Regression Tree Analysis for Determination of the Effective Factors on Birth Weight in Holstein Calves

Yalçın TAHTALI[1*]   Emine BERBEROĞLU[1]

[1]Gaziosmanpaşa University, Faculty of Agriculture, Department of Animal Science, 60250, Tokat, Turkey.
[*]e-mail:yalcin.tahtali@gop.edu.tr

Abstract: The aim of this paper was to describe the effects of calf sex, birth month and type on birth weight using Regression Tree (RT) analysis. For this purpose, 894 Holstein calves data raised in Polatlı State Farm were analyzed. The birth weight of calves averaged 38.478 ± 2.487 kg of total calves born, 95.5 % were single born. Twin born calves weights were lower (35.125 ± 1.652 kg) than single born (38.635±2.408 kg). Male calves were significantly (P<0.05) heavier than females by 1.28 kg. The mean birth weight of twin calves was 3.58 kg lower than that of single. Effects of calving month, sex of calf, birth type on birth weight were all significant (P<0.05).

Keywords: Birth month, birth weight, Holstein calf, regression tree analysis,

Siyah Alaca Buzağılarında Doğum Ağırlığı Üzerinde Etkili Faktörlerin Belirlenmesi için Regresyon Ağacı Analizi

Öz: Bu çalışmanın amacı regresyon ağacı (RT) analizi ile buzağı doğum ağırlığı üzerine, buzağı cinsiyeti, doğum tipi ve doğum ayının etkilerini belirlemektir. Bu amaçla, Polatlı Tarım İşletmesinde yetiştirilen 894 adet siyah alaca buzağılarına ait veriler analiz edilmiştir. Buzağılara ait ortalama doğum ağırlığı 38,478 ± 2.487 kg olarak bulunmuştur. Toplam buzağılamanın % 95.5 tek doğum şeklinde oluşmuştur. İkiz doğanların ağırlığı (35,125 ± 1.652 kg) tek doğanlardan (38,635 ± 2.408 kg) daha düşük bulunmuştur Erkek buzağılar, dişi buzağılardan 1.28 kg daha ağır ve önemli bulunmuştur (P<0.05). İkiz doğanların doğum ağırlığı ortalaması, tekiz doğanlardan 3.58 kg daha düşük bulunmuştur. Doğum ağırlığı üzerine buzağılama ayı, buzağı cinsiyeti ve doğum tipi etkileri anlamlı bulunmuştur (P<0.05).

Anahtar Kelimeler: Doğum ağırlığı, doğum ayı, analizi, Siyah Alaca buzağı, regresyon ağacı

1.Introduction

Birth weight is used as a first important measure of growth performance for animal science and birth weight is the basic and most reliable measure of growth during the prenatal period (Kaygısız, 1998). Environmental effects are also known to be very important in growth and the trait is under the polygenic control. There are many researches about the effects of sex, calving season, breed, year and birth type on birth weight in calves using the general linear model (Kaygısız, 1998; Akbulut et al., 2001; Koçak et al., 2007; Tilki et al., 2008; Aksakal and Bayram,

2009). Logistic regression analysis can be practical when the dependent variable is a nominal scale. This regression is a technique for modelling the probability of an occurrence in terms of appropriate explanatory or determiner variables. Logistic regression technique can also be applied to ordered categories (ordinal data). If the dependent variable has several unordered categories, then the use of discriminant analysis is preferred.

Regression Tree analysis is a nonparametric separating method, which determines the interaction between the independent and

dependent variables. This analysis is a nonparametric alternative method to logistic regression method and the least squares method. This method does not include the assumptions required for the regression analysis. In regression tree analysis, independent and dependent variables can be continuous, ordinal or nominal (Doğan ve Özdamar, 2003; Topal et al., 2010).

Regression Tree analysis has a few advantages compared with other statistical methods as variance analysis and logistic regression. These advantages; Regression Trees are invariants under transformations of independent variables and nonparametric methods. The purpose of the regression tree algorithm includes the important variables explaining the dependent variables and eliminates insignificant variables; Interactions within the data set can be determined and the graphical explication is easier (Yohannes et al., 1999; Timofeev., 2004).

In general, previous research used variance analysis and regression analysis methods to identify the factors affecting birth weight in calves. Therefore, in this paper was aimed to determine the effects of calf sex, birth month and birth type on birth weight using Regression Tree (RT) analysis. In this study, compound categories and subgroups of independent variables were estimated by CHAID analysis algorithms.

2. Material and Methods

Data included birth weights and pedigree information of Holstein calves born over a period of 16 years from 1995 to 2005 (11 years) from the Polatlı State Farm in Ankara (Turkey). Firstly, abnormal records were excluded from the data. The data set of this paper were composed of birth weight (kg), birth type (single and twin), sex (male and female) and birth month of 894 Holstein calves raised at Polatlı state farm. In this paper, birth weight was used as the dependent variable. The others (birth type, birth month, sex

of calf,) were used as independent variables. The birth weights of calves were measured within a few hours following birth. In this study analyzed the effects of variables on birth weight using Regression Tree analysis. Statistical analysis and evaluations were performed using SPSS 17.0 (SPSS Inc, 2011) statistical package software.

This analysis is a nonparametric approach which is a model that describes the relationship between independent and dependent variables (Chang and Wang, 2006; Topal et al., 2010). Specifications of the data for regression tree analysis are in Table 1.and Table 2 gives the number of observations by category.

Table 1. Characteristics of data used for regression tree analysis

Çizelge 1. *Regresyon ağacı analizi için kullanılan verilerin özellikleri*

Description	Number of records
Number of records	894
Year (1995-2005)	10
Number of animals	256
Months	12
Number of sires	86
Number of dams	244

The Regression Tree analysis creates a tree-based classification model. This procedure classifies cases into groups or predicted values of a dependent variable based on values of independent variables. This analysis provides validation tools for confirmatory and exploratory classification analysis.

The purpose of regression tree analysis, is to compose a structure according to the independent variables that built the most homogeneous nodes (Larsen and Speckman, 2004; Zheng et al., 2009).

Table 2. Independent categorical variables with the number in each category
***Çizelge 2.** Her kategorideki veriler için bağımsız kategorik değişkenler*

	Variables																
	Birth Type		Sex of Calf		Birth Months												
Categories	Single	Twin	Male	Female	January	February	March	April	May	June	July	August	September	October	November	December
N	854	40	461	393	101	104	72	62	64	48	50	55	67	98	82	91
%	95.5	4.5	53.99	46.01	11.30	11.63	8.05	6.94	7.16	5.37	5.59	6.15	7.49	10.96	9.17	10.18

The dependent variable is located at the top of the regression trees which is called root node. Dependent variable data is split into a category of right and left child nodes derived from the primary nodes. When the split is completed, child nodes are defie as terminal nodes. Depending on the type of independent variables, numerous splitting occurs (De'ath and Fabricius, 2000). If the dependent variable is numerical, the residual sum of squares method is used for regression tree analysis. In this case, impurity is defined as the sum of squares of response values (observations) around each node (Topal et al., 2010; Questier et al., 2004). Splits are selected on principle of minimizing the sum of the squared deviations from the mean in each node and observation. The function is shown as follows:

$$D(T) = \sum_{i \in N} \left\{ y_i - \overline{y}(N) \right\}^2$$

Where, $\overline{y}(N)$ is the mean over the observations in node N. The root node consists of all observations. At each step, an attempt is made to divide a parent node, N, into child nodes, a

"left" node N_{left} and "right" node N_{right}, so as to minimize $D(N_{left}) + D(N_{right})$. Splits of the following forms are considered:

1. If x_j is a continuous variable, regard to all splits of the form $N_{left} = \{i \in N : x_{ij} \le t\}, N_{right} = \{i \in N : x_{ij} > t\}$ for constant t.
2. If x_j is ordinal, regard to all splits as in (Larsen and Speckman, 2004).
3. If x_j is categorical with L levels, regard to all 2^L subsets. We need not consider the empty set. Moreover for each split into left and right subsets, there is an equivalent split with the subsets reversed. Thus there are actually 2^{L-1} -1 splits to consider (Larsen and Speckman, 2004).

3. Results and Discussion

A regression tree diagram of the characters which are expected to affect birth weight in Holstein calves are given in Figure 1.
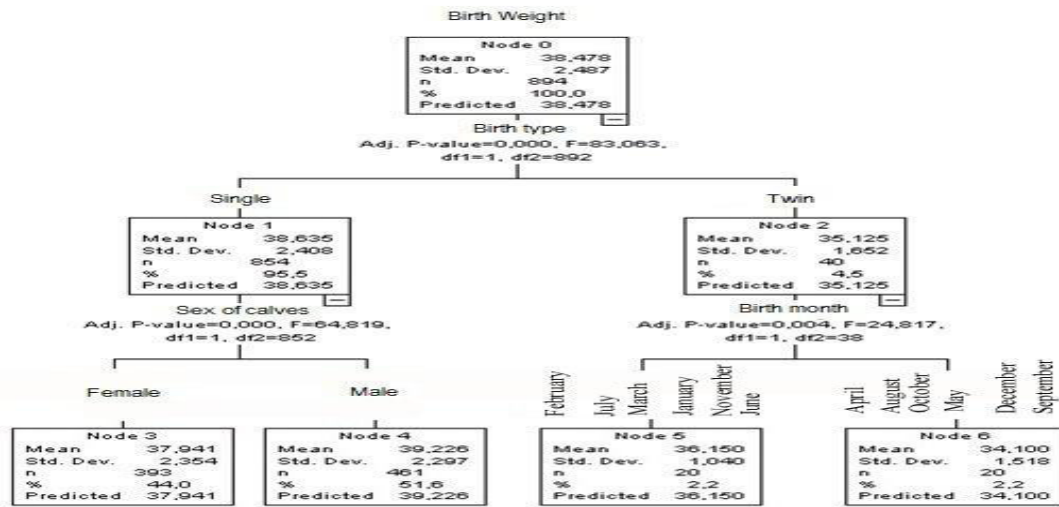
Figure 1. Regression tree diagram for birth weight
Şekil 1. Doğum ağırlığı için regresyon ağacı diyagramı

Descriptive statistics of birth weight are given in the main node of the regression tree. If the structure of regression tree is examined, it is seen that Node 0 which is the root node and the descriptive statistics of dependent variables are placed at top of tree (Figure 1). The effects of birth type (single and twin), sex (female and male), birth month (February, July, March, January, November, June, April, August, September, December, October and May) on birth weight were found significantly important in this

study (P<0.001; P<0.005). Average birth weight and standard deviation in Holstein calves were found to be 38.478 ± 2.487 kg. The first child node is called Node 1. This primary node was divided into two child nodes according to the birth type variable, which indicates that birth type is the most influential variable on birth weight in Holstein calves. It was found that the cattle in birth type node 1 had a total of 854 calves; average birth weight of these calves was 38.635 ± 2.408 kg. Descriptive statistics of nodes are given by Table 3.

Table 3. Descriptive statistics of nodes
Çizelge 3. Boğumlar için Tanımlayıcı İstatistikler

Node	N	Percent	Mean	Std.Dev.	Predicted
4	461	51,6 %	39,23	2.297	39,226
3	393	44,0 %	37,94	2,354	37,941
5	20	2,2 %	36,15	1,040	36,150
6	20	2,2 %	34,10	1,518	34,100

Growing Method: Exhaustive chaid

The ratio of single born calves count to total count was found 95.5 %. The second child node named as Node 2 was composed of 40 twin born calves. The twin born calves within the enterprise were 4.5 % and birth weight was 35.125 ± 1.652 kg. Single born calves were of higher birth weight

(38.635 ± 2.408 kg) than twin born calves (35.125 ± 1.652 kg). The mean birth weight of twin calves was 3.58 kg lower than that of single calves. Node 1was further split into two child nodes (Node 3 and Node 4) according to the sex of the calf. The regression tree diagram shows that the sex of the

calf is the secondary variable affecting the birth weight of the calves. Node 1 was further split into 2 nodes by the sex variable. In Node 3, average birth weight of female calves was 37.941±2.354 kg. In node 4, average birth weight of male calves was 39.226 ± 2.297 kg. The number of calves in node 3 and node 4 were 393 (44.0 %) and 461 (51.6 %). Female calves birth weight was significantly (P<0.001) lower than males by 1.28 kg. Node 2 was further split into two child nodes (Node 5 and Node 6) according to birth month. According to the regression tree diagram, birth month is the secondary variable affecting the birth weight of the calves. Average birth weight of the calves born in February, July, March, January, November and June was found 36.150 ± 1.040 kg. Average birth weight of the calves born in April, August, September, December, October and May was 34.100 ± 1.518 kg, respectively. The calves born in February, July, March, January, November and June were 2.05 kg higher than that born in April, August, September, December, October and May. Node 4 had the heaviest (39.226 ± 2.297 kg) mean birth weight of the six nodes contain the regression tree.

Recorded birth weights of 894 Holstein calves born on the Polatlı State Farm in Turkey were analyzed. It is important to estimate the effects of independent variables (birth type, sex of calf, birth month) particularly in specific hypothetical studies related to the effects of independent variables or variable groups on a dependent variable. Various statistical analysis methods can be used for this purpose. In this study, alternative to general methods (Least squares regression analysis, Logistic regression, etc.)regression tree analysis was used to determine the effects of various factors on birth weight of calves. In this study, type of birth, sex of calf and month of birth effects on calf birth weight is determined by the method of regression tree analysis. The most important advantage of regression tree analysis over other methods is that it does not include assumptions, as it is a nonparametric method. In this study according to the regression tree diagram shows that birth type was the most effective variable on birth weight, followed by sex of calf

and birth month. Rates of twin and single born calves were 4.5 % and 95.5 %, respectively.

Twin calves had a mean birth weight of 1.128 kg (3.21 %) lower than single born calves. These results are in agreement with those reported by Tilki et al. (2003), Bakır et al. (2004), and Tilki et al. (2008). In addition, similar results were reported by Aksakal and Bayram (2009), and Topal et al. (2010) working on different breeds in Turkey. This variation was reported by Aksakal and Bayram (2009) as 6.26 kg for Holstein calves. Also, Bakır *et al.* (2004) and Kertz *et al.* (1997) reported that among dairy cows, twin calves had a birth weight of approximately 11.2 % or 15.9 % lower than those of single born calves.

In the present study, the second most important variable affecting the birth weight of calves was found to be sex of calves. Birth weights of female calves were lower than males (1.28 kg). Similarly Aksakal and Bayram (2009) reported that male calves had 2.69 kg higher birth weight than females and this differences was found significant (p<0.01).Sex of calf was found to have a significant effect on birth weigth. Kaygısız (1998), Akbulut et al. (2001), Tilki et al. (2003), and Tilki et al. (2008) on research different herds of Brown Swiss calves in Turkey, found a significant effect of calf sex on birth weight. The other hand in parallel to the results obtained, (Akbulut et al., 2001;Bakır et al., 2004, Kertz et al., 1997) reported that male calves had 1.3 to 3.6 kg higher birth weight than females.

The second most important variable affecting the birth weight of calves was found to be calving month. Average birth weight of the calves born in February, July, March, January, November and June was 36.150 ± 1.040 kg, average birth weight of the calves born in April, August, September, December, October and May was 34.100 ± 1.518 kg, respectively. The calves born in February, July, March, January, November and June were 2.05 kg higher than those born in April, August, September, December, October and May. Node 4 had the heaviest (39.226 ± 2.297 kg) mean birth weight of the six nodes comprising the regression tree.

170

One of studies (Akbulut et al., 2001; Koçak et al., 2007) also reported that season had a significant effect on birth weight. But effect of birth month on birth weight of calves was not found.

Regression tree is a strong tool for analysis of complex data and this method is a valuable addition to the statistical tools of every researcher (Doğan ve Özdamar, 2003).

The regression tree analysis method can be used to analyze the effect of other variables on a particular variable in dairy cows. Only a limited number of studies have used regression tree analysis in studies of dairy calves. Since it is a nonparametric method, regression tree analysis does not include the assumptions required for test. The method has important advantages for application to continuous, categorical and grading measurement scales and easy interpretation of the results. Using this method, the significance of the factors affecting economic characters such as birth weight was determined, and this method served as a guide for future improvement studies for these two characteristics. It is important that define the relationship between independent and dependent (birth weight) variables measured at early time, since early selection is one of the modern selection programs applied birth weight in dairy calves.

References

Akbulut, Ö., Bayram, B. and Yanar, M (2001). Estimates of phenotypic and genetic parameters on birth weight of brown swiss and holstein friesian calves raised in semi entansif conditions. Lalahan Hay Arşt Derg, 41(2): 11-20.

Aksakal V and Bayram B (2009). Estimates of genetic and phenotypic parameters for the birth weight of calves of Holstein friesian cattle reared organically. J Anim Vet Adv, 8(3): 568-572.

Bakır G, Kaygısız A and Ulker H (2004). Estimates of genetic and phenotypic parameters for birth weight in Holstein Friesian cattle. Pakistan J Biol Sci, 7(7): 1221-1224.

Bakır G, Keskin S and Mirtagioğlu H (2010). Determination of the effective factors for 305 days milk yield by regression tree (RT) method, J Anim Vet Adv, 9 (1): 55-59.

Chang LY and Wang HW (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis and Prevention 38: 1019-1027.

De'ath G and Fabricius KE (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology, 81(11): 3178-3192.

Doğan N and Özdamar K (2003). CHAID analysis and an aplication related with family planning. Türkiye Klinikleri J. Med. Sci., 23: 392-397.

Doğan, İ (2003). Investigation of the factors which are affecting the milk yield in Holstein by CHAID analysis. Ankara Univ Vet Fak Derg, 50: 65-70.

Kaygısız, A. (1998). Estimates of genetic and phenotypic parameters for birth weight in brown swiss and simmental calves raised at Altındere State Farm. Turk J Vet Anim Sci, 22: 527-535.

Kertz AF, Reutzel LF, Barton BA and Ely RL (1997). Body weight, body condition score and wither height of prepartum Holstein cows and birth weight and sex of calves by parity: A data base and summary. J Dairy Sci, 80: 525-529.

Koçak, S., Tekerli, M., Ozbeyaz, C. and Yüceer, B (2007). Environmental and genetic effects on birth weight and survival rate in Holstein calves. Turk J Vet Anim Sci, 31(4): 241-246.

Larsen DR and Speckman PL (2004) Multivariate regression trees for analysis of abundance data. Biometrics, 60: 543-549.

Questier F, Put R, Coomans D, Walczak B and Vander HY (2004). The use of CART and multivariate regression trees for supervised and unsupervised feature selection. Chemometrics and Intelligent Laboratory Systems, 76: 45-54.

SPSS 17. (2011). Statictical package for Social Sciences (SPSS) for Windows Release 17.0 SPSS Inc,

Tilki M, İnal Ş, Tekin ME and Çolak M (2003). The Estimation of Phenotypic and Genetic Parameters for calf birth weight and gestation length of brown swiss cows reared at the Bahri Dağdaş International Agricultural Research Institute, Turk J Vet Anim Sci., 27, 1343-1348.

Tilki M, Saatcı M and Çolak M (2008). Genetic parameters for direct and maternal effects and estimation of breeding values for birth weight in Brown Swiss cattle, Turk J Vet Anim Sci, 2008; 32(4): 287-292.

Timofeev R (2004). Clasification and Regression Trees (CART) theory and applications. Center of Applied Statistics and Economics, Humboldt University, Unpublished Master Thesis, Berlin.

Topal M, Aksakal V, Bayram B and Yağanoğlu AM (2010). An Analysis of the factors Affecting Birth Weight and Actual Milk yields in Swedish Red cattle using regression tree analysis. J Anim Plant Sci, 20(2): 63-69.

Yohannes Y and Hoddinott J (1999). Classification and Regression Tress: An Introduction. International Food Policy research Institute, Washington, D.C. 20006 U.S.A.

Zheng H, Chen L, Han X, Zhao X and Ma Y (2009). 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. Agriculture, Ecosystems & Environment, 132: 98-105.