COMSTA: 6219 Model 3Gsc pp. 1–13 (col. fig: NIL)

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Multivariate frailty models for multi-type recurrent event data and its application to cancer prevention trial

Khaled Bedair[a,b], Yili Hong[c,*], Jie Li[c], Hussein R. Al-Khalidi[d]

[a] *Faculty of Commerce, Tanta University, Tanta, Egypt*
[b] *School of Medicine, University of Dundee, Dundee, UK*
[c] *Department of Statistics, Virginia Tech, Blacksburg, VA, USA*
[d] *Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA*

ARTICLE INFO

Article history:
Received 2 February 2015
Received in revised form 15 October 2015
Accepted 29 January 2016
Available online xxxx

Keywords:
Correlated frailty
MCMC
Proportional hazards model
Random effects
EM algorithm
Skin cancer

ABSTRACT

Multi-type recurrent event data arise in many situations when two or more different event types may occur repeatedly over an observation period. For example, in a randomized controlled clinical trial to study the efficacy of nutritional supplements for skin cancer prevention, there can be two types of skin cancer events occur repeatedly over time. The research objectives of analyzing such data often include characterizing the event rate of different event types, estimating the treatment effects on each event process, and understanding the correlation structure among different event types. In this paper, we propose the use of a proportional intensity model with multivariate random effects to model such data. The proposed model can take into account the dependence among different event types within a subject as well as the treatment effects. Maximum likelihood estimates of the regression coefficients, variance–covariance components, and the nonparametric baseline intensity function are obtained via a Monte Carlo Expectation–Maximization (MCEM) algorithm. The expectation step of the algorithm involves the calculation of the conditional expectations of the random effects by using the Metropolis–Hastings sampling. Our proposed method can easily handle recurrent event data that have more than two types of events. Simulation studies were used to validate the performance of the proposed method, followed by an application to the skin cancer prevention data.

1. Introduction

In many research fields, it is common to observe processes that generate events repeatedly over the follow-up time for a given subject. Such processes are called recurrent event processes and the generated data are referred to as recurrent event data. In clinical studies, patients may experience transient clinical events repeatedly over an observation period, such as occurrences of heart attack in cardiovascular studies, epileptic seizures in neurology studies, fractures in osteoporosis studies, and recurrence of bladder cancer tumors in oncology studies. Multi-type recurrent event data arise when two or more different kinds of events may occur repeatedly over an observation period. For example, in bone marrow transplantation, different types of recurrent infections (e.g., bacterial, fungal, and viral infections) can occur after the surgery.

* Corresponding author.
E-mail address: yilihong@vt.edu (Y. Hong).

http://dx.doi.org/10.1016/j.csda.2016.01.018

Please cite this article in press as: Bedair, K., et al., Multivariate frailty models for multi-type recurrent event data and its application to cancer prevention trial. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.01.018

The analyses of such multi-type recurrent event data often aim to answer scientific questions such as: what are the effects of explanatory variables (e.g., treatment) on the event process of different types, how to characterize the individual-to-individual difference in the event processes, and what is the correlation among the event process of different types. For example, in the Nutritional Prevention of Cancer Trial conducted by Arizona Cancer Center (see Section 5 for more details), the investigators were interested in the efficacy of nutritional supplement of selenium on preventing skin cancer of different types, as well as the heterogeneity and correlation among the recurrence of skin cancer of different types. Thus, a general method to analyze multi-type recurrent event data needs to be developed to address these scientific questions.

Multivariate frailty model is widely used to model recurrent event data (e.g., Duchateau et al., 2003, Manda and Meyer, 2005, and McGilchrist and Yau, 2008). Although there has been many developments on recurrent event data analysis, existing methods are not general or flexible enough to model and characterize multi-type recurrent event data. In this paper, we propose the use of a semiparametric proportional hazards model with correlated random effect to characterize multi-type recurrent event data. Monte Carlo Expectation–Maximization (MCEM) algorithm is used to estimate unknown parameters in the model.

For a literature review, a general description of models for recurrent event data can be found in literature, for example, Cook and Lawless (2007). In terms of baseline modeling, Abu-Libdeh et al. (1990), and Cook et al. (1999) considered a frailty model with parametric baseline functions. Chen et al. (2005), Moreno (2008), and Chen and Cook (2009) used a piecewise constant baseline intensity function to analyze multi-type recurrent event data. For piece-wise constant function models, one needs to specify the locations and number of pieces for the intensity function. As it is pointed out by Friedman (1982), the estimator will be biased if the locations and the number of pieces are not correctly specified. Thus, in comparison to the parametric and piecewise constant assumptions, the nonparametric baseline assumption used in this paper is more flexible.

For modeling of multiple types of events, Cai and Schaubel (2004) proposed a class of semi-parametric marginal means/rates models for multiple type recurrent event data, without random effects. Cook et al. (2010) introduced a bivariate mixed Poisson model using a copula function to describe the correlation between frailties of the two types of events. They applied the Expectation–Maximization (EM) algorithm and used numerical integration in the expectation step (E step) to calculate the conditional expectations. However, numerical integration can be unstable, even when the number of event types is two. When the number of event types is more than two, it is challenging to make numerical integration work. The Monte Carlo sampling used in our method is more flexible than numerical integration in dealing with multi-type events, even for large number of event types. Rondeau et al. (2012) developed an R package for the analysis of correlated survival data with frailty models. Mazroui et al. (2013) considered multivariate frailty models for two types of recurrent events with parametric baseline and spline based baseline functions. Mazroui et al. (2015) considered multivariate frailty models for two types of recurrent events with time-varying coefficients.

Although MCEM algorithms were used in parameter estimation for frailty models of single-type recurrent event data analysis (Vaida and Xu, 2000; Ripatti et al., 2002), the generalization of the MCEM algorithm to multi-type recurrent event data is not a trivial task. The challenges mainly arise from the need to estimate multiple baseline functions, correlated frailty, and parameters in the frailty distribution, which are incorporated in this paper. In summary, the proposed method provides a flexible and general approach that can be directly implemented to analyze multi-type recurrent event data.

The remainder of this paper is organized as follows. Section 2 describes the data setup, the model for the intensity function, and the model for the random effects. Section 3 describes the MCEM algorithm that is used to estimate the unknown parameter in the model and statistical inference procedures. Section 4 uses simulations to validate the estimation and inference procedure developed in this paper. Section 5 applies the proposed method to the skin cancer dataset. Section 6 gives some concluding remarks and areas for further research.

2. Data and model

In this section, we introduce the notation for the data and the models to describe the data.

2.1. Data

The kth event time for subject i of event type j is denoted by t_{ijk}, $i = 1, \ldots, m$, $j = 1, \ldots, J$, and $k = 1, \ldots, N_{ij}(\tau_i)$. Here m is the number of subjects under the study, J is the total number of event types, τ_i is the length of follow-up time for subject i, and $N_{ij}(t)$ is defined to be the number of type j events occurred over time interval $(0, t]$ for subject i. The censoring indicator δ_{ijk} equals to 1 if event type j is observed for subject i at time t_{ijk}, and $\delta_{ijk} = 0$ otherwise. We also have information on covariates for subject i, denoted by $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, where p is the number of covariates. Let $\boldsymbol{t}_i$ denote the data for subject i which includes observed event times, censoring indicator and the covariates. We use $\boldsymbol{t} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_m\}$ to denote the dataset for all subjects.

2.2. Model for event intensities

Let $Y_i(t) = I(t \le \tau_i)$ be the at-risk process for subject i, $i = 1, \ldots, m$. The event history for subject i up to time t is defined as $\mathcal{H}_i(t) = \left\{N_{ij}(s), \boldsymbol{x}_i, Y_i(s); j = 1, \ldots, J, 0 \le s < t\right\}$. The intensity function for the type j events of subject i is defined as