Survival Kit v6 - a Software Package for Survival Analysis

V. Ducrocq[1], J. Sölkner[2] and *G. Mészáros*[2]

Introduction

The "Survival Kit" was historically intended to fill the gap in the software available to animal breeders who generally tend to analyze large data sets and want to estimate random (genetic) effects. The first version written in Fortran77 was released in 1994 (Ducrocq and Sölkner, 1994) during the 5th WCGALP conference, and later updated to version 3.0 in 1998 (Ducrocq and Sölkner, 1998). One year later, more features were added to v3.12 (Sölkner and Ducrocq, 1999). This version available online has been the most used version so far. In 2004, version 5 was released and shared only with geneticists working on national evaluations for longevity. The new features of v5 were not published officially and mainly dealt with faster convergence for very large data sets and production of by-products of genetic evaluations.
The new Survival Kit version 6 is built on the preceding versions, while keeping all previous functionalities and adding new ones, making full use of the possibilities given by Fortran90 programming language.

General description

The Survival Kit is intended for survival analysis using proportional hazards model with a single response time. These models describe the hazard function of each individual (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 baseline hazard function can be parametric or left unspecified (Cox model, Cox, 1972).
It is possible to leave the baseline hazard function unspecified using the Cox model (Cox, 1972). When the failure time variable is discrete (i.e. has very few values, for example when length of life is measured in number of parities) and many observations with the same failure times, many ties occur between failure times. In such case, the Cox model is not valid any more, but the „grouped data“ approach (Prentice and Gloeckler, 1978) should be used. This is possible with the Survival Kit v3.12 and later (Sölkner and Ducrocq, 1999). If the baseline hazard function follows a Weibull distribution, the computationally more advantageous Weibull model can be used.
Stratification is supported, i.e. the use of different baseline hazard functions for individuals distributed in different strtata. In the case of a Weibull distribution, there is a possibility to define separate origin for each stratum, for example, for each parity, which leads to a piecewise Weibull model.

[1] UMR 1313 Génétique Animale et Biologie Intégrative INRA, F-78352 Jouy-en-Josas, France

[2] Division of Livestock Sciences, Univ. of Natural Resources and Applied Life Sciences Gregor-Mendel-Str. 33, A-1180 Vienna, Austria

The Survival Kit handles any number of fixed (possibly time dependent) continuous or discrete covariates with any number of classes. Records with time dependent effects should include the time of change and the new value of the covariate, as many times as the covariate changes. Consequently, survival records are splitted to so called elementary records covering only the time span from one change to the next.
The size of the database and number of covariates/classes is limited only by the size of the computer memory. The real computational constraint comes when using (possibly time dependent) correlated random effects. In genetic evaluations in livestock, the pedigree information is usually available and desirable to use. It should be noted, that the structure of the pedigree file does *not* restrict the possibilities of genetic models, i.e., they can be modelled separately.
There are several options to include the genetic effect (so called „frailty term“) into the model, all possible with the Survival Kit. When using a sire model, sires are assumed to be solely responsible for the genetic difference among the animals. This could be extended to a sire-maternal grandsire model, when dams of the animals are supposed to be related only via their sire (i.e. through the maternal grandsire of the progeny on which survival time is measured). In case of sire-dam models, both parents account for half of the genetic variance and full-sibs are therefore recognized as being more similar than half-sibs. Using animal models, the animal itself is considered to be responsible for the entire genetic variance. This is the most advanced, but computationally most demanding model, not possible for very large datasets. In case of the sire - maternal grandsire – dam within maternal grandsire models, the sire, maternal grandsire and dam of the animal are included as separate random effects in the evaluation.
For each of these models, the relationship matrix accounting for the correlation between estimated effects – whether they correspond to animals, sires, dams or maternal grandsire – may involve all parents (both male and female parents) or only male parents (sire and maternal grand sire) of the individuals included in the genetic model.
Technically, the hyperparameters of the prior distribution of random effects (e.g., genetic variance) are estimated from their marginal posterior density (Ducrocq and Casella, 1996). The latter is obtained either through the exact algebraic integration of the joint posterior density when the random effect is assumed to follow a log-gamma distribution (incompatible with the use of a relationship matrix), or via a Laplace approximation. Then, assuming the hyperparameters known, the estimates of all other parameters are obtained maximizing the joint posterior density. This is done using a limited memory quasi-Newton approach (Liu and Nocedal, 1989) which only requires the computation of the vector of first derivatives of the function to maximize. If required, the negative Hessian is computed and with the possibility to account for its sparse nature using FSPAK subroutines (Perez-Enciso *et al*., 1994).

New functionalities

For very large applications and models involving correlated random effects, the quasi-Newton approach may converge very slowly. From version 5, a full Newton-Raphson algorithm (using both the first and the second derivatives of the function to maximize) was integrated to the Survival Kit, 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.
The stratification variable might now be time dependent, and it is possible to "restart" the hazard from 0 for each stratum. This eventually leads to a model with a different baseline hazard for each stratum, to the so called piecewise Weibull model.
Approximate animal solutions can be estimated with the Weibull program extending the sire-maternal grandsire model (Ducrocq, 2001). This approximation is implemented using a two-step approach. The sire (or sire-maternal grand-sire) effects are estimated during the first step together with the baseline parameters and the other effects of the model. The second step estimates the remaining part of the additive genetic value of each individual, assuming that all the other parameters are known. This step also requires solutions for the random effects which were integrated out (typically herd-year or herd-year-season effects), which is also possible with v5 and later.
From version 6.0 onwards, it is also possible to estimate the variance of two random effects at the same time using a derivative free algorithm (Chandler, 1991). Normal distribution of the random effects is assumed, with the possibility to include the relationship matrix of the individuals. So far, these random effects are supposed to be independent. Extension to correlated effects is under study (Rondeau *et al.*, 2008).
The Survival Kit is usable on any operation system after compilation of the source code. Special care was given to improve the user friendliness of the software. Unlike the previous releases, version 6 takes advantage of memory allocation feature of Fortran 90.
In the past, it was necessary to recompile all programs after each change in a small file defining parameters. In version 6, once the source code is compiled, the variables defining the database size can be changed with new keywords. Moreover the naming of input files and keywords was changed to a more intuitive form, which also contributes to an easier usage.

Support material

The manual was updated clarifying the existing content and adding a special section dealing with the database structure needed for the Survival Kit. The data preparation in case of time dependent effects was identified as one of the major challenges for the users. In order to ease their labor, a section was added describing the requirements for inclusion of time dependent effects using the so called "triplets".
There are also several support programs, which are not an integral part of the Survival Kit, but might be useful in everyday work. One of these is the data simulation program *simul.f* intended to simulate data as a test database for genetic evaluation. The R (R Development Core Team, 2009) function *sortSKit* serves as an alternative possibility to sort the output files as required by the Cox program. Finally the Java application *SKonstructor* offers a graphical point-and-click opportunity to create the input text files for the Survival Kit.

Availability

The Survival Kit can be freely used (including genetic evaluations) provided its use is being credited. Use it at your own risk. The source code, compiled executable files, manual and support programs can be found at: http://www.nas.boku.ac.at/1897.html or http://www4.jouy.inra.fr/gabi/les-Recherches/Developpement-d-outils.

References

Cox, D.R. (1972) J. R. *Stat. Soc., Series B*, 34: 187-220.

Ducrocq, V., (2001) *Interbull bulletin*, 27: 147-152.

Ducrocq, V., Casella, G. (1996) *Genet. Sel. Evol.*, 28: 505-529.

Ducrocq, V., Sölkner J. (1994) *Proc. 5th World Cong. on Genet. Appl. To Livest. Prod.*, 22: 51-52.

Ducrocq, V., Sölkner J.(1998) *Proc. 6th World Cong. on Genet. Appl. To Livest. Prod.*, 27: 447-448.

Liu, D. C. and Nocedal, J. (1989) *Math. Program.*, 45: 503 - 528.

Perez-Enciso, M., Misztal, I. and Elzo, M. A. (1994) *Proc. 5th World Cong. on Genet. Appl. To Livest. Prod.*, Vol 22: 87-88.

Prentice, R. and Gloeckler, L. (1978) *Biometrics*, 34:57-67.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rondeau,V., Michiels, S., Liquet, B. *et al.* (2008) *Statist. Med.*, 27: 1894-1910.

Sölkner J., Ducrocq, V., (1999) 'Workshop on Genetic Improvement of Functional Traits in Cattle - Longevity', Jouy-en-Josas, France.