Logistic Regression Method in Interval-Censored Data

Eunyoung Yun¹ · Jinmi Kim² · Choongrak Kim³

¹Department of Statistics, Pusan National University
²Department of Statistics, Pusan National University
³Department of Statistics, Pusan National University

(Received July 2011; accepted August 2011)

Abstract

In this paper we propose a logistic regression method to estimate the survival function and the median survival time in interval-censored data. The proposed method is motivated by the data augmentation technique with no sacrifice in augmenting data. In addition, we develop a cross validation criterion to determine the size of data augmentation. We compare the proposed estimator with other existing methods such as the parametric method, the single point imputation method, and the nonparametric maximum likelihood estimator through extensive numerical studies to show that the proposed estimator performs better than others in the sense of the mean squared error. An illustrative example based on a real data set is given.

Keywords: Cross validation, imputation method, Kaplan-Meier estimator, median survival time, nonparametric maximum likelihood estimation, survival function.

1. Introduction

Interval-censored data are incomplete data, where the exact time of the event is not available, however, the event is known to occur between two defined times. Interval censoring is a general concept of censoring that includes right censoring and left censoring as special cases. Interval-censored data are more difficult to deal with than the right-censored data because of the complexity and special structure of the interval censoring. Thus, existing methods in right-censored data cannot be directly applicable to the interval-censored data. For example, the Kaplan-Meier type estimator for the interval-censored data is not analytically given. Among studies in interval-censored data, Peto (1973) derived survival curve for the interval-censored data, Finkelstein (1986) and Huang (1996) suggested a proportional hazards model for interval-censored data; in addition, Lindsey and Ryan (1998) provide an excellent review of this area. Recently, Fang et al. (2002) made nonparametric comparisons, Hudgens (2005) suggested a nonparametric maximum likelihood estimation with interval censoring and left truncation, and Lawless and Babineau (2006) proposed a simulation-based inference for

Here we consider the case II interval-censored data (a.k.a general case of interval-censored data) among four types of interval censoring (see Sun (2006) for details) because most of interval-censored data belong to the case II. As in the right-censored data, the estimation of survival function is of prime interest when we are given interval-censored data; other issues such as estimation of the median survival time, comparison of survival functions, relationship between survival times and the seemingly relevant covariates will be followed. For the parametric estimation of survival function a specific family of distributions such as Weibull is assumed a priori. There are many methods to estimate the survival function nonparametrically. Among them imputation method (Rubin, 1987) and nonparametric maximum likelihood estimator (Turnbull, 1976; Groeneboom and Wellner, 1992; Jongbloed, 1998) are often used.

In this paper we propose a method of estimating survival function and the median survival time using a logistic regression method in interval-censored data. The proposed method is motivated because the fact that the event of interest never occurred before the interval and it surely occurred after the interval. Therefore, we can augment arbitrary many samples both before and after the interval. If we regard the event before the interval failure and after the interval success, then by treating time as covariate a parametric logistic regression can be used to estimate the survival function. A cross-validation criterion to estimate the number of augmentation is proposed. Through simulation studies we compare the proposed method with imputation method and nonparametric maximum likelihood estimator in estimating the survival function and the median survival time.

2. Existing Methods

Let $T_1, T_2, \ldots, T_n$ be true survival times, then $T_i$ is called interval-censored if instead of observing $T_i$ exactly, only an interval $(L_i, R_i]$ is observed such that $T_i \in (L_i, R_i]$, where $L_i \leq R_i$. Here, $L_i$ and $R_i$ indicate the left and right endpoints of the observed event time. Recall that $L_i = R_i$ means an exact observation, $R_i = \infty$ represents a right-censored observation, and $L_i = 0$ represents a left-censored observation.

Imputation method (Rubin, 1987) is originally intended for handling missing data problems, and the method can be adapted to the analysis of interval-censored data. The idea is to impute exact survival times from interval-censored data and to take advantage of many standard methods for right-censored data. For subject $i$, the underlying true failure time $T_i$ is equal to a value within the observed interval $(L_i, R_i]$, $i = 1, \ldots, n$. Mean imputation is to let $T_i$ be the middle point of the interval for a finite interval, i.e., $R_i < \infty$. For intervals with $R_i = \infty$, they are regarded as right-censored observations. Taking $T_i$ to be $L_i$ or $R_i$ is the left end point imputation or the right end point imputation, respectively. Then, the interval-censored data are transformed to the right-censored data, and therefore, many standard methods for right-censored data can be used. This imputation method is particularly called a single imputation method. On the other hand, multiple imputation methods (Tanner and Wong, 1987; Tanner, 1991) are also available in analyzing the interval-censored data.

For the right-censored data, the nonparametric maximum likelihood estimator (NPMLE) of a survival function is just the Kaplan-Meier estimator (Kaplan and Meier, 1958). For the interval-