RESEARCH ARTICLE

Study on Multiple Linear Regression and Principal Component Analysis for Prediction of Lifetime Performance of Kankrej Cattle

Radha Rani Sawami[1]*, Virendra Kumar[2], Urmila Pannu[3]

Abstract

The present investigation was conducted on 274 Kankrej cattle maintained at Livestock Research Station, Kodamdesar, Bikaner, calved between 2012 to 2022 with the objectives to study principal component analysis (PCA) and multiple linear regression analysis (MLRA) for prediction of lifetime performance of Kankrej cattle. Six early lactation traits (First lactation length- FLL, First lactation dry period- FDP, First lactation 305 days' milk yield- F305DMY, Second lactation length- SLL, Second lactation dry period- SDP, and Second lactation 305 days' milk yield- S305DMY) were used to analyze the lifetime milk yield upto 5th and 7th lactations. The MLR analysis revealed that the model containing F305DMY, SLL and S305DMY for upto 5th lactations' lifetime milk yield, and model containing F305DMY, FDP and SLL for upto 7th lactations' lifetime milk yield having R^2 = 68.3% and 68.5%, respectively, were found to be optimal models. PCA revealed that the first 2 principal components (FLL, F305DMY) explained more than 78% of the total variation for LTMY5 and more than 81% variation for LTMY7. In this study F305DMY was found most important early trait in prediction of lifetime production of Kankrej cattle on the basis of PCA and MLR analysis, out of which PCA was found to be better. Significant finding of this study may be helpful in developing selection methodology for Kankrej cattle after validation in a large population.

Key words: Kankrej, Lifetime performance, Multiple regression analysis, Principal component analysis.

Ind J Vet Sci and Biotech (2024): 10.48165/ijvsbt.20.2.12

Introduction

In India, mixed crop and livestock farming has been a way of life since the beginning of civilization, in which cattle play a key role. Among the indigenous cattle, Kankrej is one of the important breeds of cattle in India which is mainly found in the region of north Gujarat and neighboring districts of Rajasthan. Though, Kankrej is dual purpose breed, but also good milk producer (Ekka *et al.,* 2014; Gupta *et al*., 2019). The average milk yield of Kankrej cattle is around 1,738 kg with fat content around minimum 2.9 to maximum 4.2% (NBAGR, 2019). The lifetime performance of dairy animals is used to determine their overall productivity rather than their performance during a single lactation. Because dairy farmer's main objective is to optimize milk output and profitability, lifetime milk production is a significant economic factor. The generation interval and expenses related to maintaining less productive animals may be reduced if animals are selected for lifelong production on the basis of qualities expressed in their early lifespan.

Multiple linear regression analysis generally explains the relationship between multiple independent or predictor variables and one dependent or criterion variable. A dependent variable is modeled as the function of a number of independent variables, each with a corresponding coefficient, as well as a constant term. The prediction models are being fitted using multiple linear regression (MLR) analysis, with coefficient of determination (R^2) being used as a criterion to assess the prediction accuracy of the models. Principal component analysis (PCA) is a mathematical technique by which a number of correlated variables are converted into uncorrelated variables called principal components (Lukibisi *et al.,* 2008; Dangar and Vataliya, 2022). The factor from Principal component analysis (PCA) can be applied to breeding programmes with a sufficient reduction in the number of first lactation traits to be recorded

[1,3]Department of Animal Genetics and Breeding, College of Veterinary and Animal Science, Rajasthan University of Veterinary and Animal Sciences, Bikaner -334001, India

[2]Livestock Research Station, College of Veterinary and Animal Science, Rajasthan University of Veterinary and Animal Sciences, Bikaner -334001, India

Corresponding Author: Radha Rani Sawami, Department of Animal Genetics and Breeding, College of Veterinary and Animal Science, Rajasthan University of Veterinary and Animal Sciences, Bikaner -334001, India. e-mail: radharani7737156601@gmail.com

How to cite this article: Sawami, R. R., Kumar, V., & Pannu, U. (2024). Study on Multiple Linear Regression and Principal Component Analysis for Prediction of Lifetime Performance of Kankrej Cattle. Ind J Vet Sci and Biotech. 20(2), 59-63.

Source of support: Nil

Conflict of interest: None

Submitted 26/10/2023 **Accepted** 19/11/2023 **Published** 10/03/2024

for an explanation of the maximum variability for lifetime performance traits (Ratwan *et al.,* 2017). Bhattacharya and Gandhi (2005) predicted LMY up to 6 years and 8 years in Karan Fries cattle using AFC, FL305MY, FCI and FDP. Kumar and Hooda (2013) used multiple linear regression model for prediction of milk yield in crossbred cattle, while Dangar and Vataliya (2022) used PCA model for prediction of milk yield in Gir cattle. The objective of this study was to evaluate principal component analysis (PCA) and multiple linear regression (MLR) analysis for prediction of lifetime performance of Kankrej cattle.

Materials and Methods

The performance records of Kankrej cattle maintained at Livestock Research Station, RAJUVAS, Bikaner (India), were used for the present investigation. Bikaner has a geographical location of East Longitude 28°1′ and North Latitude 73°19′, situated at an average altitude of 797 Feet, witnesses' extreme temperatures. The climate of the region is characterized as semi-arid where temperature reaches up to 48°C during the summer. Six early lactation traits (FLL, FDP, F305DMY, SLL, SDP, and S305DMY) were used to analyze the lifetime milk yield upto 5th and 7th lactations. Computer package programme, General Linear Model (Univariate) procedure of IBM SPSS (2005) version 26.0 was used for data analysis. Multivariate regression analysis model formulated is as follows;

$$y= \beta_0+ \beta_1x_1+\ldots. + \beta_n x_n + \varepsilon$$

where, y =dependent variable; $x_1, x_2\ldots, x_n$ =independent variables; β_0 =intercept, β_1, $\beta_2\ldots\beta_n$=regression coefficients; ε =error.

The regression analysis was carried out by using "Backward Elimination Procedure" to determine the best combination of the independent variables contributing significantly towards the prediction of dependent variable as described by Draper and Smith (1987).

We get the principal components by ranking the Eigen vectors in order of their Eigen values, highest to lowest, 'in order of significance' and then choose the top Eigen vectors. The principal components scores were calculated by applying the formula:

$$Z_i = A_i X$$

where, $Z_i = i^{th}$ principal component score, $A_i = i^{th}$ latent vector (weight vector), and X = Original explanatory variables.

Results and Discussion

In prediction of lifetime production using MLRA and PCA, two traits were considered, *viz.* lifetime total milk yield upto 5 lactations (LTMY5) and lifetime total milk yield upto 7 lactations (LTMY7) and prediction was made on the basis of six early traits (FLL, FL305DMY, FDP, SLL, SL305DMY and SDP) using backward elimination procedure of multiple regression analysis.

Prediction of LTMY5 by MLRA

To predict LTMY5, all the traits under study were incorporated in the model. The accuracy of prediction was found 68.9%. The analysis revealed that the equation fourth (F305DMY, SLL, S305DMY) was found best for the prediction of LTMY5 on the basis of R^2 value (68.3%) and number of traits in optimum equation (Table 1).

Gandhi (1986) predicted LTP-5 on the basis of AFC, FCI, FL305DMY and FLL with 75.29 % R^2 value. Dalal *et al.* (2004) observed higher R^2 value by using first lactation length, first lactation milk yield and second lactation milk yield as independent traits for predicting LT5MY in Hariana cattle. Chander (1977) had observed lower coefficient of determination (R^2) in Tharparkar cattle by using AFC, FL305DMY and BE as independent traits for predicting LT5MY. Kumar (2003) observed lower R^2 in Sahiwal cattle by using AFC, F305DMY and FL305DFY as independent traits for predicting LT5MY.

Prediction of LTMY7 by MLRA

All six variables (FLL, FL305DMY, FDP, SLL, SL305DMY and SDP) were considered as independent variables to predict LTMY7 (equation I). This combination explained 68.9 % of total variation in LTMY7. The analysis revealed that the equation fourth (FDP, SLL, F305DMY) was found best for the prediction of LTMY7 on the basis of R^2 value (68.5%) and number of traits in optimum equation (Table 2).

Gopal and Bhatnagar (1972) predicted LTP-8 on the basis of AFC and FL305DMY and reported 51.12 % R^2 value in Sahiwal cattle. Similarly, Chander (1977) predicted LTP-8 in Tharparkar cattle on the basis of AFC, FL305DMY and BE and found that 57 % of total variation in LTP-8 was explained by this combination. The reported values were lower than the present estimated value. Gupta and Gurnani (1984) predicted LTP-8 in Tharparkar cattle on the basis of AFC FLTMY and FCI. The accuracy of prediction observed by them was 65.73 %; which was slightly lower than the R^2 value estimated under present investigation. Gandhi (1986) predicted LTP-8 on the

Table 1: Different models for prediction of LTMY5 by MLRA

S.No.	Model No.	Prediction equation	R^2 (%)
1.	1	LTMY5 = 1100.69 + 3.61 FLL + 1.48 F305DMY + 1.33 FDP + 6.12 SLL + 1.21 S305DMY -0.658SDP	68.9
2.	2	LTMY5 = 898.95 + 3.48 FLL + 1.48 F305DMY + 1.19 FDP + 6.74 SLL + 1.21 S305DMY	68.8
3.	3	LTMY5 = 1346.96 + 2.41 FLL + 1.49 F305DMY + 7.05 SLL + 1.18 S305DMY	68.5
4.	4	LTMY5 = 1492.20 + 1.82 F305DMY + 6.80 SLL + 1.21 S305DMY	68.3

Table 2: Different models for prediction of LTMY7 by MLRA

S.No.	Model No.	Prediction equation	R^2 (%)
1.	1	LTMY7 = 634.22 + 5.68 FLL + 2.83 F305DMY + 4.87 FDP + 19.79 SLL – 0.279 S305DMY - 2.02SDP	68.9
2.	2	LTMY7 = 687.24 + 5.05 FLL + 2.88 F305DMY + 4.93 FDP + 18.09 SLL – 1.94 SDP	68.9
3.	3	LTMY7 = 93.69 + 4.11 FLL + 2.94 F305DMY + 4.22 FDP + 18.09 SLL – 1.94 SDP	68.6
4.	4	LTMY7 = 372.16 + 3.46 F305DMY + 3.91 FDP + 20.82 SLL	68.5
5.	5	LTMY7 = 1757.11 + 3.07 F305DMY + 20.15 SLL	67.0

basis of AFC, FCI, FL305DMY, FLL, ASC, SCI, SL305DMY and SLL explaining 80.73 % of total variation in LTP-8.

Prediction using PCA

With principal components analysis a large number of independent variables can be systematically reduced to a smaller, conceptually more coherent set of variables.

Prediction of LTMY5 by PCA

In model, the first principal component was explained by maximum Eigen value (2.966) and this component expressed 49.43 % of the total variance under multivariate model (Table 3). Out of six principal components the first five principal components explained 98.61 % of the total variance revealing that these five components could be used for further analysis (Table 3). When all the six studied traits (FLL, FDP, F305DMY, SLL, SDP, S305DMY) were included 2 PC were extracted. The percent variance explained by each of the first four PC were 49.43%, 29.13%, 11.30% and 6.81%, respectively. Also, together with the two PC explained more than 78.57% of the total variance of the explanatory variables was explained (Table 3). Similar results were observed by applying Bartlett's Chi-square criterion.

Table 3: Eigen values, variance and cumulative % by different PC with studied traits for LTMY5

PC	Total Eigen values	Variance %	Cumulative %
1	2.966	49.43	49.43
2	1.74	29.13	78.57
3	0.67	11.30	89.87
4	0.40	6.81	96.69
5	0.11	1.91	98.61
6	0.08	1.39	100.00

Result of Kaiser-Meyer Olkin (KMO) measure of sampling adequacy (0.655) was suitable for the data evaluated statistically as Chi-square value was 591.599 with significant ($p \leq 0.001$) factor analysis application. The same model has been reported by many authors (Tolenkhomba *et al.,* 2013; Shah *et al.,* 2018; Romero *et al.,* 2018; Mujibi *et al.,* 2019; Dangar and Vataliya, 2022) studying different productive traits in cattle.

The same principal component analysis (PCA) model was used by Barbosa *et al.* (2006) and Santos *et al.* (2010) to evaluate the formation of predictive equation for the purpose to discriminate the traits most important for milk yield. The obtained three components accounted for 82.3% of the total variation. These results showed that most of the variations are explained by the first 3 principal components (PC) for all traits, whereas more than 90% (90.8%) of the total variations were explained by Mello *et al.* (2020). Furthermore, Khan *et al.* (2013) observed that the first PC showed 61.9% variation followed by second PC (26.1%), whereas more than 90.8% of the total variations were explained. In Gir cattle, first lactation milk yield, lactation length and peak milk yield, second lactation milk yield, lactation length and peak milk yield contributed 98% of the variance for lifetime production of milk yield. So based on these six records, prediction of an animal's production potential may give a better base of selection at an early age (Dangar and Vataliya, 2022).

There were two factors extracted with Eigen values greater than 1 and accounted for 78.57% of total variance. Investigation of scree plots showed that the first 2 principal components were informative enough (Fig. 1).

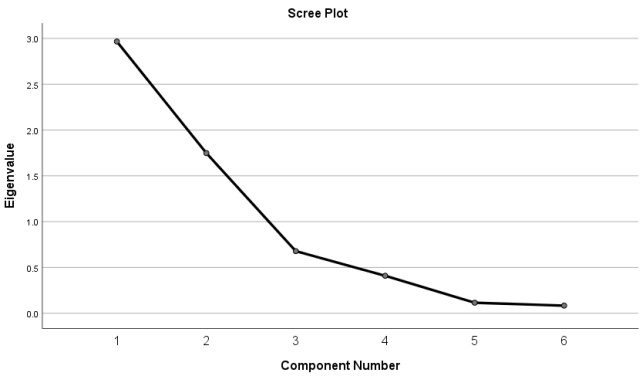

Fig. 1: Scree plot showing component number with Eigen values for LTMY5.

Varimax rotation, the widely used and accepted method was applied as it maximizes the sum of the variances of the squared loadings (squared correlations between variables and components). The coefficients of the PCA of the rotated component matrix of the two extracted principal components are given in Table 4. The component weights varied from -0.047 to 0.946 for first component for SDP and FLL, respectively. While, the second component weights varied from -0.045 to 0.939 for FDP and SLL, respectively.

Table 4: Estimates of principal component for studied traits using varimax rotation for LTMY5

Traits	Principal component	
	1	2
FLL	0.94	0.12
F305DMY	0.92	0.17
FDP	-0.80	-0.04
SLL	0.09	0.94
S305DMY	0.19	0.90
SDP	-0.04	-0.74

Prediction of LTMY7 by PCA

In this model, the first principal component was explained by maximum Eigen value (3.42) and this component expressed 57.09 % of the total variance under multivariate model. Out of six principal components, the first five principal components explained 99.05 % of the total variance revealing that these five components could be used for further analysis (Table 5). Similar results were observed by applying Bartlett's Chi-square. Result of Kaiser-Meyer Olkin (KMO) Measure of Sampling Adequacy (0.676) is suitable for the data evaluated statistically, and Chi-square = 205.047 with significant ($p \leq 0.001$) factor analysis application.

In present study, when all the six studied traits (FLL, FDP, F305DMY, SLL, SDP, S305DMY) were included 2 PC were extracted. The percent variance explained by each of the first four PC were 57.09%, 24.36%, 10.66% and 5.11%, respectively. Also, together with the two PC explained more than 81% of the total variance of the explanatory variables was explained (Table 5).

Table 5: Eigen values, variance and cumulative % by different PC with studied traits for LTMY7

PC	Total Eigen values	Variance %	Cumulative %
1	3.42	57.09	57.09
2	1.46	24.37	81.46
3	0.64	10.67	92.13
4	0.30	5.11	97.24
5	0.11	1.80	99.05
6	0.05	0.95	100.00

There were two factors extracted with Eigen values greater than 1 and accounted for 81.46% of total variance. Investigation of scree plots and cumulative explanation of principal components showed that the first 2 principal components were informative enough (Fig. 2).

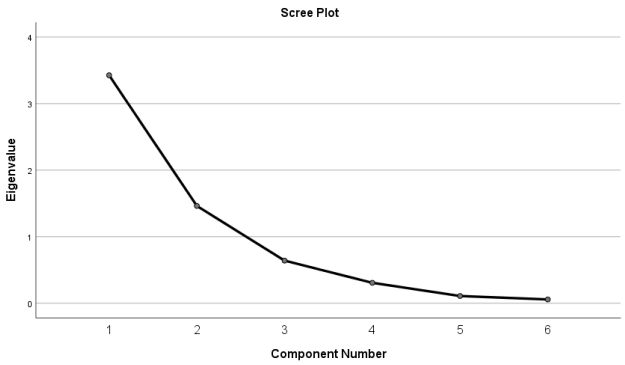

Fig. 2: Scree plot showing component number with Eigen values for LTMY7.

The coefficients of the PCA of the rotated component matrix of the two extracted principal components are given in Table 6. The component weights varied from -0.109 to 0.947 for first component for SDP to FLL, respectively. While, the second component weights varied from -0.243 to 0.928 for FDP and SLL, respectively. The correlation coefficients among various performance traits studied for lifetime milk yield upto both 5th and 7th lactation were highly significant (Table 7).

Table 6: Estimates of principal component for studied traits using varimax rotation for LTMY7

Traits	Principal component	
	1	2
FLL	0.947	0.170
F305DMY	0.941	0.128
FDP	-0.726	-0.243
SLL	0.152	0.928
S305DMY	0.337	0.866
SDP	-0.109	-0.846

Conclusion

On the basis of findings in the present investigation, it may be concluded that the prediction of lifetime milk production should be done only on the basis of first and second lactation production and reproduction traits having significant association and a favourable correlation with lifetime milk production. F305DMY was found most important early trait in prediction of lifetime production of Kankrej cattle on the basis of PCA and MLR analysis.

Table 7: Correlations among early traits for LTMY5 and LTMY7

Traits	FLL	F305DMY	FDP	SLL	S305DMY	SDP
Correlations for LTMY5, Determinant = 0.014						
FLL	1.000	0.912	-0.631	0.216	0.299	-0.110
F305DMY	0.931	1.000	-0.568	0.255	0.328	-0.132
FDP	-0.593	-0.560	1.000	-0.085	-0.166	0.185
SLL	0.329	0.292	-0.245	1.000	0.876	-0.544
S305DMY	0.496	0.422	-0.377	0.876	1.000	-0.478
SDP	-0.206	-0.208	0.395	-0.669	-0.622	1.000
Correlations for LTMY7, Determinant = 0.006						

Acknowledgements

The authors are thankful to the Director of Research, RAJUVAS, Bikaner and Dean, Veterinary College, Bikaner to provide the facility for research.

References

Barbosa, L., Lopes, P.S., Regazzi, A.J., Guimarães, S.E.F., & Torres, R.A. (2006). Evaluation of swine meat quality by principal component analysis. *Revista Brasileira de Zootecnia, 35,* 1639-1645.

Bhattacharya, T.K., & Gandhi, R.S. (2005). Principal components versus multiple regression analysis to predict lifetime production of Karan Fries cattle. *Indian Journal of Animal Sciences, 75*(11), 1317-1320.

Chander, S. (1977). Genetic evaluation of lifetime production and reproduction in Tharparkar cattle. *M.Sc. Thesis.* Kurukshetra University, Kurukshetra (Haryana), India.

Dalal, D.S., Malik, Z.S., Chhikara, B.S., & Chander, R. (2004). Prediction of lifetime milk production from early lactation trait in Hariana cattle. *Indian Journal of Animal Sciences, 74*(11), 1145-1149.

Dangar, N.S., & Vataliya, P.H. (2022). Prediction of lifetime milk yield using principal component analysis in Gir cattle. *Indian Journal of Veterinary Science and Biotechnology, 18*(4), 92-96.

Draper, N.R., & Smith, H. (1987). *Applied Regression Analysis*. John Wiley & Sons, Inc., New York.

Ekka, P., Gupta, J.P., Pandey, D.P., & Shah, R.R. (2014). Effect of non-genetic factors on first lactation and reproduction traits in Kankrej cattle of North Gujarat. *Indian Veterinary Journal, 91*(12), 18-20.

Gandhi, R.S. (1986). Selection of optimum combinations of early traits for maximising genetic improvement in dairy cattle. *Ph.D. Thesis*. Kurukshetra University, Kurukshetra (Haryana), India.

Gopal, D., & Bhatnagar, D.S. (1972). The effect of age at first calving and first lactation yield on lifetime production in Sahiwal cattle. *Indian Journal of Dairy Science, 25*, 129-133.

Gupta, A.K., & Gurnani, M. (1984). Prediction of lifetime production of milk on the basis of early economic traits in Tharparkar cattle. *Asian Journal of Dairy Research 3*(4), 201-207.

Gupta, J.P., Prajapati, B.M., Chaudhari, J.D., Pandey, D.P., Panchasara, H.H., & Prajapati, K.B. (2019). Impact of environmental trend in relation to genotypic and phenotypic trend on traits of economic interest in Kankrej cattle. *Indian Journal of Animal Sciences, 89*(11), 1255-1261.

Khan, T.A., Tomar, A.K.S., Dutt, T., & Bhushan, B. (2013). Principal component regression analysis in lifetime milk yield prediction of crossbred cattle strain Vrindavani of North India. *Indian Journal of Animal Sciences*, *83* (12), 1288-1291.

Kumar, D. (2003). Genetic studies on breeding efficiency in crossbred cows. *Indian Journal of Animal Sciences, 73*(10), 1180-1181.

Kumar, H., & Hooda, B.K. (2013). Prediction of lifetime milk production from early lactation traits in crossbred cattle. *Trends in Biosciences, 6*(1), 95- 96.

Lukibisi FB, Muhuyi WB, Muia JMK, Ole Sinkeet SN., & Wekesa WF. (2008). Statistical use and interpretation of principal component analysis in applied research. *Egerton University's 3rd Annual Research Week and International Conference,* p. 16-18.

Mello, R.R.C., Sinedino, L.D.P., Ferreira, J.E., de Sousa, S.L.G., & de Mello, M.R.B. (2020). Principal component and cluster analyses of production and fertility traits in Red Sindhi dairy cattle breed in Brazil. *Tropical Animal Health and Production, 52*(1), 273-281.

Mujibi, F., Rao, J., Agaba, M., Nyambo, D., Cheruiyot, E.K., Kihara, A., Zhang, Y., & Mrode, R. (2019). Performance evaluation of highly admixed Tanzanian smallholder dairy cattle using SNP derived Kinship Matrix. *Frontiers in Genetics, 10*, 375-387.

NBAGR (2019). National Bureau of Animal Genetic Resources, Annual Report 2019. Karnal, Haryana, India.

Ratwan, P., Mandal, A., Kumar, M., & Chakravarty, A.K. (2017). Prediction of lifetime performance traits by principal component analysis in Jersey crossbred cattle at an organized farm of eastern India. *Indian Journal of Animal Sciences 87*(9), 1163-1167.

Romero, J., Benavides, E., & Meza, C. (2018). Assessing financial impacts of subclinical mastitis on Colombian dairy farms. *Frontiers in Veterinary Science, 5,* 273-284.

Santos, E.F.N., Santoro, K.R., Ferreira, R.L.C., Santos, E.S., & Santos, G.R. (2010). Formation of productive genetic groups in dairy cows through principal components. *Revista Brasileira de Biometria, 28,* 15-22.

Shah, W.A., Ahmad, N., Javed, K., Saadullah, M., Babar, M.E., Pasha, T.N., & Saleem, A.H. (2018). Multivariate analysis of Cholistani cattle in Punjab Pakistan. *The Journal of Animal & Plant Sciences, 28,* 940-944.

Tolenkhomba, T.C., Singh, N.S., & Konsam, D.C. (2013). Principal component analysis of body measurements of bulls of local cattle of Manipur, India. *Indian Journal of Animal Sciences, 83*(3), 281- 84.