Computational Statistics and Data Analysis 87 (2015) 34–45

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Regression analysis of multivariate current status data with auxiliary covariates under the additive hazards model

Yurong Chen [a], Yanqin Feng [a,*], Jianguo Sun [b]

[a] School of Mathematics and Statistics, Wuhan University, Wuhan, 430072, PR China
[b] Department of Statistics, University of Missouri, Columbia, MO, USA

ARTICLE INFO

Article history:
Received 4 April 2014
Received in revised form 7 November 2014
Accepted 14 January 2015
Available online 22 January 2015

Keywords:
Auxiliary covariates
Interval-censored data
Partial likelihood function
Validation sample

ABSTRACT

In a biomedical study, it often occurs that some covariates of interest are not measured exactly and only some auxiliary information on them is available. In this case, a question of interest is how to make use of the available auxiliary information for statistical analysis. This paper discusses this problem in the context of regression analysis of multivariate current status failure time data arising from the additive hazards model. More specifically, we consider the situation where the covariates of interest are assessed only for the subjects in a validation set and a continuous auxiliary covariate is available for all subjects. For the problem, by employing the marginal model approach, we propose two procedures for estimation of regression parameters. The methods can be easily implemented and the asymptotic properties of the resulting estimators are established. Also an extensive simulation study is conducted for the evaluation of the proposed methods and indicates that they work well in practice. An illustrative example is provided.

1. Introduction

This paper discusses regression analysis of multivariate current status or case I interval-censored failure time data arising from the additive hazards model (Lin et al., 1998; Martinussen and Scheike, 2002; Sun, 2006). By case I interval-censored data, we mean that the failure time of interest is not exactly observed but the observation on it is either left- or right-censored. A typical example of such data is given by a tumorigenicity study and in this case, the time to tumor onset is often of interest. However, it is usually not observable as the presence or absence of tumors in animals is usually known only at their death or sacrifice. One can find current status data in many other areas too including demographical studies, economics, medical studies, reliability studies and social sciences (Jewell and van der Laan, 1995; Huang, 1996; Rossini and Tsiatis, 1996).

For any regression analysis, it is apparent that it would be helpful to have information about the covariates of interest on all study subjects. In a biomedical study, however, it is often the case that the true covariates may be observed or measured only on a subset of the study cohort, which is often referred to as the validation set, instead of the whole cohort for various reasons. One reason may be that the measurements on some covariates are expensive and to save the cost, one may instead collect the information on some related covariates that can be obtained relatively cheaply. Such an example is given in SOLVD (1991), which discussed a left ventricular dysfunction prevention trial to assess the effects of some covariates on the risk of heart failure and the first myocardial infarction. In the study, an ideal approach would be to employ a standardized

* Corresponding author.
E-mail addresses: yqfeng.math@whu.edu.cn, yanqf2008@aliyun.com (Y. Feng).

http://dx.doi.org/10.1016/j.csda.2015.01.005
0167-9473/

radionucleotide technique, which is usually quite expensive. To save the cost, the study used it only on a subset of the study subjects but used a relatively cheaper and easily obtained nonstandardized technique on all the patients to measure an auxiliary covariate. By auxiliary covariates, we usually mean the surrogate variables that are related to the true covariates but provide no additional information when the true covariates are known. In this case, a naive analysis approach is to base the analysis only on the subjects in the validation set. However, it is obvious that this would lose some efficiency and thus one may prefer to employ some methods that take into account the available auxiliary covariates. The literature discussing this problem includes Hu and Lin (2002), Huang and Wang (2000), Lin and Ying (1993), Liu et al. (2009), Wang et al. (1998), Zhou and Pepe (1995) and Zhou and Wang (2000).

Many procedures have been developed for regression analysis of interval-censored failure time data under various models. For example, Huang (1996) developed the maximum likelihood approach for fitting the proportional hazards model to case I interval-censored data and Chen et al. (2009) and Sun and Shen (2009) discussed the same problem in the presence of clustering and competing risks, respectively. Also Hu and Xiang (2013) considered the efficient estimation for semiparametric cure models when one faces case II interval-censored data, and Lin et al. (1998) and Chen and Sun (2009) discussed the fitting of the additive hazards model to case I interval-censored data. It is well-known that the additive hazards model describes a different aspect of the association between the failure time and covariates compared to the proportional hazards model and could be more plausible than the latter in many applications. One such case is epidemiological studies, in which epidemiologists are often interested in the risk difference as it can be more relevant to public health (Kulich and Lin, 2000). However, it does not seem to exist an established method for regression analysis of multivariate case I interval-censored data arising from the additive hazards model in the presence of auxiliary covariates. In the following, we present two approaches for the problem in the case of continuous auxiliary covariates.

The remainder of this paper is organized as follows. We will begin in Section 2 with introducing some notation, models and assumptions that will be used throughout the paper. Section 3 presents the estimation procedures for estimation of regression parameters in the additive hazards model based on current status data with continuous auxiliary covariates. In addition, the asymptotic properties of the proposed estimators, including the consistency and the asymptotic normality, are established. In Section 4, an extensive simulation study is conducted to evaluate the finite sample performance of the proposed methods and the results indicate that they work well for practical situations. An illustrative example is presented in Section 5 and Section 6 contains some discussions and concluding remarks.

2. Notation, models and assumptions

Consider a failure time study that consists of n independent subjects and in which there exist K different types of correlated failure times. For subject i, let T_{ik} denote the type k failure time and suppose that there exists a vector of covariates $Z_{ik}(t)$ that may depend on time t. For the relationship between T_{ik} and $Z_{ik}(t)$, in the following, we assume that given the history of covariates up to time t, the hazard function of T_{ik} has the form

$$\lambda_{T_k}(t|Z_{ik}(s), s \le t) = \lambda_{0k}(t) + \beta_0' Z_{ik}(t). \tag{1}$$

That is, T_{ik} follows the additive hazards model (Lin and Ying, 1994). In the above, $\lambda_{0k}(t)$ denotes an unknown marginal baseline hazard function and β_0 a vector of unknown regression parameters. Note that here for the simplicity, we assume the same covariate effects and it is straightforward to generalize the methods proposed below to the situation where the effects may be different.

In the following, we assume that for some subjects, covariates $Z_{ik}(t)$ are missing or not observed but there exists a vector of auxiliary covariates denoted by $X_{ik}(t)$ that are known or observed for all subjects. It will be assumed that the relationship between $X_{ik}(t)$ and $Z_{ik}(t)$ can be arbitrary, but conditional on $Z_{ik}(t)$, $X_{ik}(t)$ provides no additional information. In other words, we have

$$\lambda_{T_k}(t|Z_{ik}(t), X_{ik}(t)) = \lambda_{T_k}(t|Z_{ik}(t)).$$

Let V_k denote the set of indices of the subjects whose true covariates $Z_{ik}(t)$ are known, $\bar{V}_k$ the complement of V_k, and n_{v_k} and $n_{\bar{v}_k}$ the sizes of V_k and $\bar{V}_k$, respectively. The set V_k is usually referred to as the validation set. Note that here again for the simplicity, we assume that all components of the covariates are either known or missing together and some comments will be given below for the situation where the missing happens only on some of the components. Also we will assume that V_k is a simple random sub-sample of the whole set of study subjects. For the data on the failure times T_{ik}'s of interest, it will be supposed that each subject is observed only once at time C_{ik} and the observed information consists only of C_{ik} and $\delta_{ik} = I(T_{ik} \ge C_{ik})$, $i = 1, \ldots, n$; $k = 1, \ldots, K$. That is, we have current status data on the T_{ik}'s. In the following, we will assume that C_{ik} is independent of both Z_{ik} and T_{ik} and some comments on this will be given below.

For each (i, k), define $N_{ik}(t) = I(C_{ik} \le \min(t, T_{ik}))$ and $Y_{ik}(t) = I(C_{ik} \ge t)$. Then $N_{ik}(t)$ is a counting process with the intensity process

$$\lambda_{ik}(t|Z_{ik}(s), s \le t) = e^{-\Lambda_{0k}(t)}\, e^{-\beta_0' Z_{ik}^*(t)} \lambda_k^c(t) \triangleq \lambda_{0k}^c(t)\, e^{-\beta_0' Z_{ik}^*(t)} \tag{2}$$

(Lin et al., 1998), where $\lambda_k^c(t)$ denote the hazard of the event $\{C_{ik} = t\}$, $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(s)ds$ and $Z_{ik}^*(t) = \int_0^t Z_{ik}(s)ds$. Note that the equation above says that $\lambda_{ik}(t|Z_{ik}(s), s \le t)$ satisfies the Cox proportional hazards model. Based on this, if all true