http://dx.doi.org/10.1590/1678-4162-12346

S. Genç
https://orcid.org/0000-0002-1512-9072
M. Mendeş
https://orcid.org/0000-0003-0109-9914

Arq. Bras. Med. Vet. Zootec., **v.73, n.4, p.949-954, 2021**

Linear modeling analysis using for determining the factors affecting 305-day milk yield

[*Análise de modelagem linear usado para determinar os fatores que afetam o rendimento leiteiro de 305 dias*]

S. *Genç*[1], M. *Mendeş*[2]

[1]Kirsehir Ahi Evran University - Faculty of Agriculture - Kirsehir, Turkey
[2]Canakkale Onsekiz Mart University - Faculty of Agriculture - Canakkale, Turkey

ABSTRACT

The purpose of this study was to model the factors affecting the 305-day milk yield of dairy cows by using Automatic Linear Modeling Technique (ALM). The data set of this study consisted of eight different cow breeds grown in eight province of Turkey. Results of ALM showed that the accuracy of the model was 64.2 % means that 64.2% of the variation in the 305-day milk yield could be explained by the constructed model. Created model was consisted of four factors namely the Breed, Lactation Length, Parity, and Province. Therefore, those selected factors were more efficient than the others in predicting the 305-day milk yield

Keywords: 305-day milk yield, automatic linear modeling, prediction, dairy cows

RESUMO

O objetivo deste estudo foi modelar os fatores que afetam a produção de leite das vacas leiteiras em 305 dias, utilizando a Técnica de Modelagem Linear Automática (ALM). O conjunto de dados deste estudo consistia em oito raças diferentes de vacas cultivadas em oito províncias da Turquia. Os resultados da ALM mostraram que a precisão do modelo era de 64,2% significa que 64,2% da variação na produção de leite de 305 dias poderia ser explicada pelo modelo construído. O modelo criado consistia de quatro fatores: Raça, Comprimento da Lactação, Paridade e Província. Portanto, esses fatores selecionados foram mais eficientes do que os outros na previsão da produção de leite de 305 dias.

Palavras-chave: 305 dias de produção de leite, modelagem linear automática, previsão, vacas leiteiras

INTRODUCTION

Milk yield is one of the major concerns especially for the scientists in the field of animal breeding. Therefore, the researchers try to increase genetic progress by selecting higher milk yielding animals for the next generation (Berry *et al*., 2007; Mirtagioglu *et al*., 2008). Milk yield of dairy cattle, as in the other farm animals (i.e. sheep, goat, and water buffalo), may be affected by different genetic and environmental factors and the relations between those factors (Mendes and Akkartal, 2009). In order to estimate genetic parameters, it is necessary to get pedigree record of all cows. Since milk yield is also affected by different environmental factors such as lactation length, calving interval, service period, calving age, calving month, herds etc. these kind environmental factors should also be considered for selection programs along with genetic factors (Khalid *et al*., 2007; Kuthu *et al*., 2007).

Therefore, determining the factors that will be able to affect the milk yield of dairy cattle is very important. There are different tests and approaches and mathematical models have been proposed for estimating milk yield of dairy cattle (Van Vleck and Henderson, 1961, Ashmawy *et al.,* 1985). In practice, in many cases, various mathematical models are used by the researchers to estimate milk yield and genetic progress in the future lactations. However, the reliability of those mathematical models depends on many

Recebido em 23 de fevereiro de 2021
Aceito em 6 de abril de 2021
E-mail: serdargenc1983@gmail.com

biological factors and thus those models will not be useful when these effects are not included in the model or not used correctly (Olori *et al*., 1999; De'ath and Katharina, 2000; David and Paul, 2004; Kocak *et al*., 2007; Zheng *et al*., 2009). However, different data mining techniques and regression-based methods have been developed and these techniques may be effectively used in determining the factors that affect milk yield (Lacroix *et al*., 1995). In this study, it has been aimed at determining important factors that can affect the 305-day milk yield of different dairy cattle breeds by using Automatic Linear Modeling (ALM).

MATERIAL AND METHODS

The data sets of this study were consisted from lactation records obtained from the Cattle Breeder Association of Turkey. 10 different factors (Breed, Lactation Length, Service Period, Dry Period, Parity, Calving, Calving Year, Calving Age, Province, and Calving Month) of different dairy cattle were considered in investigating relations between 305 day milk yield and those factors. The model for ALM is the same as multiple linear regression model. However, the ALM is more effective especially when there is a large and complex data set (if there are a large numbers of predictors) in terms of determining the factors or variables that affect the outcome.

Preparation of the Data for Analyzing. Firstly, in the lactation and year groups, the animals whose numbers was less than 100, which gave a stillbirth, dismissed birth, and left from the herd because of the reasons such as illness and disability were excluded from the evaluation during the data preparation. The animals whose lactation length was longer than 600 days and shorter than 220 days, those whose calving age was younger than 20 months and older than 45 months for first lactation and the subsequent lactations, and those except for the ones with 12 months added to previous lowest level and 14 months added to the highest level were excluded from the analysis. Besides, the data concerning calving interval which was below 300 days and above 675 days were not used. As a conclusion, in this study, 3808 lactation records among 305 day milk yield records. In this study, Automatic Linear Modeling approach (ALM) was used in modeling the factors affecting the 305-day milk yield of dairy cows (Statistical…, 2008). This method is also very useful for selecting variables and classifying:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon$$

Where Y_i is dependent or outcome variable, X_i's predictor or independent variables, β_0 is intercept or constant, β_p is the slope coefficients for each predictor and ϵ is the error term. Automatic Linear Modeling (ALM), is considered a relatively new method, introduced in SPSS software (version 19 and up), enabling researchers to select the best subset automatically especially when there are a large numbers of variables. In ALM, the predictor variables are automatically transformed in order to provide an improved data fit, and SPSS uses rescaling of time and other measurement values, outlier trimming, category merging and other methods for the purpose (Breiman *et al*., 1984; Bevilacqua *et al*., 2003; David and Paul, 2004; Camdeviren *et al*., 2005; Mendes and Akkartal, 2009; Karabag *et al*., 2010).

RESULTS AND DISCUSSION

Results of the Automatic Linear Model (ALM) are presented in Figure 1, 2, 3, and 4. In determining an appropriate model for fit our data set many models have been run (not discussed here) and it has been observed that except the model has been used in this study the other models have large information criterion values and above). The accuracy level of the model which is equivalent to the Adjusted R-squared value was found to be 64.2%, which means that the selected model might be accepted as a sufficient model which can be able to use in fitting and estimating process (Figure 1).

The lower the information criterion (AIC) is, the better the model is compared to models with a higher information criterion. Since the model used here has been the lowest information criterion value compared to the many other models, this model has been preferred in investigating the relations between the 305-day milk yield and independent variables (experimental conditions) (Camdeviren *et al*., 2005; Mendes and Akkartal, 2009; Karabag *et al*., 2010). The use of this method in animal science is not common when compared to the other fields of science.

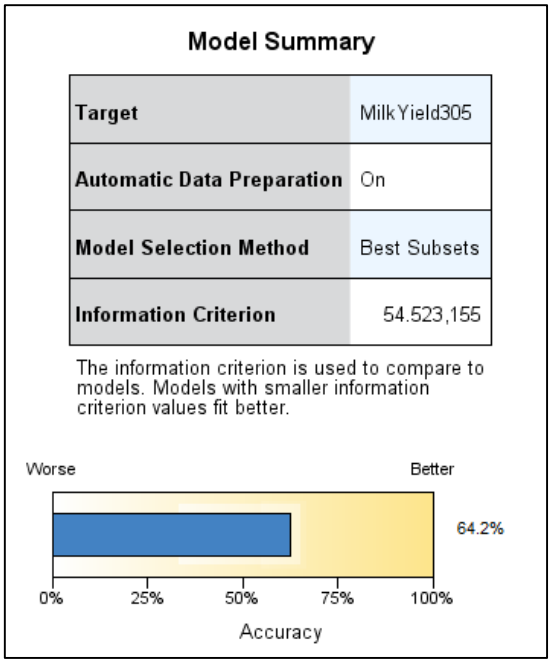

Figure 1**.** Accuracy Level of the Model. Importance levels of the predictors have been presented in Figure 2.

Figure 2 shows the predictors in the final model in rank order of importance. For linear models, the importance of a predictor is the residual sum of squares with the predictor removed from the model, normalized so that the importance values sum to 1. When Figure 2 is examined, it is seen that the most importance variables or factors that affect 305-day estimates are Breed, Lactation Length, Parity, and Province. Therefore, it can be concluded that the factors related to Breed, Lactation Length, Parity, and Province should be taken into consideration in order to get reliable and stable estimations.

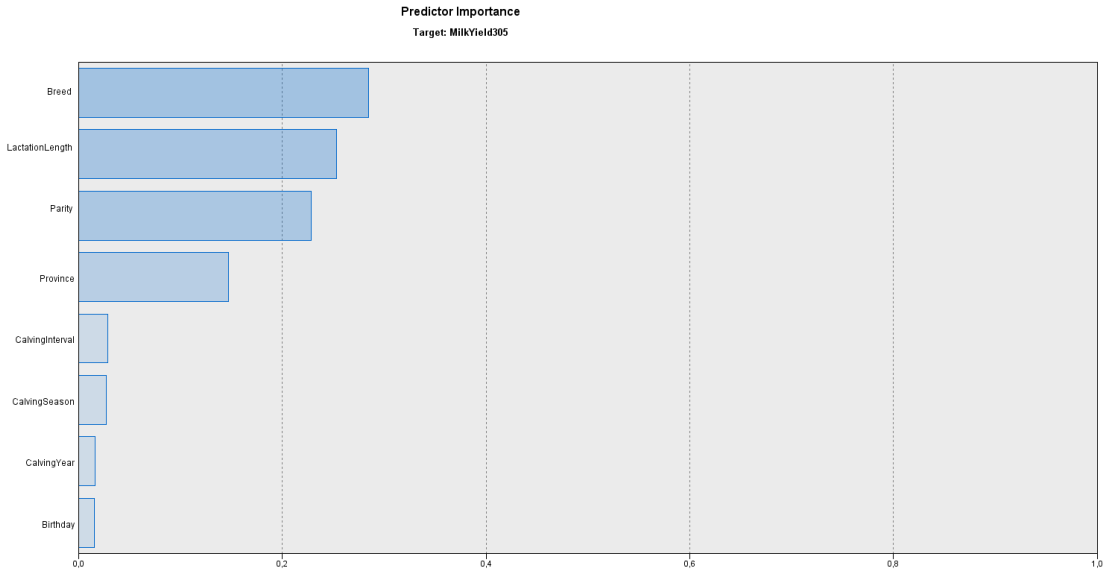

Figure 2. Importance Levels of the Variables or Predictors. After the most important factors or variables are determined (Figure 2), the Automatic Linear Modeling is re-run using just importance variables namely Breed, Lactation Length, Parity, and Province. The final results of Automatic Linear Modeling are presented in Figure 3 and 4.

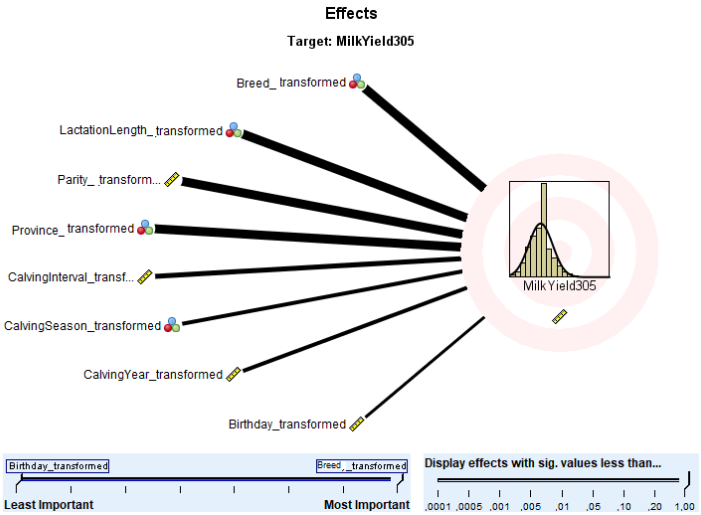

Figure 3. Diagram for showing statistical significance of variables or factors. This diagram is corresponding to the ANOVA table after multiple linear regression analysis. The predictors are ordered from top to bottom in order of importance levels, and the thickness of each line shows the statistical significance (P-values) of the relevant effect. As it is seen from the diagram in Figure 3, the Breed is the most important predictor or the factor affect the 305-day milk yield estimates, and it is followed by lactation length, parity, and province based on importance levels of them. The effect of all four predictors were statistically significant (P=0.000).

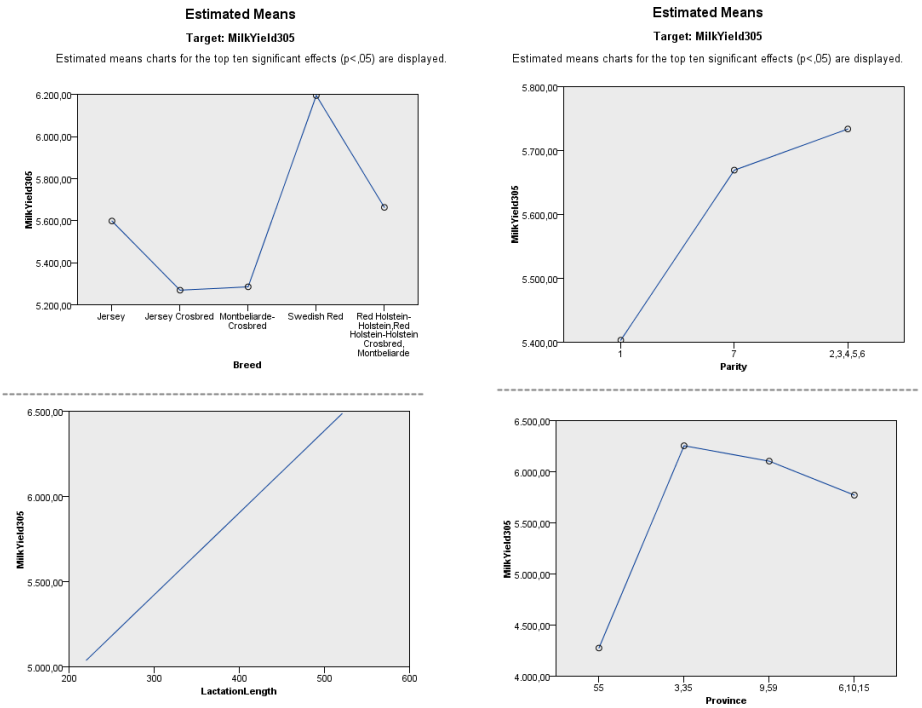

Figure 4. Diagram for showing the effect of significant factors. Especially in recent years, regression-based methods which can present the results graphically have become very popular (Berry *et al*., 2007; Mendes and Akkartal, 2009). Although the Automatic Linear Modeling is not commonly used regression-based method, we think that this method will become increasingly use for investigating relations between a dependent and many independent variables and determining the factors affecting interested variable. As a result of the ALM, the four most important factors were determined as breed, lactation length, parity, and province. Therefore, these factors should be included to the model in predicting the 305-day milk yield. With this model, 64.2% of the variation in the 305-day milk yield could be explained (Mendes and Akkartal, 2009; Topal *et al*., 2010).

According to the results of the ALM, it is possible to conclude that:

a) ALM results showed that 305-day milk yield is affected by different factors (breed, lactation length, parity, and province) and thus these factors should be considered in order to get more reliable and stable estimates.
b) Using the graphical methods instead of classical techniques makes it possible to investigate the complex and latent relations between dependent and independent variables especially when we have a large and complex data set.
c) Using the ALM in analyzing large and complex data sets might enable us to interpret the results easily.
d) Using ALM enables us to evaluate higher order interactions among the independent variables.
e) The ALM method can be efficiently used in all branches of sciences especially when there is a large and complex data set. However, it should not be ignored a potential threat of misuse of the ALM due to its simplicity.

Results of this study showed that the most important factors affecting the 305-day milk yield were the Breed, Lactation Length, Province, and Parity. Therefore, those selected factors were more efficient than the others in predicting the 305-day milk yield. It is thought that the effect of above factors on 305-day milk yield may change based on herd management, breeding systems, and maintenance and feeding. Furthermore, it is thought that the observed variation for the lactation length can be brought closer to the normal acceptable length (305 days) by arrangements to be made in production and marketing (Genc and Soysal, 2019).

CONCLUSİON

One of the other important factors that caused different results was the differences in the statistical techniques which are used in analyzing data sets. In this study, the Regression Tree Method was used in determining important factors on 305-day milk yield. In this way, it was possible to investigate the effect of latent and interrelated factors on milk yield estimation. Mendeş (2021) used ALM for evaluating results of Monte Carlo Simulation Studies and he informed that the ALM could be used efficiently to determine the factors that affect the response variable when there is a large and complex data set.

REFERENCES

ASHMAWY, A.A.; KHATTAB, A.S.; HAMED, M.K. Ratio and regression factors for predicting 305-day production from part-lactation milk records in Friesian cattle in Egypt. *Bull. Fac. Agric*., v.36, p.789-802, 1985.

BERRY, D.P.; BUCKLEY, F.; DILLON, P. Body condition score and live-weight effects on milk production in Irish Holstein-Friesian dairy cows. *Animal*, v.1, p.1351-1359, 2007.

BEVILACQUA, M.; BRAGLIA, M.; MONTANARI, R. The classification and regression tree approach to pump failure rate analysis. *Reliabil. Eng. Syst. Saf*., v.79, p.59-67, 2003.

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A. *et al*. *Classification and regression trees*. New York: Chapman and Hall, Wadsworth Inc., 1984. 368p.

CAMDEVIREN, H.; MENDES, M.; OZKAN, M.M. *et al*. Determination of depression risk factors in children and adolescents by regression tree methodology. *Acta Med*., v.59, p.19-26, 2005.

DAVID, R.L.; PAUL, L.S. Multivariate regression trees for analysis of abundance data. *Biometrics*, v.60, p.543-549, 2004.

DE'ATH, G.; KATHARINA, E.F. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, v.81, p.3178-3192, 2000.

GENC, S.; SOYSAL, M.I. Estimation of genetic parameters and genetic trend of Holstein Friesian cattle population in Turkey. *Fresenius Environ. Bull*., v.28, p.2617-2624, 2019.

KARABAG, K.; MENDES, M.; ALKAN, S. *et al*. An assessment of embryonic mortality stages in Chukar partridge (Alectoris chukar) by means of classification tree method. *Arch. Geflugelk*., v.74, p.269-273, 2010.

KHALID, J.; MASROOR, E.B.; MUHAMMAD, A. Within herd phenotypic and genetic trend lines for milk yield in Holstein Friesian dairy cows. *J. Anim. Biol*., v.1, p.66-70, 2007.

KOCAK, S.; TEKERLI, M.; OZBEYAZ, C. *et al*. Environmental and genetic effects on birth weight and survival rate in Holstein calves. Turk. *J. Vet. Anim. Sci*., v.31, p.241-246, 2007.

KUTHU, Z.H.; JAVED, K.; AHMAD, N. Reproductive performance of indigenous cows of Azad Kashmir. *J. Anim. Plants Sci*., v.17, p.47-51, 2007.

LACROIX, R.; WADE, K.M.; KOK, R. *et al*. Prediction of cow performance with a connectionist model. *Trans. Am. Soc. Agric. Eng*., v.38, p.1573-1579, 1995.

MENDES M. Re-evaluating the Monte Carlo simulation results by using graphical techniques. *Turk. Klinikl. J. Biostat*, v.13, p.28-38, 2021.

MENDES, M.; AKKARTAL, E. Regression tree analysis for predicting slaughter weight in broilers. *Ital. J. Anim. Sci*., v.8, p.615-624, 2009.

MIRTAGIOGLU, H.; KESKIN, S.; BAKIR, G. Regression tree analysis for 305 day milk yield in Holstein cows. *Indian Vet. J*., v.85, p.943-945, 2008.

OLORI, V.E.; HILL, W.G.; MC GUIRK, B.J. *et al*. Estimating variance components for test day milk records by restricted maximum likelihood with a random regression animal model. *Livest. Prod. Sci*., v.61, p.53-63, 1999.

STATISTICAL package social science: SPSS for windows release 17.0. Armonk: SPSS Inc., 2008.

TOPAL, M.; AKSAKAL, V.; BAYRAM, B. *et al*. An analysis of the factors affecting birth weight and actual milk yield in swedish red cattle using regression tree analysis. *J. Anim. Plant Sci*., v.20, p.63-69, 2010.

VAN VLECK, L.D.; HENDERSON, C.R. Estimates of genetic parameters of some functions of part lactation milk records. *J. Dairy Sci.*, v.44, p.1073-1084, 1961.

ZHENG, H.; CHEN, L.; HAN, X. *et al*. Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: the importance of phosphorus application rates under drought conditions. Agric. *Ecosys. Environ*., v.132, p.98-105, 2009.