Modeling Lactation Curves and Estimation of Genetic Parameters for First Lactation Test-Day Records of French Holstein Cows

T. Druet¹, F. Jaffrézic, D. Boichard, and V. Ducrocq

Station de Génétique Quantitative et Appliquée, INRA, Jouy-en-Josas 78352, France

ABSTRACT

Several functions were used to model the fixed part of the lactation curve and genetic parameters of milk test-day records to estimate using French Holstein data. Parametric curves (Legendre polynomials, Ali-Schaeffer curve, Wilmink curve), fixed classes curves (5-d classes), and regression splines were tested. The latter were appealing because they adjusted the data well, were relatively insensitive to outliers, were flexible, and resulted in smooth curves without requiring the estimation of a large number of parameters.

Genetic parameters were estimated with an Average Information REML algorithm where the average information matrix and the first derivatives of the likelihood functions were pooled over 10 samples. This approach made it possible to handle larger data sets. The residual variance was modeled as a quadratic function of days in milk.

Quartic Legendre polynomials were used to estimate (co)variances of random effects. The estimates were within the range of most other studies. The greatest genetic variance was in the middle of the lactation while residual and permanent environmental variances mostly decreased during the lactation. The resulting heritability ranged from 0.15 to 0.40. The genetic correlation between the extreme parts of the lactation was 0.35 but genetic correlations were higher than 0.90 for a large part of the lactation. The use of the pooling approach resulted in smaller standard errors for the genetic parameters when compared to those obtained with a single sample.

(Key words: genetic parameters, lactation curve, test-day model)

Abbreviation key: AIC = Akaike's Information Criterion, BIC = Schwarz' Bayesian Information Criterion, DCC = days carried calf, DO = days open, MSSE = mean sums of squares of residuals, TD = test-day.

INTRODUCTION

In recent years, numerous studies have investigated the topic of genetic evaluation of dairy cattle using test-day (TD) models. Advantages of the TD model over an approach using 305-d lactation yields are now widely acknowledged. France, like many other countries, plans to implement a TD model for routine genetic evaluation. However, some issues must still be investigated because there is a large variation in models among countries.

Indeed, several approaches have been used for modeling the fixed part of the lactation curve (Jensen, 2001) such as fixed classes curves using classes of DIM (Pool et al., 2000), or parametric curves such as the Ali-Schaeffer curve (Ali and Schaeffer, 1987), the Wilmink curve (Wilmink, 1987) or orthogonal polynomials (Olori et al., 1999). For the modeling of random effects, this variability has ranged from parametric curves to the use of eigenvectors to reduce the rank of the (co)variance matrices. Recently, White et al. (1999) proposed the use of natural cubic splines.

The first step in implementing a routine evaluation with TD models is to estimate variance components. A large heterogeneity of the estimated genetic parameters (Misztal et al., 2000) indicated the need to use large data sets. Unfortunately, the TD models are computationally very expensive and therefore data sets of reduced size typically have been used. Several studies relied on Gibbs sampling but the size of the data sets was still limited and chains took very long to converge (Jamrozik et al., 1998).

In addition to the large data sets required, some authors have found it important to model the heterogeneity of the residual variance across the lactation (Brotherstone et al., 2000; Jaffrezic et al., 2000). Most methods work with different classes of residual variance across the lactation but these require a large number of parameters or do not result in a smooth residual variance curve.

The objective of this study was to develop tools for the genetic evaluation of dairy cattle with TD models in France. The specific tasks were: 1) to compare spline functions to more traditional curves for the fixed part of the lactation curve, 2) to develop a strategy of variance

Received December 23, 2002.

Accepted January 21, 2003.

Corresponding author: T. Druet; e-mail: tom.druet@dg.jouy.inra.fr.

¹On leave from Animal Science Unit, Gembloux Agricultural University, Gembloux B-5030, Belgium.

Table 1. Numbers of test-day records, animals with records, animals in pedigree files and herd-test-date effects for the 10 analyzed samples.

Data sample	Number of test-day records	Number of animals with records	Number of animals in pedigree file	Number of herd-test-date effects
1	72,545	9149	14,591	8500
2	82,321	10,362	16,386	9632
3	80,551	10,149	15,871	8963
4	83,220	10,533	16,493	8894
5	81,507	10,382	16,054	8475
6	75,061	8648	13,881	8327
7	79,993	9213	14,718	9131
8	69,489	8066	13,003	8112
9	85,519	10,835	16,247	9953
10	66,652	8486	13,150	7249
Total	776,858	95,823	150,394	87,236

component estimation for random regression applicable to very large data sets, 3) to develop a method for taking into account the continuous heterogeneity of the residual variance over the lactation period that does not require a large number of parameters, and 4) to estimate variance components for milk yield of first parity cows.

MATERIALS AND METHODS

Data

The data were selected from 1,699,925 first lactation records of Holstein cows from the four French administrative regions of Bretagne. The following edits were applied to the data: calving date was required to be in the period from August 1994 to July 2000, and age at first calving had to be from 20 through 38 mo. Records with fewer than 40 d open, from animals with unknown parents, or with unknown calving age were also discarded. Only lactations with at least four TD records known were considered and lactation stage of TD records had to be between 5 and 305 d. Herds with an average of at least 72 TD records per year were kept. Finally a random selection on herd number was applied to create 10 data sets, each with approximately 10,000 cows with records. Those data sets are characterized in Table 1.

Lactation length was limited to 305 d for several reasons: 1) 305 d are the present official lactation length in France and for ICAR, 2) TD records after 305 d might bring only limited additional information and might not reflect the ability to produce during the 305-d lactation period, 3) any TD records after 305 d might be more difficult to model and problems in their modeling might affect the evaluation for the entire lactation, 4) taking those TD records into account might penalize cows with good fertility and shorter lactations. However, we plan

to investigate the use of longer lactation length in the future.

Model

Lactation curves. Five different functions were compared to model the fixed part of the lactation curve: a fifth order (quartic) Legendre polynomial, the Wilink curve (Wilink, 1987), the Ali-Schaeffer curve (Ali and Schaeffer, 1987), a fixed classes curve with 5-d classes for DIM (30 classes) and regression splines as defined by White et al. (1999). The regression splines are equivalent to the natural cubic spline without roughness penalty. In this study, six knots were chosen at 5, 20, 50, 130, 230 and 305 DIM.

To compare these functions, the same fixed effect model was assumed:

$$y_{ijk} = \text{HTD}_i + f_1(\text{age}_j, \text{dim}) + f_1(\text{month}_k, \text{dim}) + e_{ijk} \quad [1]$$

where y_{ijk} is the milk record, HTD_i is the herd by test-date effect, i is the HTD level, f_1 is the tested function, age_j represents the age at calving in class j (nine classes: 20–22, 23–24, 25–26, 27–28, 29–30, 31–32, 33–34, 35–36, and 37–38 mo), month_k represents the month at calving k (12 classes) and e_{ijk} is the residual term. In total, there were 21 different curves. Curves for the age effect were added to the curves for the month at calving effect because both were assumed independent.

Effect of gestation. After comparison of the lactation curves, three functions were tested to model the effect of the gestation or days carried calf (DCC). $\text{DCC} = \text{DIM} - \text{DO}$ where **DO** is the number of days from calving to successful insemination. When there is no successful insemination $\text{DO} = \text{DIM}$ and $\text{DCC} = 0$. If DCC was fewer than 100 d, the effect of the gestation on milk production was assumed to be nil. The DCC effect ranged from 100 to 265 d because the lactation

length was limited to 305 d and minimum DO was 40 d. The effect of DCC was not allowed to differ with DIM because large DCC values could only be observed in late lactation stages while small DCC values were assumed to have small effects not varying much with DIM. The three tested functions were Legendre polynomials (order five), a fixed classes curve with 5-d classes and regression splines with four, five and six knots at DCC 100/150/200/265, 100/140/180/220/265 and 100/133/166/199/232/265, respectively. The new model was:

$$y_{ijk} = \text{HTD}_i + f_1(\text{age}_j, \text{dim}) + f_1(\text{month}_k, \text{dim}) + f_2(\text{dcc}) + e_{ijk} \quad [2]$$

where f_1 is the function chosen for modeling the lactation curve (regression splines—see results) and f_2 is the tested function for modeling the gestation effect.

Model for the estimation of the genetic parameters. In addition to the fixed effects tested earlier, random effects were added to the model:

$$y_{ijkln} = \text{HTD}_i + f_1(\text{age}_j, \text{dim}) + f_1(\text{month}_k, \text{dim}) + f_2(\text{dcc}) + \sum_{m=0}^4 (a_{lm} + p_{lm})\varphi_{(m,t)} + \sum_{m=0}^3 h_{ynm}\varphi_{(m,t)} + e_{ijkln} \quad [3]$$

where f_2 is a regression spline as described earlier with DCC ranging from 0 to 265 d of gestation and with five knots chosen at 0, 100, 150, 200 and 265 DCC, a_{lm} and p_{lm} are the random additive genetic and permanent environmental regression coefficients of animal l for the m^{th} term of the Legendre polynomial of order 5, respectively. Parameter $\varphi_{(m,t)}$ is the value of the m^{th} term of the Legendre polynomial at time t (DIM standardized from -1 to 1) as in Kirkpatrick et al. (1990), and h_{ynm} is the random regression coefficient of the herd by year of calving n for the m^{th} term of the Legendre polynomial of order four. Legendre polynomials were chosen because in preliminary studies, it was found that the variance component structures (eigenvectors and eigenvalues) obtained by this method and by an unstructured model (with 10 by 10 unstructured covariance matrices, corresponding to a multi-trait model) were very similar.

The phenotypic covariance matrix $\mathbf{V}$ of the observations is given by:

$$\mathbf{V} = \begin{pmatrix} \mathbf{G} \otimes \mathbf{A} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P} \otimes \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H} \otimes \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R} \end{pmatrix} \quad [4]$$

where $\mathbf{G}$, $\mathbf{P}$, and $\mathbf{H}$ are the random regression covariance matrices for the genetic, permanent environmen-

tal and herd by year of calving effects, respectively, $\mathbf{A}$ is the additive genetic relationship matrix, and $\mathbf{R}$ is the diagonal matrix of the residual variance that depends on DIM: $\mathbf{R} = \text{diag}[\sigma_{e_q}^2]$ where q denotes DIM and

$$\sigma_{e_q}^2 = \exp(a + bt_q + ct_q^2) \quad [5]$$

where $t_q = (-1 + 2 \frac{\text{dim}-5}{305-5})\sqrt{3}$ which is proportional to the second term of the Legendre polynomial. Parameters a , b , and c describe the variation of the residual variance over DIM, and the exponential function ensures that the residual variance is always positive. Other sources of heterogeneity of variance were not included in the model for genetic parameter evaluation. However, they are planned to be modeled in the national genetic evaluation system and are under study.

Method

Lactation curves comparison. Lactation curves were compared using five criteria. First, the mean sum of squares of the residuals (**MSSE**) were computed for the first data sample:

$$\text{MSSE} = \frac{1}{n} \sum_{i=1}^n (y_{ijk} - \text{HTD}_i - f_1(\text{age}_j, \text{dim}) - f_1(\text{month}_k, \text{dim}))^2 \quad [6]$$

where n is the number of records. This parameter indicates the overall fit of the curve. Second, the mean residual was computed for each DIM:

$$\bar{r}_{(\text{dim}=d)} = \frac{1}{n_{d,i=1}} \sum_{i=1}^{n_d} (y_{ijk} - \text{HTD}_i - f_1(\text{age}_j, d) - f_1(\text{month}_k, d)) \quad [7]$$

where n_d is the number of records at DIM = d . By plotting this second parameter over DIM, it is possible to check whether the fit is adequate or if there is a bias at any lactation stage.

In addition to those two criteria, twice the opposite of the logarithm of the likelihood function and the REML forms of Akaike's Information Criterion (**AIC**) and Schwarz' Bayesian Information Criterion (**BIC**) were computed (e.g., Meyer, 2001).

Effect of gestation. The three functions were already compared in the previous part. The effect of DCC were plotted for each function for the first sample.

Method for estimating the genetic parameters. A program from Ignacy Misztal and Shogo Tsuruta

(Misztal et al., 2002) based on the Average Information REML algorithm and relying on the work of Jensen et al. (1996) was modified in order to accommodate all the features of the models to be tested. The program was extended to random regression models and to correlated effects and traits. Computation of the $\mathbf{f}$ vectors (see Jensen et al., 1996 and below) and combination of Average Information-REML with an EM-REML algorithm (in the case of non-positive definite matrices) were rewritten.

Model for the estimation of the residual variance. The residual variance was described with a function as in equation [5]. For this purpose, first derivatives of the log-likelihood and of the $\mathbf{f}$ vectors had to be computed for the three parameters (a, b and c) of σ_e^2 .

In Jensen et al. (1996), $\mathbf{V}$ is equal to $\text{var}[\mathbf{y}]$ or to $\mathbf{ZGZ}' + \mathbf{R}$, and $\mathbf{P}$ is a projection matrix mapping observations into weighted residuals:

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} \quad [8]$$

The average information matrix $\mathbf{I}_A(\theta)$ (where θ is the vector of parameters) can be computed as (Jensen et al., 1996):

$$\mathbf{I}_A(\theta) = \mathbf{F}'\mathbf{P}\mathbf{F} = \mathbf{F}'\mathbf{R}^{-1}\mathbf{F} - \mathbf{C}'\mathbf{W}'\mathbf{R}^{-1}\mathbf{F}\mathbf{W}'\mathbf{R}^{-1}\mathbf{F} \quad [9]$$

where $\mathbf{W} = [\mathbf{X} \ \mathbf{Z}]$ and $\mathbf{F}$ is a matrix whose j^{th} column $\mathbf{f}_j$ ($\mathbf{f}$ vector) consists of the vector $\frac{\partial \mathbf{V}}{\partial \theta_j} \mathbf{P}\mathbf{y}$.

One needs to compute $\frac{\partial \mathbf{V}}{\partial \theta_{\mathbf{R}(j,k)}}$ which is equal to $\mathbf{R}_{jk}$, a symmetric indicator matrix containing ones in positions corresponding to parameter $\theta_{\mathbf{R}(j,k)}$ (which is the parameter of the residual variance matrix corresponding to (co)variance between traits j and k) and zeros elsewhere. In the case of a constant variance and one lactation, there is one single parameter in θ corresponding to the residual variance ($\theta_{\mathbf{R}(1,1)}$). In the case of a residual variance described by three parameters a, b and c, θ contains three parameters related to the residual variance and:

$$\begin{aligned} \frac{\partial \mathbf{V}}{\partial a} &= \frac{\partial (\mathbf{ZGZ}' + \mathbf{R})}{\partial a} = \frac{\partial \mathbf{R}}{\partial a} = \frac{\partial \text{diag}\{e^a + bt_q + ct_q^2\}}{\partial a} \\ &= \text{diag}\{e^a + bt_q + ct_q^2\} = \mathbf{R} \end{aligned} \quad [10]$$

Similarly,

$$\frac{\partial \mathbf{V}}{\partial b} = \mathbf{R} \text{diag}\{t_q\} \text{ and } \frac{\partial \mathbf{V}}{\partial c} = \mathbf{R} \text{diag}\{t_q^2\} \quad [11]$$

Derivatives obtained from equations [10] and [11] are easy to implement in the program because the computa-

tion of the first derivative of the log-likelihood and of the $\mathbf{f}$ vectors are (Jensen et al., 1996):

$$\begin{aligned} \frac{\partial \mathbf{L}(\theta)}{\partial \mathbf{R}_{(i,j)}} &= \text{tr}[\mathbf{R}_{ij}\mathbf{R}^{-1}] - \text{tr}[\mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{R}_{ij}\mathbf{R}^{-1}\mathbf{W}] \\ &\quad - \hat{\mathbf{e}}'\mathbf{R}^{-1}\mathbf{R}_{ij}\mathbf{R}^{-1}\hat{\mathbf{e}} \end{aligned} \quad [12]$$

$$\mathbf{f}(\theta_{\mathbf{R}(i,j)}) = \mathbf{R}_{ij}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}}) \quad [13]$$

Here, $\mathbf{R}_{ij}$ is replaced by values given in equations [10] and [11]. This leads to some simplifications of equations [12] and [13] because $\mathbf{R}$ is multiplied by its inverse. The Average Information-REML algorithm can easily be extended to other parametric functions for the residual variance.

Pooling method for the estimation of the genetic parameters. In addition, the Average Information-REML program was transformed for the simultaneous estimation of the genetic parameters of several samples. Instead of combining the genetic parameters obtained from different samples *a posteriori*, the approach cumulated the likelihood, the first derivatives and the average information matrix over several samples (see Babb, 1986; Yerex, 1988; Ducrocq, 1993) such that:

$$\mathbf{L}(\theta) = \sum_{i=1}^n \mathbf{L}_i(\theta) \quad [14]$$

where $\mathbf{L}(\theta)$ is the log-likelihood for all samples for the set of parameters θ , and $\mathbf{L}_i(\theta)$ is the log-likelihood for sample i for the same set of parameters and n is the number of samples.

$$\frac{\partial \mathbf{L}(\theta)}{\partial \theta_j} = \sum_{i=1}^n \frac{\partial \mathbf{L}_i(\theta)}{\partial \theta_j} \quad [15]$$

where θ_j is the j^{th} parameter and

$$\mathbf{I}_A(\theta) = \sum_{i=1}^n \mathbf{I}_{A_i}(\theta) \quad [16]$$

The set of parameters θ is updated after these terms are computed for all the samples using:

$$\hat{\theta}^{(nr+1)} = \hat{\theta}^{(nr)} - \mathbf{I}_A^{-1} \frac{\partial \mathbf{L}(\theta)}{\partial \theta} \quad [17]$$

where $\hat{\theta}^{(nr)}$ is the parameter estimation at iteration nr .

In addition, with the Average Information-REML algorithm, one can obtain standard deviation of the estimated parameters through the inverse of the average information matrix. The comparison of these standard deviations obtained from one single sample or from the

Table 2. Mean sum of squares of residuals (MSSE), $-2 \times$ logarithm of the likelihood function ($-2\log L$), AIC and BIC values for the fifth order Legendre polynomial, the Wilmlink function, the Ali-Schaeffer curve, a fixed classes curve and a regression spline function.

Curve	Number of parameters	MSSE	$-2\log L^{1,2}$	AIC ²	BIC ²
Legendre polynomial	8605	1213	5152	2842	521
Wilmlink	8563	1212	5148	2754	318
Ali-Schaeffer	8605	1209	4631	2321	0
Fixed classes	9760	1187	0	0	846
Regression spline	8626	1206	4633	2365	102

¹ $-2\log L$: $-$ twice the logarithm of the likelihood function.

²Values are expressed as a deviation of the best value.

ten samples together was used to measure the gain of precision with the pooling method. This comparison was applied for the estimates of (co)variances of the coefficient of the random regression model. The variances (and thus the standard deviations) of the correlations were also computed with the following approach, known as the “delta method” (Oehlert, 1992):

$$\text{var}(f(\theta)) = \frac{\partial f(\theta)'}{\partial(\theta)} \text{var}(\theta) \frac{\partial f(\theta)}{\partial(\theta)} \quad [18]$$

where $f(\theta)$ here is the function computing the correlation between two given DIM and $\text{var}(\theta)$ is the asymptotic variance of the parameters obtained from the inverse of the Average Information matrix in our case.

Transformation of the genetic parameters on a 305 DIM scale. The genetic variance matrix for all DIM was obtained as:

$$\mathbf{G}^* = \Phi \mathbf{G} \Phi' \quad [19]$$

where $\mathbf{G}^*$ is a 301 by 301 genetic (co)variance matrix for all DIMs ranging from 5 to 305 d. Φ is a 301 by 5 matrix with the values of the five coefficients of the fifth order Legendre polynomial for each DIM from 5 to 305 d. The same operation was applied to $\mathbf{P}$ and $\mathbf{H}$. The eigenvectors of $\mathbf{G}^*$ and $\mathbf{P}^*$ were computed subsequently.

RESULTS AND DISCUSSION

Lactation Curves

The ranking of the functions with different criteria are presented in Table 2. According to the MSSE criterion, the fixed classes curve performed better than the others for which the differences in MSSE were marginal. Except for the fifth order Legendre polynomial, the ranking of the functions was simply the reverse of the ranking of the number of parameters per curve, 60, 6, 5, 3 and 5, respectively. This clearly explained why

the fixed classes curve had the best fit. These results were consistent with our previous studies on other samples. The ranking based on the logarithm of the likelihood function or on the AIC criterion was the same as for the MSSE, with a permutation between the spline function and the Ali-Schaeffer curve. However, with the BIC criterion, with a stronger penalty for large number of parameters, the fixed classes curve ranked last. Also for these three criteria, some differences between curves were marginal. With criteria based on goodness-of-fit, it is not trivial to conclude which of these curves is the best and different rankings might be obtained across samples or studies. Moreover, a compromise, often based on subjective grounds, must be accepted between goodness-of-fit and other properties such as flexibility, robustness and computational considerations. This explains the large variation in models across studies and countries.

Figure 1 shows the trend of the mean residuals over DIM. For the non-parametric curve, as expected, the mean residuals were zero for all parts of the curve. This was also true for the regression spline except for a small deviation at the beginning and the end of the lactation. The last three curves showed biases throughout the lactation. For instance, all of them underestimated milk yield at the peak and overestimated it in early lactation indicating that they were not able to model all the variation of the curve. This was also found by Jamrozik et al. (1997) for the Ali-Schaeffer and Wilmlink curves. Because the Wilmlink curve had only three parameters, it had reduced flexibility. For instance, the slope of the curve was modeled as a constant from the peak to the end of the lactation, although there seemed to be an inflection point around 200 d. As a consequence, the Wilmlink curve was systematically too high before the inflection point and too low after it. The Ali-Schaeffer curve and the Legendre polynomial showed problems throughout the lactation. The Legendre polynomials had typical border effects at the beginning and the end of the lactation (increased bias) and showed waves in

Figure 1. Trend of mean residual in function of DIM for: — fixed classes curve; —◆— spline function; —▲— Wilmink curve; —●— Ali and Schaeffer curve; —△— fifth order Legendre polynomial.

the middle of the lactation. Parametric curves were not flexible enough to track all trends in the data.

For the fixed classes curve, the influence of the data is local: a record observed in one class does not influence the estimate for any other class. In addition, because it has a very large number of parameters, it shows great flexibility. This can be a problem, however, for classes with a small number of observations. For the three parametric curves, any record influences the whole curve and the flexibility is less important than for the fixed classes curve. As a consequence, atypical curves can be obtained when there are outliers. For instance, with the Wilmink curve, if some points at the end of the lactation are particularly high, the whole part of the curve after the peak will rise. Regression splines are semi-parametric curves. The influence of the points is not completely local but records have an influence only on the parts of the curve close to them and they show greater flexibility than the other parametric curves. Therefore, they offer a compromise between the fixed classes curve and the parametric functions with better flexibility than parametric curves and a reduced number of parameters.

Effect of Gestation

Figure 2 shows curves obtained for the effect of DCC for different functions. First, the fixed classes curve had good flexibility and local behavior. However, the curve

had random fluctuations that might be due to sensitivity to the data rather than to real biological reasons and might be undesirable. When the number of records within classes decreases, the sampling variance of the estimate of the mean of the class increases. This creates more local variation that might result in inconsistent curves because of few unexpected records. For instance, the effect for the last class was much higher than for the other classes. Also, there was no reason for the DCC effect to decrease after 230 d of gestation. This might arise from missing components in the models and might disappear when all effects, such as the animal genetic effect, are included. This problem was also observed for the Legendre polynomial, for which the DCC effect also increased at the end of the lactation. In fact, Meyer (1998) observed that data points at the beginning and at the end of the lactation trajectory have a relatively large impact on the regression coefficient estimates.

In contrast to fixed classes curves, splines showed little local variation. With cubic splines, there is a roughness penalty parameter which penalizes rapid variation and smoothes the curve (White et al., 1999, Green and Silverman, 1994). The purpose of smoothing is to reduce random fluctuation associated with some outlier records. In addition to the roughness penalty, the limitation of the number of knots has also a smoothing effect. Regression splines rely only on this second source of smoothing (White et al., 1999). So, as the

Figure 2. Effect of days carried calf (DCC) on test-day milk production modeled with: 1) 5 d classes curve, 2) Fifth order Legendre polynomial, 3) Spline function with 4 knots, 4) Spline function with 5 knots and 5) Spline function with 6 knots.

number of knots increases, the curves fit the data better and become rougher and more sensitive to the data. The regression splines are again a compromise between the fixed classes and the parametric curves. The balance between all these properties can be adjusted by selection of the number and the position of the knots. The regression spline with 6 knots was more sensitive to the data; and the last part of the curve was essentially influenced by the few records between knots at 232 and 265 DCC. With the second curve, data ranging from 220 to 265 DCC determined the end of the curve, resulting in a decreasing curve but with an inflection point. Points between 220 and 232 DCC were more numerous and greatly influenced the shape of the end of the curve. Finally, the last regression spline, with the last knots at 200 and 265 DCC, was influenced little by the few points after 232 DCC and resulted in a smoother and decreasing curve without an inflection point.

In conclusion, regression splines offer a good compromise between goodness-of-fit, sensitivity to the data, smoothness, local behavior and the number of parameters necessary to fit the curve. Additionally, fixed classes curves do not rely on prior assumptions about the shape of the curve. They can also be used for other traits such as fat or protein contents or for other species and traits.

Regarding the gestation effect, different curves for the effect of DCC would result in a different impact on

high milk yielding cows with fertility problems. The regression spline with four knots, with the effect of DCC increasing exponentially, seemed to be in agreement with the results from Smith and Legates (1962) and Olori et al. (1997). This makes biological sense as the growth of the embryo is exponential. However, it might be that the curve described by the Legendre polynomial or the spline with 6 knots corresponds to biological reality. In all cases, the magnitude of the effect must be more properly estimated with all the other terms (e.g., additive genetic, permanent environment) included in the model.

Estimation of the Genetic Parameters

The method to model the residual variance as a continuous function of DIM resulted in estimates very close to the ones obtained in a multivariate analysis with 10 classes calculated with ASREML (Gilmour et al., 2000). The use of this function allowed a reduction of the number of parameters needed to fit the data (Pool and Meuwissen, 2000). The method seemed more appealing than the use of a finite number of classes as in Olori et al. (1999) and Rekaya et al. (1999) because it allows to model the continuous changes of the residual variance over time with a small number of parameters (only three here).

Variance function estimates obtained are presented in Figure 3. These variances were in agreement with

Figure 3. Variances across DIM for milk yield in first lactation: —●— residual variance, —■— genetic variance, —▲— permanent environmental variance and — herd-year variance.

multi-trait studies where lactation curves were split into different traits, as presented by Meyer et al. (1989), Pander et al. (1992), Swalve (1995), Rekaya et al. (1999), and Pool et al. (2000). According to these studies, the residual variance was found to be decreasing throughout the lactation with a slight increase at the end. Also in agreement with these authors, the genetic variance was highest in mid-lactation and estimates were lower at the beginning and the end of the lactation.

Published results obtained in different studies with random regression models were very heterogeneous (Misztal et al., 2000). The shapes obtained in this study were in agreement with Pool et al. (2000) and Mayeres (2002) both using Legendre polynomials. Olori et al. (1999) obtained an increasing genetic variance and Jakobsen et al. (2001) found a highly increasing residual variance at the end of the lactation (higher than at the beginning of the lactation) and an increasing genetic variance throughout the lactation. Variance of the random herd by year of calving effect was similar to that obtained in previous studies (Gengler and Wiggans, 2001; Auvray and Gengler, 2002; De Roos et al., 2002). The shape and amplitude were the same as well.

Maximum heritability was close to 0.39 and was found around 200 DIM. The minimum was at the beginning of the lactation and there was a decreasing heritability at the end (see Figure 4). This was a consequence of the maximum genetic variance in the middle of the lactation and the decreasing residual and permanent

environmental variances. Heritability ranged from 0.16 to 0.39 which are medium values in comparison with other studies. Most multi-trait analyses also found the highest heritability in mid-lactation. This also was true for the study from Liu et al. (2000) with the lactation separated into 6 traits with a covariance function fit in a second step. Similar results were obtained by White et al. (1999) working with cubic splines or in studies based on random regression from Pool et al. (2000), Auvray and Gengler (2002), Jakobsen et al. (2001) or Mayeres (2002). However, in some of these studies, this similar heritability was obtained from very different

Figure 4. Heritability of milk yield across DIM for first parity.

Table 3. Genetic correlations (above diagonal) and their standard deviations (below the diagonal) between daily milk yield at d 5, 20, 50, 80, 155, 230, 275, 290 and 305.

DIM	DIM								
	5	20	50	80	155	230	275	290	305
5		0.968	0.829	0.722	0.598	0.541	0.469	0.420	0.346
20	0.002		0.942	0.870	0.761	0.695	0.617	0.563	0.480
50	0.011	0.004		0.984	0.915	0.847	0.771	0.718	0.632
80	0.015	0.007	0.001		0.967	0.908	0.838	0.789	0.707
155	0.017	0.011	0.004	0.002		0.978	0.928	0.888	0.818
230	0.019	0.012	0.007	0.005	0.001		0.981	0.954	0.899
275	0.020	0.014	0.009	0.007	0.004	0.001		0.994	0.966
290	0.022	0.016	0.011	0.009	0.006	0.003	0.001		0.989
305	0.025	0.020	0.015	0.013	0.010	0.007	0.003	0.001	

variances, especially with a very high residual variance both at the beginning and the end of the lactation. Jamrozik and Schaeffer (1997), Kettunen et al. (1998), Samoré et al. (2002) and Strabel and Misztal (1999) estimated highest heritability at both extremes of the lactation curve. Ranges of heritabilities across the lactation varied considerably among studies, ranging from as low as 0.10 (e.g., Strabel and Misztal, 1999) to 0.60 (e.g., Jamrozik and Schaeffer, 1997). The values in Figure 4 seemed moderate to high but not extreme. The range of heritabilities of most multi-trait studies was from 0.20 to 0.35 and higher in some cases. Our results are close to those of Liu et al. (2000) and Jakobsen et al. (2001) and lower than results from Jamrozik and Schaeffer (1997), Kettunen et al. (1998) and Olori et al. (1999) who found heritabilities higher than 0.50 for some parts of the curve.

Genetic correlations between effects at different DIM are presented in Table 3. The genetic effects for the TD yields in the middle of lactation were highly correlated, with correlations higher or equal to 0.90 between the genetic effect at DIM 155 and genetic effects of a large part of the lactation (from 45 to 280 DIM). The genetic effects at the beginning of the lactation were less correlated and seemed to be a somewhat different trait. Genetic correlation was equal to 0.35 for the extreme parts of the lactation, which seemed reasonable. To compare these genetic correlations with estimates from multi-trait analyses, the effect at DIM 20 should be compared to class 1 (the median of the first 30 d class) and the effect at DIM 290 to class 10, between these two extremes all correlations were higher than 0.55. The correlations obtained are in agreement with those found by multi-trait analyses by Meyer et al. (1989), Pander et al. (1992) and Kettunen et al. (1998) and with estimated correlations obtained from random regression analysis by Brotherstone et al. (2000) and Olori et al. (1999). Rekaya et al. (1999) with multi-trait analysis and White et al. (1999) found even larger genetic correlations while Strabel and Misztal (1999) found lower correlations.

Finally, Jamrozik and Schaeffer (1997), Kettunen et al. (1998) or Rekaya et al. (1999) obtained negative correlations for extreme parts of the lactation when working with random regressions.

Differences between studies might be explained by the modeling of the fixed part of the lactation curve. In the present study, effect of age at calving and month of calving are not constant but can vary with the lactation stage. The gestation effect might have an important impact on the end of the lactation, creating large variations between pregnant and non-pregnant cows. In some studies, this effect was ignored, which might cause a larger increase of the residual variance at the end of the lactation.

Differences might also be explained by the size of the data sets. Indeed, in preliminary studies, we found that there was a large variation (in magnitude and shape) among genetic parameters obtained between our 10 distinct samples, each containing 80,000 TD records, for milk production in first lactation. Working simultaneously on these ten samples certainly improved our estimation of the genetic parameters and made it more reliable. This was confirmed by comparison of standard deviations of coefficients of random regressions obtained with one single sample and the pooling approach. The standard deviations of (co)variances were 1.16 to 3.19, 2.54 to 3.19 and 2.87 to 3.25 times larger with the single sample for the genetic, permanent environmental and herd-year variances, respectively. For the genetic effects, the standard deviations of the variances of the first three coefficients of the Legendre polynomial were more than three times larger with the single sample. For the genetic correlations presented in Table 3, the standard deviations obtained with the single sample were 1.25 to 2.84 times larger than with the pooling method. Smaller differences were noted for very close DIM with high correlation at the end of the lactation.

Regarding the selection strategy, if we give equal weight to each DIM in the objective, the estimated genetic parameters would result in an emphasis on test-

Figure 5. Coordinates of eigenvectors of the genetic covariance matrix: — first eigenvector, - - - second eigenvector.

day records in mid-lactation, after the peak. Early TD records would be given lower weight. This is especially useful for animals with production dropping in mid or late lactation.

In order to reduce the number of parameters and genetic values to be estimated with a fifth order Legendre polynomial, the eigenvectors of the obtained covariance matrices were calculated. In preliminary studies, it was found that the first three eigenvectors for both the genetic and permanent environmental parts were very similar when estimated under an unstructured model (with 10 by 10 unstructured covariance matrices, that corresponds to a multi-trait model) as well as under a fifth order random regression model on the 10 tests. The eigenvectors estimated in this study should be close to the ones we would have obtained with other methods.

Analysis of these eigenvalues and eigenvectors confirmed some previous studies (Van der Werf et al., 1998; Olori et al., 1999; Pool et al., 2000). For the genetic covariance matrix, the two first eigenvalues represented more than 98% of the total variation (91.6 and 6.6%, respectively). The associated eigenvectors (see Figure 5) represented approximately a constant term and a term varying linearly throughout the lactation. These terms seemed to make sense biologically as the first eigenvector might represent the average lactation potential of an animal and the second would be its persistency. For the permanent environmental effect, three eigenvectors were necessary to explain more than 95% of the total variation (78.3, 12.5, and 5.3%, respectively).

Because these eigenvectors seem to make sense biologically, it is appealing to use them in the routine evaluation. The obtained breeding values would be independent and interpretable. In addition, use of the

eigenvectors as proposed by Van der Werf et al. (1998) would result in fewer genetic parameters, fewer equations in the MME, diagonal variance matrices (as the eigentraits are independent) and better convergence properties for the solution of large MME. If such a strategy is implemented, the coefficients of the Legendre polynomial in the random regression would be replaced by the coefficients specified by the eigenvectors in Figure 5 and equation [3] would be changed to:

$$y_{ijkln} = \text{HTD}_i + f_1(\text{age}_j, \text{dim}) + f_1(\text{month}_k, \text{dim}) + f_2(\text{dcc}) + \sum_{m=1}^2 a_{lm} \nu_{(m,t)} + \sum_{m=1}^3 p_{lm} \omega_{(m,t)} + \sum_{m=0}^3 h_{ym} \varphi_{(m,t)} + e_{ijkln} \quad [20]$$

where $\nu_{(m,t)}$ and $\omega_{(m,t)}$ are the value of the m^{th} eigenvector of the genetic and permanent environment covariance matrix at DIM equal to t . For instance, $\nu_{(1,100)}$ and $\nu_{(2,200)}$ can be obtained from the information used to plot Figure 5 and would be equal to 0.0617 and 0.0293, respectively.

CONCLUSIONS

The use of regression splines for modeling the fixed effects seemed appealing because their properties are a compromise between those of fixed classes and parametric curves. Regression splines require a limited number of parameters, have good flexibility, are smooth, and have limited sensitivity to the data. Furthermore, they can be relatively easily implemented in the MME.

Tools were developed to work with relatively large data sets and to model a continuous heterogeneity for the residual variance, ensuring better estimates of genetic parameters.

Models using eigenvectors seemed appealing because they can reduce the computational difficulty of the model and improve its convergence properties.

ACKNOWLEDGMENTS

Tom Druet, who is Chargé de Recherches of the National Fund for Scientific Research (Brussels, Belgium), acknowledges his financial support. We are grateful to Anne Barbat, Bernard Bonaiti, Jean-Jacques Colleau, and Nicolas Gengler for valuable discussions. Ignacy Misztal and Shogo Tsuruta are gratefully acknowledged for sharing their programs. Finally, we express special thanks to Ian White, Robin Thompson, and Arthur Gilmour for discussions on splines.

REFERENCES

- Ali, T. E., and L. R. Schaeffer. 1987. Accounting for covariances among test day milk yields in dairy cows. *Can. J. Anim. Sci.* 67:637.

- Auvray, B., and N. Gengler. 2002. Feasibility of a Walloon test-day model and study of its potential as tool for selection and management. *INTERBULL Bull.* 29:123–127.
- Babb, J. S. 1986. Pooling maximum likelihood estimates of variance components obtained from subsets of unbalanced data. M. Sc. Thesis, Cornell University, Ithaca, NY.
- Brotherstone, S., I. M. S. White, and K. Meyer. 2000. Genetic modeling of daily milk yield using orthogonal polynomials and parametric curves. *Anim. Sci.* 70:407–415.
- De Roos, A. P. W., A. G. F. Harbers, and G. de Jong. 2002. Herd specific random regression curves in a test-day model for protein yield in dairy cattle. *Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France. CD-ROM. Communication No. 01–05.*
- Ducrocq, V. 1993. Genetic parameters for type traits in the French Holstein breed based on a multiple-trait animal model. *Livest. Prod. Sci.* 36:143–156.
- Gengler, N., and G. Wiggans. 2001. Heterogeneity in (co)variances structures of test-day yields. *INTERBULL Bull.* 27:179–184.
- Gilmour, A. R., B. R. Cullis, S. J. Welham, and R. Thompson. 2000. *ASREML Manual*. New South Wales Dep. Agric., Orange, Australia.
- Green, P. J., and B. W. Silverman. 1994. *Nonparametric regression and generalized linear models*. Chapman & Hall, London.
- Jaffrezic, F., I. M. S. White, R. Thompson, and W. G. Hill. 2000. A link function approach to model heterogeneity of residual variances over time in lactation curve analyses. *J. Dairy Sci.* 83:1089–1093.
- Jakobsen, J. H., P. Madsen, J. Jensen, J. Pedersen, L. G. Christensen, and D. A. Sorensen. 2002. Genetic parameters for milk production and persistency for Danish Holsteins estimated in random regression models using REML. *J. Dairy Sci.* 85:1607–1616.
- Jamrozik, J., and L. R. Schaeffer. 1997. Estimates of genetic parameters for a test day model with random regressions for yield traits of first lactation Holsteins. *J. Dairy Sci.* 80:762–770.
- Jamrozik, J., L. R. Schaeffer, and F. Grignola. 1998. Genetic parameters for production traits and somatic cell score of Canadian Holsteins with multiple trait random regression model. *Proc. 6th World Congr. Genet. Appl. Livest. Prod., Armidale, Australia XXIII:303–306.*
- Jensen, J. 2001. Genetic evaluation of dairy cattle using test-day models. *J. Dairy Sci.* 84:2803–2812.
- Jensen, J., E. A. Mäntysaari, P. Madsen, and R. Thompson. 1996. Residual maximum likelihood estimation of (co)variance components in multivariate mixed linear models using average information. *J. of the Indian Society of Agricultural Statistics.* 215–236.
- Kettunen, A., E. A. Mäntysaari, I. Strandén, J. Pöso, and M. Lidauer. 1998. Estimation of genetic parameters for first lactation test day milk production using random regression models. *Proc. 6th World Congr. Genet. Appl. Livest. Prod., Armidale, Australia XXIII:307–310.*
- Kirkpatrick, M., D. Lofsvold, and M. Bulmer. 1990. Analysis of inheritance, selection and evolution of growth trajectories. *Genetics* 124:979–993.
- Liu, Z., F. Reinhardt, and R. Reents. 2000. Estimating parameters of a random regression test day model for first three lactation milk production traits using the covariance function approach. *INTERBULL Bull.* 25:74–80.
- Mayeres, P. 2002. Appui technique à la gestion des troupeaux de bovins laitiers à travers la mise en évidence d'influences spécifiques au niveau troupeaux et des vaches sur les performances et l'efficacité biologique. Master thesis, Gembloux Agricultural University, Gembloux, Belgium.
- Meyer, K., H.-U. Graser, and K. Hammond. 1989. Estimates of genetic parameters for first lactation test day production of Australian Black and White cows. *Livest. Prod. Sci.* 21:177–199.
- Meyer, K. 1998. Estimating covariance functions for longitudinal data using a random regression model. *Genet. Sel. Evol.* 30:221–240.
- Meyer, K. 2001. Estimates of direct and maternal covariance functions for growth of Australian beef calves from birth to weaning. *Genet. Sel. Evol.* 33:487–514.
- Misztal, I., T. Strabel, J. Jamrozik, E. A. Mäntysaari, and T. H. E. Meuwissen. 2000. Strategies for estimating the parameters needed for different test-day models. *J. Dairy Sci.* 83:1125–1134.
- Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. H. Lee. 2002. BLUF90 and related programs (BGF90). *Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France. CD-ROM. Communication No. 28–07.*
- Oehlert, G. W. 1992. A note on the delta method. *American Statistician* 46:27–29.
- Olori, V. E., S. Brotherstone, W. G. Hill, B. J. McGuirk. 1997. Effect of gestation stage on milk yield and composition in Holstein Friesian dairy cattle. *Livest. Prod. Sci.* 52:167–176.
- Olori, V. E., W. G. Hill, B. J. McGuirk, and S. Brotherstone. 1999. Estimating variance components for test day milk records by restricted maximum likelihood with a random regression animal model. *Livest. Prod. Sci.* 61:53–63.
- Pander, B. L., W. G. Hill, and R. Thompson. 1992. Genetic parameters of test day records of British Holstein-Friesian heifers. *Anim. Prod.* 53:11–21.
- Pool, M. H., and T. H. E. Meuwissen. 2000. Reduction of the number of parameters needed for a polynomial random regression test day model. *Livest. Prod. Sci.* 64:133–145.
- Pool, M. H., L. L. G. Janss, and T. H. E. Meuwissen. 2000. Genetic parameters of Legendre polynomials for first parity lactation curves. *J. Dairy Sci.* 83:2640–2649.
- Rekaya, R., M. J. Carabaño, and M. A. Toro. 1999. Use of test day yields for the genetic evaluation of production traits in Holstein-Friesian cattle. *Livest. Prod. Sci.* 57:203–217.
- Samoré, A. B., P. Boettcher, J. Jamrozik, A. Bagnato, and A. F. Groen. 2002. Genetic parameters for production traits and somatic cell scores estimated with a multiple trait random regression model in Italian Holsteins. *Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France. CD-ROM. Communication No. 01–07.*
- Smith, J. W., and J. E. Legates. 1962. Relation of days open and days dry to lactation milk and fat yields. *J. Dairy Sci.* 45:1192–1198.
- Strabel, T., and I. Misztal. 1999. Genetic parameters for first and second lactation milk yields of Polish Black and White cattle with random regression test-day models. *J. Dairy Sci.* 82:2805–2810.
- Swalve, H. H. 1995. The effect of test day models on the estimation of genetic parameters and breeding values for dairy yield traits. *J. Dairy Sci.* 78:929–938.
- Van der Werf, J. H. J., M. E. Goddard, and K. Meyer. 1998. The use of covariance functions and random regressions for genetic evaluation of milk production based on test day records. *J. Dairy Sci.* 81:3300–3308.
- White, I. M. S., R. Thompson, and S. Brotherstone. 1999. Genetic and environmental smoothing of lactation curves with cubic splines. *J. Dairy Sci.* 82:632–638.
- Wilmink, J. B. M. 1987. Adjustment of test-day milk, fat, and protein yields for age season and stage of lactation. *Livest. Prod. Sci.* 16:335–348.
- Yerex, R. P. 1988. Pooling restricted maximum likelihood estimates from data subsets under the animal model. Ph.D. thesis, Cornell University, Ithaca, NY.