Topics that may deserve further attention in survival analysis applied to dairy cattle breeding – some suggestions

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

Abstract

For the past ten years, the need to account for specificities of survival data (censoring, skewed distributions, time-dependent covariates) has lead to the development of methods, models, programs and routine genetic evaluations that are becoming standards. In this paper, I try to present some potential directions for further research, concentrating on modelling. These directions include the computation of cow EBVs, the use of time-dependent sire effects, the implementation of multivariate analyses, the definition of survival traits on a lactation basis and the investigation of culling components. Obviously, this list is not exhaustive and some topics may be perhaps premature. They should be considered as the basis for discussion and for projects of collaborations.

1. Introduction

Survival analysis applied to animal breeding has gained popularity since the pioneering work of Smith (1983). From my obviously biased perspective, the main steps of its development were: the extension of the Cox model to a mixed model for sire evaluation (Smith, 1983; Smith and Quaas, 1984), the use and justification of a Weibull model with time-dependent covariates (Ducrocq, 1987), the availability of a general program for applications of mixed models involving time-dependent covariates to large data sets (Ducrocq and Sölkner, 1994), the design of proper methods for the estimation of genetic parameters (Ducrocq and Casella, 1996; Korsgaard et al., 1998) and the implementation of routine genetic evaluations (see these proceedings). The same general evolution as for other related fields in animal breeding (e.g., for discrete data) can be observed: advanced methods were developed and used because they more precisely describe the statistical and biological characteristics of the data at hand. Initially, they were restricted to simple models or limited size data sets but a better understanding of their nature, a constant increase in computing power, and the use of more efficient algorithms made them applicable to larger and more complex problems. There is no reason to believe that such a trend will stop in the near future. Here, I will concentrate on some potential directions for model improvement. I will voluntarily exclude other important research topics, such as on economic aspects of longevity, on improvement of evaluations using early predictors, on its use in selection schemes, etc…

2. Animal models versus sire - maternal grand sire models

2.1. Context

Genetic evaluations based on survival analysis have been developed so far considering sire or sire-maternal grand sire models only. Some people are concerned because only bull EBVs are computed, when they consider as essential to supply breeders with cow EBVs too. They are very reluctant to use pedigree values for cows, although the low heritability traits implies that the own performance of the cow –especially if she is still alive - would probably not influence much her EBV.

At the same time, *one* particular approach for estimating genetic variance in frailty models (the Laplacian integration technique used to find an approximate marginal posterior density of this genetic variance) was found to give biased results when a survival analysis model was used on a simulated data set (Ducrocq and Casella, 1996). One of the explanations given was that on that simulated data set, a very simple pedigree structure was assumed with no information at all coming from female relationships.

Unfortunately, this lead to the general belief that survival analysis could not be applied to animal models and that other methods perhaps less adapted to survival data should be preferred.

It is important to restate that there is nothing in frailty (mixed) models theory that prevents the use of an animal model. Such models have been applied in other contexts (Korsgaard et al., 1998; Ducrocq et al., 1999). The main problem is computational: large scale applications based on sire models are already computationally very demanding and national evaluations based on an animal model cannot be envisioned in the near future. However, it will be showed that approximate animal model EBVs can be obtained.

2.2. The animal survival model:

The approximate procedure to get cow EBVs for longevity requires a formal presentation of the "correct" model (model without any approximation). Using the classical mixed model notations, let $\mathbf{x}_m$ and $\mathbf{z}_m$ be the vectors of explanatory variables relating the failure time of animal m to the fixed and random effects vectors $\boldsymbol{\beta}$ and $\mathbf{a}$. For the time being and without loss of generality, $\mathbf{x}_m$ and $\mathbf{z}_m$ will be considered as time-independent and we will assume that there is only one random effect: vector $\mathbf{a}$ represents the additive genetic value of all animals with observations, and their ancestors. Let:

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

The hazard function h(t) and $\mathbf{w}_m$ are related assuming a Weibull proportional hazards model:

$$h(t\,;\mathbf{w}_m) = \rho t^{\rho-1} \exp\{\,\mathbf{w}_m{}'\boldsymbol{\theta}\} \qquad [1]$$

$\boldsymbol{\theta}$ includes an "intercept term" $\rho\log\lambda$, where ρ and λ are the parameters of the baseline Weibull distribution. A multivariate normal distribution is a natural choice for the random effects $\mathbf{a}$:

$$\mathbf{a} \sim MVN(\mathbf{0}, \mathbf{A}\sigma_a^2)$$

where σ_a^2 is the additive genetic variance and $\mathbf{A}$ is the relationship matrix between all animals.

If $\mathbf{y}$ is the vector of observations (failure times + censoring codes), and if effects other than $\mathbf{a}$ have flat priors, the joint posterior density of all parameters is (Ducrocq and Casella, 1996):

$$\log p(\theta,\rho,\sigma_a^2|\mathbf{y}) = \left(N\log\rho + (\rho-1)\sum_{\{\text{unc.}\}}\log y_m + \sum_{\{\text{unc.}\}}\mathbf{w}'_m\theta \right) - \sum_{\{\text{unc., cens.}\}} y_m^{\rho} e^{\mathbf{w}'_m\theta} - \frac{N_a}{2}\log\left(2\pi\sigma_a^2\right) - \frac{1}{2}\log|\mathbf{A}| - \frac{1}{2\sigma_a^2}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} \qquad [2]$$

{unc.} and {cens.} represent the sets of uncensored and censored observations, respectively. If σ_a^2 is assumed known (e.g., $\sigma_a^2 = 4\hat{\sigma}_s^2$ and $\hat{\sigma}_s^2$ is the sire variance used in the sire-maternal grand sire model):

$$\log p(\theta,\rho|\mathbf{y},\ \sigma_a^2) = \left(N\log\rho + (\rho-1)\sum_{\{\text{unc.}\}}\log y_m + \sum_{\{\text{unc.}\}}\mathbf{w}'_m\theta \right) - \sum_{\{\text{unc., cens.}\}} y_m^{\rho} e^{\mathbf{w}'_m\theta} - \frac{1}{2\sigma_a^2}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} + \text{cons tan t} \qquad [3]$$

Estimates of $\boldsymbol{\theta}$ and ρ are obtained at the mode of the log posterior density. At the mode, the vector of its first derivatives with respect to each parameter is $\mathbf{0}$. This maximisation is *exactly* what is currently implemented in the Survival Kit, *without any modification for an animal model.* Again, the only limitation is the resulting slow convergence, This lead to often prohibitive CPU requirements. This is the motivation for looking for an approximate procedure.

2.3. A two-step procedure:

From [2], we have at the mode, for a particular animal m:

$$0 = \frac{\partial \log p(\theta,\rho|\mathbf{y},\ \sigma_a^2)}{\partial a_m} = \frac{\partial}{\partial a_m}\left(\sum_{\{\text{unc.}\}}\mathbf{w}'_m\theta - \sum_{\{\text{unc., cens.}\}} y_m^{\rho} e^{\mathbf{w}'_m\theta} - \frac{1}{2\sigma_a^2}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} \right)$$

Hence, if δ_m is the censoring code ($\delta_m = 1$ if animal m is uncensored; $\delta_m = 0$ if m is censored):

$$0 = \delta_m - y_m^{\rho} e^{\mathbf{w}'_m\boldsymbol{\theta}} - \frac{1}{\sigma_a^2}\left(\mathbf{A}^{-1}\mathbf{a}\right)_m \qquad [4]$$

We will assume that all fixed effects and additive genetic effects for males are known and equal to their estimates obtained from the sire-maternal grand-sire models. We will also assume that cow m does not have any progeny, that her own dam does not have any observation and that only her sire (the maternal-grand sire of m) is known. Then:

$$\left(\mathbf{A}^{-1}\mathbf{a}\right)_m = d_m^{-1}\left(a_m - \frac{1}{2}a_s - \frac{1}{4}a_{mgs} \right) = d_m^{-1}\,\phi_m \qquad [5]$$

where d_m represents the fraction of total genetic variance in ϕ_m (11/16 if the sire and maternal grand-sire are known). If we find an approximation of $\hat{\phi}_m$, we could approximate a_m as:

$$a_m = \frac{1}{2}\hat{a}_s + \frac{1}{4}\hat{a}_{mgs} + \hat{\phi}_m \qquad [6]$$

Combining expressions [4] and [5], the MAP estimate of ϕ_m is the solution of the equation:

$$\delta_m - y_m^{\hat{\rho}} e^{(x'_m\hat{\beta} + \frac{1}{2}\hat{a}_s + \frac{1}{4}\hat{a}_{mgs} + \phi_m)} - \frac{d_m^{-1}}{\sigma_a^2}\phi_m = 0 \qquad [7]$$

or, if $\hat{r}_m = y_m^{\hat{\rho}} e^{(x'_m\hat{\beta} + \frac{1}{2}\hat{a}_s + \frac{1}{4}\hat{a}_{mgs})}$ [8]

$$\hat{r}_m e^{\phi_m} + \frac{d_m^{-1}}{\sigma_a^2}\phi_m - \delta_m = f(\phi_m) = 0 \qquad [9]$$

The nonlinear equation $f(\phi_m)=0$ can be easily solved iteratively, for example using Newton's algorithm. Take, e.g., $\phi_m^{(0)} = 0$. At iteration k:

$$\phi_m^{(k+1)} = \phi_m^{(k)} - \frac{f(\phi_m^{(k)})}{f'(\phi_m^{(k)})}$$

If ϕ_m is small, one can use the approximation $e^{\phi_m} \approx (1+\phi_m)$ and expression [9] leads to:

$$\hat{\phi}_m = \frac{\delta_m - \hat{r}_m}{\frac{d_m^{-1}}{\sigma_a^2} + \hat{r}_m} = \frac{\delta_m - \hat{r}_m}{d_m^{-1} + \hat{r}_m\sigma_a^2}\ \sigma_a^2 \qquad [10]$$

Although this formula is not correct if ϕ_m is large, it illustrates a few points:

- It is the result of the first iteration of Newton's algorithm when $\phi_m^{(0)} = 0$.
- Expression [8] for $\hat{r}_m$ is the estimate of generalised residual (Cox and Snell, 1966) of the observation on animal m. If the Weibull sire-maternal grand-sire model is correct, the generalised residuals are distributed as a unit (censored) exponential, of mean and variance 1 (Cox and Oakes, 1984). When an animal dies (δ_m =1) with $\hat{r}_m$ equal to the mean value $\hat{r}_m$ =1, then $\hat{\phi}_m$=0. If animal m dies very quickly and $\hat{r}_m$ is very small, $\hat{\phi}_m \approx d_m\ \sigma_a^2$. This is the largest positive value it can take.
- More importantly, the evolution of $\hat{\phi}_m$ for censored records (δ_m =1) is of interest: initially, $\hat{r}_m$ is very small: $\hat{\phi}_m \approx 0$, so $\hat{a}_m \approx \frac{1}{2}\hat{a}_s + \frac{1}{4}\hat{a}_{mgs}$, i.e., its pedigree value. Then, as time goes, $\hat{\phi}_m$ becomes more and more negative, corresponding to a better EBV $\hat{a}_m$ (less risk of being culled). *But as soon as the record is uncensored (δ_m=1), the cow EBV jumps up by a value of* $\frac{1}{d_m^{-1} + r_m\sigma_a^2}\ \sigma_a^2$ *!!*
- On average, this jump brings back the cow's EBV to her pedigree value.

If needed, this whole derivation can be started again to lead to more precise EBVs, by exactly solving the nonlinear equation [9] or by relaxing some of the assumptions, e.g., on the knowledge of grand-dams or the existence of daughters. But this does not elude another question: does it make sense to publish proofs for censored cows? These proofs will change over time until the animal is censored and are not equal to the expected value of what they will be if the animal dies the next day.

Evidently, more work is needed on this topic and more generally, on the accuracy, usage and relevance of cow EBVs.

3. Time-dependent sire effects

Classically, in survival models used for genetic evaluation, it is assumed that sire effects are time-independent random effects, i.e., constant over time. But functional longevity is a complex trait: there is a long list of events leading to involuntary cullings. These events do not have the same probability of occurring during the life of the cow: e.g., fertility or mastitis problems are more likely to occur later in life. At the same time, the respective contribution of their genetic component to the genetic merit for functional longevity may be very different from one sire to the other. For many bulls, this may not be so crucial: for example, the Kaplan-Meier survivor curves of the daughters of the three bulls in figure 1 do not suggest that their ranking on functional longevity EBV may change as time goes, at least during, say, the first three lactations. If one is concerned about sire reranking later in life, a simple strategy would be to censor all records of cows still alive after 3 lactations or after 1000 days. It would be interesting to have a look at the impact of such a censoring rule, for example on estimates of genetic variances and on sire ranking. I have never seen this strategy applied but it is attractive in the sense that it would explicitly look at what we want to improve most: the ability to delay *early* cullings which are the most costly ones.

Figure 1: Kaplan-Meier survivor curves of the daughters of three bulls which seem to have a constant effect on survival

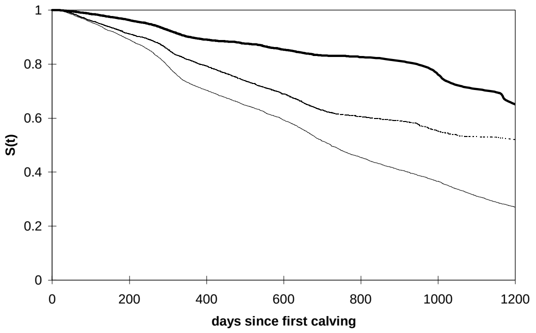


Figure 2 displays the survival curve of the 27971 daughters of a bull with a good EBV for functional longevity, but which went down (-0.7σ_g) when a large number of these daughters finished their second lactation: within lactation, culling rate is quite homogeneous but clearly, it differs between the first and the second lactation. For another bull with very poor EBVs on most functional traits (for functional longevity: -1.9σ_g, on somatic cell score: -1.7σ_g, on udder type: -1.8σ_g, on milking speed: -1.2σ_g) despite a good production EBV, the survivor curve of his daughters (figure 3; for first crop daughters only, as he did not get any second crop ones) displays a similar pattern across lactations but clearly shows within-lactation differences with 2 visible bumps, after about 100-150 days and at the end of each lactation. Although based on raw data, these two examples are illustrations of potentially time-dependent sire effects on survival.

Figure 2: Kaplan-Meier survivor curve of the daughters of a bull which seems to have a different effect on survival during the first and second lactations (dotted line: 95% confidence interval)

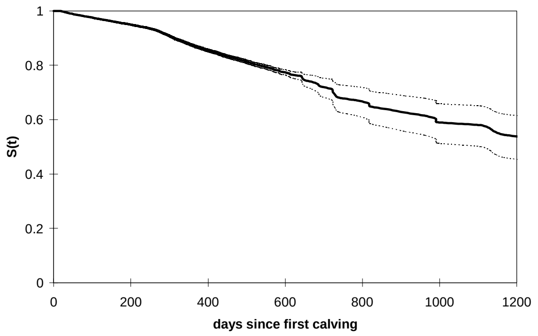
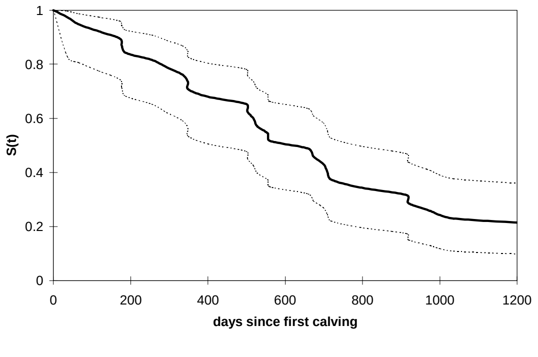


Figure 3: Kaplan-Meier survivor curve of the daughters of a bull which seems to have a different effect on survival within lactation (dotted line: 95% confidence interval)



Indeed, nothing prevents us from extending the present evaluation models by assuming *time-dependent* sire effects. For example, a different sire effect could be assumed in first, second and later lactations. The same methodology as described in Ducrocq and Casella (1996) could be applied, using as a prior distribution for sire effects:

$$\mathbf{s}(\mathrm{t}) = \begin{pmatrix} \mathbf{s_1} \\ \mathbf{s_2} \\ \mathbf{s_3} \end{pmatrix} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{G} \otimes \mathbf{A}) \qquad [11]$$

where $\mathbf{s}_\mathrm{i} = \{ \mathrm{s_{qi}} \}$ represents the vector of sire effects in lactation i and **G** is the variance-covariance matrix between sire x lactation effects. Obviously, there are drawbacks to such an approach:

- How many and on what basis should the periods be defined ? Too many of them would lead to very imprecise sire x period EBV. A definition on a lactation basis makes sense but does not reflect situations such as for the bull in figure 3. Note however that within lactation changes in sire effects could be described with, say, random regression models, but that does not make the problem simpler…
- How do we estimate the genetic (co)variances in **G** ?
- In practice, how much do we gain in terms of efficiency of selection or on accuracy, in a situation already characterised by lack of interest from the breeders, because longevity evaluations come too late and have poor reliabilities? In other words, is it worth the effort ? I am not too optimistic about it, but from a research perspective, this may help to understand what we are doing.

4. Multivariate analyses

So far, only *univariate* survival analyses of longevity traits have been performed in animal breeding. Again, as for the animal model situation described in section 2, the apparent inability to extend survival models to multivariate ones has been put forward to use less sophisticated approaches, that *may* perform well in some real life situations but that do not account for peculiarities of survival data (censoring, skewed distributions, time-dependent explanatory variables, etc..).

Let's first see where the "problem" lies: in the case of a Weibull regression model and in absence of censoring, the typical Weibull hazard model [1] can be redefined as a log-linear mixed model (Kalbfleich and Prentice, 1980): if "1" refers to the survival trait and $\mathrm{y_{1m}} = \log \mathrm{T_{1m}}$ is the failure time of animal m, this log-linear model states that:

$$\mathrm{y_{1m}} = \mathbf{x}'_{1m}\beta^*_1 + \mathbf{z}'_{1m}\mathbf{s}^*_1 + \mathrm{e_{1m}} \qquad [12]$$

with $\beta^*_1 = -\rho_1^{-1}\beta$, $\mathbf{s}^*_1 = -\rho_1^{-1}\mathbf{s}$ and $\mathrm{e_{1m}} = \rho_1^{-1}\omega_{1m}$,

and ω_{1m} follows an extreme value distribution. In other words :

$$\mathrm{y_{1m}} = -\frac{1}{\rho_1}\mathbf{x}'_{1m}\beta_1 - \frac{1}{\rho_1}\mathbf{z}'_{1m}\mathbf{s}_1 + \frac{1}{\rho_1}\omega_{1m} \qquad [13]$$

We would like to analyse this first trait simultaneously with a second trait described by the following mixed model :

$$y_{2m} = \mathbf{x}'_{2m}\beta^*_2 + \mathbf{z}'_{2m}\mathbf{s}^*_2 + e_{2m} \qquad [14]$$

Note that this second trait can be a "linear" trait, such as a type trait or milk production, in which case e_{2m} is supposed to be normally distributed, or it can be another survival trait. Then, we will assume that y_{2m} = log $T_{2m,}$ that a different Weibull parameter ρ_2 is involved and that e_{2m} is also proportional to an extreme value distribution.

Let $\mathbf{\Sigma}$ be the set of dispersion parameters and $\boldsymbol{\theta}_\mathbf{i}$=($\boldsymbol{\beta}_\mathbf{i}$, ρ_i , $\mathbf{s_i}$). A direct application of Bayes' theorem leads to the joint posterior density (with obvious extensions of notations and removing the star '*' for clarity):

$$\begin{aligned} &p(\theta_1,\theta_2,\Sigma \mid \mathbf{y_1},\mathbf{y_2}) \\ &\quad \propto p(\mathbf{y}_1,\mathbf{y_2} \mid \theta_1,\theta_2,\Sigma)\ p(\theta_1,\theta_2 \mid \Sigma)\, p(\Sigma) \end{aligned} \qquad [15]$$

Traditionally, a flat prior is assumed for fixed effects and the Weibull parameters in $\boldsymbol{\theta}$, and often (but this is not an obligation) for $\mathbf{\Sigma}$. Also, a joint multivariate normal distribution is chosen for the genetic effects. The difficulty lies in the derivation of the joint density $p(\mathbf{y}_1,\mathbf{y_2} \mid \theta_1,\theta_2,\Sigma)$. If the two traits are observed in independent environments (e.g., on different sets of daughters), we can write:

$$p(\mathbf{y}_1,\mathbf{y_2} \mid \theta_1,\theta_2,\Sigma) = p(\mathbf{y}_1 \mid \theta_1,\Sigma)\, p(\mathbf{y}_2 \mid \theta_2,\Sigma) \qquad [16]$$

With censoring, expression [16] must be modified to take into account that the contribution of censored records is not the density function at failure time but the survivor function at censoring time (Kalbfleisch and Prentice, 1980).

If traits 1 and 2 are observed on the same animal, the independence of residuals can no longer be supposed. *I don't know any bivariate distribution whose marginals are either a normal and an extreme value distributions or two extreme value distributions!*

Another approach must be used. Let's review these two cases:

- Traits in different environments: indeed, there are quite a few situations when we may be interested in the joint analysis of two survival traits in two distinct environments and the estimation of the genetic correlations between these two traits. For example, sire effects on functional longevity are likely to differ between temperate and tropical or harsh environments, between mountainous and flat regions or between intensive and extensive management systems.

Using equation [16], it is possible to combine the likelihood contributions of both traits and to maximise the resulting posterior density [15] or its logarithm to get estimates of $\boldsymbol{\theta}$. To estimate the genetic parameters (genetic variances and covariances in $\mathbf{\Sigma}$), Laplacian integration could be applied as in Ducrocq and Casella (1996) to get an approximate marginal posterior density p($\mathbf{\Sigma}$ | $\mathbf{y}$). Modal estimates of this approximate density can be obtained or, if the interest is on the genetic correlation ρ between the two traits, Laplacian integration can be applied directly to obtain the approximate marginal posterior density p(ρ | $\mathbf{y}$) along the same lines as in Hofer and Ducrocq (1997). Of course, other techniques such as Gibbs sampling (as in Korsgaard et al, 1998) can be used to obtain the exact marginal posterior densities of these parameters. In all cases, as in most multivariate analyses, computing costs will increase dramatically but there is no conceptual difficulty involved.

As a final remark, note that often, the data can be sampled in such a way that a residual correlation is forced to be 0 (Larroque and Ducrocq, 1999).

- Traits with nonzero residual correlations

A typical example is the joint analysis of a survival measure and a type trait, in particular in order to estimate the genetic correlation between the two traits, without the approximation of, e.g., Larroque and Ducrocq (1999). Perhaps the best that we can do is to adopt the same approach as Foulley et al. (1983), which was also used with some adaptations in Janss and Foulley (1993) or Le Bihan-Duval et al. (1997):

Assume that for each animal, both traits are present (no missing trait). We will rewrite [12] as:

$$y_{1m} = \mathbf{x}'_{1m}\beta^*_1 + \mathbf{z}'_{1m}\mathbf{s}^*_1 + E(e_{1m} \mid e_{2m}) + e^*_{1m} \qquad [17]$$

where e^*_{1m} is an error term uncorrelated with the error term for the other trait e_{2m}.

$E(e_{1m} \mid e_{2m}) = b\,\hat{e}_{2m}$ is the regression of e_{1m} on $\hat{e}_{2m}$ and takes into account the residual correlation ρ_e through the residual regression coefficient b= $\rho_e\,\sigma_{e1}/\sigma_{e2}$. Note that Var($e^*_{1m}$) = $(1-\rho^2_e)\sigma^2_{e1}$. If ω_{1m} still has an extreme value distribution, the rescaling of e^*_{1m} will be through a new estimate of the Weibull parameter ρ_1 (ρ_1 and the rescaling factor are confounded). Expression [17] allows the following decomposition of the joint likelihood :

$$\begin{aligned} &p(\mathbf{y}_1,\mathbf{y_2} \mid \theta_1,\theta_2,\Sigma) \\ &\quad = p(\mathbf{y}_1 \mid \theta_1,\theta_2,\mathbf{y}_2,\Sigma)\ p(\mathbf{y}_2 \mid \theta_2,\Sigma) \end{aligned} \qquad [18]$$

As previously, censoring should be accounted for, by a proper calculation of the likelihood contributions. The posterior density [15] can be computed either to jointly estimate θ_1 and θ_2 assuming that $\boldsymbol{\Sigma}$ is known, or as the starting point for applying, e.g., Laplacian integration to get an approximate marginal posterior density of the parameters in $\boldsymbol{\Sigma}$. Again, the main challenge lies in the computational implementation of this approach

When for some animals, one or the other of the traits is missing, it is possible to distinguish subsets with homogeneous information available (only one trait or both) as in Janss and Foulley (1993) or Le Bihan-Duval et al. (1997). The contributions of the likelihood of each subset can be combined since they are conditionally independent given $\boldsymbol{\theta}$.

When the second trait is also a longevity trait (e.g., two longevity measures determined by different culling reasons), a joint analysis seems even more difficult, in particular because definition of generalised residuals (Cox and Snell, 1968) does not seem to fit here for the computation of $\hat{e}_{2m}$. Care must also be taken in the computation of likelihood contributions of censored records (censored for one or the other or both traits).

5. A model for survival on a lactation basis

In most of the current genetic evaluation models that are based on survival analysis, stage of lactation x lactation number (SLLN) effects are included as time dependent covariates (these proceedings). The stage of lactation effect is included to account for changes in culling policy during the lactation: it is assumed (and verified) that culling is more intense at the end of the lactation, when production is lower, when it is known whether the cow is pregnant or not and when her carcass value is better. Classes of stage of lactation are defined somewhat empirically in order to model these changes by a piecewise constant function. However, as SLLN effects are changing according to a predefined sequence as time goes, the exact interpretation of its effect must be done after combining it with the baseline as (Ducrocq, 1999):

$$\hat{h}(t) = \hat{\rho}\, t^{\hat{\rho}-1} * \exp\{\, \hat{l}_j(t,t')\,\} \qquad [19]$$

where $l_j(t,t')$ represents the j^{th} SLLN effect. Figure 4 shows a typical plot of the change in hazard during the lactation. It is tempting to conclude from such a graph that the hazard pattern is more or less the same for each lactation, with a regular (convex) increase of the risk during the lactation, and that this regularity is partly broken by the arbitrary choice of boundaries for SLLN effects. A more regular curve would imply a thinner definition – i.e., more classes- of stages of lactation, but this has already a huge drawback : at least with the approach used in the Survival Kit (Ducrocq and Sölkner, 1994), each change in SLLN induces the creation of a new « elementary record », a record covering an interval on which no time-dependent covariate changes. This makes the recoded file bigger and bigger.

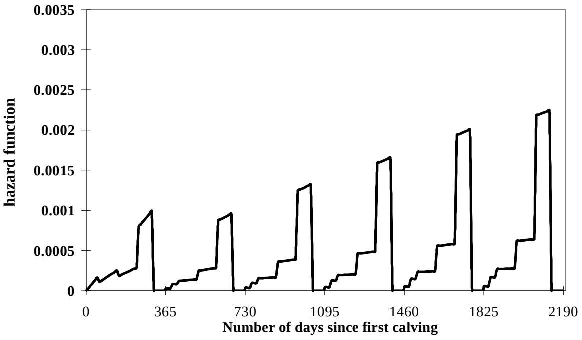
This leads to the suggestion of an alternative model, for which survival is looked at on a lactation basis rather than on the whole length of productive life.

The current longevity measure uses the date of first calving as starting point and the date of culling as end point. The proportional hazard model is written as:

$$h(t) = h_0(t)\exp\{\, hy_r(\tau) + \textstyle\sum_k f_k(t) + l_j(t,t') + s_u + 0.5 s_{gs}\,\} \qquad [20]$$

where $hy_r(\tau)$ is the herd year effect, s_u and s_{mgs} are the sire and maternal grand sire effects and the f_k's are the other fixed effects, some of them being time dependent (see, e.g., Ducrocq, 1999).

Figure 4: Hazard function of a reference cow with constant lactation length and calving interval, in the Normande breed



For the alternative model, each lactation of each animal is treated as *one* survival measure : the origin point is date of calving, the end point is date of culling or date of next calving (whichever comes first) and of course, the record is censored if the cow starts a new lactation. The hazard of cow m is:

$$h(t) = h_{0,j}(t)\exp\{\, hy_r + \textstyle\sum_k f_k + c_m + s_u + 0.5 s_{gs}\,\} \qquad [21]$$

where $h_{0,j}(t)$ is a Weibull baseline hazard specific to each lactation, hy_r is the herd year effect when the lactation was started (i.e., time-independent), and the f_k's are the other fixed effects, which can be time-dependent but most often are not (e.g., variation in herd size or deviation in milk production can be considered as time-independent within lactation). c_m is an extra random effect (cow effect), which describes the shared unobservable genetic and non genetic characteristics specific to cow m that affect her hazard during all her lactations. This definition is along the lines of the definition of the random terms in the frailty models advocated by Vaupel et al.(1979) or Clayton and Cuzick (1985).

I see some important attractive features for such a model:

- It allows a continuous description of the SLLN effect in figure 4, avoiding the definition of arbitrary stage of lactation intervals. If a Weibull baseline hazard stratified by lactation is chosen, [19] is replaced by:

$$h(t) = \text{constant} \times t^{\rho_j - 1} \qquad [22]$$

The term $t^{\rho_j -1}$ should account for the convexity of the lactation curves in figure 4. At the beginning of the lactation, h(t)=0 : this model cannot describe a « bathtub » hazard, with a phase with decreasing hazard followed by an increasing hazard. If culling risk is important at the very beginning of the lactation, this may be a drawback. Then, a time dependent stage of lactation effect may have to be added… Note however that if survival information comes from milk recoding schemes, very early cullings occur before the first 'potential' test date and the 'apparent' culling is considered as occurring late during the previous lactation.

- From a computational point of view, this model leads to a drastic reduction of the size of recoded data file: with the current models, the average number of elementary records per cow is large (e.g., 19.4 for the French evaluation, see Ducrocq, 1999). With the alternative model, this number would go down to about 3, i.e., a reduction by a factor of about 6, maybe making the analysis possible without the time-consuming compression - decompression approach (Ducrocq, 1999).
- This approach may offer a much nicer framework for multiple trait analyses (e.g., within lactation milk production and survival) or for time-dependent sire x lactation effects.

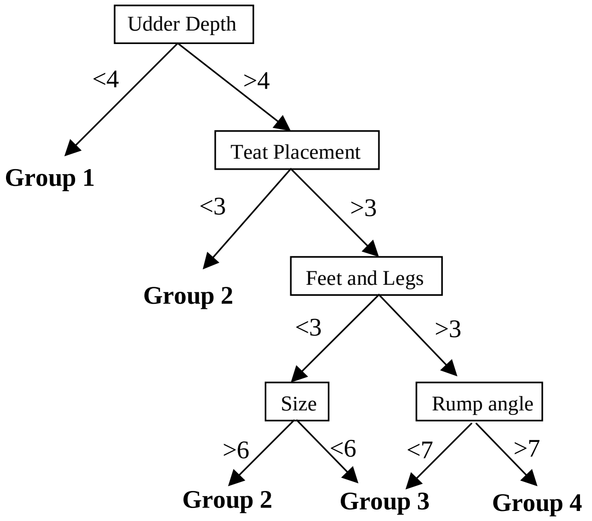
The main drawback of this approach is, I think, the need to include an extra random effect: the cow effect. What does it change? Do we loose accuracy by including it in the model? What distribution do we choose for it? What variance does it have? How do we estimate it? Can the effect be integrated out? What is its consequence on convergence rate? There are a lot of interesting questions to work on!

Note that successful applications of survival analysis on a lactation basis (but with only one lactation considered per cow, i.e., without cow (c_m) effect) can be found in epidemiology (e.g., Gröhn et al.,1997, 1998).

6. Towards a better understanding of the components of longevity

A better knowledge of the relationships between traits will be obtained by getting more precise estimates of the genetic and residual correlations between them, either from true multivariate analyses or from approximate approaches (Druet et al; 1999; Larroque and Ducrocq, 1999). But my impression is that our perception of the factors influencing culling rates is to some extent limited when it is only based on the estimates of regression parameters or of correlations. To go one step further, I suggest to try to decompose the culling decisions into different components in a more didactic way. For example, can we identify groups of animals with, say, different type characteristics that are homogeneous (within group) but with very different survivor curves from one group to another? Specific tools have been developed by statisticians to tackle such questions. Discriminant analysis is one such tool. Another one is classification and regression tree (CART) models (Breiman et al., 1984, Chipman et al., 1998). These are used to create binary trees that recursively partition the space of explanatory variables into subsets in which the distribution of the dependant variable (survival time in our case) is more homogeneous. Figure 5 illustrates what kind of results CART models can lead to, using phenotypic scores on type traits as predictors.

Figure 5: a completely invented regression tree which defines 4 groups of animals with different type phenotypes that have distinct, but homogeneous (within group) survivor curves



There are quite a few difficulties involved: at each stage of the recursive partitioning, groups are divided into two "nodes" as distinct as possible from each other. This is decided according to the value of some function of the data, for example a likelihood function or a posterior density in a

Bayesian analysis (Chipman et al., 1998). As many such functions may be computed when the limit separating nodes is changed, the resulting algorithm can be extremely time consuming. Also, there is a balance to be found between maximising homogeneity within group and minimising the number of groups. It seems recommended to start by constructing a large tree with many terminal nodes and then to reduce it by "pruning", combining the groups that are most similar. Finally, cross-validation is advisable, if not compulsory when such a technique is used.

Note that CART models could be also useful to tackle one of the research topics indicated in Beaudeau et al. (1999), regarding the interpretation of herd-year effects and their components.

Conclusion

As it can be seen, most of the topics described here are in their early stages of development and some of them may be unworkable, unpractical or very premature. My hope is simply that they could be the basis for discussions and for future development of fruitful collaborations.

References

Beaudeau, F., Seegers, H., Ducrocq, V., Fourichon, C. Effect of health disorders on culling in dairy cows: a review and critical discussion. *These proceedings.*

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. Wadsworth, Belmont, CA, USA.

Chipman H.A., George, E.I., McCulloch, R.E. 1998. Bayesian CART model search. *J. Amer. Stat. Assoc.*, 93:935-947.

Clayton, D. G., Cuzick, J., 1985. Multivariate generalizations of the proportional hazards model. *J. Royal. Stat. Soc.* (Series A), 148:82-117.

Cox, DR, 1972. Regression models and life-tables. *J. Royal. Stat. Soc.* (Series B), 34: 187-220 (with discussion).

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

Cox, D. and Snell, E. (1968). A general definition of residuals. *J. Royal Stat. Soc.* (Series B}, 30:248-275 (with discussion).

Druet, T., Solkner, J., Groen, A.F., Gengler, N., 1999. Improved genetic evaluation of survival using MACE to combine direct and correlated information from yield and functional traits. *These proceedings.*

Ducrocq, V., 1987. An analysis of length of productive life in dairy cattle. PhD thesis, Cornell University, N.Y., U.S.A.

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

Ducrocq, V., 1999. Two years of experience with the French genetic evaluation of dairy bulls on production-adjusted longevity of their daughters. *These proceedings.*

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

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

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., 22: 51-52*. Guelph, Ontario, Canada.

Ducrocq, V., Besbes, B. and Protais, M., 1999. Genetic improvement of laying hens using survival analysis. *Submitted to Genet. Sel. Evol.*

Foulley J.L., Gianola D., Thompson R., 1983. Prediction of genetic merit from data on binary and quantitative variates with an application to calving difficulty, birth weight and pelvic opening. *Genet. Sel. Evol., 15 : 401-424.*

Gröhn, Y.T., Ducrocq, V. and Hertl, J.A. 1997. Modeling the effect of a disease on culling: an illustration of the use of time-dependent covariates for survival analysis. *J. Dairy Sci.* 80: 1755-1766.

Gröhn, Y.T., Eicker, S.W., Ducrocq, V. and Hertl, J.A. 1998. Effect of diseases on the culling of Holstein dairy cows. *J. Dairy Sci* 81: 966-978.

Hofer, A., Ducrocq, V, 1997. Computing marginal posterior densities of genetic parameters of a multiple trait animal model using Laplace approximation or Gibbs sampling. *Genet. Sel. Evol, 29, 427- 450.*

Janss, L.L.G., Foulley, J.L., 1993. Bivariate analysis for one continuous and one threshold dichotomous trait with unequal design matrices and an application to birth weight and calving difficulty. *Livest. Prod. Sci., 33 : 183-198.*

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

Korsgaard, I.R., Madsen, P. and Jensen, J., 1998. Bayesian inference in the semiparametric lognormal frailty model using Gibbs sampling. *Genet. Sel. Evol.*, 30: 241-256.

Larroque, H., Ducrocq, V., 1999. An indirect approach for the estimation of genetic correlations between longevity and other traits. *These proceedings.*

Le Bihan-Duval E., C. Beaumont, J.J. Colleau, 1997. Estimation of the genetic correlation between twisted legs and growth on conformation traits in broiler chickens. *J. Anim. Breed. Genet., 114:239-259.*

Smith, S.P. (1983): The extension of failure time analysis to problems of animal breeding. Ph.D. Thesis, Cornell Univ., Ithaca, NY.

Smith, S.P., Quaas, R.L. (1984): Productive lifespan of bull progeny groups: failure time analysis. *J. Dairy Sci. 67: 2999-3007.*

Vaupel, J., Manton, K.G., Stallard, E., 1979. The impact of heterogeneity in individual frailty and the dynamics of mortality. *Demography,* 16:439-454