PAPER

Comparison of parametric, orthogonal, and spline functions to model individual lactation curves for milk yield in Canadian Holsteins

Nicolò P.P. Macciotta,[1] Filippo Miglior,[1,2] Corrado Dimauro, [1] Larry R. Schaeffer[3]

[1]Dipartimento di Scienze Zootecniche, Universita' di Sassari, Italy
[2]Agriculture and Agri-Food Canada, Dairy and Swine Research and Development Center, Sherbrooke, QC, Canada
[3]Centre for the Genetic Improvement of Livestock, University of Guelph, ON, Canada

Abstract

Test day records for milk yield of 57,390 first lactation Canadian Holsteins were analyzed with a linear model that included the fixed effects of herd-test date and days in milk (DIM) interval nested within age and calving season. Residuals from this model were analyzed as a new variable and fitted with a five parameter model, fourth-order Legendre polynomials, with linear, quadratic and cubic spline models with three knots. The fit of the models was rather poor, with about 30%-40% of the curves showing an adjusted R-square lower than 0.20 across all models. Results underline a great difficulty in modelling individual deviations around the mean curve for milk yield. However, the Ali and Schaeffer (5 parameter) model and the fourth-order Legendre polynomials were able to detect two basic shapes of individual deviations among the mean curve. Quadratic and, especially, cubic spline functions had better fitting performances but a poor predictive ability due to their great flexibility that results in an abrupt change of the estimated curve when data are missing. Parametric and orthogonal polynomials seem to be robust and affordable under this standpoint.

Introduction

Random regression models (RRM) are currently used in the prediction of breeding values and in the estimation of variance components for milk production traits of dairy cattle in several countries. Direct modeling of test day (TD) records instead of 305-d yields allows the shape of the lactation curve to be modelled with subsequently more precise adjustment for temporary environmental effects, avoidance of extended records for culled cows or lactations in progress, and evaluation of lactation persistency (Jamrozik and Schaeffer, 1997).

In the basic structure of a RRM, the fixed part includes effects peculiar to all cows on the same test day and effects specific to cows on a given test day, such as pregnant or diseased, plus a factor accounting for the yield level on a specific day in milk (Ptak and Schaeffer, 1993) whereas individual lactation curves are fitted by random regression coefficients (Jamrozik and Schaeffer, 1997; Schaeffer and Dekkers, 1994). This feature of RRM allows for the prediction of breeding values and the estimation of (co)variance functions throughout the whole lactation.

Mean lactation curves are usually estimated on a large number of records and are characterized by quite regular patterns. As a consequence the use of either mathematical functions or fixed intervals of days in milk will generally lead to the same results (Schaeffer, 2004). On the contrary, the small number of records and a high sensitivity to outliers makes the choice of the function used to fit individual curves rather cumbersome. The evaluation of functions to fit individual effects (genetic and permanent environmental) is usually based on: i) fit diagnostics such as Akaike's or Bayesian Information Criterion (Lopez-Romero and Carabano, 2003; Liu *et al*., 2006); ii) predictive ability of the model (Pool and Meuwissen, 2000); and iii) mathematical features such as correlations among parameters, and scale of parameters (Misztal, 2006). Parametric functions specifically conceived to model lactation curves, such as the Wilmink (WIL) (1987) or the Ali and Schaeffer (AS) (1987) models, have been generally abandoned because estimated (co)variance matrices usually show very high correlations between random regression coefficients which can hinder the estimation process (Schaeffer, 2004). Legendre orthogonal polynomials (LP) (Kirkpatrick *et al*., 1990) are used to model a variety of curves for variances and covariances, and several papers have reported their advantages in comparison with more traditional models (Pool and Meuwissen, 2000; Lopez-Romero and Carabano, 2003; Odegard *et al*., 2003; Strabel *et al*., 2005). However, LP usually yield very large estimates of variances at the beginning and at the end of lactation that tend to increase with the order of the polynomials (Lopez-Romero and Carabano, 2003). This seems to be a characteristic of polynomial covariates, which has also been observed in modeling growth curves for beef cattle (Meyer, 2005). Recently, cubic splines (CSPL) have been proposed to provide extra flexibility in fitting lactation curves (White *et al*., 1999; Silvestre *et al*., 2005; Misztal, 2006). Splines are a type of segmented regression in which the curve is divided into different segments of the dependent variable, joined at points named knots, each fitted with different polynomials (Guo and White, 2005). Problems in using CSPL are the increase of computational complexity and the optimization of the number and placement of knots.

The shift towards more flexible functions able to explain the most variation in the observations is necessarily accompanied by a sensible increase in the number of possible shapes of individual lactation curves detected. As a consequence, it may become difficult to discriminate variations that can be ascribed to genetic or permanent environmental causes, and that have scientific and technical relevance, from random perturbations. Moreover, a high flexibility of a function allows for modeling all the variations among individual patterns but may compromise the predictive power of the model. The evaluation of the ability of a mathematical function to adequately fit the data together with the generation of meaningful results represents a critical point in finding a suitable RRM. In this work, this ability is investigated at the phenotypic level by a

Corresponding author: Prof. Nicolò Pietro Paolo Macciotta, Dipartimento di Scienze Zootecniche, Università di Sassari, via De Nicola 9, 07100 Sassari, Italy.
Tel. +39.079.229299 - Fax: +39.079.229302.
E-mail: macciott@uniss.it

Key words: Test day model, Residuals, Legendre polynomials, Splines.

Acknowledgments: this research was funded by the Ministry of University and Research, Italy (Grant PRIN 2005).

Received for publication: 25 April 2010.
Revision received: 8 November 2010.
Accepted for publication: 11 November 2010.

Italian Journal of Animal Science 2010; 9:e87
doi:10.4081/ijas.2010.e87

fixed regression analysis of individual deviations around the mean curves for milk yield of first lactation Canadian Holsteins using some of the functions proposed to fit random effects in RRM for milk production traits in cattle.

Materials and methods

Data were 496,745 TD records for milk yield of 57,390 first lactation Canadian Holsteins of Ontario. Data were edited based on the number of TD records per cow (>7), calving year (>1997), days in milk (DIM) (between 5 and 305), age at calving (between 22 and 31 months), milk yield (between 1.5 and 90 kg), DIM at first test (<50 d) and number of records within a herd test date (>11). In order to avoid all possible influence on the estimated individual curves that may arise form the function chosen to fit the animal effect in RRM, a two-step approach was followed. Data were analyzed with a linear model including fixed factors recently proposed for Canadian Holsteins (Liu *et al*., 2006):

$$Y = HTD + DIM*SEA*AGE + e \quad [1]$$

where Y = test day milk yield;
HTD = fixed effect of herd test date class (33,977);
DIM*SEA*AGE = fixed effect of DIM interval (60 intervals of 5 days each) nested within calving season (1=April-September, 2=October-March) and age at calving (10 classes from 22 to 31 months) for a total of 1200 levels.

The term DIM*SEA*AGE of model [1] fits the average lactation curve for cows calving at the same age and in the same season (Druet *et al*., 2005; Jamrozik *et al*., 2002).

Residuals of model [1], which can be approximately considered as the result of genetic, permanent environmental and measurement error effects were stored as a new variable (Z). Patterns of Z for each individual cow were fitted separately with the following fixed regression models:

1) Ali and Schaeffer regression (AS) (Ali and Schaeffer, 1987)

$$Z_t = a + b(t/340) + c(t/340)^2 + d[\log(340/t)] + f[\log(340/t)]^2 \quad [2]$$

2) A fourth-order Legendre orthogonal polynomials (LP4) (Kirkpatrick *et al*., 1990)

$$Zt = \alpha_0 * P_0 + \alpha_1 * P_1 + \alpha_2 * P2 + \alpha_3 * P_3 + \alpha_4 * P_4 \quad [3]$$

where functions of time (P_j) were calculated according to Schaeffer (2004).

3) A linear spline regression model (LSPL) (Guo and White, 2005)

$$Z_t = a + b(t) + \sum_{i=1}^{k} k(t - t_i) \quad [4]$$

4) A quadratic spline regression model (QSPL) (Guo and White, 2005)

$$Z_t = a + b(t) + c(t)^2 + \sum_{i=1}^{k} k(t - t_i)^2 \quad [5]$$

5) A cubic spline regression model (CSPL) (Guo and White, 2005)

$$Z_t = a + b(t) + c(t)^2 + d(t)^3 + \sum_{i=1}^{k} k(t - t_i)^3 \quad [6]$$

In all functions, t refers to DIM at the test; a, b, c, d, f and α_i are parameters to be estimated, and k is the number of knots in the splines. AS and LP4 functions have been selected in order to compare differences in model flexibility due to the degree of correlation among parameters. Moreover, cubic splines were compared with splines of lower order. In fact cubic splines are the preferred functions for milk production traits in RRM both for the fixed (Druet *et al*., 2003; 2005) and random (White *et al*., 1999; Silvestre *et al*., 2005) curves. However, lower order splines may offer a good compromise between model complexity and plausibility of results (Meyer, 2005).

A preliminary issue in fitting splines is the number and location of knots. Actually, few guidelines are available. Some authors suggest that the numbers of knots should be as large as possible, especially when splines are used to model fixed average curves (Druet *et al*., 2005; Silvestre *et al*., 2005), and placed at points of maximum concentration of records (Misztal, 2006). Other authors argued that fewer knots tend to yield smoother curves, even though together with a less detailed local fit (Meyer 2005). In the present paper the choice of number and locations of knots was performed using a two-step procedure. First, the average Z curve was fitted with the different types of splines using a non linear estimation procedure where knot positions are considered additional independent variables (Fadel, 2004). The number of knots (k) was set to three for all the three types of splines because larger values (four and five) resulted in convergence problems. Estimated positions were at 32, 64 and 240 DIM respectively and were then used in fixed regressions for modeling individual curves. Also in this case, no substantial differences in adjusted R-squared values obtained by fitting splines with three, four or five knots were observed, probably due to the homogeneous distribution of data across days in milk (on average of 1639 records per day in milk, with a standard deviation of 79 per day in milk).

Goodness of fit was assessed by examining the adjusted R-square (ADJ-R2) and curves were classified according to five levels of ADJ-R2 (<0.20, from 0.20 to 0.40, from 0.40 to 0.60, from 0.60 to 0.80, >0.80). Moreover, different models were compared on the basis of their ability to predict missing records by using the mean square error of predictions of missing observations (MSEP), calculated as

$$MSEPi = \sum_{j=1}^{n} \frac{\left(Z_{ij} - \hat{Z}_{ij}\right)^2}{n}$$

where each Z predicted value ($\hat{Z}_i$) is obtained by deleting the i-th record from all the n lactations of the complete data set, estimating the regression function from the remaining records, and then using the fitted regression function to obtain the predicted value of the omitted records (Neter *et al*., 1996). In the present study, eight subsets of missing records were generated deleting one test per lactation in turn.

Results and discussion

Average values of Z along the lactation tend to remain quite close to zero (Figure 1), indicating an acceptable fit of the mean lactation curves with DIM intervals (Druet *et al*., 2003), apart from a perturbation that occurs within the first 50 days of lactation (that actually cor-

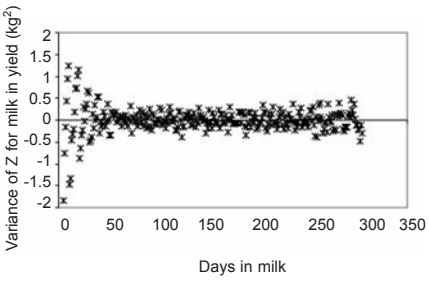

Figure 1. Lactation pattern of individual deviations (Z) around the mean lactation curve values averaged for all animals.

responds to the location of the first record). The variance of Z along the lactation indicated the variability among animal Z curves (Figure 2). Maximum variances were reached around 50 days from parturition (Pool and Meuwissen, 1999; 2000). Differences among shapes were related to the stage of lactation, as evidenced by the continuous trend over time, although the role of random fluctuations between individual cows should not be neglected.

The functions AS and LP4 were theoretically able to describe 25 types of curves on the basis of combination of signs of the estimated parameters (Macciotta *et al*., 2005). About 90% of individual Z patterns were classified by the AS function into two main shapes (Figure 3), characterized by an opposite succession of curvatures across lactation that corresponds to an opposite combination of parameter signs.

The function LP4 resulted in a balanced distribution of curves among classes of different combination of parameter signs (curves were uniformly distributed among all 32 theoretical shapes, with a frequency ranging from 0.01 to 0.05). However, each of these shapes can be regarded as a subclass, i.e. the result of a specific deformation, of the two basic patterns detected by the AS function. Parametric functions and orthogonal polynomials were able to model differences among individual shapes related to the stage of lactation. However, LP4 showed a greater ability to fit random variations that can be ascribed to the lower degree of correlation among parameters in comparison with AS. A more accurate description of random variations was expected when using spline functions, although an *a priori* classification of shapes on the basis of parameter signs can not be realised due to their specific mathematical features.

The fitting of Z was rather poor (Table 1), with about 30-40% of curves showing an ADJ-R2 lower than 0.20 in all the models. Differences between the functions can be highlighted by considering both the distributions of fits among ADJ-R2 classes and the magnitude of standard deviation of residuals of the different models (Table 2). An increase in the number of curves showing an ADJ-R2>0.80 can be observed with the quadratic and cubic splines, with the CSPL having almost double the number in comparison with AS and LP4 (Table 1). There was a parallel reduction in the standard deviation of residuals (Table 2). The specific ability of CSPL to fit patterns characterised by marked oscillations can be clearly observed in Figures 4a and 4b where examples of individual curves fitted and actual data points for Z are reported. In particular, Figure 4a refers to an individual curve that has been well fitted only by the CSPL (ADJ-R2>0.90), whereas Figure 4b shows a curve that has been adequately fitted by all models. CSPL shows a tendency to produce very large estimates of Z at the extremes of the trajectory. In any case when the magnitude of individual deviations becomes rather high, i.e. in more than 40% of cases in this work, the cubic spline was no longer able to fit the pattern, as evidenced by Figure 4c where a curve poorly fitted by all models (ADJ-R2<0.20) is shown.

The comparison between the predictive ability of the different models is summarized in Table 3. Poor results can be observed in the prediction of records at the beginning and at the end of the lactation trajectory, with extremely high values of MSEP for quadratic and cubic splines. In general all three orders of splines showed higher values of MSEP in comparison with the other two functions. Such a marked contrast between fitting performances, higher in splines than in AS and LP4 models, and predictive ability could be explained with the great flexibility of spline functions. These models were able to follow closely individual fluctuations and, therefore, when some data are missing, the shape of the estimated curve could change markedly. An example of the effect of data missing in the

Table 1. Absolute and relative (below, in italics) frequencies of individual Z curve fits among different classes of adjusted R^2.

ADJ-R2 class	Model				
	AS	LP4	LSPL	QSPL	CSPL
<0.20	19,816	20,266	20,438	19,095	17,776
	0.34	*0.35*	*0.37*	*0.33*	*0.31*
0.20-0.40	7786	8011	8049	6653	4907
	0.14	*0.14*	*0.14*	*0.12*	*0.09*
0.40-0.60	9492	9383	9322	8219	6505
	0.16	*0.16*	*0.16*	*0.14*	*0.11*
0.60-0.80	10,912	10,703	10,738	10,437	9,068
	0.20	*0.19*	*0.19*	*0.18*	*0.16*
>0.80	9384	9027	8843	12,986	19,134
	0.16	*0.16*	*0.14*	*0.23*	*0.33*

ADJ-R2, ajusted R-square; AS, Ali and Schaeffer regression; LP4, fourth-order Legendre polynomials; LSPL, linear spline; QSPL, quadratic spline; CSPL, cubic spline.

Table 2. Standard deviation of residuals (kg) of different models used to fit the variable Z.

Variable	Milk
AS residuals	1.70
LP4 residuals	1.71
LSPL residuals	1.72
QSPL residuals	1.46
CSPL residuals	1.17

AS, Ali and Schaeffer regression; LP4, fourth-order Legendre polynomials; LSPL, linear spline; QSPL, quadratic spline; CSPL, cubic spline.

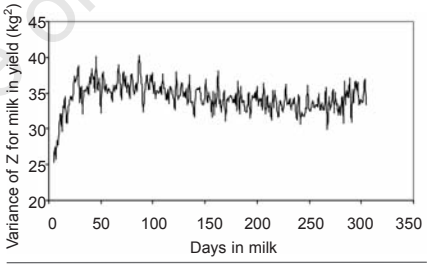

Figure 2. Lactation pattern of variance of individual deviations (Z) around the mean lactation curve.

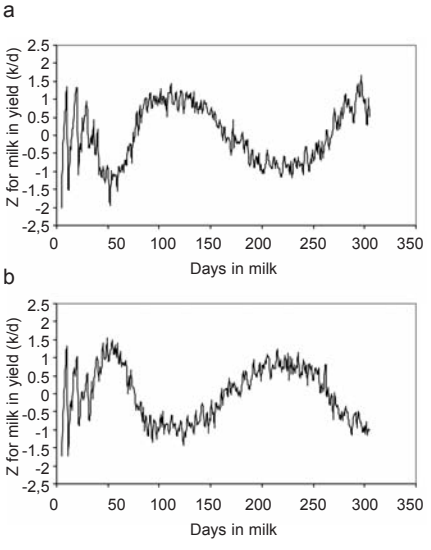

Figure 3. (a) Most frequent average pattern (about 48% of curves) of individual deviations (Z) around the mean lactation curve Z by DIM. (b) Second most frequent average pattern (about 43% of curves) of individual deviations (Z) around the mean lactation curve Z by DIM.

Table 3. Mean square of prediction of missing records for the different models.

Missing record°	Model				
	AS	LP4	LSPL	QSPL	CSPL
1	26968	260	2391	610607	1001305
2	99	19	39	15920	116082
3	14	13	14	20	526
4	12	11	14	14	21
5	10	11	12	13	18
6	11	11	11	13	53
7	11	15	15	67	4904
8	42	154	774	4043371	127418585771

°Record to be predicted within each individual lactation; AS, Ali and Schaeffer regression; LP4, fourth-order Legendre polynomials; LSPL, linear spline; QSPL, quadratic spline; CSPL, cubic spline.

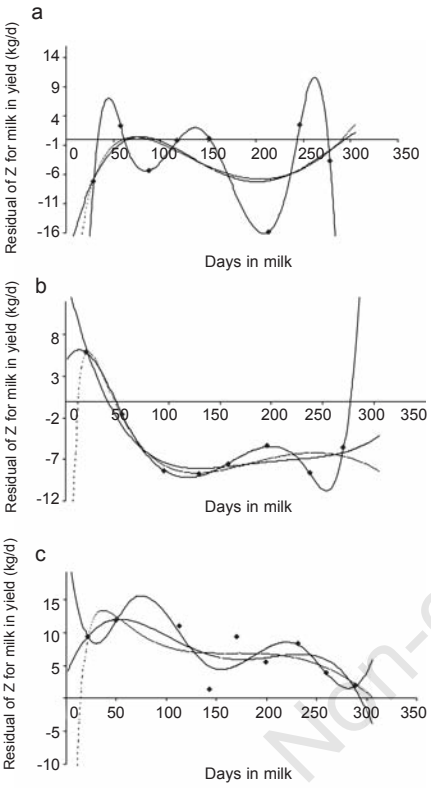

Figure 4. (a) Example of individual curve of deviations (Z) around the mean lactation curve (◆) that showed a good fit only for the cubic spline (------, Ali and Schaeffer model; — - — - —, fourth-order legendre polynomials; ——, cubic spline). (b) Example of individual curve of deviations (Z) around the mean lactation curve (◆) that has been well fitted by all models (------, Ali and Schaeffer model; — - — - —, fourth-order legendre polynomials; ——, cubic spline). (c) Example of individual curve of deviations (Z) around the mean lactation curve (◆) that showed a poor fit (adjusted r-square<0.20) for all the models (------, Ali and Schaeffer model; — - — - —, fourth-order legendre polynomials; ——, cubic spline).

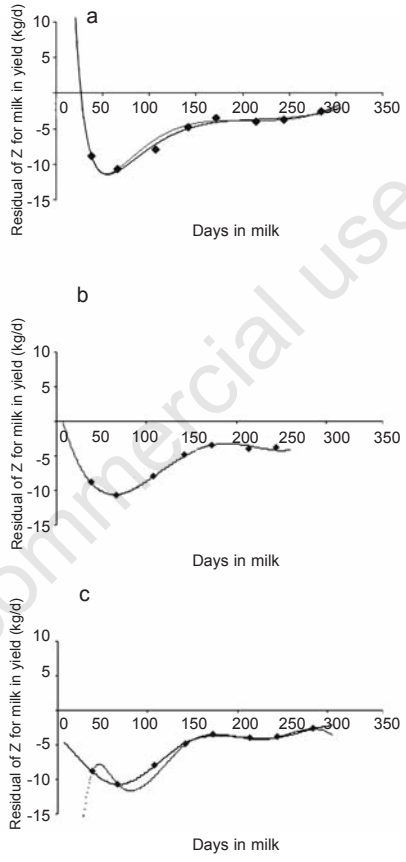

Figure 5. (a) Example of individual curve of deviations (Z) around the mean lactation curve (actual data = ◆) estimated with the Ali and Schaeffer model using all the n records available (——) or using n-1 records (------). (b) Example of individual curve of deviations (Z) around the mean lactation curve (actual data = ◆) estimated with the Legendre fourth order polynomials using all the n records available (——) or using n-1 records (------). (c) Example of individual curve of deviations (Z) around the mean lactation curve (actual data = ◆) estimated with the cubic spline using all the n records available (——) or using n-1 records (------).

shape of an estimated individual curve by AS, LP4 and CSPL is reported in Figure 5. Curves estimated using all eight available records or seven (the third is missing) were very similar in the case of AS and LP4 functions (for LP4 the two curves are almost the same) (Figures 5a and 5b), whereas a marked change of shape occurs for the CSPL (Figure 5c).

Conclusions

Flexible functions are preferred to fit the great variability of individual shapes. On the other hand, the ability to recognise a continuous and somewhat regular pattern is essential for a consistent interpretation of results in terms of genetic and permanent environmental effects. Most previous investigations on RRM have recommended patterns and functions characterised by a high flexibility, such as orthogonal polynomials of high order or cubic splines, as more suitable models.

The present study, developed at the phenotypic level, highlights the difficulty to adequately model individual patterns around the mean curve for milk yield, due to a large variability of shapes among cows. Functions such as the Ali and Schaeffer model and fourth-order Legendre polynomials were able to reduce the wide range of shapes into two basic forms, even if there were differences in the ability to follow random fluctuations. Although the goodness of fit was in general rather poor for all models considered, a certain superiority of the quadratic and, especially, cubic spline functions was observed. These results are in agreement with the suggestions of some authors to use non-parametric regressions as sub-models in RRM. However, the greater fitting performance of cubic splines was offset by a poor predictive ability due to the great flexibility that results in an abrupt change of the estimated curve when some data was missing. Parametric and orthogonal polynomials seem to be more robust and affordable under this criterion.

References

Ali, T.E., Schaeffer, L.R., 1987. Accounting for covariances among test day milk yields in dairy cows. Can. J. Anim. Sci. 67:637-644.

Druet, T., Jaffrezic, F., Boichard, D., Ducroq, V., 2003. Modeling lactation curves and estimation of genetic parameters for first lactation test-day records of French Holstein

cows. J. Dairy Sci. 86:2480-2490.
Druet, T., Jaffrezic, F., Ducroq, V., 2005. Estimation of genetic parameters for thest day records of dairy traits in the first three lactations. Genet. Sel. Evol. 37:257-271.
Fadel, J.G., 2004. Technical note: estimating parameters of nonlinear segmented models. J. Dairy Sci. 87:169-173.
Guo, Q., White, R.E., 2005. Cubic spline regression for the open-circuit potential curves of a lithium-ion battery. J. Electrochem. Soc. 152:A343-A350.
Jamrozik, J., Schaeffer, L.R., 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., Schaeffer, L.R., Weigel, K.A., 2002. Estimates of genetic parameters for single- and multiple-Country test day models. J. Dairy Sci. 85:3131-3141.
Jensen, J., 2001. Genetic evaluation of dairy cattle using test day models. J. Dairy Sci. 84:2803-2812.
Kirkpatrick, M., Lofsvold, D., Bulmer, M., 1990. Analysis of inheritance, selection and evolution of growth trajectories. Genetics 124: 979-993.
Liu, Y.X., Zhang, J., Schaeffer, L.R., Yang, R.Q., Zhang, W.L., 2006. Optimal random regression models for milk production in dairy cattle. J. Dairy Sci. 89:2233-2235.
Lopez-Romero, P., Carabano, M.J., 2003. Comparing alternative random regression models to analyse first lactation daily milk yield data in Holstein Friesian cattle. Livest. Prod. Sci. 82:81-98.
Macciotta, N.P.P., Vicario D., Cappio-Borlino A., 2005. Detection of different shapes of lactation curve for milk yield in dairy cattle by empirical mathematical models. J. Dairy Sci. 88:1178-1191.
Meyer, K., 2005. Random regression analyses using B-splines to model growth of Australian Angus cattle. Genet. Sel. Evol. 37:473-500.
Misztal, I., 2006. Properties of random regression models using linear splines. J. Anim. Breed. Genet. 123:74-80.
Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1996. Applied Linear Statistical Models. McGraw-Hill Co., Inc., Chicago, IL, USA.
Odegard, J., Jensen, J., Klemetsdal, G., Madsen, P., Heringstad, B., 2003. Genetic analysis of somatic cell score in Norwegian cattle using random regression test day models. J. Dairy Sci. 86:4103-4114.
Pool, M.H., Meuwissen, T.H.E., 1999. Prediction of daily milk yields from a limited number of test days using test day models. J. Dairy Sci. 82:1555-1564.
Pool, M.H., Meuwissen, T.H.E., 2000. Reduction of the number of parameters needed for a polynomial random regression test day model. Livest. Prod. Sci. 64:133-145.
Ptak, E., Schaeffer, L.R., 1993. Use of test day yields for genetic evaluations of dairy sires and cows. Livest. Prod. Sci. 34:23-34.
Schaeffer, L.R., 2004. Applications of Random Regression models in animal breeding. Livest. Prod. Sci. 86:35-45.
Schaeffer, L.R., Dekkers, J.C.M., 1994. Random regressions in animal models for test day production in dairy cattle. Proc 5th World Congr. Genet. Appl. Livest. Prod., Guelph, ON, Canada, 18:443-446.
Silvestre, A.M., Petim-Batista, M.F., Colaco, J., 2005. Genetic parameter estimates for milk, fat and protein using a spline test day model. J. Dairy Sci. 88:1225-1230.
Strabel, T., Szyda, J., Ptak, E., Jamrozik, J., 2005. Comparison of random regression test-day models for Polish black and white cattle. J. Dairy Sci. 88:3688-3699.
White, I.M.S., Thompson, R., Brotherstone, S., 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 yield for age, season and stage of lactation. Livest. Prod. Sci. 16:335-348.