Animal Science 2000, **70**: 197-206

1357-7298/00/96180197$20·00

A comparison of a linear and proportional hazards approach to analyse discrete longevity data in dairy cows

R. Lubbers[1,2], S. Brotherstone[3], V. P. Ducrocq[4] and P. M. Visscher[1]†

[1]*University of Edinburgh, Institute of Ecology and Resource Management, West Mains Road, Edinburgh EH9 3JG*
[2]*Wageningen Agricultural University, PO Box 338, 6700 AH Wageningen, The Netherlands*
[3]*University of Edinburgh, Institute of Cell, Animal and Population Biology, West Mains Road, Edinburgh EH9 3JT*
[4]*Station de Génétique Quantitative et Appliquée, Institut National de la Recherche Agronomique, 78352 Jouy-en-Josas, France*

† Corresponding author.

Abstract

The objective of this study was to compare two methods for analysis of longevity in dairy cattle. The first method, currently used for routine genetic evaluation in the UK, uses a linear model to analyse lifespan, i.e. the number of lactations a cow has survived or is expected to survive. The second method was based on the concept of proportional hazard, i.e. modelling the conditional survival probability of a cow as a function of time. Comparisons were based on estimated heritabilities, ranking of estimated breeding values of sires, estimated effects of covariates used in the final models, and the distribution of residuals. The same data set, 21497 observations on the number of lactations cows had survived, was used for both analyses, even in the presence of censored observations. Cows in the data were progeny of 487 sires. Heritability estimates for lifespan or survival were approximately 0·06 for both methods, using the definition of heritability on a logarithmic scale for the proportional hazards model. Correlations between breeding values for sires were high, with absolute values ranging from 0·93 to 0·98, depending on the model fitted. It was concluded that it may be justified to use the standard Weibull model even for discrete time measures such as the number of completed lactations, but that more research is needed in the area of discrete time variates.

Keywords: *dairy cattle, heritability, lifespan, longevity, survival.*

Introduction

Length of productive life, or herd life, is an important trait affecting dairy farm profitability (Rendel and Robertson, 1950). Decreasing culling due to involuntary causes (e.g. related to disease, infertility, lameness, etc.) by genetic or non-genetic means has a positive effect on economic performance, mainly through decreased replacement costs and increased opportunity for voluntary culling (e.g. van Arendonk, 1985).

Many definitions of survival have been used, including survival to a particular age, number of completed lactations and lifetime production (e.g. Dekkers and Jairath, 1994). Until recently, data available in the United Kingdom (UK) have comprised 305-day lactation yields with a minimum qualifying lactation length of 200 days, with only data from five lactations reliably stored. Although monthly test data are now available, they do not yet cover a sufficient time period to be useful for survival analysis. In the past, analyses of the UK population have therefore been of survival through two, three, four and five completed lactations, expressed as binomial traits (Brotherstone and Hill, 1991). One drawback with an endpoint measure of survival is that information before the specified endpoint is lost. Another drawback with those endpoints is that all animals must have had the opportunity to reach that stage, by which time the information is often less useful for selection procedures (Brotherstone *et al.*, 1997). To overcome these problems Brotherstone *et al.* (1997) defined another measure of longevity, lifespan, i.e. the number of lactations an animal completes or is expected to complete prior to culling. For animals deemed still to be in the herd, having completed lactation n but not had time to complete lactation

$n + 1$, the lifespan reflects the number of lactations completed and the number of further lactations the animal is expected to complete. These sources of information can be combined for animals whose survival is unknown, and using population-based average survival probabilities from lactation to lactation a lifespan can be calculated which predicts the expected longevity of the animal. In this way, all animals in the data set are allocated one of a predefined set of lifespans depending on whether the animal is deemed to have survived or not (Brotherstone *et al.*, 1997).

Instead of considering traits such as 'survival until a certain age' or 'lifespan' defined by Brotherstone *et al.* (1997), alternatives to the well known linear model may be preferable. Cox (1972) and Kalbfleisch and Prentice (1980) described the method of survival analysis in which the risk of failure instead of the actual longevity of an animal is modelled. It relies on the concept of hazard rate, the limiting probability of being culled among animals still alive. The hazard rate can be modelled for all records, whether censored or not. Another advantage of the method is the possibility of modelling effects in a time-dependent way, thus it is expected that such models mimic reality better. Famula (1981) introduced the method in animal breeding, and Smith and Quaas (1984) were the first to estimate genetic parameters with survival analysis. Following the studies of Ducrocq (1987) and Ducrocq *et al.* (1988), Ducrocq and Sölkner (1994) introduced a program for survival analyses which is more generally applicable, the Survival Kit. The Survival Kit has been updated continuously since then and used by various researchers (e.g. Gröhn *et al.*, 1997; Ringmar-Cederberg *et al.*, 1997; Vukasinovic *et al.*, 1997; Vollema and Groen, 1998).

The objective of this study was to compare the method based on the concept of lifespan (Brotherstone *et al.*, 1997), with the method based on the concept of proportional hazard (Cox, 1972; Kalbfleisch and Prentice, 1980; Ducrocq, 1987; Ducrocq *et al.*, 1988; Ducrocq and Sölkner, 1994) using the same data set. Comparisons were based on estimated heritabilities, ranking of estimated breeding values of sires, estimated effects of covariates used in the final models, and the distribution of residuals. As far as we know, this is the first example of the comparison of a linear and non-linear model for survival using both (i) a discrete time variable (failure times of 1 to 5) and (ii) using the same data for both analyses, even in the presence of censored observations.

Material and methods

Data

A data set of 53450 first lactation records of pedigree Holstein-Friesian (HF) cows calving between November, 1986 and October, 1987, inclusive, was extracted from the National Milk Records (NMR) files. The reason for choosing this data set was that similar data had been used before by Brotherstone *et al.* (1997), and the use of a single cohort of cows would simplify model fitting. All cows were registered in the UK, and had the opportunity to complete five lactations (the maximum number used in genetic evaluations for yield in the UK). The NMR files were searched for each cow's lactation two to lactation five records and the cow was deemed to have completed lactation n if her record of lactation n and all previous lactations were found in the files. Any cow that changed herd or did not complete a qualifying, minimum 200-day, lactation was judged not to have survived. Lower and upper bounds for age at calving of 20 and 40, 30 and 60, 40 and 75, 50 and 85, and 60 and 100 months, were set for lactations 1 to 5, respectively. Herds and sires were required to have at least 10 observations in the data. After editing, 21497 records were left. The number of different herds at first calving equalled 1265. The number of different sires was equal to 487, with a minimum of 10 daughters and a maximum of 863 daughters per sire. For the cohort of cows in this study, the average herd life was at least 3·03 lactations, which is the value calculated assuming that cows were culled after lactation 5 (Table 1). Records of cows that completed a fifth lactation were censored, i.e. it was not known whether these cows were culled or remained in the herd for further lactations.

Concept of lifespan

If p_n is the probability of survival to lactation $n + 1$ of an animal that has survived to complete lactation n, the expected lifespan of a random animal that has completed n lactations but has not had time to complete $n + 1$ lactations is (Brotherstone *et al.*, 1997):

Table 1 *Survival rate and survival probabilities*

Lactation (t)	No. of records	$S(t)$†	$p(t,t+1)$‡
1	21497	1·00	0·77
2	16520	0·77	0·73
3	12091	0·56	0·73
4	8812	0·41	0·71
5	6252	0·29	
		$\Sigma S(t) = 3{\cdot}03$§	

† Stayability: survival until time = t.
‡ Conditional probability of survival from lactation t to lactation $t + 1$.
§ Average herd life until lactation 5.

$$n + p_n + p_n \times p_{n+1} + p_n \times p_{n+1} \times p_{n+2} + \ldots.$$

The survival probabilities p for lactation 1 to 5 are given in Table 1. It was assumed that p remains constant after lactation 5, and that the maximum number of lactations that could have been completed equalled N. Thus, $p\ (5, 6) = p\ (6, 7) = \ldots = p\ (N\text{-}1, N) = p\ (4, 5) = 0{\cdot}71$. A predefined set of lifespans was set up. Cows that completed n lactations but not $(n + 1)$ were given lifespan n. Cows that completed lactation n but had no time to complete $(n + 1)$ were given lifespans of 3·7, 4·6, 5·5, 6·4, or 7·4, for n being 1, 2, 3, 4, or 5, respectively. For example, if an animal had only time to finish three lactations and was censored at $t = 3$ (i.e. had no time for lactation 4), then the assigned lifespan (LS) was:

$$\text{LS} = 3 + 0{\cdot}73 + 0{\cdot}73 \times 0{\cdot}71 + 0{\cdot}73 \times 0{\cdot}71 \times 0{\cdot}71 + 0{\cdot}73 \times 0{\cdot}71 \times 0{\cdot}71 \times 0{\cdot}71 + \ldots.$$

$$= 3 + 0{\cdot}73 \times (1/(1\text{–}0{\cdot}71))$$

$$= 5{\cdot}5.$$

Hence, prediction of expected lifespan is based upon population-wide average conditional survival probabilities. Note that there can be cows with censored records which have larger (predicted) values for lifespan than the largest value for an uncensored observation (i.e. LS = 4). This is based upon the fact that we know that on average cows which have survived until lactation 3 and are still in the herd will have a lifespan greater than four lactations.

Concept of proportional hazard

The concept of hazard function, defined as the limiting rate of culling or death at time t, conditional upon survival to time t, allows a natural modelling of the relationship between longevity and specific covariates. In proportional hazards, the hazard of an animal (i.e. its risk of being culled) at time t is described as the product of a baseline hazard function $\lambda_0(t)$, which is either left completely arbitrary (Cox model) or has a parametric form (e.g. exponential, Weibull or gamma) and of a positive term which is an exponential function of a vector of covariates $\mathbf{x}'$ multiplied by a vector of regression parameters β (e.g. Kalbfleisch and Prentice, 1980):

$$\lambda\ (t; \mathbf{x}) = \lambda_0\ (t) \exp\ \{\mathbf{x}'\beta\};$$

or, when x is a vector of time dependent covariates:

$$\lambda\ (t\ ; \mathbf{x}(t)) = \lambda_0\ (t) \exp\ \{\mathbf{x}\ (t)'\beta\}.$$

Proportional hazard models can be extended to include random (e.g. genetic) effects, as in the regular mixed linear models that are used for genetic evaluations worldwide (Ducrocq and Casella, 1996). In the current study the baseline hazard function had a parametric form and was assumed to follow a Weibull distribution, so:

$$\lambda_0\ (t) = \lambda\ \rho\ (\lambda\ t)^{\rho-1}$$

where λ and ρ are location and shape parameters of the baseline Weibull hazard function.

Model definition

Basically, the following two types of models were compared:

$$\text{LS} = \mathbf{x}'\beta \textit{ versus } \lambda\ (t; \mathbf{x}) = \lambda_0\ (t) \exp\ \{\mathbf{x}'\beta\}$$

LS indicates the lifespan whereas $\lambda(t; \mathbf{x})$ indicates the hazard, t is the number of lactations completed and β is a vector of explanatory (independent) variables (with corresponding design vector $\mathbf{x}$) including: herd at first calving, month at first calving, age at first calving (in months), squared age at first calving (in months^2), proportion of HF genes, first lactation milk yield, deviated from the herd mean first lactation milk yield (in s.d. from the herd mean) and a sire effect.

Herd was fitted as a random effect with an assumed log-gamma distribution (e.g. Ducrocq and Casella, 1996) in the proportional hazards models, which implies that a single parameter (γ) needs to be estimated. In a preliminary analysis for which there was no restriction on the number of records per sire and herd, herds were fitted as fixed effects. However, the survival analysis had problems with convergence, because some herds had only one or two uncensored observations. Therefore, it was decided to set a lower limit for the number of records per herd, and to treat herds as random. For the analyses on lifespan, herd was fitted as a normally distributed random effect.

Month of first calving (MC) was fitted as a fixed classified effect, and age at first calving as a linear and quadratic covariate. As an association between breed effect and survival has been demonstrated (Brotherstone and Hill, 1994), the proportion of HF genes was included in the model as a linear covariate. To adjust for the effect of first lactation milk yield on the decision to retain or cull an animal, the analysis included first lactation milk yield in the model as a covariate. This variable (ΔM) was expressed in standard deviation units from the herd mean. Survival models that include production

Table 2 *Overview of the different models used*

Model	Lifespan/Hazard	Time dependency of covariates
1	LS	No
2	ln(LS)	No
3	Hazard	No
4	Hazard	Yes

variables as covariates are considered to describe functional herdlife (Ducrocq, 1987). Genetic parameters were estimated using sire as a random class effect normally distributed, ignoring relationships between sires.

Four models were studied (Table 2) (1) analysis of lifespan (LS), (2) analysis of ln(LS), (3) proportional hazard model without time dependent variates, and (4) proportional hazard model with time dependent variates. The latter was extended from model 3 to investigate the value of time dependent traits in survival analysis. Only the variables MC and ΔM could be made time dependent. By treating MC and ΔM as time dependent, the month of calving and milk yield deviation for each lactation were fitted in the model, instead of the month of first calving and the milk yield deviation in the first lactation. Due to the fact that only heifers calving between November 1986 and October 1987 are under study, it was not possible to introduce an adequate time dependent herd-year(season) variable. Ducrocq (1994) introduced a random time-dependent herd-year-season effect to improve the modelling of the environmental part. Neither could the effect of (variation in) herd size be included due to limited information. Ducrocq (1994) pointed out the influence of variation in herd size on culling in the Normande breed. Due to the fact that only information about the number of completed lactations was available in this study, no stage-of-lactation effect could be included in the model, as suggested by Ducrocq (1987) and Ducrocq *et al.* (1988).

For our data, the observations on all cows which had completed five lactations were censored, because we did not know whether these cows were culled after their fifth lactation or completed more than five lactations. Censoring was taken into account in the lifespan analysis by adding the number of additional lactations a cow which has survived five lactations is expected to complete (Brotherstone *et al.*, 1997) and in the survival analysis by adequately treating the different contribution of censored and uncensored records in the estimation of parameters (e.g. Ducrocq, 1987).

VCE (Groeneveld, 1995) was used to obtain REML estimates of the heritability of lifespan, and PEST (Groeneveld, 1990) was used to predict breeding values. Survival analysis relying on the concept of proportional hazard rate was performed using the Survival Kit (Ducrocq and Sölkner, 1994, 1998a and b).

Estimates of covariates. The estimates β_i can be expressed in relative risk ratios (RR) by the simple transformation $RR(\beta_i) = e^{\beta i}$. This expression gives the relative risk of culling due to that effect, and follows from assuming a proportional hazards model:

$$\lambda(t;\mathbf{x}_a)/\lambda(t;\mathbf{x}_b) = exp((\mathbf{x}_a' - \mathbf{x}_b')\beta),$$

i.e. the relative hazard for two individuals with covariates described by $\mathbf{x}_a$ and $\mathbf{x}_b$, respectively, is independent of time and of other covariates. For a continuous covariate with $\mathbf{x}_a = 1$ and $\mathbf{x}_b = 0$, the relative hazard is e^{β}. For example, consider two cows which differ only in the proportion of HF, and assume that the estimate of the regression parameter on proportion HF from the proportional hazards model is –0·10. Then, the RR of a 100% HF cow relative to a 0% HF cow is, $\exp(c–0{\cdot}10 \times 1)/\exp(c–0{\cdot}10 \times 0) = \exp(–0{\cdot}10) = 0{\cdot}90$, where c refers to the other effects in the model. Hence, the 100% HF cow has a 10% reduction of the risk of being culled.

Genetic parameters. For proportional hazard models which assume a Weibull survival function (or the exponential distribution as its special case) the heritability estimate on the log-scale is,

$$h^2_{log} \sim 4\,\sigma_s^2/(\pi^2/6 + \sigma_h^2 + \sigma_s^2),$$

with σ_h^2 the herd variance, and σ_s^2 the sire variance (Ducrocq and Casella, 1996). Here, the herd variance is equal to $\Psi^{(1)}(\gamma)$, where $\Psi^{(1)}(.)$ is the trigamma function.

An approximation of the heritability on the original scale using the sire variance from the Weibull model and based on a Taylor series expansion around the mean of ln(t) is,

$$h^2_{orig} \sim 4\,\sigma_s^2/[\exp\{k/\rho\}^2\,(\pi^2/6 + \sigma_h^2 + \sigma_s^2)],$$

with $k = \Psi(\gamma) - \ln(\gamma) - \nu$ where $\Psi(.)$ is the digamma function and ν is Euler's constant (~0·5772). (Vollema and Groen, 1998; Ducrocq, 1999; Yazdi *et al.*, 2000). There is some controversy about the utility of the heritability transformed to the original scale (Korsgaard *et al.*, 1999). However, empirical results (Ducrocq, 1999) and extensive simulations (unpublished results) indicate that, at least in the

context envisioned here, a clear relationship between the heritability on the original scale and the repeatability of sire proofs.

Relationships between the two methods

To be able to compare the estimates from the Weibull model with those from the lifespan models, it is necessary to explore the relationships between the methods. In many dairy cattle populations, conditional survival probabilities for lactations are roughly similar over time, with typical values of 0·7 to 0·8. If we assume that these conditional probabilities are constant with value of p, and cows have had no time restriction in the opportunity to express LS, then,

$$\text{Prob}(LS = x) = (1-p)p^{x-1}[x = 1, 2, 3, \ldots]$$

or,

$$\text{Prob}(LS = x) = [(1-p)/p]p^{x}$$

i.e. LS has a geometric distribution (Brotherstone *et al.*, 1997), with mean and variance

$$E(LS) = 1/(1-p) = 1 + p/(1-p)$$

$$\text{var}(LS) = p/(1-p)^2$$

The probability that LS is equal or larger than value T is,

$$\text{Prob}(LS \geq T) = 1 - \text{Prob}(LS < T) = 1 - (1 - p^{T-1}) = p^{T-1}.$$

This corresponds to the definition of 'stayability until lactation T' which has been used in the literature (e.g. Visscher *et al.*, 1994). The survival function is simply,

$$S(t) = p^t, \text{ and } \ln(S(t)) = t \ln(p).$$

The probability density function (pdf) of t from the survival function, assuming a continuous t, is,

$$f(t) = -dS(t)/dt = -\ln(p)\, p^t = \lambda e^{-\lambda t},$$

i.e. the exponential pdf, with $\lambda = -\ln(p)$. The actual probability function of discrete t was given above, $\text{Prob}(t) = [(1-p)/p]\, p^t$. For p close to unity, $(1-p)/p \sim -\ln(p)$. (For $p = 0{\cdot}75$ the error in this approximation is about 10%. For $p = 0{\cdot}70$, the error is 15%). This suggests that even with simple discrete dairy cattle data an exponential model could be appropriate, since failure time at lactation t, under the assumption of constant conditional survival probabilities, approximately follows an exponential distribution. Including covariates in the model, and assuming an exponential baseline hazard function, leads to an exponential regression model.

Note however that in most studies, including the present one, the conditional survival probabilities are slowly decreasing over time. With a continuous time scale, a typical model to accommodate such a decrease is the Weibull model ($S(t) = \exp(-(\lambda t)^{\rho})$) with ρ greater than 1. Indeed, the discrete-time interpretation of the increasing failure rate $\lambda(t)$ of the continuous Weibull model would be that the conditional survival probabilities $p_n = S(n+1)/S(n)$ are decreasing.

The Weibull model can also be interpreted as a particular case of an accelerated failure time model (Kalbfleisch and Prentice, 1980): if we assume that the decreasing conditional survival probabilities are merely the consequence of the fact that on average, cows 'wear out' at a faster rate as time goes, a change in time scale $t \rightarrow t^* = t^{\rho}$ would lead to the exponential model $S(t^*) = \exp(-\lambda^* t)$.

For the Weibull regression model (i.e. including covariates in the model),

$$S(t,\mathbf{x}) = \exp(-t^{\rho})^{\exp(\mathbf{x}'\beta)} = \exp((-t^*)^{\rho})$$

with $t^* = t\, \exp(\mathbf{x}'\beta/\rho)$ and $\mathbf{x}'\beta$ containing a term for the intercept ($\rho\ln(\lambda)$).

If we assume no censoring, i.e., $\ln(t) = \ln(LS)$, the expectation of the logarithm of t (or ln(lifespan)) is,

$$E(\ln(t)) = E(\ln(LS)) = E(\ln[t^*/\exp(\mathbf{x}'\beta/\rho)]) = E(\ln(t^*)) - \mathbf{x}'\beta/\rho.$$

It follows that,

$$\ln(t) = \ln(LS) = E(\ln(t^*)) - \mathbf{x}'\beta/\rho + \omega/\rho$$

where ω has an extreme value distribution (e.g. Kalbfleisch and Prentice, 1980). This implies that if all assumptions hold, the log of lifespan is linear in the covariates and that the effects of the covariates from a Weibull model are a simple linear transformation of the effect of a ln(LS) model. That is, if we ignore the difference in the distributions assumed for the residual part of the models (normal for LS or ln(LS), extreme value for the Weibull model) we expect the effects from the log-linear model to be $-\beta/\rho$, with β and ρ being the regression and shape parameters from the Weibull regression model.

The above transformations were investigated for our data set by fitting a linear model to ln(LS) (model

(2)), and by transforming the estimates from the Weibull models (3 and 4) to those on the log-scale.

Transformation of covariates. From the analytic results above, a transformation of the covariates from the Weibull analysis (model 3) to the log(LS) analysis is,

$$\beta_{log(LS)} \sim -\beta_{Weibull}/\rho$$

A simple transformation of the effects from the proportional hazards model (3) ($\beta_{Weibull}$) to the linear model (1) (β_{linear}) was made by considering the expected lifespans for two cows with different covariates:

$$\beta_{linear} \sim E(LS(x_a)) - E(LS(x_b))/(x_a - x_b)$$

with

$$E(LS(x_a)) = \Sigma\,[t\ S_0(t)^{\exp(x_a'\beta_{Weibull})}], \text{ and}$$

$$E(LS(x_b)) = \Sigma\ t\ S_0(t)^{\exp(x_b'\beta_{Weibull})}]$$

where $S_0(t)$ is the baseline survival function. For a single covariate, using $x_a = 1$ and $x_b = 0$, the transformation becomes,

$$\beta_{linear} \sim \Sigma\,[S_0(t)^{exp(\beta_{Weibull})} - S_0(t)],$$

which was used in this study.

Model validation

The assumptions in the various models were tested by estimating the distribution of the residuals for each model. For the linear models (models 1 and 2), the assumption is that residuals are normally distributed. For the Weibull models (models 3 and 4), or any proportional hazards model, generalized residuals follow a unit censored exponential distribution (exponential distribution with parameter $\lambda = 1$, see Cox and Snell (1968)).

For models 1 and 2 (LS and ln(LS)), the residuals were calculated by PEST, using the estimated variance components in the analysis. For models 3 and 4, residuals were calculated using the Survival Kit. From these 'observed' residuals, a survival curve of the residuals (S(e)) was calculated, and $-\ln(S(e))$ plotted against the residuals. If the assumption of the Weibull model is correct, this curve should have an intercept of zero and a slope of one (e.g. Kalbfleisch and Prentice, 1980).

Table 3 *Heritabilities of the length of productive life on the diagonal (with h^2 on the transformed original scale in brackets), and Pearson correlations (below diagonal) between predicted breeding values for sires from the various models*†

Model‡	1	2	3	4
1	0·06			
2	0·97	0·05		
3	–0·97	–0·94	0·07 (0·17)	
4	–0·96	–0·93	0·98	0·07 (0·17)

† Approximate s.e. of heritability estimates were <0·01.
‡ See Table 2 for model definition.

Results

Table 3 gives an overview of the heritabilities of length of productive life for all four models, and correlations between sire breeding values predicted by the different models. Heritabilities of lifespan, whether on the natural scale or the log scale and for both uncensored and censored data were around 5%, consistent with previous analysis of this trait (Brotherstone *et al.*, 1997). For the proportional hazard models, the heritabilities on the log scale were of similar magnitude (7%). When heritability estimates from the Weibull model were transformed to the original scale, the values were substantially higher than the estimates from the linear model (17% *v.* 5%).

The correlations between sire breeding values from the different Weibull models were very high, 0·98 Correlations between breeding values predicted from the two linear models (models 1 and 2) were also very high, 0·97. Correlations between breeding values from the linear model and from the Weibull model were negative, because a positive breeding value for lifespan indicates a higher longevity, whereas a positive value for a breeding value from the Weibull model indicates a higher risk of culling and hence a lower longevity. The absolute values of the correlations between breeding values from the linear model and the Weibull model were high, varying from 0·93 to 0·97. The lowest value was between the linear model analysis of ln(LS) and the Weibull model with time-dependent variates, with a correlation of –0·93.

All included covariates in the four different models were significant ($P < 0·05$). The values of the covariates are given in Table 4. For model 3, the transformed estimates for HF and ΔM were 0·23 and 0·28, respectively, which compares to values of 0·31 and 0·38 from the LS linear model (model 1). The transformed values of HF and ΔM from the Weibull analysis to the ln(LS) scale were 0·11 and 0·14 for model 3. These values compare with the estimated effects using ln(LS) of 0·10 and 0·12 (model 2), respectively.

HF proportion had a constant effect of a RR of around exp(–0·17) ~ 84% in both proportional hazard

Table 4 *Parameter estimates for milk deviation (ΔM), proportion of Holstein-Friesian (HF), and age at first calving (AC, in months) for all models, and transformed estimates of the parameters from the Weibull model (3)*

	Model†					
Effect	1	3‡	2	3§	3	4
HF	0·31	0·23	0·10	0·11	−0·18	−0·17
ΔM	0·38	0·28	0·12	0·14	−0·22	−0·32
AC	0·06	0·05	0·02	0·03	−0·04	−0·05
AC^2	−0·002	−0·001	−0·0005	−0·0007	0·001	0·001
ρ					1·6	1·6
Intercept					−1·80	−1·85
γ					5·0	5·0

† Models defined in Table 2.
‡ Transformed from Weibull to linear scale.
§ Transformed from Weibull to log-linear scale.

models, with cows with larger proportion HF having a lower risk of culling and hence a longer herd life. The effect of milk yield deviation (untransformed) varied from −0·22 to −0·32, depending on whether this variate was fitted as time independent or time dependent. This smaller value clearly shows that milk deviation in the first lactation (which was fitted when ΔM was treated as time independent) is a predictor that underestimated the risk of culling compared with ΔM in the current lactation.

Table 5 shows that the calving pattern is one of autumn calving, with 3269 (uncensored records only) cows having their first lactation starting in September. The relative risk ratios from the data analysis with time-independent variates only (model 3) show a marginal increased rate of culling for autumn and winter calving, although the pattern is not clear. However, when ΔM and month of calving

Table 5 *Number of uncensored observations (N) and risk ratios (RR) of separate classes of the fixed variable month of calving for models 3 and 4*†

Month of calving	N	RR (model 3)	RR (model 4)
January	865	1·07	1·28
February	471	0·89	1·14
March	252	0·93	1·11
April	136	0·98	1·25
May	141	0·95	1·28
June	346	1·05	1·10
July	1240	0·96	1·01
August	2719	1·01	0·95
September	3269	1·00‡	1·00‡
October	2355	1·06	1·05
November	2018	1·02	1·26
December	1433	1·04	1·35

† Models defined in Table 2.
‡ The estimates are relative to the month of September.

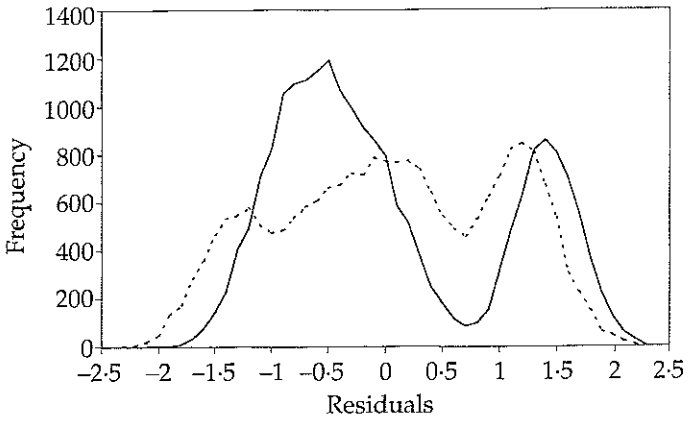

Figure 1 Frequency of residuals for model 1 (——— LS) and model 2 (– – – ln (LS)).

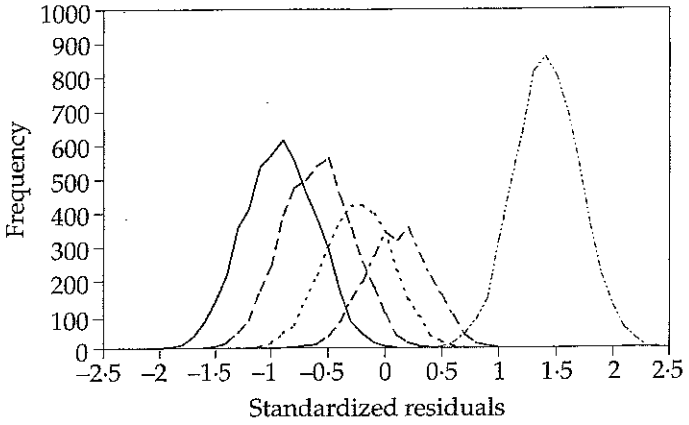

Figure 2 Frequency of residuals per lactation, for LS (model 1). (——— lactation 1; — — lactation 1; ----- lactation 3; ——·—lactation 4; —··— lactation 5).

are fitted as time dependent variables (model 4), the results are clearer: autumn calvers (August to October) are favoured, with cows calving in the winter or spring having up to 35% higher probability of being culled.

Figure 1 shows the frequency distribution of the residuals from the LS and ln(LS) models. The shape of the distribution for both models is similar, with a clear bimodal distribution. The explanation for the shape of the distribution is the discrete nature of the data and the fact that only those cows that lived longest (five completed lactations) had censored observations. This is shown in Figure 2, which gives the residuals for LS, for each of the five failure times. There are five distinct peaks in the distribution, ranging from −1·0 (cows culled after lactation 1) to +1·5 (cows still in the herd after five lactations). Residuals for LS and ln(LS) were clearly not normally distributed (Figure 1), with the residual

distribution for LS being more skewed than the distribution of residuals from the analysis of the logarithm of LS (skewness –0·48 *v.* –0·10, median/s.d. –0·28 *v.* 0·00).

Figures 3 and 4 show the plot of residuals against minus the log of the survival curve of the residuals, for models 3 and 4, respectively. The intercept and slope are close to the expected values of 0·0 and 1·0, with values of 0·03 and 0·98 (model 3) and 0·03 and 0·95 (model 4), although there is a clear departure from expectation for small and large residuals, that may be attributed to the discrete nature of the data.

Discussion

We have shown the similarities and differences between a simple linear approach and a complex proportional hazards model when analysing longevity data in a small sample of dairy cattle using discrete time measures. Because we have used a linear approach method which can utilize censored observations, this is a first, albeit simple, comparison of the two methods for exactly the same data set which includes censoring. Our results suggest that

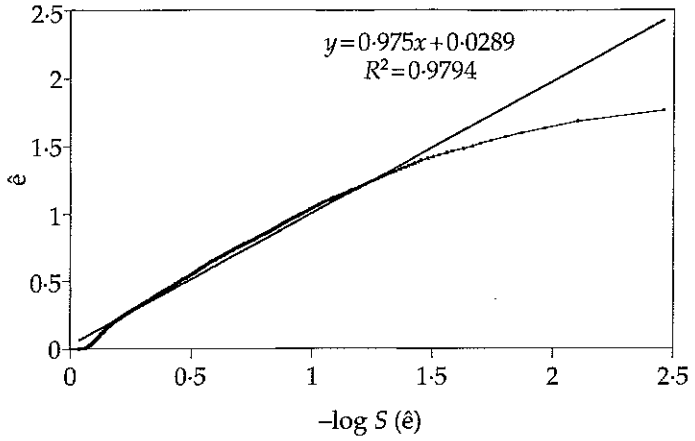

Figure 3 Graphical test of proportional hazard assumption for model 3, using generalized residuals, (ê).

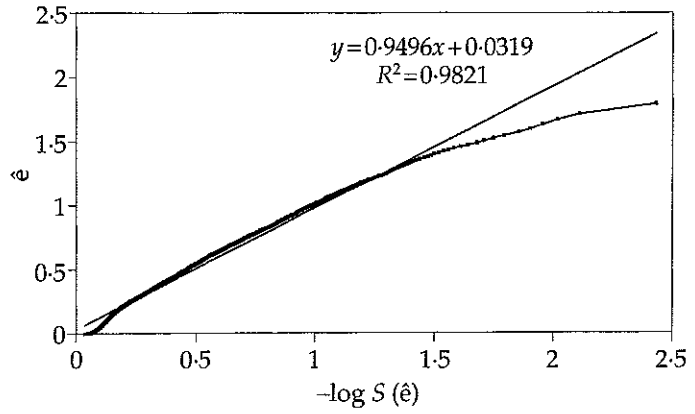

Figure 4 Graphical test of proportional hazard assumption for model 4, using generalized residuals (ê).

the main advantage in the complex model may be in modelling the fixed effects structure, because for simple models with time independent variates only, the correlations between breeding values for linear and non-linear models were high.

An assumption in the current genetic evaluation for lifespan in the UK is that the conditional survival probabilities for lactation 6 and higher is constant ($\rho = 0{\cdot}71$). From this study and studies elsewhere, it is probably more realistic to assume that these probabilities would decline. For example, it is unlikely that the survival probability of a cow in her 10^{th} lactation remains at 0·71. This is confirmed by the estimate of $\rho = 1{\cdot}6$ from the Weibull model (Table 4) which indicates an increasing hazard rate over time. The (phenotypic) prediction of the number of further lactations a cow is expected to remain in the herd could be adapted easily to reflect the decreasing conditional survival probabilities. However, for the analysis of lifespan, this is unlikely to change the ranking of sires, unless the prediction of remaining herdlife (after lactation 5) is severely biased upwards. In that case, breeding values of sires with a large proportion of censored daughters would be biased upwards. In the present study this potential bias was not a problem, because censored records were all censored at lactation 5.

In retrospect, the fact that we are considering only one cohort may lead to problems for 'old' cows when including a time-dependent covariate such as ΔM. During their fifth lactation, these cows are compared with other cows of their own group, and the deviation from herd mean is then the deviation from all other fifth lactation cows. It could be argued that calculating the deviation of these cows relative to all other cows in the herd would be more appropriate, since it is likely that the farmer will compare the production of these cows relative to the production of cows in their second, third, and fourth lactation. However, it is not clear whether the definition of 'herd-mean' to calculate milk deviations should include cows of all parities or cows of similar parities, since it is unlikely that a farmer will directly compare the milk yield of heifers to cows of higher parities.

The absence of a clear pattern in the risk ratios for months of calving (Table 5) with the time-independent model may reflect the fact that the month of calving in first lactation may be a poor indicator of the actual month of calving for the lactation in which the cow was culled. The difference between the RR of models 3 and 4 shows the advantage of fitting time dependent variates, i.e. fitting an appropriate fixed effects structure.

However, the model fitted was incomplete (e.g. no herd-year effect, only one cohort), and care should be taken in the interpretation of the actual solutions.

Vollema and Groen (1998) recently compared breeding values and genetic parameters between analyses of longevity from a linear model and a Weibull model. When the two methods were used on the same data set (uncensored records only), the correlation between sires' breeding values (or risk ratios) was high (0·93 to 0·94). Since the genetic values were calculated from the same data, sires had very large progeny groups (> 150 progeny, average between 600 and 700 daughters per sire), and data were uncensored, it is surprising that the correlation is not closer to unity. The differences in breeding values were presumably from fitting a herd-year-season, yield deviation, and stage of lactation effect in the Weibull model which were time-dependent. The correlations were much lower (~0·60) when censored data was included for the Weibull model.

The risk ratios for %HF classes decreased with increasing proportion of HF (from 1·37 to 0·81 in large herds) in the Vollema and Groen (1998) study, so our assumption of fitting a single continuous covariate may be justified, assuming HF proportion has a similar effect on longevity in UK dairy herds.

Vollema and Groen (1998) found that for their data, the heritabilities from the linear model and the transformed heritabilities from the proportional hazards model were very similar (~0·06). In our study, we found that the heritabilities on the log scale (Weibull and ln(LS)) were similar, but that the transformed heritability to the original scale was much larger than the estimated heritability on the LS scale (17% *v.* 5%). However, care should be taken in a comparison between the two heritabilities. The heritability on the log-scale has a standard definition, whereas the transformed heritability is an approximation which was originally derived as a tool to compute better approximations of reliabilities using standard selection index theory. The latter heritability has not been 'tested' thoroughly, but was found to be most useful in the prediction of the reliability of sire proofs (Ducrocq, 1999).

The Weibull model is defined for a continuous time variable, since the relationship between the hazard function $\lambda(t)$ and survival function $S(t)$ is, $\lambda(t) = -d\ln S(t)/dt$. In our case, we only have failure times of $T = 1, 2, 3, 4$, and 5. In addition, all cows had a qualifying first lactation ($S(0) = S(1) = 1$), so essentially we have four data points. However, the conditional survivals from one lactation to the next were similar (Table 1), and a plot of $\ln(-\ln(S(t)))$ against $\ln(t)$ shows a fairly straight line (results not shown). A more thorough investigation into the distribution of the residuals showed that for both the LS and ln(LS) linear model, residuals had a bi- or multi-modal distribution (Figures 1 and 2), due to the discrete nature of the time variates and the fact that only records of older cows were censored. The residuals from the Weibull analyses (Figures 3 and 4) showed a slight departure from expectation, but overall a good fit. Given this apparent fit, and the high correlation between breeding values from linear models and proportional hazard models, we are cautiously optimistic that it may be justified to use the standard Weibull model even for discrete time measures such as the number of completed lactations.

Proportional hazard models for discrete time variates have been proposed (Prentice and Gloeckler, 1978), but more research is needed in applications to large data sets with random genetic effects. We will pursue this matter in a subsequent study in which a larger data set from multiple cohorts will be analysed.

Acknowledgements

We acknowledge support from the BBSRC and the EU (GIFT Concerted Action), and thank the reviewers for their helpful suggestions and comments.

References

Arendonk, J. A. M. van. 1985. Studies on the replacement policies in dairy cattle. II. Optimum policy and influence of changes in production and prices. *Livestock Production Science* **13:** 101-121.

Brotherstone, S. and Hill, W. G. 1991. Dairy herd life in relation to linear type traits and production. 2. Genetic analyses for pedigree and non-pedigree cows. *Animal Production* **53:** 289-297.

Brotherstone, S. and Hill, W. G. 1994. Estimation of non-additive genetic parameters for lactations 1 to 5 and for survival in Holstein-Friesian dairy cattle. *Livestock Production Science* **40:** 115-122.

Brotherstone, S., Veerkamp, R. F. and Hill, W. G. 1997. Genetic parameters for a simple predictor of the lifespan of Holstein-Friesian dairy cattle and its relationship to production. *Animal Science* **65:** 31-37.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society (Series B)* **34:** 187-220.

Cox, D. R. and Snell, E. 1968. A general definition of residuals (with discussion). *Journal of the Royal Statistical Society (Series B)* **30:** 248-275.

Dekkers, J. C. M. and Jairath, L. K. 1994. Requirements and uses of genetic evaluations for conformation and herd life. In *Proceedings of the fifth world congress on genetics applied to livestock production, Guelph, vol. 17,* pp. 61-64.

Ducrocq, V. P. 1987. An analysis of length of productive life in dairy cattle. *Ph.D. thesis, Cornell University, New York.*

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

Ducrocq, V. P. 1999. Two years of experience with the French genetic evaluation of dairy bulls on production-adjusted longevity of their daughters. *Proceedings of the international workshop on EU concerted action for genetic improvement of functional traits in cattle; longevity. Interbull Bulletin* **21:** 60-67.

Ducrocq, V. P. and Casella, G. 1996. A Bayesian analysis of mixed survival models. *Genetics, Selection, Evolution* **28:** 505-529.

Ducrocq, V. P. and Sölkner, J. 1994. The Survival Kit, a FORTRAN package for the analysis of survival data. *Proceedings of the fifth world congress on genetics applied to livestock production, Guelph, vol. 22* pp. 51-55.

Ducrocq, V. P. and Sölkner, J. 1998a. *The Survival Kit V3.0, user's manual, 31 March, 1998.* Institut National de la Recherche Agronomique, Paris.

Ducrocq, V. P. and Sölkner, J. 1998b. The Survival Kit V3·0, a package for large analyses of survival data. *Proceedings of the sixth world congress on genetics applied to livestock production, Armidale, vol. 27,* pp. 447-448.

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

Famula, T. R. 1981. Exponential stayability model with censoring and covariates. *Journal of Dairy Science* **64:** 538-545.

Groeneveld, E. 1990. *PEST, user's manual.* Institute of Animal Husbandry and Animal Ethology, Federal Agricultural Research Centre, Mariensee, Germany.

Groeneveld, E. 1995. *REML VCE — A multivariate multimodel restricted maximum likelihood (co)variance component estimation package, version 3.1. User's guide.* Institute of Animal Husbandry and Animal Ethology, Federal Agricultural Research Centre, Mariensee, Germany.

Gröhn, Y. T., Ducrocq, V. P. and Hertl, J. A. 1997. Modelling the effect of diseases on culling in New York state Holstein dairy cows. *Journal of Dairy Science* **80:** 1755-1766.

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

Korsgaard, I. R., Andersen, A. H. and Jensen, J. 1999. Discussion of heritability of survival traits. *Proceedings of an international workshop on EU concerted action for genetic improvement of functional traits in cattle; longevity. Interbull Bulletin* **21:** 31-35.

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 cows. *Empire Journal of Experimental Agriculture* **18:** 49-56.

Ringmar-Cederberg, E., Johansson, K., Lundeheim, N. and Rydhmer, L. 1997. Longevity of Large White and Swedish Landrace sows. *Proceedings of the 48th annual meeting of the European Association for Animal Production, Vienna, Austria,* paper G3.6.

Smith, S. P. and Quaas, R. L. 1984. Productive lifespan of bull progeny groups: failure time analysis. *Journal of Dairy Science* **67:** 2999-3007.

Visscher, P. M., Bowman, P. and Goddard, M. E. 1994. Breeding objectives for pasture based production systems. *Livestock Production Science* **40:** 123-137.

Vollema, A. R. and Groen, A. 1998. A comparison of breeding value predictors for longevity using a linear model and survival analysis. *Journal of Dairy Science* **81:** 3315-3320.

Vukasinovic, N. J., Moll, J. and Künzi, N. 1997. Analysis of productive life in Swiss Brown cattle. *Journal of Dairy Science* **80:** 2572-2579.

Yazdi, M. H., Rydhmer, L., Ringmar-Cederberg, E., Lundeheim, N. and Johansson, K. 2000. Genetic study of longevity in Swedish Landrace sows. *Livestock Production Science* In press.

(*Received 7 May* 1999—*Accepted* 12 *October* 1999)