# Survival analysis using random regression models

***R. F. Veerkamp*[1]*, S. Brotherstone*[2] *and T. H. E. Meuwissen*[1]**

***[1]Dep. of Animal Breeding and Genetics,***
***DLO-Institute for Animal Science and Health,***
***P.O. Box 65, 8200 AB Lelystad, The Netherlands***
***and***
***[2]ICAPB, Univ. of Edinburgh,***
***West Mains Road, Edinburgh EH93JT, Scotland***

## Abstract

Censoring of records is a problem in the estimation of breeding values for longevity, because breeding values are required earlier in life than realised longevity can be measured. In this study we investigate the use of random regression models to analyse survival data, because this method combines some of the advantage of a multi-trait approach and the more sophisticated survival analysis. Production records in lactation 1 to 5 were available on 6320 cows in the UK, all having had the opportunity to survive 5 lactations. The random regression model contained a fixed cubic polynomial for lactation number (1 to 4), a herd effect (n = 167), a quadratic regression on milk yield within herd, a quadratic regression on age at calving within herd, Holstein percentage and year-season of calving effect (n=66). The additive animal genetic effects were modelled using a orthogonal polynomial of order 3 with random coefficients, and the error term was fitted as a diagonal matrix with different uncorrelated variances in each lactation. Variance components from the full (i.e. uncensored) data set, were used to estimate breeding values for survival in each lactation from both uncensored and censored data. Random censoring was applied proportionally to 0.3 or 0.6 of the cows, equal proportions of these censored animals had their last, last two or last three lactations set to missing. Two different procedures were applied: censoring or not censoring first lactation information. In the uncensored data, estimates of the residual variances were 0.15, 0.17, 0.17 and 0.18, and heritabilities were 0.03, 0.07, 0.05 and 0.01 for culling probability at the end of lactation 1 to 4, respectively. Breeding values for lifespan (calculated from the survival breeding values) had a range of 2.8 to 4.5 lactations and a standard deviation of 0.18. Correlations between predicted breeding values for 60 bulls, each with more than 20 daughters, from the various analyses ranged from 0.84 to 0.97. It is concluded that random regression analysis might be an alternative procedure to analyse censored survival data.

## 1. Introduction

Several traits associated with longevity have been considered for breeding value estimation (for reviews see Dekkers & Jairath, 1994, Essl, 1998). This is because longevity information on a cow is required earlier in life than real longevity can be measured. When actual longevity is considered, information on cows still alive is ignored, and therefore methods have been proposed that include cows still alive.

For instance, survival to a certain endpoint is proposed as a binary trait. However, information on herdlife before (and after) the endpoint is also ignored. Another method of accounting for censored records is to extent the records for cows still alive (VanRaden & Klaaskate, 1993). This is common practise when extending part lactation production records. A geometric distribution to expand lifespan of cows still alive to their predicted lifespan has also been used (Brotherstone *et al.*, 1997).

A multi-trait analysis was proposed where survival in each lactation was treated as a different trait (Madgwick & Goddard, 1989). Information from living cows can be treated as missing observations (i.e. for later lactations than the current one), and hence all information is taken account of.

A more sophisticated method of handling survival data is using a proportional hazard model, which has been adopted for animal breeding purposes (Ducrocq & Solkner, 1994). This method deals with censoring and the distribution of survival data, and another advantage of the survival kit is that time-dependant environmental effects can be included in the model. Apart from the complexity of proportional hazard models, there are, however, other disadvantages to the method: i) there is no multivariate implementation yet, which is particular important as most of the information during early life will come from predictor traits, e.g. linear type score (Brotherstone *et al.*, 1998) and ii) only one genetic effect is fitted for each animal during its whole life, i.e. the culling probability of two contemporaries have a constant ratio during their life. Although, in theory, this could be solved by using time-dependant sire-effects (Ducrocq, personal communication).

The objective of this study was to investigate the use of random regression for survival analysis, because it is expected that random regression models encompass some of the advantages of multitrait and survival analysis, especially the ability to include censored data and time dependant fixed effects in breeding value estimation.

## 2. Material and Methods

### *2.1. Data*

A subset of the data described by Brotherstone et al. (1997) was used: herds were selected that had more than 30 animals present which had the opportunity to survive 5 lactations. Lactation production records were available for 6320 animals in 167 herds. Records in lactation 1 to 4 were coded 0 or 1 depending on whether a next lactation was present or not. After culling (or removal from milk recording data set) lactation records were presented as missing. In this data set 1627 animals had a record for lactation 5 present, hence, these were censored (Table 1). This full data set was used to estimate variance components and breeding values.

**Table 1: Description of the full data set and the four data sets that were artificially censored.**

| | Full | Dat1 | Dat2 | Dat3 | Dat4 |
|---|---|---|---|---|---|
| Culled | 4693 | 3776 | 3259 | 2815 | 1855 |
| Censored | 1627 | 2544 | 3061 | 3505 | 4465 |
| Total | 6320 | 6320 | 6320 | 6320 | 6320 |
| | Records identified as missing per lactation | | | | |
| 1 | 0 | 0 | 913 | 0 | 1715 |
| 2 | 1561 | 2242 | 2242 | 2945 | 2945 |
| 3 | 2974 | 3565 | 3565 | 4187 | 4187 |
| 4 | 3977 | 4519 | 4519 | 5054 | 5054 |
| | Mean culling per lactation | | | | |
| 1 | 0.25 | 0.25 | 0.19 | 0.25 | 0.13 |
| 2 | 0.30 | 0.24 | 0.24 | 0.17 | 0.17 |
| 3 | 0.30 | 0.26 | 0.26 | 0.19 | 0.19 |
| 4 | 0.31 | 0.28 | 0.28 | 0.23 | 0.23 |

To investigate the effect of censoring on breeding value estimation, censoring was applied proportionally to 0.3 or 0.6 of the cows. Equal proportions of these censored animals had their last, last two or last three non-missing lactations set to missing. Two different procedures were applied: censoring or not censoring first lactation information. Hence, four data sets were created with increasing levels of censoring (Table 1).

### *2.2 Analysis*

### *2.2.1 Variance components*

Variance components were estimated using ASREML (Gilmour *et al.*, 1998). The random regression model contained a fixed cubic polynomial for lactation number (1 to 4), a herd effect (n = 167) and an effect for year-season of calving (n=66). Furthermore, a quadratic regression on milk yield within herd, a quadratic regression on age at calving within herd and a linear regression on Holstein percentage were included. The additive animal genetic effects were modelled using orthogonal polynomials of order 3 with random coefficients, and the error term was fitted as a diagonal matrix with different uncorrelated variance in each lactation. Cows and their (grand-) parents were included in the relationship matrix that contained 15372 individual animals; 818 sires sired the 6320 cows in the datafile, of which 177 had 10 daughters or more included.

### *2.2.2 Breeding values*

Breeding values for the three random regression coefficients were estimated for each animal in the pedigree file using ASREML. Variance components were fixed and the same model was used as was used to estimate the variance components.

As breeding values for the random components are not easy interpretable, these were transformed to culling probabilities in each of the four lactations by multiplying the breeding values with the appropriate coefficients of the polynomials. Breeding values for survival till the end of each lactation were calculated as follows:

$$surv_{il+1} = surv_{il} * (1 - (cull_{il} + \overline{cull_l}))$$

where $surv_{il}$ is the breeding value for survival and $cull_{il}$ the breeding value for culling of animal i in lactation l; $\overline{cull_l}$ is the mean culling probability in lactation l (Table 1). Finally, summing survival probabilities in each lactation gives a breeding value for lifespan for each animal. This procedure was repeated for all five data sets described in Table 1.

## 3. Results

The full data set of 6320 cows was used to estimate variance components. Of all these cows 1627 had a milk record present in lactation 5 and therefore are censored records (Table 1). Nearly half of the animals (2974) were culled before lactation 3, and subsequently were classified as missing in lactation 3. Culling probability was lowest at the end of the first lactation (i.e. no second lactation present conditional on a first lactation being present). Censoring had an obvious effect on mean culling probabilities, as relatively large proportions of records coded as 1 are deleted.

**Table 2: Estimates of the genetic variances of the coefficients of the quadratic polynomial (a,b,c), and the genetic correlations between them.**

| | a | b | c |
|---|---|---|---|
| a | 0.0208 | -0.52 | -0.99 |
| b | | 0.0040 | 0.59 |
| c | | | 0.0036 |

### *3. 1. Variance components*

The variance components for the additive genetic effect are given in Table 2. The third component was closely correlated to the first component. Estimates of the residual variances were 0.15, 0.17, 0.17 and 0.18.

The covariance function in Table 2 can be used to calculate the additive genetic (co)variances in each of the four lactations (Table 3). Heritabilities were small 0.03, 0.07, 0.05 and 0.01 in lactation 1 to 4, respectively. Genetic correlations between the first three lactations were above 0.83, however, genetic correlations with lactation four ranged from 0.0 to 0.55. That is probably related to only 2343 ‘none-missing’ records being present in lactation four, and hence there is probably too little information to estimate this genetic variance and genetic correlations accurately.

**Table 3: Estimate for the additive genetic variance for culling probability (diagonal) and the genetic correlations between culling at the end of each lactation, derived from the covariance function in Table 2.**

| | Culling 1 | Culling 2 | Culling 3 | Culling 4 |
|---|---|---|---|---|
| Culling 1 | 0.005 | 0.94 | 0.83 | 0.00 |
| Culling 2 | | 0.012 | 0.97 | 0.34 |
| Culling 3 | | | 0.009 | 0.55 |
| Culling 4 | | | | 0.002 |

### *3. 2. Breeding values*

Breeding values for the regression components are not very informative, and therefore summary statistics are given for breeding values for culling probability in each lactation (Table 4). These were derived for 60 sires that had at least 20 daughters in the full data set. Mean breeding values were close to zero in all lactations. The largest range was found in the second lactation where breeding values differed by as much as 0.31.

**Table 4. Summary statistics for breeding values for culling at the end of each lactation for 60 sires with at least 20 daughters.**

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Mean | -0.01 | -0.01 | -0.01 | 0.00 |
| Sd | 0.043 | 0.067 | 0.056 | 0.015 |
| Min | -0.11 | -0.18 | -0.15 | -0.03 |
| Max | 0.07 | 0.13 | 0.12 | 0.04 |
| Range | 0.18 | 0.31 | 0.27 | 0.07 |

Although these differences appear moderate, when bulls are compared for expected survival at the end of each lactation (Figure 1) reasonable difference can be observed in their genetic merit. All animals in the data set had survived lactation 1 (i.e. records were conditional on having a first lactation record), but at the end of lactation 5, proportionally 0.15 or 0.43 of the daughters survived of the worst and best sire, respectively.

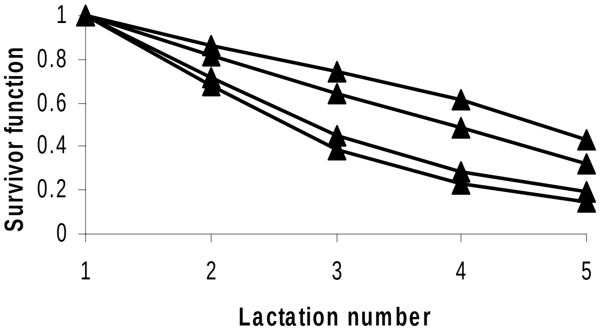

**Figure 1 Estimated survival function for the additive genetic merit of four extreme sires.**

**Table 5: Simple correlation between estimated breeding values for 60 sires from 5 data sets with increasing levels of censoring.**

| | Full | Dat1 | Dat2 | Dat3 |
|---|---|---|---|---|
| Dat1 | 0.95 | | | |
| Dat2 | 0.93 | 0.97 | | |
| Dat3 | 0.87 | 0.91 | 0.87 | |
| Dat4 | 0.84 | 0.86 | 0.86 | 0.92 |

For the sixty bulls considered in Table 4, breeding values for lifespan ranged from 2.4 to 3.7 lactations, i.e. there was a range of 1.3 lactation between the extreme bulls. The standard deviation of the breeding values was 0.28 lactations.

More importantly was the question of how robust these breeding values are to censoring of the data.

Correlations in Table 5 indicate that lifespan breeding values for the 60 sires were robust to censoring. Even after severe censoring (i.e. dat4) the correlation between the full data set and the censored data set remained above 0.83. Also, no obvious bias was apparent after censoring (Figure 2).

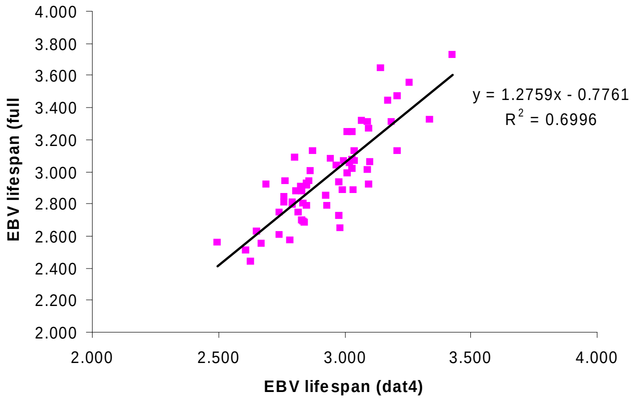

**Figure 2: Regression of breeding value for lifespan from the full data set on the same breeding value from the data set that is most heavily censored (Dat4).**

## 4. Discussion

Methodology to estimate breeding values for longevity should be able to cope with censored data and time dependant variables. Proportional hazard models seem to be the most appropriate method to handle this type of data. However the random regression methodology seems relatively robust to censoring of data also, at least as used in this study on discreet data. Furthermore, time-dependant variables can be included in this model as is illustrated by the year-season effect. Whilst these are modelled for each lactation record separately, account is taken of changing culling intensities over time in the data set. Further extension is straightforward by fitting interactions between time and other dependant variables, for example, herd year season effects. These effects would allow for change in culling policies in herds over time. Further advantages of the approach used herein are: i) Model and software are relatively similar to that under development for analysis of test day records for milk yield in some countries, and ii) it should be relatively straightforward to obtain breeding values for each cow. Although information for a cow is relatively limited, it should be relatively straightforward to include predictors of longevity in this analysis, e.g. linear type traits (Brotherstone *et al*., 1998, Jairath *et al*., 1998). This would utilise the two sources of information optimally (using appropriately estimated genetic and error correlations), and would enable combination of cows breeding values for longevity with other traits of economic importance in an index, e.g. like ITEM in the UK (Veerkamp *et al*., 1995).

Of course there are weaknesses in this approach and in the analysis used here. For example, we have treated binary data as if it was continuous, and assumed uncorrelated normally distributed error terms in each lactation. The latter while there are repeated records for each cow. Hence, more appropriate error structures are required, although these might not be obvious given that each animal has a string of zeros ended with a single 1. It might be necessary to define the error structure depending on which records are available for a cow (Jairath *et al*., 1998).

In this analysis there seems to be a problem with culling at the end of lactation four. Low heritabilities, unrealistically low genetic correlations and little variation in the breeding values were obtained. These problems have been subscribed to too little data present in the last lactation. Initial analysis on another larger data set confirm this, but there might still be other systematic effects caused by the random regression model, or the used legendre polynomials. These require further investigation, and non-parametric curves might give better solutions. Given the little information in later lactation, it is also difficult to get a clear picture on differences between survivor functions of bulls. Also, results from this study, i.e. variance components and breeding values, have not been tested against other methods dealing with censored longevity records. Therefore the conclusion from this study might be that random regression models are an alternative to proportional hazard model, because time dependant variables can be fitted and, at least in the data used here, breeding value estimation appears relatively robust to censoring of the data.

## Acknowledgements

Vincent Ducrocq provided many useful points for the discussion, his contribution is acknowledged.

## References

Brotherstone, S., Veerkamp, R. F. & 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.

Brotherstone, S., Veerkamp, R. F. & Hill, W. G. (1998). Predicting breeding values for herd life of Holstein-Friesian dairy cattle from lifespan and type. *Animal Science* **67**, 405-411.

Dekkers, J. C. M. & Jairath, L. K. (1994). Requirements and uses of genetic evaluations for conformation and herd life. In *Proceedings, 5th*

*World Congress on Genetics Applied to Livestock Production*.

Ducrocq, V. P. & Solkner, J. (1994). "The survival kit", a Fortran package for the analysis of survival data. In *Proceedings, 5th World Congress on Genetics Applied to Livestock Production.*

Essl, A. (1998). Longevity in dairy cattle breeding: A review. *Livestock Production Science. Dec.* **57**, 79-89.

Gilmour, A. R., Cullis, B. R., Welham, S. J. & Thompson, R. (1998). ASREML. Program user manual. *NSW Agriculture, Orange Agricultural Institute, Forest Road, Orange, NSW, 2800, Australia.*

Jairath, L., Dekkers, J. C. M., Schaeffer, L. R., Liu, Z., Burnside, E. B. & Kolstad, B. (1998). Genetic evaluation for herd life in Canada. *Journal of Dairy Science* **81**, 550-562.

Madgwick, P. A. & Goddard, M. E. (1989). Genetic and phenotypic parameters of longevity in Australian dairy cattle. *Journal of Dairy Science* **72**, 2624-2632.

VanRaden, P. M. & Klaaskate, E. J. H. (1993). Genetic evaluation of length of productive life including predicted longevity of live cows. *Journal of Dairy Science* **76**, 2758-2764.

Veerkamp, R. F., Hill, W. G., Stott, A. W., Brotherstone, S. & Simm, G. (1995). Selection for longevity and yield in dairy cows using transmitting abilities for type and yield. *Animal Science* **61**, 189-197.