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

A Gamma-frailty proportional hazards model for bivariate interval-censored data

Prabhashi W. Withana Gamage [a], Christopher S. McMahan [a,*], Lianming Wang [b], Wanzhu Tu [c]

[a] Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, USA
[b] Department of Statistics, University of South Carolina, SC 29208, USA
[c] Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

ARTICLE INFO

Article history:
Received 16 June 2017
Received in revised form 11 April 2018
Accepted 26 July 2018
Available online xxxx

Keywords:
EM algorithm
Gamma-frailty
Interval-censored data
Monotone splines
Multivariate regression
Poisson latent variables
Proportional hazards model
Survival analysis

ABSTRACT

Correlated survival data naturally arise from many clinical and epidemiological studies. For the analysis of such data, the Gamma-frailty proportional hazards (PH) model is a popular choice because the regression parameters have marginal interpretations and the statistical association between the failure times can be explicitly quantified via Kendall's tau. Despite their popularity, Gamma-frailty PH models for correlated interval-censored data have not received as much attention as analogous models for right-censored data. In this work, a Gamma-frailty PH model for bivariate interval-censored data is presented and an easy to implement expectation–maximization (EM) algorithm for model fitting is developed. The proposed model adopts a monotone spline representation for the purposes of approximating the unknown conditional cumulative baseline hazard functions, significantly reducing the number of unknown parameters while retaining modeling flexibility. The EM algorithm was derived from a data augmentation procedure involving latent Poisson random variables. Extensive numerical studies illustrate that the proposed method can provide reliable estimation and valid inference, and is moreover robust to the misspecification of the frailty distribution. To further illustrate its use, the proposed method is used to analyze data from an epidemiological study of sexually transmitted infections.

1. Introduction

Interval-censored data frequently arise from clinical and epidemiological studies, where outcome events are periodically assessed. In studies of sexually transmitted infections (STIs), for example, participants are often followed prospectively with predetermined testing schedules. As a result, the precise timing of infection acquisition is generally unavailable, except for the rare situations where tests are prompted by emergence of symptoms. Interval-censored data are particularly common in investigation of STIs with no or mild symptoms. For example, the motivating example considered herein involves a cohort study of young women aimed at assessing the association between certain risk factors and the contraction of STIs. In particular, this study considers *Chlamydia trachomatis* and *Trichomonas vaginalis*, two organisms that cause clinical diseases of chlamydia and trichomoniasis, respectively. Moreover, individuals infected with *C. trachomatis* and *T. vaginalis* can often be asymptomatic, thus preventing knowledge of the exact acquisition time. Herein, a joint modeling approach to accommodate the known synergy between these two pathogens (Workowski and Bolan, 2015) is developed. The primary objectives of

* Corresponding author.
E-mail address: mcmaha2@clemson.edu (C.S. McMahan).

https://doi.org/10.1016/j.csda.2018.07.016
0167-9473/

this analysis are to estimate the organism-specific survival functions, and quantify the associations between participant characteristics and risks of STI acquisition.

For correlated survival times, there are two basic modeling approaches; i.e., marginal or frailty modeling. The marginal approach specifies a marginal model for each failure time, adopts a working independence assumption in the likelihood construction, obtains point estimates of the regression parameters under this assumption, and then uses the so-called sandwich estimator to obtain standard error estimates (Wei et al., 1989). Various marginal models have been proposed along the lines of this general approach for multivariate interval-censored data; e.g., the proportional hazards (PH) model (Goggins and Finkelstein, 2000; Kim and Xue, 2002), the proportional odds (PO) model (Chen et al., 2007), the additive hazards model (Tong et al., 2008), the linear transformation model (Chen et al., 2013), and the additive transformation model (Shen, 2015). Moreover, a goodness-of-fit test for assessing the appropriateness of the marginal Cox model for multivariate interval-censored data was proposed by Wang et al. (2006). Even though the marginal approach provides robust inference, it does not adequately account for the correlation that naturally exists between the multiple failure times.

In contrast, frailty models directly acknowledge the correlation structure and introduce frailty terms in order to model the dependence between multiple responses. For this reason frailty modeling has become quite popular in survival analysis (Hougaard, 2000; Ibrahim et al., 2008; Wienke, 2012). For analyzing multivariate case 1 interval-censored data (i.e., current status data), several frailty models have been previously proposed; e.g., a probit model with normal frailty (Dunson and Dinse, 2002), a PH model with a normal frailty (Chen et al., 2009), and a PO model with a gamma-frailty (Lin and Wang, 2011). Extending to multivariate general interval-censored data, Komarek and Lessaffre (2007) proposed a frailty accelerated failure time model and Zuma (2007) explored the Gamma-frailty Weibull model.

For the analysis of correlated survival data, the Gamma-frailty proportional hazards (PH) model has proven to be a popular choice among practitioners. One advantage of this model is that the statistical association between the failure times can be explicitly (in closed-form) quantified via Kendall's τ. Research based on the Gamma-frailty PH model for multivariate right-censored data include (Klein, 1992; Andersen et al., 1997; Rondeau et al., 2003; Cui and Sun, 2004; Yin and Ibrahim, 2005) among many others. Related work on multivariate or clustered current status data include (Chang et al., 2007; Hens et al., 2009; Wen and Chen, 2011; Wang et al., 2015). In contrast, very few works have considered extending the Gamma-frailty PH model to allow for the analysis of multivariate interval-censored data, within the context studied here. For analyzing clustered interval-censored data, Lam et al. (2010) proposes a multiple imputation approach under the Gamma-frailty PH model. Similarly, Henschel et al. (2009) and Yavuz and Lambert (2016) propose frailty models for clustered interval-censored data within a Bayesian framework. To our knowledge, the work most closely related to that presented here is of Wen and Chen (2013). These authors developed an algorithm which could be used to maximize the full likelihood based on the Gamma-frailty PH model and established the asymptotic properties of their proposed estimator. However, the proposed algorithm is rather arduous to implement, even for experts in the area, and software is not readily available. In particular, the algorithm involves iteratively updating the regression parameters and the frailty variance parameter through a Newton–Raphson algorithm and solving self-consistency equations for the conditional cumulative baseline hazard functions.

Seeking to generalize (Wang et al., 2015), this paper focuses on developing methods for analyzing correlated bivariate interval-censored data under the Gamma-frailty PH model. In the proposed model formulation, a monotone spline representation (Ramsay, 1988) is used to approximate the unknown conditional cumulative baseline hazard functions, thus greatly reducing the number of unknown parameters while retaining a great deal of modeling flexibility. To complete model fitting, an expectation–maximization (EM) algorithm is developed through a carefully structured data augmentation scheme involving latent Poisson random variables. This scheme leads to both straightforward parameter updates in the M-step as well as closed-form expectations in the E-step. These features make the algorithm easy to implement and computationally efficient. Moreover, through an extensive Monte Carlo simulation study, the proposed approach is shown to provide reliable estimates of all model parameters as well as valid inference, and further, is robust to the misspecification of the frailty distribution. As a companion to this work, a set of functions (coded in R) which implement all aspects of the proposed methodology have been developed and are being added to the next release of the `ICsurv` package, which is freely available from the CRAN (i.e., http://cran.us.r-project.org/).

The remainder of this article is organized as follows. In Section 2, the details of the proposed model and approach are presented, including but not limited to the use of monotone splines, the data augmentation steps, and the derivation of the EM algorithm. In Section 3, the results of an extensive simulation study designed to evaluate the finite sample performance of the proposed approach are provided. Section 4 provides the results of the analysis of the motivating data application; i.e., the STI data collected as a part of the Young Women's Project. Section 5 concludes with a summary discussion.

2. Model and methodology

Let T_1 and T_2 denote the two unobserved failure/event times of interest; e.g., the time at which a patient becomes infected with *C. trachomatis* or *T. vaginalis*, respectively. To jointly model these two failure times, a Gamma-frailty proportional hazards model is considered; i.e., as in Wang et al. (2015) it is assumed that the conditional cumulative hazard function for T_j, given the frailty η, is given by

$$\Lambda_j(t|\mathbf{x}, \eta) = \Lambda_{0j}(t)\exp(\mathbf{x}'\boldsymbol{\beta}_j)\eta, \text{ for } j = 1, 2, \quad (1)$$

Please cite this article in press as: Gamage P.W.W., et al., A Gamma-frailty proportional hazards model for bivariate interval-censored data. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.07.016.