Statistical Inference of Interval-censored Failure Time Data

Jinheum Kim

Department of Applied Statistics, University of Suwon

1. Introduction

Interval censoring means that the time to some event such as death is known only to lie within an interval instead of being observed exactly. Interval-censored data might be generated in many clinical trials and longitudinal studies (Finkelstein, 1986; Kalbfleisch and Prentice, 2002; Sun, 2006). One common example occurs in medical or health studies that involve periodic follow-up. In this situation, an individual may miss some observations and return with a changed status. Accordingly, we only know that the true event time is greater than the last observation time at which the change has not occurred and less than or equal to the first observation time at which the change has been observed to occur.

An important special case of interval-censored data is the current status data (Jewell and van der Laan, 1995; Sun and Kalbfleisch, 1993). This type of censoring means that each subject is observed only once for the status of the occurrence of the event of interest. In other words, we do not directly observe the survival endpoint but, instead, we only know the observation time and whether or not the event of interest has occurred at the time. Thus, the survival time is either left- or right-censored. One such example is the data arising from cross-sectional studies on survival events. Another example is given by the tumourigenicity study in which the time to tumor onset is of interest but not directly observable. As a matter of fact, we only have the exact measurement of the observation time, which is often the death or sacrifice time of the subject. Sometimes we also refer current status data to as case I interval-censored data and the general case as case II interval-censored data.

We now introduce some notation for interval censoring. Let $T$ denote the survival time of interest. When $T$ is interval-censored, $I = (L, R]$ is used to denote the interval containing $T$. Using this notation, we see that current status data correspond to the situation where either $L = 0$ or $R = \infty$. Interval censoring also contains right censoring and left censoring as special cases. That is, if $R = \infty$, we have a right censored observation, while if $L = 0$ we obtain a left censored observation.

To this point, the survival time has been defined as the time between a fixed starting time point zero and the event time. One can apply a more general framework that defines the survival time as the time between two related events whose occurrence times are random variables and both could suffer censoring. If there is right or interval censoring on both occurrence times, the resulting data are commonly referred to as doubly-censored data (De Gruttola and Lagakos, 1989; Sun, 2004). An example of such complicated type of data is provided by the AIDS studies when the variable of interest is AIDS incubation time (Sun, 2004), the time from HIV infection to AIDS diagnosis, with both the HIV infection time and the AIDS diagnosis time being right- or interval-censored.

In applications, interval-censored data can be easily confused with grouped survival data. There is actually a fundamental difference between these two data structures although both usually appear in the form of intervals. The grouped survival data can be seen as a special case of interval-censored data and commonly mean that the intervals for any two subjects either are completely identical or have no overlapping. In contrast, the intervals for interval censored...
data may overlap in any way. As a consequence of this structure difference, statistical methods for grouped survival data are much more straightforward than those for interval-censored data.

2. An illustrative example

To illustrate the concepts described above, consider a set of well-known case II interval censored data on breast cancer, which can be found in Finkelstein and Wolfe (1985) and Sun (2006) among others. The data consist of 94 early breast cancer patients treated at the Joint Center for Radiation Therapy in Boston between 1976 and 1980. For their treatments, the patients were given either radiation therapy alone (RT, 46 patients) supposed to have clinic visits every 4-6 months to be examined for cosmetic appearance such as breast retraction. However, actual visit times differ from patient to patient and the times between the visits also vary. As a consequence, only interval-censored data are observed for breast retraction times. Specifically, among the 94 patients, 38 of them did not experience breast retraction during the study, giving right-censored observations for the breast retraction times. For other patients, intervals such as (25, 37] were observed for their breast retraction times. Here, the interval (25, 37] means that the patient had a clinic visit at month 25 and no breast retraction was detected at the visit, while at the next visit at month 37, breast retraction was found to be present already. There are five patients for whom the breast retraction was detected at their first clinical visits, giving the observed intervals with the left end points being zero or left-censored observations. One objective of the study was to compare the two treatments through their effects on breast retraction.

3. Selected topics

In general, this talk will follow Sun (2006) with the focus on medical applications, some basic issues and available methods for them, and the recent development in the literature for interval-censored data.

- We will describe a fundamental and important assumption behind most methodologies dealing with interval censoring: non-informative interval censoring. It basically states that the censoring mechanism does not contribute to the likelihood function. In the case of right-censored failure time data, a common assumption is that the censoring time is independent of the survival time of interest marginally or conditionally given external covariates. It is clear that this assumption cannot be directly generalized to interval censoring since the endpoints of the interval, \( L \) and \( R \), together with the survival time \( T \), have a natural relationship \( L < T \leq R \). Instead, for interval-censored data, the non-informative interval censoring assumption specified as

\[
Pr(T \leq t | L = l, R = r, L < T \leq R) = Pr(T \leq t | l < T \leq r)
\]

is used (Sun, 2006; Oller, Gomez, and Calle, 2004). In other words, except for the fact that \( T \) lies between \( l \) and \( r \) which are the realizations of \( L \) and \( R \), the interval \((L, R)\) does not provide any extra information for \( T \).

- For the analysis of interval-censored data, we will first discuss non-parametric estimation of a survival function as well as a hazard function. Among many methods, we will briefly describe three algorithms commonly used for case II interval-censored data. The first and simplest one is the self-consistency algorithm given in Turnbull (1976). In fact, the algorithm is essentially an application of the EM algorithm. The second approach is the iterative convex minorant (ICM) algorithm introduced by Groeneboom and Wellner (1992), which converges faster than the self-consistency algorithm. The third commonly used algorithm is the EM-ICM algorithm presented in Wellner and Zhan (1997). As suggested by the name, it is a hybrid and the fastest algorithm which combines the first two approaches. All three algorithms are iterative and there is no closed form for the non-parametric maximum likelihood estimator of the survival function.
The methods for comparing several survival functions will be reviewed. With right censoring, many non-parametric test procedures have been developed and most of them can be classified into two categories: rank-based tests and survival-based tests. The fundamental difference between them is that the former relies on the differences between the estimated hazard functions while the latter bases the comparison on the differences between the estimated survival functions. Among them, the log-rank test is perhaps the most widely used method. A few of these rank-based or survival-based test procedures have been generalized to the case of interval-censored data. Here, we review three procedures that are direct generalizations of the corresponding methods for right-censored data. First we discuss a rank-based approach (Zhao and Sun, 2004) that is a direct generalization of the log-rank test. Instead of the generalizations of the rank-based tests for right-censored data, there also exist a few generalizations of the survival-based tests. The third method (Sun, Zhao, and Zhao, 2005) that we will briefly discuss is a generalization of the method given in Peto and Peto (1972).

Regression analysis of interval-censored data under various semi-parametric models is then considered. Unlike most methods developed for right-censored data, estimating regression parameters under interval censoring usually involves estimation of both the parametric and the non-parametric parts. In other words, for interval-censored data, one has to deal with estimation of some unknown baseline functions in order to estimate regression parameters. In the case of case II interval-censored data, Finkelstein (1985) proposed to apply the Newton-Raphson algorithm to determine the MLE of the vector of unknown regression parameters in Cox proportional hazards model and the baseline cumulative hazard function together. The approach actually simplifies the situation to a finite-dimensional parametric estimation problem. Alternatively, one can apply the marginal likelihood approach and the stochastic approximation algorithm given in Satten (1996), the Markov Chain Monte Carlo EM algorithm developed by Goggins et al. (1998) or the multiple imputation-based method presented in Pun (2000). We will also discuss the proportional odds model, the accelerated failure time model, the additive hazards model, and the linear transformation model commonly used in survival analysis.

We briefly cover a few other topics including parametric approaches, interval-censored data with truncation, multivariate interval-censored data, competing risks interval-censored data and informative interval censoring. Multivariate interval-censored data arise if a survival study involves several related survival variables of interest and each of them suffers interval censoring. It is apparent that in this case, one needs different inference procedures than those discussed above and one key and important feature of these different procedures is that they need to take into account the correlation among the survival variables. In addition to the basic issues discussed before, a new and unique issue for multivariate data is to make inference about the association between the survival variables. For this, one of the tools that are commonly used is the copula model, which provides a very flexible way to model the joint survival function. Competing risks analysis is needed when the failure on an individual may be one of several distinct failure types. For example, death of a cancer patient may be classified as disease-related or non-disease-related. As mentioned above, all methods discussed so far require the non-informative interval censoring assumption. Several inference procedures have been developed in the literature for situations where the censoring may be informative (Sun, 2006). For this, a common way is to jointly model the survival variable and the variables representing interval censoring by using the latent variable approach (Zhang, Sun, and Sun, 2005; Zhang et al., 2007).

References