Applied Mathematical Modelling 39 (2015) 4630–4645

Contents lists available at ScienceDirect

Applied Mathematical Modelling

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

Local linear smoothing to estimate accelerated lifetime model with censoring and truncation

Antonio Jesús López-Montoya [a], María Luz Gámiz-Pérez [a,*], María Dolores Martínez-Miranda [a,b]

[a] University of Granada, Spain
[b] Cass Business School, City University London, United Kingdom

ARTICLE INFO

Article history:
Received 22 July 2014
Received in revised form 25 February 2015
Accepted 20 March 2015
Available online 25 April 2015

Keywords:
Semi-parametric
Nonparametric
Survival function
Local linear
Water supply system
Bootstrap

ABSTRACT

A nonparametric accelerated failure time (*AFT*) model is considered to evaluate the dependability measures of a water supply system laid in a city of the Mediterranean Sea. To do it, a set of breakdown data of sections of pipe of the system is available.

Unlike the usual methods that assume a parametric family for the underlying lifetime, we propose an *AFT* model with an unspecified distribution for the underlying lifetime.

To carry out the model fitting we suggest a two-stage procedure. Firstly, we estimate the influence of certain factors over the lifetime of the system. Secondly, we propose modern and flexible statistical tools based on counting processes to construct a smooth estimate the reliability measure of a specific system. We prove the good asymptotic properties of the estimator.

1. Introduction

In reliability analysis, right-censoring and left-truncation are common features arising in lifetime datasets. Consequently, specific models and statistical procedures to analyze failure times data have been developed in the specialized literature. The proportional hazards (*PH*) model [1] and the accelerated failure time (*AFT*) model [2,3]are frequently used in applied studies (see for example [4–6]).

The main advantage of the *PH* model is that no additional assumptions about the baseline hazard function are needed to evaluate the effect of the covariates on the lifetime distribution. However, the basic *PH* assumption may not hold in many practical cases. In such situations, the *AFT* model has proven to be a convenient and intuitive alternative.

The *AFT* model establishes a direct relationship between the failure time and the covariates. The estimation of the model is usually carried out by assuming a parametric distribution for the lifetime. In this paper we refer to this model as parametric *AFT*. Several approaches have been proposed for the estimation of a parametric *AFT* model, see [2,7]. However, such parametric *AFT* model is very restrictive in most cases. As an alternative, semi-parametric models, where no assumptions are specified for the underlying survival distribution, can be more convenient in practice. Ritov [8] studied the general linear square estimation method, rank-based methods for censored data have been proposed by Tsiatis [9], Lai and Ying [10] or Jin et al. [11], and least squares based methods for censored data have been explored, for example, by Miller [12], Buckley and James [13] or Stute [14]. Stute [15,16] considered a semi-parametric *AFT* model and introduced a procedure

* Corresponding author at: Department of Statistics and O.R., Faculty of Sciences, University of Granada, 18071 Granada, Spain. Tel.: +34 958243156; fax: +34 958243267.

E-mail address: mgamiz@ugr.es (M.L. Gámiz-Pérez).

http://dx.doi.org/10.1016/j.apm.2015.03.063
0307-904X/

to estimate the regression coefficients under random censorship. Gross and Lai [17] developed a regression analysis under the presence of left-truncation in addition to right-censoring. Their approach is based on the estimation of a trimmed functional of the survival distribution, given that, with this sampling scheme, the lifetime is only observable within the range between the lower boundary of the support of the truncation variable and the upper boundary of the support of the censoring variable. It leads to estimators of the regression parameters which are relatively simple to obtain and useful to explore the relationship between the response variable and the covariate vector.

The interest in this paper goes further the estimation of the regression parameters since our practical motivation is to evaluate the dependability measures of a water supply system laid in a city of the Mediterranean Sea. So we aim to provide a reliable estimation of the probability of survival of a particular pipe beyond a specific period of time. To this goal we propose the use of nonparametric tools for evaluating the risk of failure in the water supply system. With the term "nonparametric" we mean that we do not consider any particular family of distributions for the underlying lifetime. In other words, a semi-parametric *AFT* that directly links the failure time of a pipe to its particular characteristics has been taken into account.

We follow a sequential procedure. Firstly, we consider the methods suggested by Gross and Lai [17] to estimate the parameters involved in the regression problem. These estimates are used to transform the data into the baseline scale of time. Then we conduct a nonparametric procedure to estimate the baseline survival function. Finally, a back transformation provides the estimator of the survival function for a specific subject.

The baseline survival function is estimated using a weighted-least-squares minimization approach that provides a local linear estimator. The estimator is closely related to the hazard estimator suggested by Nielsen and Tanggaard [18] and the density estimator by Nielsen et al. [19]. In these papers, the authors consider a counting process formulation to represent survival data. The advantage of this formulation is that complicated truncation and/or censoring schemes may be easily incorporated to the model. We adopt this point of view in this paper since the dataset that we analyze consists of breakdown data of sections of pipe of the system, where left-truncation and right-censoring are present.

The rest of the paper is structured as follows. In Section 2 it is described the model under a counting process formulation which comprises the important cases of left-truncation or/and right-censoring. Our proposal is fully described in Section 3. The sequential procedure to derive the semi-parametric estimator of the survival function is introduced in Section 3.1. Section 3.2 describes the estimation of the regression coefficients in the semi-parametric model. Section 3.3 describes the nonparametric local linear estimator of the baseline survival. The asymptotic properties of the local linear estimator are derived in Appendix A. Section 4 shows a simulation study to report the performance of the proposed model and methods. The analysis of the water supply system dataset is described in Section 5. Section 6 concludes the paper.

2. The model

For $i = 1, \dots, n$, let T_i denote the event time for the ith subject. We assume that the subjects are independent. Define $N_i(t)$ as the number of events that have occurred on the ith subject by time t in the absence of filtering. That is, $N_i(t) = I(T_i \leqslant t)$, where $I(\cdot)$ is the indicator function. Also define $Y_i(t)$ as the indicator function that takes the value 1 if the ith individual is at risk at time t, which means that it has not failed and it is under observation. So, $Y^{(n)}(t) = \sum_{i=1}^{n} Y_i(t)$ counts the number of individuals at risk at time t and $\{Y^{(n)}(t), t \geqslant 0\}$ is called the *risk process*.

Suppose that the mean function of the counting process $N_i(t)$ associated with a k-vector of covariates X_i takes the form

$$E\{N_i(t)|X_i\} = Pr\{T_i \leqslant t|X_i\} = \Phi_0(t\exp(-\beta' X_i)), \tag{1}$$

where β is a k-vector of unknown regression parameters, and Φ_0 is an unspecified continuous function. If we write $T_{0,i} = T_i \exp(-\beta' X_i)$ and define $N_{0,i}(t) = I(T_{0,i} \leqslant t)$, then, clearly $N_{0,i}(t) = N_i(t\exp(\beta' X_i))$, and also, under (1) we have that

$$E\{N_{0,i}(t)\} = Pr\{T_{0,i} \leqslant t\} = \Phi_0(t).$$

It means that the probability of failure by time t at the level $X_i = x$ equals the probability of failure by time $t\exp(-\beta' x)$ at the level $X_i = 0$. In other words, the set of covariates X_i affects the probability of occurrence of the failure by expanding or contracting the time scale on which this event occurs by a multiplicative factor $\exp(-\beta' x)$ relative to that of a zero-valued covariate vector, that is we have the following direct and intuitive relationship

$$T_i = \exp(\beta' X_i) T_{0,i},$$

or, equivalently

$$\log T_i = \beta' X_i + \epsilon_i, \tag{2}$$

where the error terms $\epsilon_i = \log T_{0,i}$, for $i = 1, \dots, n$, have a common distribution with survival function S_0. This function is called the baseline survival and it represents the survival function of a subject at the zero-valued level of covariates.

The important situation of right-censoring and/or left-truncation can be described from the general counting process formulation given above just as particular cases. Many papers in the literature which are relevant to our purposes are described just in the case of right-censoring or both right-censoring and left-truncation (e.g. [15–17]). To facilitate the reference to these previous papers we conclude this section by describing the data formulation commonly used in the case of