice, simply adopting the exponential relative risk function without justification may be misleading. For instance, other function forms such as the linear relative risk model (Prentice and Mason, 1986) were found to be more meaningful. Guerrero and Johnson (1982) proposed the use of a general relative risk function. With the use of Box–Cox transformation, the exponential relative risk assumption was relaxed and could be considered as a special case instead. Following the same line, Yau and McGilchrist (1999), Zeng et al. (2005) and Yin and Ibrahim (2005) adopted the power family of transformations to provide a class of hazard functions in analyzing survival data. Influence diagnostics for survival models with power family of transformations in the hazard function can be found in Xiang et al. (2007).

Our work in this paper considers the long-term survivor model with random effects (Yau and Ng, 2001) but assuming the general relative risk function in the hazard function component of the model. Specifically, it extends Yau and Ng's (2001) model by incorporating a power family of transformation in the hazard function for the at-risk sub-population. Such extension allows us to flexibly base the inferences on various hazard function assumptions, particularly taking exponential and linear relative risk as two special cases. The parameter governing the power transformation could be determined by means of a modified Akaike information criterion (AIC).

In Section 2, the random effects cure model with a general relative risk function form is described. The estimation procedure within the GLMM framework is outlined in Section 3. For illustration, in Section 4, the proposed model is applied to analyze two sets of data. In Section 5, a simulation study is carried out to examine the performance of the approximate REML estimators for the proposed random effects cure model. Further discussions are presented in Section 6.

2. Random effects cure model with general relative risk function

Let t_{ij} be the observed survival time for individual j in clinic i, and x_{ij} is the corresponding vector of covariates, $j = 1, \ldots, n_i$, $i = 1, \ldots, M$, $\sum_{i=1}^{M} n_i = N$. Then the mixture cure model can be written as

$$S(t_{ij}; x_{ij}) = 1 - \pi(x_{ij}) + \pi(x_{ij}) S_u(t_{ij}; x_{ij}) \tag{1}$$

where $\pi(x_{ij})$ is the probability of belonging to the at-risk group and can be specified by a logistic form (Farewell, 1982)

$$\pi(x_{ij}) = \frac{\exp \xi_{ij}(x_{ij})}{1 + \exp \xi_{ij}(x_{ij})}, \qquad \xi_{ij}(x_{ij}) = \gamma_0 + x_{ij}'\gamma + V_i \tag{2}$$

V_i corresponds to the random effect of clinic i associated with the cured probability. For the at-risk (uncured) group, the survival function $S_u(t_{ij}; x_{ij})$ is defined via the hazard function which is assumed to follow a power family of functions

$$h(t_{ij}) = h_0(t_{ij}) \exp\{f(\eta_{ij})\}, \qquad f(\eta_{ij}) = \begin{cases} \ln(1 + k\eta_{ij})^{1/k}, & k \neq 0 \\ \eta_{ij} & k = 0 \end{cases} \tag{3}$$

where $h_0(t)$is the unspecified baseline hazard function and the linear predictor η_{ij} relates to both fixed and random effects: $\eta_{ij} = x_{ij}'\beta + U_i$. Model (3) becomes the Cox proportional hazards model $h(t_{ij}) = h_0(t_{ij}) \exp(\eta_{ij})$ when $k \to 0$, and it reduces to the linear relative risk function form $h(t_{ij}) = h_0(t_{ij})(1 + \eta_{ij})$if $k = 1$. Note that the transformation parameter k could mathematically take any value on the real line. However our primary interests will focus only on $k \in [0, 1]$ since this restriction represents the most commonly used relative risk functions in practice. As a remark, the term "relative risk" is not exactly suitable since the interpretation of β changes as the value of k changes. Nevertheless, for the sake of convenience and easier exposition, the term "relative risk" is used in the paper. Within the GLMM framework, the random effects of clinic i in the hazard function and the cured probability, denoted by U_i and V_i, are typically assumed to follow the Normal distribution $N(0, \theta_1)$ and $N(0, \theta_2)$ respectively.

We arrange the N failure/censoring times in increasing order and denote the k distinct event times by $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$. Following Kuk and Chen's (1992) argument, the conditional log-likelihood with random effects being fixed, l_1, can be rewritten as

$$l_1 = \log \prod_{i=1}^{N} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i} + \log \prod_{i=1}^{k} \frac{\exp f(\eta_{(i)})}{\sum_{h \in R(t_{(i)})} y_h \exp f(\eta_h)} \tag{4a}$$

where $\eta_{(i)}$ is the linear predictor corresponding to $t_{(i)}$. y is the latent indicator which equals to 0 if individual being cured or 1 otherwise. Define $u = (U_1, \ldots, U_M)'$ and $v = (V_1, \ldots, V_M)'$. The logarithm of the joint probability density function of the random effects is

$$l_2 = -\frac{1}{2}\left(M \log 2\pi\theta_1 + \frac{u'u}{\theta_1}\right) - \frac{1}{2}\left(M \log 2\pi\theta_2 + \frac{v'v}{\theta_2}\right). \tag{4b}$$