A comparison of two survival analysis methods with the number of lactations as a discrete time variate

M.Hossein Yazdi[1], Robin Thompson[2], Vincent Ducrocq[3] and Peter Visscher[1]

[1] University of Edinburgh, IERM, West Mains Road, Edinburgh EH9 3JG, UK,
[2] Station de Génétique Quantitative et Appliquée – Institut National de la Recherche
Agronomique, Jouy-en-Josas, France,
[3] Rothamsted Experimental Station, Institute of Arable Crops Research, Harpenden, Hertfordshire AL5 2JQ, UK, and Roslin Institute
(Edinburgh), Roslin, Midlothian EH25 9PS

Abstract

The objective of this study was to compare two survival analysis methods with the number of lactations in dairy cattle as the time variate. A data set of 85367 records of UK pedigree Holstein-Friesian cows calving from 1984 through 1993 was used. The last completed lactation was taken as the total number of lactations for each cow. The records of those cows which were still in the herd in 1993, as well as cows with five completed lactations were treated as censored (31.7%).

Analyses were performed using a parametric Weibull regression model and a model which accounted for the discrete nature of the time variable. Sire effect was the source of genetic variation and treated as a random effect. Correlation between solutions for fixed effect of herd was close to unity (0.99). Estimates of linear regression for age at first calving from the two models were very similar in all analyses. Estimates of linear regression for milk yield from the Weibull model were higher than estimates from the discrete model and these effects were higher when milk at different lactations was treated as a time-dependent covariate.

Estimates of sire variance from the Weibull model were higher than estimates from the discrete model. Correlations between estimated breeding values from the two models were high (0.98 and 0.99). Based on our data set, the estimates of fixed effects and sires' breeding values on risk of culling did not differ and it appears that the Weibull model may be robust enough to be used even on discrete data.

1. Introduction

It is well documented that length of animal's productive life affects her profitability as well as the dairy farm production efficiency through increasing proportion of higher yielding cows, decreasing replacement cost, and more opportunity for voluntary culling (e.g. Rendel and Robertson, 1950; Brotherstone et al., 1998).

Depending on the purpose of investigation, several definitions of length of productive life in dairy cattle have been used. To investigate the association between lifespan and the type traits, Brotherstone et al. (1998) used the number of completed lactations as the length of productive life for those cows that were culled and number of expected lactations to be completed for those still in the herd. VanRaden and Klaaskate (1993) investigated total months in milk by 84 month of age as a measure of length of useful life. Ducrocq (1987, 1994) defined the length of productive life as the number of days from first calving to death or culling which is a continuous variable in the survival analysis.

Among the several measurements of length of productive life (e.g. age of cow at disposal, number of months at lactations) in dairy cattle, number of completed lactations is clearly defined and easily recorded. Most statistical methods of dealing with lifetime data are based on the assumption that observed failures are described on a continuous scale. In some cases (e.g., the Cox model), a further requirement is that the observed failure times must be distinct. When the observations are expressed on a discrete scale, for example, when longevity of animals is measured in the number of years, parities, lactations, it is theoretically more proper to model the longevity accordingly. This can be done for example by assuming an underlying continuous time scale on which intervals are defined and only the interval in which the individual died is known: this is the so-called grouped data model (Prentice and Gloeckler, 1978; Ducrocq, 1999a).

The objective of this study was to compare the two survival analysis approaches; the one based on the continuity of the failure time variable assuming a Weibull baseline distribution and the other one for grouped data without any assumption for the baseline distribution (the discrete model). The survival time was the number of lactations and the same data set was used in both analyses.

Table 1. **Number of censored and uncensored records per lactation**

Lactation	Censored	Uncensored	Total	% of total records
1	7381	22539	29920	35.05
2	5219	17112	22331	26.16
3	2715	11070	13785	16.15
4	11783	7548	19331	22.64
Total	27098	58269	85367	
% of total records	31.74	68.26		100

2. Material and Methods

2.1 Data

The time of first calving was the origin of the survival time variable in this study. The records of UK pedigree Holstein-Friesian cows having first calved between 1984 and 1993 and their subsequent lactations records were extracted from the original files (Brotherstone and Hill, 1991). The data files restricted to be records of herds with at least 10 observations and sires with at least 100 daughters. Records of those cows with missing subsequent records were deleted. After editing, 85367 records were left. The number of different herds was 1254 and the number of cows in herd ranged from 10 to 256. The number of different sires was 284 and the number of daughters per sire ranged from 100 to 2157. The number of cows with censored and uncensored records is presented in Table 1.

2.2 Statistical Methods

The statistical model was based on the concept of a hazard function, the animals' limiting risk of being culling at time t, conditional upon survival to time t (Ducrocq, 1987; Ducrocq and Casella, 1996). The number of completed lactations of cow, measured as length of productive life, was considered as the dependent survival time (longevity). Two statistical models were used. First, a Weibull mixed model, treating longevity of cows as a continuous time variable and assuming that the baseline, representing the ageing process, has a Weibull distribution. Second, the longevity of a cow expressed as the number of completed lactations was treated as a discrete variable (using the grouped data approach) with no assumption for the baseline distribution (Ducrocq, 1999a, b). In both cases the hazard function of a cow was modelled as follows:

$$h(t, x(t)) = h_0 (t) \exp\{he_i + y_j + b_1(age) + b_2(milk(t)) + s_k\}$$

where h(t, x(t)) is the hazard function of an animal depending on time t. In the Weibull model, $h_0 (t) = \lambda\rho(\lambda t)^{\rho-1}$ is the baseline hazard function (related to the ageing process) where λ and ρ are the location and shape parameters of the Weibull distribution. In the discrete model h_0 (t) is a piecewise constant function which describes the discrete hazard for each particular interval. he_i is the i^th herd effect, y_j is the j^th year effect, b_1 (age) is the partial regression coefficient of age effect, b_2 (milk(t)) is the partial regression coefficient on milk production where milk yield is a number of standard deviation from herd mean, s_k was the random additive genetic effect of the k^th sire. The sire additive genetic effect was assumed to have a normal distribution, $s_q \sim N(0, \sigma_s^2)$, where subscript q refers to sire q, and σ_s^2 is the sire variance.

Several analyses, treating milk yield as a time-independent or time-dependent covariate, were carried out. The herd and year effects were class effects and treated as time-independent covariate in all analyses.

The heritability of longevity in the Weibull model was calculated from the sire variance component as a proportion of phenotypic variance on the logarithmic scale as described by Ducrocq and Casella (1996) : $h^2_{\log} = 4\sigma_s^2/(\pi^2/6+\sigma_s^2)$, where $\pi^2/6$ is the variance of the standard extreme value distribution (Lawless, 1982).

3. Results and Discussion

Results of the different analyses are given in Tables 2 and 3.

Table 2. Results of comparison when milk yield treated as a time-independent in the model

Covariate	Weibull model	Discrete model	Correlation
Herd			0.99
Year			0.63
Age	0.026	0.025	
Milk	-0.185	-0.180	
Sire			0.99
σ^2_s	0.019	0.018	
h^2	0.05		
Rho (ρ)	2.09		
Lambda (λ)	0.541		

Table 3. Results of comparison when milk yield at different lactations treated as a time-dependent covariate in the model

Covariate	Weibull model	Discrete model	Correlation
Herd			0.99
Year			0.67
Age	0.023	0.022	
Milk	-0.214	-0.218	
Sire			0.98
σ^2_s	0.017	0.008	
h^2	0.04		
Rho (ρ)	2.07		
Lambda (λ)	0.524		

Comparisons of fixed effects included in the models were based on correlations between estimates from two models. The correlations between estimates of relative culling risk in different herds from the two models were close to unity, indicating the choice of model does not have any influence on the herd estimate. The positive regression coefficient for age indicates that with increasing age at first calving the animal’s risk of being culled is increased. The estimates of regression coefficient were similar in all analyses (on average 0.024 per month). The regression coefficients of milk yield on animal’s risk of being culled were negative in all analyses and it was higher when the milk production was treated as a time-dependent covariate. Animals with higher milk production had lower risk of culling and vice versa. Since there was no assumption for the baseline distribution in the discrete model, it was not possible to compare heritabilities for the two models. Only the sire variances from two models were compared and the correlation between estimated breeding values for sires were calculated (Table 2 and 3). The correlations between solutions for sires were high (0.98 and 0.99). These high correlations indicate that ignoring the discrete nature of the data as in the (continuous) Weibull mixed models does not have an impact on the ranking of sires, at least when these sires have a lot of daughters (>100 as here). The estimates of sire variances ranged from 0.008 to 0.019 in the different analyses. The estimates from the Weibull models were similar and in general they were higher than estimates from the discrete model. The baseline survival rate in the discrete model from lactation 1 to • 4 ranged from 0.65 to 0.09 and from 0.67 to 0.20 for model when milk yield was time-independent and time-dependent, respectively.

The longevity measure time in this survival analysis was the number of completed lactations during the animal's productive life. It is theoretically better to use the grouped data model because it does not violate the fact that the time scale is clearly discrete. On our data set, however, the different approaches did not influence the estimates of fixed effect or sires' breeding values substantially, and it appears that the Weibull model may be robust enough to be used even on discrete data. However, further work is necessary using different data, and in particular considering situations where sires have less daughters to verify these claims.

Acknowledgements

We thank the BBSRC and EU (GIFT programme) for support.

References

Brotherstone, S., and Hill, W. G. 1991. Dairy herd life in relation to linear type traits and production. 1. Phenotypic and genetic analyses in pedigree type classified herds. Anim. Prod. 53:279-287.

Brotherstone, S., Veerkamp, R. F. and Hill, W. G. 1998. Predicting breeding values for herd life of Holstein-Friesian dairy cattle from lifespan and type. Anim. Sci. 67:405-411.

Ducrocq, V. P. 1987. An analysis of length of productive life in dairy cattle. Ph.D. dissertation, Cornell University, Ithaca, NY, USA.

Ducrocq, V. P. 1994. Statistical analysis of length of productive life for dairy cows of the Normande breed. J. Dairy Sci. 77:855-866.

Ducrocq, V. P. 1999a. *Survival analysis applied to animal breeding and Epidemiology*. (Lecture notes) University of New England, Armidale, NSW, Australia

Ducrocq, V. P. 1999b. Extension of survival analysis models to discrete measures of longevity (these proceedings).

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

Lawless, J. 1982. *Statistical models and methods for lifetime data*. John Wiley & Sons. New York, NY.

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

Rendel, J. M. and Robertson, A. 1950. Some aspects of longevity in dairy cattle. Empire Journal of Experimental Agriculture, 18:49-56.

Van Raden, P. M. and Klaaskate, E. J. H. 1993. Genetic Evaluation of length of productive life including predicted longevity of live cows. J. Dairy Sci. 76:2758-2764.