Estimation of Protein Amount in Milk by Ridge Regression

Cem Tırınk[1*], Dariusz Piwczyński[2]

***Abstract* —** In this study, it was aimed to compare the performance of ridge regression in the presence of multicollinearity that will be used in regression analysis as an alternative to Least Squares. The content of protein in milk was estimated by using with various milk components such as: fat (%), urea (mg/l), dry matter (%), milk yield (kg) for Polish Holstein-Friesian cattle breed. In the presence of multicollinearity, more reliable models can be obtained by using some regression algorithms. We used one of these algorithms and the obtained model performances of the ridge regression were compared with the linear regression. To compare the performances of these obtained models, MSE, RMSE, rRMSE, MAPE, R^2, and AIC were used. As a result of, in the presence of multicollinearity, ridge estimator is recommended as an alternative method to linear regression technique.

***Keywords* —** linear regression, ridge regression, prediction, milk components, protein, multicollinearity.

I. Introduction

Milk and milk products are essential foods of animal origin in human nutrition. In this context, according to FAO data, 90% of the total milk production in the world is cattle [1]. The quality of the milk produced by the breeders for the expectation of economic income also affects its price.

Milk protein is also an important component in determining the quality of milk. Milk protein has an important place not only in milk but also in terms of products produced from milk, in terms of the processability of milk. When evaluated according to milk dry matter, the share of milk protein in milk content is approximately 25% [2]. For this reason, milk protein is a very important component in terms of milk processing.

Coagulation properties of milk are an important and necessary factor for cheese production industries (Duchemin et al., 2020) [3]. Coagulation properties of milk, which has an important place in milk processing, are affected by many factors such as somatic cell count, fat content, dry matter of milk and milk protein composition. In this context, it is also very important to determine the factors affecting the protein content, which has the highest share of quality criteria for both the quality of milk in natural consumption and the quality of the processed products of milk.

In this context, multivariate statistical methods are used as a common method. Regression analysis is one of the multivariate statistical methods that can be used to reveal the relationships between protein content of milk and milk production, fat content, urea content and dry matter. Regression analysis, which is one of the multivariate statistical modeling methods, is a process for estimating the relationship between explanatory variables and response variable [4]. The least squares method (LS) is a widely used procedure to estimate response variables in the regression model [5]. In addition, the LS method is an unbiased method that not only estimates response variables but also minimizes the error for the obtained model. However, the LS method requires several assumptions that must be satisfied for the model to be trustworthy. If the assumptions are not provided as desired, the reliability of the model will decrease. For this reason, it would be wrong to explain with models obtained from methods whose assumptions are not provided. Therefore, in order to guarantee the usability of the LS method, assumptions such as that the errors are independent and normally distributed and that they are independent among the explanatory variables must be valid [5]. If there is a linear relationship between the explanatory variables that called a multicollinearity problem. The multicollinearity problem causes the regression coefficients to be estimated inaccurately and decreases the predictability of the model.

The aim of this study is to compare ridge regression and least squares methods to estimate the protein content of milk by using parameters such as milk urea content, fat content, dry matter content and milk production amount in case of multiple correlations between explanatory variables.

II. Material and Methods

The study included a total of 200 Polish Holstein-Friesian cows kept in a dairy farm located in Kuyavian-Pomeranian voivodeship (Poland). Cows were milked using automatic milking system (AMS; Astronaut A4 by Lely East) and were fed a Partial Mixed Ration (PMR), with concentrate feed that was given to animals individually in the milking box depending on their milk yields. The barns had a free-stall system. The following milking performance variables were tested in this study: protein content (PC), %; fat content (FC), %; Dry matter (DM), %; Urea content (UC), mg/l, milking day (MD), days.

Cem Tırınk, Igdir University, Faculty of Agriculture, Department of Animal Science, Türkiye

Dariusz Piwczyński, Bydgoszcz University of Science and Technology, Faculty of Animal Breeding and Biology, Department of Animal Biotechnology and Genetics, Poland

https://doi.org/10.17758/EIRAI17.F1222226

All statistical analyzes were performed using R software [6]. For this aim, some packages such as “psych” and “lmridge” were used for the calculating the descriptive statistics and estimating the milk protein content from the explanatory variables, respectively [7, 8]. As a goodness of fit criteria were used to compare of the prediction models. For this aim, the “ehaGoF” package were used [9].

The common methods to explain the relationship between response and explanatory variables is regression analysis. In the matrices form of the multiple regression analysis as given below:

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \tag{1}$$

Regression analysis adopts the LS method. The main purpose of the least squares method (LS) is to minimize the sum of the squares of the error terms. In addition, the LS method must provide some assumptions such as having a normal distribution and having homogeneous variance and thus optimizing the model. In the presence of multicollinearity, the obtained model gave trustworthy results for the researchers. To overcome this problem, it can be used ridge regression.

To determine the multicollinearity, variance inflation factor (VIF) can be used. As the VIF value is more than the 10, it can be stated that there is a multicollinearity problem between explanatory variables. VIF value can be calculated the equation as given below.

$$VIF = c_{ij} = \frac{1}{1-R_j^2} \tag{2}$$

Ridge regression (RR) was proposed by Hoerl and Kennard (1970) [10]. The matrix notation used to estimate the coefficients with the LS method and ridge regression is given below, respectively.

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

$$\hat{\beta} = (X'X + kI)^{-1}X'Y \tag{4}$$

According to the equation 4, in the case of multicollinearity, the variance and covariance of the regression coefficients increase in the $X'X$ [5, 11]. However, to overcome this problem, ridge trace (k) adds the diagonal elements of the $X'X$ matrix. k must be between 0 and 1. If the k=0, the parameter estimation is the same as LS [5]. The parameter estimation of the ridge regression in matrix notation is with equation 5 as given above. The main purpose of the ridge regression method is getting a smaller variance estimation than the LS method in the $X'X$ matrix.

III. Results and Discussion

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

TABLE I: Descriptive statistics

	N	Mean±Std. Dev.	Min-Max	CV (%)
Protein, %	200	3.57±0.33	2.90-4.37	9.27
Fat, %	200	4.23±0.68	2.71-6.05	15.96
Dry matter %	200	13.30±0.85	11.39-15.35	6.37
Urea, mg/l	200	222.05±53.69	98.00-398.00	24.18
Milking days, days	200	30.25±8.29	9.60-55.50	27.41

Pearson correlation coefficient and the significance of the correlation coefficients between milk protein content and explanatory variables are given Table 2.

TALE II: Correlation matrix

	PC	FC	DM	UC	MD
PC	1*				
FC	0.66*	1*			
DM	0.82*	0.95*	1*		
UC	0.06	-0.07	-0.06	1*	
MD	-0.68*	-0.65*	-0.70*	0.12	1*

PC: Protein content, FC: Fat content, DM: Dry matter, UC: Urea content, MD: milking day

The estimated regression coefficients, standard error, test statistics (t-test) and collinearity statistic for VIF values obtained from the linear regression method are given in Table 3.

TABLE III. Linear regression and collinearity statistics

	β	Std. Error	t	Sig.	VIF
Intercept	-3.212	0.475	-6.750	0.000	-
MD	-0.008	0.002	-3.598	0.000	2.074
FD	-0.550	0.055	-9.938	0.000	9.316
DM	0.692	0.048	14.372	0.000	11.075
UC	0.001	0.0002	2.594	0.010	1.019

PC: Protein content, FC: Fat content, DM: Dry matter, UC: Urea content, MD: milking day

According to the Table 3, all regression coefficients were determined statistically significant (p<0.05). However, there was determined a multicollinearity problem for dry matter content, due to the VIF was greater than the 10.

For this reason, it has been determined that there is a need for a new method that will eliminate multicollinearity and provide more reliable results. For this aim, ridge regression was used to overcome this problem. The first step in the ridge regression is determine the optimal ridge trace (k) value. Table 4 shows the optimal k value.

According to the Table 4, When the value of k is 0, it means that the analysis method is performed according to LS. In addition, the optimal k value was determined as k=0.0028 when the VIF value is lower than 10. The model obtained when performing the ridge regression with the k values determined as 0.0028 is given in Table 5.

According to the Table 5, all regression coefficients were determined statistically significant for ridge regression with the k=0.0028 (p<0.05). In addition, it can be seen that the significance of the regression coefficients was the same. Furthermore, the model was given a trustworthy result.

TABLE IV: Optimum k values selection according to the VIF

k values	MD	FC	DM	UC
k=0	2.07413	9.31581	11.07466	1.01921
k=0.001	2.05705	8.98257	10.66456	1.01678
k=0.002	2.0405	8.66791	10.27742	1.01438
k=0.0021	2.03887	8.63741	10.23991	1.01414
k=0.0022	2.03725	8.60708	10.2026	1.0139
k=0.0023	2.03563	8.57693	10.1655	1.01366
k=0.0024	2.03402	8.54694	10.12861	1.01343
k=0.0025	2.03241	8.51711	10.09193	1.01319
k=0.0026	2.03081	8.48746	10.05545	1.01295
k=0.0027	2.02921	8.45796	10.01918	1.01271
k=0.0028	**2.02762**	**8.42863**	**9.98311**	**1.01248**
k=0.003	2.02445	8.37046	9.91156	1.012

TABLE V:. The summary of the ridge regression

	Estimate	Std. Error	t-value	Sig.
Intercept	-3.2082	0.6853	-4.6816	0.000
MD	-0.0076	0.0021	-3.6148	0.000
FC	-0.5499	0.0551	-9.9759	0.000
DM	0.6915	0.0479	14.4295	0.000
UC	0.0006	0.0002	2.6050	0.010

TABLE VI:. Goodness of fit criteria from the obtained models

	Linear regression		Ridge regression	
	Train	Test	Train	Test
RMSE	0.139	0.126	0.139	0.125
PC	0.915	0.910	0.915	0.910
R^2	0.836	0.825	0.836	0.825
AIC	-511.515	-277.239	-511.514	-277.206

According to the model comparison criteria used in Table 6, it is possible to interpret the best model for the lowest RMSE and AIC and the highest Pearson’s correlation coefficient (PC) and R^2 value [12]. In table 6 examined, ridge regression, used as an alternative method to LS, has the lowest RMSE and AIC and the highest PC and R^2 value.

IV. Conclusion

In our studies, we have shown that in predicting protein content based on variables between which there is collinearity, ridge regression should be used instead of linear regression. The results showed that ridge regression is more reliable model estimation than LS.

References

[1] FAO, (2019). Crops and livestock products. (Last Accessed Time: 21/11/2021). URL: https://www.fao.org/faostat/en/#data/QCL.

[2] Ozek, K. (2015). Factors Affecting Composition of Milk in Dairy Cattle and Relation between Nutrition and Milk Composition. Journal of Bahri Dagdas Animal Research 4 (2):37-45.

[3] Duchemin, S.I., Nilsson, K., Fikse, W.F., Stålhammar, H., Johansen, L.B., Hansen, M.S., Lindmark-Månsson, H., de Koning, D.J., Paulsson, M., Glantz, M., 2020. Genetic parameters for noncoagulating milk, milk coagulation properties, and detailed milk composition in Swedish Red Dairy Cattle. Journal of Dairy Science, 103(9), 8330-8342. https://doi.org/10.3168/jds.2020-18315

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

[5] 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.

[6] R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[7] Revelle W. (2020) psych: Procedures for Personality and PsychologicalResearch, Northwestern University, Evanston, Illinois, USA. https://CRAN.R-project.org/package=psych Version = 2.0.12.

[8] Imdad MU, Aslam M. 2018a. lmridge: linear ridge regression with ridge penalty and ridge statistics. URL: https://CRAN.R-project.org/package=lmridge, R package version 1.2.

[9] Eyduran E. 2020. ehaGoF: calculates goodness of fit statistics. R package version 0.1.1. URL: https://CRAN.R-project.org/package=ehaGoF.

[10] Hoerl, A.E. and Kennard, R., 1970. Ridge regression: Biased estimation for non-orthogonal problems. Technometrics, 12: 55-67. https://doi.org/10.1080/00401706.1970.10488634

[11] Vupa ve Alma, 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.

[12] Tatliyer A. 2020. The effects of raising type on performances of some data mining algorithms in lambs. KSU J Agric Nat, 23(3): 772-780.

+About Author (s):

Dr. Cem Tırınk

Scientific interests: Biometry, determinants of production and functional traits, genetic parameters, breeding value.

Prof. Dr. Hab. Dariusz Piwczyński

Scientific interests: Biometry; Sheep and dairy cattle breeding – determinants of production and functional traits, genetic parameters, breeding value; assessment of the influence of the automatic milking system on production and functional traits