Genetic analysis of survival data - a Gibbs sampling approach.

Inge Riis Korsgaard
Danish Institute of Animal Science, Dept. of Breeding and Genetics, Research Centre Foulum, P.O.Box 39, DK-8830 Tjele

ABSTRACT

Survival analysis accommodate e.g. left truncation and right censoring. In this paper concepts dealing with survival data are introduced and an animal model for survival data is proposed. Its potential use in animal breeding is discussed briefly.

1. INTRODUCTION

The classical linear model cannot accommodate records that are left truncated and/or right censored. Often incomplete records are discarded or some derived quantity is considered, thereby loosing information, or incomplete records are projected to obtain predicted records, which are considered observed in the analysis. The purpose of this paper is to introduce concepts dealing with analysis of survival data and to illustrate different mixed models for survival data. Section 2 gives an introduction to survival data

and classical ways to model these. Section 3 deals with gamma frailty models, that are mixed models for survival data. Section 4 deals with log normal frailty models, another type of mixed models for survival data, that from a genetic point of view seems appealing. For Bayesian inference about the parameters in this model a Gibbs sampling approach is outlined. Section 5 briefly comments potential uses in animal breeding.

2. INTRODUCTION TO SURVIVAL ANALYSIS

Survival analysis deals with survival data. These data are characterised by special patterns of incompleteness of data, of which right censoring and left truncation are two important examples.

Example 1. Assume that it is decided to study time from birth until first calving in a five year period from January 1'st, 1996 until January 1'st, 2001. It is assumed that date of birth is recorded but date of first calving is recorded from January 1'st, 1996 only. Heifer calves that are born and have their first calving during this five year period have records. Heifer calves born in the five year period, but having their first calf after January 1'st, 2001 have right censored records. Concerning these animals, it is only known that age at first calving is greater than age at January 1'st, 2001, i.e. age at (right) censoring. Heifer calves, born before January 1'st 1996 and having their first calf in the five year period, have left truncated records, i.e. these animals belong to the dataset conditional that they were alive on January 1'st 1996.

Example 2. Nielsen et al (1992) and Korsgaard and Andersen (1995) analysed length of life of Danish adopted children and of their biological parents. Only families, where the child was born between 1924 and 1926 and where the child was alive at his or her sixteenth birthday, were included. Individuals, that did not die before analysing data, have right censored records. Beyond being right censored, data are left truncated, because individuals are not at risk of being observed to die from age zero: children belong to the dataset conditional on being alive at their 16'th birthday and are considered to be at risk from that age. Mothers are considered to be at risk from delivery and fathers from conception assumed to take place 280 days before delivery.

In many examples only right censoring is present and the rest of this paper deals with right censoring only.

Assume that data are right censored, i.e. for some individuals it is only known that lifetime T_i is greater than age at censoring C_i. Data of individual i is (Y_i, δ_i), i=1,...,n, where $Y_i = \min\{T_i, C_i\}$ and δ_i is an indicator random variable equal to 1 if a lifetime is observed $(Y_i = T_i)$ and δ_i is equal to 0 if a censoring time is observed $(Y_i = C_i)$.

The distribution of lifetime can be uniquely determined by each of the following interrelated quantities (Kalbfleisch and Prentice, 1980): the density function $f_i(t)$, the distribution function $F_i(t)$, the survival function $S_i(t) = 1 - F_i(t) = P(T_i > t)$, the hazard function $\lambda_i(t)$ or the cumulative hazard function $\Lambda_i(t) = \int_0^t \lambda_i(u)du$. The hazard function is defined by

$$\lambda_i(t) = \lim_{\Delta t \to 0} P(T_i \le t + \Delta t | T_i > t)/(\Delta t)$$

i.e. for Δt small $\lambda_i(t)\Delta t$ is approximately the conditional probability of individual i of dying in the interval $(t, t + \Delta t]$ given it was alive at time t.

In survival analysis most often the hazard function is modelled. The hazard function can be non-parametric, semiparametric or fully parametric. The Cox model is a semiparametric model. The hazard function is given by $\lambda_i(t) = \lambda_0(t)\exp\{x_i'\beta\}$, where $\lambda_0(t)$ is an underlying hazard function, x_i is a vector of covariates of individual i, due to which individual i's hazard functions deviate from the underlying hazard function, β is the corresponding vector of regression parameters. In the Cox model, the ratio of two different individuals hazard functions is independent of time t, i.e. for $i \neq j$, $\lambda_i(t)/\lambda_j(t) = \exp\{(x_i' - x_j')\beta\} = \alpha_{ij}$, where α_{ij} is a constant independent of time t. The Cox model can be generalised in different ways, e.g. to allow for stratification, and to

allow for time dependent covariates. An example, of a model with time dependent covariates, is given in the next section. In fully parametric models also $\lambda_0(t)$ is parameterised. E.g. $\lambda_i(t) = \lambda_0(t)\exp\{x_i'\beta\}$, with $\lambda_0(t) = \lambda\,;\ \lambda > 0$, is the exponential regression model.

3. GAMMA FRAILTY MODELS

The models described in the preceding section were all fixed effects models for survival data. Frailty models are mixed models for survival data. In the gamma frailty model it is assumed that there is a random variable, that is gamma distributed (or a linear combination of gamma distributions), that acts multiplicatively on the hazard function. The simplest case is the shared gamma frailty model - a sire model for survival data. In the shared frailty model, groups of individuals (or several survival times on the same individual) share the same frailty variable; e.g. daughters of a given sire could share the same frailty variable.

A fully parametric shared gamma frailty model with time dependent covariates was used to model true stayability of dairy cows in Ducrocq et al. (1988). True stayability is defined as the aptitude of a cow to delay culling, and the models was

$$\lambda_{ij}(t) = \lambda_0(t) z_i \exp\{h_m(t) + g_{kl}(t)\}$$

with $\lambda_0(t) = \lambda p(\lambda t)^{p-1}\,;\ \lambda, p > 0$. $\lambda_{ij}(t)$ is the hazard function of the j'th daughter of sire i, conditional on the frailty of sire i: $Z_i = z_i\,;$ i=1,...,s; j=1,..., n_i. The underlying hazard function, $\lambda_0(t)$, is that of a Weibull distribution. $h_m(t)$ and $g_{kl}(t)$ are time dependent covariates. $h_m(t)$ is the m'th time dependent herd × year effect, which changes on January 1'st each year, i.e. it is a function of calender time. $g_{kl}(t)$ is the time dependent stage of lactation × lactation number effect corresponding to the k'th stage of lactation (from 0 to 29 days after parturition, from 30 to 249 days or from 250 to the beginning of

the next lactation) and the l'th lactation number (lactations 1,2 and 3 or more). g is a function of biological time.

4. LOG NORMAL FRAILTY MODELS

The frailty variable is often assumed to follow a gamma distribution or a linear combination of gamma distributions, mainly for reasons of mathematical convenience.

In the log normal frailty model, the frailty variable is assumed to follow a log normal distribution, i.e. conditional that frailty Z_i of individual i equals z_i, the hazard function of individual i; i=1,...,n, is given by

$$\lambda_i(t) = \lambda_0(t) z_i \exp\{x_i'\beta\} \tag{1}$$

where $Z_i = \exp(W_i)$, $W \sim N_n(0,\Sigma)$ where $W = (W_1,...,W_n)$. If the baseline hazard function, $\lambda_0(\cdot)$, is unknown, this is a semiparametric model. The model given by (1) can be generalised, like the Cox model, to allow for stratification and for time dependent covariates.

In the rest of this section, a special case of (1), that essentially is an animal model for survival data, is focused on.

Let $W_i = a_i + e_i$, i=1,...,n, where $a = (a_1,...,a_n)$ given σ_a^2 is $N_n(0, A\sigma_a^2)$ distributed and independent of $e = (e_1,...,e_n)$, that given σ_e^2 is $N_n(0, I\sigma_e^2)$ distributed. **A** is the additive genetic relation matrix. In this case (1) becomes

$$\lambda_i(t) = \lambda_0(t)\exp\{a_i + e_i\}\exp\{x_i'\beta\} \tag{2}$$

or

$$\log\lambda_i(t) = \log\lambda_0(t) + a_i + e_i + x_i'\beta \tag{3}$$

i.e. there is an additive genetic part a_i and an environmental part e_i affecting frailty in the multiplicative way given by (2) or in the log additive way given by (3).

Inference is approached in a Bayesian way. The Bayesian approach requires computation of the joint and marginal posterior distributions of parameters and hyperparameters. These are intractable, but the full conditional distributions of the parameters are either known distributions or proportional to some log concave distribution. This fact makes it possible to use e.g. the Gibbs sampler, a MCMC method, to sample from the joint posterior distribution of the parameters (e.g. Gelfand et al. (1990)). The empirical distribution functions of the parameters can be used to make inferences about the parameters in the model.

It is assumed that each component in β are independent and have an uniform improper prior on $[-\infty;\infty]$ and that the prior for Λ_0, the integrated underlying hazard function, is an independent increment gamma processes with mean $E(\Lambda_0(t)) = \Lambda_0^*(t)$ and variance $Var(\Lambda_0(t)) = \Lambda_0^*(t)/c$ where $\Lambda_0^*(\cdot)$ is a known function. The increments are independent and gamma distributed: $d\Lambda_0(t) \sim \Gamma(cd\Lambda^*(t), c^{-1})$. An improper vague prior is incorporated for Λ_0 by letting $c = 0$. As stated already $a|\sigma_a^2 \sim N_n(0, A\sigma_a^2)$ and $e|\sigma_e^2 \sim N_n(0, I\sigma_e^2)$. A priori σ_a^2 is assumed to be $IG(a_1, b_1)$ distributed and σ_e^2 is assumed to be $IG(a_2, b_2)$ distributed, that is inverted gamma distributions. These prior distributions are chosen in close agreement with those used in earlier Bayesian analysis of survival data (e.g. Kalbfleisch (1978), Clayton (1991), Sinha (1993) and Gauderman and Thomas (1994)), to the extend that the models agree.

A priori, it is assumed that that β, (a, σ_a^2), (e, σ_e^2) and the process $\Lambda_0(\cdot)$ are mutually independent. By this assumption, the joint posterior distribution of parameters and hyperparameters $\theta = (\Lambda_0(\cdot), \beta, a, \sigma_a^2, e, \sigma_e^2)$ given data (y, δ) is given by

$$p(\theta|y,\delta) \propto p(y,\delta|\theta)p(\Lambda_0(\cdot))p(\beta)p(a|\sigma_a^2)p(\sigma_a^2)p(e|\sigma_e^2)p(\sigma_e^2)$$

where $p(y,\delta|\theta)$ is the conditional likelihood of $\psi=(\Lambda_0(\cdot),\beta,\sigma_a^2,\sigma_e^2)$ given a and e. The product $p(\Lambda_0(\cdot))p(\beta)p(a|\sigma_a^2)p(\sigma_a^2)p(e|\sigma_e^2)p(\sigma_e^2)$ is the prior distribution of θ.

Under the assumptions, that conditional on a and e censoring is independent and non-informative on θ, $p(y,\delta|\theta)$ is given by (e.g. Andersen et al. (1992))

$$p(y,\delta|\theta)=\prod_{i=1}^{n}\left[\left(\exp\{a_i+e_i+x_i'\beta\}\lambda_0(y_i)\right)^{\delta_i}\exp\left\{-\left(\exp\{a_i+e_i+x_i'\beta\}\right)\Lambda_0(y_i)\right\}\right]$$

Given assumptions on censoring and on prior distributions, the full conditional posterior distributions are given as follows: $\Lambda_0(\cdot)$ is another independent increment gamma process, that jumps at failure times only. The full conditional distribution of each component in β is proportional to a log concave distribution, similarly for each a_i and adaptive rejection sampling (Gilks and Wild, 1992) can be used to sample from these distributions. σ_a^2 follows an $IG\left(n/2+a_1, b_1+2/(a'A^{-1}a)\right)$ distribution and σ_e^2 an $IG\left(n/2+a_2, b_2+2/(e'A^{-1}e)\right)$ distribution. Further details are given in a paper under preparation.

5. POTENTIAL USE OF THE LOG NORMAL FRAILTY MODEL - BRIEFLY

Whether animals with high or low values of a_i should be selected, depend on the nature of the problem; this is exemplified in the following.

In the context of example 1, assume that the breeding goal is to shorten time from birth to first calving and that the model for time from birth until first calving conditional on a and e is given by $\lambda_i(t)=\lambda_0(t)z_i\exp\{x_i'\beta\}$, where $z_i=\exp\{a_i\}\exp\{e_i\}$. Considering the genetic contribution only, a high value of $\exp\{a_i\}$ or identical a high value of a_i is preferential. In a breeding program, animals with the highest values of a_i are selected.

In other cases a low value of a_i is preferential. Consider an example where the breeding goal is to prolong length of life and that the model is given, conditional on a and e, by $\lambda_i(t) = \lambda_0(t) z_i \exp\{x_i'\beta\}$, where again $z_i = \exp\{a_i\}\exp\{e_i\}$. Considering the genetic part of frailty, a low value of a_i is desirable. A low value of a_i gives a low conditional probability of dying in the next small time interval given alive right now - over the whole time interval. As an example consider two animals i and j at time t with exactly the same covariates and by assumption of the model the same underlying hazard function. Given $e_i = e_j = 0$ and $a_i = -a_j = 1$ then $\lambda_j(t)/\lambda_i(t) = \exp\{-2\}$, i.e. the hazard of individual j is only 0.14 of that of individual i's for all $t \in R_+$.

REFERENCES

Andersen P.K., Borgan, Ø., Gill, R.D., Keiding, N. (1992). Statistical models based on counting processes. Springer.

Clayton, D.G. (1991). A Monte Carlo method for Bayesian inference in frailty models. Biometrics 47: 467-485.

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

Gauderman, W.J. and Thomas, D.C. (1994). Censored survival models for genetic epidemiology: A Gibbs sampling approach. Genetic Epidemiology 11: 171-188.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. J. Am. Stat. Assoc. 85: 972-985.

Gilks, W.R., Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Appl. Statist. 41, No. 2.:337-348.

Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data. J. R. Stat. Soc. B, 40: 214-221.

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

Korsgaard, I.R. and Andersen, A.H. (1995). The additive genetic gamma frailty model. Research Report No. 315, Department of Theoretical Statistics, University of Aarhus, Denmark.

Nielsen, G.G, Gill, R.D., Andersen, P.K. and Sørensen, T.I.A. (1992). A counting process approach to maximum likelihood estimation in frailty models. Scand. J. Statist. 19: 25-43.

Sinha, D. (1993). Semiparametric Bayesian analysis of multiple event time data. J. Am. Stat. Assoc. 88: 979-983.