Turkish Journal of Agriculture - Food Science and Technology, 10(4): 791-797, 202
DOI: https://doi.org/10.24925/turjaf.v10i4.791-797.4524

Turkish Journal of Agriculture - Food Science and Technology

Available online, ISSN: 2148-127X | www.agrifoodscience.com | Turkish Science and Technology Publishing (TURSTEP)

# The Solution of Multicollinearity Problem via Biased Regression Analysis in Southern Anatolian Red Cattle

**Hatice Hızlı**[1,a,*]

[1]*Eastern Mediterranean Agricultural Research Institute 01375 Adana, Turkey*
[*]*Corresponding author*

ARTICLE INFO

*Research Article*

Received : 14/06/2021
Accepted : 28/04/2022

*Keywords:*
Multicollinearity
Statistical bias
Ridge regression
Principal component regression
Least squares regressin

ABSTRACT

The aim of this study is to investigate the effectiveness of biased estimation methods, principal component regression (PC) and ridge regression (RR) methods, according to unbiased the least squares (LS) method, against the multiple linearity problem (multicollinearity) encountered in regression methods. For this purpose to fit a model on account to predict body weight from some body measurements of 32 South Anatolian Red Kilis (SAR) cattle of different ages. $R^2$, RMSE, MSE, and CV were used as the goodness of fit criteria for the performance of the models. According to these criteria respectively, 0.9970, 0.0224, 0.0005, 0.0099 for LS; 0.9970, 0.0224, 0.0005, 0.0099 for PC; and 0.9876, 0.0455, 0.0021, 0.0201 of k=0.02 for RR gave the best fit values. According to these results, LR and PC showed the best fit. But RR and PC techniques from biased prediction techniques provided more consistent, valid, stable, and theoretical expectations than LS technique, since LR did not provide the necessary assumptions.

[a] *haticehizli@gmail.com* *http://orcid.org/0000-0002-5451-1397*



## Introduction

Regression analysis is a statistical method that is used in estimating the relationships between the dependent variable and one or more independent variables, and develops models of these relationships (linear or unlinear) to determine the relationships between the variables and thanks to this usefulness is, frequently used in almost all branches of science. The purpose of regression analysis is to create the best model that can predict the dependent variable from the independent variables or to determine which independent variables are more affected by the dependent variable.

The least squares regression (LS) estimation technique allows the parameters in the model to be estimated so that the error sum of squares is minimal (Draper and Smith, 1981; Kleinbaum et al., 1988). According to LS technique in order to estimate the parameters for the multiple regression analysis to be applied some assumptions must be valid. Otherwise the predictions are biased and thus the relevant significance tests lose their validity (Orhunbilge, 2017; Topal et al., 2010). These assumptions; the expected value (mean) of the errors is zero E (e) = 0, the errors are independent of each other. That is, there is no order correlation between unit values (absence of serial correlation) Cov ($e_i$, $e_j$) = 0, the variance of errors is constant (no different variance), Var ($e_i$) = $\sigma^2$, there is no correlation between errors ($e_i$) and dependent variable (Y) (absence of simultaneous equation bias), Cov ($e_i$, $Y_i$) = 0, errors and independent variables are independent of each other, Cov ($e_i$, $X_i$) = 0, between the independent variables no significant relationship (absence of multicolinearity), Cov ($X_i$, $X_j$) = 0 (Draper and Smith, 1981; Kleinbaum et al., 1988).

In the case of multicollinearity, parameter estimates made with the LS technique are unbiased, but the results are invalid because the assumptions cannot be met and so it is not reliable (Shrestha, 2020). In order to solve the multicollinearity problem, it may be suggested to add new independent variables to the existing data or to remove some of the related independent variables from the model. However, this causes information can lost in the model. Another suggestion is to use the biased estimation techniques ridge regression (RR) or principal component (PC) techniques instead of the LS technique.

By adding a small bias constant (k) with biased estimation techniques, standard error values become smaller due to the variances of the parameters and much more meaningful models can be obtained ($P<0.05$) (Büyükuysal and Öz, 2016; Hoerl and Kennard, 1970a,b; Kleinbaum et al., 1988; Orhunbilge, 2017). The assumptions of biased estimation techniques are the same as those of LS: linearity, covariance, and independence. However, since confidence intervals are not calculated in biased estimation techniques, normality assumption is not made (Rawlings, 1998). Larger samples are needed to obtain effective, valid and stable results with the LS technique in multicollinearity data (Maxwell, 2000). In such cases, biased estimation techniques yield more effective, valid and stable results by using smaller samples (Vinod, 1995).

The aim of this study is to compare the relative predictive validity of LS and biased estimation techniques (RR and PC) in estimating the linear relationship between body weight and explanatory variables. Based on the high multicollinearity problem between the explanatory variables as withers height (WH), body length (BL), chest girth (CG), ankle girth (AG), rump width (RW), it was expected that RR and PC techniques would provide predictions with lower standard errors, and stationary and theoretical expectations than the LS technique.

## Material and Methods

### *Materials*

This study was conducted using the data obtained from 32 Southern Anatolian Red Kilis (SAR) cattle raised within the scope of the "National Project for Conservation of Local Genetic Resources in the Hands of the Public". The project was carried out by the Eastern Mediterranean Agricultural Research Institute in Hatay. WH, BL, CG, AG, RW, and BW measurements were taken with the cattle standing, using a measuring stick and tape measure in cm and a digital balance in kg, respectively. In the regression models, WH, BL, CG, AG, and RW body measurements were assigned as the independent variable and BW as the dependent variable. All statistical analysis was performed using the NCSS trial version (NCSS, 2021). Shapiro-Wilk's test of normality was performed on all variables for statistical analysis. Since the normality test was found to be significant at $P<0.05$, first of all, all extreme observations were removed and statistical analyzes were conducted by performing Logarithmic transformation in all variables. Details of body measurements for descriptive statistics are given in Table 1. For clarity, Table 1 results were given real results by taking their anti-logarithms.

### *Methods*

In this study was to investigate the effectiveness with principal component regression (PC) and ridge regression (RR) models of biased estimation techniques, according to unbiased the least squares regression (LS) technique, to against the multicollinearity encountered in regression methods. Regression analysis is a method that reveals the cause-effect relationship between a dependent and one or more independent variables, and the matrix notation of the estimation equation belonging to the multiple regression model with more than one independent variable is shown in equation (1) (NCSS Inc, 2001).

$$\underline{Y} = X\underline{B} + \underline{e} \qquad (1)$$

where, $\underline{Y}$, the $n\times 1$ dimensional dependent variable vector; X, the $n\times(p+1)$ dimensional independent variable matrix; $\underline{B}$, $(p+1)\times 1$ dimensional vector of the regression coefficients; $\underline{e}$ denotes the $n\times 1$ error vector.

The presence of a high level of the linear relationship between the independent variables in the multiple regression model is important as it may cause the problem of multicollinearity, since the variances get larger, the estimations may diverge from their true values. There are several approaches used to determine multicollinearity. These are simple correlation matrix, changes in $R^2$ when new independent variables are added to the model, partial correlation coefficients, Variance Inflation Factor (VIF), Tolerance Value (TV), F values calculated using auxiliary regression equations, condition number (CN) and condition index (CI) are examined sequentially (Neter et al. 1990; Kim, 2019).

*Least Squares regression (LS)*

The least squares technique is a optimumm model that estimates the relationship between data points with the most appropriate linear line and minimizes the sum of squares of error terms, has homogeneous variance (Çankaya et al. 2019; Kayaalp et al. 2015; Şahinler, 2000). In the multiple regression analysis, equation (2) is shown for the estimation of the coefficients vector by the LS technique (NCSS Inc, 2001).

$$\hat{B} = (X'X)^{-1}X'Y \qquad (2)$$

In the equation; $\hat{B}$, $(p-1)\times 1$ dimensional vector of regression coefficients; $X'X = R = r_{XX}$, Correlation matrix of $(p-1)\times(p-1)$ dimensional independent variables; $X'Y = r_{YX}$, it is the correlation vector between the $(p-1)\times 1$ dimensional dependent $Y$ variable and the independent $X$ variables.

Table 1. The data structure of the body measurements for SAR cattle

| Variable | Count | Mean | Standard Deviation | Minimum | Maximum | CV % |
|---|---|---|---|---|---|---|
| Withers height | 32 | 107.77 | 1.23 | 66.07 | 147.91 | 1.14 |
| Body length | 32 | 115.06 | 1.29 | 64.57 | 173.78 | 1.12 |
| Chest girth | 32 | 127.35 | 1.37 | 64.57 | 194.98 | 1.07 |
| Ankle girth | 32 | 14.75 | 1.26 | 8.51 | 19.95 | 8.54 |
| Rump width | 32 | 36.60 | 1.38 | 18.20 | 57.54 | 3.76 |
| Body weight | 32 | 184.74 | 2.37 | 34.67 | 630.96 | 1.28 |

*Ridge regression (RR)*

In ridge regression technique, firstly, the dependent and independent variables are standardized by taking the difference from their mean and dividing them by their standard deviations. When the final regression coefficients are obtained, the coefficients are converted to the original measurement units. Since the variables are standardized and R shows the correlation matrix between the independent variables is $X'X = R$. These estimates are unbiased as the expected values of these estimates will be equal to the population values, it is shown in equation (3):

$$E(\underline{\hat{B}}) = \underline{B} \quad (3)$$

Variance-covariance matrix of estimates is shown equation (4):

$$V(\underline{\hat{B}}) = \sigma^2 \mathrm{R}^{-1} \quad (4)$$

Since the y's are standardized $\sigma^2 = 1$ and thus equality (5) is written:

$$V(\hat{b}_j) = r^{jj} = \frac{1}{1-\mathrm{R}_j^2} = VIF \quad (5)$$

In Ridge regression technique, by adding a small bias constant to the diagonal values of the correlation matrix, the biased standardized regression coefficients are calculated as follows:

$$\underline{\tilde{B}} = (X'X + kI)^{-1}X'Y \text{ or } \underline{\tilde{B}} = (R + kI)^{-1}X'Y \quad (6)$$

where k is a positive numerical value less than 1 (usually k≤ 3). The expected value of the bias of this estimate is as follows equation (7).

$$E(\underline{\tilde{B}} - \underline{B}) = [(X'X + kI)^{-1}X'Y - I]\underline{B} \quad (7)$$

and the covariance matrix is obtained by the following equation (8):

$$V(\underline{\tilde{B}}) = (X'X + kI)^{-1}X'X(X'X + kI)^{-1} \quad (8)$$

In order to research for the optimumm k constant, a graph called Ridge Trace, which is calculated between biased standardized regression coefficients and k is used (Hoerl and Kennard, 1970a,b).

*Principal Component regression (PC)*

Independent variables are converted to principal components to perform PC regression technique (NCSS Inc., 2001). This is mathematically expressing the equation (2) as equation (9).

$$X'X = PDP' = Z'Z \quad (9)$$

where, *D* describing the PC model in the equation is the diagonal matrix of *X'X* eigenvalues; *P* is the *X'X* eigenvector matrix; and *Z* (similar to the *X* structure) is the data matrix. Since the principal components (*P*) are orthogonal, *P'P=I.*

Thus, there is no multicollinearity anymore in the model, where *Y* variable is dependent and *Z* components are independent variables. The results are then converted back to the X scale to obtain estimates of *B*. These estimates will be biased. However, it is hoped that the magnitude of this bias will be offset by reducing the variance. In other words, the mean squares of error of the PC estimates are expected to be smaller than the LS estimates. Mathematically, due to the special nature of the prime components, the estimation of the regression coefficients is as follows (NCSS Inc., 2001):

$$\underline{\hat{A}} = (Z'Z)^{-1}Z'Y = (D)^{-1}Z'Y \quad (10)$$

This equation is the LS regression applied on different independent variable set. Thus, the relationships between two sets of regression coefficients, A and B, are written as equation (11):

$$\underline{\hat{A}} = P'\underline{\hat{B}} \text{ and } \underline{\hat{B}} = P'\underline{\hat{A}} \quad (11)$$

Principal components (*Z*) are obtained by applying PC analysis for the *X* matrix. In contrast to the uncertainty experienced in the selection of the k-bias constant in the RR technique, the number of principal components to be eliminated in the PC analysis is relatively certain (NCSS Inc., 2001).

*Comparison of LS, PC and RR*

In order to compare the performance of predicted PC, RR and LS regression models, determination coefficient $R^2$, root mean square error RMSE, mean squere error MSE and coefficient of variation CV of models were used as the goodness of fit criteria, and represented by Equations (12), (13), (14) and (15), respectively (Erzin and Cetin, 2017)

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}\hat{Y}_i^{\,2}}\right) \quad (12)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \quad (13)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad (14)$$

$$CV = \left[\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}}{(\bar{Y}_i)}\right].100 \quad (15)$$

where $Y_i$, $\hat{Y}_i$ , $\bar{Y}_i$ n is i.th measured, i.th predicted, the total number of the data, and the average of the data, respectively

## Results and Discussion

The results of multiple regression analysis of the examined variables are given according to LS, RR, and PC techniques, respectively.

***Least Squares regression (LS)***

Table 2 shows that according to the LS analysis, the linear relationship between body weight and explanatory variables is 0.998%, and 0.997% of the changes in body weight are explained by the independent variables and the model is significant (P<0.001).

Table 2. LS ANOVA

| Source | DF | If Term(s) Removed | Sum of Squares | Mean Square | F-Ratio | Prob Level |
|---|---|---|---|---|---|---|
| Regression | 5 | 0.997 | 4.335726 | 0.867145 | 1734.862 | 0 |
| Error | 26 | 0.003 | 0.01299572 | 0.0005 | | |
| Total(Adjusted) | 31 | | 4.348722 | 0.140281 | | |
| R | $R^2$ | $R^2_{adj.}$ | SE | | | |
| 0.998 | 0.997 | 0.996 | 0.0005 | | | |

Table 3. Results of LS analysis

| Independent Variable | Regression Coefficient b(i) | Standard Error Sb(i) | Standard-ized Coefficient | T-Statistic to Test H0: β(i)=0 | Prob Level | Lower 95.0% Conf. Limit of β(i) | Upper 95.0% Conf. Limit of β(i) |
|---|---|---|---|---|---|---|---|
| Intercept | -2.89272 | 0.312669 | 0 | -9.25169 | 0 | -3.53542 | -2.25002 |
| WH | -0.70854 | 0.538444 | -0.1723 | -1.31591 | 0.1997 | -1.81533 | 0.398246 |
| BL | 0.313219 | 0.520737 | 0.0933 | 0.601492 | 0.5527 | -0.75717 | 1.38361 |
| CG | 2.45879 | 0.329459 | 0.8902 | 7.463118 | 0 | 1.781577 | 3.136002 |
| AG | -0.10753 | 0.223373 | -0.0288 | -0.48138 | 0.6343 | -0.56668 | 0.351624 |
| RW | 0.578075 | 0.376698 | 0.2142 | 1.534587 | 0.137 | -0.19624 | 1.352388 |

Table 4. Multicollinearity Report. LS

| Indep. Var. | Correlation Matrix | | | | | | $R^2$ Versus Other I.V.'s | Tolerance Value TV | Diagonal of X'X Inverse |
|---|---|---|---|---|---|---|---|---|---|
| | WH | BL | CG | AG | RW | VIF | | | |
| WH | 1.000 | | | | | 149.1851 | $R^2_{1.2345}$ =0.9933 | 0.0067 | 580.0354 |
| BL | 0.996 | 1.000 | | | | 209.2321 | $R^2_{2.1345}$ =0.9952 | 0.0048 | 542.5133 |
| CG | 0.981 | 0.986 | 1.000 | | | 123.7799 | $R^2_{3.1245}$ =0.9919 | 0.0081 | 217.1577 |
| AG | 0.961 | 0.969 | 0.979 | 1.000 | | 31.0803 | $R^2_{4.1235}$ =0.9678 | 0.0322 | 99.82425 |
| RW | 0.983 | 0.988 | 0.996 | 0.983 | 1.000 | 169.4634 | $R^2_{5.1234}$ =0.9941 | 0.0059 | 283.8955 |

Table 5. Eigenvalues of Correlations

| No. | Eigenvalue | Incremental Percent | Cumulative Percent | Condition Number CN | Condition Index CI |
|---|---|---|---|---|---|
| 1 | 4.929 | 98.580 | 98.580 | 1.000 | 1.000 |
| 2 | 0.047 | 0.950 | 99.530 | 103.970 | 10.197 |
| 3 | 0.017 | 0.340 | 99.860 | 293.840 | 17.142 |
| 4 | 0.004 | 0.080 | 99.940 | 1272.400 | 35.671 |
| 5 | 0.003 | 0.060 | 100.000 | 1694.040 | 41.159 |

Table 3 shows, the regression coefficients, standardized coefficients of the regression, standard errors, t-test, significance level, and confidence intervals of the independent variables, and only CG from the explanatory variables was found to be significant according to the t-test ($P<0.05$). If the $R^2$ value of the model is high, but none or very few of the independent variables are significant according to the partial t-test, it is an indication of the multicollinearity problem.

In Table 4, there is a positive correlation between the body weight of the SAR cattle and the examined body characteristics, all correlations are over 95% and significant, while the highest correlation is found between BL and WH ($r=0.996$, $P<0.01$), the lowest correlation was found between AG and WH ($r=0.961$, $P<0.01$). If the correlation coefficients between the examined independent variables are close to 1, it can be mentioned that there is a multicollinearity problem. The fact that VIF values are greater than 10 in all variables indicates the existence of multicollinearity problem. When each of the independent variables is taken as a dependent variable and the relationships between the remaining independent variables are examined, it is seen that these relationships are greater than 0.90 ($R^2_{1.2345}$, $R^2_{2.1345}$, $R^2_{3.1245}$, $R^2_{4.1235}$, $R^2_{5.1234}$) (Table 4). In addition, the number of conditions number and condition index (CN=1694.040, CI=41.159) in Table 5 is greater than the critical value (1000 and 30). All of the statistics given in Table 4 and Table 5 show that there is a very strong linear connection problem in the data.

***Ridge regression (RR)***

Table 6 shows the k bias estimator and the VIF selection table. It is determined as the selection value for the k constant, in which the regression coefficients standardized for the determination of the k constant in the RR technique become stationary and the VIF values of these coefficients approach 1. As a result of the analysis, k=0.02 was determined.

In Table 7, at a value of k=0.02, it is seen that the multicollinearity problem between body measurements used for BW estimation is eliminated by the RR technique.

Looking at the graph Figure 1 the bias constant (k) and the biased standardized regression coefficients, it is seen that the regression coefficients become stationary after a very small bias constant (k=0.02) (Hastie, 2020).

As a result of the multicollinearity problem in the study, it is seen that the standard errors of the estimators are high and the sign of the coefficient of the WH and AG variables contradicts the theoretical and empirical expectations. In addition, the magnitudes of the regression coefficients are negatively affected by multicollinearity. Iterations are started by choosing an approximate k value in the region where the biased regression coefficients become stationary and the VIF values of these coefficients approach 1 together by using VIF and ridge graphs in order to search for the optimum bias constant. As a result of the iterations, the sign of the coefficient of the WH and AG variables changes and the standard errors of the estimators decrease in the results obtained with the optimum k=0.02 bias constant chosen for RR.

Table 6. Ridge Regression, VIF and k Analysis Section

| k | Variance Inflation Factor Section | | | | | k Analysis Section | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | WH | BL | CG | AG | RW | R2 | Sigma | B'B | Ave VIF | Max VIF |
| 0 | 149.185 | 209.232 | 123.780 | 31.080 | 169.463 | 0.997 | 0.022 | 0.878 | 136.548 | 209.232 |
| 0.0001 | 140.146 | 196.009 | 118.463 | 30.633 | 161.407 | 0.997 | 0.023 | 0.864 | 129.332 | 196.009 |
| 0.0002 | 131.957 | 184.031 | 113.527 | 30.207 | 153.932 | 0.997 | 0.023 | 0.851 | 122.731 | 184.031 |
| 0.0003 | 124.514 | 173.146 | 108.935 | 29.800 | 146.985 | 0.997 | 0.023 | 0.838 | 116.676 | 173.146 |
| 0.0004 | 117.727 | 163.223 | 104.654 | 29.411 | 140.515 | 0.997 | 0.024 | 0.826 | 111.106 | 163.223 |
| 0.0005 | 111.521 | 154.154 | 100.655 | 29.037 | 134.481 | 0.997 | 0.024 | 0.815 | 105.969 | 154.154 |
| 0.0006 | 105.832 | 145.841 | 96.913 | 28.677 | 128.842 | 0.997 | 0.024 | 0.804 | 101.221 | 145.841 |
| 0.0007 | 100.603 | 138.203 | 93.407 | 28.330 | 123.565 | 0.996 | 0.024 | 0.794 | 96.822 | 138.203 |
| 0.0008 | 95.784 | 131.168 | 90.114 | 27.996 | 118.620 | 0.996 | 0.025 | 0.784 | 92.737 | 131.168 |
| 0.0009 | 91.334 | 124.675 | 87.019 | 27.673 | 113.979 | 0.996 | 0.025 | 0.774 | 88.936 | 124.675 |
| 0.001 | 87.216 | 118.668 | 84.104 | 27.361 | 109.616 | 0.996 | 0.025 | 0.765 | 85.393 | 118.668 |
| 0.002 | 58.677 | 77.168 | 62.307 | 24.687 | 77.361 | 0.996 | 0.028 | 0.687 | 60.040 | 77.361 |
| 0.003 | 43.082 | 54.682 | 48.809 | 22.576 | 57.910 | 0.995 | 0.030 | 0.629 | 45.412 | 57.910 |
| 0.004 | 33.552 | 41.097 | 39.735 | 20.826 | 45.225 | 0.994 | 0.031 | 0.583 | 36.087 | 45.225 |
| 0.005 | 27.252 | 32.238 | 33.261 | 19.335 | 36.458 | 0.994 | 0.033 | 0.545 | 29.709 | 36.458 |
| 0.006 | 22.834 | 26.122 | 28.431 | 18.040 | 30.125 | 0.993 | 0.034 | 0.513 | 25.110 | 30.125 |
| 0.007 | 19.592 | 21.710 | 24.700 | 16.900 | 25.385 | 0.993 | 0.035 | 0.486 | 21.657 | 25.385 |
| 0.008 | 17.124 | 18.412 | 21.738 | 15.887 | 21.735 | 0.992 | 0.036 | 0.463 | 18.979 | 21.738 |
| 0.009 | 15.191 | 15.877 | 19.334 | 14.980 | 18.858 | 0.992 | 0.037 | 0.442 | 16.848 | 19.334 |
| 0.01 | 13.638 | 13.880 | 17.347 | 14.161 | 16.545 | 0.991 | 0.038 | 0.424 | 15.114 | 17.347 |
| 0.02 | 6.628 | 5.590 | 7.789 | 8.915 | 6.458 | 0.988 | 0.046 | 0.319 | 7.076 | 8.915 |
| 0.02 | 6.628 | 5.590 | 7.789 | 8.915 | 6.458 | 0.988 | 0.046 | 0.319 | 7.076 | 8.915 |

Table 7. Ridge Regression Coefficient Section for k = 0.02

| Independent Variable | Regression Coefficient | Standard Error | Standardized Regression Coefficient | VIF |
|---|---|---|---|---|
| Intercept | -3.061 | | | |
| WH | 0.118 | 0.231 | 0.029 | 6.628 |
| BL | 0.336 | 0.173 | 0.100 | 5.590 |
| CG | 1.254 | 0.168 | 0.454 | 7.789 |
| AG | 0.425 | 0.243 | 0.114 | 8.915 |
| RW | 0.804 | 0.150 | 0.298 | 6.458 |

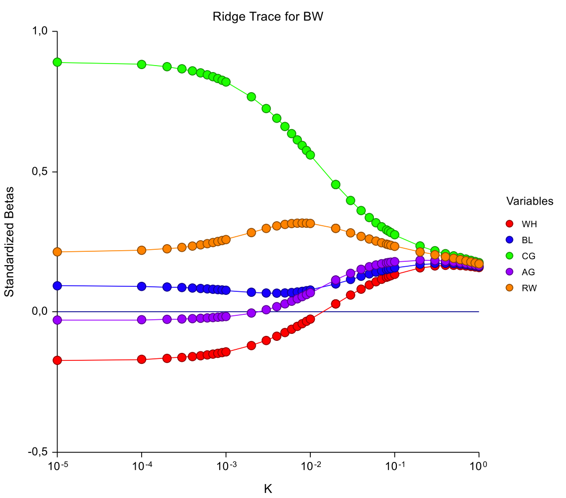


Figure 1. RR graphic

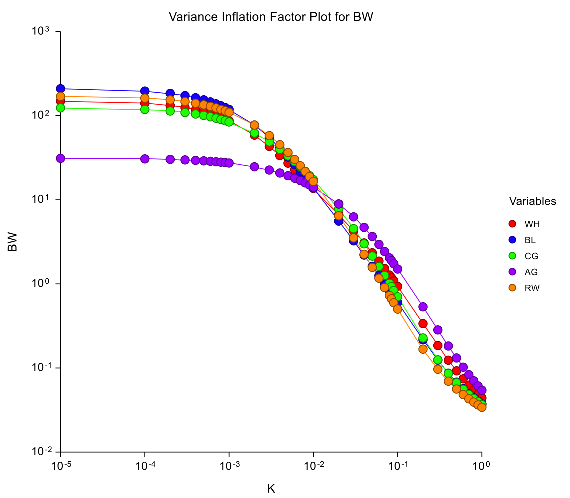


Figure 2. VIF graphic

***Principal Component regression (PC)***

The results of PC analysis of some body features of the SAR are shown in Tables 8 and 9 respectively. When Table 8 is examined, the eigenvalues of the 5 basic components are seen. VIF values according to PC analysis results are given in Table 9. It is seen that the multicollinearity problem and multiple correlations are detected between them, are resolved.

***Comparison of LS, PC and RR***

In Table 10 shows comparison of PC, RR, and LS regression models, by the goodness of fit criteria R2, RMSE, MSE and CV, respectively. According to these results, LR and PC showed the best perpormance whit critical values R2=0.9970, RMSE=0.0224, MSE=0.0005, and CV=0.0099. But RR and PC techniques from biased prediction techniques provided more consistent, valid, stable, and theoretical expectations than LS technique, since LR did not provide the necessary assumptions. Many researchers on different subjects, Topal, et al. (2010) in carp fish, Shafey et al. (2015) and Akçay and Sarıözkan (2015) in chickens, Çiftsüren and Akkol (2018) in eggs, Çelik Ş et al. (2018) in white turkeys and Yılmaz et al. (2020), Tırınk et al (2020) Saanen kids, compared LS with RR and PC and other unbiased techniques against multicollinearity problem in and suggested biased estimation methods.

Table 8. PC Regression Analaysis. Descriptive statistics

| Principal Component | PC Coefficient | Individual R-Squared | Eigenvalue |
|---|---|---|---|
| PC1 | -0.168 | 0.987 | 4.929 |
| PC2 | 0.060 | 0.001 | 0.047 |
| PC3 | 0.264 | 0.008 | 0.017 |
| PC4 | 0.145 | 0.001 | 0.004 |
| PC5 | -0.027 | 0.000 | 0.003 |

Table 9. PC regression analysis results

| PC's | $R^2$ | SSE | B'B | Ave VIF | Max VIF |
|---|---|---|---|---|---|
| 1 | 0.987 | 0.047 | 0.200 | 0.041 | 0.041 |
| 2 | 0.988 | 0.045 | 0.226 | 4.259 | 10.840 |
| 3 | 0.996 | 0.025 | 0.723 | 16.182 | 27.434 |
| 4 | 0.997 | 0.022 | 0.872 | 67.811 | 167.488 |
| 5 | 0.997 | 0.022 | 0.878 | 136.548 | 209.232 |

Table 10. Comparison of PC, RR, and LS regression models, by the goodness of fit criteria

| Goodness of fit criteria | LS | RR, k=0.02 | PC |
|---|---|---|---|
| $R^2$ | 0.9970 | 0.9876 | 0.9970 |
| RMSE | 0.0224 | 0.0455 | 0.0224 |
| MSE | 0.0005 | 0.0021 | 0.0005 |
| CV | 0.0099 | 0.0201 | 0.0099 |

R2: Coefficient of determination; RMSE: Root Mean Square Error; MSE: Mean Square Error; CV: Coefficient of Variation

## Conclusion

In this study, LS, RR and PC techniques were used for body weight estimation in SAR cattle. Problems related to multicollinearity of the investigated independent variables (features) were identified. For the solution of the problem, RR and PC techniques have been used since they provide more stable and theoretical (or empirical) results compared to the LS technique. Although the choice of LS and one of the biased estimation techniques (RR and PC) implies the choice of either biased or unbiased estimators, in reality this is not the case. As is known, in practical terms, LS estimators are unbiased only if the model is defined without error. For this reason, it is generally accepted that LS estimators will be biased in practice. In short, explanatory variables that are in a meaningful relationship with each other can be analyzed together in order to reduce the multicollinearity problem with biased estimation techniques.

As a result, since there is a multicollinearity between the independent variables explaining body weight, the RR technique, which is one of the biased estimation techniques, is more consistent, valid, stable and in line with the theoretical expectations than the LS technique, and also the PC technique gives the closest estimates to the LS technique with a smaller standard error. Thus, since there is a multicollinearity between the variables explaining body weight, the RR and PC techniques, which is one of the biased estimation techniques, provided more consistent, valid, stable and theoretical expectations than the LS technique.

## Author contributions

Hatice Hızlı performed project administration, supervision, conceptualization, dataset curation, formal analysis, methodology, writing original draft, review and editing, read and approved the final article.

## Conflict of Interest

The authors declared that there is no conflict of interest.

## References

Akçay A, Sarıözkan S. 2015. Yumurta tavukçuluğunda gelirin ridge regresyon analizi ile tahmini. Ankara Üniversitesi Veteriner Fakültesi Dergisi, 62, 69-74.

Büyükuysal MÇ, Öz İİ. 2016. Çoklu doğrusal bağıntı varlığında en küçük kareler alternatif yaklaşım: Ridge regresyonu. Düzce Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi, 6(2), 110-114.

Çankaya S, Eker S, Abacı 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.

Çelik Ş, Şengül T, Söğüt B, İnci H, Şengül AY, Kayaokay A, Ayaşan T. 2018. Analysis of variables affecting carcass weight of white turkeys by regression analysis based on factor analysis scores and ridge regression. Brazilian Journal Poultry Science, 273-280.

Çiftsüren MN, Akkol S. 2018. Prediction of internal egg quality characteristics and variable selection using regularization methods: Ridge, LASSO and Elastic Net. Arch. Anim. Breed, 61, 279-284.

Draper NR., and Smith H. 1981. Applied Regression Analysis, John Willey, NY.

Erzin Y, Çetin T. 2017. The prediction of the critical factor of safety of homogeneous finite slopes using neural networks and multiple regressions, Computers & Geosciences, 51:305-313, 2017geliştirilmiş 3. Basım, VIII, 386 S, ISBN:978-605-320-644-6, İstanbul.

Hastie T. 2020. Ridge Regularizaton: an Essential Concept in Data Science, https://arxiv.org/pdf/2006.00371.pdf

Hoerl AE. and Kennard RW.1970a. Ridge regression: biased estimation for nonortogonal problems. Technometrics, 12(1), 55-56.

Hoerl AE, Kennard RW. 1970b. Ridge regression: applications to nonortogonal problems, Technometrics, 12(1), 69-82.

Kayaalp GT, Güney O, Çelik M, Cebeci Z. 2015. Çoklu doğrusal regresyon modelinde değişken seçiminin zootekniye uygulanışı. Çukurova Üniv. Zir. Fak. Der., 30(1), 1-8.

Kim J. H. 2019. Multicollinearity and misleading statistical results. Korean journal of anesthesiology, 72(6): 558–569. https://doi.org/10.4097/kja.19087

Kleinbaum DG, Lawrence L, Kupper and Keith E. Muller 1988. Applied Regression Analysis and Other Multivariable Methods, Duxbury Press, New Jersey.

Maxwell, Scott E. 2000. Sample Size in Multiple Regression Analysis, Psychological Methods, Vol. 5, No: 4, s. 435-458.

NCSS 2021 Statistical Software 2021. NCSS, LLC. Kaysville, Utah, USA, ncss.com/software/ncss.

NCSS Inc. 2001. NCSS User Guide 2001, Kaysville, NCSS Inc.

Neter J, Wasserman W, Kunter M. 1990. Applied Linear Statistical Models, 3rd Ed., New Jersey.

Orhunbilge N. 2017. Uygulamalı regresyon ve korelasyon analizi. Nobel Yayınevi, gözden gözden geçirilmiş 3. Basım, VIII, 386 S, ISBN:978-605-320-644-6, İstanbul.

Rawlings JO. 1998. Applied Regression Analysis: A Research Tool, California.

Shafey TM, Hussein ES, Mahmoud AH, Abouheif MA, Al-Batshan HA. 2015. Managing collinearity in modeling the effect of age in the prediction of egg components of laying hens using stepwise and ridge regression analysis. Brazilian Journal Poultry Science, 473-482.

Shrestha N. 2020. Detecting Multicollinearity in Regression Analysis. American Journal of Applied Mathematics and Statistics, vol. 8(2):39-42. https://doi:10.12691/ajams-8-2-1

Şahinler S. 2000. En Küçük Kareler Yöntemi ile Doğrusal Regresyon Modeli Oluşturmanın Temel Prensipleri. MKÜ. Ziraat Fakültesi Dergisi 5 (1-2). Hatay, 57–73.

Tırınk C, Abacı Sh, Önder H. 2020. Comparison of Ridge Regression and Least Squares Methods in the Presence of Multicollinearity for Body Measurements in Saanen Kids, Journal of the Institute of Science and Technology, 10(2): 1429-1437, https://DOI:10.21597/jist.671662

Topal M, Eyduran E, Yağanoğlu AM, Sönmez A, Keskin S. 2010. Çoklu Doğrusal Bağlantı Durumunda Ridge ve Temel Bileşenler Regresyon Analiz Yöntemlerinin Kullanımı. Atatürk Üniversitesi, Ziraat Fakültesi Dergisi, 41(1): 53-57.

Vinod HD. 1995. Double Bootstrap for Shrinkage Estimators, Journal of Econometrics.

Yılmaz F, Bayyurt L, Abacı SH, Tahtalı Y. 2020. Comparison of Least Squares and Some Bias Estimators in Multicollinearity, Turkish Journal of Agriculture-Food Science and Technology, 8(3): 793-799, DOI: https://doi.org/10.24925/turjaf.v8i3.793-799.3405