International Research Journal of Mathematics, Engineering and IT
Vol. 3, Issue 6, June 2016 IF- 3.563 ISSN: (2349-0322)
© Associated Asia Research Foundation (AARF)
Website: www.aarf.asia Email : editor@aarf.asia , editoraarf@gmail.com

EXAMINATION OF MULTIVARIATE MULTIPLE LINEAR REGRESSION ANALYSIS[1]

Nazlı Elif GÜNAŞDI, Mehmet TOPAL*

Biometry and Genetics Unit, Department of Animal Science, Faculty of Agriculture,

Ataturk University, 25240 Erzurum,Turkey.

ABSTRACT

Regression analysis is a statistical method determining the functional relationship between dependent (Y_1, Y_2, . . . , Y_q) and independent (X_1, X_2, . . . , X_p) variables. In regression model, if there are one dependent (Y_1) and one independent (X_1) variables, the simple liner regression is used, if there are one dependent variable (Y_1) and more than one independent variables (X_1, X_2, . . . , X_p), multiple linear regression model is used, and if there are more than one dependent (Y_1,Y_2, . . . , Y_q) and more than one independent variables (X_1, X_2, . . . , X_p) multivariate multiple linear regression model is used. The fundamental purpose of regression analysis is to determine the best model in order to predict the dependent variable or variables. Besides canonical correlation used for analyzing the relationship among data sets, the function obtained by multivariate multiple linear regression analysis can also determine the effect of which independent variable on dependent variable. Multivariate multiple linear regression models were formed when body weight (LW) and cold carcass weight (CCW) were dependent variable, chest depth (CD), height at withers (HAW), pump width (PW), forehead length (FL), and head width (HW) were independent variable. Canonical correlation coefficient and multiple coefficient of determination R_M^2 have been used in order to determine the relationship between dependent and independent variables.

[1] This paper was presented as poster presentation at the VII. Balkan Conference on Animal Science (BALNIMALCON-2015) Sarajevo

*: **Corresponding author:** Mehmet TOPAL.

Atatürk University, Faculty of Agriculture, Departmant of Animal Science, Biometry and Genetics Unit, Erzurum, Turkey.

Key Words: Multivariate multiple linear regression, Awassi sheep, Multiple coefficient of determination

INTRODUCTION

The regression analysis was used to determine the functional relation between two or more variables that have a cause-effect relation, and to be able to make estimations about the topic by using this relation. The mathematical model which is used to explain the functional relation between the variables in the regression analysis is called the regression Model (Özdamar 2004; Alpar 2011). Regression is a functional relation and defines which way and to what extent the change in the independent variables affects the dependent variables. The basic purpose of the regression is finding the mathematical equation which expresses the functional relation between the variables in the best way and using this equation in estimating the values of the dependent variables in statistical analyses (Yıldız and Bircan, 1994; Neter et al. 1996). If the number of the dependent and independent variables in the regression equation is one, the simple Regression model is formed; if there are one dependent and more than one independent variables, the multivariate regression model is formed; and if the number of the dependent and independent variables is more than one, the multivariate multiple regression model is formed.

In this study, the multivariate multiple linear regression models between the variables of the body weight, cold carcass weights, chest depth, height at withers, pump width, forehead length, and the head width of the Awassi sheep has been determined.

MATERIALS AND METHODS

The data from a study conducted in 2013 at the Research and Application Farm of Atatürk University have been used in this study. The body weight and cold carcass weights of the Awassi sheep have been taken as the dependent variables; and chest depth, height at withers, pump width, forehead length, and width of the head were taken as the independent variables; and the multivariate multiple linear regression analysis was used in determining the functional relation between the dependent and independent variables.

Multivariate Multiple Linear Regression

In multivariate multiple linear regression model is used in examining the linear relation between the dependent and independent variable sets when the number of dependent and independent variables is more than one (Dattalo 2013). In order to the multivariate multiple

linear regression to be applied, the distribution of the dependent variables must fit the multivariate normal distribution, the sampling must be made based on chance, and and also there must not be multicollinearity among themselves and between the dependent and independent variables, and the variance-covariance matrix must be homogenous. The multivariate multiple linear regression equation is written in the matrix notation as follows;

$$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$$

the regression coefficients matrix $\boldsymbol{\beta}$ is estimated as follows,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{I}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{I}}\mathbf{Y})$$

Here, $\widehat{\boldsymbol{\beta}}$; is the least squares estimator of the regression coefficients matrix $\boldsymbol{\beta}$, $\boldsymbol{X}^{'}$; is the transpose of the independent variable matrix. As the first column of the $\widehat{\boldsymbol{\beta}}$ matrix gives the regression coefficients of the $Y_1,\ldots, Y_q$ over the $X_1,\ldots,X_p$ variables.

Finding the coefficients in multivariate multiple linear regression with the matrix system

In a sample where there are **n** observations, if the number of the dependent variable is **q**, and the number of the independent variable is **p**; then the regression models for each dependent variable may be written as follows (Quick 2013; Dattalo 2013);

$$Y_{i1} = \beta_{01} + \beta_{11}X_{i1} + \beta_{21}X_{i2} + \cdots + \beta_{p1}X_{1p} + \varepsilon_{i1}$$
$$Y_{i2} = \beta_{02} + \beta_{12}X_{i1} + \beta_{22}X_{i2} + \cdots + \beta_{p2}X_{1p} + \varepsilon_{i2}$$
$$\vdots$$
$$Y_{iq} = \beta_{0q} + \beta_{1q}X_{i1} + \beta_{2q}X_{i2} + \cdots + \beta_{pq}X_{1p} + \varepsilon_{iq}$$

$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$**;** the multivariate multiple linear regression equation in matrix notation written as;

$$\begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1q} \\ Y_{21} & Y_{22} & & Y_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nq} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & & x_{2p} \\ & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0q} \\ \beta_{11} & \beta_{12} & & \beta_{1q} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pq} \end{bmatrix} +$$

$$\begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1q} \\ \varepsilon_{21} & \varepsilon_{22} & & \varepsilon_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nq} \end{bmatrix}$$

where, Y: $n \times q$ dimensional dependent variable matrix; X: $n \times (p+1)$ dimensional independent variable matrix; β: $(p+1) \times q$ dimensional coefficient matrix; ε: $n \times q$ dimensional error matrix (Rencher 2002).

In the significance test of the multivariate multiple linear regression coefficients, the null hypothesis is established as if there is no linear relation between the dependent and independent variable sets or as if all the slope coefficients in the β matrix are equal to null (Dattalo 2013). The Wilks ∧ statistics is used in testing the significance of multivariate multiple linear regression coefficients (Rencher 2002).

The most frequently used measure of association in determining the relation between the dependent and independent variables is the canonic correlation (Çankaya 2005; Timm 2002; Johnson and Wichern 2002). In addition, the multiple coefficient of determination (R_M^2) may also be used in determining the relation between the variables (Rencher 2002).

RESULT AND DISCUSSION

The live weight and cold carcass weights, chest depths, height at withers, thigh circumference, forehead length, and the width of the head of the Awassı Sheep have been used in the study. The body weight (LW) and cold carcass weights (CCW) have been taken as the dependent variable; and the chest depths (CD), height at withers (HAW), pump width (PW), forehead length (FL), and the head width (HW) have been taken as the independent variable. In order to the multivariate multiple linear regression analysis to be applied, the dependent variable matrix which is needed **(Y)** and the independent variable matrix **(X)** have been formed as follows. There is the live weight in the first column and the cold carcass weight values in the second column in the **Y** matrix. In the first column of the **X** matrix, there is the column vector which consists of the fixed value **1**, and the chest depth is in the second column, the height at withers in the third column, the pump width in the fourth column, the forehead length in the fifth column, and the head width in the sixth column.

$$\mathbf{Y} = \begin{bmatrix} 62{,}0 & 27{,}0 \\ 59{,}0 & 26{,}2 \\ 58{,}5 & 27{,}6 \\ 58{,}0 & 26{,}2 \\ 58{,}0 & 28{,}1 \\ 62{,}0 & 28{,}6 \\ 60{,}0 & 27{,}6 \\ 55{,}5 & 25{,}6 \\ 58{,}0 & 27{,}8 \\ 58{,}5 & 26{,}2 \end{bmatrix} \text{ and } \mathbf{X} = \begin{bmatrix} 1{,}0 & 45{,}0 & 71{,}0 & 22{,}0 & 8{,}0 & 15{,}0 \\ 1{,}0 & 44{,}0 & 76{,}0 & 21{,}5 & 9{,}0 & 11{,}0 \\ 1{,}0 & 42{,}0 & 72{,}0 & 20{,}0 & 8{,}0 & 10{,}0 \\ 1{,}0 & 40{,}0 & 69{,}0 & 22{,}0 & 7{,}0 & 9{,}0 \\ 1{,}0 & 40{,}0 & 70{,}0 & 20{,}0 & 7{,}0 & 8{,}0 \\ 1{,}0 & 41{,}0 & 78{,}0 & 22{,}0 & 7{,}0 & 9{,}0 \\ 1{,}0 & 41{,}0 & 73{,}0 & 20{,}5 & 8{,}0 & 9{,}0 \\ 1{,}0 & 40{,}0 & 70{,}0 & 23{,}0 & 8{,}0 & 9{,}0 \\ 1{,}0 & 39{,}0 & 71{,}0 & 20{,}0 & 10{,}0 & 10{,}0 \\ 1{,}0 & 42{,}0 & 72{,}0 & 21{,}0 & 7{,}0 & 9{,}0 \end{bmatrix}$$

The inverse value of the $\mathbf{(X'X)}$ is taken by multiplying the $\mathbf{X'}$ matrix and the $\mathbf{X}$ matrix.

$$(\mathbf{X^tX})^{-1} = \begin{bmatrix} 222{,}4620 & -3{,}8231 & -0{,}0478 & -3{,}0996 & -3{,}9281 & 3{,}6470 \\ -3{,}8231 & 0{,}1524 & -0{,}0346 & 0{,}0281 & 0{,}0934 & -0{,}1333 \\ -0{,}0478 & -0{,}0346 & 0{,}0220 & -0{,}0103 & -0{,}0245 & 0{,}0308 \\ -3{,}0996 & 0{,}0281 & -0{,}0103 & 0{,}1246 & 0{,}0632 & -0{,}0466 \\ -3{,}9281 & 0{,}0934 & -0{,}0245 & 0{,}0632 & 0{,}2017 & -0{,}1115 \\ 3{,}6470 & -0{,}1333 & 0{,}0308 & -0{,}0466 & -0{,}1115 & 0{,}1529 \end{bmatrix}$$

The $\mathbf{(X'Y)}$ is obtained by multiplying the transpose of the X matrix and the Y matrix.

$$\mathbf{(X^tY)} = \begin{bmatrix} 590 & 271 \\ 24424 & 11212 \\ 42591 & 19568 \\ 12497 & 5738 \\ 4654 & 2140 \\ 5854 & 2680 \end{bmatrix}$$

The regression coefficients matrix is obtained as follows by using the $\hat{\beta} = (X'X)^{-1}(X'Y)$ equation;

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 61{,}2391 & 49{,}6927 \\ -0{,}7857 & -0{,}7293 \\ 0{,}6072 & 0{,}3137 \\ -0{,}7588 & -0{,}7753 \\ -1{,}4763 & -0{,}6913 \\ 1{,}4286 & 0{,}6906 \end{bmatrix}$$

According to this, the regression equations as follows;

BW= 61,2391 -0,7857CD + 0.6072HAW -0.7588PW – 1,4763FL +1,4286HW

CCW= 49,6927 -0,7293CD + 0.313HAW -0.7753PW – 0,6913FL +0,6906HW

The HAW and HW have statistically positive effects at a very significant level on the body weight;, and the CD, PW and FL have negative significant effects. For this reason, an increase in the HAW and HW values lead to an increase in the body weight; and an increase in CD, PW and FL variables lead to decrease in body weight. When the beta coefficients are examined, it is observed that the most significant effect was made by the HW, HAW, CD, FL and PW, respectively.

It may be suggested that the HAW and HW had statistically significant effects at extremely important level on cold carcass weight; and the CD, PW and FL have significant negative effects and therefore there is a linear relation between the HAW and HW values, and it may be claimed that there is a reverse relation among the CD, PW and FL variables in cold carcass weight. When the beta coefficients are examined, it may be observed that the most significant effect was made by HW, CD, HAW, PW and FL, respectively. The error matrix **ε (ε = Y-$\widehat{Y}$)** is obtained by taking the difference between the observed dependent variables matrix **(Y)** and the expected values matrix ($\widehat{Y}$) obtained according to the applied models. The error sum of squares (ESS) and mean square error (MSE) of each model have been found to be

ESS=2,4135 and MSE= 0,6034 for CA

ESS= 0,7163 and MSE=0,1791 for SKA

The error values being small is a display of the explainer of the real values of the adopted model. The mean square error obtained by taking the sum of squares of the error values and dividing them by error values is used as a criterion in comparing the models. If the mean square error is low, the model adopted is accepted as being good. When the models are compared, it may be suggested that the model obtained for cold carcass weight is more explanatory than the model obtained for body weight.

The most frequently used measure of association in determining the relation between the dependent and independent variables is the canonic correlation. The canonic correlation coefficient between the dependent and independent variables has been found to be 0.96 and as being statistically significant (P<0.05). According to this situation, it can be suggested that the linear relation between the two variable groups is very important..

The R_M^2 relation coefficient, which is calculated by breaking the multi identification coefficient and covariance matrix may also be used in determining the relation between the variables. The R_M^2 relation coefficient may be calculated as follows

$$R_M^2 = \frac{|S_{YX}\, S_{XX}^{-1} S_{XY}|}{|S_{YY}|}$$

In the equation, $|S_{YX}S_{XX}^{-1}S_{XY}|$; S_{YX} , S_{XX}^{-1} , S_{XY} are the determinant of the multiplication of the matrices, and the determinant of the $|S_{YY}|$; S_{YY} matrix. The covariance matrix of the dataset has been found as;

$$\mathbf{S} = \begin{bmatrix} 3{,}858 & 1{,}038 & 2{,}078 & 3{,}233 & 0{,}011 & -0{,}283 & 1{,}994 \\ 1{,}038 & 0{,}992 & -0{,}384 & 1{,}002 & -0{,}598 & -0{,}023 & -0{,}234 \\ 2{,}078 & -0{,}384 & 3{,}600 & 1{,}800 & 0{,}411 & 0{,}044 & 2{,}933 \\ 3{,}233 & 1{,}002 & 1{,}800 & 7{,}956 & 0{,}233 & 0{,}133 & 0{,}133 \\ 0{,}011 & -0{,}598 & 0{,}411 & 0{,}233 & 1{,}122 & -0{,}256 & 0{,}467 \\ -0{,}283 & -0{,}023 & 0{,}044 & 0{,}133 & -0{,}256 & 0{,}989 & 0{,}656 \\ 1{,}994 & -0{,}234 & 2{,}933 & 0{,}133 & 0{,}467 & 0{,}656 & 3{,}878 \end{bmatrix}$$

$$\mathbf{S_{YY}} = \begin{bmatrix} 3{,}858 & 1{,}038 \\ 1{,}038 & 0{,}992 \end{bmatrix}; \qquad \mathbf{S_{YX}} = \begin{bmatrix} 2{,}078 & 3{,}233 & 0{,}011 & -0{,}283 & 1{,}994 \\ -0{,}384 & 1{,}002 & -0{,}598 & -0{,}023 & -0{,}234 \end{bmatrix}$$

$$\mathbf{S_{XY}} = \begin{bmatrix} 2{,}078 & -0{,}384 \\ 3{,}233 & 1{,}002 \\ 0{,}011 & -0{,}5980 \\ -0{,}2830 & -0{,}023 \\ 1{,}9940 & -0{,}234 \end{bmatrix}; \quad \mathbf{S_{XX}} = \begin{bmatrix} 3{,}600 & 1{,}800 & 0{,}411 & 0{,}044 & 2{,}933 \\ 1{,}800 & 7{,}956 & 0{,}233 & 0{,}133 & 0{,}133 \\ 0{,}411 & 0{,}233 & 1{,}122 & -0{,}256 & 0{,}467 \\ 0{,}440 & 0{,}133 & -0{,}256 & 0{,}989 & 0{,}656 \\ 2{,}933 & 0{,}133 & 0{,}467 & 0{,}656 & 3{,}878 \end{bmatrix} \mathbf{S_{YX}S_{XX}^{-1}S_{XY}}$$

$$= \begin{bmatrix} 3{,}5903 & 1{,}0648 \\ 1{,}0648 & 0{,}9135 \end{bmatrix} \qquad |S_{YX}\, S_{XX}^{-1} S_{XY}| = 2.1459 \text{ and } |S_{YY}| = 2.7497$$

$$R_M^2 = \frac{2.1459}{2.7497} = 0.7804$$

R_M^2 and the value as 0.7804, and this value being closer to 1 shows that the relation between the dependent and independent variables is significant. In other words, it may be suggested that the 78.04% of the variation in the dependent variables is explained by the independent variable. In testing the significance of the multivariate multiple linear regression coefficients, the hypotheses are established as H_0:$\boldsymbol{\beta}_1$=**0** and H_1:$\boldsymbol{\beta}_1 \neq$**0**, with **0** matrix. The Wilks ∧ value is calculated as following;

$$\wedge = \frac{|\boldsymbol{S}|}{|\boldsymbol{S}_{XX}||\boldsymbol{S}_{YY}|} = \frac{0.3714}{18.2993 * 2.7497} = 0.0074$$

Since the ∧=0.0074 < $\wedge_{0.01,2,5,4}$=0.017, the H_0 hypothesis is rejected and it is claimed that the regression coefficients are significant. In this context, it is concluded that the multivariate multiple linear models which are suggested explain the body weight increase and cold carcass

weight at a significant level. Cam et al. (2010) found higher correlations between body weight and cold carcass weight in Karayaka sheep.

CONCLUSION

It may be concluded that the multivariate multiple linear regression method is a proper method in examining the functional relation between data sets. As well as the canonic correlation which is used in examining the relation between the datasets, the function, which is obtained with the multivariate multiple linear regression method, may also be used in determining the effect of a variable on the dependent variable.

REFERENCES

Alpar, R., 2011. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. 3.Baskı, Detay Yayıncılık, 405-608 s, Ankara.

Cam, M.A., Olfaz,M., Soydan, E. (2010). Body Measurements Reflect Body Weights and carcass Yields in Karayaka Sheep. Asian Journal of Animal and Veterinary Advances, 5(2), 120-127.

Çankaya, S., 2005. Kanonik Korelasyon Analizi ve Hayvancılık Kullanımı. Doktora Tezi, Fen Bilimleri Enstitüsü, Adana.

Dattalo, P., 2013. Analysis of Multiple Dependent Variables. Oxford University, 180 p, Oxford.

Johnson, A.R. and Wichern, D.W., 2002. Applied Multivariate Statistical Analysis: Canonical Corelation Analysis. Fifth Edition, Prentice Hall, New Jersey, 543-580

Neter J., Kunter M.H., Nachtsheim C.J. ve Wasserman W., 1996. Applied Linear Regression Models. The Mc Graw-Hill Companies, Inc., 206 p, Chicago.

Özdamar, K., 2004. Paket Programlar ile İstatistiksel Veri Analizi. Kaan Kitapevi, 649 s, Eskişehir.

Quick, C., 2013. Multivariate Multiple Regression with Applications to Powerlifting Data. 48 p, Minnesota.

Rencher, A.C., 2002. Methods Multivariate Analysis. 2. Baskı, John Wiley & Sons, Inc. 322-361s, Canada,

Tımm N.H., 2002. Applied Multivariate Analysis. Springer-Verlag, Inc. 693 p, New York.

Yıldız, N. ve Bircan, H., 1994. Uygulamalı İstatistik. 4. Baskı, Atatürk Üniversitesi Yayınları No:704, 174-183 s, Erzurum.