The Survival Kit: a tool for analysis of survival data

Johann Sölkner and Vincent Ducrocq
Department of Livestock Science,
University of Agricultural Sciences Vienna,
Gregor-Mendel-Strasse 33; A-1180 Vienna, Austria
and
Station de Génétique Quantitative et Appliquée,
Institut National de la Recherche Agronomique,
78352 Jouy en Josas, France

Abstract

The Survival Kit is a set of programs that allows analysis of data which are measuring time untilsome defined event. Nonlinear Cox and Weibull regression models are featured. The programs are specifically adapted to the needs of animal breeders who tend to use very large data sets and want to estimate random effects. The covariance structure between observations based on genetic relationship can also be included. The programs are useful for analysis of (fixed) effects on longevity or any other time variable as well as for genetic evaluations on a national scale. As mixed or random effects models (called frailty models in survival analysis terminology) are not provided by standard statistical software packages, the Survival Kit can also be useful to researchers outside of animal science.

1 Introduction

Longevity is a trait of increasing interest to animal breeders (Essl, 1998). However, the statistical analysis of survival data is not straightforward for several reasons: : 1) the distribution of survival time is rarely known and in most cases, extremely skewed; 2) for part of the observations, only a lower bound of survival time is known (a phenomenon called censoring) e.g., for individuals still alive at the end of the study period; 3) the independent variables influencing survival time may themselves vary with time (e.g., current milk production, herd size, disease occurrence).

Methods dealing with these problems were developed in several areas, including biomedical statistics and the area of testing physical components (reliability testing). In animal breeding, the first person to use such methods extensively was Smith (1983, see also Smith and Quaas, 1984, Smith and Allaire, 1986). Having both some practical experience with formal methods of survival analysis (Ducrocq ,1987, ,Ducrocq et al., 1988ab, Sölkner ,1988, Sölkner and Essl, 1990) we decided in 1992 to collaborate on the development of a set of programs that would satisfy the needs of animal breeders. There seemed to be scope for such a program as animal breeders tend to use extremely large data sets and want to estimate random effects.

2 The Survival Kit

Although developed byanimal breeders for animal breeders, the programs of the Survival Kit should also be interesting for people from other areas encountering similar problems of large models and random effects. Frailty models (a term used for models including random effects in the area of survival analysis) are not supported by any of the well-known software packages (like SAS or BMDP). To make the Survival Kit user-friendly, commands used in the parameter files mimic the SAS command language. The use of the programs is currently wide-spread and several countries are using or intending to use the program for national genetic evaluation of longevity in cattle. Programs are under continuous development and the current version is 3.1.

2.1 Class of models supported

The models supported by the Survival Kit belong to the following class of univariate

proportional hazards models with a single response time:

$$h(t, \mathbf{x}(t), \mathbf{z}(t)) = h_{0j}(t) \exp(\mathbf{x}(t)'\mathbf{b} + \mathbf{z}(t)'\mathbf{u}) \quad (1)$$

where $h(t, \mathbf{x}(t), \mathbf{z}(t))$ is the hazard function of an individual depending on time *t*, a vector of (possibly) time-dependent fixed covariates **x**(*t*) with corresponding parameter vector **b** and a vector of (possibly) time-dependent random covariates **z**(t) with corresponding parameter vector **u.** For a detailed statistical presentation of the methodology of survival analysis, see Cox, 1972, Prentice and Gloeckler, 1978, Cox and Oakes, 1984, Kalbfleisch and Prentice, 1980, Klein and Moeschberger, 1997.

2.2 Features supported

baseline hazard function: the (possibly stratified) baseline hazard function $h_{0j}(t)$ may either be unspecified or follow a Weibull hazard distribution :

$$h_{0j}(t) = \lambda_j \rho_j (\lambda_j t)^{\rho_j - 1}$$

In the first case, if the failure time variable *t* is continuous, (1) defines a Cox model. Estimates of **b** and **u** are obtained using what is known as a partial likelihood, a part of the full likelihood in which the baseline hazard function does not appear. When the failure time variable is discrete with few categories and many observations with the same failure time (ties), the Cox's partial likelihood approach is no longer valid. Because the baseline hazard function can then be described with few parameters, these can be estimated together with **b** and **u** using an approach due to Prentice and Gloeckler (1978). The second case corresponds to a Weibull model, a common type of regression model that has been shown to be flexible and often adequate for biological data.

fixed covariates: any number of fixed covariates is supported. They may either be discrete (class) variables or continuous. There is no explicit limit to the number of levels of a discrete covariate (the limit will usually be a function of the computer memory available).

An intercept (grand mean) is always implicitly included in the Weibull mode (when there is only one stratum and no covariate specified, this intercept is equal to $\rho \log \lambda$, where ρ and λ are the two Weibull parameters). Covariates may be time-dependent (**x**(*t*)), where the dependency is modelled through "piecewise" constant hazard functions with jumps at times corresponding to calendar dates (e.g., January 1st) or linked to the individual itself (e.g., beginning and end of a disease if the effect of the disease on survival is investigated).

random covariates: the random covariates in vector **u** may be defined to follow a log-gamma or a normal distribution. They may also follow a multivariate normal distribution where the covariance structure between individuals is modelled by the matrix of genetic relationships (a typical application to describe the additive genetic values of individuals in animal breeding). The log-gamma was chosen because exp(**u**) is then gamma distributed, which is a frequent assumption for frailty models. The two parameters of the Gamma distribution are taken to be equal so that the expectation E(exp(**u**)) = 1. With the normal distribution, E(**u**) = 0. The distribution parameters (Gamma parameter for the log-gamma and variance for the normal or multivariate normal distribution) may either be prespecified or estimated alongside with the effects in the model (Ducrocq and Casella, 1996). Several random effects may be specified in the same model but currently only one may involve the relationship matrix. Random effects may also be time-dependent. Although the expression *frailty term* is used to generally describe random effects in survival analysis, it was originally introduced to account for individual heterogeneity of observations. Such a term may also follow one of the distributions mentioned above and is technically treated in the same way as the other random effects.

strata: stratification may be used to separate groups of individuals with different baseline hazard functions $h_{0j}(t)$ with *j* being the group indicator. Together with time-dependent covariates, this is another means of relaxing the required assumption of proportional hazards for all individuals over the total observational period. Only one variable may be chosen as strata variable. The number of strata is not restricted.

In addition to estimation of fixed and random effects, the Survival Kit offers options for calculating asymptotic standard deviations of effects (only for moderate size models, where the matrix of second derivatives may be actually set up), a sequence of likelihood ratio tests and different ways of setting constraints to deal with dependencies in the model. As a special feature, different values of the survivor function may be estimated for individuals with pre-set covariates. This way, it is possible to calculate estimated median survival time (for example) or survival probability to a specified age for any combination of covariate values. Generalised, martingale and deviance residuals (Cox and Snell, 1966, Klein and Moeschberger, 1997) can also be computed.

2.3 The programs

The Survival Kit mainly consists of a set of three Fortran programs, called PREPARE, COX and WEIBULL and a file *parinclu* holding parameter definitions that is included in each of the programs (via a Fortran *include* statement). The package works stand-alone, i.e. does not rely on any subroutines from mathematical subroutine libraries. The optimisation routines used are partly taken from public domain subroutine libraries (Liu and Nocedal, 1989, Perez-Enciso et al., 1994) integrated in the programs.

PREPARE is used to prepare the data for the actual analysis done with either COX or WEIBULL. Data preparation includes recoding of class variables and in the presence of time-dependent covariates splitting up individual records into so-called elementary records with each elementary record covering only the time span from one change in any time-dependent covariate to the next. The recoded file may therefore have many more records than the original one.

The estimation of effects under the proportional hazards model described above is then performed by COX or WEIBULL, depending on whether the baseline hazard function is assumed to be unspecified in the Cox model or it is assumed to follow the two-parameter Weibull hazard distribution. Specifications for both models are similar, but it is computationally easier and less time consuming to estimate the parameters of distributions of the random effects under the Weibull model.

For extremely large applications, e.g., national genetic evaluations, the number of elementary records may become huge (sometimes > 100 millions) when time-dependent covariates are used in the model. In the Survival Kit version 3.0 and higher, modified versions of programs PREPARE and WEIBULL, PREPAREC and WEIBULLC were written using public domain C subroutines for compressing and decompressing data during I/O operations (zlib general purpose compression library, version 1.0.4, J. Gailly and M. Adler, web-page: http://quest.jpl.nasa.gov/zlib/.). Our experience is that the resulting programs are about 3 times slower but compression is extremely efficient since the compressed files may take up to 20 or 30 times less disk space. A similar version for the Cox model was not implemented as it is not really suited for huge applications.

2.4 Hardware requirements

The programs have been written in Fortran 77 and have been tested on PC (using Lahey's Fortran compiler) and on several UNIX platforms. No system routines are used, except a timing subroutine second() for UNIX platform which can be replaced or switched off without any consequence. The size of the program may be varied through changes in parameters affecting the maximum number of records and maximum number of levels of effects to be estimated. These parameters are defined in a single file called *parinclu* and included in each program using a Fortran INCLUDE statement. To make the changes effective (and only then), it is necessary to recompile the programs. Programs PREPAREC and WEIBULLC require the compiled C subroutines of *zlib*. These are supplied with the Survival Kit.

2.5 Latest Changes

In version 3.0, the programs PREPAREC and WEIBULLC were made available for very large applications. A few bugs were corrected and some features (like choice of input/output formats, inclusion of groups of unknown parents, use of left-truncated records with COX) were added.

In version 3.1, the main addition was the possibility to properly analyse *discrete* failure time data. Discrete failure times with very few distinct observed values occur for example when survival time is expressed in years or parities. Then, it becomes more difficult to find proper parametric proportional hazards models and the semi-parametric approach of Cox (Cox model) is no longer suitable (an exact calculation of the partial likelihood is not feasible in general in presence of many "ties"). The approach of Prentice and Gloeckler (1978) for grouped data is much more satisfying: a full likelihood is written, involving a limited number of parameters describing the baseline hazard function. Using a reparameterisation of Prentice and Gloeckler's model, it is possible to transform the grouped data model into a form close to an exponential regression model including a particular time-dependent covariate, with changes at every time point. The PREPARE and WEIBULL programs have been modified to accommodate such a model (note however that the resulting model is *not* a Weibull model).

The use of D6 (ddmmyy) formats with years larger or equal to year 2000 is now possible (option NBUG_2000 in the *parinclu* file).

2.6 Disclaimer

The "Survival Kit" can be freely used for non-commercial purposes provided its use is being credited (Ducrocq and Sölkner, 1994, 1998) and can be freely distributed for use at your own risk. There is no technical support, but questions can be directed to the authors.

Dr Vincent Ducrocq, Station de Génétique Quantitative et Appliquée, Institut National de la Recherche Agronomique, F-78352 Jouy-en-Josas, France. Phone: + 33 1 34 65 22 04, Fax: + 33 1 34 65 22 10,
email:Vincent.Ducrocq@dga.jouy.inra.fr

Dr Johann Sölkner, Universität für Bodenkultur, Gregor-Mendel-Strasse 33, A-1180 Vienna, Austria. Tel: + 43 1 47654 3272, Fax:+43 1 3105175,
email: soelkner@mail.boku.ac.at

The programs and manual can be retrieved on the Web in compressed form at:

http://www.boku.ac.at/nuwi/popgen

References

Cox, D. (1972). Regression models and life tables. J. Royal Stat. Soc., Series B}, 34:187-20.

Cox, D.R. and Oakes, D. (1984) Analysis of survival data. Chapman and Hall, London, UK.

Cox, D. and Snell, E. J. (1966). A general definition of residuals. J. Royal Stat. Soc. , Series B}, 30:248-275.

Ducrocq, V. (1987). An analysis of length of productive life in dairy cattle. Dissertation, Cornell University, Ithaca, New York, USA.

Ducrocq, V. and Casella G. (1996). A Bayesian analysis of mixed survival models. Genet. Sel. Evol., 28: 505-529.

Ducrocq, V.P., Quaas, R.L., Pollak, E.J. and Casella, G. (1988a). Length of productive life of dairy cows. 1: Justification of a Weibull model. J. Dairy Sci. 71:3061-3070.

Ducrocq, V.P., Quaas, R.L., Pollak, E.J. and Casella G. (1988b). Length of productive life of dairy cows. 2: Variance component estimation and sire evaluation. J. Dairy Sci. 71: 3071-3079.

Ducrocq, V. and Sölkner, J. (1994)."The Survival Kit", a FORTRAN package for the analysis of survival data. In: 5th World Cong. Genet. Appl. Livest. Prod., Volume 22, pages 51-52. Dep. Anim. Poultry Sci., Univ. of Guelph, Guelph, Ontario, Canada.

Ducrocq, V. and Sölkner, J. (1998). "The Survival Kit - V3.0, a package for large analyses of survival data. In: 6th World Cong. Genet. Appl. Livest. Prod., Volume 27, pages 447-448. Anim. Genetics and Breeding Unit, Univ. of New England, Armidale, Australia.

Essl, A. (1998). Longevity in dairy cattle breeding: a review. Livest. Prod. Sci., 57:79-89

Kalbfleisch, J.D. and Prentice, R.L. (1980). The statistical analysis of failure time data. John Wiley and sons, New-York, USA.

Klein, J.P. and Moeschberger, M. (1997). Survival analysis. Springer-Verlag, New-York, USA.

Liu, D.C. and Nocedal, J. (1989). On the limited memory {BFGS} method for large scale optimization. Mathematical Programming, 45:503-528.

Perez-Enciso, M., Mizstal, I., and Elzo, M.A. (1994). Fspak: an interface for public

domain sparse matrix subroutine.In: 5th World Cong. Genet. Appl. Livest. Prod., Volume22, pages 87--88. Dep. Anim. Poultry Sci., Univ. of Guelph, Guelph, Ontario, Canada.

Prentice, R. and Gloeckler, L. (1978). Regression analysis of grouped survival data with application to breast cancer data. Biometrics, 34:57-67.

Smith, S.P. (1983). The extension of failure time analysis to problems of animal breeding. Dissertation, Cornell University, Ithaca, New York, USA.

Smith, S.P. and Allaire, F.R. (1986). Analysis of failure time measured on dairy cows: Theoretical considerations in animal breeding. J. Dairy Sci. 69:1156-1165.

Smith, S.P. and Quaas, R.L. (1994). Productive life span of bull progeny groups: failure time analysis. J. Dairy Sci. 67:2999-3007.

Sölkner, J. (1988). Analyse möglicher Ursachen für den Rückgang der Nutzungsdauer der österreichischen Milchkühe. Dissertation Universität für Bodenkultur Wien, Austria.

Sölkner, J.and EssL, A. (1990): Einfluß verschiedener Formen der Anbinde-haltung auf die Nutzungsdauer von Kühen. Züchtungskunde 62: 222-233.