Computational Statistics and Data Analysis 56 (2012) 645–655

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Marginal analysis of multivariate failure time data with a surviving fraction based on semiparametric transformation cure models

Chyong-Mei Chen *, Tai-Fang C. Lu

Department of Statistics and Informatics Science, Providence University, Taiwan, ROC

ARTICLE INFO

Article history:
Received 8 February 2010
Received in revised form 4 August 2011
Accepted 12 September 2011
Available online 18 September 2011

Keywords:
Cure models
Transformation models
Estimating equations
Marginal approach
Stochastic process

ABSTRACT

In biomedical, genetic and social studies, there may exist a fraction of individuals not experiencing the event of interest such that the survival curves eventually level off to nonzero proportions. These people are referred to as "cured" or "nonsusceptible" individuals. Models that have been developed to address this issue are known as cured models. The mixture model, which consists of a model for the binary cure status and a survival model for the event times of the noncured individuals, is one of the widely used cure models. In this paper, we propose a class of semiparametric transformation cure models for multivariate survival data with a surviving fraction by fitting a logistic regression model to the cure status and a semiparametric transformation model to the event time of the noncured individual. Both models allow incorporating covariates and do not require any assumption of the association structure. The statistical inference is based on the marginal approach by constructing a system of estimating equations. The asymptotic properties of the proposed estimators are proved, and the performance of the estimation is demonstrated via simulations. In addition, the approach is illustrated by analyzing the smoking cessation data.

1. Introduction

Typically, models for survival analysis assume that all study subjects will eventually experience the defined event if the follow-up is sufficiently long, e.g. Cox's proportional hazards model (Cox, 1972). However, in biomedical, genetic and social studies, there may exist a fraction of individuals not experiencing the event such that the survival curves level off to nonzero proportions eventually. These are referred to as cured or nonsusceptible individuals. The terminology "cure" in clinical study corresponds to the term "nonsusceptibility" in biomedical or genetic studies.

Models for addressing this problem, called cure models, have attracted much attention in the past few decades. One of the widely used cure models is the mixture model (Farewell, 1982; Kuk and Chen, 1992; Peng and Dear, 2000; Sy and Taylor, 2000; Lu and Ying, 2004). The mixture model consists of a model for the binary cure status (e.g. the logistic regression model), and a survival distribution, referred to as the latency distribution, for the event times of the noncured individuals. Another class of cure models, bounded cumulative hazard models (Yakovlev, 1994; Tsodikov, 1998; Tsodikov et al., 2003), assumes that the survival time follows an improper survival function.

In the study of familial diseases, such as breast cancer, colon-rectal cancer or inherited diseases, there exists a significant proportion of cured or nonsusceptible patients. Furthermore, since the familial association exists within the family or cluster, the study data are best viewed as multivariate survival data. Also, it is better to extend the cure model to a multivariate survival model with a cure fraction. Yau and Ng (2001) and Peng and Taylor (2011) extended the mixture model by adding random effects to both the logistic regression model for the cure status and the latency distribution.

---

* Corresponding author. Tel.: +886 4 26328001; fax: +886 4 26324653.
E-mail address: cmchen2@pu.edu.tw (C.-M. Chen).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2011.09.013

Chatterjee and Shih (2001) constructed a quasi-likelihood function based on pairwise odds ratio with a copula model to estimate the incidence rate and the age-at-onset distribution for familial breast cancer. The pairwise odds ratio was adopted as an association measure between two individuals for the susceptibility to breast cancer, and the copula model was used for the dependency structure between the event times of two susceptible individuals. Wienke et al. (2003) considered a bivariate frailty model with cure for twin siblings. They extended the mixture model to breast cancer data by applying the correlated gamma-frailty model to bivariate lifetimes of two susceptible twin pairs. Chen et al. (2002) proposed an extension of the bounded cumulative hazard model by using a frailty term with a positive stable distribution to measure the correlation. Peng et al. (2007) analyzed multi-institutional data of the local recurrences of tonsil cancer patients based on the mixture cure model combining logistic regression with proportional hazards regression. They proposed a marginal approach for estimating the marginal, population-averaged measures of effect. Yu and Peng (2008) adopted the logistic-Weibull mixture regression as the marginal cure model and used the sandwich variance estimate. Yin (2008) proposed a class of the transformation cure frailty models based on the bounded cumulative hazard model for multivariate failure time data and used the Bayesian paradigm for the inference procedure.

Only a few papers in the literature concern the correlation structure for the multivariate survival data with a cure fraction. This is largely due to the difficulty in verifying the associations of both the cure statuses of all individuals and the survival times of the susceptible within the same cluster. Furthermore, the correlation assumptions are liable to be misspecified. In some circumstances, more interest lies in the covariate effects in the marginal model than the association within a cluster. Hence, we tend to leave the dependency structures of related individuals unspecified. As an example, for the smoking cessation data from the 51 zip codes in and around Rochester, Minnesota (Murray et al., 1988), which was analyzed by Banerjee and Carlin (2004) and Yu and Peng (2008), the primary interest is in inferring the effect of the smoking intervention, rather than the dependency structure with the same zip code. Therefore, the marginal approach is feasible (Wei et al., 1989; Lin, 1994; Spiekerman and Lin, 1998) as it is robust in the misspecification of the association structure, particularly in the multivariate survival data with a cure fraction.

This paper presents a semiparametric transformation cure model for multivariate survival data with a cure fraction. The approach fits the cure status with a logistic regression model, and the survival time of the susceptible individual with a semiparametric transformation model. Both submodels allow incorporating covariates and do not require any assumption of the association structure. This cure model includes the marginal semiparametric proportional hazards mixture cure model as a special case, which has been considered by Peng et al. (2007). The statistical inference is based on the marginal approach by constructing a system of estimating equations. The rest of the paper is organized as follows. In Section 2, we present the proposed cure model, and the estimating functions are derived. The computational algorithm and the asymptotic properties of the estimators are presented in Section 3. The performance of the proposed approach via simulations is reported in Section 4. To illustrate the method, the analysis of the smoking cessation data is presented in Section 5. The article is concluded with a discussion on the extension and future work.

2. Multivariate semiparametric transformation cure model

In this proposed cure model, the study population consists of both cured and noncured subpopulations indicated by a binary random variable D, in which $D = 1$ means noncured. The noncured individual has a finite survival time T^* for the event of interest. Let D_{ij} be the cure status for individual j in cluster i with the marginal model of logistic regression type

$$\Pr(D_{ij} = 1|X_{ij}) = \exp(\gamma' X_{ij})/(1 + \exp(\gamma' X_{ij})), \tag{1}$$

where X_{ij} is a time-independent vector of covariates with dimension p, and γ is the corresponding covariate effect. When $D_{ij} = 1$, the marginal survival distribution of T^*_{ij} is assumed to follow the semiparametric transformation model,

$$H(T^*_{ij}) = -\beta' Z_{ij} + \epsilon_{ij}, \tag{2}$$

where Z_{ij} is a time-independent vector of covariates with dimension q and covariate effect β. The transformation function $H(\cdot)$ is unknown and monotone increasing, and ϵ_{ij} is the error term with a prespecified marginal cumulative hazard function $\Lambda(\cdot)$. This two-component mixture cure model allows X_{ij} and Z_{ij} to share common covariates. If $\Lambda(u) = \exp(u)$, then (2) reduces to the proportional hazards model. Therefore, this cure model becomes the semiparametric proportional hazards mixture cure model for multivariate survival data considered by Peng et al. (2007). However, if $\Lambda(u) = \log(1 + \exp(u))$, then (2) reduces to the proportional odds model, and this cure model becomes a semiparametric proportional odds mixture cure model for multivariate survival data.

The waiting time, T_{ij}, for the jth individual in the ith cluster, is defined as $T_{ij} = D_{ij}T^*_{ij} + (1 - D_{ij}) \times \infty$. The observed multivariate survival data with N clusters are $\{(\tilde{T}_{ij}, \delta_{ij}, X_{ij}, Z_{ij}),\ j = 1, \ldots, n_i,\ i = 1, \ldots, N\}$, in which $\tilde{T}_{ij} = \min(T_{ij}, C_{ij})$ and $\delta_{ij} = I(T_{ij} \le C_{ij})$, where C_{ij} is the right censoring time and $I(.)$ is the usual indicator function. Conditional independence of T^*_{ij} and C_{ij} given covariates is assumed in this study. Emphasis is placed on estimating the marginal parameters (H, β, γ) and treating the association structure as a nuisance. For simplicity, let θ be the interesting parameters (H, β, γ). Given the true parameter $\theta_0 = (H_0, \beta_0, \gamma_0)$, the marginal survival function for the waiting time T_{ij}, derived from (1) and (2), is

$$S_T(t|X_{ij}, Z_{ij}) = \bar{G}(\gamma_0' X_{ij}) + G(\gamma_0' X_{ij}) \exp\bigl(-\Lambda\{H_0(t) + \beta_0' Z_{ij}\}\bigr),$$

where $G(u) = \exp(u)/(1 + \exp(u))$ and $\bar{G}(u) = 1 - G(u)$.