Journal of Statistical Planning and Inference 139 (2009) 1381 – 1393

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Semiparametric inference of proportional odds model based on randomly truncated data

Rajeshwari Sundaram*,[1]

Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS, 6100 Executive Boulevard, Rockville, MD 20852, USA

ARTICLE INFO

Article history:
Received 29 March 2007
Received in revised form
19 May 2008
Accepted 13 August 2008
Available online 20 August 2008

MSC:
Primary 62N02, 62F12
Secondary 62F35, 62E20

Keywords:
Proportional odds model
Weighted empirical processes
Minimum distance
Truncation
Survival

ABSTRACT

This paper studies the estimation in the proportional odds model based on randomly truncated data. The proposed estimators for the regression coefficients include a class of minimum distance estimators defined through weighted empirical odds function. We have investigated the asymptotic properties like the consistency and the limiting distribution of the proposed estimators under mild conditions. The finite sample propertieswere investigated through simulation study making comparison of some of the estimators in the class. We conclude with an illustration of our proposed method to a well-known AIDS data.

Published by Elsevier B.V

1. Introduction

In many prospective and retrospective studies, survival data encountered are often incomplete due to censoring and/or truncation. Examples of randomly truncated data can be found in various areas like astronomy, economics (e.g. Woodroofe, 1985; Fiegelson and Babu, 1992), epidemiology and biometry (e.g. Keiding et al., 1987; Lagakos et al., 1988). For instance, random truncation occurs naturally in retrospective studies of a disease when only data available are those pertaining to individuals who have experienced some ascertaining event (E_2), prior to some time point t^*. So, individuals are enrolled into the study only if their time of occurrence t_2 of the event E_2 is less than or equal to t^*. For these individuals, the time t_1 of an initiating event (E_1) is determined retrospectively. Note that the survival time of interest $T = t_2 - t_1$ is observed only if $T \leqslant Y$, with $Y = t^* - t_1$ denoting the right truncating variable. An example of such a data was encountered in acquired immune deficiency syndrome (AIDS) study (Kalbfleisch and Lawless, 1989), where an individual is infected with human immuno-deficiency virus (HIV) at time t_1 and diagnosed with AIDS at time t_2. If the observation period is terminated at t^*, then only those individuals for whom the incubation time $T = t_2 - t_1$ is less than or equal to $Y = t^* - t_1$ can be observed and right truncation occurs. Another example can be found in AIDS cohort studies, where a group of patients who are infected with HIV at time s but have not yet developed AIDS (at some time $t > s$) are selected. If the recruitment starts at Y_0 then only those individuals for whom $T = t - s \geqslant Y = Y_0 - s$ are observed and random left truncation occurs. So, under random left truncation, one observes a pair (T, Y) only if $T \geqslant Y$, where

* Tel.: +1 301 435 6946; fax: +1 301 402 2084.
E-mail address: sundaramr2@mail.nih.gov.

[1] This research was supported by the Intramural Research Program of the NIH, NICHD.

0378-3758/$ - see front matter Published by Elsevier B.V
doi:10.1016/j.jspi.2008.08.006

T is the true variable of interest and Y is the truncating random variable. One can view random right truncation as random left truncation by just multiplying the survival time T and the truncating time Y by -1.

Non-parametric inference of distribution function based on randomly truncated data has been studied by Woodroofe (1985), Wang et al. (1986), Keiding and Gill (1990), Lai and Ying (1991)among others; regression analysis of such data has been studied by Gross and Lai (1996), Grigoletto and Akritas (1999), He and Yang (2003)among others. In this paper, our goal is to develop methodology to analyze the semiparametric proportional odds model based on randomly truncated data. The proportional odds model has been suggested as an alternative to the Cox's proportional hazards model in analyzing the effect of covariates on event times. Bennett (1983a,b) and Pettitt (1982)were among the first to generalize proportional odds model to the context of survival analysis. In the two sample setup, the proportional odds model is useful for fitting the data whose hazard rates converge asymptotically. Bennett (1983a) has suggested using proportional odds structure to model effective cure as the mortality rate of a disease group would converge to the mortality rate of a control group as time progresses under this model. Moreover, as pointed out by Murphy et al. (1997), this model can be used to demonstrate that the morbidity rate for a group halting a deleterious habit, such as smoking, converges to the morbidity rate of similar never-smokers as time progresses. Other examples of application of this model can be found in the analysis of environmental health data and in analysis of non-lethal tumors. In studying harmful effects of chemical exposure, experiments are performed in which groups of animals are exposed to different levels of chemicals. The resulting tumor distributions between groups are often compared using the proportional odds structure.

The proportional odds model specifies that the distribution function F_Z for failure time T associated with a p-dimensional covariate Z is parameterized by

$$\frac{1-F_Z(t)}{F_Z(t)} = \exp(Z'\beta)\frac{1-F(t)}{F(t)}. \tag{1.1}$$

Inference procedures under the above model has been studied by Pettitt (1982), Bennett (1983a,b), Bickel (1986), Dabrowska and Doksum (1988a) among others for complete data. Cheng et al. (1995), Dabrowska and Doksum (1988b), Murphy et al. (1997), Yang and Prentice (1999), Sundaram (2006a,b) have studied the inference for the above model based on right censored data. Huang and Rossini (1997) studied the above model for interval censored data and Rossini and Tsiatis (1996) for current status data. Grigoletto and Akritas (1999) have studied inference when the data are possibly left truncated. However, their method is only applicable for data with just one continuous covariate and the rest of the covariates have to be categorical. This may be too restrictive in practice. In this paper, we propose a methodology for regression analysis of semiparametric proportional odds model based on randomly truncated data with no such restriction on the covariates.

Motivated by robustness properties and ease of computation, Yang and Prentice (1999) proposed a weighted empiricals approach for analyzing the proportional odds model based on right censored data. They showed that their estimators had comparable performance in terms of efficiency, to the profile likelihood estimator of Murphy et al. (1997). In this paper, we generalize the weighted empiricals approach to study estimation of model (1.1) under random truncation. The organization of the paper is as follows: in Section 2, we discuss the inference procedure, the large sample properties of the proposed estimators are discussed in Section 3 and in Section 4, we present a numerical study and analyze the AIDS incubation time data presented in Kalbfleisch and Lawless (1989) to illustrate the proposed methodology.

2. Preliminaries

Let (T,Z) denote the survival time and the corresponding p-covariate. Let $F_Z(S_Z)$ denote the distribution (survival) function of T given Z, and $F(S)$ denote the distribution (survival) function of T given $Z=0$. Let $R_Z = F_Z/(1-F_Z)$ denote the odds function for failure given Z and let $R = F/(1-F)$ denote the baseline odds function for failure, i.e., the odds function for failure given $Z=0$. Under proportional odds model, the distribution of T given the covariate Z is modeled by

$$R_Z(t) = \exp(-Z'\beta)R(t). \tag{2.2}$$

The unknown parameters are R, the baseline odds function and β, a p-dimensional vector of regression coefficients. This model (2.2) has been studied by Yang and Prentice (1999) and Murphy et al. (1997)when data are possibly right censored. This paper studies inference of the above model (2.2) when the variable of interest T is possibly randomly left truncated. Under random truncation, the observation (T,Z) is interfered by another independent variable Y such that all three quantities are observed only if $T \geqslant Y$. In absence of random truncation, one could think of the observations as an infinite sequence of independent random vectors $\{(T_k, Y_k, Z_k); k = 1,2,\ldots\}$. Under random truncation, however, some of these vectors would be missing and only a subsequence $\{(T_{k_i}, Y_{k_i}, Z_{k_i}); i = 1,2,\ldots\}$ can be observed. We will denote the observable subsequence by

$$(U_i, V_i, W_i), \quad i = 1,2,\ldots \text{subject to } U_i \geqslant V_i, \tag{2.3}$$

i.e., $U_i = T_{k_i}, V_i = Y_{k_i}, W_i = Z_{k_i}$. So, the observed randomly truncated sample of size n will be denoted by $\{(U_i, V_i, W_i), i = 1,2,\ldots,n\}$.

Denote the cumulative distribution (survival) function of the survival time T and the random truncating variable Y by $F_Z(S_Z)$ and $G_Z(\bar{G}_Z)$ respectively, conditional on $Z=z$. Typically, under random truncation T and Y are assumed independent, conditional on the covariate $Z=z$. Let H denote the cumulative distribution function of Z. Let (a_F, b_F) denote the support of F defined by

$$a_F = \inf\{x : F(x) > 0\} \quad \text{and} \quad b_F = \sup\{x : F(x) < 1\}.$$