Journal of Statistical Planning and Inference 156 (2015) 48–63

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A general class of semiparametric models for recurrent event data

Akim Adekpedjou [a], Russell Stocker [b,*]

[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 12 February 2014
Received in revised form 27 June 2014
Accepted 24 July 2014
Available online 8 August 2014

Keywords:
Recurrent events
Effective age process
Imperfect repair
Counting process

ABSTRACT

We propose a general class of semiparametric models for analyzing recurrent event data that takes into account the change in age of a unit due to interventions; allows for the possibility of the unit receiving a life supplement after being repaired; and provides a mechanism for researchers to incorporate time-dependent covariates. The class of models includes as special cases many other models that have been proposed for analyzing recurrent event data. Models belonging to the class can be easily generalized and new models can be created to accommodate a variety of practical considerations. A partial maximum likelihood estimator of the regression parameter and a Nelson–Aalen type estimator of the baseline cumulative intensity are given. Asymptotic properties of the estimators are established and the finite sample properties are investigated via a simulation study. The statistical analysis of a real data set is used to illustrate the class of models.

1. Introduction

Recurrent event data is a special case of multivariate lifetime data wherein the events are ordered and of the same nature. It is prevalent in a variety of disciplines and settings including the biomedical, engineering, and social sciences. Examples include the recurrence of tumors in cancer patients, multiple breakdowns of a car's brake system, absenteeism rate of employees, and the recurrence of war and conflict in geographical regions. Due to its high prevalence and importance in many diverse areas, it is essential there exists appropriate statistical methodology to analyze these types of data. This includes the creation of highly flexible models that allow researchers to account for a variety of different conditions that are unique to the structural challenges found in recurrent event data. The asymptotic properties of estimators should be established under a set of mild regularity conditions. This will allow for the creation of accurate statistical inference procedures for a wide range of situations.

Full specification and marginal modeling are commonly used for recurrent event data. Full specification requires a practitioner to make assumptions about the joint distribution of the gap times. Models such as the homogeneous Poisson process, renewal models, and those that assume an effective age process are examples of full specification. These models may be fully parametric or semiparametric (including non-parametric). Marginal modeling requires the researcher to stratify on the kth recurrent event and then assign a marginal probability measure to each stratum. The proponents of these models advocate for their use due to the lack of assumptions on the joint distribution of gap times and the interpretability of the model parameters.

Marginal modeling does have some important disadvantages that need to be considered. Since full specified models contain the marginal models, it could be of particular interest to find a full specified model that is consistent for a particular

* Corresponding author.
E-mail addresses: akima@mst.edu (A. Adekpedjou), rstocker@iup.edu (R. Stocker).

http://dx.doi.org/10.1016/j.jspi.2014.07.012
0378-3758/

marginal model used in practice. This can be a difficult task that may result in several consistent full models for the marginal model being employed. Additionally, the perceived advantage of a non-specified joint distribution for the gap times is called into question when a consistent model is found. Peña et al. (2007) considered a full specified general class of models for recurrent event data that uses a very general effective age process. They analyzed the bladder cancer data set of Byar (1980) and observed that the parameter estimates and associated standard errors for different choices of the effective age process were close to those obtained using the marginal models of Prentice et al. (1981) and Wei et al. (1989). Although a mathematically rigorous connection was not obtained, the results of Peña et al. (2007) do point to particular choices of the effective age process explaining the differences between the models of Prentice et al. (1981) and Wei et al. (1989). This leads to the hypothesis that the full specified models are good approximations to the aforementioned models.

We propose a general class of full specified semiparametric models that uses a general effective age process that describes the effects of interventions applied to units after experiencing a recurrent event and allows for life supplements that increase units' remaining lifetimes. Our process considers imperfect repair that is somewhere between the extremes of perfect and minimal repair. We also incorporate the effects of important time-dependent covariate processes that can substantially impact a unit's baseline survival.

A key component of our class of models is it contains a large number of previously proposed models. A very important set of examples of subsumed models includes the general Cox proportional hazards model that was investigated by Prentice et al. (1981), Lawless (1987), and Aalen and Husebye (1991); the imperfect repair models of Brown and Proschan (1983) and Block et al. (1985); and the non-parametric model of Dorado et al. (1997) that also includes those given by Kijima (1989). Many of the aforementioned models can be generalized to incorporate a variety of considerations found in recurrent event data. These generalizations along with newly created models will often be special cases of our class of models that can be used in a variety of different situations find in both reliability and survival analysis. We provide both estimators and asymptotic properties in a broad framework that can be easily adapted to these other models. Our asymptotic properties are based on a mild set of regularity conditions that hold for a diverse set of models. Additionally, we avoid using martingale approximations or overreaching assumptions about tightness that may be difficult to verify for particular models.

Many other modeling strategies have been proposed for recurrent event data. The texts of Hougaard (2000), Rigdon and Basu (2000), and Cook and Lawless (2007) provide some approaches. The class of models we consider is a special case of those given in Peña and Hollander (2004). Their class of models uses an effective age process; takes into account the effects of accumulating events on a unit; and allows for the inclusion of possibly time-dependent covariates. Peña et al. (2007) proposed estimators of the unknown parameters and baseline functions for the class of models given by Peña and Hollander (2004) under a semiparametric model specification. Finite sampling properties of the estimators were investigated via a computer simulation study. They did not provide the asymptotic properties of the resulting estimators and their simulation studies dealt with a general repair model that does not supersede our class of models. Stocker and Peña (2007) studied the class of models under a fully parametric specification and provided the necessary reformulation of the regularity conditions given in Borgan (1984) in terms of gap time to obtain asymptotic properties.

The article is organized as follows. In Section 2 we describe the class of imperfect repair models and then reformulate them into functions of both calendar and gap time. Using this reformulation, we derive estimators of the regression parameter and the baseline functions. We next derive the asymptotic properties of the estimators in Section 3. In Section 4 we study the finite sample properties of our proposed estimators via a computer simulation study. We also compare the asymptotic approximations obtained in Section 3 to the results of the simulation study. Finally in Section 5 we use the class of models to analyze the load–haul–dump data given in Kumar and Klefsjö (1992).

Before proceeding with the article, we introduce some notation that is used throughout this manuscript. All random entities are assumed to be defined on a complete probability space $(\Omega, \mathcal{F}, \mathbf{P})$. The function $I\{\cdot\}$ will denote the indicator function, so that $I\{A\} = 1$ if A occurs and 0 otherwise. The space $D[0, t]$ denotes the cadlag functions on $[0, t]$ equipped with the supremum norm $\|\cdot\|_\infty$. All asymptotic results are taken as $n \to \infty$ and the notations $\Rightarrow$, $\xrightarrow{d}$, $\xrightarrow{as}$, and $\xrightarrow{p}$ will denote weak convergence, convergence in distribution, almost sure convergence, and convergence in probability respectively. For a vector $\mathbf{a} = (a_1, \ldots, a_q)$, $\mathbf{a}^t$ denotes its transpose and if $\mathbf{b}$ is also a q-dimensional vector, then $\mathbf{a} \otimes \mathbf{b}$ is the $q \times q$ matrix $\mathbf{ab}^t$ with (i, j)th element $a_i b_j$. In addition, $\mathbf{a}^{\otimes 2} = \mathbf{a} \otimes \mathbf{a}$. For a matrix $\mathbf{A}$, $\|\mathbf{A}\| = \sup_{i,j} |A_{i,j}|$ and $\mathrm{Dg}(\mathbf{A})$ is its diagonal matrix. For a vector $\mathbf{a}$, $\|\mathbf{a}\| = \sup_i |a_i|$. Finally, for a q-dimensional vector $\boldsymbol{\beta}$, we define the operator $\nabla_{\boldsymbol{\beta}}$ by $\nabla_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \equiv (\partial/\partial\beta_j, j = 1, 2, \ldots, q)$.

2. The model and associated estimators

2.1. Description of the model

We consider a study with n units where each ith unit is monitored over the interval $[0, \tau_i]$. The τ_is are iid right-censoring random variables with common absolutely continuous distribution function $G(t) = P(\tau_i \le t)$. For the ith unit, we observe events at calendar times $0 \equiv S_{i,0} < S_{i,1} < S_{i,2} < \cdots$. Associated with these calendar times are the interocurrence times, $T_{i,k} = S_{i,k} - S_{i,k-1}$. We assume the τ_is are noninformative about the $S_{i,j}$s and they satisfy the independent censoring condition. For each unit, there is an associated q-dimensional covariate process $(\mathbf{X}_i(s) : 0 \le s \le \tau_i)$ and a random variable $K_i = \max\{k : S_{i,k} \le \tau_i\}$. K_i is the number of recurrent events observed for unit i. Therefore, the observable data for the ith