Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 995–1009
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Functional approach of flexibly modelling generalized longitudinal data and survival time

Fang Yao*

Department of Statistics, University of Toronto, 100 St. George Street, Toronto, Ont., Canada M5S 3G3

Received 1 February 2006; received in revised form 20 October 2006; accepted 11 March 2007
Available online 7 June 2007

Abstract

We propose a flexible functional approach for modelling generalized longitudinal data and survival time using principal components. In the proposed model the longitudinal observations can be continuous or categorical data, such as Gaussian, binomial or Poisson outcomes. We generalize the traditional joint models that treat categorical data as continuous data by using some transformations, such as CD4 counts. The proposed model is data-adaptive, which does not require pre-specified functional forms for longitudinal trajectories and automatically detects characteristic patterns. The longitudinal trajectories observed with measurement error or random error are represented by flexible basis functions through a possibly nonlinear link function, combining dimension reduction techniques resulting from functional principal component (FPC) analysis. The relationship between the longitudinal process and event history is assessed using a Cox regression model. Although the proposed model inherits the flexibility of non-parametric methods, the estimation procedure based on the EM algorithm is still parametric in computation, and thus simple and easy to implement. The computation is simplified by dimension reduction for random coefficients or FPC scores. An iterative selection procedure based on Akaike information criterion (AIC) is proposed to choose the tuning parameters, such as the knots of spline basis and the number of FPCs, so that appropriate degree of smoothness and fluctuation can be addressed. The effectiveness of the proposed approach is illustrated through a simulation study, followed by an application to longitudinal CD4 counts and survival data which were collected in a recent clinical trial to compare the efficiency and safety of two antiretroviral drugs.

Keywords: Cox regression; EM algorithm; Functional principal components; Generalized linear mixed effects model; Longitudinal data; Survival

1. Introduction

Many scientific investigations, such as clinical trials, collect longitudinal data with repeated measurements for a sample of subjects, and event history data that are possibly censored time-to-event, i.e., "failure" or "survival". Additional covariate information may be also recorded. A complication that often occurs is that the longitudinal process is usually unobservable due to measurement error or random error. Because of this, if the Cox regression model is used to analyze the survival process, the required longitudinal information at each failure time for all members in the corresponding risk set is not available. It is well known that conventional partial likelihood approaches used for the Cox model cannot avoid biased inference by using some sort of imputation of the latent longitudinal process, such as last-value-carried-forward method (Prentice, 1982), smoothing techniques (Raboud et al., 1993), or "two-stage"

* Tel.: +1 416 9465790; fax: +1 416 9785133.
E-mail address: fyao@utstat.toronto.edu.

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.03.053

approaches (Tsiatis et al., 1995). This invokes the consideration of longitudinal and event processes simultaneously, i.e., the "so-called" joint models, that have attracted substantial research interest and make more efficient use of data by jointly maximizing the likelihood of both processes.

There has been substantial recent work on jointly modelling a continuous longitudinal process and survival history. Typical examples are HIV trials, where a biomarker such as CD4 lymphocyte counts are measured intermittently and time to progression to AIDS or death is recorded, with possible early dropout or failing to experience event by the end of study. Tsiatis et al. (1995), Faucett and Thomas (1996), Wulfsohn and Tsiatis (1997), Bycott and Taylor (1998) and Dafni and Tsiatis (1998) characterized the longitudinal process by parametric random effects models focusing on smooth trends determined by a small number of random effects. Alternative models consisting of random effects and some mean-zero stochastic processes were proposed by Taylor et al. (1994), Henderson et al. (2000), Wang and Taylor (2001) and Xu and Zeger (2001), and investigated "wiggly" fluctuations that may be caused by a biological mechanism.

The research of this paper is motivated by the fact that in many clinical trials longitudinal observations are not necessarily continuous, but may be categorical, such as binomial or Poisson outcomes. For example, in some cancer clinical trials smoking status, a binary longitudinal covariate of interest, is often related to the progression to cancer. The binary observations are obviously subject to random error that yields dichotomous outcomes and plays a significant role to mask the latent process from observed values. In such cases the mean or probability of the binary observations is more appropriate to be used as the covariate process for modelling the survival time instead of the observed dichotomous outcomes. Another important example is the CD4 count data. It is well known that CD4 counts are usually transformed by fourth-root power or logarithm to achieve normality and homogeneity of within-subject variation (Taylor et al., 1991). The transformed data are then modelled as continuous variable using linear mixed effects models, and the inference is usually based on Gaussian assumption of the transformed data. An alternative approach proposed in this paper is to model the original CD4 counts as Poisson outcomes. It is clear that linear mixed effects models may not be adequate for modelling a variety of outcome measures. Therefore it is natural and necessary to extend the regression analysis to a general class of models, and simultaneously take the event process into account.

For the modelling of generalized longitudinal outcomes that can be continuous or categorial outcomes, generalized linear mixed models (GLMMs, see Diggle et al., 2002 for introduction and Section 2.2 for general formulation) are a natural outgrowth of both linear mixed effects models and generalized linear models (McCullagh and Nelder, 1989). They are of wide applicability and practical importance (Breslow and Clayton, 1993). GLMMs enable the accommodation of non-normally distributed outcomes. Specifically they can model within-subject correlation by incorporating random effects for longitudinally measured outcomes. Although GLMM is a rich class of models, its use in practice has been limited by the complexity of the likelihood function. The approaches involving analytical approximation to the likelihood (Goldstein, 1991; Schall, 1991; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) are known to be inconsistent under standard (small domain) asymptotic assumptions (Breslow and Lin, 1995; Lin and Breslow, 1996). McCulloch (1994, 1997) explored Monte Carlo EM algorithms using Gibbs chain and Metroplis–Hastings steps to approximate the E-step, while Zeger and Karim (1991) employed a Gibbs sampling approach. An automated Monte Carlo EM algorithm was developed by Booth and Hobert (1999) that used rejection or importance sampling to simulate random samples in the E-step and yields approximately unbiased estimation.

Due to the numerical challenges of evaluating the intractable integrals in both GLMM and the joint model, there is a lack of research in such a generalized joint model framework. Molenberghs et al. (1997) combined the multivariate Dale model for longitudinal ordinal data with a logistic regression model for drop-out instead of a survival model. Faucett et al. (1998) proposed a joint model to analyze the survival data with binary longitudinal covariate using a Markov model that is not capable of incorporating random effects. Larsen (2005) proposed a joint approach with a two-parameter logistic model that was applied to the Women's Health and Aging Study. In this paper we develop a framework which jointly models generalized longitudinal outcomes and survival time by combining GLMM and Cox regression into a mega-model. The joint models that consider continuous longitudinal covariate can be viewed as a special case. The joint likelihood is maximized using the Monte Carlo EM algorithm which yields approximately unbiased parameter estimation, as illustrated by the simulation study in Section 4.

As a parametric model will only find those features in the data that have been pre-specified, this may not be adequate if the time course is not well defined and does not fall into the preconceived class. In such situation an analysis through semi- or non-parametric methods is advisable. Functional data analysis attracted substantial interest recently for modelling a sample of trajectories semi- or non-parametrically, see Ramsay and Silverman (1997, 2002) for a summary. In particular, functional principal component (FPC) analysis attempts to find the dominant modes of