Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 1998–2016
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Pseudo-full likelihood estimation for prospective survival analysis with a general semiparametric shared frailty model: Asymptotic theory

David M. Zucker[a,*], Malka Gorfine[b], Li Hsu[c]

[a] *Department of Statistics, Hebrew University, Mt. Scopus, Jerusalem 91905, Israel*
[b] *Faculty of Industrial Engineering and Management, Technion, Technion City, Haifa 32000, Israel*
[c] *Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle WA 98109-1024, USA*

Received 9 August 2006; received in revised form 8 August 2007; accepted 15 August 2007
Available online 12 October 2007

Abstract

In this work we present a simple estimation procedure for a general frailty model for analysis of prospective correlated failure times. Earlier work showed this method to perform well in a simulation study. Here we provide rigorous large-sample theory for the proposed estimators of both the regression coefficient vector and the dependence parameter, including consistent variance estimators.

Keywords: Correlated failure times; EM algorithm; Frailty model; Prospective family study; Survival analysis

1. Introduction

Many epidemiological studies involve failure times that are clustered into groups, such as families or schools. Unobserved characteristics shared by members of the same cluster (e.g. genetic information or unmeasured shared environmental exposures) could influence time to the studied event. Frailty models express within-cluster dependence through a shared unobservable random effect. Estimation in the frailty model has received much attention under various frailty distributions, including gamma (Gill, 1985, 1989; Nielsen et al., 1992; Klein, 1992, among others), positive stable (Hougaard, 1986; Fine et al., 2003), inverse Gaussian, compound Poisson (Henderson and Oman, 1999) and log-normal (McGilchrist, 1993; Ripatti and Palmgren, 2000; Vaida and Xu, 2000, among others). Hougaard (2000) provides a comprehensive review of the properties of the various frailty distributions. In a frailty model, the parameters of interest typically are the regression coefficients, the cumulative baseline hazard function, and the dependence parameters in the random effect distribution.

Since the frailties are latent covariates, the expectation–maximization (EM) algorithm is a natural estimation tool, with the latent covariates estimated in the E-step and the likelihood maximized in the M-step after substituting in the

* Corresponding author. Tel.: +972 2 588 1291; fax: +972 2 588 3549.
E-mail addresses: mszucker@mscc.huji.ac.il (D.M. Zucker), gorfinm@ie.technion.ac.il (M. Gorfine), lih@fhcrc.org (L. Hsu).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.08.005

estimated latent quantities. Gill (1985), Nielsen et al. (1992), and Klein (1992) discussed EM-based maximum likelihood estimation for the semiparametric gamma frailty model. One problem with the EM algorithm is that variance estimates for the estimated parameters are not readily available (Louis, 1982; Gill, 1989; Nielsen et al., 1992; Andersen et al., 1997). It has been suggested (Gill, 1989; Nielsen et al., 1992) that a nonparametric information calculation could yield consistent variance estimators. Parner (1998), building on Murphy (1994, 1995), proved the consistency and asymptotic normality of the maximum likelihood estimator in the gamma frailty model. Parner also presented a consistent estimator of the limiting covariance matrix of the estimator, based on inverting a discrete observed information matrix. He noted that since the dimension of the observed information matrix grows with the number of observed survival times, inverting the matrix is practically infeasible for a large data set with many distinct failure times. He therefore suggested an alternate approach to estimating the covariance, based on solving a discrete version of a second order Sturm–Liouville equation, along the lines of Bickel (1985). This covariance estimator requires less computational effort, but still is not so simple to implement.

We (Gorfine et al., 2006) developed a new method that can handle any parametric frailty distribution with finite moments. Nonconjugate frailty distributions can be handled by a simple univariate numerical integration over the frailty distribution. Our new method possesses a number of desirable properties: a noniterative procedure for estimating the cumulative hazard function; consistency and asymptotic normality of the parameter estimates; a direct consistent covariance estimator; and easy computation and implementation. The method was found to perform well in a simulation study and the results are very similar to those of the EM-based method. Indeed, on a data set-by-data set basis, the correlation between our estimator and the EM estimator was found to be 95% for the covariate regression parameter and 98–99% for the within-cluster dependence parameter. The purpose of the current paper is to present in detail the theoretical justification for the method.

Our technical approach resembles that of Bagdonavicius and Nikulin (1999) and Dabrowska (2006a, b). These works, however, dealt with a univariate data context, whereas we deal with a clustered data context. Dabrowska works with a transformation model with unknown transformation. She discusses the univariate gamma frailty model, but assumes that the shape parameter of the frailty distribution is known. Indeed, as discussed in Dabrowska (2006a, pp. 147–148), identifiability problems arise in the univariate gamma frailty model with unknown shape parameter when an unknown transformation is involved. In fact, even when the transformation is known, if there are no covariate effects on the hazard rate (i.e. in the model (1), the regression parameter vector $\boldsymbol{\beta}$ is equal to zero), the shape parameter cannot be identified from univariate data (Lancaster and Nickell, 1980). In our setting, there is no unknown transformation, and we have clustered data. In this case, the shape parameter is identifiable irrespective of whether $\boldsymbol{\beta}$ is zero or nonzero. In our work, we are specifically interested in estimating the shape parameter, which expresses the within-cluster dependence. In genetic research and other contexts, this cluster dependence parameter is itself of significant scientific interest, because it provides insight into the impact of genetic and environmental factors on the disease incidence. Dabrowska (2006b) discusses a one-step method for converting a consistent estimator into a semiparametric efficient estimator. In principle, this approach could be applied to our estimator as well. In our simulations, however, we found that our estimator was comparable in efficiency to the full nonparametric MLE. Thus, although our estimator is not theoretically semiparametric efficient, in practical terms it closely approaches semiparametric efficiency.

The plan of the paper is as follows. Section 2 presents the estimation procedure. Section 3 presents the consistency and asymptotic normality results, along with the covariance estimator for the parameter estimates. Section 4 presents a simulation study. Section 5 presents the technical conditions required for our theoretical results and the proofs of these results. The proofs are patterned after Zucker (2005), but with a number of significant differences, which are described at the beginning of Section 5.

2. The proposed approach

Consider n families, with family i containing m_i members, $i = 1, \ldots, n$. Following Parner (1998, p.187), we regard m_i as a random variable over $\{1, \ldots, m\}$ for some m, and build up the remainder of the model conditional on m_i. Let T_{ij}^0 and C_{ij} denote the failure and censoring times, respectively, for individual ij. The observed follow-up time is $T_{ij} = \min(T_{ij}^0, C_{ij})$, and the failure indicator is $\delta_{ij} = I(T_{ij}^0 \leqslant C_{ij})$. On each individual, we observe a p-vector of covariates $\mathbf{Z}_{ij}$. In addition, we associate with family i an unobservable family-level covariate W_i, the "frailty", which induces dependence among family members. The conditional hazard function for individual ij, given the family frailty