Journal of Multivariate Analysis 123 (2014) 129–142

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Marginal regression analysis of clustered failure time data with a cure fraction

Yi Niu [a,c], Yingwei Peng [a,b,c,*]

[a] Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada K7L 3N6
[b] Department of Public Health Sciences, Queen's University, Kingston, ON, Canada K7L 3N6
[c] Cancer Care and Epidemiology, Queen's Cancer Research Institute, Kingston, ON, Canada K7L 3N6

ARTICLE INFO

Article history:
Received 27 March 2013
Available online 17 September 2013

AMS 2000 subject classifications:
primary 62H30
62N02
secondary 62F12

Keywords:
Clustered survival time
EM algorithm
ES algorithm
Estimating equations
Marginal method
Mixture cure model

ABSTRACT

We consider a marginal mixture cure model for clustered survival data in which there may be a cure fraction and the survival times may be correlated. We propose a generalized estimating equation approach by incorporating working correlation matrices into an EM algorithm to estimate the regression coefficients and the baseline hazard function in the marginal model. The estimators of the regression parameters and the baseline hazard function are shown to be consistent and asymptotically normal, and their variances can be consistently estimated by a sandwich estimator. The proposed method is simple to use, and our simulation study shows the proposed method is more efficient than the existing marginal methods that ignore the correlation within clusters. An application of the proposed method to data from a smoking cessation study is demonstrated.

1. Introduction

Modeling failure time data with long-term survivors is an important issue in biomedical and clinical research. For example, in breast cancer [11] and prostate cancer [26] studies, a fraction of patients may respond favorably to a treatment and are often considered cured in the sense that they would never experience the event of interest even if they had sufficiently long follow-up. As a commonly used approach to modeling survival data with a cure proportion, the mixture cure model [2] has been extensively studied by many authors in the past few decades. See, for example, [11,14,25,24,19,17], among others. All these studies assumed independence among subjects.

In practical applications, correlation often arises in some cancer clinical studies due to clustering. For example, ages at diagnosis of breast cancer of female siblings and failure times of patients from the same hospital form clusters, and the ages or the failure times in each cluster may potentially be correlated because of the shared genetic environments and treatment resources. Ignoring correlation within clusters may incur a loss in estimation efficiency when the correlation is substantial. Therefore, it is important to take the correlation into account when analyzing clustered failure times.

Random effects models and marginal models are the two most common approaches to modeling clustered data with a cure fraction. Random effects models explicitly formulate the underlying dependence via random effects. Yau and Ng [28] generalized the proportional hazards mixture cure model by using two independent normal random effects to model

* Corresponding author at: Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada K7L 3N6.
E-mail addresses: yingwei.peng@queensu.ca, pengp@queensu.ca (Y. Peng).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.09.003

the correlation among cure statuses and the correlation among the failure times of uncured patients in a cluster. They proposed a best linear unbiased prediction (BLUP) method to estimate the parameters in the model. Lai and Yau [15] extended this method by allowing dependent random effects and a nonparametric baseline distribution in the model. Peng and Taylor [20] considered maximum likelihood estimation for the mixture cure model with random effects. Their method provides flexibility in specifying distribution for the random effects and is computationally intensive because of the numerical integration involved in the method. In the Bayesian paradigm, Chen et al. [6] extended the parametric bounded cumulative hazard cure model by using a positive stable frailty term to model the correlation within a cluster. Yin [30] further extended this work by considering a transformation cure model with a frailty effect for multivariate failure time data. Banerjee and Carlin [1] extended the parametric bounded cumulative hazard cure model to interval-censored and spatially correlated data.

Chatterjee and Shih [5] also generalized the univariate mixture cure models to bivariate survival data. They modeled the correlation among the cure statuses and the failure times of uncured subjects in a familial cluster in a breast cancer study using pairwise odds ratios and a copula model respectively, and proposed a quasi-likelihood method to estimate the parameters in the model. Wienke et al. [27] considered a full likelihood method with a similar model for bivariate data. Both methods do not consider covariate effects and the estimation methods become infeasible when cluster size is large.

Marginal models focus on the population average on the marginals of the joint distribution of data from one cluster, and the correlation is often treated as a nuisance parameter in the model to reduce the dependence of the marginal models on the specification of the unobservable correlation structure of clustered data. Peng et al. [21] proposed a semiparametric marginal proportional hazards mixture cure model to analyze survival data from a multi-institutional study of tonsil cancer and provided robust variance estimates of parameters. Yu and Peng [29] also considered a marginal mixture cure model with Weibull baseline distribution for a smoking cessation study and provided jackknife variance estimates of the parameters in the model. Chen and Lu [7] further extended the work of Peng et al. [21] by considering a transformation model for uncured patients and proved the asymptotic properties of the proposed estimators. All the models above are robust to misspecification of the correlation structure. However, when there is information available for the correlation structure, the estimates of the parameters in the above marginal methods may have substantial efficiency loss.

In this paper, we propose a new estimating equation approach for the marginal proportional hazards mixture cure model. Working correlation structures for the cure statuses and for the failure times of uncured subjects within a cluster are employed in the estimating equations. We investigate the efficiency improvement in the proposed estimation method compared to the existing methods for the marginal models when the correlations are present in the data.

The paper is organized as follows. In Section 2, we introduce the marginal proportional hazards mixture cure model for clustered survival data with a cure fraction. A set of estimating equations for the model is proposed, and the asymptotic properties of the estimators are investigated in this section. We conduct a simulation study to evaluate the finite sample performance of the proposed estimation method in Section 3, and illustrate this method by analyzing the smoking cessation data in Section 4. Conclusions and discussions are presented in Section 5.

2. Marginal model with correlation structures

Let $\tilde{T}_{ij}$ and C_{ij} be the failure and censoring times for the jth subject in the ith cluster ($i = 1, \ldots, K, j = 1, \ldots, n_i$), and $N = \sum_{i=1}^{K} n_i$. The observed failure time is $T_{ij} = \min(\tilde{T}_{ij}, C_{ij})$, its censoring status is denoted as $\delta_{ij} = I(\tilde{T}_{ij} \leq C_{ij})$ ($I(A) = 1$ if A is true and 0 otherwise), X_{ij} and Z_{ij} are two vectors of p_X and p_Z covariates (the two vectors may share some covariates) that may have effects on the failure time distribution of uncured subjects and the cure probability. We assume that C_{ij} is independent of $\tilde{T}_{ij}$ given X_{ij} and Z_{ij}. Let Y_{ij} denote the cure status of subject j in cluster i, that is, $Y_{ij} = 0$ if the subject is cured and 1 otherwise. It is obvious that if $\delta_{ij} = 1$, then $Y_{ij} = 1$. However, if $\delta_{ij} = 0$, the value of Y_{ij} is unobservable and Y_{ij} is a latent variable.

We assume that the marginal survival function of $\tilde{T}_{ij}$ is from the mixture cure model

$$S(t; X_{ij}, Z_{ij}) = P(\tilde{T}_{ij} > t; X_{ij}, Z_{ij}) = 1 - \pi(Z_{ij}) + \pi(Z_{ij}) S_u(t; X_{ij}), \tag{1}$$

where $\pi(Z_{ij}) = P(Y_{ij} = 1; Z_{ij})$ is in a logistic regression form

$$\pi(Z_{ij}) = \frac{\exp(\gamma' Z_{ij})}{1 + \exp(\gamma' Z_{ij})}, \tag{2}$$

and $S_u(t; X_{ij}) = P(\tilde{T}_{ij} > t | Y_{ij} = 1; X_{ij})$ is specified by the proportional hazards (PH) model

$$S_u(t; X_{ij}) = S_{u0}(t)^{\exp(\beta' X_{ij})}, \tag{3}$$

and $S_{u0}(t)$, the baseline survival function of $\tilde{T}_{ij} | \{Y_{ij} = 1\}$ when $X_{ij} = 0$, is assumed to follow the Weibull distribution with $S_{u0}(t) = \exp(-t^{\alpha})$. Note that we use the two-parameter Weibull distribution where the scale parameter is considered as an intercept term in the PH model. Here β and γ are $p_X + 1$ and $p_Z + 1$ unknown regression parameters for X_{ij} and Z_{ij} (both vectors are expanded to include 1 for the intercept terms) and α is an unknown parameter in the baseline distribution.