Computational Statistics and Data Analysis 125 (2018) 1–9

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Semiparametric analysis of the additive hazards model with informatively interval-censored failure time data

Shuying Wang [a], Chunjie Wang [b], Peijie Wang [a,*], Jianguo Sun [c]

[a] *Center for Applied Statistical Research, School of Mathematics, Jilin University, Changchun, 130012, China*
[b] *The College of Basic Science, Changchun University of Technology, Changchun, 130012, China*
[c] *Department of Statistics, University of Missouri, Columbia, MO, 65211, USA*

ARTICLE INFO

Article history:
Received 8 August 2017
Received in revised form 3 January 2018
Accepted 18 March 2018
Available online 28 March 2018

Keywords:
Additive hazards model
Case K interval-censored data
Informative censoring
Sieve maximum likelihood estimation

ABSTRACT

Regression analysis of failure time data has been discussed by many authors and for this, one of the commonly used models is the additive hazards model, for which some inference procedures have been developed for various types of censored data. In this paper, a much general type of censored data, case K informatively interval-censored data, is considered for which there does not seem to exist an established inference procedure. For the problem, a joint modeling approach that involves a two-step estimation procedure and the sieve maximum likelihood estimation is presented. The proposed estimators of regression parameters are shown to be consistent and asymptotically normal, and a simulation study conducted suggests that the proposed procedure works well for practical situations. In addition, an application is provided.

1. Introduction

Regression analysis of failure time data has been discussed by many authors and for this, one of the commonly used models is the additive hazards model, which assumes that the covariates of interest have additive effects on the failure time of interest and is often used when such types of effects are of interest such as in social sciences (Lin and Ying, 1994). Correspondingly some inference procedures for it have been developed for various types of censored data (Lin and Ying, 1994; Lin et al., 1998; Sun et al., 2006; Wang et al., 2010; Yin and Cai, 2004). Among them, a highly cited one is the estimating equation method given by Lin and Ying (1994) for right-censored failure time data. In this paper, we consider a much general type of censored data, case K informatively interval-censored data, for which there does not seem to exist an established inference procedure.

Interval-censored failure time data have recently attracted much attention due to their general structure and common occurrence in many areas such as demographical, financial and medical studies among others (Sun, 2006). In these situations, the exact occurrence time of failure event is not observed and instead is known only to belong to a window or an interval. It is apparent that interval-censored data include right-censored data as a special case and can occur in different forms. Among them, one that has been discussed by many is case I interval-censored data, which is also often referred to as current status data (Huang, 1996; Lin et al., 1998; Ma et al., 2015). By them, we usually mean that each study subject is observed only once and the only observed information for the failure event of interest is whether the event has occurred or not before the observation time. In other words, the failure time is either left- or right-censored. Case K interval-censored data can be seen

* Corresponding author.
E-mail address: wangpeijie@jlu.edu.cn (P. Wang).

https://doi.org/10.1016/j.csda.2018.03.011
0167-9473/

as a generalization of case I interval-censored data, and they mean that for each study subject, there exists a sequence of observation time points and one only observes between which two time points the failure event occurs.

Many methods have been developed for regression analysis of interval-censored failure time data for the situation where the censoring mechanism is independent of the failure time of interest (Chen et al., 2013; Huang, 1996; Sun, 2006). Some approaches have also been proposed for the case where the censoring mechanism may be related to the failure time of interest, which is often said to have informative censoring (Ma et al., 2015; Wang et al., 2016; Zhang et al., 2005, 2007). In particular, Wang et al. (2010) and Zhao et al. (2015) discussed regression analysis of informatively interval-censored data arising from the additive hazards model. The former considered case II interval-censored data, a special case of case K interval-censored data, while the latter investigated current status data. To deal with informative censoring or model the relationship between the failure variable of interest and censoring variables, two commonly used methods are the copula model approach and the frailty model approach. In the following, we will present a frailty model-based inference procedure.

The remainder of this article is organized as follows. First we will begin in Section 2 with introducing some notation and describing the assumed models as well as the structure of the observed data. The resulting likelihood function is then presented. In Section 3, a two-step estimation procedure is proposed for inference and in the procedure, the estimated sieve maximum likelihood estimation approach is used. The proposed estimators of regression parameters are shown to be consistent and asymptotically normal. Section 4 presents some results obtained from an extensive simulation study, which indicate that the proposed method seems to work well for practical situations. In Section 5, the proposed approach is applied to a set of case K interval-censored data arising from an AIDS study and Section 6 contains some discussions and concluding remarks.

2. Notation, assumptions and models

Consider a failure time study that involves n independent subjects and let T_i denote the failure time of interest for subject i. Also for subject i, suppose that there exists a p-dimensional vector of covariates denoted by x_i and there exists a sequence of observation times $U_{i0} = 0 < U_{i1} < U_{i2} < \cdots < U_{iK_i}$, where K_i denotes the number of observations on the subject. Define $\widetilde{N}_i(t) = \sum_{j=1}^{K_i} I(U_{ij} \le t)$ and $\delta_{ij} = I(U_{ij-1} < T_i \le U_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, K_i$. Then $\widetilde{N}_i(t)$ denotes the total number of observation times up to time t for the ith subject and jumps only at each observation time, and the observed data have the form

$$O = \Big\{ O_i = (\tau_i, U_{ij}, \delta_{ij}, x_i, j = 1, \ldots, K_i), i = 1, \ldots, n \Big\},$$

which are usually referred to as case K interval-censored data. In the above, τ_i denotes the follow-up time on the ith subject that will be assumed to be independent of T_i.

To describe the covariate effects and the relationship between the failure time of interest and the censoring mechanism, we will assume that there exists a latent variable b_i and given x_i and b_i, T_i and $\widetilde{N}_i(t)$ are independent. Also we will assume that given x_i and b_i, T_i follows the additive hazards frailty model

$$\lambda_i(t|x_i, b_i) = \lambda_0(t) + x_i^\top \beta_1 + b_i \beta_2 \,, \tag{1}$$

where $\lambda_0(t)$ denotes an unknown baseline hazard function and β_1 and β_2 are unknown regression parameters. Furthermore it will be assumed that given x_i and b_i, $\widetilde{N}_i(t)$ is a nonhomogeneous Poisson process with the intensity function

$$\lambda_{ih}(t|x_i, b_i) = \lambda_{0h}(t) \exp(x_i^\top \alpha + b_i) \,, \tag{2}$$

where $\lambda_{0h}(t)$ is an unknown continuous baseline intensity function and α a vector of regression parameters as β_1 and β_2. Note that models (1) and (2) with $b_i = 0$ have been commonly used in the analysis of failure time data (Klein and Moeschberger, 2003) and event history data (Cook and Lawless, 2007), respectively. It is apparent that the parameter β_2 represents the extent of the association between the failure time and the observation process. The two will be independent if $\beta_2 = 0$.

Define $\beta = (\beta_1^\top, \beta_2)^\top$ and $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. For inference about models (1) and (2), if the distribution of the b_i's is known, it is apparent that one could employ the observed likelihood function that would involve their distribution and

$$L(\beta, \Lambda_0 | b_i's) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{K_i} \Big(S_i(U_{ij-1}) - S_i(U_{ij}) \Big)^{\delta_{ij}} S_i(U_{iK_i})^{1-\sum_{j=1}^{K_i} \delta_{ij}} \right\}, \tag{3}$$

the conditional likelihood function given the U_{ij}'s, b_i's and x_i's, where $S_i(t) = \exp(-\Lambda_0(t) - (x_i^\top \beta_1 + b_i \beta_2)t)$. On the other hand, the likelihood would involve some difficult integrations and also the distribution of the b_i's is usually unknown. To avoid these issues, in the next section, we present a two-step estimation procedure that can be easily implemented.

3. A two-step estimation procedure

Now we will present a two-step estimation procedure for inference about models (1) and (2) by following the ideas discussed in Huang and Wang (2004) and Wang et al. (2016). The former considered regression analysis of recurrent event data and the latter discussed regression analysis of case K interval-censored data under the Cox model. The main idea behind