Computational Statistics and Data Analysis 62 (2013) 83–92

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

A class of inference procedures for validating the generalized Koziol–Green model with recurrent events

Akim Adekpedjou [a,*], Russell Stocker [b], Withanage A. De Mel [a]

[a] Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO 65409, United States

[b] Department of Mathematics, Indiana University of Pennsylvania, Indiana, PA 15705, United States

ARTICLE INFO

Article history:
Received 31 May 2012
Received in revised form 16 October 2012
Accepted 20 December 2012
Available online 2 January 2013

Keywords:
Recurrent events
Informative monitoring
Weak convergence
Goodness of fit
Generalized Koziol–Green model
Kolmogorov–Smirnov type test
Empirical processes

ABSTRACT

The problem of validity of a model on the informativeness of the right-censoring random variable on the inter-event time with recurrent events is considered. The generalized Koziol–Green model for recurrent events has been used in the literature to account for informativeness in the estimation of the gap time distribution or the cumulative hazard rate function. No formal procedure for validating such assumption has been developed for a recurrent failure time data. In this manuscript, we propose procedures for assessing the validity of the assumed model with recurrent events. Our tests are based on the scaled difference of two competing estimators of the cumulative hazard rate possessing nice asymptotic properties. Large sample properties of the proposed procedures are presented. The asymptotic results are applied for the construction of χ^2 and Kolmogorov–Smirnov type tests. Results of a simulation study on Type-I error probabilities and powers are presented. The procedures are also applied to real recurrent event data.

1. Introduction

In survival analysis, informative monitoring occurs when the distribution function of the end of the monitoring period random variable is informative about the distribution function of the failure times. The right censored data also known as the random censorship model is defined by a sequence of independent and identically distributed (i.i.d.) pair of random variables (X_i, Y_i), $i = 1, \ldots, n$ where the X_is are the *unique* failure times for each unit and the Y_is are the right-censoring random variables. In the literature, the model is always described by the pair $\{Z_i, \delta_i\}$, $i = 1, \ldots, n$ where $Z_i = X_i \wedge Y_i$ and $\delta_i = I\{X_i \le Y_i\}$. Here the symbols $\wedge$ and $I\{\cdot\}$ indicate minimum and indicator functions respectively. The Y_is can also be viewed as the end of the monitoring period which occurs due to competing risks (risks not related to that under study) or loss to follow up. If the distribution function of the Y_is is related to that of the X_is, then we have informative monitoring which is equivalent to informative censoring in the single event case.

The issue is different with recurrent events. We assume that n independent units are available. Each is monitored over a random period $[0, \tau_i]$ for the occurrence of recurrent events. Denote by $S_{i,j}$, $j = 1, \ldots$ the successive calendar times of event occurrences and by $T_{i,j} = S_{i,j} - S_{i,j-1}$ the inter-occurrence times. The $T_{i,j}$s are assumed to be i.i.d. nonnegative random variables with common absolutely continuous distribution function $F(t) = P(T_{i,j} \le t)$ and cumulative hazard rate function $\Lambda(\cdot) = \int_0^{\cdot} dF(t)/(1 - F(t))$. Furthermore, the τ_is are assumed to be i.i.d. nonnegative random variables with distribution function $G(t) = P(\tau_i \le t)$. Over the monitoring period, $K_i = \max\{k \in \aleph : S_{i,k} \le \tau_i\}$ denote the total number of event

* Corresponding author. Tel.: +1 5733414649.
E-mail addresses: akima@mst.edu (A. Adekpedjou), rstocker@iup.edu (R. Stocker), wadc96@mst.edu (W.A. De Mel).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.12.014

occurrences per unit. Then, the observable random variables over the monitoring period are

$$\mathbf{D}_i = (K_i, \tau_i, T_{i,1}, \ldots, T_{i,K_i}, \tau_i - S_{i,K_i}), \quad i = 1, \ldots, n, \tag{1}$$

where $\tau_i - S_{i,K_i}$ is a right-censoring random variable. We point out that, with recurrent event data, it is obvious that the right-censoring random variable does not coincide with the end of monitoring period τ_i. However, the τ_is may be viewed as the right-censoring variables, though the right-censoring structure is somewhat different from the usual right-censoring structure in the single event settings.

Examining the censoring structure for recurrent events, note that, theoretically, there will always be one right-censored inter-event time per unit. However, as one of the reviewers mentioned, in practice τ_i could coincide with a recurrence. The models we consider are constructed using a counting process formulation that models recurrent event data in a dynamic fashion that is well equipped to handle such situations. If the last event time for unit i coincides with τ_i then $S_{i,K_i} = \tau_i$ leading to $\tau_i - S_{i,K_i} = 0$. Therefore, the counting process for unit i, $N_i^{\dagger}(s) = \sum_{j=1}^{\infty} I\{S_{i,j} \le s \wedge \tau_i\} = \sum_{j=1}^{K_i} I\{S_{i,j} \le s \wedge S_{i,K_i}\}$ easily adapts to this situation. In the computational point of view, there will be no right-censored observation for that particular subject. The potential mistake is that we leave $\tau_i - S_{i,K_i}$ to be right-censored when in fact it is the observable T_{i,K_i+1}. For example, if for the ith unit there were 4 observed events (and the 4th event is coinciding with the end of the observation period), then K_i should be 4 and T_{i,K_i+1} or $T_{i,5}$ is completely unobserved. In general, the random variable $\tau_i - S_{i,K_i}$ is the right-censoring variable for the inter-event time T_{i,K_i+1} which depends on the previous inter-event times through S_{i,K_i}. Since both $T_{i,K_{i+1}}$ and S_{i,K_i} depend on the random variable K_i, they are dependent. More interestingly, observe that T_{i,K_i+1} has a distribution different from $T_{i,1}$ owing to the randomness of K_i. Therefore, $\tau_i - S_{i,K_i}$ is informative about the $T_{i,j}$s. This property of recurrent event data, if not taken into account for statistical inference can lead to biases and/or under or over-estimation of parameters. The severity and consequences of not taking it into account are discussed in Adekpedjou, Peña, and Quiton (2010) (henceforth APQ) and Adekpedjou et al. (2013).

One major concern for researchers is how to model informative monitoring. There have been several models suggested in the literature for dealing with the property. Link (1989) proposed a model where the censoring variable is related to the frailty of the individual. Wang et al. (2001) proposed various models where the occurrence of recurrent events is modeled by a subject specific non-stationary Poisson process via a latent variable. Siannis (2004) considered a parametric model where the parameter represents the level of dependence between the failure and censoring process. In this article, we employ a generalization to recurrent events of the model studied in Koziol and Green (1976), the so-called Koziol–Green (KG) model or proportional hazards model. The generalized KG model (henceforth GKG) for recurrent events postulates that there exists a $\beta > 0$ such that $\bar{G}(t) = \bar{F}(t)^{\beta}$, where $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$ are the survivor functions of F and G respectively. The GKG model is equivalent to $\Lambda_G(\cdot) = \beta\Lambda_F(\cdot)$, where Λ_G and Λ_F are the cumulative hazard rate functions of G and F respectively. In single event settings, the parameter β is referred to as the censoring parameter since $P(\tau_i < T_i) = \beta/(1 + \beta)$ and τ_i right-censors T_i. In contrast, in recurrent events, β determines the length of the monitoring period relative to the inter-event times, and a better interpretation is a monitoring parameter. More details on the GKG model can be found in APQ where it was first introduced.

The KG model has been utilized in studying efficiency aspects under informative censoring in single event settings. Chen et al. (1982) obtained exact properties of the Kaplan–Meier estimator (Kaplan and Meier, 1958) under the KG model, and Cheng and Lin (1987) derived an estimator of the survivor function utilizing the informative structure. In recurrent event settings, the reference is APQ where the GKG model has been used for modeling informative monitoring with recurrent events thereby enabling the derivation of an estimator of the cumulative hazard function and assessing efficiency loss when it is ignored. In both settings, the conclusion that transpired is that ignoring informative censoring/monitoring in the estimation process can lead to loss in statistical efficiency and/or biased estimators. Although the model might seem only of technical relevance, in many applications, such as biomedical studies, it is of substantial importance. See for instance Koziol and Green (1976) where the model was used to develop a Cramér–von Mises type statistic to check cancer deaths among oestrogen patients.

Henze (1993), Herbst (1992) and Kirmani and Dauxois (2004), to name a few, proposed procedures for checking the assumption of the KG model in the single event settings. For a review of some of their proposed procedures, see Kay (1984). To the best of our knowledge, no formal procedures have been proposed in the literature for assessing the validity of the assumption of the GKG model with recurrent gap time data. In this manuscript, we develop procedures for checking the validity of the informative monitoring model based on the difference of two competing estimators of the cumulative hazard rate $\Lambda(t)$. These procedures are based on the asymptotic properties of a scaled difference of these estimators. The asymptotic property of the properly scaled process is used to construct several inferential procedures for validating the aforementioned model.

The content of the article is as follows. In Section 2, we give a brief overview of random entities of interest and the two competing estimators along with some important asymptotic results. This will be followed in Section 3 by the large sample properties of our proposed test statistic. The class of inference procedures based upon the asymptotic properties in Section 3 will be discussed in Section 4. Section 5 presents the results of a simulation study and an application to the bladder cancer recurrent event data given by Byar (1980). We conclude with a discussion of our results and future research topics.