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

Q1 Regression analysis of current status data in the presence of dependent censoring with applications to tumorigenicity experiments

Q2 Shuwei Li [a], Tao Hu [b], Peijie Wang [a,*], Jianguo Sun [a,c]

[a] Center for Applied Statistical Research, School of Mathematics, Jilin University, Changchun, China
[b] School of Mathematical Sciences, Capital Normal University, Beijing, China
[c] Department of Statistics, University of Missouri, Columbia, MO, USA

ARTICLE INFO

Article history:
Received 29 March 2016
Received in revised form 23 December 2016
Accepted 29 December 2016
Available online xxxx

Keywords:
Dependent censoring
Frailty model approach
Monotone splines
Proportional hazards model

ABSTRACT

Current status data frequently occur in many fields including demographic studies and tumorigenicity experiments. In these cases, the censoring or observation time may be correlated to the failure time of interest, the situation that is often referred to as dependent or informative censoring. Although several semiparametric methods have been developed in the literature for the situation, they either only apply to limited situations or may be computationally unstable. To address these, a frailty model-based maximum likelihood approach is proposed with the use of monotone splines to approximate the unknown baseline cumulative hazard function of the failure time. Also a novel EM algorithm, which is based on a three-stage data augmentation and can be easily implemented, is presented. The proposed estimators are proved to be consistent and asymptotically normally distributed. An extensive simulation study is performed to assess the finite sample performance of the proposed approach and suggests that it works well for practical situations. An application to a tumorigenicity study is provided.

1. Introduction

Current status data are a type of failure time data that consist of only either left- or right-censored observations and often occur in many fields such as demographical investigations, cross-sectional studies and tumorigenicity experiments. For the situation, each study subject is observed only once and the failure time of interest is known only to be either smaller or larger than the observation time. One feature that often associates with current status data is that the observation time may be related to the failure time of interest, which is usually referred to as dependent or informative censoring. Although many authors have discussed regression analysis of current status data (Huang, 1996; Rossini and Tsiatis, 1996; Lin et al., 1998; Martinussen and Scheike, 2002; Sun, 2006; Chen et al., 2009; Hu et al., 2009; Wen and Chen, 2011), there exists only limited literature on semiparametric regression analysis of current status data with dependent censoring. In the following, we will discuss this latter situation and present a frailty model-based maximum likelihood approach.

One area that often produces current status data with dependent censoring is tumorigenicity experiments on the types of tumors that are between lethal and non-lethal, meaning that their occurrences have some effects on the animal death rate.

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

http://dx.doi.org/10.1016/j.csda.2016.12.011
0167-9473/

Please cite this article in press as: Li, S., et al., Regression analysis of current status data in the presence of dependent censoring with applications to tumorigenicity experiments. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2016.12.011

In these situations, current status data occur because all study animals are usually followed up to their death by either nature or sacrifice and thus the only relevant information available on the tumor occurrence is the tumor presence or absence at the death. A specific example on liver tumor, which motivated this study, is discussed in Section 6. Note that a great deal of literature has been developed on the analysis of such tumor data in the context of tumorigenicity experiments, but most of them are parametric procedures (Dinse and Lagakos, 1983; Lagakos and Louis, 1988; Rai et al., 2002). Recently several semiparametric methods have been developed in the literature for regression analysis of current status data with dependent censoring or observation times. For example, Ma et al. (2015) and Zhao et al. (2015) proposed some copula model-based estimation procedures that jointly model both the failure time of interest and the observation time. The former considered the situation where the failure time marginally follows the proportional hazards model, while the latter discussed the case where the failure time can be marginally described by the additive hazards model.

Instead of the copula model-based approach, Zhang et al. (2005) and Chen et al. (2012) gave some frailty model-based procedures that employ some latent variables to describe the association between the failure time of interest and the observation time. The former considered the marginal additive hazards model as Zhao et al. (2015) and the latter modeled the failure time using the transformation frailty model. For estimation, Chen et al. (2012) developed an EM algorithm as one usually does under the frailty model framework. In this paper, we will develop an efficient and stable EM algorithm based on a three-stage data augmentation with the use of Poisson latent variables. Additionally, the new algorithm allows one to estimate the covariance matrix of the estimated parameters by using the profile likelihood approach.

The remainder of this paper is organized as follows. We first begin in Section 2 with describing the notation, assumptions and models that to be used throughout the paper and then present the resulting likelihood function. In particular, we assume that the failure time of interest follows the proportional hazards frailty model. Section 3 discusses the proposed sieve maximum likelihood estimation procedure and describes the three-stage data augmentation-based EM algorithm. In the procedure, we employ monotone splines to approximate the baseline cumulative hazard function of the failure time of interest. In Section 4, the asymptotic properties of the proposed estimators are established. Some results obtained from an extensive simulation study conducted to evaluate the finite sample performance of the proposed approach are presented in Section 5 and suggest that it works well for practical situations. In Section 6, we apply the proposed methodology to a tumorigenicity experiment and Section 7 contains some discussion and concluding remarks.

2. Assumptions, models and the likelihood function

Consider a failure time study that consists of n independent subjects and only gives current status data. For subject i, let T_i denote the failure time of interest and suppose that there exists a p-dimensional vector of covariates denoted by X_i. Also for the subject, suppose that there exist two other associated times C_i and C_i^c related to the observation on T_i, where C_i may be related to T_i and C^c is independent of T_i. In the tumorigenicity experiment example described above, the former represents the animal natural death time, while the latter denotes the sacrifice time. For the information on T_i, one only observes $\tilde{C}_i = \min(C_i, C_i^c)$, $\Delta_i = I(\tilde{C}_i = C_i)$ and $\delta_i = I(T_i \le \tilde{C}_i)$. That is, the observed current status data have the form

$$\{ (\tilde{C}_i, \Delta_i, \delta_i, X_i); i = 1, \ldots, n \}.$$

In the following, we assume that the main goal is to estimate the covariate effects.

To describe the covariate effects, suppose that there exists a latent variable b_i with mean one and variance η and given X_i and b_i, the cumulative hazard function of T_i has the form

$$\Lambda(t|X_i, b_i) = \Lambda_1(t) \exp(X_i' \beta_1) b_i. \tag{1}$$

Here $\Lambda_1(t)$ denotes an unknown baseline cumulative hazard function and β_1 is a p-dimensional vector of regression parameters. As T_i, in practice, the observation time C_i may depend on covariates too. For this, we assume that given X_i and b_i, C_i follows the same model given by

$$\Lambda^c(t|X_i, b_i) = \Lambda_2(t) \exp(X_i' \beta_2) b_i, \tag{2}$$

where $\Lambda_2(t)$ and β_2 are defined as $\Lambda_1(t)$ and β_1, respectively. Note that Zhang et al. (2005) considered similar models with T_i following the additive hazards frailty model instead of model (1). More comments on the models above will be given below. In the following, as usual, we will assume that T_i and C_i are conditionally independent given the latent variable b_i.

Let $S(t) = \exp\{-\Lambda_1(t) \exp(X_i' \beta_1) b_i\}$, the survival function of T_i given X_i and b_i, $S^c(t) = \exp\{-\Lambda_2(t) \exp(X_i' \beta_2) b_i\}$ and $f^c(t) = \mathrm{d}\Lambda_2(t) \exp(X_i' \beta_2) b_i S^c(t)$, the survival and density functions of C_i given X_i and b_i, respectively. Then under the assumptions above, the likelihood function has the form

$$L_n(\beta_1, \beta_2, \Lambda_1, \Lambda_2, \eta) = \prod_{i=1}^{n} \int_0^\infty \left\{ \left[(1 - S(\tilde{c}_i)) f^c(\tilde{c}_i) \right]^{\delta_i} \left[S(\tilde{c}_i) f^c(\tilde{c}_i) \right]^{1-\delta_i} \right\}^{\Delta_i}$$
$$\times \left\{ \left[(1 - S(\tilde{c}_i)) S^c(\tilde{c}_i) \right]^{\delta_i} \left[S(\tilde{c}_i) S^c(\tilde{c}_i) \right]^{1-\delta_i} \right\}^{1-\Delta_i} p(b_i; \eta) \, \mathrm{d}b_i, \tag{3}$$

Please cite this article in press as: Li, S., et al., Regression analysis of current status data in the presence of dependent censoring with applications to tumorigenicity experiments. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2016.12.011