Journal of Statistical Planning and Inference 140 (2010) 597 – 615

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Estimation and efficiency with recurrent event data under informative monitoring

Akim Adekpedjou[a,*], Edsel A. Peña[b], Jonathan Quiton[c]

[a] Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO 65409, USA
[b] Department of Statistics, University of South Carolina, Columbia, SC 29208, USA
[c] Department of Mathematics and Computer Science, Western Kentucky University, Bowling Green, KY 42101-1078, USA

ARTICLE INFO

Article history:
Received 15 September 2008
Received in revised form
5 August 2009
Accepted 5 August 2009
Available online 19 August 2009

MSC:
primary 62N05
secondary 62F12

Keywords:
Counting processes
Efficiency comparisons
Exponential inter-event times
Generalized Koziol–Green model
Martingales
Weibull inter-event times

ABSTRACT

This article deals with studies that monitor occurrences of a recurrent event for n subjects or experimental units. It is assumed that the i th unit is monitored over a random period $[0, \tau_i]$. The successive inter-event times $T_{i1}, T_{i2}, \ldots$, are assumed independent of τ_i. The random number of event occurrences over the monitoring period is $K_i = \max\{k \in \{0, 1, 2, \ldots\} : T_{i1} + T_{i2} + \cdots + T_{ik} \leq \tau_i\}$. The T_{ij}'s are assumed to be i.i.d. from an unknown distribution function F which belongs to a parametric family of distributions $\mathscr{C} = \{F(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subset \Re^p\}$. The τ_i's are assumed to be i.i.d. from an unknown distribution function G. The problem of estimating $\boldsymbol{\theta}$, and consequently the distribution F, is considered under the assumption that the τ_i's are informative about the inter-event distribution. Specifically, $1 - G = (1 - F)^\beta$ for some unknown $\beta > 0$, a generalized Koziol–Green [cf., Koziol, J., Green, S., 1976. A Cramer–von Mises statistic for randomly censored data. Biometrika 63, 139–156; Chen, Y., Hollander, M., Langberg, N., 1982. Small-sample results for the Kaplan–Meier estimator. J. Amer. Statist. Assoc. 77, 141–144] model. Asymptotic properties of estimators of $\boldsymbol{\theta}$, β, and F are presented. Efficiencies of estimators of θ and F are ascertained relative to estimators which ignore the informative monitoring aspect. These comparisons reveal the gain in efficiency when the informative structure of the model is exploited. Concrete demonstrations were performed for F exponential and a two-parameter Weibull.

1. Introduction

The parametric, semiparametric, and nonparametric estimation of the distribution function of an event time has been the subject of intense research in past decades, especially in settings where there is at most one observed event (so-called *single-event* settings) during the monitoring period per experimental unit. Among the seminal papers dealing with this problem are those of Kaplan and Meier (1958), Efron (1967), Cox (1972), Breslow and Crowley (1974), Aalen (1978), and Borgan (1984); see the books of Fleming and Harrington (1991), Andersen et al. (1993), Kalbfleisch and Prentice (2002), and Aalen et al. (2008). The situation where the event is recurrent so there could be more than one event occurrence per unit has also been dealt with, albeit not as thoroughly yet as the single-event case. In the recurrent event setting, the estimation problem has been considered by Gill (1980, 1981), Vardi (1982a, b), Wang and Chang (1999), and Peña et al. (2001). Gill (1981) dealt with the problem of nonparametric inference for renewal processes in a life testing setting. Vardi (1982a) presented an algorithm for obtaining the

* Corresponding author.
E-mail addresses: akima@mst.edu (A. Adekpedjou), pena@stat.sc.edu (E.A. Peña), Jonathan.Quiton@wku.edu (J. Quiton).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2009.08.007

maximum likelihood (ML) estimator of the survivor function when the underlying interoccurrence times are arithmetic. Sellke (1988), in the case of a single unit, considered the problem of establishing weak convergence of a Nelson–Aalen-type estimator when the length of the monitoring period increases without bound. Peña et al. (2001) proposed Nelson–Aalen and Kaplan–Meier-type estimators and derived their asymptotic properties when the number of units increases but with the monitoring time for each unit being finite with probability one, in contrast to the setting in Sellke (1988) where the monitoring time increases to infinity.

An important issue that arose in the single-event setting is the impact of an informative censoring mechanism. An analytically tractable informative random censorship model was proposed by Koziol and Green (1976) which assumes proportional hazards for the event time and the censoring time. This model was utilized by Chen et al. (1982) to study exact properties of the Kaplan–Meier estimator. Cheng and Lin (1987) also utilized this model to derive an estimator of the survivor function which exploits the informative censoring structure, and showed that their estimator is more efficient than the Kaplan and Meier (1958) estimator, especially under heavy censoring. Hollander and Peña (1989) also used this Koziol–Green model to obtain a more efficient class of confidence bands for the survivor function.

There are many situations, however, in the engineering, actuarial, biomedical, public health, social and economic sciences, as well as in business, where the event of interest is recurrent. Examples of such events are machine (mechanical or electronic) malfunction, nonlife insurance claim, onset of depression, heart attack, economic recession, marital strife, and commission of a criminal act. In this recurrent event setting, the impact of an informative monitoring period has not been examined extensively. This article is for the purpose of studying the impact of an informative monitoring period especially in the context of efficiency gains and losses in the estimation of the inter-event time parameter and distribution. As pointed out in Peña et al. (2001) and Pena and Hollander (2004), recurrent event data have additional features that require attention in performing statistical inference. Two of these important features are: (i) because of the sum-quota data accrual scheme, the number of observed event occurrences is informative about the inter-event distribution even if G is unrelated to F; and (ii) the variable that right-censors the last inter-event time at the end of the monitoring period is dependent on the previous inter-event times. Thus, there is both informative and dependent censoring in recurrent event data. Because of these additional features for recurrent event data, there is a need to study the additional impact of having a G informative about F in the estimation of F or its parameters, in particular, in the efficiency gain when the informative structure is exploited.

There has been several models that have been proposed to model informative censoring. William (1989) proposed a model where the censoring variable is related to the frailty of the individual. He showed in particular that in the case of exponential frailty the use of the Kaplan–Meier estimator can lead to errors in estimating the survivor probability. Wang et al. (2001) proposed various models where the occurrence of recurrent events is modeled by a subject specific nonstationary Poisson process via a latent variable. Siannis (2004) considered a parametric model where the parameter represents the level of dependence between the failure and the censoring process. In this article we employ a generalization to the recurrent event setting of the model studied in Koziol and Green (1976), the so-called Koziol–Green (KG) model. This KG model has been most utilized in studying efficiency aspects under informative censoring in single-event settings; see for instance Chen et al. (1982) which obtains exact properties of the Kaplan–Meier estimator under this model, and Cheng and Lin (1987) which derives an estimator of the survivor function utilizing the informative structure. We point out that, just as in the case of the single-event setting, the utility of the proposed generalized KG model is not primarily to provide a practical and realistic model, but rather to provide a medium in which to examine analytically properties of inference procedures with recurrent event data.

The major goal of this article is to obtain estimators of the inter-event time distribution and its parameter for this generalized KG model and to ascertain the loss in efficiency if one ignores the informative structure. An outline of this article is as follows. Section 2 introduces relevant processes, describes the generalized KG model and its properties, and develops the estimators. The framework of stochastic processes is adopted to gain generality. Section 3 deals with asymptotic properties of the estimators under the KG model and those estimators derived by ignoring the KG assumption. Section 4 performs efficiency comparisons of the estimators that exploits the informative structure relative to those which were derived ignoring the structure. In particular, the efficiency of a fully nonparametric estimator of the inter-event distribution is examined. Section 5 presents the results of simulation studies which studies small- to moderate-sample properties of estimators for models in which closed-form analytical expressions are not possible, specifically when the inter-event distribution is a two-parameter Weibull. Finally, Section 6 provides some concluding remarks.

2. Model of interest and estimators

2.1. Random entities

All random entities are defined on a basic probability space $(\Omega, \mathscr{F}, P)$. We suppose that there are n subjects in the study. For the i th subject, $\{S_{ij},\ j=1,2,\ldots,\}$ are the successive calendar times of event occurrences, while $\{T_{ij},\ j=1,2,\ldots\}$ are the successive inter-event times. Thus, we have $S_{i0}=0$, $S_{ij}=\sum_{k=1}^{j} T_{ik}$ and $T_{ij}=S_{ij}-S_{i,j-1}$. The T_{ij}'s are assumed to be i.i.d. nonnegative r.v.s. with a common absolutely continuous distribution function F. In this paper we restrict to the i.i.d. inter-event times setting, while the possibly more relevant model for biostatistical applications with correlated inter-event times, specifically with the association induced by frailty components, will be dealt with in a separate paper.