Frailty Models in Survival Analysis

Johann Sölkner

Department of Livestock Sciences, University of Agricultural Sciences, Vienna, Austria

The use of frailty models to account for unobserved individual heterogeneity and other random effects in survival analysis are discussed. After a description of proportional hazards models including random effects, the biological meaning of frailty and the resulting problems for analysis are pointed out. The Survival Kit, a set of Fortran programs for the analysis of univariate frailty models has been developed by V. Ducrocq (Fr) and the author. Features of the Survival Kit are presented.

Keywords: Survival analysis, random effects, frailty model, idividual heterogeneity.

1. Introduction

Survival analysis is a vastly expanding discipline in biostatistics. The common feature of data subjected to survival analysis is that the outcome or response of interest is one (or more) event time(s), the time from some appropriate starting point until some event of interest occurs. Examples for such event ,times" include the lifetimes of machine components in industrial reliability, the duration of strikes or periods of unemployment in economics, the times taken by persons to complete specific tasks in psychological experimentation, the survival of patients in clinical trials and lifetime production of farm animals or price money earned by race horses. Apart from the fact that this definition implies that the observed variable is always positive, two special sources of difficulty are often present: censoring and time-dependent covariates. Censoring means that we are not able to observe an individual over the full time, from the starting point to the event, e.g. because we terminate the study after a limited time period, irrespective of whether all events have been observed or not. The term "time-dependent covariates" describes (observable) factors that influence the time to the event, but may change their values during the period of observation. As an example, smoking is a known factor influencing the risk of lung cancer. A person stops smoking during the period of observation, and starts again after some time. The status of the covariate "smoking" changes its value twice in this case.

The most common model to estimate the effect of factors of influence (covariates) on event times is the proportional hazards model (Cox, 1972). In this paper a review is given on the extension of the proportional hazards model to include random effects, commonly called frailty terms in survival analysis (the idea was introduced by Vaupel et al., 1979). The review will be restricted to univariate survival models with only one response event. In addition, a computer software package to analyse frailty models will be presented (The Survival Kit, Ducrocq and Sölkner, 1994).

2. The model

In development of frailty models we start from the classical proportional hazards model, where the hazard function $h(t, \mathbf{x})$ for an individual associated with a vector of $\mathbf{x}$ is:

$$h(t, \mathbf{x}) = h_0(t)\psi(\mathbf{x}) \qquad (1)$$

The hazard function is composed of two parts, a time dependent function, $h_0(t)$, assumed to be identical for all individuals, called *baseline hazard function*, and a "stress-dependent" multiplier $\psi(\mathbf{x})$, expressing how the covariates in $\mathbf{x}$ modify the hazard (independent of time). As $\psi(\mathbf{x})$ has to be positive for all values of $\mathbf{x}$ (the hazard function is a non-negative function),

$\psi(\mathbf{x})$ is very often represented by $\exp(\mathbf{x}'\mathbf{b})$, so that the most common form of the hazard function is

$$h(t, \mathbf{x}) = h_0(t)\exp(\mathbf{x}'\mathbf{b}) \qquad (2)$$

where **b** is the vector of (regression) parameters to be estimated. This formulation of the model implies that the observed covariate represented in **x** fully determines the hazard function, additional (probably unobservable) covariates are not represented in the model which does not include a residual term. Vaupel et al. (1979) introduced a multiplicative term w to the hazard rate, to account for unobserved population heterogeneity. The model is therefore

$$h(t, \mathbf{x}, w) = h(t, \mathbf{x}) \cdot w \qquad (3)$$

The term w (also called frailty term) varies from individual to individual and is not observable. The distribution of w of the population $G(w)$, must be specified. Again, as the hazard function is non-negative, w must be restricted to non-negative values. Frequently used distributions (e.g. the Gamma distribution) and some of the problems related to the use of specific distributions are discussed by Klein et al. 1992. An alternative way of writing the frailty model stated above is

$$h(t, \mathbf{x}, \varepsilon) = h_0(t)\exp(\mathbf{x}'\mathbf{b} + \varepsilon) \qquad (4)$$

where $\varepsilon = \log(w)$, i.e. ε is log-Gamma distributed when w is assumed to be Gamma distributed. In the same way as unobservable population heterogeneity, effects of a common environment or genetic effects may also be modelled as random effects with a given distribution. The model therefore extends to

$$h(t, \mathbf{x}, \mathbf{z}, \varepsilon) = h_0(t)\exp(\mathbf{x}'\mathbf{b} + \mathbf{z}'\mathbf{u} + \varepsilon) \qquad (5)$$

with **z** being the vector of random effects in the model and **u** the corresponding vector of parameters to estimate. Again, the random effects may be assumed to follow different distributions to be specified.

3. The biological meaning of frailty

Individual effects, not included in the vector of covariates observed, may lead to systematic effects and misinterpretation of the results of survival analyses. A simple example taken from Blossfeld et al. (1986) is used to illustrate the problem. Assume a population is composed of two subpopulations of equal size with different constant hazard functions $h(t)_1 = 0.4$ and $h(t)_2 = 0.1$. In this case the hazard function of the total population $h(t)$ is decreasing, as shown in Figure 1.

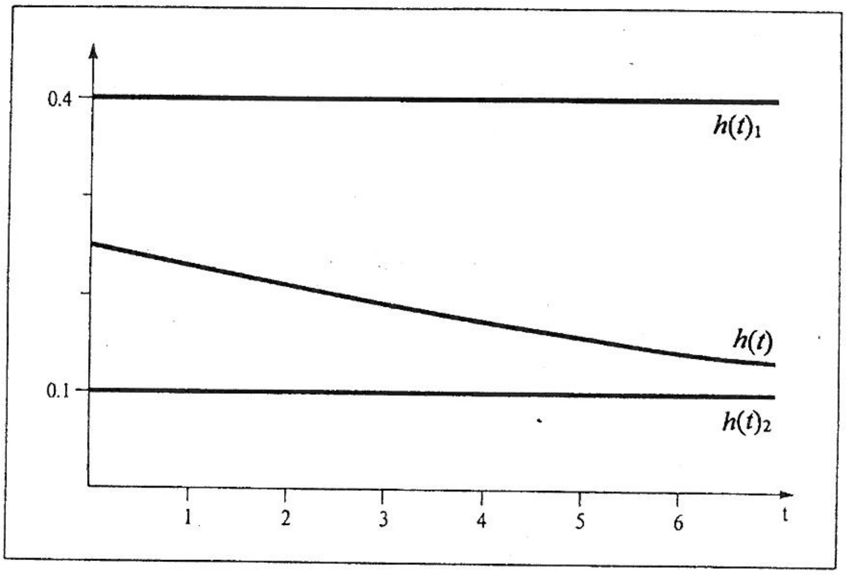

Fig. 1. Hazard functions for subpopulations with different frailty ($h(t)_1$ and $h(t)_2$) and resulting hazard function for the total population ($h(t)$).

This result is intuitively plausible. As time proceeds, individuals with the high hazard rate (subpopulation 1) on an average die earlier and the composition of the (mixed) population still at risk changes towards a higher proportion of low risk individuals of subpopulation 2. This simple result of a change of the hazard rate in negative direction extends to any situation where individuals have different hazard rates and these differences are not accounted for in the model, as shown by Heckman and Singer (1984).

An example where this phenomenon may play a role are reports on the effect of blood-pressure on survival rate in old people. Various studies were presented where persons with low pressure showed a higher risk of mortality than persons with high pressure. One possible reason for that may be, that frail persons with high pressure had died earlier, before reaching the age to enter the study and those individuals with high pressure were in the study, therefore, a selected sample of robust persons. A very good review of the effects of individual heterogeneity on survival analysis is given by Aalen (1994).

Another problem with ignoring individual heterogeneity was pointed out by Henderson (1996). In a study he presented exclusion or inclusion of a frailty term scarcely changed the size of the effects of the other covariates in the model. The level of significance, however, was severely influenced. Inclusion of a frailty term led to grossly reduced significance of the other effects in the model suggesting a large number of "false positive" effects found in many studies ignoring frailty.

4. The Survival Kit

The Survival Kit is a set of Fortran programs that was primarily written with the intention to fill a gap in the software available to animal breeders who generally tend to use extremely large data sets and want to estimate random effects. Although developed from animal breeders for animal breeders, the programs of the Survival Kit should be interesting for people from other areas encountering similar problems of large models and random effects. Frailty models 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 mimick the SAS command language.

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:

$$h(t, \mathbf{x}(t), \mathbf{z}(t), \varepsilon) = h_{0j}(t) \exp(\mathbf{x}(t)'\mathbf{b}+\mathbf{z}(t)'\mathbf{u}+\varepsilon) \tag{6}$$

where $h(t, \mathbf{x}(t), \mathbf{z}(t), \varepsilon)$ is the hazard function of an individual depending on time t, a vector of (possibly) time-dependent fixed covariates $\mathbf{x}(t)$ with corresponding parameter vector $\mathbf{b}$, a vector of (possibly) time-dependent random covariates $\mathbf{z}(t)$ with corresponding parameter vector u and a term ε, describing individual heterogeneity of observations.

Features supported

baseline hazard function: the (possibly) stratified baseline hazard function $h_{0j}(t)$ may either be unspecified so that we are actually dealing with the Cox model (this is possible due to a special way of deriving maximum likelihood estimates based on partial likelihoods, where the baseline hazard function cancels out; Cox, 1972), or it may follow a Weibull distribution, a common assumption that has been shown to be flexible and often adequate for biological data (e.g. Ducrocq et al., 1988).

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 (will depend on machine size). Covariates may be time-dependent ($\mathbf{x}(t)$), 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. starting or stopping smoking).

random covariates: the random covariates in vector $\mathbf{z}(t)$ 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 $\exp(\mathbf{z}(t)'\mathbf{b})$ is then Gamma distributed, which is a usual assumption for frailty models (see above). The two parameters of the Gamma distribution are taken to be equal so that the expectation $\mathrm{E}(\exp(\mathbf{z}(t)'\mathbf{b})) = 1$. With the Normal distribution, $\mathrm{E}(\mathbf{z}(t)'\mathbf{b}) = 0$. The distributional 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. Several random effects may be specified in the same model, they may be time-dependent.

individual heterogeneity (the frailty term): although the expression "frailty term" is used to generally denote random effects in survival analysis, it was originally introduced to account for individual heterogeneity of observations, as is done by the error term in the linear model. This term may also follow one of the distributions mentioned above and is technically treated in the same way as other random effects described above.

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 assumption of proportional hazards for all individuals over the total observational period. One variable may be chosen as strata variable, the number of strata is not restricted.

In addition to the estimation of fixed and random effects, the Survival Kit offers options for calculating asymptotic standard deviations of effects (only for smaller models, where the matrix of second derivatives may be set up), a suite of likelihood ratio tests and different ways of setting constraints to deal with dependecies in the model. As a special feature, different values of the survivor function may be estimated for individuals with preset covariate sructure. In this way it is possible to calculate estimated median survival time or survival probability to a specified age for any combination of covariate values.

The programs

The Survival Kit currently consists of a set of three Fortran programs, called prepare.f, cox.f and weibull.f (denoted as PREPARE, COX and WEIBULL subsequently).The package works stand-alone, i.e. does not rely on any subroutines from mathematical subroutine libraries like NAG. The optimisation routines used are partly taken from public domail libraries and are integrated in the programs.

PREPARE is used to prepare the data for the actual analysis with either COX or WEIBULL. Data preparation includes recoding of class variables and (more importantly) 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 and 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 distribution. Specifications for both models are similar, but it is computationally easier (and less time consuming) to estimate the distributional parameters of the random effects under the Weibull model.

An example

Just to give an idea about the command structure in the parameter files needed for performing analyses, the parameter files for an analysis of the effect of maternal age (AGEDAM) at birth on the longevity (ND) of daughters in a cattle are given. The analysis includes several fixed effects (HERD, U1, U2) relating to the environment of the daughter, a strata effect year of birth (YOB) as well as the random effect of the mother (DAM) to account for the fact that mothers had usually had more than one daughter and these daughters share the genes of their mother. The parameter files will not be described in detail, the statements can be looked up in the manual.

Parameter file for PREPARE

```
FILES datcox1 datcox2 newcode ad.plo;
INPUT ND I4 CENS I4 YOB I4 HERD R4 U1
R4 U2 R4 AGEDAM R4 DAM I4;
TIME ND;
CENSCODE CENS 0;
CLASS YOB;
FUTURE datfu datfu2;
OUTPUT ND CENS YOB HERD U1**2 U2
AGEDAM DAM;
```

Parameter file for COX

```
TITLE Effect of age of dam at birth on survival
of daughter;
FILES datcox2 newcode ad.out;
STRATA YOB;
MODEL HERD U1 U2 AD YOB DAM;
RANDOM DAM NORMAL .2;
CONSTRAINT LARGEST;
TEST LAST;
BASELINE;
SURVIVAL DATFU2 ADFU.OUT EQUAL_SP
25 5000;
```

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. The size of the program may be varied through changes in paramaters affecting the maximum number of records and maximum number of levels of effects to be estimated. For PC compilers making use of extended memory is favorable.

How to get the Survival Kit

The Survival Kit may be requested from its authors:

Vincent Ducrocq: Vincent.ducrocq@dga.jouy.inra.fr
Johann Sölkner: Soelkner@mail.boku.ac.at

It may also be retrieved from the following web page: http://www.boku.ac.at/nuwi/popgen/

References

AALEN, O. O., (1994) "Effects of frailty on survival analysis", Stat. Meth. in Medical Res. 3, 227–243

BLOSSFELD, H. P.O., HAMERLE, A. AND K. U. MEYER, (1986) Ereignisanalyse. Statistische Theorie und Anwendung in den Wirtschafts- und Sozialwissenschaften, Campus, Frankfurt.

COX, D. R., (1972) "Regression models and life tables (with discussion)", J. R. Statist. Soc. Series B 34, 187–220.

DUCROCQ, V. P., QUAAS, R. L., POLLAK, E. J. AND G. CASELLA, (1988) "Length of productive life of dairy cows. 1: Justification of a Weibull model", J. Dairy Sci. 71, 3071–3079.

DUCROCQ, V. P. AND J. SÖLKNER, 1994: "The Survival Kit — a Fortran package for the analysis of survival data.", Proc. of the 5th World Congress on Genetics Applied to Livestock Production 22, 51–52.

HECKMAN, J. J. AND B. SINGER , (1984) "Econometric duration analysis", Journal of Econometrics, 24, 63–132.

HENDERSON, R., (1996) "Frailty in survival and point process analysis", Paper presented at the Meeting of the Viennese Section of the International Biometrical Society, Vienna, 14 May.

KLEIN, J. P., MOESCHBERGER, M., LI, Y. H. AND S. T. WANG, (1992) "Estimating random effects in the Framingham heart study", in Survival analysis: state of the art, ed. by J. P. Klein and P. K. Goel, Kluwer Academic Publishers, pp. 99–120.

VAUPEL, J. W., MANTON, K. G. AND E. STALLARD, (1979) "The impact of heterogeneity in individual frailty and the dynamics of mortality", Demography 16, 439–454.

Received: October, 1996
Accepted: November, 1996

Contact address:
Johann Sölkner
Department of Livestock Sciences
University of Agricultural Sciences
Gregor–Mendel–Str. 33
A–1180 Vienna
Austria
e-mail: `soelkner@mail.boku.ac.at`

JOHANN SÖLKNER Associate professor of Population Genetics and Animal Breeding at the University of Agricultural Sciences in Vienna, Austria. Interest in methods of Survival Analysis stems from studies about the longevity of dairy cows. Other areas of interest are experimental design and modelling of genetic systems.