Available online at www.sciencedirect.com
SCIENCE DIRECT®

Computational Statistics & Data Analysis 49 (2005) 169 – 186

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Penalized spline smoothing in multivariable survival models with varying coefficients

Göran Kauermann*

Department of Economics, Universität Bielefeld, Postfach 10031, 33501 Bielefeld, Germany

Received 28 July 2003; received in revised form 5 May 2004; accepted 6 May 2004

Abstract

Penalized spline (P-spline) smoothing is discussed for hazard regression of multivariable survival data. Non-proportional hazard functions are fitted in a numerically handy manner by employing Poisson regression which results from numerical integration of the cumulative hazard function. Multivariate smoothing parameters are selected by utilizing the connection between P-spline smoothing and generalized linear mixed models. A hybrid routine is suggested which combines the mixed model idea with a classical Akaike information criteria. The model is evaluated with simulations and applied to data on the success and failure of newly founded companies.

Keywords: Generalized linear mixed model; P-spline smoothing; Survival model; Varying coefficient model

1. Introduction

Modeling of survival data is largely dominated by the proportional hazard (PH) model introduced by Cox (1972). Even though the PH model appeals by simple numerical fitting based on the partial likelihood, the PH assumption often restricts the model in applications since it means that covariate effects remain constant over survival time. This assumption has been under major investigation and numerous papers suggest extensions and testing procedures, see for instance O'Sullivan (1988), O'Quigley and Pessione (1989), Hastie and Tibshirani (1990), Gray (1994), Hess (1994) or Abrahamowicz et al. (1996). For a general overview of estimation and tests in proportional hazard models, we also refer to Lin

* Corresponding author. Tel.: +495211066930; fax: +4952110689004
E-mail address: gkauermann@wiwi.uni-bielefeld.de (G. Kauermann).

doi:10.1016/j.csda.2004.05.006

and Wei (1991), Sasieni (1999) or Grambsch and Therneau (2000). Allowing covariate effects to be dynamic in time leads to a varying coefficient model as generally introduced by Hastie and Tibshirani (1993). Here, constant covariate effects are replaced by smooth but unknown functions. Smooth estimation can then be carried out using e.g. spline fitting, as in Hastie and Tibshirani (1993), see also Kooperberg et al. (1995) or by applying local techniques, see e.g. Fan et al. (1997) or Cai and Sun (2003).

Smooth estimation in survival models is usually based on the partial likelihood function. There are, however, two points of criticism which should be raised against the use of the partial likelihood in the context of smoothing. First, in the simple case that covariate effects are in fact constant over time, that is if the PH assumption holds, the cumulative (integrated) hazard function in the likelihood function factorizes to the cumulative baseline hazard multiplied by the covariate effects. If the baseline hazard is then estimated by the empirical survivor function, the resulting profile likelihood for the parameters is equivalent to the partial likelihood suggested by Cox. This justification of the partial likelihood is due to Breslow (1972) (see also Cox, 1975 or Wong, 1986). However, if covariate effects do vary with time, that is if the PH assumption is violated, such factorization of the cumulative hazard does not exist and, consequently, the partial likelihood does not have any justification as profile likelihood function. Secondly, in partial likelihood estimation the baseline hazard is treated as nuisance component and not explicitly estimated. In applications, however, knowledge about the baseline hazard can be of interest, in particular if smooth, non-parametric regression is pursued. For this reason, it seems worthwhile to work directly with the likelihood function. This approach is pursued in this paper in order to fit a smooth, non-proportional hazard model. The integrated hazard function in the likelihood is thereby approximated using numerical integration based on a trapezoid approximation. This in turn leads to a simple likelihood function which resembles a Poisson model.

As smoothing technique we employ penalized spline fitting (*P*-spline). The approach was originally introduced by O’Sullivan (1986), but the procedure finally achieved general recognition with the paper by Eilers and Marx (1996). A comprehensive overview about the current state of the art is found in Ruppert et al. (2003). *P*-spline smoothing in survival models has been studied in Cai et al. (2002) for baseline hazard smoothing. The underlying idea of *P*-spline smoothing is to fit a smooth curve by using a high-dimensional basis. But instead of simple parametric fitting a penalized version is pursued to provide a smooth fit. The approach resembles standard spline smoothing as discussed, e.g. in Wahba (1978), or in its generalized form in Green and Silverman (1994). The major difference is that for spline smoothing the dimension of the corresponding spline basis grows with the sample size. In contrast, for *P*-spline smoothing a finite-dimensional basis is used, where the dimension is chosen in a rich and generous manner. The approach is numerically very handy. It also has strong links to linear mixed models (see Wand, 2003) and to penalized quasi-likelihood (PQL) estimation in generalized linear mixed models (GLMM), as discussed in Breslow and Clayton (1993). The connection becomes obvious if the penalty is rewritten as a priori distribution on the coefficients of the basis. In fact, the smoothing parameter steering the amount of penalization is then playing the role of the a priori variance in the resulting GLMM. We utilize this link for smoothing parameter estimation. It will be demonstrated that the PQL approach is numerically simple but fails to estimate reasonable smoothing