COMSTA: 5880 Model 3Gsc pp. 1–10 (col. fig: NIL)

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Semiparametric analysis of clustered interval-censored survival data with a cure fraction

Q1 K.F. Lam [a,*], Kin-Yau Wong [b]

[a] *Department of Statistics & Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong*

[b] *Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA*

ARTICLE INFO

Article history:
Received 19 August 2013
Received in revised form 29 January 2014
Accepted 27 May 2014
Available online xxxx

Keywords:
Asymptotic normal data augmentation
Clustered interval-censored data
Cure model
Frailty
Survival analysis

ABSTRACT

A generalization of the semiparametric Cox's proportional hazards model by means of a random effect or frailty approach to accommodate clustered survival data with a cure fraction is considered. The frailty serves as a quantification of the health condition of the subjects under study and may depend on some observed covariates like age. One single individual-specific frailty that acts on the hazard function is adopted to determine the cure status of an individual and the heterogeneity on the time to event if the individual is not cured. Under this formulation, an individual who has a high propensity to be cured would tend to have a longer time to event if he is not cured. Within a cluster, both the cure statuses and the times to event of the individuals would be correlated. In contrast to some models proposed in the literature, the model accommodates the correlations among the observations in a more natural way. A multiple imputation estimation method is proposed for both right-censored and interval-censored data. Simulation studies show that the performance of the proposed estimation method is highly satisfactory. The proposed model and method are applied to the National Aeronautics and Space Administration's hypobaric decompression sickness data to investigate the factors associated with the occurrence and the time to onset of grade IV venous gas emboli under hypobaric environments.

1. Introduction

Interval-censored data arise naturally in many epidemiological and biomedical research when the effects of interventions are assessed using surrogate outcomes based on the periodically observed laboratory measurements such as the determination of seropositivity in HIV and AIDS. The surrogate outcomes are only known to have occurred between two time points making the measurements and hence these data are interval censored. Interest in regression analysis of clustered interval-censored data is growing rapidly. A popular approach is to introduce a cluster-specific unobserved random effect term called frailty to the Cox proportional hazards model to accommodate the intra-cluster association. Bellamy et al. (2004) and Goethals et al. (2009) performed fully parametric analyses on a community-based study in asthma and clustered cow udder quarter infection times, respectively, using frailty models. Semiparametric analysis based on the frailty–Cox model with clustered interval-censored data is not straightforward. Kim (2010) approximated the nonparametric cumulative baseline hazard function using a piecewise linear function and the model was applied to the lymphatic filariasis data. Lam et al. (2010) considered a multiple imputation approach by converting the interval-censored data to right-censored data to allow a semiparametric analysis of the diabetic retinopathy study data using the well-known EM algorithm of Klein (1992).

* Corresponding author. Fax: +852 2858 9041.
E-mail address: hrntlkf@hku.hk (K.F. Lam).

http://dx.doi.org/10.1016/j.csda.2014.05.019
0167-9473/© 2014 Published by Elsevier B.V.

Except for fully parametric analysis, there is yet no established approach for the regression analysis of clustered interval-censored data, not to mention clustered interval-censored data with a cure fraction.

This study was motivated by the National Aeronautics and Space Administration's hypobaric decompression sickness data (HDSD) (Conkin et al., 1992; Conkin and Powell, 2001). The variable of interest was the time to onset of grade IV venous gas emboli (VGE) when volunteer individuals were exposed to different hypobaric environments. The main objective of the study was to understand the mechanism of the onset of grade IV VGE and its association with some explanatory variables like age and gender, as well as some experimentally manipulated variables like stress and posture. Individuals underwent the experiment possibly more than once with a maximum of 13 measurements from the same individual. Subjects were given enough time to wash out the possible carry-over effects from the previous experiments before undergoing another experiment. For each trial, the time to onset of grade IV VGE was only recorded as within a certain time interval (interval censoring) or beyond a certain time point (right censoring). Repeated measurements from the same individual were possibly dependent because the development of grade IV VGE was likely to depend on the build of the individual that could not be quantified easily. It was noted that 77.6% (425/548) of the observations were right censored and most censoring times were quite large relative to the upper limit of the interval-censored times. This was a strong evidence that subjects might not necessarily develop grade IV VGE at a particular trial and we simply referred this event as a *cure*. Moreover, a cure under one trial did not imply that the individual would not develop grade IV VGE in other trials.

In the past decade, extension of cure models had been considered widely to accommodate associations among the observations. Chatterjee and Shih (2001) considered a bivariate model for right-censored data using a copula to accommodate the associations based on the two marginal distributions. A fully parametric marginal survival distribution was assumed but no covariate was allowed in their formulation. Chen et al. (2002) extended the promotion time cure model (see Yakovlev et al. (1993); Chen et al. (1999); Ibrahim et al. (2001a,b)) to analyze right-censored clustered survival data. Bayesian inference was adopted. Banerjee and Carlin (2004) considered a cure model for interval-censored data which accommodated spatial correlations under the Bayesian paradigm, but covariate effect on the incidence probability was not allowed. Yu and Peng (2008) considered a marginal approach for clustered right-censored survival data. In their approach, the correlation structure could not be estimated and the sandwich variance estimator was used to correct for the variance estimate. Lai and Yau (2009) considered a logistic–Cox model that included two separate cluster-specific random effects with one in the logistic regression part for the cure probability and the other in the survival function part for the susceptible individuals. Their method was restricted to right-censored data only, but was further extended by Xiang et al. (2011) to accommodate interval censoring using a piecewise linear approximation to the nonparametric cumulative baseline hazard function. The restricted maximum likelihood estimation (REML) was adopted. However, it was assumed in the model that the two random effects responsible for modeling correlations among the cure statuses and the finite times to event of subjects in the same cluster were independent, which might not be realistic. Li and Ma (2010) assumed a fully parametric model with a bivariate random effect to accommodate the association between the cure status and the finite time to event of an individual in a trial. They applied the model to analyze the HDSD and remarked that there is yet no clear way to extend the existing semiparametric models to the two-part model.

As remarked by Kim and Jhun (2008), a subject with high propensity of occurrence inclined to have earlier occurrence time. For example, in a cancer trial, a more frail subject is less likely to be cured and is expected to experience a tumor recurrence at an earlier time. In many situations, it is natural that the cure probability is influenced by the underlying health condition, which also affects the time to event if the subject is not cured. The health condition of a subject may depend on some explanatory variables like age and sex. A model with a subject specific covariate-dependent frailty is proposed in this paper to handle the above mentioned complicated data structure. The frailty can be interpreted as a quantification of the health condition of a subject that determines the cure status and also affects his/her time to event. The clustering effect is induced by a cluster-specific latent variable. Details of the proposed model are described in Section 2. In Section 3, an estimation method using multiple imputation is proposed, first for right-censored data and then extended to accommodate interval-censored data. Simulation results are presented in Section 4 and the proposed model and estimation method are illustrated using the HDSD in Section 5. Some concluding remarks appear in Section 6.

2. Model formulation

Suppose we have a random sample of m clusters with n_i individuals in the ith cluster ($i = 1, \ldots, m$). Let T_{ij} be the actual time to event and $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ be a set of observed covariates for member j in cluster i, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$. Define $\boldsymbol{x}_{ij}^{(0)} = \left(1, x_{ij1}^{(0)}, \ldots, x_{ijp_0}^{(0)}\right)'$ and $\boldsymbol{x}_{ij}^{(1)} = \left(x_{ij1}^{(1)}, \ldots, x_{ijp_1}^{(1)}\right)'$ be the sets of covariates associated with the frailty and the time to event, respectively, where $p_0, p_1 \le p$ and that $x_{ijk}^{(l)}$ is an element in $\boldsymbol{x}_{ij}$ ($l = 0, 1$). Unlike the cure models for clustered data proposed by Lai and Yau (2009); Li and Ma (2010); Xiang et al. (2011) that involve two separate cluster-specific random effects, we suggest using only one frailty term U_{ij} for each individual to account for a cure and the association among the members of the same cluster. Moreover, for cluster i, the values of U_{ij}'s, $j = 1, \ldots, n_i$ are derived from a cluster-specific latent variable Ξ_i. Conditioned on $\Xi_i = \xi_i$, U_{ij}'s are independent and so are T_{ij}'s. The conditional hazard function of the semiparametric frailty–Cox model is specified as

$$\lambda\left(t \mid u_{ij}, \boldsymbol{x}_{ij}^{(1)}\right) = u_{ij}\lambda_0(t)\,\mu_{ij},$$