Zootekni / *Animal Science* **DOI: 10.21597/jist.671662**
Araştırma Makalesi / *Research Article*
Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 10(2): 1429-1437, 2020
Journal of the Institute of Science and Technology, 10(2): 1429-1437, 2020
ISSN: 2146-0574, eISSN: 2536-4618

Comparison of Ridge Regression and Least Squares Methods in the Presence of Multicollinearity for Body Measurements in Saanen Kids

Cem TIRINK[1*], Samet Hasan ABACI[2], Hasan ÖNDER[2]

ABSTRACT: Least square (LS) method is a common method used to estimate the coefficients in multiple regression models. The least square multiple regression models produce biased regression coefficients when the multicollinearity is encountered in the studied data sets. Multicollinearity problem can be solved by using some methods. As one of the methods, Ridge Regression (RR) is a biased estimation method that enables to obtain models having more reliable coefficient of determination (R^2). This study was conducted on 40 Saanen kids in order to determine some morphological measurements (withers height, rump height, body length, chest width, chest girth and chest depth) affecting body weight. In this study, usability of ridge regression method in the presence of multicollinearity was evaluated. Variance Inflation Factor (VIF) values higher than 10 were detected for withers height and rump height. Coefficient of determination (R^2) was obtained as 0.88 from LS method and R^2 was obtained 0.875 with k=0.0136 from RR method. As a result, the model obtained from RR is more reliable than that obtained from LS.

Keywords: Least squares, multiple regression, multicollinearity, ridge regression

[1] Cem TIRINK (**Orcid ID:** 0000-0001-6902-5837), Igdir University, Faculty of Agriculture, Department of Animal Science, Igdir, Turkey,
[2] Samet Hasan ABACI (**Orcid ID:** 0000-0002-1341-4056), Hasan ÖNDER (**Orcid ID:** 0000-0002-8404-8700), Ondokuz Mayis University, Faculty of Agriculture, Department of Animal Science, Samsun, Turkey

*Sorumlu Yazar/Corresponding Author: Cem TIRINK, e-mail: cem.tirink@gmail.com

This research was presented as an oral presentation at the International 8th Balkan Animal Science Conference (BALNIMALCON) held on 6-8 September 2017 in Prizren.

Geliş tarihi / *Received:* 07-01-2020
Kabul tarihi / *Accepted:* 01-02-2020

INTRODUCTION

Body weight prediction provides a great convenience for the breeders to make the right decision for determining suitable feed portion, medical dose and correct market price of farm animals. (Eyduran et al., 2017). Therefore, in the case of the scarcity of weighing scales in rural conditions, it is an important and easy way to predict body weight to breeders. It is a very important strategy for determining the relationship between body weight and body measurement for the different viewpoints of animal breeding (Aytekin et al., 2018; Lukuyu et al., 2016). A great number of statistical methods can be used for determining the relationship between body weight and various body measurements.

To reveal the cause and effect relationship of economic characters that vary depending on many factors in animal husbandry, there are a lot of statistical methods to execute. One of these statistical methods is multiple regression analysis (Akcay and Sariozkan, 2015). Regression analysis can be explained as a function between interested response variable and explanatory variables thought to be related on response (Ari and Onder, 2013). Least square method (LS) is a common method to estimate parameters in the regression model (Uckardes et al., 2012). Besides, the LS method is an unbiased method that is not only estimate parameter but also minimizing the error of the model. However, the LS method needs some assumptions which should be provided for the model reliable. If assumptions aren't provided, the reliability of the model will decrease. Therefore, it will cause misinterpretations. To guarantee the usability of this method, the assumptions must be valid such as that the errors are independent and normally distributed, and independent among explanatory variables (Uckardes et al, 2012). If the linear relationship (multicollinearity) exists between the explanatory variables, this situation causes to occur wrong estimates of the regression coefficients and it reduces the model's predictability.

The aim of this study was to compare the ridge regression and the least squares methods to estimate body weight from withers height, rump height, body length, chest depth, chest width and chest girth parameters in the case of multicollinearity.

MATERIALS AND METHOD

Material

In this study, 40 Saanen kids at Ondokuz Mayis University research farm unit were used for predicting body weight. For this aim, some body measurements (withers height, rump height, body length, chest depth, chest width, chest girth, rump width) and body weight were used. Body weight were used as a response variable, and body measurements were used as explanatory variables. All statistical analysis was performed using IBM SPSS 21.0 via Ondokuz Mayis University license (IBM, 2012) and NCSS trial version (NCSS, 2016).

Method

Regression analysis is the most common statistical method used to explain the relationship between response and explanatory variables. In the matrix form of the equation to multiple regression is:

$$Y = X\beta + \varepsilon \quad (1)$$

Y is a vector (n×1) of response variable (observations), X is a n×(p+1) matrix of explanatory variables and β is a vector (p+1)×1 which is estimated regression coefficients (Alpar, 2011). The error terms in the ε have n×1 dimension of a vector that have a normal distribution with $E(\varepsilon)=0$ and $Var(\varepsilon)=\boldsymbol{I}\sigma^2$.

Least squares method

The main purpose of least squares method (LS) is to minimize the sum of squares of error terms, in case of error terms having a normal distribution and having homogeneous variance and thus to optimize the model (Kutner et al., 2004).

$$Q_{LS} = \sum_{i=1}^{n} e_i^2 \quad (2)$$

In estimating regression coefficients, LS is generally used. In the calculation of the coefficients vector $\hat{\beta}$, the following equation is used for the LS method (Alpar, 2011).

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (3)$$

The LS method requires certain assumptions to obtain reliable results. The first assumption is that the regression model must be expressed in a linear way. The expected average error of the regression model must be zero. The variance of the errors must be constant and the errors must be independent (no autocorrelation) (Sarstedt and Mooi, 2014).

The LS method needs to require some assumptions such as the absence of a linear relationship between explanatory variables. The model's reliability depends on the realization of the assumptions of the LS method. If there is a linear relationship between the explanatory variables, this problem called a multicollinearity. In the case of multicollinearity, the variance and covariance of the regression coefficients increase, although the R^2 value of the model is influenced, none or some of the independent variables will be significant. Thus, the model will be misinterpreted in case of the multicollinearity (Cankaya et al., 2019).

- **Determination approaches of multicollinearity**
  - *Simple correlation coefficient*

Determination of the multicollinearity, if the correlation coefficient between the explanatory variables is close to 1, there may be interpreted as multicollinearity is exist (Albayrak, 2005; Sahin et al., 2018).

  - *Determination of coefficient*

The second method for determining the multicollinearity can be by examining the changes in R^2. If there is no change in R^2 as a result of adding an independent variable or observation to the model, this may indicate multicollinearity (Sahin et al., 2018).

  - *Partial correlation coefficient*

One approach to determining multicollinearity is to analyze the partial correlation coefficients. If the simple correlation coefficient between the explanatory variables is significant and the partial correlation coefficient is insignificant, this can be demonstrated for the multicollinearity problem. However, the approach based on the partial correlation coefficient is not always effective. On the other hand, even if the partial correlation coefficients are high, problems of multicollinearity can still arise (Albayrak, 2005).

- *Tolerance value*

Another approach in determining of multicollinearity is calculated to tolerance value for explanatory variables. Tolerance value can be calculated by the following equation:

$$\mathrm{TV} = 1 - \mathrm{R}_j^2 \quad (4)$$

If the tolerance value is small (TV<0.10), this means a larger VIF value (VIF>10), which means it has multicollinearity (Hair et al., 2014; Albayrak, 2005; Sahin et al., 2018).

- *Variance Inflation Factor (VIF)*

The fifth approach to determining of multicollinearity is the use of Variance Inflation Factor (VIF). If the VIF value is greater than 10, multicollinearity can be mentioned (Topal et al., 2010; Albayrak, 2005).

$$\mathrm{VIF} = \mathrm{c}_{ij} = \frac{1}{1-\mathrm{R}_j^2} \quad (5)$$

- *Condition number (CN) and Condition Index (CI)*

One of the methods to the determination of multicollinearity is to calculate the condition number (CN) and condition index (CI).

$$\mathrm{CN} = \frac{\lambda_{max}}{\lambda_{min}} \quad (6)$$

$$\mathrm{CI} = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \quad (7)$$

If the CI is under 10 (CI<10), there is no multicollinearity problem in this model. Furthermore, it is possible to mention that multicollinearity for both 10≤CI ≤30 (medium-multicollinearity) and CI≥30 (high-multicollinearity) (Pagel and Lunneborg, 1985; Gujarati, 1995; Rathert et al., 2011; Sahin et al., 2018; Akcay and Sariozkan, 2015). If the CN is under 100, there is no multicollinearity problem in this model. Furthermore, it is possible to mention that multicollinearity for both 100≤CN≤1000 (medium-multicollinearity) and CN≥ 1000 (high-multicollinearity) (high-multicollinearity) (Akcay and Sariozkan, 2015).

Ridge regression

Ridge Regression (RR) was proposed to analyze multivariate data that suffer from multicollinearity by Hoerl and Kennard (1970). In the case of multicollinearity, the variance and covariance of the regression coefficients increases in the $X'X$ matrix (Vupa ve Gurunlu Alma, 2008; Uckardes et al., 2012). In order to eliminate this problem, variance and covariance can be decreased by adding ridge trace (k) parameter to the diagonal elements of $X'X$ matrix. Ridge trace value should be $0 \leq k \leq 1$. If the k is zero, parameter estimation is the same as LS (Uckardes et al., 2012). Parameter estimation of RR in matrix notation can be given as;

$$\hat{\beta} = (\mathrm{X'X} + \mathrm{kI})^{-1}\mathrm{X'Y} \quad (8)$$

The main purposes of ridge regression method are eliminating the multicollinearity between explanatory variables, getting a smaller variance prediction than LS, decreasing mean square error and display the instability that occurs in coefficients on graphs in case of the multicollinearity (Cankaya et al., 2019).

Many researchers suggested different equations to calculate optimum k value. Kurtulus (2001) have suggested an equation to calculate the optimum k value based on eigenvalue and obtained the following equation:

$$k \leq \frac{\lambda_{max} - 100\lambda_{min}}{99}, k \neq 0 \quad (9)$$

By calculating the optimal k value using this equation, the VIF value approaches to 1 (Anderson, 1998; Uckardes et al., 2012; Cankaya et al., 2019).

RESULTS AND DISCUSSION

Descriptive statistics for response and explanatory variables are given in Table 1. When the data is examined, the coefficient of variation is less than 30% and is reliable. The Kolmogorov Smirnov normality test yield that the data of the variables examined were compatible with the normal distribution (P> 0.05).

Table 1. Descriptive Statistics

	Mean±Std. Deviation	Min-Max	CV (%)
Body weight	16.18 ± 3.22	9.39-23.6	19.90
Chest girth	56.06 ± 4.39	42.5-66	7.81
Chest width	9.88 ± 1.15	8.5-13	11.65
Chest depth	17.77 ± 1.64	13.5-22	9.23
Withers height	49.67 ± 3.63	42.5-60	7.33
Rump height	51.01 ± 3.63	43-61	7.12
Body length	46.27 ± 4.12	36.5-55	8.91

Pearson correlation coefficients and significance test results between body weight and some body measurements taken from Saanen kids are given in Table 2. There is a positive correlation between body weights and live weights of Saanen kids during the weaning period. The highest correlation was found between rump height and withers height (r=0.98, P<0.01). If the correlation coefficient between the variables is greater than 90%, there may be multicollinearity problems (Topal et al., 2010).

Table 2. Correlation matrix

	Chest girth	Chest width	Chest depth	Withers height	Rump height	Body length
Chest width	0.67**					
Chest depth	0.83**	0.56**				
Withers height	0.56**	0.36*	0.70**			
Rump height	0.56**	0.40*	0.73	0.98**		
Body length	0.74**	0.66**	0.71**	0.63**	0.65**	
Body weight	0.87**	0.73**	0.84**	0.70**	0.73**	0.80**

**P<0.01

Results of the Least Squares

Estimated regression coefficients, standard error, test statistics and collinearity statistics (tolerance and VIF values) obtained from the LS method are given in Table 3. According to the results of multiple regression analyses using the LS method, the regression coefficients of chest girth and chest width were found to be statistically significant (P<0.05). However, chest depth, withers height, rump height and body length were found to be statistically insignificant (P>0.05). In addition, the multicollinearity problem was determined between explanatory variables withers height and rump height (VIF>10) in

Table 3. Accordingly, these results showed that standard error increased, so that predicted regression coefficients with the LS method was inconsistent parameter predictions.

Table 3. Regression coefficients and collinearity statistics obtained from the LS method

	β	Std. Error	t	Sig.	Collinearity Statistics	
					Tolerance	VIF
Constant	-25.684	3.142	-8.175	0.000	-	-
Chest girth	0.279	0.095	2.937	0.006	0.217	4.608
Chest width	0.584	0.252	2.316	0.027	0.445	2.250
Chest depth	0.312	0.258	1.212	0.234	0.211	4.732
Withers height	-0.105	0.319	-0.330	0.743	0.028	35.753*
Rump height	0.321	0.332	0.966	0.341	0.026	38.628*
Body length	0.082	0.083	0.983	0.333	0.321	3.112

*VIF's are greater than 10, so multicollinearity problem was detected

The condition number (CN) coefficients calculated for Equation 6 were examined. In table 4, the multicollinearity problem was determined because of the calculated CN (100≤CN≤1000) (Table 4).

Table 4. Correlation eigenvalue and condition number

Number	Eigenvalue	Condition Number
1	4.27	1.00
2	0.93	4.56
3	0.38	11.24
4	0.26	16.30
5	0.13	31.89
6	0.01	315.63

Results of the Ridge Regression

In table 5 shows the ridge regression analysis results for each of k values which eliminates the multicollinearity problem and gives the highest R^2 value.

In table 6, VIF values are given for each k. k=0 represents the VIF values for the LS method. When all VIF values were examined, it was detected that the VIF<10 for k = 0.0136.

Table 5. Optimum k values selection

k	R^2	Sigma	B'B	Ave VIF	Max VIF
0	0.8800	1.21164	0.3700	14.8471	38.6278
0.01	0.8768	1.22753	0.2854	6.3689	13.1059
0.0136	0.8758	1.23253	0.2742	5.2838	9.9764
0.02	0.8741	1.24099	0.2616	4.1014	6.6907
0.03	0.8716	1.25352	0.2503	3.1147	4.1488
0.04	0.8691	1.26550	0.2434	2.5642	2.8851
0.05	0.8667	1.27708	0.2384	2.2078	2.6150
0.06	0.8643	1.28835	0.2344	1.9537	2.3935
0.07	0.8620	1.29935	0.2310	1.7604	2.2018
0.08	0.8597	1.31012	0.2281	1.6065	2.0345
0.09	0.8574	1.32067	0.2255	1.48	1.8874

Table 6. VIF values for detection of k values

k	Chest girth	Chest width	Chest depth	Withers height	Rump height	Body length
0.0000	4.608	2.250	4.732	35.753	38.628	3.113
0.0100	3.894	2.029	4.023	12.278	13.159	2.883
0.0136	3.711	1.978	3.838	9.393	9.976	2.807
0.0200	3.430	1.898	3.552	6.359	6.691	2.680
0.0300	3.066	1.792	3.181	4.002	4.149	2.499
0.0400	2.768	1.700	2.874	2.823	2.885	2.337
0.0500	2.517	1.616	2.615	2.145	2.164	2.191
0.0600	2.302	1.541	2.394	1.716	1.711	2.058
0.0700	2.117	1.471	2.202	1.426	1.408	1.938
0.0800	1.956	1.407	2.035	1.220	1.193	1.829
0.0900	1.815	1.348	1.887	1.066	1.035	1.729

When the VIF values were examined in Table 7, the multicollinearity problem between withers height and rump height measurements, which was used for body measurement, was eliminated by the RR method.

Table 7. Regression analysis results according to ridge trace and ridge regression method (*k = 0.0136*)

Parameters	Coefficients	Standard Error	t-Values	Sig.	VIF
Intercept	-25.273	-	-	-	-
Chest girth	0.263	0.087	3.023	*	3.711
Chest width	0.603	0.241	2.502	*	1.978
Chest depth	0.342	0.236	1.449	-	3.838
Withers height	-0.001	0.166	-0.006	-	9.393
Rump height	0.210	0.172	1.221	-	9.976
Body length	0.086	0.080	1.075	-	2.807

In Table 8 is examined, the models were statistically significant for LS and RR methods (P<0.001). The least square method had lower MSE value than RR method. However, RR method had higher R^2 value than LS method. In many studies such as Ergunes (2004), Topal et al (2010), Uckardes et al (2012) and Cankaya et al (2019) were similar results.

Table 8. Comparison of LS and RR analysis results

Methods	MSE	R^2	Sig.
Least Squares	1.47	0.880	<0.001
Ridge Regression	1.52	0.876	<0.001

CONCLUSION

In the assessment of the data for agricultural studies based on cause and effect relationships, multiple regression is used based on the LS method. In this study, the LS method used in the presence of multicollinearity determined by VIF value was compared with ridge regression method for estimating the body weight. In estimating the body weight from some body measurements (withers height, rump height, body length, chest depth, chest width, chest girth, rump width), the RR method is more reliable than the LS method. On account of the presence of a multicollinearity problem, Ridge Regression is an alternative method according to the multiple regression method.

REFERENCES

Akcay A, Sariozkan A, 2015. Estimation of income with using Ridge Regression analysis in layer hen industry. Ankara Üniversitesi Veteriner Fakültesi Dergisi, 62, 69-74, 2015.

Albayrak AS, 2005. An alternative bias estimation technique and an application of the least-squares technique in multiple linear connections. Zonguldak Karaelmas University Journal Social Sciences1: 105-126.

Alpar R, 2011. Uygulamalı çok değişkenli istatistiksel yöntemler. 3. Baskı. Kızılay/Ankara. Detay Yayımcılık. ISBN:978-605-5437-42-8

Anderson B, 1998. Scandinavian evidence on growth and age structure, ESPE 1997 Conference at Uppsala University.

Ari A, Onder H, 2013. Regression Models Used for Different Data Structures. Anadolu Journal of Agricultural Sciences, 2013,28(3):168-174. doi: 10.7161/anajas.2013.28.3.168.

Aytekin I, Eyduran E, Karadas K, Aksahan R, Keskin I, 2018. Prediction of Fattening Final Live Weight from some Body Measurements and Fattening Period in Young Bulls of Crossbred and Exotic Breeds using MARS Data Mining Algorithm. Pakistan Journal of Zoology, vol. 50(1), pp 189-195, 2018.

Cankaya S, Eker S, Abaci, SH, 2019. Comparison of Least Squares, Ridge Regression and Principal Component Approaches in the Presence of Multicollinearity in Regression Analysis. Turkish Journal of Agriculture-Food Science and Technology, 7(8): 1166-1172. DOI: 10.24925/turjaf.v7i8.1166-1172.2515

Ergüneş E, 2004. The examing least square method and Ridge regression method by comparison. Cukurova University Graduate School of Natural and Applied Sciences, Master Thesis (Printed).

Eyduran E, Zaborski D, Waheed A, Celik S, Karadas K, Grzesiak W, 2017. Comparison of the predictive Capabilities of several data mining algorithms and multiple linear regression in the prediction of body weight by means of body measurements in the indigenous beetal goat of Pakistan. Pakistan Journal of Zoology, 49: 257-265. https://doi.org/10.17582/journal.pjz/2017.49.1.257.265.

Gujarati DN, 1995. Basic econometrics, 3rd ed. McGraw-Hill, New York, USA.

Hair JF Jr, Black WC, Babin BJ, Anderson RE, (2014). Multivariate Data Analysis (7th edition), pp. 200, ISBN 10: 1-292-02190-X, Pearson Education Limited -England.

Hoerl AE, Kennard R, 1970. Ridge regression: Biased estimation for non-orthogonal problems. Technometrics, 12: 55-67. https://doi.org/10.1080/00401706.1970.10488635

IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Kurtuluş M, 2001. A study on ridge regression. Gazi University Graduate School of Natural and Applied Sciences, Master Thesis (Printed).

Kutner MH, Nachtsheim CJ, Neter J, Li W, 2004. Applied Linear Statistical Models, 5th edition, pp. 15, McGraw-Hill/Irwin.

Lukuyu MN, Gipson JP, Savage DB, Duncan AJ, Mujibi FBN, Okeyo AM, 2016. Use of body linear measurements to estimate live weight of crossbred dairy cattle in small holder farms in Kenya. SpringerPlus, 5:3-14. https://doi.org/10.1186/s40064-016-1698-3.

NCSS 11 Statistical Software trial version, 2016. NCSS, LLC. Kaysville, Utah, USA, ncss.com/software/ncss.

Pagel MU, Lunneborg CE, 1985. Empirical evaluation of Ridge Regression. Psychological Bulletin, 97: 342-355. https://doi.org/10.1037/0033-2909.97.2.342.

Rathert ÇT, Üçkardeş F, Narinç D, Aksoy T, 2011. Comparison of principal component regression with the least square method in prediction of internal egg quality characteristics in Japanese quails. Kafkas Universitesi Veteriner Fakültesi Dergisi, 17:687-692.

Sahin M, Yavuz E, Uckardes F, 2018. Multicollinearity Problem and Bias Estimates in Japanese Quail. Pakistan Journal of Zoology, vol. 50(2), pp 757-761, 2018. DOI: 10.17582/journal.pjz/2018.50.2.757.761.

Sarstedt M, Mooi E, 2014. A Concise Guide to Market Research, Springer Texts in Business and Economics: Chapter 7, Regression Analysis pp. 193-233. DOI 10.1007/978-3-642-53965-7_7, #Springer-Verlag Berlin Heidelberg.

Topal M, Eyduran E, Yaganoglu AM, Sonmez AY, Keskin S, 2010. Use of Ridge and Principal Component Regression Analysis Methods in Multicollinearity. Journal of Agricultural Faculty of Atatürk University, 41 (1), 53-57. ISSN: 1300-9036.

Uckardes F, Efe E, Narinc D, Aksoy T, 2012. Estimation of the egg albumen index in the Japanese quails with ridge regression method. Akademik Ziraat Dergisi 1(1): 11-20. ISSN: 2147-6403.

Vupa O, Gurunlu Alma O, 2008. Investigation of Multicollinearity Problem in Small Samples Included Outlier Value in Linear Regression Analysis. Selcuk University Journal of Science Faculty. Vol.31, 97-107.