Original Article

Buffalo Bulletin (July-September 2024) Vol.43 No.3
https://doi.org/10.56825/bufbu.2024.4334036

PRINCIPAL COMPONENT REGRESSION ANALYSIS TO PREDICT LIFETIME MILK YIELD OF JAFFARABADI BUFFALOES

Nikhil Dangar[1,*] and Pravin Vataliya[2]

Received: 04 December 2020
Accepted: 13 September 2024

ABSTRACT

The study aims to devise most appropriate prediction model for lifetime milk production of Jaffarabadi Buffalo, based on principal components formulated on initially expressed lactation records as predictors. Lactation milk yield, lactation period and peak milk yield records of first, second and third lactations of animals under study were used of 24 years (1987 to 2010). Principal components (PCs) were derived from data set using principal component regression analysis (PCRA), the principal components were used as predictors for predicting lifetime milk yield (LTMY). Multiple linear regression models were fitted to identify the best fitted model for prediction of lifetime milk yield with the first principal component to all principal component as a predictor. The equation LTMY = 7825.8768+2.8118 (PC1) - 13.7098 (PC2) - 599.0908 (PC3) + 3.0266 (PC4) - 8.8196 (PC5) - 257.9315 (PC6) + 2.6042 (PC7) explained 98.9% variation in the estimated values with adjusted R^2= 59.09% variation in the estimated values. The curve estimation analysis showing the appropriateness of first seven principal components as predictor was the most appropriate model for lifetime milk yield. These prediction equations may be helpful in selection at an early stage of Jaffarabadi Buffalo based on early part lactation records.

Keywords: *Bubalus bubalis*, buffaloes, principal component analysis, prediction, Jaffarabadi buffalo

INTRODUCTION

Livestock with high production efficiency is desirable attribute in economical view that ultimately targets for genetic up gradation. In fact, the performance parameters of dairy animals directly connected with the economy of dairy industry; therefore, development of particular guideline for selection becomes more relevant to tackle out the means for ameliorating the performance efficiencies. Maximum genetic gain per unit of time for various traits of economic important is main aim of an animal breeder in a breed improvement programme. In dairy buffalo breeding, this implies maximizing genetic gain mainly for milk yield and production efficiency traits. This calls for assessing change in the genetic constitution as well as environmental

[1]College of Veterinary Science and Animal Husbandry, Navsari Agricultural University, Gujarat, India, *E-mail: drnik2487@gmail.com

[2]Faculty of Extension Education, Kamdhenu University, Gujarat, India

(managemental) conditions over time in organized herds for a particular breed evaluation and its breeding programme. Effectiveness of breeding programme one can see by magnitude and direction of production trends in a herd and help in bringing further improvement by developing or modifying appropriate strategies. Therefore, in early age prediction of most important production potential of an animal is at prime importance now a day. In this era of genomics, prediction of economic traits at early age is on aim for various scientists working with it. Various analytical methods such as multiple linear regression and principal component analysis like some of the basic breeding methodology also providing some assumptions to do the same by using it. Looking these facts present study was designed to predict the production potential of an animal at early age using part production records.

MATERIALS AND METHODS

In order to achieve the objective, the data pertinent to production traits on 118 Jaffarabadi buffaloes calving from 1987 to 2010, progeny of 28 sires maintained at Cattle Breeding Farm, Junagadh, Gujarat, India were considered. The duration of 24 years was classified into 6 periods of four years each. Winter (November-February), summer (March to June) and monsoon (July to October) on the basis of geo-climatic conditions prevailing in the region were delineated as three seasons. Total 12 Parity was considered for the study. First lactation milk yield (FTLY), first lactation length (FLL), first lactation peak milk yield (FPMY), second lactation milk yield (STLY), second lactation length (SLL), second lactation peak milk yield (SPMY), third lactation milk yield (TTLY), third lactation length (TLL), third lactation peak milk yield (TPMY), were recorded with lifetime milk production of the same buffalo for the principal component analysis. Buffalos with some specific or non-specific diseases, reproductive disorder and physical injury records were excluded from the present investigation. To find the explanatory variables for the highest determination of coefficients forward selection strategy was used and that worked as new explanatory variables added to the model to achieve maximum determination of coefficients. The model equation is given as

$$Yijkmn = \mu + Pi + Cj + Lk + Sm + eijkmn$$

Where, Yijkmn-Performance trait of individual animal (n), calved in $(i)^{th}$ period and $(j)^{th}$ season, of the $(k)^{th}$ parity, born to $(m)^{th}$ sire, μ-overall population mean, Pi fixed effect of period of calving (i = 1 to 6), Cj-fixed effect of season of calving (j = 1 to 3), L_k-fixed effect of parity (k = 1 to 12), Sm-random effect of sire (m = 1 to 52), ijkmn-random error with mean zero and variance σ^2E.

Principal component analysis is a method for transforming the variables in a multivariate data set, x1, x2, ..., xp, into new variables, y1, y2, ..., yp which are uncorrelated with each other and account for decreasing proportions of the total variance of the original variables defined as:

$$y1 = a11x1 + a12x2 + ... + a1pxp$$
$$y2 = a21x1 + a22x2 + ... + a2pxp$$
$$yp = ap1x1 + ap2x2 + ... + appxp$$

with the coefficients being chosen so that y1,y2,...,yp account for decreasing proportions of the total variance of the original variables, x1,x2,...,xp. Principal component analysis with correlation matrix was used to find the relationship among

FTLY, FLL, FPMY, STLY, SLL, SPMY, TTLY, TLL and TPMY and other fixed effects including breed, year at calving, season, and parity. Since scales of measurements of the production traits were different; correlation matrix was used instead of covariance matrix. According to cumulative explanatory proportions number of principal components was chosen and corresponding scores were estimated. Then based on these scores, regression analyses were done again; coefficients of determination based on explanatory variables regression and based on principal component regression scores were compared.

Eigen values (> 1) and scree plot methods were used to identify the principal components to be retained as predictors. Residual and diagnostics plots were examined to find out appropriate fitted model. Curve estimation analysis was undertaken to know the appropriateness of the models (lifetime milk yields with first principal as predictor). The adequacy of the best fitted model was adjudged based on adjusted R^2-value and significance of coefficients.

RESULTS AND DISCUSSIONS

The main objective of this study was to conduct a phenotypic analysis exploring relationships and dependencies among a group of traits that significantly affects the production of Jaffarabadi buffalo (Production traits). This investigation was based on data collected from Jaffarabadi buffalo kept in an experimental research farm for mostly scientific experiments, using exploratory analysis by principal components analysis. These statistical methods and concepts such as principal components regression are equally applicable to large volumes of data collected by research organizations. Although limited in data size, some of the exploratory analyses were statistically significant and will be useful. The following sections provide results of our findings. Table 1 showing the descriptive analysis of data under present study.

Principal component analysis was performed using the explanatory variables based on model. Investigation of scree plots (Figure 1) shows that most of the variations are explained by the first seven principal components. For the first seven principal components of all production traits, eigenvalues, proportions, and cumulative are given in Table 2. Table 3 shows results of determination coefficients based on regression analysis with explanatory variables, RSquare. However, dimension was reduced from 7 explanatory variables to 4 principal components using principal components instead of explanatory variables and it has changed the collinearity due to which variance inflation factors found one for all the traits on principal component regression. Looking to this, use of principal components instead of explanatory variables gained both reduction of the explanatory data set and broke the collinearity. After use of principal components and actual measurements comparison for predictions of observations shown in Figure 2 for first second and third lactation. Visual inspection of Figure 2 showed reasonably accurate predictions for all the traits since most of the points were lying around the straight line with a slope.

Principal component analysis could retain following first seven components explained 98.9% variation of the original variables. The first principal component showed 45.2% variation followed by second principal component 17.3%, third principal component 14.5% fourth principal component 8.9%, fifth principal component

Table 1. Number of records, N, means, standard deviations, and minimum, maximum for Milk Yield, lactation length and peak milk yield for selected lactation.

Variable	Mean	Std. Dev.	N	Min	Max
First Lactation Milk Yield	1804	810.91	116	566	4904
First Lactation Length	340.5	101.31	116	148.0	624.0
First Lactation Peak Milk Yield	8.087	2.51	116	4.650	16.400
Second Lactation Milk Yield	2095.4	861.97	116	762.9	5083.2
Second Lactation Length	344.3	94.16	116	165.0	690.0
Second Lactation Peak Milk Yield	9.746	3.25	116	5.200	21.400
Third Lactation Milk Yield	2108.4	778.69	116	758.5	4878.3
Third Lactation Length	328.9	80.71	116	135.0	581.0
Third Lactation Peak Milk Yield	10.25	2.75	116	5.30	18.80

Table 2. Eigen values and proportion of the variance of principal components (PC) of the correlation matrix of original variables.

PC	Eigen Value	Difference	Proportion	Cumulative
1	2.017	0.769	0.452	0.452
2	1.248	0.106	0.173	0.625
3	1.142	0.244	0.145	0.770
4	0.898	0.147	0.089	0.860
5	0.751	0.076	0.062	0.923
6	0.675	0.305	0.050	0.974
7	0.370	0.118	0.015	0.989
8	0.252	0.063	0.007	0.996
9	0.189	0.189	0.003	1.000

Table 3. Comparison of Determination Coefficients using Explanatory Variable (R square).

Variable	R^2 Value
First lactation milk yield	0.94
First lactation length	0.86
First lactation peak milk yield	0.78
Second lactation milk yield	0.91
Second lactation length	0.81
Second lactation peak milk yield	0.72
Third lactation milk yield	0.90
Third lactation length	0.80
Third lactation peak milk yield	0.76

Table 4. Significance of coefficients of fitted models and respective adjusted R2- values.

Predictor	B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	Adj R^2
First PC	5728.469*** (1339.162)	2.183** (0.678)									0.1412
Two PC	6721.637** (1967.231)	2.771* (1.088)	-6.028 (8.711)								0.1331
Three PC	14468.73*** (3863.787)	7.035** (2.129)	-27.373* (12.509)	-1010.523* (439.233)							0.1959
Four PC	7764.0613* (3320.5829)	3.9357* (1.7942)	-17.3590 (10.180)	-725.774* (355.5789)	3.142*** (0.5631)						0.4839
Five PC	7853.6674* (3825.5750)	3.9405* (1.8141)	-17.3872 (10.294)	-729.6333 (367.6443)	3.176*** (0.9096)	-0.3773 (7.761)					0.474
Six PC	10034.065* (3893.029)	3.147 (1.814)	-12.937 (10.289)	-568.340 (367.730)	4.996*** (1.290)	-10.402 (9.157)	-403.353 (207.727)				0.5006
Seven PC	7825.8768* (3579.5794)	2.8118 (1.6448)	-13.7098 (9.3152)	-599.0908 (332.9473)	3.0266* (1.2960)	-8.8196 (8.3003)	-257.9315 (192.5468)	2.6042*** (0.7439)			0.5909
Eight PC	7769.129* (3635.526)	2.857 (1.689)	-14.155 (9.880)	-601.934 (336.805)	3.067* (1.337)	-9.121 (8.627)	-258.966 (194.585)	2.506* (1.004)	1.172 (7.944)		0.5827
All PC	8341.0582* (3871.8273)	2.5526 (1.8284)	-13.1038 (10.2215)	-540.6170 (364.9623)	3.0712* (1.3482)	-9.0613 (8.6989)	-259.4029 (196.1749)	2.9214 (1.3584)	-0.9763 (9.2799)	-104.4255 (227.8601)	0.5759

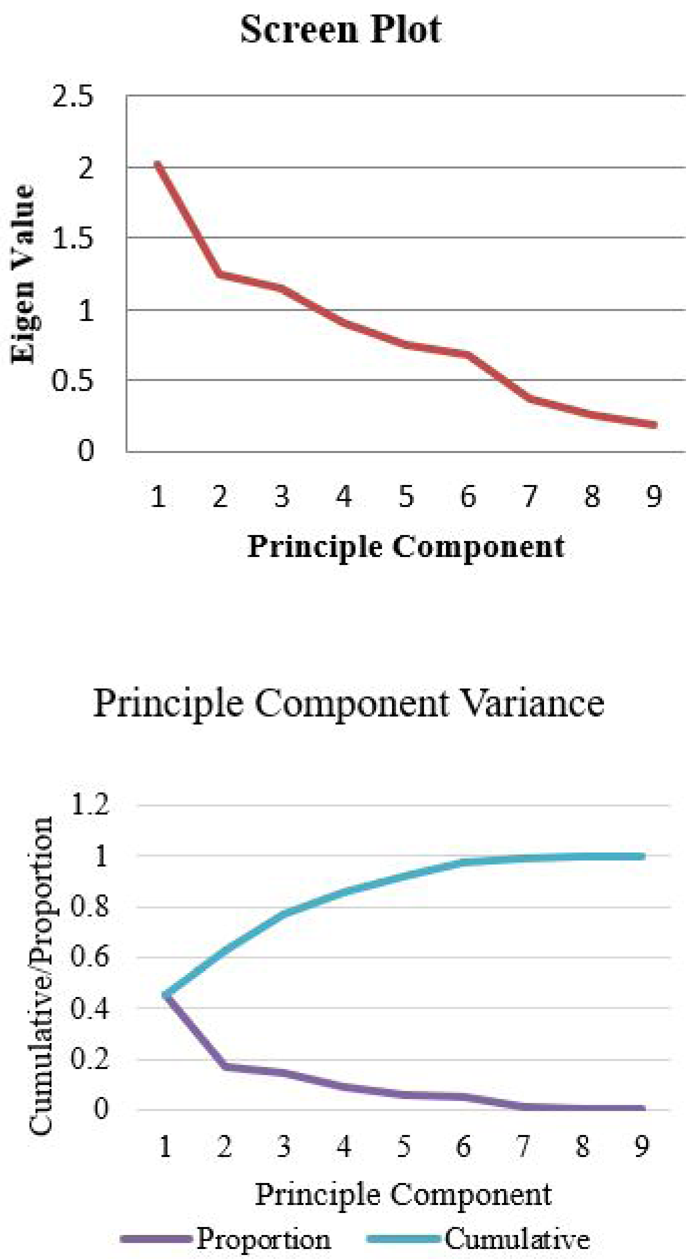

Figure 1. Screen plots of principal component analysis for lactation milk yield, lactation length and peak milk yield of first three lactation in combinations.

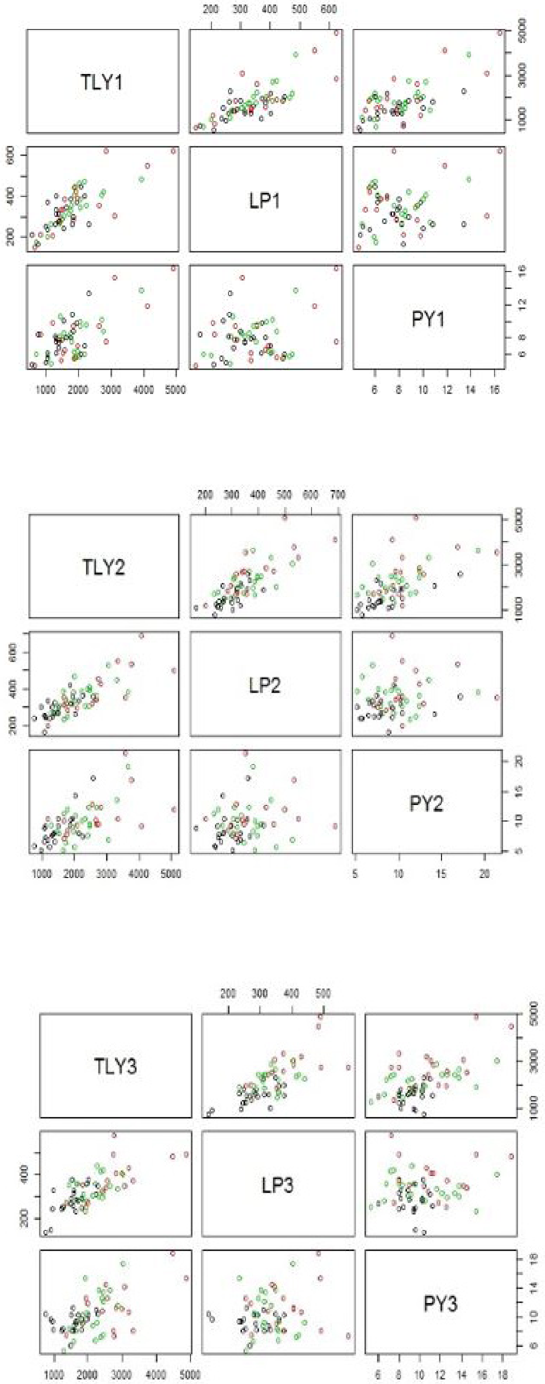

Figure 2. Observations and predictions based on the loadings of the principal components analysis with type traits for first second and third lactation milk yield, lactation length and peak milk yield. Straight middle line as base to check the prediction abilities.

6.2%, sixth principal component 5% and seventh principal component 1.5% (Table 1 and Figure 1). Here first three principal components had eigen values more than one, but scree plot showed the appropriateness of first seven components (clearly as curve bend at principal component seven). So, first seven principal components were retained to be used as predictors (Table 4). The equation LTMY = 7825.8768 + 2.8118 (PC1) - 13.7098 (PC2) - 599.0908 (PC3) + 3.0266 (PC4) - 8.8196 (PC5) - 257.9315 (PC6) + 2.6042 (PC7) explained 98.9% variation in the estimated values with adjusted R^2= 59.09% variation in the estimated values.

The study was to show the use of principal components regression analysis as principal components are orthogonal contrasts, free from problem of multicollinearity. Since expression of animals traits on growth, production and reproduction is very much complex and correlated, the information generated out of all the traits studied, can be included as null hypothesis as their contribution towards the lifetime production. We may formulate the principal components to reduce the data (into components) with the variability explained in the original set of observed traits (variables), as discussed by many workers (Khan *et al*., 2013; Hotelling 1933; Chapman *et al*., 2001; Rugoor *et al*., 2000).

Variation of 40.32% was reported in estimated lifetime yields (total of first 4 lactations-LTMY4) with initial growth, reproduction, part lactation records with stepwise procedure of regression analysis in Vrindavani cattle. Whereas variation of 54.46% was estimated in the same crossbred strain by considering part lactation records to estimate it using principal components by Khan *et al*. (2012). Comparison of multiple regression analysis and principal components analysis was done to predict lifetime milk production and found that total variance was lower from the model having PCs as compared to original variables in the regression model by Bhatacharya and Gandhi (2005). This showed the principal component regression analysis (PCRA) has much of importance for estimation of lifetime production traits.

Based on the results of this experiment, it is concluded that 98% of variance for lifetime production of milk yield was contributed by six early ages records: first lactation milk yield, lactation length and peak milk yield, second lactation milk yield, lactation length and peak milk yield. Hence, prediction based on these six production potential records of an animal may give better base for early age of selection.

REFERENCES

Bhatacharya, T.K. and R.S. Gandhi. 2005. Principal components versus multiple regression analysis to predict lifetime production of Karan Fries cattle. *Indian Journ. Anim. Sci*., **75**(11): 1317-1320.

Chapman, K.W., H.T. Lawless and K.J. Boor. 2001. Quantitative descriptive analysis and principal component analysis for sensory characterization of ultrapasteurized milk. *J. Dairy Sci*., **84**(1): 12-20. DOI: 10.3168/jds.S0022-0302(01)74446-3

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol*., **24**(6): 417-441. DOI: 10.1037/h0071325

Khan, T.A., A.K.S. Tomar and D. Triveni. 2012. Prediction of lifetime milk production in synthetic crossbred cattle strain. *Vrindavani* of North India. *Indian J. Anim.*

Sci., **82**(11): 1367-1371. DOI: 10.56093/ijans.v82i11.25156

Khan, T.A., A.K.S. Tomar, D. Triveni and B. Bhushan. 2013. Principal component regression analysis in lifetime milk yield prediction of crossbred cattle strain Vrindavani of North India. *Indian J. Anim. Sci.*, **83**(12): 1288-1291. DOI: 10.56093/ijans.v83i12.35805

Rougoor, C.W., R. Sundaram and J.A.M. van Arendonk. 2000. The relation between breeding management and 305-day milk production, determined via principal components regression and partial least squares. *Livest. Prod. Sci.*, **66**: 71-83. DOI: 10.1016/S0301-6226(00)00156-1