Ankara Univ Vet Fak Derg, **69**, 367-373, 2022
DOI: 10.33988/auvfd.803924

Different coding systems for the modeling of lactation milk yields of Awassi sheep

İrfan GÜNGÖR[1,a,✉], Fatih ATASOY[2,b]

[1]Republic of Türkiye Ministry of Agriculture and Forestry, General Directorate of Agricultural Research and Policies, Department of Livestock and Aquaculture Research Ankara; [2]Ankara University, Faculty of Veterinary Medicine, Department of Animal Science, Ankara, Türkiye
[a]ORCID: 0000-0001-6248-3464; [b]ORCID: 0000-0002-9060-3950

✉Corresponding author: igungor@ankara.edu.tr
Received date: 01.10.2020 - Accepted date: 21.08.2021

Abstract: This study evaluated the feasibility of using different coding systems for categorical variables when using continuous and categorical variables together for the modeling of the lactation milk yield of Awassi sheep. In the study, when all variables were included in the model, and Dummy Coding and Effect Coding methods were used for age, the effects of lactation duration, average daily milk yield, type of birth, and age 5 group were found to be statistically significant in addition to the constant term. When the Deviation Coding method was used for age, the effects of lactation duration and average daily milk yield were found to be statistically significant in addition to the constant term. On the other hand, when Forward and Backward Coding methods were used, the effect of the age 5 group was found to be statistically significant, along with the effects of lactation duration and average daily milk yield. The results of the study indicated that different results can be obtained depending on the various coding systems used. The results also indicated that the choice of coding system affected the interpretation of the obtained coefficients. Therefore, it can be stated that the aims of the researcher in the study should be defined clearly and the proper codding system should be selected according to the variables to be included in the model.
Keywords: Awassi sheep, milk yield, multiple regression, reference category, variable coding systems.

İvesi koyunlarında farklı kodlama sistemleri kullanılarak laktasyon süt veriminin modellemesi

Özet: Bu çalışmada, İvesi koyunlarda laktasyon süt verimi için sürekli ve kategorik değişkenlerin birlikte ele alınarak; kategorik değişkenler için farklı kodlama sistemlerinin uygulanabilirliği değerlendirilmiştir. Çalışmada ele alınan tüm değişkenlerin modele dâhil edilmesi durumunda, yaş için kukla ve etki kodlama yöntemleri kullanıldığında; sabit terim ile birlikte laktasyon süresi, günlük ortalama süt verimi, doğum tipi ve 5 yaş grubundan kaynaklanan fark istatistik olarak önemli bulunmuştur. Yaş için sapma kodlama yöntemi kullanıldığında; sabit terim ile birlikte, laktasyon süresi ve günlük ortalama süt verimine ait etkiler istatistik olarak önemli bulunmuştur. İleriye ve geriye dönük fark yöntemleri kullanıldığında ise laktasyon süresi ve günlük ortalama süt verimi ile birlikte 5 yaş grubunun negatif etkisi de istatistik olarak önemli bulunmuştur. Çalışmanın sonuçları, kullanılan kodlama sistemilerine göre farklı sonuçların elde edilebileceğini göstermiştir. Sonuçlar ayrıca, kodlama sistemi seçiminin, elde edilen katsayıların yorumlanmasını etkilediğini göstermiştir. Bu nedenle araştırmacının araştırmadaki amaçlarının açık bir şekilde belirlenmesi ve modele dahil edilecek değişkenlere göre uygun kodlama sisteminin seçilmesi gerektiği söylenebilir.
Anahtar sözcükler: Çoklu regresyon, değişken kodlama sistemleri, İvesi koyunu, referans kategori, süt verimi.

Introduction

The decision of which statistical methods are to be used for the analysis of data obtained from research in terms of variables or characteristics of interest is related directly to the variable type; in other words, the data structure and the means of acquisition. So, when deciding on the best statistical method for an analysis, a researcher should think about the types of variables and other environmental factors (11, 12, 13).

The yield and quality of economically important animal products such as meat, milk, eggs, fleece, and honey are affected by many factors, some of which are continuous, such as age and weight, while others are categorical, such as sex and birth type. When breeding to improve yield and quality, it is very important to use the right method of analysis for these economically important products (2, 9).

The direct inclusion of such categorical variables as gender, type of birth, and lactation order in standard multiple regression analysis models violates the assumptions of a regression analysis (14), and in these situations, different regression approaches may be used. The difficulties associated with the implementation of these approaches and the interpretation of their results lead few researchers to make use of them, and more often than not, these variables are included in the model after coding (3). When examining their relationship with the response variable, categorical variables can be included in the same model as continuous variables, and this makes it possible to identify the effects of the categorical and continuous variables included in the model on the response variable, as well as any potential interactions among the explanatory variables (3, 13, 14).

In the Dummy coding method, as one of the most frequently used coding approaches, the values of 1 and 0 are used to indicate whether individual observations belong to a particular group. The variables used in dummy coding are known as artificial variables, and do not exist in the original data, being created later for the transformation of categorical data into numerical data. Dummy coding is the preferred coding method when the goal is to compare multiple treatment groups with a single control group. In this case, the control group is called the reference group, and the differences between the regression coefficient of this group and those of other groups are examined. The statistical significance of the regression coefficients for these variables is tested using the t statistic (15, 18). In effect coding, dummy variables are assigned values of 1 or -1, which is a method that is similar to dummy coding, although there are differences in how the reference group is defined. The reference group is defined as "0" in dummy coding, and as "-1" in effect coding. The R^2 and F values in regression models are the same for the two coding methods, but the regression constant and regression coefficients are interpreted differently. In Alkharusi's (3) examination of the dummy coding and effect coding methods, it was reported that similar R^2 and F values were obtained through the two methods, although the interpretation was different, depending on the coding method used. When including categorical variables in a multiple regression model, the choice of the statistical software to be used is also important.

According to the literature review, it was observed that there were almost no studies about examining and interpreting different coding systems together in animal science. Therefore, in this study, various coding systems were examined together and their usability in modeling milk yield was evaluated, as well as the results obtained according to different coding systems were interpreted.

Materials and Methods

Material: Data on 287 Awassi sheep kept in the Şanlıurfa GAP Agricultural Research Institute (GAPTAEM) of the General Directorate of Agricultural Research and Policies (TAGEM) of the Ministry of Agriculture and Forestry were collected between 2013 and 2015.

Included in the study were ewes aged 3–5 years, whose lactation duration ranged from 30 days to 191 days, and lactation milk yield (LMY) varied between 27 and 248 kg. Ewe weight varied between 40 kilograms and 73 kilograms; 29% of the ewes gave birth to twins and 57% of the lambs were male. The lactation milk yields and lactation curves of ewes that lambed in different periods were examined. Lambing started in a different month in each year of the study, in November, December and January. The flock was cared for and feeding was performed with routine procedures in the Institute. Milk controls were performed on a 24-hour basis and repeated every 20 days, and the milk controls ended when two-thirds of the flock had finished lactating, upon which milking was ended for the entire flock.

Methods: The explanatory variables included in multiple regression models are usually continuous; although in many cases it is important to include also categorical variables in the model to improve its goodness of fit, to eliminate prediction errors, and to identify any potential interactions or joint effects (3, 10). Categorical variables are coded qualitatively, meaning that the assigned codes have no numerical values, and these variables can be included in standard regression analysis models as independent or explanatory variables. Regarding the coding system, if the categorical variable has "g" levels, it is possible to code "g-1" binary variables. For a gender variable with two categories (male, female), g = 2, and coding for either male or female would suffice (18, 19). The coding systems applied in the study are Dummy Coding, Effect Coding, Deviation Coding, Forward Difference Coding, Backward Difference Coding, Helmert Coding and Reverse Helmert Coding (1, 3, 6, 14). The coding schema for age categories was presented in Table 1.

The model used to examine the environmental factors that affect the observed lactation milk yield (OLMY), the estimated lactation milk yield (ELMY) and the model parameters was as follows: Y_{ijk}

$$Y_{ijk}= \mu+YA_i+ DT_j+b\ (X_{ijk})+ e_{ijk}$$

Yijk: ELMY, OLMY, model parameters, μ: Overall mean in terms of the analyzed trait, YA_i: i^{th} lambing year-Month, DT_j: J^{th} birth type, b: the partial regression coefficient of X_{ijk}, X_{ijk}: lactation length of the kth ewe, e_{ijk}: Residual associated with Y_{ijk}.

Table 1. Coding schema for age categories.

Age group	Dummy			Effect		Deviation			Forward		Backward		Helmert		R. Helmert	
	D1	D2	D3	E1	E2	S1	S2	S3	F1	F2	B1	B2	H1	H2	R1	R2
3	0	0	0	-1	-1	-1/3	-1/3	-1/3	1	0	-1	0	1	0	-1	-1/2
4	1	0	0	1	0	-1/3	2/3	-1/3	-1	1	1	-1	-1/2	1	1	-1/2
5	0	1	0	0	1	-1/3	-1/3	2/3	0	-1	0	1	-1/2	-1	0	1

R. Helmert: Reverse Helmert.

Table 2. Results of the regression model for lactation milk yield, including all variables, with Dummy coding and Effect coding for Age (kg).

Item	Dummy		Effect	
	b ± SE	t (P)	b ± SE	t (P)
Constant	-105.707 ± 5.969**	-17.708 (0.001)	-106.630 ± 6.062	-17.590 (0.001)
Birth Weight	-0.236 ± 0.751	-0.314 (0.754)	-0.204 ± 0.747	-0.273 (.785)
Ewe Weight	0.046 ± 0.081	0.571 (0.569)	0.036 ± 0.081	0.438 (.662)
Lac. Dur.	0.817 ± 0.019**	43.937 (0.001)	0.823 ± 0.019**	-43.897 (0.001)
ADMY	0.130 ± 0.002**	71.201 (0.001)	0.130 ± 0.002**	71.433 (0.001)
Sex	-0.622 ± 1.027	-0.606 (0.545)	-0.354 ± 0.511	-0.693 (0.489)
Type of Birth	-3.446 ± 1.310**	-2.630 (0.009)	-1.630 ± 0.653*	-2.494 (0.013)
Age 3	0.153 ± 1.370	0.112 (0.911)	-0.824 ± 0.903	-0.913 (0.362)
Age 4	0.763 ± 1.488	0.513 (0.608)	-0.526 ± 0.907	-0.580 (0.562)
Age 5	3.899 ± 1.795*	2.172 (0.031)	2.554 ± 1.156*	2.211 (0.028)
	R^2 = 0.977 F = 1186.665 P = 0.001		R^2 = 0.977 F = 1199.843 P = 0.001	

SE: Standard error, F: F statistic, t (P): t statistic (P value), Lac. Dur.: Lactation Duration, b: Coefficient, ADMY: Average Daily Milk Yield, R^2: Determination coefficient, *: P <0.05 , **: P<0.01.

The following model was used to analyze the factors (environmental and flock management) affecting lactation length (LD):

$$Y_{ijk} = \mu + YA_i + DT_j + e_{ijk}$$

Y_{ijkl}: ELMY, OLMY, model parameters, μ: Overall mean in terms of analyzed trait, YA_i: i^{th} lambing year-Month, DT_j: J^{th} birth type, e_{ijk}: Residual associated with Y_{ijk}.

Regression analyses were carried out for the study. The level of statistical significance was set as 5% and IBM SPSS Statistics software (Version 21.0. Armonk, NY: IBM Corp.) was used for all statistical computations.

Results

The results of the regression analysis are presented in Table 2, in which lactation duration, average daily milk yield, birth type, and age 5 can be seen to have had statistically significant effects (P<0.05), while the other variables did not. Of the variables with significant effects, all but birth type had positive coefficients. A one-day increase in lactation duration was thus predicted to increase the mean LMY by 0.817 kg, and a 1 kg increase in average daily milk yield was predicted to increase the mean LMY by 0.130 kg.

For the Dummy Coding of the birth type, the singleton category was used as the reference category. Thus, a value of 3.446 for type of birth denoted the difference between the groups of ewes that gave birth to single lambs and those that gave birth to twins. This coefficient had a negative sign, indicating that the ewes that gave birth to twins had a mean LMY of 3.446 kg lower than those that gave birth to single lambs. Similarly, for the variable of age, the "Age 3" group was used as the reference category. The regression coefficients for the differences between the mean LMY of the reference group and the mean LMY of the groups of ewes aged 3, 4, and 5 years were positive, indicating those with higher ages were associated with higher LMY. The differences between the mean LMY of the groups of ewes aged 3 and 4 and the reference category, that is to say, the group consisting of ewes aged 3, were not statistically significant (0.153 and 0.76 kg, respectively). In contrast, the difference between the mean LMY of the group Age 5 and that of the reference category (3.899 kg) was found to be statistically significant (P<0.05). Accordingly, the Age 5 group was predicted to have a mean LMY 3.899 kg higher than the mean LMY of the Age 2 group. When all variables were included in the model, the coefficient of determination (R2) was found to be 97.7%, which is higher. So, it was decided that the variables in the model could explain 97.7% of the change or variation in LMY. The other 2.3% could be explained by random environmental factors that were not part of the model.

To understand whether they affect LMY, the categorical variables were coded using the effect coding method and included in the model along with birth weight, lactation duration, and average daily milk yield. The results of the regression analysis for this model are presented in Table 2. As was the case with dummy coding, Table 2 shows that of the variables included in the model, only lactation duration, average daily milk yield, type of birth, and age 5 were found to have statistically significant differences (P<0.05). All of these variables had positive coefficients except birth type; thus, a one-day increase in lactation duration was predicted to increase the mean LMY by 0.823 kg, while a 1 kg increase in average daily milk yield was predicted to increase the mean LMY by 0.130 kg.For the effect coding of birth type, ewes that gave birth to single lambs were coded "-1" and those that gave birth to twins were coded "1". Thus, the mean LMY of ewes that gave birth to single lambs was predicted to be 106.630-1.630 = 105.00 kg, whereas the mean LMY of ewes that gave birth to twins was predicted to be 106.630 + 1.630 = 108.26 kg. For the effect coding of the variable age, the "Age 3" group was used as the reference category. Of the groups of ewes aged 3, 4, and 5, only the age 5 group had a positive regression coefficient, and this coefficient was statistically significant. This shows that the mean for the Age 5 group was higher than the overall mean in other words, the Age 5 group was predicted to have a mean LMY 2.554 kg higher than the overall mean. Similar to dummy coding, the coefficient of determination (R2) was found to be 97.7% when effect coding was used and all variables were included in the model.

Deviation Coding: In order to understand whether the age affects LMY or not, this variable was coded using the deviation coding method and included in the model together with birth weight, lactation duration and average daily milk yield. The results of the regression analysis are presented in Table 3. As Table 3 shows, only lactation duration and average daily milk yield had statistically significant effects (P<0.001), along with the constant term, whereas the effects of other variables were not (statistically) significant. All variables with significant effects had positive coefficients; thus, a one-day increase in lactation duration was predicted to increase the mean LMY by 0.816 kg, and a 1 kg increase in average daily milk yield was predicted to increase the mean LMY by 0.131 kg. Similar to the case in which all variables were included in the model together, R^2 was found to be 97.6% (P<0.001).

To understand whether it affects LMY, the age variable was coded using a forward (and backward) difference coding approach and included in the model along with birth weight, lactation duration, and average daily milk yield. The results of the regression analysis are presented in Table 3, in which it can be seen that lactation duration, average daily milk yield, and age 5 group had statistically significant effects (P<0.05), along with the constant term, as the effects of other variables were not significant. Of the variables with significant effects, all except Age 5 had positive coefficients. Thus, a one-day increase in lactation duration was predicted to increase the mean LMY by 0.809 kg, and a 1 kg increase in average daily milk yield was predicted to increase the mean LMY by 0.131 kg. Regarding the age variable, the coefficient for the category of age 3 was found to be significant, and this coefficient represented the difference between the means of the age 5 and age 4 groups. When the effects of other variables included in the model were taken into consideration, the difference between the mean LMY of the Age 5 and Age 4 groups was 2.537 kg, and this difference was statistically significant.

Backward difference coding: In backward difference coding for age, the same values obtained in forward difference coding were obtained but with opposite signs. Aside from that, the coefficients were identical.

Table 3. Results of the multiple regression model for lactation milk yield with Deviation and Forward (and backward) difference coding for Age (kg).

	Deviation		Forward (and backward) difference	
Traits	**b ± SE**	**t (P)**	**b ± SE**	**t (P)**
Constant	-109.003 ± 5.826**	-18.711 (0.001)	-108.505 ± 6.085**	-17.833 (0.001)
Birth Weight	0.542 ± 0.690	0.785 (0.433)	0.534 ± 0.695	0.788 (0.433)
Ewe Weight	0.030 ± 0.081	0.374 (0.709)	0.044 ± 0.081	0.541 (0.589)
Lac. Dur.	0.816 ± 0.019**	43.704 (0.001)	0.809 ± 0.002**	43.703 (0.001)
ADMY	0.131 ± 0.002**	72.161 (0.001)	0.131 ± 0.002**	71.867 (0.001)
Age 3	-0.689 ± 1.308	-0.527 (0.599)	-0.514 ± 0.896	-0.574 (0.567)
Age 4	-0.031 ± 1.461	-0.021 (0.983)	-2.031 ± 1.219	-1.666 (0.097)
Age 5	2.972 ± 1.760	1.689 (0.093)	-2.537 ± 1.165*	-2.178 (0.030)
	R^2 = 0.976 F = 1523.075 P = 0.001			

SE: Standard error, F: F statistic, t (P): t statistic (P value), Lac. Dur.: Lactation Duration, b: Coefficient, ADMY: Average Daily Milk Yield, R^2: Determination coefficient, *: P<0.05 , **: P<0.01.

Table 4. Results of the multiple regression model for lactation milk yield with Helmert and Reverse Helmert coding for Age (kg).

Item	Helmert		Reverse Helmert	
	b ± SE	t (P)	b ± SE	t (P)
Constant	-108.504 ± 6.085**	-17.830 (0.001)	-110.930 ± 5.862**	-18.925 (0.01)
Birth Weight	0.534 ± 0.695	0.768 (0.433)	0.610 ± 0.689	0.885 (0.377)
Ewe Weight	0.044 ± 0.081	0.541 (0.709)	0.054 ± 0.080	0.676 (0.500)
Lac. Dur.	0.809 ± 0.019**	43.743 (0.001)	0.817 ± 0.018**	45.317 (0.001)
ADMY	0.131 ± 0.002**	71.867 (0.001)	0.131 ± 0.002**	73.380 (0.001)
Age 3	-0.515 ± 0.898	-0.574 (0.567)	-0.113 ± 0.726	-0.156 (0.876)
Age 4	-1.688 ± 0.920	-1.834 (0.068)	0.238 ± 0.880	0.270 (0.787)
Age 5	-1.522 ± 0.853	-1.783 (0.076)	2.385 ± 0.870**	2.742 (0.007)
	R^2 = 0.976 F = 1517.303 P = 0.001		R^2 = 0.976 F = 1517.303 P = 0.001	

SE: Standard error, F: F statistic, t (P): t statistic (P value), Lac. Dur.: Lactation Duration, b: Coefficient, ADMY: Average Daily Milk Yield, R^2: Determination coefficient, *: P<0.05 , **: P<0.01.

To understand its effects on LMY, the age variable was coded using the Helmert coding method and included in the model along with birth weight, lactation duration, and average daily milk yield. The results of the regression analysis are presented in Table 4, in which it can be seen that of the variables included in the model, only lactation duration and average daily milk yield had statistically significant coefficients (P<0.001), along with the constant term. Thus, a one-day increase in lactation duration was predicted to increase the mean LMY by 0.809 kg, and a 1 kg increase in average daily milk yield was predicted to increase the mean LMY by 0.131 kg. The model's coefficient of determination (R^2) was found to be 97.6%.

To understand whether it affects LMY, the age variable was coded using the Reverse Helmert coding method and included in the model along with birth weight, lactation duration, and average daily milk yield. The results of the regression analysis are presented in Table 4, in which it can be seen that lactation duration, average daily milk yield, and age 5 had statistically significant effects (P<0.05), along with the constant term, while the effects of other variables were not (statistically) significant. Thus, a one-day increase in lactation duration was predicted to increase the mean LMY by 0.817 kg, and a 1 kg increase in average daily milk yield was predicted to increase the mean LMY by 0.131 kg. Regarding the age variable, the coefficient for the category of age 3 was found to be significant. In reverse Helmert coding, this figure represents the difference between the mean LMYs of the Age 5 group and the other three groups (ages 4, 3, and 2). When the effects of the other variables included in the model were taken into consideration, the mean LMYs of the Age 5 group and the other three groups was 2.385 kg. The model's coefficient of correlation (R^2) was found to be 97.6%.

Discussion and Conclusion

In the present study, different coding methods involving categorical variables that affect LMY were used. When the dummy coding and effect coding methods were used, the effects of lactation duration, average daily milk yield, type of birth and age 5 were found to be statistically significant, in addition to the constant term. After these variables were included in the model, a coefficient of determination (R^2) of 97.7% was obtained. The birth type coefficient was found to be negative in both models. When the deviation coding method was used for age, the effects of lactation duration and average daily milk yield were found to be statistically significant, in addition to the constant term. When forward and backward coding methods were used, on the other hand, the effect of age 5 was also found to be statistically significant, along with the effects of lactation duration and average daily milk yield. Both coding methods had R^2 values of 97.6%. When the Helmert and Reverse Helmert coding methods were used, the Helmert coding method was found to result in statistically significant coefficients for the variables of lactation duration and average daily milk yield, along with the constant term. Similar to the forward and backward difference coding methods, the Reverse Helmert coding method resulted in a statistically significant coefficient for the age 5 group, but with a positive sign. As directly related to our study, there are no studies in which different coding systems are used to predict lactation milk yield in sheep in the literature. However, although not directly related, in a study examining the effect of age on lactation the milk yields of Karakaş sheep kept by villagers, Gökdal et al. (8) reported 5-year old ewes to have a lactation milk yield around 20 kg higher than that of 2-year old ewes, and the difference was statistically significant. The differences between the other groups, however, were not statistically significant. Altın (5) examined the effects of sheep breeds,

type of lambing and age on the real lactation milk yields of Akkaraman sheep and Hamdani x Akkaraman hybrids (F1), and found that none of the three factors developed statistically significant differences. For the lactation milk yields, about 10-liter difference between the ewes that gave birth single (52 liters) and twins (62 liters) was not found to be statistically significant. The real lactation milk yields of the ewe groups aged 1, 2, 3, 4, and 5 (or more) years were, respectively, 58, 55, 56, 70, and 47 liters, on average, and the differences between these values were not significant. In Yılmaz et al.'s (19) examination of the effects of age, type of birth and weight on the lactation milk yields of Norduz sheep, the lactation milk yields of ewe groups aged 2, 3, 4, and 5 years were, respectively, 107, 122, 130, and 142 kg. The authors reported that the differences were statistically significant, and that each group was different from the others. However, the 8-kilogram difference between the average lactation milk yield (121 kg) of the ewes that had single births and the average lactation milk yield (129 kg) of the ewes with twins was not found to be statistically significant. In a study examining the effects of type of birth, lambing season and mother's age on lactation milk yields involving 77 ewes, Allah et al. (4) found the 5-year old ewes to have the highest mean yield at 73 kg, and those older than 5 years to have a mean yield of 68 kg, representing a statistically significant difference. The ewe group aged 2, 3 or 4 had a mean yield of are 70 kg, and the differences between this group (2, 3 or 4 years) and the other two groups (5 years or older) were not found to be significant. Allah et al. (4) also reported that ewes who gave birth to single lambs had a lactation milk yield that was some 12 kg higher than those who gave birth to twins, but this difference was not statistically significant. In Erol et al.'s (7) study of the effect of lactation order and year on lactation milk yield in Ankara goats, the reported difference of approximately 20 kilograms between the average lactation milk yield (approximately 73 kg) of the animals in the first lactation and the average lactation milk yield (approximately 92 kg) of the animals in the second lactation to be statistically significant. It was emphasized, however, that the difference of approximately 10 kg between the average lactation milk yield of the animals in the third lactation and the milk yield of the animals in the second lactation (mean lactation) was not statistically significant. The results from models that included categorical variables coded using different coding systems were similar to the findings reported previously in literature. Of the previous studies in literature that included the age variable in their models without coding, some found the effect of age on lactation milk yield to be significant (4, 7, 19), whereas others reported no significance (5, 8). This was the case with respect also to birth type.

Programs such as SAS, SPSS, and R use dummy coding, whereas JMP uses effect coding. The last category in alphabetical order is the reference category in SAS and SPSS, but the last category gets a value of "-1" in STATA and JMP (3).

The effect coding method is very similar to dummy coding, with the last group being coded as "-1". As is the case with dummy coding, this coding method is not appropriate when the goal is to make contrasts, but in such situations, effect coding is easier to understand and interpret than dummy coding, although dummy coding is the simplest of the coding systems. In dummy coding, the newly created binary variables take on values of 0 or 1, while in effect coding, different values may be assigned to categorical variables. Dummy coding only uses the numbers 1 and 0, but effect coding also uses the numbers 16 and 17.

In Helmert coding, the mean of a given category is compared with the overall mean of the following categories. As these codes are orthogonal, the regression coefficients represent the difference between the weighted means. If a matrix approach is to be used in Helmert coding, the Helmert contrasts are entered into the columns. For k common variables, a matrix of "k+1" columns and "n" rows is needed. The entry in the first row of the first column is k, and all other entries in this column are "-1". In the second column, the first entry is 0, the second entry is k+1, and all other entries are "-1". In the third column, the first two entries are 0, the third entry is-2, and all other entries are "-1". This operation continues until the kth column.

The Reverse Helmert Coding method is also known as the Difference Contrasts method, as the order of entries is the reverse of Helmert coding. In deviation coding, the mean of a given group is compared with the overall mean of the other groups. For example, when there are four groups, the mean of the first group is compared with the mean of the remaining three groups; the mean of the second group is compared with the mean of the remaining three groups; and the mean of the third group is compared with the mean of the remaining three groups.

The present study exhibited and explained various codding systems in regression models. In addition, the study also examined the usability of various codding systems for categorical variables with continuous variables in the modeling of the lactation milk yields of Awassi sheep. The results of the study indicated that different results can be obtained depending on the various coding systems used. The results also indicated that the choice of coding system affected the interpretation of the obtained coefficients. Therefore, it can be stated that the aims of the researcher in the study should be defined clearly and the proper codding system should be selected according to the variables to be included in the model.

This study is expected to make a significant contribution to the literature based on its detailed examination of the different coding systems used in regression models.

Acknowledgments

This study is a summary of the first author's PhD thesis.

Financial Support

Şanlıurfa GAP Agricultural Research Institute of the General Directorate of Agricultural Research and Policies of the Ministry of Agriculture and Forestry (TAGEM/11/08/01/01).

Ethical Statement

This study does not present any ethical concerns.

Conflict of Interest

The authors declare no conflict of interest.

References

1. **Aguinis H, Pierce CA** (2006): *Computation of effect size for moderating effects of categorical variables in multiple regression*. Appl Psychol Meas, **30**, 440–442.
2. **Akçapınar H, Özbeyaz C** (1999): Hayvan Yetiştiriciliği Temel Bilgileri, Kariyer Matbaacılık, Ankara.
3. **Alkharusi H** (2012): *Categorical variables in regression analysis: a comparison of dummy and effect coding*. Int J Educ, **4**, 202-210.
4. **Allah MA, Abass SF, Allam FM** (2011): *Factors affecting the milk yield and composition of Rahmani and Chios sheep*. Int J Livest Prod, **2**, 24-30.
5. **Altın T** (2001): *Koyunlarda süt veriminin laktasyon boyunca değişimi ve farklı yöntemlere göre tahmin edilmesi*. J Agric Sci, **11**, 1-7.
6. **Bruin, J** (2006): Newtest: command to compute new test. ucla: statistical consulting group. Available at https://stats.idre.ucla.edu/stata/ado/analysis/. (Accessed October 05, 2018).
7. **Erol H, Akçadağ Hİ, Ünal N, et al** (2012): *Ankara keçilerinde süt verimi ve oğlaklarda büyümeye etkisi*. Ankara Univ Vet Fak Derg, **59**, 129-134.
8. **Gökdal Ö, Ülker H, Oto MM, et al** (2000): *Köylü koşullarında yetiştirilen Karakaş koyunlarının çeşitli verim özellikleri ve vücut ölçüleri*. J Agric Sci, **10**, 103-111.
9. **Haenlein GFW, Wendorff LW** (2006): Sheep milk.137–194. In: Park WY, Haenlein GFW, editors. Handbook of Milk of Non-Bovine Mammals. Blackwell Publishing; Iowa.
10. **Kayaalp GT, Güney Ç, Cebeci Z** (2015): *Çoklu doğrusal regresyon modelinde değişken seçiminin zootekniye uygulanışı*. Çukurova J Agric Food Sci, **30**, 1-8.
11. **Keskin S, Boysan M, Göktaş I** (2008): *Multivariate analysis approach to relationships between perfectionism and obsessive compulsive symptoms*. Turkiye Klinikleri J Med Sci, **28**, 319-326.
12. **Keskin S** (2002): Varyansların homojenliğini test etmede kullanılan bazı yöntemlerin 1. Tip hata ve testin güç bakımından irdelenmesi. Doktora tezi. Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.
13. **Keskin S, Özsoy AN** (2004): *Kanonik korelasyonan alizi ve bir uygulaması*. J A S, **10**, 67-71.
14. **Maran BF** (2016): Çoklu regresyon analizinde kategorik değişkenlerin kullanımı. Yüksek Lisans Tezi. Yüzüncü Yıl Üniversitesi Sağlık Bilimleri Enstitüsü, Van.
15. **Sundström S** (2010): Coding in multiple regression analysis: A review of popular coding techniques. Department of Mathematics Uppsala University. Project Report 2010, 14.
16. **Wendorf CA** (2004): *Primer on multiple regression coding: Common forms and the additional case of repeated contrasts*. Underst Stat, **3**, 47-57.
17. **Wissmann M, Toutenburg H, Shalab H** (2007): *Role of categorical variables in multicollinearity in the linear regression model*. Technical Report Number, Department of Statistics University of Munich, **8,** 1–34.
18. **Yalta T** (2011): Kukla değişkenlerle bağlanım, Ekonometri ders notları. TÜBA (Türkiye Bilimler Akademisi) Açık Ders Malzemeleri Projesi, TOBB, ETU. 184p. Ankara.
19. **Yilmaz O, Denk H, Nursoy H** (2004): *Milk yield characteristics of Norduz sheep*. Van Vet J, **15**, 27-31.