Computational Statistics and Data Analysis 92 (2015) 134–147

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Regression under Cox's model for recall-based time-to-event data in observational studies

Sedigheh Mirzaei Salehabadi *, Debasis Sengupta

Applied Statistical Unit, Indian Statistical Institute, Kolkata, 700108, India

ARTICLE INFO

Article history:
Received 3 March 2015
Received in revised form 17 June 2015
Accepted 5 July 2015
Available online 13 July 2015

Keywords:
Informative censoring
Interval censoring
Proportional hazards
Related risk regression model
Retrospective study
Turnbull estimator

ABSTRACT

In some retrospective observational studies, the subject is asked to recall the age at a particular landmark event. The resulting data may be partially incomplete because of the inability of the subject to recall. This type of incompleteness may be regarded as interval censoring, where the censoring is likely to be informative. The problem of fitting Cox's relative risk regression model to such data is considered. While a partial likelihood is not available, a method of semi-parametric inference of the regression parameters as well as the baseline distribution is proposed. Monte Carlo simulations show reasonable performance of the regression parameters, compared to Cox estimators of the same parameters computed from the complete version of the data. The proposed method is illustrated through the analysis of data on age at menarche from an anthropometric study of adolescent and young adult females in Kolkata, India.

1. Introduction

Retrospective studies on a landmark event can produce dichotomous data on the current status of an individual (whether or not the event has occurred till the day of observation). From the perspective of the time to event, these data can be regarded as left or right censored. In some retrospective studies, the subject is asked to recall the time of the landmark event, in case it has already taken place. Such retrospective data can be incomplete because of the possibility that the time is forgotten. Sometimes the subject may be able to specify only a range for the time-to-event. For some other subjects, the event may be found not to have happened till the time of visit. Thus, data arising from this kind of retrospective studies are interval-censored. However, the chance of recall may depend on the time span between the occurrence of the event and the time of interview. For instance, between two young adult females interviewed at the same age, the one having experienced menarche more recently may have a higher chance of recalling the date. Thus, the censoring mechanism in this set-up is likely to be informative. Mirzaei et al. (2015) and Mirzaei and Sengupta (2015) have proposed parametric and nonparametric methods of likelihood based inference, when the data are subjected to informative interval censoring of this type. They have shown theoretically as well as through simulation that estimators of survival function that ignore the informative nature of censoring can have large bias even when the sample size is large (Mirzaei et al., 2015). On the other hand, the problem of regression with the above type of censored data has not been addressed yet.

The relative risk regression model, also known as the proportional hazards model, is widely used in the analysis of event time data with covariates. The method of analysis proposed by Cox (1972) can accommodate right-censored data which are usual in survival problems, and left-truncated data which arise when there are delayed entries in a cohort

* Corresponding author.
E-mail addresses: sedigheh_r@isical.ac.in (S. Mirzaei Salehabadi), sdebasis@isical.ac.in (D. Sengupta).

http://dx.doi.org/10.1016/j.csda.2015.07.005
0167-9473/

(Breslow et al., 1983). Other models which are used for more complex observation schemes include the accelerated failure time (AFT) model (Wei, 1992), the additive hazard regression model (Klein and Moeschberger, 2003), proportional odds ratio model (Dabrowska and Doksum, 1988) and so on (Vonta, 1996). There has also been some work on more general regression models for survival data, such as single index regression models (Chaudhuri, 2007) and models with random effect/frailty (Wienke, 2010).

Discrete-time regression models for right-truncated data have been developed and applied in the analysis of AIDS incidence and induction time distributions (Kalbfleisch and Lawless, 1991; Gross and Huber-Carol, 1992). Finkelstein (1986) and Finkelstein et al. (1993) discussed methods for fitting a discrete proportional hazards model for the case where the data are either interval-censored or right-truncated. In both cases, a score test was developed for testing the hypothesis of a zero regression coefficient. Tu et al. (1993) discussed a discrete proportional hazards model and an associated EM algorithm for data that are censored as well as truncated. Alioum and Commenges (1996) discussed a method for fitting a relative risk regression model for arbitrarily interval censored data. Their method assumes the censoring to be non-informative.

DeMasi et al. (1997) and Tanaka and Rao (2005) considered the regression problem for informatively censored data. Their model treats informative censoring as a type of risk in a competing risks setup, where the subject experiences two types of mutually exclusive events. This set-up is not meant to model the informative censoring found in recall data.

In this paper, we consider regression under Cox's model for the special type of informatively censored data arising from uncertainly recalled event time in a retrospective study. In Section 2, we develop a semiparametric maximum likelihood estimator of the regression coefficients under the model. In Section 3, we report results of a simulation study of the performance of the proposed maximum likelihood estimator. Section 4 illustrates this method with data on menarcheal age of adolescent and young adult females, collected during the course of a project undertaken by the Indian Statistical Institute, Kolkata. Proofs of all the results are given in Appendix.

2. Model and inference

2.1. Model

Consider a subject having time of occurrence of the landmark event T_i, which is a single sample from a distribution F_i with density f_i and support $[t_{\min}, t_{\max}]$, for $i = 1, \ldots, n$. Let these subjects be interviewed at times $S_1, \ldots, S_n \in [t_{\min}, t_{\max}]$, respectively. Suppose U_i is the unobservable time that the ith subject would take to forget the epoch of his/her landmark event. For the sake of simplicity, we assume that when the subject forgets the epoch, there is no recollection of an approximate range of time also. There are observable indicators δ_i and ε_i of the events $T_i \le S_i$ and $U_i > S_i - T_i \geqslant 0$, respectively. We assume that $U_1, \ldots, U_n$ are samples from a distribution with distribution function π, and that these are independent of both T_i and S_i. It follows that, given S_i and T_i, the non-recall probability depends on the time elapsed since the landmark event as

$$P(\varepsilon_i = 0 | T_i = t, S_i = s) = \pi(s - t), \quad s > t. \tag{1}$$

According to this model, the likelihood, conditional on the ages at interview, is

$$\prod_{i=1}^{n} [\bar{F}_i(S_i)]^{1-\delta_i} \left[\{f_i(T_i)(1 - \pi(S_i - T_i))\}^{\varepsilon_i} \left(\int_0^{S_i} f_i(u)\pi(S_i - u)du \right)^{1-\varepsilon_i} \right]^{\delta_i}. \tag{2}$$

Here the informativeness of the censoring mechanism is captured through the function π. If π is a constant, then the likelihood (2) becomes a multiple of the likelihood for non-informatively left- or right censored data with possibility of no censoring. As a further special case, if $\pi = 1$, then the likelihood (2) simplifies to the likelihood for dichotomous data. If $\pi = 0$, i.e., there is perfect recall with probability 1, then $\varepsilon_i = 1$ for all i such that $\delta_i = 1$, and the likelihood reduces to that for right-censored data.

Let Z_i be the r-dimensional vector of covariates, assumed to be independent of both S_i and U_i. Note that the distribution of T_i would depend on Z_i. Under Cox's relative risk regression model, the probability of the individual i, with covariate vector Z_i, having the event after time t is

$$\bar{F}_i(t) = [\bar{F}_0(t)]^{\exp(\beta^T Z_i)}, \tag{3}$$

where $\bar{F}_0$ is the baseline survival function, assumed to have a density.

2.2. Identifiability

Before embarking on developing a method of estimation, we need to check the identifiability of β, F_0 and π. By substituting (3) in the likelihood (2), after dropping the subscript i for simplicity and following Theorem 1 of Mirzaei et al. (2015), one can show that a typical factor in the product likelihood is equal to the conditional density of the observable vector (V, δ), given S and Z, where $V = (S - T)\varepsilon$. The conditional density is written alternatively as