Extension of survival analysis models to discrete measures of longevity

Vincent Ducrocq
Station de Génétique Quantitative et Appliquée,
Institut National de la Recherche Agronomique,
78352 Jouy en Josas, France

Abstract

Proportional hazards models most often assume that failure times are expressed on a continuous time scale. There are common situations in animal breeding when this is not the case, for example when longevity is expressed in years, number of parities, etc… As a result, some basic assumptions of most survival models are violated. For example, the Cox' partial likelihood is no longer correct. This paper presents a more appropriate strategy, the "grouped data model" of Prentice and Gloeckler (1978) for the analysis of discrete data, that maintains the assumption of proportional hazards. A reparameterisation of the model underlines the main differences with the Cox model or with a parametric regression model, the Weibull model. It also allows an easy modification of existing programs to make them suitable for the analysis of discrete survival data. The « grouped data model » is compared with continuous models on simulated data sets. The robustness of the Weibull model and the inadequacy of the Cox model on discrete data are illustrated.

1. Introduction

In survival analysis, the time scale used to describe failure time is most often considered to be continuous. Such an assumption seems reasonable when length of life of large domestic animals is expressed, e.g., in days. Then, it is becoming a standard practice to analyse such data using proportional hazards models (Cox, 1972, Kalbfleisch and Prentice, 1980) or, in genetic studies, their extension to mixed (frailty) models (Ducrocq, 1997). However there are situations when the time scale is obviously discrete with very few classes: this is the case when the available information is limited to a total number of parities, a number of completed lactations, etc… But in some instances, even though more precise information is available, the exact timing of culling – just after calving, 3 months later or just before the next calving - is somewhat irrelevant, e.g., what counts is that there will be no more progeny born after culling. For discrete data, a direct survival analysis using 'standard' proportional hazards models is *a priori* incorrect, as the usual approaches assume continuity of the baseline hazard distribution and/or absence of ties between ordered failure times. After a description of a technique due to Prentice and Gloeckler for the analysis of discrete survival data without rejecting the proportional hazards model, I will describe how this technique can be easily accommodated in the Survival Kit, a package developed for the use of regression and frailty models with time-dependent covariates (Ducrocq and Soelkner, 1998).

2. Background

2.1 Proportional hazards models

Let $\mathbf{x} = (x_1 \; .. \; x_n)'$ be a vector of explanatory variables upon which failure time may depend. The x_j's can be continuous or discrete covariates. In *proportional hazards models* (PHM ; Cox, 1972), the hazard function h(t) and **x** are associated through the expression:

$$h(t ; \mathbf{x}) = h_0(t) \exp\{\mathbf{x}'\boldsymbol{\beta}\} \qquad [1]$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients. $h_0(t)$ is called the *baseline hazard function* and $\exp\{\mathbf{x}'\boldsymbol{\beta}\}$, represents a stress-dependent term specific to the animals with covariates **x**. The simple intuitive interpretation of the regression coefficients in proportional hazards models greatly explains their popularity: the hazards of two individuals are proportional over time. If (piecewise constant) time-dependent covariates are used, the proportionality assumption must hold over intervals and not longer on the whole time axis.

When a parametric form is chosen for the baseline hazard function $h_0(.)$ in [1], it is relatively easy to write down, to compute and to maximise the full likelihood function, combining contributions from censored and uncensored records. This provides maximum likelihood estimates of the parameters of the baseline density and of $\boldsymbol{\beta}$. But the parametric forms most frequently used (exponential, Weibull, gamma, log-normal, Pareto, etc…) are all continuous. One can wonder about the consequences on

parameter estimation of the use of continuous functions to describe inherently discrete variables.

It is also possible to leave $h_0(.)$ completely arbitrary. Expression [1] then defines a semiparametric regression model known as a *Cox model* (Cox, 1972). The attractive feature of the Cox model is that it permits the estimation of $\boldsymbol{\beta}$ without making any assumption about the form of $h_0(.)$. The procedure developed by Cox relies on the definition of what he calls a *partial likelihood function* which is the part of the full likelihood function which does not depend on $h_0(t)$. The formal expression of the logarithm of the partial likelihood is:

$$\log \boldsymbol{L_C}(\boldsymbol{\beta}) = \sum_{k \in \boldsymbol{F}} \left[\mathbf{x}'_{[k]}\boldsymbol{\beta} - \log \sum_{j \in \boldsymbol{R}(T_{[k]})} e^{\mathbf{x}'_j \boldsymbol{\beta}} \right] \quad [2]$$

where **F** is the set of ordered <u>*distinct*</u> failure times $T_{[k]}$ and $\boldsymbol{R}(T_{[k]})$ is the set of animals at risk (i.e., alive) at time $T_{[k]}$. The partial likelihood also received other formal justifications. In particular, it can be obtained as the marginal likelihood of the ranks of failure times, i.e., it contains *all* the information about the *order* in which animals died. However, the ranking of failure times is not possible with a discrete measure of failure times, which generates a large amount of "ties". When there are few ties between failure times (at least compared with the total number of observations), approximations of Cox's partial (log-)likelihood have been proposed. In particular Peto (1972, in the discussion of Cox's paper) suggested to use:

$$\log \boldsymbol{L_P}(\boldsymbol{\beta}) = \sum_{k \in \boldsymbol{F}} \left[\left(\sum_{i \in \boldsymbol{D}(T_{[k]})} \mathbf{x}'_i \boldsymbol{\beta} \right) - d_k \left\{ \log \sum_{j \in \boldsymbol{R}(T_{[k]})} e^{\mathbf{x}'_j \boldsymbol{\beta}} \right\} \right] \quad [3]$$

where **D** is the set of the d_k dying at time $T_{[k]}$.

Once estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ have been obtained maximising [3], the baseline survivor function $S_0(t)$ is estimated assuming that $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

2.2 The grouped data model of Prentice and Gloeckler (1978)

When there are many ties among failure times, e.g., when only a few classes (say, less than 20) of a discrete measure of survival are available, the approximation [3} is no longer valid or useful and a different analysis must be performed. Prentice and Gloeckler (1978) presented another approach for such data, that I will introduce now.

Define the intervals representing the unit of measure (e.g., years):

$$[0 = \tau_0,\ \tau_1), [\tau_1,\ \tau_2), \ldots, [\tau_{k-1},\ \tau_k), \ldots$$

Implicitly, all failures occuring during the interval $[\tau_{k-1},\ \tau_k)$ will be "grouped" and the attached failure time will be k. It will also be assumed that censoring only occurs *at the end* of each interval. The survivor function at $t = \tau_{k-1}$, i.e., at the beginning of the interval k, is:

$$S(t = \tau_{k-1}; \mathbf{x}) = \exp\left\{ -\int_0^{\tau_{k-1}} h(u; \mathbf{x})\, du \right\} \quad [4]$$

which can be calculated as:

$$S(\tau_{k-1}; \mathbf{x}) = \exp\left\{ -\sum_{i<k} \int_{\tau_{i-1}}^{\tau_i} h_0(u) e^{\mathbf{x}'\boldsymbol{\beta}}\, du \right\} \quad [5]$$

$$= \prod_{i<k} \exp\left[\left\{ -\int_{\tau_{i-1}}^{\tau_i} h_0(u)\, du \right\} e^{\mathbf{x}'\boldsymbol{\beta}} \right]$$

$$= \prod_{i<k} \left(\exp\left\{ -\int_{\tau_{i-1}}^{\tau_i} h_0(u)\, du \right\} \right)^{e^{\mathbf{x}'\boldsymbol{\beta}}}$$

$$= \prod_{i=1}^{k-1} (\alpha_i)^{e^{\mathbf{x}'\boldsymbol{\beta}}} \quad [6]$$

where $\alpha_i = \exp\left\{ -\int_{\tau_{i-1}}^{\tau_i} h_0(u)\, du \right\}$ [7]

Then, adapting the definition of the hazard function, we have:

$$h(t; \mathbf{x}) = \frac{S(\tau_{k-1}; \mathbf{x}) - S(\tau_k; \mathbf{x})}{S(\tau_{k-1}; \mathbf{x})} = 1 - \alpha_k^{e^{\mathbf{x}'\boldsymbol{\beta}}} \quad [8]$$

Using the formal relationship between survivor function, hazard function and density function and combining [6] and [8], we get :

$$f(t; \mathbf{x}) = h(t; \mathbf{x})\, S(t; \mathbf{x})$$

$$= \left(1 - \alpha_k^{e^{\mathbf{x}'\boldsymbol{\beta}}} \right) \left(\prod_{i=1}^{k-1} \alpha_i^{e^{\mathbf{x}'\boldsymbol{\beta}}} \right) \quad [9]$$

Expressions [6] and [9] are used in the construction of the full likelihood (Kalbfleisch and Prentice, 1980):

$$\boldsymbol{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{m \in \{\text{unc.}\}} f(y_m) \prod_{m \in \{\text{cens.}\}} S(y_m) \quad [10]$$

In contrast with the Cox model approach, the elements of the baseline survivor curve (the α_i's in [6]) are estimated jointly with $\boldsymbol{\beta}$.

An example using this methodology for the analysis of number of years in competition in horses can be found in Ricard and Fournet-Hanocq (1997).

3. Reparameterisation of the grouped data model

The joint estimation of the α_i's and $\boldsymbol{\beta}$ in [10] requires the writing of a specific programme. In fact, a simple

reparameterisation of the model permits the use of the Survival Kit package (Ducrocq and Soelkner, 1998) with minimal modification.

By definition, the α_i's in [6] take only values between 0 and 1. This requires a constrained maximisation of [10]. As noted by Miller (1981, p139), it is more convenient to reparameterise the α_i's into ξ_i's, where $\xi_i = \log(-\log \alpha_i)$ which all take values between $-\infty$ and $+\infty$. Then:

$$\alpha_i = \exp\{-\exp \xi_i\} \qquad [11]$$

This reparameterisation leads to new expressions for equations [5] and [6] :

$$h(t=\tau_{k-1}; \mathbf{x}) = 1-\alpha_k^{e^{\mathbf{x}'\boldsymbol{\beta}}} = 1-\exp\left\{-e^{\xi_i+\mathbf{x}'\boldsymbol{\beta}}\right\} \quad [12]$$

and:

$$S(t; \mathbf{x}) = \exp\left\{-e^{\mathbf{x}'\boldsymbol{\beta}}\left(e^{\xi_1}+e^{\xi_2}+\cdots+e^{\xi_{k-1}}\right)\right\} \quad [13]$$

Assume that all interval lengths are equal and define this interval length as unity ($\tau_0 = 0$, $\tau_1 = 1,\ldots,\ \tau_i = i$). Define $\mathbf{x}^{*'}(\tau_{i-1})\,\boldsymbol{\beta}^* = \xi_i + \mathbf{x}'\boldsymbol{\beta}$, then:

$$S(t; \mathbf{x}) = \exp\left\{-e^{\mathbf{x}'\boldsymbol{\beta}}\left(e^{\xi_1}+e^{\xi_2}+\cdots+e^{\xi_{k-1}}\right)\right\}$$

$$= \exp\left\{-\sum_{i=1}^{k-1} e^{\mathbf{x}^{*'}(\tau_{i-1})\boldsymbol{\beta}^*} \times 1\right\}$$

$$= \exp\left\{-\sum_{i=1}^{k-1} e^{\mathbf{x}^{*'}(\tau_{i-1})\boldsymbol{\beta}^*} \times (\tau_i - \tau_{i-1})\right\} \quad [14]$$

Now, this equation [14] will be related to another expression obtained in a different context : consider a Weibull regression model, in a situation where the regression vector $\mathbf{x}(t)$ is time-dependent. The expression of the survivor function $S(y_m)$ for this animal is :

$$S(y_m; \mathbf{x}) = \exp\left\{-\int_0^{y_m} \lambda\rho(\lambda u)^{\rho-1} e^{\mathbf{x}'(u)\boldsymbol{\beta}}\, du\right\}$$

$$= \exp\left\{-\int_0^{y_m} \rho\, u^{\rho-1}\, e^{\rho\log\lambda + \mathbf{x}'(u)\boldsymbol{\beta}}\, du\right\}$$

$$= \exp\left\{-\int_0^{y_m} \rho\, u^{\rho-1}\, e^{\mathbf{x}^{*'}(u)\boldsymbol{\beta}^*}\, du\right\} \quad [15]$$

defining $\mathbf{x}^{*'}(u) = (1 \quad \mathbf{x}'(u))'$ and $\boldsymbol{\beta}^{*'} = (\rho\log\lambda, \boldsymbol{\beta}')$. Assume that $\mathbf{x}'(u)$ is a piecewise constant function of time : the value of at least one element of $\mathbf{x}(u)$ changes at time $q_0 = 0 < q_1 < \ldots < q_Q = y_m$, the failure time of animal m. This implies that expression [15] can be integrated explicitly:

$$S(y_m; \mathbf{x}) = \exp\left\{-\sum_{j=1}^{Q} \int_{q_{j-1}}^{q_j} \rho\, u^{\rho-1}\, e^{\mathbf{x}^{*'}(q_{j-1})\boldsymbol{\beta}^*}\, du\right\}$$

$$= \exp\left\{-\sum_{j=1}^{Q}\left[e^{\mathbf{x}^{*'}(q_{j-1})\boldsymbol{\beta}^*}\left(q_j^{\rho} - q_{j-1}^{\rho}\right)\right]\right\} \quad [16]$$

This resemblance between [14] and [16] suggests another interpretation of Prentice and Gloeckler's model: it is equivalent to an exponential ($\rho = 1$) regression model which includes a time-dependent covariate that I will call *time_unit*. This time-dependent covariate is a step function of time with changes at $\tau_0=0$, $\tau_1=1$, $\tau_2=2$, …, $\tau_k=k$,…. Then, the resulting expression for the survivor curve of such an exponential regression model is identical to [14]. However, the hazard functions differ. Indeed, the time-dependent exponential regression model leads to :

$$h(t=\tau_k; \mathbf{x}) = e^{\mathbf{x}^{*'}(\tau_{i-1})\boldsymbol{\beta}^*} = e^{\xi_k+\mathbf{x}'\boldsymbol{\beta}} \quad [17]$$

or, equivalently:

$$\log h(t; \mathbf{x}) = \xi_k + \mathbf{x}'\boldsymbol{\beta} \qquad [18]$$

instead of:

$$\log h(t; \mathbf{x}) = \log\left(1-\exp\left\{-e^{\xi_k+\mathbf{x}'\boldsymbol{\beta}}\right\}\right) \qquad [19]$$

for the grouped data model.

To illustrate the subtle difference between the xpressions [18] and [19], consider the limiting situation when there are many different (small) time intervals $[\tau_{k-1},\ \tau_k)$. Then, one can expect that the conditional survival probabilities α_k will not be very different from 1, i.e., $\alpha_k \approx 1-\varepsilon_k$ for some small ε_k or :

$$\xi_k = \log(-\log\alpha_k) \approx \log\varepsilon_k .$$

Then, starting from [19]:

$$\log h(t; \mathbf{x}) = \log\left(1-\left(\exp(-e^{\xi_k})\right)^{e^{\mathbf{x}'\boldsymbol{\beta}}}\right)$$

$$\approx \log\left(1-\exp(-\varepsilon_k)^{e^{\mathbf{x}'\boldsymbol{\beta}}}\right)$$

$$\approx \log\left(1-(1-\varepsilon_k)^{e^{\mathbf{x}'\boldsymbol{\beta}}}\right)$$

$$\approx \log\left(1-\left(1-\varepsilon_k e^{\mathbf{x}'\boldsymbol{\beta}}\right)\right)$$

$$= \log\left(\varepsilon_k e^{\mathbf{x}'\boldsymbol{\beta}}\right)$$

$$= \log(\varepsilon_k) + \mathbf{x}'\boldsymbol{\beta}$$

$$= \xi_k + \mathbf{x}'\boldsymbol{\beta} \qquad [20]$$

The latter expression is the same as definition [18]. In other words, Prentice and Gloecker's model can be viewed as a fully parametric (exponential) model, for which the baseline is estimated at every (discrete) time-point (through the use of the *time_unit* time-dependent variable) and for which the definition of the hazard is modified to take into account the discrete time scale. This property can be used to modify existing software for the analysis of grouped survival data In particular, the « Survival Kit – V3.1 » , which is available at :

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

includes such a change : in the parameter file of the program « prepare.f » of recodification, the simple keyword « DISCRETE ; » forces the definition of the time-dependent covariate *time_unit* and the creation for each animal of as many elementary records as changed in *time_unit.* Without any further indication, Prentice and Gloeckler's model will be automatically used when the next program (weibull.f – not cox.f !) is called. It should be remembered that in this case, the fitted model is *not* a Weibull model and that in the calculations, the value $\rho = 1$ will be always taken.

4. Extension to frailty models

Most of the Bayesian analysis of mixed (frailty) models developed in Ducrocq and Casella (1996) can be applied to Prentice and Gloeckler's model for discrete (or « grouped ») data. The vector β in $e^{\mathbf{x}'\beta}$ of expressions [6] and [8] can be extended to include random effects $\mathbf{s}=\{s_q\}$. Let:

$$\mathbf{w}'_m = \begin{pmatrix} \mathbf{x}'_m & \mathbf{z}'_m \end{pmatrix} \quad \text{and} \quad \theta = \begin{pmatrix} \beta \\ \mathbf{s} \end{pmatrix}$$

In the Bayesian analysis, the log-gamma or (multivariate) normal prior distributions can be used for the frailty term s_q. The analysis proceeds as in Ducrocq and Casella (1996).

One needs the following expressions of the survivor function:

$$S(t;\mathbf{w}_m) = \exp\left\{-e^{\mathbf{w}_m'\beta}\left(e^{\xi_1} + e^{\xi_2} + \cdots + e^{\xi_{k-1}}\right)\right\} \qquad [21]$$

and of the hazard function:

$$h(t;\mathbf{w}_m) = 1 - \exp\left\{-e^{\xi_k + \mathbf{w}'_m\beta}\right\} \qquad [22]$$

in order to combine the contribution of each individual to the construction of the likelihood function [10]. These formula can be adapted to accommodate time-dependent covariates. Inferences on θ can be drawn from the posterior distribution $\pi(\theta, \xi_1, \xi_2 ..., \xi_k, \text{hyperparameters} \mid \mathbf{y})$ exactly in the same way as in Ducrocq and Casella (1996) : the Laplacian integration technique to obtain the approximate marginal posterior distribution of the hyperparameter(s) of the prior distributions can be applied without any change. However, it is important to note that the algebraic integration of the frailty term from the joint posterior distribution (when the random term follows a log-gamma prior distribution is *not* possible here.

5. Numerical example

To illustrate the grouped data approach and without aiming at general conclusions, several proportional hazards (mixed) models were applied to the analysis of simulated data sets.

5.1. Data sets

Using the program simul.f (also available at the Survival Kit home page), 10000 records were simulated assuming a Weibull frailty model (data set **A**), on a continuous time scale. The true parameters of the Weibull baseline hazard distribution were ρ = 2.0 and $\rho \log\lambda$ = -14.05, which corresponds to a median failure time of 750 days. All values above 3000 days were censored at this date. (1.6% censored records). In the simulation model, the records were influenced by 2 fixed effects with 5 and 4 levels each and a random (« sire ») effect with 100 levels. The levels of both fixed effects were randomly distributed across records and each sire had 100 simulated daughters, resulting in a nearly balanced design The sire effects were assumed to be *iid* N(0, 0.05). Simulated values of fixed effects are indicated in table 1.

A discretised version of data set **A** (data set **B**) was created using the following rule : if y_m is continuous failure or censoring time, the new discrete value is $Y_m = n$ if $365*(n-1) \leq y_m < 365*n$. In other words, Y_m represents the number of « started » years of life.

Data sets $\mathbf{A_C}$ and $\mathbf{B_C}$ were obtained from **A** and **B** censoring records larger than 1095 d (= beginning of the fourth year) at t=1095 for the former and records larger than 4 (years) at t=4 for the latter. The corresponding censoring rate is 45.1% for both files. Note that the underlying model is still Weibull. To drastically force a different (not Weibull) model data set $\mathbf{B_C}$ was modified assuming that all records with Y_m= 4 were in fact *not* censored (data set $\mathbf{B_D}$).

5.2. Analyses

These data sets were analysed using the « Survival Kit – V3.1 » and fitting a Cox or a Weibull model (i.e., ignoring the discrete scale for data sets **B**, $\mathbf{B_C}$ and $\mathbf{B_D}$) or the grouped data model described in this paper. Solutions of fixed effects, characteristics of the approximate marginal posterior distribution of the sire variance were compared to the true values.

Table 1 : true values used in the simulation and estimates from the Cox model and the Weibull model when the time scale is really continuous

Model **Data set**[(a)] **Time scale**	**(True)** **(A/B)** **continuous**	**Cox** **A** **continuous**	**Weibull** **A** **continuous**
Intercept ($\rho \log\lambda$)	-14.0503		-14.007
ρ	2.0		1.994+/-0.016
fixed effect 1 : $\beta 1$	0	0	0
$\beta 2$	0.5296	0.5636	0.5675
$\beta 3$	-0.3456	-0.3073	-0.3080
$\beta 4$	0.3803	0.4046	0.4066
$\beta 5$	-0.5136	-0.5203	-0.5201
fixed effect 2 : $\gamma 1$	0	0	0
$\gamma 2$	-0.0510	-0.0480	-0.0485
$\gamma 3$	-0.3917	-0.4229	-0.4246
$\gamma 4$	-0.4071	-0.4764	-0.4785
sire variance : mode	0.05	0.04915	0.04948
mean		0.05169	0.05203
std		0.00917	0.00917
skewness		0.596	0.596

(a) see text

Table 2 : true values used in the simulation and estimates from the Cox model, the Weibull model and the drouped Data model when the time scale is discrete

Model **Data set**[(a)] **Time scale**	**(True)** **(A/B)** **continuous**	**Cox** **B** **discrete**	**Weibull** **B** **discrete**	**Grouped data** **B** **discrete**
intercept	-14.0503		-3.0075	
ρ	2.0		2.330/-0.018	
fixed effect 1 : $\beta 1$	0	0	0	0
$\beta 2$	0.5296	0.4270	0.5705	0.5578
$\beta 3$	-0.3456	-0.2428	-0.3253	-0.3102
$\beta 4$	0.3803	0.3061	0.4084	0.3983
$\beta 5$	-0.5136	-0.4210	-0.5542	-0.5296
fixed effect 2 : $\gamma 1$	0	0	0	0
$\gamma 2$	-0.0510	-0.0457	-0.0598	-0.0595
$\gamma 3$	-0.3917	-0.3287	-0.4377	-0.4244
$\gamma 4$	-0.4071	-0.3725	-0.4937	-0.4787
sire variance : mode	0.05	0.02756	0.05452	0.05078
mean		0.02924	0.05719	0.05341
std		0.00583	0.00993	0.00944
skewness		0.600	0595	0.596

(a) see text

5.3. Results

Table 1 illustrates the excellent behaviour of both the Cox model and the Weibull model when the time scale used for the analysis is continuous, in this idealised situation (almost no censoring, balanced design, true underlying Weibull model). The similarity of the estimates of the Cox and the Weibull models is striking : there is virtually no information lost when the partial likelihood is used. None of the estimates is significantly different from its true value. The approximate estimation procedure of the sire variance also gives very satisfying results

Table 3 : estimates from the Weibull model (correct underlying model) and the grouped data model when the time scale is either continuous or discrete, in presence of censoring

Model **Data set**[(a)] **Time scale**	**(True)** **(A/B)** **continuous**	**Weibull** **A_C** **continuous**	**Weibull** **B_C** **discrete**	**Grouped data** **B_C** **discrete**
intercept	-14.0503	-14.174	-2.8643	
ρ	2.0	2.028+/-0.0247	2.123+/-0.026	
fixed effect 1 : β1	0	0	0	0
β2	0.5296	0.5642	0.5734	0.5493
β3	-0.3456	-0.2898	-0.3109	-0.2952
β4	0.3803	0.4079	0.4206	0.4026
β5	-0.5136	-0.5111	-0.5510	-0.5238
fixed effect 2 : γ1	0	0	0	0
γ2	-0.0510	-0.0737	-0.0952	-0.0873
γ3	-0.3917	-0.4329	-0.4662	-0.4407
γ4	-0.4071	-0.5067	-0.5338	-0.5069
sire variance : mode	0.05	0.05551	0.06345	0.05698
mean		0.05867	0.06696	0.06018
std		0.01142	0.01264	0.01169
skewness		0.603	0.601	0.603

(a) see text

Table 4 : estimates from the Weibull model (incorrect underlying model) and the grouped data model when the time scale is discrete

Model **Data set**[(a)] **Time scale**	**(True)** **(A/B)** **continuous**	**Weibull** **B_D** **discrete**	**Weibull** **+ *time_unit* effect** **B_D** **discrete**	**Grouped data** **B_D** **discrete**
intercept	-14.0503	-4.3852	-2.341	
ρ	2.0	3.696+/-0.32	1 (constrained)	
fixed effect 1 : β1	0	0	0	0
β2	0.5296	0.3719	0.3061	0.5494
β3	-0.3456	-0.1429	-0.1228	-0.2951
β4	0.3803	0.2578	0.2113	0.4026
β5	-0.5136	-0.2323	-0.1992	-0.5237
fixed effect 2 : γ1	0	0	0	0
γ2	-0.0510	-0.0566	-0.0463	-0.0873
γ3	-0.3917	-0.2528	-0.2107	-0.4407
γ4	-0.4071	-0.2868	-0.2386	-0.5069
sire variance : mode	0.05	0.01532	0.00733	0.05693
mean		0.01633	0.00778	0.06018
std		0.00389	0.00264	0.01169
skewness		0.598	0.583	0.603

(a) see text

In table 2, the time scale is changed. Then the Cox model gives poor results : all fixed effects solutions are shrinking towards 0, compared with the true value (by 20 to 30% for fixed effect 1).

More importantly, the sire variance is strongly under-estimated. This reflects the inadequacy of the approximation [3] of the partial log-likelihood in presence of many ties. The solutions of the Weibull model are reasonably correct, except for the Weibull parameter ρ. Indeed, it seems that the biases in β may be a direct consequence of the overestimation of ρ (the ratios $\hat{\beta}/\hat{\rho}$ and β/ρ are very similar). The grouped data model gives the best results, with no significant bias for the fixed effects as well as for the sire variance.

Table 3 reports the analyses of data sets $\mathbf{A_C}$ and $\mathbf{B_C}$, obtained from **A** and **B** after censoring records

of 4 years and more. When the time scale is still continuous, the solutions from the Weibull model and also from the Cox model (not shown) are almost unchanged : censoring has very limited impact. When failure and censoring times can take only one out of 4 values (1, 2, 3 or 4), solutions of fixed effects are only marginally affected. However, the sire variance is overestimated, although the true value 0.05 is still in the 95% credible set of its marginal posterior density. The grouped data model gives better results than the Weibull (discrete) model, for the fixed effects as well as for the sire variance component.

Finally, in data set $\mathbf{B_D}$, the underlying distribution is forced to be not distributed as Weibull by treating all censored records of data set $\mathbf{B_C}$ as uncensored. Note that this has no impact on the « true » values of β and of the sire variance used to simulate the data nor on the validity of the proportional hazards assumption [1]. Table 4 shows that applying a Weibull model to such data set is incorrect : solutions for fixed effects are even more shrunk towards 0 than for the Cox model in table 2 and the sire variance is grossly underestimated (close to 0). Once again, the grouped data set gives excellent results, unchanged with respect to the analysis of data set $\mathbf{B_C}$ in table 3. In contrast with the Weibull model, the grouped data model can accommodate a conditional survival probability of 0 at time 4, and reveals the proper genetic variability.

6. Conclusion

It is not possible to generalise inferences drawn from this small, idealised numerical example. Nevertheless, it illustrates some characteristics that are well known from methodological considerations, or from real life examples obtained elsewhere. These characteristics have important consequences on the appropriate strategy that should be used for the analysis of survival data :

- When the time scale is continuous and the Weibull assumption is reasonable, Cox and Weibull models give almost identical results (see Ducrocq and Casella, 1996, for other simulated examples).
- This is true even when censoring rate is substantial, at least in relatively balanced situations.
- When the time scale is discrete, the Cox model is no longer adequate.
- However, at least when the « underlying continuous baseline » remains distributed as Weibull, the Weibull model seems remarkably robust (see Lubbers et al., 1999 for another example of such robustness).
- When this is not the case, the grouped data model should be the method of choice, as it does not require any particular assumption about the shape of the baseline distribution.

References

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

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

Ducrocq, V., 1997. Survival analysis, a statistical tool for longevity data. *48th Annual meeting of the EAAP*, Vienna, Austria.

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

Ducrocq, V., and Sölkner, J. 1998. «The Survival Kit V3.0», a package for large analyses of survival data. *Proc. 6th World Congr. on Genet. Appl. to Livest. Prod*., 27: 447-448.

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

Lubbers, R., Brotherstone, S., Ducrocq, V. and Visscher, P., 1999. A simple comparison of a linear and proportional hazards approach to analyse longevity in dairy cows using discrete data. Submitted to *Animal. Science.*

Miller, R. 1981. *Survival analysis*. Wiley, New-York, USA

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

Ricard, A and Fournet-Hanocq, F., 1997. Analysis of factors affecting length of competitive life of jumping horses. *Genet. Sel. Evol.*, 29 : 251-267.