



Random regression models using Legendre orthogonal polynomials to evaluate the milk production of Alpine goats

F.G. Silva¹, R.A. Torres¹, L.F. Brito¹, R.F. Euclides¹, A.L.P. Melo¹,
N.O. Souza¹, J.I. Ribeiro Jr.² and M.T. Rodrigues¹

¹Departamento de Zootecnia, Universidade Federal de Viçosa,
Viçosa, MG, Brasil

²Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, MG, Brasil

Corresponding author: F.G. Silva

E-mail: felipe.melhoramento@hotmail.com

Genet. Mol. Res. 12 (4): 6502-6511 (2013)

Received August 9, 2013

Accepted September 20, 2013

Published December 11, 2013

DOI <http://dx.doi.org/10.4238/2013.December.11.1>

ABSTRACT. The objective of this study was to identify the best random regression model using Legendre orthogonal polynomials to evaluate Alpine goats genetically and to estimate the parameters for test day milk yield. On the test day, we analyzed 20,710 records of milk yield of 667 goats from the Goat Sector of the Universidade Federal de Viçosa. The evaluated models had combinations of distinct fitting orders for polynomials (2-5), random genetic (1-7), and permanent environmental (1-7) fixed curves and a number of classes for residual variance (2, 4, 5, and 6). WOMBAT software was used for all genetic analyses. A random regression model using the best Legendre orthogonal polynomial for genetic evaluation of milk yield on the test day of Alpine goats considered a fixed curve of order 4, curve of genetic additive effects of order 2, curve of permanent environmental effects of order 7, and a minimum of 5 classes of residual variance because it was the most economical model among those that were equivalent to the complete model by the likelihood ratio test. Phenotypic variance and heritability were higher at the end of the lactation period, indicating that the length

of lactation has more genetic components in relation to the production peak and persistence. It is very important that the evaluation utilizes the best combination of fixed, genetic additive and permanent environmental regressions, and number of classes of heterogeneous residual variance for genetic evaluation using random regression models, thereby enhancing the precision and accuracy of the estimates of parameters and prediction of genetic values.

Key words: Genetic grouping; Heterogeneity of variance; Model selection

INTRODUCTION

Techniques for the selection of milk goats have been in constant progress. Animal selection was traditionally based on morphological aspects that had little relation to milk production until more organized producers started using the accumulated milk production (AMP) trait under individual selection methodology, which consists of a trait estimated using the milk yield on the test day (MYTD).

Due to the use of mixed models (Henderson, 1963) and the improvement of computer resources, selection through best linear unbiased prediction (BLUP) brought great genetic progress to breeding programs in the last few decades. Repeatability and multitrait models brought advantages as they can consider more than one lactation period in the same analysis.

Henderson Jr. (1982) showed that the use of random regression models (RRMs) associated to the methodology of mixed models was adequate to evaluate longitudinal data, making it possible to select milk goats using MYTD instead of AMP, which brings additional advantages for data modeling besides not requiring an estimation of the trait to be evaluated. According to Araújo et al. (2006), Ptack and Schaeffer (1993) were the first to use RRM in genetic evaluations of MYTD using data from the Quebec Dairy Herd Analysis Service.

Many studies have been performed to define the best RRM to evaluate animals of different breeds in different herds to estimate genetic parameters, and to study the behavior of milk production curves. Menezes et al. (2008a), working with Saanen goats, tested different adaptations of the Wilmink model, while Menezes et al. (2010) compared the two best models using the Wilmink function with four models using Legendre orthogonal polynomials (LOP). The authors found that the best model, in relation to all of the criteria evaluated, was the one that used LOP with a third order for the fixed curve, a fourth order for the curve of additive genetic variance, and a sixth order for the permanent environmental curve in addition to considering six classes of residual variance.

Freitas et al. (2008) used LOP, the Wilmink function, and the Ali and Schaeffer parametric functions for the MYTD of Guzerá cows. Of the models evaluated, the worst adjustment was given by the Wilmink function, while the best one was given by the Ali and Schaeffer function. However, the authors reported some difficulties of convergence in their analysis using this function and they recommend the use of fourth order LOP because they present the second best adjustment, according to Bayesian's information criterion (BIC), and because they do not result in convergence difficulties.

The objective of this study was to define the best RRM using LOP to evaluate Alpine goats genetically and to estimate parameters for the MYTD.

MATERIAL AND METHODS

MYTD data from the Goat Sector of Universidade Federal de Viçosa (Viçosa, MG, Brazil) were used. After editing and checking the database for errors, 20,710 records of MYTD for the 667 first lactations of Alpine goats remained in the experiment. The animals were raised in a free stall system and fed diets based on corn and hay as roughage and a concentrated mixture provided according to the nutritional needs of the animals.

Milk controls were carried out in a milking machine twice a day on a weekly basis. The morning collection started at 6:00 am and the afternoon collection started at 2:00 pm. Milk production on the test day was the sum of the milk collected in these two periods.

The tested models had combinations of distinct adjustment orders for polynomials of fixed (F_—), random genetic (A_—), and permanent environmental (EP_—) curves, besides a number of classes for residual variance (H_—). For example, a model with order 4 fixed regression, order 3 additive genetic regression, and order 6 permanent environmental regression, besides considering a heterogeneous residual variance with five classes, can be summarized by the code F4A3EP6H5. The one-trait animal model for random regression is described in a generalized manner as follows:

$$y_{ij} = EF_i + \sum_{m=0}^{k_b-1} b_m \Phi_m(t_{ij}) + \sum_{m=0}^{k_a-1} \alpha_{im} \Phi_m(t_{ij}) + \sum_{m=0}^{k_y-1} \gamma_{im} \Phi_m(t_{ij}) + \varepsilon_{ij}$$

in which, y_{ij} is the milk production on the j control of the i^{th} goat; EF_i is the fixed effect set, which consisted of genetic grouping (Alpine > 90% = 1; 90% ≥ Alpine > 80% = 2; 80% ≥ Alpine > 70% = 3; 70% ≥ Alpine > 60% = 4; and 60% ≥ Alpine > 50% = 5), year-season, and type of kidding. The age of the goat at kidding was included as a covariable with linear and quadratic effects; b_m is the m^{th} regression coefficient of the MYTD on the Legendre polynomial for modeling the average curve of the population; α_{im} and γ_{im} are the m^{th} regression coefficient of the additive genetic and permanent environmental regression, respectively, for the i^{th} goat; k_b , k_a , and k_y are the degree of LOP; t_{ij} is the control variable, lactation week of the i^{th} goat standardized for the -1 to 1 interval, as described by Kirkpatrick et al. (1990); $\Phi_m(t_{ij})$ is the Legendre polynomial function for parameter m evaluated for age t_{ij} ; and ε_{ij} is the specific effect for each observation that was not explained by the correction factors or the regression of the model.

Criteria for choosing models in this study were the following: logarithm of the restricted maximum likelihood function (LogL), modified Akaike's Information Criterion (AICm), modified BIC (BICm), likelihood ratio test (LRT), convergence through the AI-REML algorithm, and condition number (CN) lower than 100 for additive genetic and permanent environmental regressions. The other criteria and a combinatory index of the criteria may be seen in Liu et al. (2006).

AIC can be calculated as follows: $AIC = -2\text{LogL} + 2p$, in which p is the number of parameters in the model. More details can be found in Akaike (1973). BIC can be calculated as follows: $BIC = -2\ln L + p \ln(N - r)$, in which p is the number of parameters in the model, N shows the total amount of information, and r is the matrix X rank (i.e., the matrix of the incidence of the fixed effects). More details can be seen in Schwarz (1978). The LRT can be calculated as follows: $\text{LRT} = -2(\ln L_1 - \ln L_2)$. More details can be seen in Rao (1973). The AICs and BICs presented in this study were modified (AICm and BICm, respectively) by multiplying them by the term (-1/2) because this makes these criteria on a similar scale to LogL and also ensures that their

greatest values correspond to the best model, which makes it easier to compare the models and the behavior of the criteria following alterations to the models. The number of conditions may be obtained in the following manner: $NC = \lambda_{max} / \lambda_{min}$, which is the division of the greatest eigenvalue by the lowest one associated with the estimated parameters. High results indicate problems of multicollinearity. More details can be seen in Montgomery and Peck (1981).

WOMBAT software (Meyer, 2009) was used in all genetic analyses, giving estimates of the parameters and values needed to calculate the criteria for selecting the models.

Liu et al. (2006) described residues as non-correlated effects among and within individuals. This implies that such effects are dependent on a measurement and not influenced by adjacent measurements, but there are effects that may influence the residual variance, being more commonly related to the scale effect.

However, the scale effect is not the only factor; the phenotypic variance of the trait can be used because it is expected that the phenotype is a good representative of residual variation. In addition to being influenced by scale effects, residual variance is relatively higher than additive genetic variance in economically interesting traits in husbandry animals. This concept was used to organize residual variance into different classes, which are shown in Figure 1.

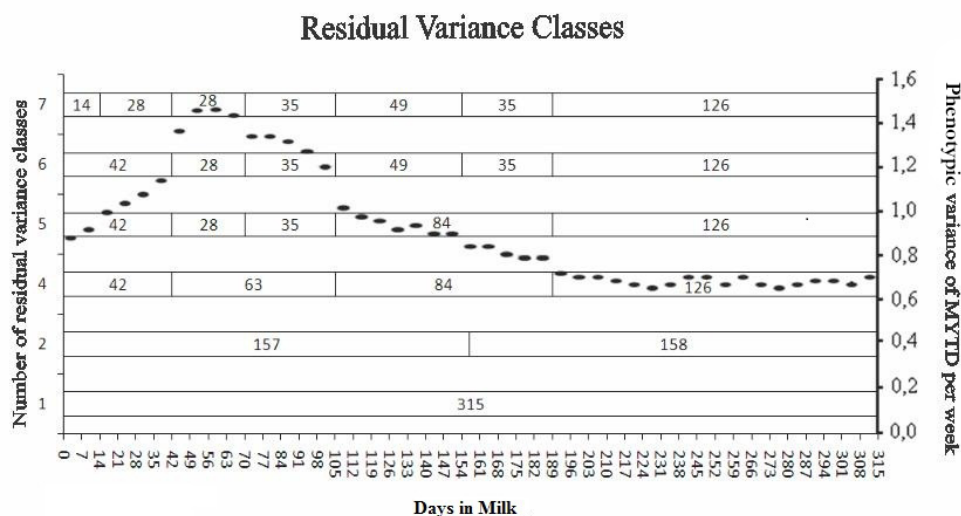


Figure 1. Residual variance divided into 2, 4, 5, 6, and 7 classes, using phenotypic variance.

RESULTS AND DISCUSSION

The results for LogL, AICm, and BICm in the function of the number of classes of the residual variance and polynomial degree for the permanent environmental, additive genetic effects, and fixed curve are expressed in Figures 2, 3, and 4, respectively. The figures show that alterations in the orders of regression for the permanent environmental and genetic parts resulted in greater increments of LogL, AICm, and BICm. In addition, the increments become progressively smaller as the order increases, whereas the increments in the fixed curve and number of classes of the residual variance showed smaller increases in the used criteria.

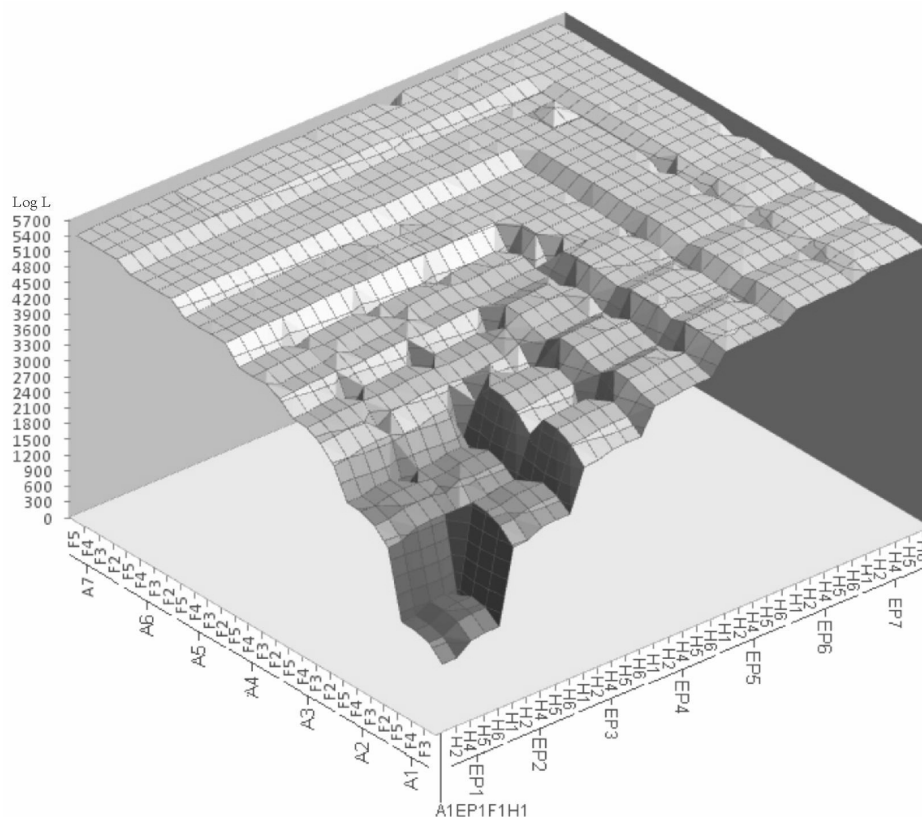


Figure 2. Results of maximum likelihood function in relation to number of classes of residual variance (H) and polynomial degrees of additive genetic (A), permanent environment (PE) and fixed (F) regression effects.

The Figures also show that there were negative interaction effects among the orders of adjustments for the genetic and permanent environmental regression, which may be linked to the difficulty in separating genetic variance from permanent environmental variance. This task may be complicated in as much as the genetic effect of the milk production curve of an animal is estimated not only based on information from their parents but also by considering information from the animal itself.

Overall, models with a greater order of regression for the genetic effects in relation to the regression order of permanent environmental effects presented higher heritability estimates over the production curve, which does not show that those models are better fitted because the estimates may be wrong, i.e., not correcting correctly for environment variations on the individual and attributing them as genetic effects.

Models with orders of higher and nearer genetic and permanent environmental regressions converged more difficultly, and the more parameterized models could only reach convergence through the parameter expanded-expectation maximization (PX-EM) algorithm proposed by Meyer (2006).

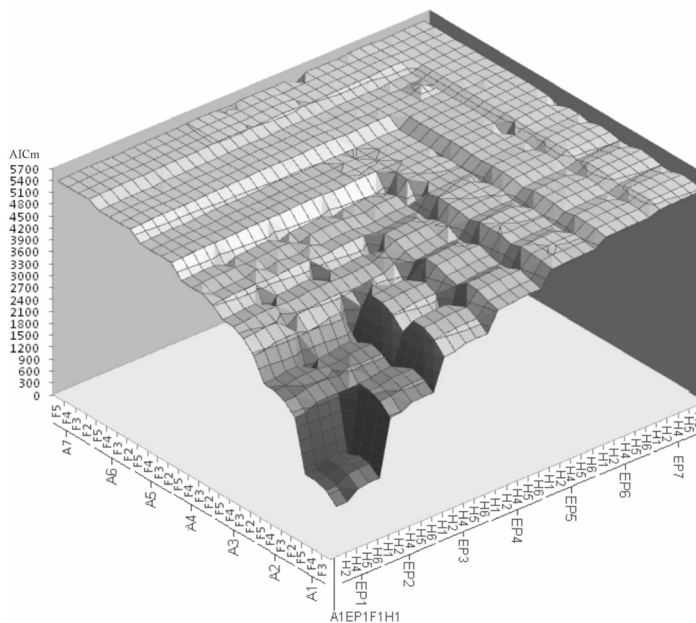


Figure 3. Results of modified Akaike's information criterion ($AIC_m = -1/2 \cdot AIC$) in function of the number of classes of residual variance (H) and polynomial degrees for additive genetic (A), permanent environment (EP) and fixed (F) regression effects.

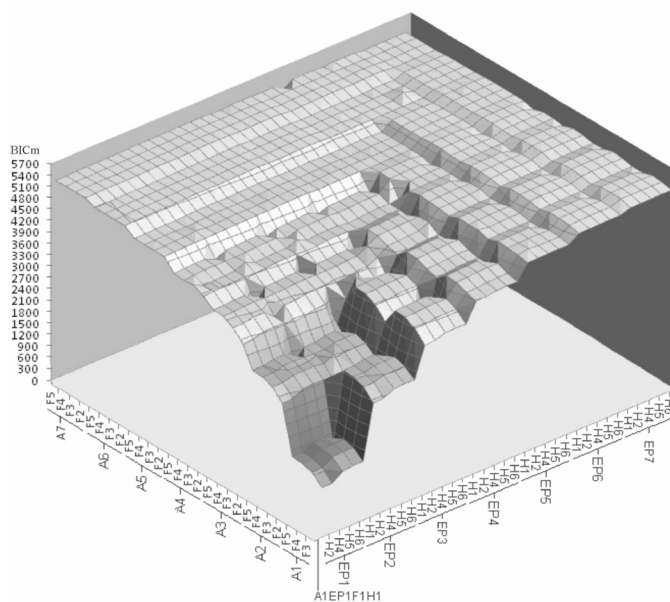


Figure 4. Results of modified Bayesian's information criterion ($BIC_m = -1/2 \cdot BIC$) in function of the number of classes of residual variance (H) and polynomial degrees for additive genetic (A), permanent environmental (PE) and fixed (F) regression effects.

According to Foulley and van Dyk (2000), in RRM, the expectation-maximization algorithm proposed by Dempster et al. (1977) may converge on different results according to the start values used; however, PX-EM algorithms are more reliable and also faster in analyses using RRM.

According to Misztal (2008), in several models, convergence using the average information (AI-REML) algorithm is reached only in some turns; however, poor starting values may make convergence impractical or generate inaccurate estimates, even from within the parametric space. Thus, analyses that converged through the PX-EM algorithm and used estimates as start values for the AI-REML algorithm, but did not converge (F5A7EP7H6), may indicate that the estimates were inaccurate.

Figure 5 shows the results of the LRT ($P < 0.01$) together with an indication of the models that were statistically equal to the complete model (F5A7EP7H6) with their respective parameter numbers. According to the results of the test, 25 of the 980 evaluated models were statistically equal to the complete model. Of those 25 models, we recommended the use of F4A2EP7H5 because it was simpler and converged easily using the AI-REML algorithm without requiring much in relation to the quality of the start values.

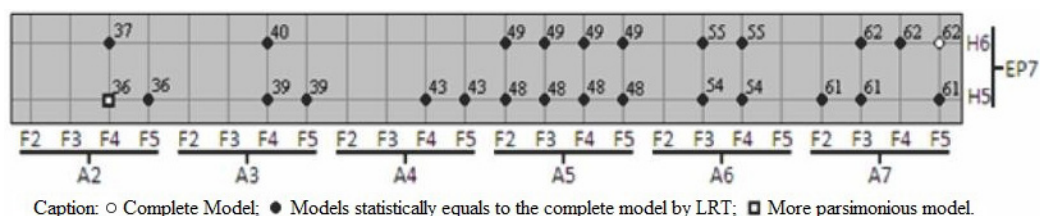


Figure 5. Results of likelihood ratio test, showing the complete model, the models statistically equal to the complete model and among them, distinctly represented from the others, the most economical model. The models tested had combinations of distinct adjustment orders for polynomials of fixed (F_{_}), random genetic (A_{_}), and permanent environmental (EP_{_}) curves, besides a number of classes for residual variance (H_{_}).

The results of the present study diverge from those collected from 1997 to 2004 by Sarmiento et al. (2006) who studied Alpine goats from the same herd. The authors indicated that the models using the fifth order for additive genetic effects and the seventh order for permanent environmental effects were the best ones. Nevertheless, they suggested that the use of a less parameterized model of the fourth order for additive genetic effects and of the sixth order for permanent environmental effects would present similar results.

Figure 6 presents the phenotypic, permanent environment, residual, and additive genetic variances with an effect on lactation days for the F4A2EP7H5 model. The results show greater permanent environmental variance close to the turning point of the lactation curve instead of close to the peak, which would be expected from the scale effect. These results show that the period and intensity of the decline of the lactation curve are very susceptible to environmental conditions. Conversely, additive genetic variance was small up to the turning point of the lactation curve. These results are in agreement with those from Menezes et al. (2008b), who evaluated different lactation measures of persistence for Saanen goats using RRM. They found low heritability for all of the evaluated measures.

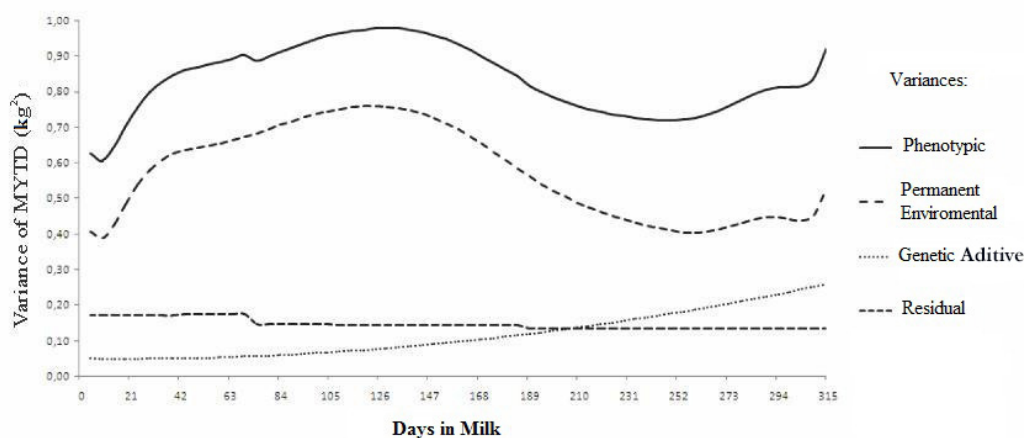


Figure 6. Permanent environmental, additive genetic and residual phenotypic variances in function of lactation days for F4A2EP7H5 model.

Additive genetic variance increased as the number of lactation days increased, i.e., small at the beginning of the lactation period and intermediate at the end, indicating the presence of a greater genetic variability with respect to lactation length than to production. Heritability presented with similar behavior to additive genetic variance, i.e., greater in the final third of lactation, which is in contrast to the results found by Menezes et al. (2010) who found higher values of additive genetic variance and heritability at the beginning of the lactation curve and greater values for the permanent environment at the end of lactation using an F3A4EP6H6 model.

Sarmento et al. (2008) reported higher heritability estimates in the middle of the lactation curve. According to them, the animals were more susceptible to environmental changes in the tests at the beginning and end of the lactation period. Moreover, they found that permanent environmental effect modeling may interfere with the modeling of the additive genetic effect. In addition, estimates of these effects conflict with each other because although genetic value is also influenced by information from the parents, more expressive information is collected from the individual animal. Such information is also used to estimate the permanent environmental effect. This causes multicollinearity problems, with higher degrees for genetic regression; they also increase multicollinearity among the parameters for permanent environment regression and vice versa.

Overall, it was possible to see that, among the additive genetic curves and permanent environmental curves, the use of higher degrees for one in relation to the other also results in relatively higher effects, and the use of the same degree for both curves makes convergence of the analyses more difficult and is negatively reflected in the criteria for choosing the models. Increases in the degree of the fixed curve and in the number of residual variance classes cause fewer problems in the analyses; however, they result in only a small improvement in the quality of the model in relation to the increases in the number of parameters of the additive genetic and permanent environmental curves, as can be seen in Figures 2, 3, and 4.

Multicollinearity may impair the estimates of the regression parameters, and its diagnoses are useful for detecting problems in heavily parameterized models. Hoerl and Kennard (1970) showed that the multicollinearity effect affects the distances between the estimator of the

least squares and the parameter. Cruz and Carneiro (2006) reported the following methods for the detection of multicollinearity: informal methods, analysis of elements of the correlation matrix, examination of the correlation matrix determinant, and analyses of eigenvalues and eigenvectors.

Table 1 shows the correlation values among the parameters for the regression of the additive genetic part (above the main diagonal) and permanent environmental part (below the main diagonal). It is possible to see that there was no significant correlation, which indicates that there were no multicollinearity problems, although they might exist without causing any problems in parameter estimates without causing significant correlations between them.

Table 1. Correlation among parameters of regression of parts, additive genetic (above the principal diagonal) and permanent environmental (below the principal diagonal).

Coefficient	1	2	3	4	5	6	7
1	1	0.6986	-	-	-	-	-
2	-0.2334	1	-	-	-	-	-
3	-0.5667	0.2881	1	-	-	-	-
4	0.3435	-0.3747	-0.1763	1	-	-	-
5	0.0798	-0.0867	-0.4176	0.0387	1	-	-
6	0.0134	-0.2461	-0.0261	-0.2029	-0.2960	1	-
7	-0.0743	0.1942	-0.0871	-0.0827	-0.2275	-0.1126	1

The determinant of a correlation matrix varies from zero to one; values close to zero indicate a linear dependence among the set of variables. The determinants of matrices of correlation between the genetic parameters and correlations among the permanent environmental parameters were 0.51 and 0.21, respectively, which does not indicate that there are no problems with respect to multicollinearity in the model.

The values of the condition numbers for genetic additive and permanent environmental regressions were 10.5 and 85, respectively, showing a weak effect of multicollinearity. The results for the genetic and environmental correlations between different lactation days ranged from 0.98 to 1.00, being even greater at the end of lactation. Such results show that animal selection for milk production on any day of the lactation curve will result in an improvement for all of the other points on the curve. Therefore, the greatest heritability point would be the most adequate one for practicing selection because it would provide the greatest correlated response (Falconer, 1981) for all of the other points inasmuch as genetic correlation is not reduced. However, selection using a correlated response may lose efficiency over time due to the occurrence of crossing-over.

Another way of practicing selection by using the results from RRM is based on the calculation of the area below the genetic curve of each animal, thereby selecting higher genetic value animals for all days of lactation. Such practice may bring advantages, but it is rarely used because it is very laborious. However, RENPED software (Silva, 2011) performs this calculation in a simple manner, even by directly using the RnSoln.dat file from WOMBAT (Meyer, 2006), to generate the area below the additive genetic curve of each animal.

CONCLUSIONS

The RRM using the most indicated LOP for the MYTD genetic evaluation of Alpine goats was the one that considered a fixed curve of order 4, curve of the additive genetic effect of order 2, curve of the permanent environmental effect of order 7, and at least 5 classes of residual variance.

Studies aiming to choose the best analysis model before genetic evaluation are very important because they generate more precise and accurate parameter estimates and predictions of genetic values.

REFERENCES

- Akaike H (1973). Information Theory and an Extension of the Maximum Likelihood Principle. Proceedings of 2nd International Symposium on Information Theory, Budapest.
- Araújo CV, Torres RA, Costa CN, Torres Filho RA, et al. (2006). Uso de funções ortogonais para descrever a produção de leite no dia de controle pro meio de modelos de regressão aleatória. *Rev. Bras. Zootec.* 35: 967-974.
- Cruz CD and Carneiro PCS (2006). Modelos Biométricos Aplicados ao Melhoramento Genético. 2nd edn. Editora UFV, Viçosa.
- Dempster AP, Laird NM and Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B* 39: 1-38.
- Falconer DS (1981). Introdução à Genética Quantitativa. UFV, Viçosa.
- Foulley JL and van Dyk DA (2000). The PX-EM algorithm for fast stable fitting of Henderson's mixed model. *Genet. Sel. Evol.* 32: 143-163.
- Freitas LS, Silva MA, Verneque RS, Santos GG, et al. (2008). Comparação de Modelos de Regressão Aleatória, para Avaliação Genética da Produção de Leite no Dia do Controle de Vacas Guzerá. VII Simpósio Brasileiro de Melhoramento Animal, São Carlos.
- Henderson, CR (1963). Selection index and expected genetic advance. In: Statistical Genetics and Plant Breeding. National Academy of Sciences-National Research Council, Washington.
- Henderson Jr CR (1982). Analysis of covariance in the mixed model: higher-level, nonhomogeneous, and random regressions. *Biometrics* 38: 623-640.
- Hoerl AE and Kennard RW (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics* 12: 69-82.
- Kirkpatrick M, Lofsvold D and Bulmer M (1990). Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics* 124: 979-993.
- Liu YX, Zhang J, Schaeffer LR, Yang RQ, et al. (2006). Short communication: Optimal random regression models for milk production in dairy cattle. *J. Dairy Sci.* 89: 2233-2235.
- Menezes GRO, Brito LF, Melo ALP, Silva FG, et al. (2008a). Comparação de Modelos de Regressão Aleatória Utilizados na Avaliação da Produção de Leite em Cabras da Raça Saanen. VII Simpósio Brasileiro de Melhoramento Animal, São Carlos.
- Menezes GRO, Melo ALP, Brito LF, Angelini MS, et al. (2008b). Avaliações e Diferentes Medidas da Persistência da Lactação de Cabras da Raça Saanen Utilizando Modelos de Regressão Aleatória. VII Simpósio Brasileiro de Melhoramento Animal, São Carlos.
- Menezes GRO, Caetano GC, Brito LF, Silva FG, et al. (2010). Estudo da Produção de Leite em Cabras da Raça Saanen Utilizando Modelos de Regressão Aleatória. VIII Simpósio Brasileiro de Melhoramento Animal - Melhoramento Animal: Uma Visão Crítica, Maringá.
- Meyer K (2006). PX x AI: Algorithmics for Better Convergence in Restricted Maximum Likelihood Estimation. Proceedings of the Eighth World Congress of Genetics Applied to Livestock Production, Belo Horizonte, 24-15.
- Meyer K (2009). WOMBAT - A program for Mixed Model Analyses by Restricted Maximum Likelihood. Manual Version 1.0. Armindale.
- Misztal I (2008). Reliable computing in estimation of variance components. *J. Anim. Breed. Genet.* 125: 363-370.
- Montgomery DC and Peck EA (1981). Introduction to Linear Regression Analysis. John Wiley & Sons, New York.
- Ptak E and Schaeffer LR (1993). Use of test day yields for genetic evaluation of dairy sires and cows. *Livestock. Prod. Sci.* 34: 23-34.
- Rao CR (1973). Linear Statistical Inference and its Applications. 2nd edn. John Wiley & Sons, New York.
- Sarmiento JLR, Reis Filho JC, Albuquerque LG, Lopes PS, et al. (2006). Avaliação genética de caprinos usando a produção de leite no dia de controle. *Rev. Bras. Zootec.* 35: 443-451.
- Sarmiento JLR, Albuquerque LG, Torres RA, Rodrigues MT, et al. (2008). Comparação de modelos de regressão aleatória para estimação de parâmetros genéticos em caprinos leiteiros. *Rev. Bras. Zootec.* 37: 1788-1796.
- Schwarz G (1978). Estimating the dimension of a model. *Ann. Stat.* 6: 461-464.
- Silva FG (2011). Estudo da curva de produção de leite de caprinos utilizando modelos de regressão aleatória. Master's thesis, Universidade Federal de Viçosa, Viçosa.