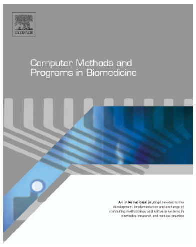

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

The Survival Kit: Software to analyze survival data including possibly correlated random effects

G. Mészáros[a,*], J. Sölkner[a], V. Ducrocq[b]

[a] Division of Livestock Sciences, University of Natural Resources and Life Sciences, Vienna, Gregor-Mendel-Str. 33, A-1180 Vienna, Austria
[b] INRA, UMR 1313 Génétique Animale et Biologie Intégrative, F-78352 Jouy-en-Josas, France

ARTICLE INFO

Article history:
Received 17 November 2011
Received in revised form
10 January 2013
Accepted 13 January 2013

Keywords:
Survival analysis
Proportional hazards
Frailty model
Correlated random effects

ABSTRACT

The Survival Kit is a Fortran 90 Software intended for survival analysis using proportional hazards models and their extension to frailty models with a single response time. The hazard function is described as the product of a baseline hazard function and a positive (exponential) function of possibly time-dependent fixed and random covariates. Stratified Cox, grouped data and Weibull models can be used. Random effects can be either log-gamma or normally distributed and can account for a pedigree structure. Variance parameters are estimated in a Bayesian context. It is possible to account for the correlated nature of two random effects either by specifying a known correlation coefficient or estimating it from the data. An R interface of the Survival Kit provides a user friendly way to run the software.

1. Introduction

The most popular class of survival models is the class of proportional hazard models [1,2], where the hazard of an individual at time t is described as the product of the baseline function and of a positive term which is an exponential function of a vector of covariates $\mathbf{w}'$ multiplied by vector of regression parameters $\boldsymbol{\theta}$. Frailty models are an extension of standard survival analysis models which allows to account for unobserved random heterogeneity [3] or equivalently, to include random effects. These account for an unobserved environmental or genetic effect affecting the hazard of the individual. When two random effects are included (e.g., [4]), these can be independent from each other or related to some degree, leading to the need to estimate correlated random effects. Analyses failing to account for this underlying correlation in survival times are likely to underestimate the variances of parameters [5].

The aim of this paper is to introduce “the Survival Kit”, software for survival analysis capable to handle very large amounts of data, with the possibility to account for their right censored or left truncated status in proportional hazards models. The fixed, random and stratification variables can be time dependent. The estimation of variance components is done in a Bayesian framework and is based on a Laplace approximation of the marginal posterior density of these parameters, from which a modal point estimate can be obtained. Various modeling possibilities are shown in [4], including stratified and frailty survival models with simultaneous estimation of variances for two random effects, center and interaction of treatment by center. The first random effect corrected for deviation centers from the overall baseline hazard, while the second was to deal with deviation of each center from the overall treatment effect. When required, the first three moments of this posterior density can be estimated and the full posterior density can be approximately constructed and visualized. The program was originally written in Fortran 90 for computational

[*] *Corresponding author*. Tel.: +43 1 47654 3259.
E-mail address: gabor.meszaros@boku.ac.at (G. Mészáros).

0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.01.010

efficiency on very large datasets. An R interface was added to provide easier usage and graphical capabilities.

In Section 2 we present the statistical model. In Section 3, the Survival Kit is described and in Section 4 two illustrative applications are presented, one using real infant mortality data, the other using simulated data, and computing variances and covariance of correlated random effects.

2. Theoretical background and computational methods

This section presents a brief overview of the methods used in the Survival Kit. More detailed information can be found in [4,6].

Proportional hazard models and their extension to include random effects describe the hazard function of each individual $\lambda(t)$ (i.e., its limiting probability of dying at time t, given it is still alive just prior to t) as the product of a baseline hazard function and a positive (exponential) function of explanatory covariates.

The model is specified as:

$$\lambda(t; \mathbf{x}(t), \mathbf{z}(t)) = \lambda_0(t) \exp\{\mathbf{x}(t)'\boldsymbol{\beta} + \mathbf{z}(t)'\mathbf{s}\} \quad (1)$$

where $\boldsymbol{\beta}$ and $\mathbf{s}$ are vectors of fixed regression coefficients and random effects. The second part, $\exp\{\mathbf{x}(t)'\boldsymbol{\beta} + \mathbf{z}(t)'\mathbf{s}\}$, represents a stress-dependent term specific to the animals with fixed covariates $\mathbf{x}$ and random covariates $\mathbf{z}$. Both the fixed and random covariates can be time dependent. Only stepwise functions of time are considered for $\mathbf{x}(t)$ or $\mathbf{z}(t)$, i.e., $\mathbf{x}(t)$ and $\mathbf{z}(t)$ are supposed to remain constant over intervals $[t_i, t_i + 1[$. The first part $\lambda_0(t)$ is the baseline hazard function. It is left unspecified in the Cox model [1] or it can take a parametric form as in Weibull model shown in (2).

$$\lambda_0(t) = \lambda\rho(\lambda t)^{\rho-1} \quad (2)$$

where λ and ρ are the shape and scale parameters of the Weibull distribution [2].

The baseline hazard function can be unique or can differ between groups of individuals. The time scale can be divided into several intervals using stratified models, with specific baseline hazards with a separate origin for each, defining a piecewise (e.g., piecewise Weibull) model. This is useful to evaluate hazards with a repetitive pattern. One example is the modeling of culling in dairy cows which clearly follows a particular within lactation pattern [7].

In case of discrete time scale (i.e. with very few distinct time values), there are often many failures occurring at the same time, leading to “ties” between failure times. In such case, the Cox model is no longer valid: if m failure times are tied at time l and n individuals are at risk just prior to l, the partial likelihood contribution involves a summation over all possible subsets of size m from the n at risk, which makes the choice of a Cox model for the discrete time measures inadequate and computationally demanding. Prentice and Gloeckler [8] proposed another approach, the “grouped data model” based on [9]. They assumed that the actual failure times occur in a number of intervals (e.g., years) $[0 = \tau_0, \tau_1)$, $[\tau_1, \tau_2)$, ... $[\tau_{k-1}, \tau_k)$, ... and that the risk of failure is constant within each interval. All failures occurring in the same interval $[\tau_{k-1}, \tau_k)$ are “grouped”, and the attached failure time is k. They also assumed that censoring occurs at the end of each interval. The estimation procedure they proposed was included in the Survival Kit, using a reparameterization described in [10]. Indeed, it is possible to rewrite the model as an exponential regression model including an additional time-dependent effect changing at the beginning of each new interval (see [10] for details).

Technically, the hyperparameters of the prior distribution of random effects (e.g., genetic variance) are estimated from their marginal posterior density [6]. The latter can be obtained through the exact algebraic integration of the random effect out of the joint posterior density when the random effect is assumed to follow a log-gamma distribution. However this is not possible when a normal (or multivariate normal) distribution is used for random effects, for example genetic effects of related animals. Instead, an approximate integration can be implemented using a Laplace approximation. Then, assuming the hyperparameters known, the estimates of all other parameters are obtained as the mode of their joint posterior density. This maximization is done using a limited memory quasi-Newton approach [11] which only requires the computation of the vector of first derivatives of the function to maximize.

For very large applications and models involving correlated random effects, the quasi-Newton approach may converge very slowly. In this case, a full Newton–Raphson algorithm (using both the first and the second derivatives of the function to maximize) can be used to guarantee convergence in a much smaller number of (computationally more expensive) iterations. Also a combination of both quasi-Newton and full Newton–Raphson algorithms is possible and even advisable when good starting values are not available.

Finally, it is also possible to jointly estimate the variance of two random effects using a derivative free algorithm. A normal distribution can be assumed for each level of both random effects. When individuals are (genetically) related, all relationships can be accounted for, assuming a multivariate normal distribution with a (co)variance matrix proportional to $\mathbf{A}$, their relationship matrix [12]. These random effects can be independent from each other [4], but it is also possible to account for their correlated nature as in [5], for example when they correspond to time-dependent effects, for example two genetic effects influencing differently the trait of interest in early and late life. In this case, the two random effects should have the same number of levels. The variances of the random effects and their correlation coefficient could be specified (in case of availability of good prior estimates) or estimated simultaneously with the program.

3. Computer program

3.1. *General description*

The Survival Kit has been developed since its first release in 1994, gradually adding possibilities of stratification and different model types, notably the possibility to model correlated random effects as its latest feature. It is heavily used mostly in the animal breeding community, demonstrated by over