Computational Statistics and Data Analysis 103 (2016) 242–249

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A multiple imputation approach to the analysis of clustered interval-censored failure time data with the additive hazards model

Ling Chen [a,*], Jianguo Sun [b], Chengjie Xiong [a]

[a] Division of Biostatistics, Washington University School of Medicine, Campus Box 8067, 660 S. Euclid Ave, St. Louis, MO 63110, United States

[b] Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, United States

ARTICLE INFO

Article history:
Received 10 September 2014
Received in revised form 17 May 2016
Accepted 23 May 2016
Available online 28 May 2016

Keywords:
Additive hazards model
Clustered interval-censored data
Multiple imputation
Within-cluster resampling

ABSTRACT

Clustered interval-censored failure time data can occur when the failure time of interest is collected from several clusters and known only within certain time intervals. Regression analysis of clustered interval-censored failure time data is discussed assuming that the data arise from the semiparametric additive hazards model. A multiple imputation approach is proposed for inference. A major advantage of the approach is its simplicity because it avoids estimating the correlation within clusters by implementing a resampling-based method. The presented approach can be easily implemented by using the existing software packages for right-censored failure time data. Extensive simulation studies are conducted, indicating that the proposed imputation approach performs well for practical situations. The proposed approach also performs well compared to the existing methods and can be more conveniently applied to various types of data representation. The proposed methodology is further demonstrated by applying it to a lymphatic filariasis study.

1. Introduction

Correlated or clustered failure time data occur in many areas such as medical studies (Cai and Prentice, 1997). In many situations, the exact time of failure cannot be observed but is known to fall between two observation times. Such data are usually referred to as interval-censored failure time data. Clustered interval-censored failure time data may naturally arise in periodic follow-up studies where each study subject is repeatedly measured at discrete time points and some subjects come from the same cluster or group such as sibling, family and community. In this case, the interval-censored observations from the same cluster share certain unknown characteristics and are correlated as a result.

Several methods have been proposed for regression analysis of clustered interval-censored failure time data. These include parametric approaches (Bellamy et al., 2004; Zhang and Sun, 2010; Lam et al., 2010) and semiparametric approaches (Xiang et al., 2011; Zhang and Sun, 2013; Li et al., 2014). For most of the existing semiparametric methods, it is assumed that the failure time of interest follows the proportional hazards model (Pan, 2000). It is known that the proportional hazards model may not fit failure time data well sometimes and in this case, one of the alternatives is the additive hazards model. The additive hazards model describes a different aspect of the relationship between survival time and covariates and could be more plausible than the proportional hazards model in many situations (Lin and Ying, 1994). For example, in public

* Corresponding author. Tel.: +1 314 747 2373; fax: +1 314 362 2693.
E-mail addresses: lingchen@wustl.edu (L. Chen), sunj@missouri.edu (J. Sun), chengjie@wustl.edu (C. Xiong).

http://dx.doi.org/10.1016/j.csda.2016.05.011

health studies, people may be more interested in the risk difference described by the additive hazards model than the risk ratio described by the proportional hazards model (Breslow and Day, 1987). Li et al. (2012) investigated regression analysis of clustered interval-censored failure time data. However, they considered only the simple situation where each subject is observed only twice, or the failure time is left-, interval-, or right-censored. In the following, we will discuss general clustered interval-censored failure time data.

We develop a multiple imputation approach to analyze clustered interval-censored failure time data. Our method reduces the analysis of clustered interval-censored data to that of right-censored failure time data. The algorithm is implemented in two steps. First, we apply a within-cluster resampling (WCR) procedure (Hoffman et al., 2001) to sample a single subject from each cluster and obtain independent interval-censored failure time data. In other words, one observation is randomly sampled with replacement from each of the N clusters and an independent interval-censored dataset of size N is formed. Then a multiple imputation procedure is applied to convert the interval-censored dataset to right-censored failure time data. Inference is made based on the right-censored failure time data taking advantage of an existing estimation procedure developed by Lin and Ying (1994). If the resampling process is repeated a large number of times, say Q, where each of the Q analyses provides a consistent estimator of parameter of interest, one can obtain the WCR estimate by taking the average of the Q resampling-based estimates.

The multiple imputation approach has been used in many fields. In particular, Lam et al. (2010) developed such a procedure for the same type of the data considered here but under a parametric gamma frailty model. In addition, their method requires some strong assumptions on the distribution of the unobserved frailty and the use of the EM algorithm for the implementation. In contrast, the presented procedure has the advantage of simplicity because it can be easily implemented by using a standard software for the additive hazards model. Another advantage of the proposed approach is the use of a semiparametric additive hazards model, which provides more flexibility in describing the relationship between the failure time and covariates than a fully parametric model. In addition, the procedure given below does not need the estimation of the unobserved frailty.

In the following, before presenting the multiple imputation approach, we will first briefly introduce in Section 2 the additive hazards model and the inference procedure proposed by Lin and Ying (1994) for independent right-censored failure time data. The proposed estimation procedure is then presented in Section 3 and as mentioned above, it makes use of the method given in Lin and Ying (1994). Results from an extensive simulation study are reported in Section 4 for assessing the performance of our proposed approach and comparing to the existing method. Section 5 applies the proposed method to a set of well-known clustered interval-censored failure time data arising from a lymphatic filariasis study. Section 6 contains some concluding remarks.

2. The additive hazards model and right-censored failure time data with independent samples

Consider a survival study and let T denote the failure time of interest and Z a vector of covariates that may depend on time t. We assume that given Z, the hazard function of T has the form

$$\lambda(t; Z) = \lambda_0(t) + \beta' Z(t), \tag{1}$$

where $\lambda_0(t)$ denotes the unknown baseline function and β is the vector of unknown regression coefficients. That is, T follows the additive hazards model (Cox and Oakes, 1984). In the following, as most authors, it will be assumed that the covariate $Z(t)$ is known or can be observed at any time.

In this section, we will assume that instead of clustered data, one observes right-censored failure time data given by $\{X_i, \delta_i, Z_i,\ i = 1, \ldots, n\}$ from n independent subjects. Here X_i denotes the observed failure time defined as the minimum of the true failure time T_i and the censoring time for subject i and $\delta_i = 1$ if the true failure time is observed and 0 otherwise. Also it will be assumed that the failure time and the censoring time are independent given covariates. Define $Y_i(t) = I(X_i \geq t)$, the risk indicator process, and $N_i(t) = I(X_i \leq t, \delta_i = 1)$, a counting process, $i = 1, \ldots, n$.

To estimate β in model (1), Lin and Ying (1994) proposed to use the following estimating equation

$$U(\beta) = \sum_{i=1}^{n} \int_0^\infty \left\{Z_i(t) - \bar{Z}(t)\right\} \left\{dN_i(t) - Y_i(t)\beta' Z_i(t)dt\right\} = 0,$$

where

$$\bar{Z}(t) = \frac{\sum_{i=1}^{n} Y_i(t) Z_i(t)}{\sum_{i=1}^{n} Y_i(t)}.$$

It can be easily shown that the solution to the equation above has the explicit form

$$\hat{\beta} = \left[\sum_{i=1}^{n} \int_0^\infty Y_i(t)\{Z_i(t) - \bar{Z}(t)\}^{\otimes 2}\, dt\right]^{-1} \left[\sum_{i=1}^{n} \int_0^\infty \{Z_i(t) - \bar{Z}(t)\}\, dN_i(t)\right], \tag{2}$$