Journal of the Korean Statistical Society 45 (2016) 89–100

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Estimation of Kendall's tau for bivariate doubly truncated data

Pao-sheng Shen

Department of Statistics, Tunghai University, Taichung, 40704, Taiwan

ARTICLE INFO

Article history:
Received 5 January 2015
Accepted 25 July 2015
Available online 12 August 2015

AMS 2000 subject classifications:
primary 62N01
secondary 62N02

Keywords:
Association of variables
Bivariate truncation
Inverse-probability-weighted

ABSTRACT

In this article, we consider the estimation of Kendall's tau for bivariate doubly truncated data, where two correlated event times are potentially observed only if both fall within subject specific intervals of times. Using the inverse-probability-weighted (IPW) approach, we propose two nonparametric estimators of Kendall's tau for bivariate doubly truncated data. The first estimator is based on V-statistics and the second estimator is based on weighted comparable pairs. The asymptotic properties of the proposed estimators are established. Simulation studies are conducted to investigate their finite sample performance.

1. Introduction

Double truncation of survival data occurs when only those individuals whose event times lie within a certain subject-specific observational window are observed. Doubly truncated data play an important role in the statistical analysis of survival times (Bilker & Wang, 1996; Moreira & de Uña-Álvarez, 2010a,b; Shen, 2010a,b; Zhu & Wang, 2012, 2014) as well as in other fields such as astronomy (Efron & Petrosian, 1999; Lynden-Bell, 1971) or economy. In this article, we consider bivariate double-truncated data. Consider the following application:

Example: Age-of-onset anticipation (AOA)

The clinical phenomenon called age-of-onset anticipation or AOA is defined as a decrease in age at onset and/or an increase in disease severity in successive generations of afflicted families. The cause of AOA has been identified as DNA instability, such as nucleotide repeats that change in length in subsequent generations, which alters the phenotype of the disease. Reports of AOA exist in literatures, such as bipolar disorder (McInnis, McMahon, Stine, & Ross, 1993), facioscapulohumeral muscular dystrophy (Zatz et al., 1995), schizophrenia (Bassett & Honer, 1994), rheumatoid arthritis (Deighton, Heslop, McDonagh, Walker, & Thomson, 1994). Recently, changes in disease phenotype in subsequent generations also have been identified in other disorders, such as in colon cancer, breast cancer, Alzheimer disease and diabetes (Nilbert, Timshel, Bernstein, & Larsen, 2009; Paterson, Kennedy, & Petronis, 1996).

The data set used in testing for AOA usually consist of affected parent–child pairs between two calender times, say τ_1 and τ_2. Hence, the age of onset distribution in parents and children, respectively, is doubly truncated relative to the population distribution. For example, for the period 1992–2003 the Odense Pharmaco-epidemiological Database (OPED) (see Støvring, Andersen, Beck-Nielsen, Geen, & Vach, 2003, Støvring & Wang, 2007) contains subject information on all prescriptions for subsidized medications redeemed at any pharmacy in the County of Fyn, as well as information on births, deaths and

E-mail address: psshen@thu.edu.tw.

http://dx.doi.org/10.1016/j.jkss.2015.07.005
1226-3192/

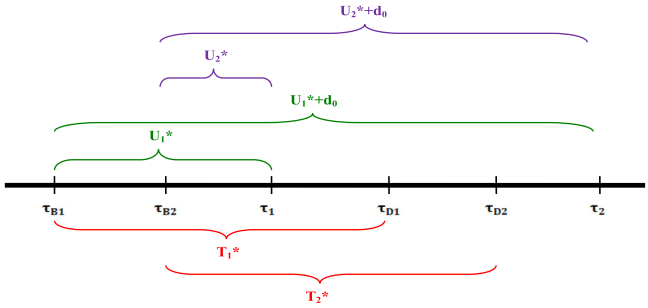

Fig. 1. Schematic depiction of bivariate doubly truncated data.

migration into and out of the County of Fyn. The tracking of individuals is based on the Civil Registration Number (CRN) which is assigned to all at birth or first immigration into Denmark. Incident events (occurrence of diabetic) are defined to be the first treatment event observed in the time window for subjects who did not have any previous events during a one year run-in period. Assume that the minimal and maximal observable age (in years) at diabetic onset before death is known and denoted by τ_0 and τ_M, respectively. Let $\tau_1 = 1992$ and $\tau_2 = 2003$. Define a target population as the individuals who were born after the calendar time (in years) $\tau_1 - \tau_M$ (i.e. 1992-maximum age = year of birth for the oldest person), and before $\tau_2 - \tau_0$ (i.e. 2003-minimum age = year of birth for the youngest person) and will be treated with the diabetic before death. For a pair of parents and children of the population defined above, let τ_{B1} and τ_{B2} be the calendar time (in years) of the initiating events (birth) for parents and children, respectively. Similarly, let τ_{D1} and τ_{D2} be the calendar time (in years) at diabetic onset for parents and children, respectively. Let $T_1^* = \tau_{D1} - \tau_{B1}$ and $T_2^* = \tau_{D2} - \tau_{B2}$ be the age (in years) at diabetic onset for parents and children. For $i = 1, 2$, let $U_i^* = \tau_1 - \tau_{Bi}$ and Let $d_0 = \tau_2 - \tau_1$. Notice that U_i^* and $U_i^* + d_0$ denote the age (in years) at τ_1 and τ_2, respectively. Hence, we observe (T_1^*, T_2^*) if and only if $U_1^* \le T_1^* \le U_1^* + d_0$ and $U_2^* \le T_2^* \le U_2^* + d_0$. To assess AOA, the researchers are interested in finding the association between age of onset of parents and that of their children. Thus, we need to estimate association between T_1^* and T_2^* using bivariate doubly truncated data. Fig. 1 highlights all the different times for bivariate doubly truncated data as described in example.

For any distribution function W denote the left and right endpoints of its support by $a_W = \inf\{t : W(t) > 0\}$ and $b_W = \inf\{t : W(t) = 1\}$, respectively. For $i = 1, 2$, let $G_i(u) = P(U_i^* \le u)$ denote the distribution function of U_i^*. Throughout this article we assume that $a_{G_1} = a_{G_2} = a_G$, $a_{F_1} = a_{F_2} = a_F$ and

$$a_G \le a_F \le a_G + d_0 \quad \text{and} \quad b_G \le b_F \le b_G + d_0. \tag{1.1}$$

Under assumption (1.1), $F_i(t) = P(T_i^* \le t)$ and $G_i(u)$ are both identifiable (see Woodroofe, 1985). Furthermore, we assume that (T_1^*, T_2^*) is independent of (U_1^*, U_2^*).

The measurement of association has been a major topic in bivariate survival analysis. Kendall's tau (Kendall & Gibbons, 1990) is a popular measure of association and is suitable for lifetime data since it is rank invariant. Let (T_{1i}^*, T_{2i}^*) and (T_{1j}^*, T_{2j}^*) $(i \ne j)$ be two independent realizations from (T_1^*, T_2^*). The (i, j)th pair is called concordant if $(T_{1i}^* - T_{1j}^*)(T_{2i}^* - T_{2j}^*) > 0$ and discordant if $(T_{1i}^* - T_{1j}^*)(T_{2i}^* - T_{2j}^*) < 0$. The untruncated version of Kendall's tau is (denoted by τ) defined as the difference of concordance and discordance probabilities between the (i, j)th pair, i.e.

$$\begin{aligned} \tau &= P((T_{1i}^* - T_{1j}^*)(T_{2i}^* - T_{2j}^*) > 0) - P((T_{1i}^* - T_{1j}^*)(T_{2i}^* - T_{2j}^*) < 0) \\ &= 2P((T_{1i}^* - T_{1j}^*)(T_{2i}^* - T_{2j}^*) > 0) - 1. \end{aligned} \tag{1.2}$$

In Section 2, using inverse-probability-weighted (IPW) approach, we propose two nonparametric estimators of Kendall's tau for bivariate doubly truncated data. The first estimator is based on V-statistics and the second estimator is based on weighted comparable pairs. The asymptotic properties of the proposed estimators are established. In Section 3, a simulation study is conducted to investigate their performance in finite samples.

2. The proposed estimators

2.1. The V-statistics approach

When (T_{1i}^*, T_{2i}^*)'s are continuous positive random variables, τ can be written as

$$\tau = 4P(T_{1i}^* > T_{1j}^*, T_{2i}^* > T_{2j}^*) - 1 = 4\int_0^\infty \int_0^\infty S(x, y)S(dx, dy) - 1,$$

where $S(x, y) = P(T_1^* > x, T_2^* > y)$ is the joint survival function of T_1^* and T_2^*. Therefore, we can write $\tau = \Gamma(S)$, where $\Gamma : \mathcal{D}[\mathcal{S}] \to \mathcal{R}$, $\mathcal{S} = \{(x, y) : S(x, y) > 0\}$ and $\mathcal{D}[\mathcal{S}]$ is the space of cadlag functions on $\mathcal{S}$. When there is no truncation, i.e. (T_{1i}^*, T_{2i}^*)'s are observable, τ can be estimated by $\Gamma(\hat{S})$, where $\hat{S}$ is the empirical estimator of S. $\Gamma(\hat{S})$ has the form of a