A Two-Step Procedure to get Animal Model Solutions in Weibull Survival Models Used for Genetic Evaluations on Length of Productive Life

Vincent Ducrocq
Station de Génétique Quantitative et Appliquée, Département de Génétique Animale
Institut National de la Recherche Agronomique,
78352 Jouy en Josas, France
email: vincent.ducrocq@dga.jouy.inra.fr

Abstract

Using results from sire-maternal grand sire survival models, a strategy was developed and tested to compute accurate approximations of animal model solutions for length of productive life of dairy cows. The procedure was incorporated into the "Survival Kit" software. Applications to simulated and real data sets indicate that the formal inconsistency of the Weibull sire-maternal grand sire model currently used in national evaluations has no real impact on sire EBVs. The main restriction of such sire-maternal grand-sire models comes from the ignorance of relationships between females. The article also presents the computation of individual "pre-adjusted records" and weights for length of productive life that could be used for BLUP multiple trait animal model evaluations naturally combining direct and indirect information on longevity.

1. Introduction

In several countries, a routine genetic evaluation of bulls on length of productive life of theirs daughters has been implemented assuming a proportional hazards model (Ducrocq and Sölkner, 1998a). The evaluation relies on the modelling of a hazard function, which describes the limiting probability for a cow alive just prior to time t of being culled at time t. This allows a conceptually natural analysis of records from animals that are still alive (censored records) together with already culled animals (uncensored records).

The non-linear model used to describe the hazard function involves time-dependent fixed effects (e.g., herd-year-season, stage of lactation) which permit to precisely accounts for changes in culling policies over time. Not accounting for these effects would neglect important environmental parts, would reduce our capability to detect genetic differences between animals and would result in biased genetic evaluations.

So far, the French genetic evaluation is based on a sire-maternal grand-sire model (Ducrocq and Sölkner, 1998a; Ducrocq, 1999a). Similar evaluations have been (or are being) implemented in Germany, The Netherlands, Italy, Denmark, Switzerland and Austria. Cow EBVs are not available. It is believed that direct information on longevity from a single (often still alive) animal is not sufficient to reach cow EBVs with a satisfactory reliability. The development (Ducrocq et al., 2001) of a strategy to combine longevity data with indirect information from early predictors - type traits, somatic cell count and female fertility- leads to reconsider this assertion.

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., 2000). The main problem is computational: because they require the joint maximisation of a non-linear function of tens of thousands of parameters, large scale applications based on sire models are already computationally very demanding. National evaluations based on an animal model cannot be envisioned in the near future.

This paper extends and illustrates an approximate two-step procedure to get cow EBVs for longevity that we proposed earlier (Ducrocq, 1999b). The first step is the application of the

current sire-maternal grand sire frailty model. The second step aims at the estimation of the component of the animal's own additive genetic value which is not already known after the first step. It assumes that the estimates of male EBVs as well as estimates of all environmental effects are available.

2. The "exact" approach

To derive the proposed approximation, consider first the "correct" model (animal model without any approximation). Using classical notations, let $\mathbf{x}_m$ and $\mathbf{z}_m$ be the vectors of explanatory variables relating the length of productive life of animal m to the fixed and random effects vectors $\boldsymbol{\beta}$ and $\mathbf{a}$. To facilitate the presentation and without loss of generality, $\mathbf{x}_m$ and $\mathbf{z}_m$ will be supposed to be time-independent. Only one random effect is included: the additive genetic value. This limitation will be later relaxed in section 4. Let $\mathbf{a} \sim MVN(\mathbf{0}, \mathbf{A}\sigma_a^2)$ be the vector including the additive genetic values of all animals with observations and their ancestors. $\mathbf{A}$ is the relationship matrix between all animals. σ_a^2 is the additive genetic variance and is assumed to be 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 Let:

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

If a Weibull proportional hazards model is assumed, the hazard function h(t) is written:

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

In [1], $\boldsymbol{\theta}$ includes an "intercept term" $\rho \log \lambda$, where ρ and λ are the parameters of the baseline Weibull distribution.

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 can be written as (Ducrocq and Casella, 1996):

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

Here, {unc.} and {cens.} are the sets of uncensored and censored observations, respectively. Estimates of $\boldsymbol{\theta}$ and ρ are obtained at the mode of the log posterior density [2]. Then, the vector of its first derivatives with respect to each parameter is $\mathbf{0}$. Software, such as the Survival Kit (Ducrocq and Sölkner, 1998b) perform this maximisation, whether the considered model is an animal model or not.

3. A two-step procedure

Let's consider the particular equation that the additive genetic value of a particular animal m has to fulfil at the joint mode of [2]. Equating to 0 the first derivative of [2] with respect to a_m, we obtain:

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

i.e.,

$$\delta_m - y_m^{\rho} e^{\mathbf{w}'_m \boldsymbol{\theta}} - \frac{1}{\sigma_a^2} \left(\mathbf{A}^{-1} \mathbf{a} \right)_m = 0 \qquad [3]$$

where δ_m is the censoring code (δ_m =1 if animal m is uncensored; δ_m =0 if m is censored):

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 (first step). Let's also assume that animal m does not have any progeny, that its own dam does not have any observation and that only its sire (the maternal-grand sire of m) is known. Then:

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

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

$$\hat{a}_m = \frac{1}{2}\hat{a}_s + \frac{1}{4}\hat{a}_{mgs} + \hat{\phi}_m \qquad [5]$$

To find $\hat{\phi}_m$, combine expressions [3] and [4]: the MAP estimate of ϕ_m is solution of the equation:

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

Denote $\hat{r}_m = y_m \hat{\rho} e^{(x'_m \hat{\boldsymbol{\beta}} + \frac{1}{2}\hat{a}_s + \frac{1}{4}\hat{a}_{mgs})}$ [6]

Then, $\hat{\phi}_m$ is the solution of :

$$\hat{r}_m e^{\phi_m} + \frac{d_m^{-1}}{\sigma_a^2}\phi_m - \delta_m = f(\phi_m) = 0 \qquad [7]$$

This non-linear 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:

$$\phi_m^{(k+1)} = \phi_m^{(k)} - \frac{f(\phi_m^{(k)})}{f'(\phi_m^{(k)})} \qquad [8]$$

In practice, very few iterations of [8] are needed (typically 2 or 3) to get an exact solution of [7].

If ϕ_m is small, one can use the approximation $e^{\phi_m} \approx 1 + \phi_m$ and expression [7] leads to:

$$\hat{\phi}_m = \frac{\delta_m - \hat{r}_m}{\frac{d_m^{-1}}{\sigma_a^2} + \hat{r}_m} = \frac{\delta_m - \hat{r}_m}{d_m^{-1} + \hat{r}_m \sigma_a^2}\,\sigma_a^2 \qquad [9]$$

which is also the result of the first iteration of Newton's algorithm when $\phi_m^{(0)} = 0$.

It is interesting to note that expression [6] 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, with mean and variance 1 (Cox and Oakes, 1984). Expression [9] indicates that 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.

The evolution of $\hat{\phi}_m$ for censored records (δ_m =0) is also 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 animal's 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 animal's EBV to its pedigree value. Therefore, one has to be aware that the proofs of a censored animal naturally change over time, until the animal is uncensored.

4. A particular case: herd-year–season estimates are not available

As in other countries, the French genetic evaluation model for length of productive life includes a time-dependent random herd-year-season effect which accounts for local changes in culling policies with time (Ducrocq and Sölkner, 1998a; Ducrocq, 1999a). Its distribution is assumed to be log-gamma (γ, γ) with a constant value for γ equal to 4 (determined from preliminary analyses).

In the evaluation process, the herd-year-season effect is integrated out of the joint posterior distribution in [2] and is not explicitly computed. Therefore, the herd-year-season estimates which are part of $\hat{\boldsymbol{\beta}}$ in [7] are not directly available to compute $\hat{r}_m$.

Again, an approximation is necessary. Equating to 0 the first derivative of [2] with respect to a particular herd-year-season effect b_j leads to the approximate estimate:

$$\hat{b}_j = \log\left(\frac{n_j + \gamma}{\hat{R}_j + \gamma}\right) \qquad [10]$$

where n_j is the total number of uncensored records in herd-year-season j and $\hat{R}_j$ is the sum of the cumulative hazard functions over all animals at risk in this particular herd-year-season. Then, the value $\hat{b}_j$ obtained in [10] can be used to compute $\hat{r}_m$ in [7].

5. Use of an animal EBV for functional longevity in a total merit index

One of the motivations to implement an animal model evaluation for functional longevity was the inclusion of EBVs for this trait in a total merit index (TMI – ISU in French). The strategy that was chosen in France to compute the ISU was to approximate a multiple trait BLUP animal model evaluation on production and functional traits. The functional traits considered are somatic cells score, female fertility, functional longevity and some type traits believed to be early predictors of other functional traits. The main approximation relies on the replacement of raw data by pre-adjusted records, free of environmental effects and summarising repeated records (when such records exist) of a same animal into a single value. The general characteristics of this approach is described in Ducrocq et al. (2001).

In contrast with traits described by linear models, there is no obvious way to obtain the equivalent of "pre-adjusted records" for functional longevity: the nonlinearity of the model, the existence of time dependent effects as well as the presence of censored records impose again the use of an approximation.

The derivation of such an approximation was undertaken with the following objective in mind: the use of the pre-adjusted records y^*_m's (with appropriate weights ω_m's) in the simplest univariate BLUP evaluation based on an animal model:

$$y^*_m = \mu + a_m + e_m \qquad [11]$$

$$\text{with} \quad \text{var}(e_m) = \left(\frac{1}{\omega_m}\right)\sigma^2_e$$

should lead to animal EBVs as close as possible to the approximate EBVs obtained in the (non-linear) Weibull analysis.

Let $\mathbf{W}$=diag$\{\omega_m\}$ and take σ^2_e=1. We want:

$$\left[\mathbf{W} + \frac{1}{\sigma^2_a}\mathbf{A^{-1}}\right]\hat{\mathbf{a}} = \mathbf{W}\,\mathbf{y}^* \qquad [12]$$

or, for the equation of animal m:

$$\left[\omega_m \hat{a}_m + \frac{1}{\sigma^2_a}\left(\mathbf{A^{-1}}\hat{\mathbf{a}}\right)_m\right] = \omega_m\, y^*_m \qquad [13]$$

Starting back from equation [3] and partitioning the exponential term as $\mathbf{w}^{*'}_m\boldsymbol{\theta}^* = \mathbf{w}'_m\boldsymbol{\theta} + a_m$, one obtains at the mode of the posterior distribution of fixed and random effects:

$$\left(y^{\hat{\rho}}_m e^{\mathbf{w}^{*'}_m\hat{\boldsymbol{\theta}}^*}\right)e^{\hat{a}_m} + \frac{1}{\sigma^2_a}\left(\mathbf{A^{-1}}\hat{\mathbf{a}}\right)_m = \delta_m \qquad [14]$$

Writing $e^{\hat{a}_m} = \hat{a}_m + (e^{\hat{a}_m} - \hat{a}_m)$ and defining $\hat{r}^*_m = \left(y^{\hat{\rho}}_m e^{\mathbf{w}^{*'}_m\hat{\boldsymbol{\theta}}^*}\right)$, expression [14] becomes:

$$\hat{r}^*_m \hat{a}_m + \frac{1}{\sigma^2_a}\left(\mathbf{A^{-1}}\hat{\mathbf{a}}\right)_m = \hat{r}^*_m\left(\frac{\delta_m}{\hat{r}^*_m} - e^{\hat{a}_m} + \hat{a}_m\right) \qquad [15]$$

The resemblance with equation [13] suggests the use of:

$$y^*_m = \frac{\delta_m}{\hat{r}^*_m} - e^{\hat{a}_m} + \hat{a}_m \qquad [16]$$

as pre-adjusted record, with weight $\omega_m = \hat{r}^*_m$.

6. Numerical illustration

To illustrate the proposed procedures, a moderate size data set of 10000 records was simulated under the following model:

$$h(t, m) = h_0(t) \exp\{f_i + b_j + a_m\} \qquad [17]$$

where $h_0(.)$ is a Weibull hazard function with parameters ρ=1.8 and λ such that the median lifetime is 800 days; f_i is a fixed effect with 15 levels and corresponding relative risks varying between 1 and 3; b_j mimics a herd-year-season effect (100 levels), generated as a fixed effect also with relative risks in the range of 1 to 3, but later treated as a random, loggamma distributed effect; a_m is an additive genetic value with variance σ^2_a=0.16 (corresponding to a sire variance of 0.04). These a_m values were generated assuming that 5000 animals were daughters of 5 unrelated sires (with 1000 daughters each) and the rest were daughters of 50 young sires (with 100 daughters each), sons of 5 unrelated sires of sires. 27% of the records were censored, either at 1200 days (46% of daughters of young sires) or at 3000 days (9% of daughters of "old" sires).

Several analyses were performed using a extended version of the Survival Kit V3.1. First, the Weibull animal model [17] used to generate the data was applied to get "exact" animal EBVs. Second, a procedure similar to the current national genetic evaluation was applied: herd-year-season effects were integrated out and male EBVs were computed using a sire model. Then, the approximations described first in section 4 (to get herd-year-season solutions) and then in section 3 were implemented to get "approximate Weibull animal solutions". Finally, pre-adjusted records and weights were calculated as indicated in section 5 and were analysed based on a univariate BLUP model. Table 1 presents for *one* particular simulation the correlations between exact Weibull animal solutions and the various approximations, separately for males and females. The lowest correlation (0.984) is observed between sire EBVs from the Weibull sire and animal models. This results from the fact that a fraction $\sigma^2 = 3/4\,\sigma^2_a$ of the additive genetic variance is ignored in the sire model. This part is implicitly included in the residual part of the sire model. But the residuals of *both* Weibull (sire and animal) models are assumed to follow the *same* extreme value distribution. This formal inconsistency was several times indicated in the literature (Ducrocq and Casella, 1996; Korsgaard et al., 2000). However, this example, confirmed in several other situations, both with simulated and field data, shows that the Weibull sire model still provides nearly optimal male EBVs. This is easily explained by the fact that the σ^2 component which is ignored is only a small fraction (7% here) of the residual variance ($\pi^2/6$) .

The correlations between animal models are all very high, both for males and females. The slightly lower correlation observed for the approximate Weibull animal model (0.993) is the direct consequence of the use of sire EBVs coming from the Weibull sire model.

Table 1. Raw correlations between EBVs calculated under the "exact" Weibull animal model and under three approximate procedures

	Male EBVs	Female EBVs
Weibull sire model	0.984	
Approximate Weibull animal model [(a)]	0.9996	0.993
BLUP animal model on pre-adjusted records [(b)]	0.9994	0.998

(a) Two-step procedure described in sections 3 and 4
(b) approximate BLUP procedure described in section 5.

An obvious drawback of the simulated data set we used is that no relationship was assumed between females other than through their sires. The differences between animal and sire models are not fully displayed. Another test was performed based on the same data and the same model used for the routine genetic evaluation of the Normande breed in October 2000. The number of records (1203345) was incompatible with the computation of true Weibull animal model EBVs. Therefore, solutions of the approximate BLUP animal model based on pre-adjusted records were compared with solutions from the current Weibull sire-maternal grand-sire model for bulls and from the approximate Weibull animal model (as in sections 3 and 4) for cows. The correlations between these solutions were 0.982 for bulls and 0.995 for cows. If relationships between females were ignored in the approximate BLUP animal model evaluation,

which was equivalent to restrict in this BLUP evaluation the pedigree information to be exactly the same as in the Weibull sire model, the correlations between EBVs became 0.994 for males and 0.971 for females. This illustrates two important results: first, the Weibull sire model, despite its formal inconsistencies, leads to bull EBVs that were even a better approximations than for the simulated data set(s) analysis indicated: Second, accounting for relationships between females did contribute to some slight reranking between bulls.

7. Conclusion

The limited numerical examples presented here demonstrate the feasibility of the two-step procedure proposed to obtain female EBVs for length of productive life. The procedure is now incorporated in the Survival Kit software with minimal change for the user (only one extra keyword involved) and minimal computing costs (CPU time equivalent to two iterations of the maximisation process). It also reveals that in practical situations, the main drawback of the sire-maternal grand sire model is not the fact that part of the additive genetic value is pushed into the residual while maintaining the variance of this residual at a constant level but the fact that relationships between females are ignored.

The usefulness of a cow EBV for longevity in breeding programs has not been tackled here. We proposed here a possible strategy to enrich direct information with indirect early predictors of functional longevity such as type traits. The use of the equivalent of “pre-adjusted records” with appropriate weights in a univariate BLUP animal model gave promising results for the combined analysis of direct and indirect information. Obviously, a critical conjecture is that such an approach would also lead to reasonably good results when used in a multivariate BLUP setting, together with pre-adjusted records from other traits.

References

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

Cox, D.R. & Snell, E.J. 1966. A generalized definition of residuals. *J. Royal Stat. Soc., Series B. 30,* 248-275.

Ducrocq, V. 1999a. Two years of experience with the French genetic evaluation of dairy bulls on production-adjusted longevity of their daughters. ***In:*** 4th GIFT Workshop: Longevity, Jouy-en-Josas, May 9-11 1999, *Interbull Bulletin 20,* 60-67. Uppsala, Sweden**.**

Ducrocq, V. 1999b. Topics that may deserve further attention in survival analysis applied to dairy cattle breeding : some suggestions. ***In:*** 4th GIFT Workshop: Longevity, Jouy-en-Josas, May 9-11 1999, *Interbull Bulletin 20,* 181-189. Uppsala, Sweden.

Ducrocq V., Boichard D., Barbat A. & Larroque H. 2001. Implementation of an approximate multitrait BLUP evaluation to combine production traits and functional traits into a total merit index. *52nd EAAP Annual Meeting.* Budapest, Hungary, 7 (Abstract).

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

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

Ducrocq, V. & Sölkner, J. 1998a. Implementation of a routine breeding value evaluation for longevity of dairy cows using survival analysis techniques. ***In:*** 6th World Cong. Genet. Appl. *Livest. Prod. 23,* 359-362, Armidale, Australia.

Ducrocq, V. & Sölkner, J. 1998b. "The survival kit - V3.0" A package for large analyses of survival data. ***In:*** *6th World Cong. Genet. Appl. Livest. Prod. 27,* 447-448, Armidale, Australia.

Ducrocq, V., Besbes, B. & Protais, M. 2000. Genetic improvement of laying hens using survival analysis. *Genet. Sel. Evol. 32,* 23-40.

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

Korsgaard, I.R., Andersen, A.H. & Jensen, J. 2000. On different models on heritability, reliability and related quantities of survival traits. ***In:*** *51st EAAP Annual Meeting*, The Hague, The Netherlands, 6:80 (Abstract).