Indian Journal of Animal Sciences **83** (12): 1288–1291, December 2013/Article

Principal component regression analysis in lifetime milk yield prediction of crossbred cattle strain Vrindavani of North India

T A KHAN[1], A K S TOMAR[2], TRIVENI DUTT[3] and BHARAT BHUSHAN[4]

Indian Veterinary Research Institute, Izatnagar, Uttar Pradesh 243 122 India

Received: 20 April 2013; Accepted: 28 August 2013

ABSTRACT

The study aims to devise most appropriate prediction model for lifetime milk production of Vrindavani, a crossbred cattle strain developed and maintained at Institute, based on principal components formulated on initially expressed part lactation records as predictors. Principal components (PCs) were derived on a data set pertaining to 10-year period (1999–2009). Part lactation records of 100, 170 and 240 days of first and second lactations and their respective total milk yields were used. Using principal component regression analysis (PCRA), the principal components were used as predictors for predicting lifetime milk yield as total milk yield up to 4 lactations (LTMY4), and up to 5 lactations (LTMY5). Eight types of model were fitted to identify the best fitted model for both the traits (LTMY4 and LTMY5) with the first principal component (PC_1) as a predictor. The equation LTMY4=4410.305+0.596PC_1-1.171PC_3 and LTMY5 = 7987.560– 2.301PC_3 explained 54.46% and 39.74% variation in the estimated values. The curve estimation analysis showing the appropriateness of power function was the most appropriate model for both the lifetime traits. The model LTMY4=32.609 $(PC\_1)^{0.671}$ was found best fit and explained 53.50% variation in estimation. In the other lifetime trait, model LTMY5==103.769 $(PC\_1)^{0.566}$, explained 38.40% variation in estimated values. These prediction equations may be helpful in selection at an early stage of Vrindavani cattle based on early part lactation records.

Key words: Vrindavani cattle, Lifetime milk production, Part lactation, Principal component regression analysis

India with 210.2 million head of cattle contributes to 14.7% of the total cattle population of the world. The country's per capita milk availability in 2010–11 was at 281 g/day (compared to the World average of 279.4 g/day) (BAHS 2012). To increase milk availability, efficient planning and management tools are to be developed. Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components (Pearson 1901, Lukibisi *et al.* 2008a, Hotelling, 1933, Rao 1964). Chapman *et al.* (2001) used principal component analysis for sensory characterization of ultra pasteurized milk quality and found as much as 81.0% variation in estimation. Rugoor *et al*. (2000) investigated advantages and disadvantages of principal components regression (PCR) and partial least squares (PLS) for livestock management research. Bhatacharya and Gandhi (2005) compared multiple regression analysis and principal components to predict lifetime production of Karan Fries cattle and found advantage of principal components analysis.

Present address: [1]Senior Scientist (takhan@ivri.res.in), [2,4]Principal Scientist [2](drakstomar@gmail.com, [4]bhushan.drbharat @gmail.com), [3]Jt. Director (jdee@ivri.up.nic.in).

Lifetime production is an important economic parameter when defining the breeding objective. Khan *et al*. (2012), formulated the lifetime prediction model based on birth weight, 5 initially expressed reproductive traits age at first calving (AFC), first service period (FSP), first dry period (FDP), first calving interval (FCI) and first lactation length (FLL)) and part lactation records of 100, 170 and 240 days of first lactation, second lactation and their respective total milk yields and explained 40.32% variation in estimated lifetime yields (total of first 4 lactations) in Vrindavani cattle. Malhotra and Singh (1980) predicted lifetime production (total milk yield in the first 3 lactations) for Red Sindhi cows on the basis of traits available in early life. Puri and Sharma (1965) studied first lactation yield and age at first calving on lifetime production and determines the relative importance of them for selection purposes in Red Sindhi and crossbred cows. Shinde *et al.* (2010) predicted lifetime milk production up to third lactation in Phule Triveni cows. Vrindavani cattle is a synthetic crossbred cattle strain with exotic inheritance (50 to 75%) of Holstein-Friesian, Brown Swiss, Jersey and indigenous inheritance (25 to 50%) of Hariana cattle, developed at Indian Veterinary Research Institute, Izatnagar (Singh *et al. 2011)*. The climate of Izatnagar is semi tropical and conditions are hot humid for many months of the year, especially during summer and monsoon seasons.

Since data on traits expressed by animals are correlated and huge in size, present study was initiated to formulate prediction models for lifetime production based on principal components formulated on initially expressed part lactation records as predictors. These models may be used in early selection of animals.

MATRIALS AND METHODS

Ten year data records (1/4/1999 to 31/3/2009) of Vrindavani were used to formulate lifetime prediction models. The data set consisted of parity-wise part milk yields of 100, 170 and 240 days in first lactation (MY100_1, MY170_1, MY240_1), in second lactation (MY100_2, MY170_2, MY240_2) and their respective total milk yields (TLMY1, TLMY2) along with total milk yield of first 2 lactations (LTLMY2) and principal components (PCs) were derived using the *princomp* procedure of SAS 9.2. Principal component regression analysis (PCRA) was performed to predict lifetime milk yield up to 4 (LTMY4) and 5 parities (LTMY5) using principal components as predictors (Draper and Smith 1986, *Proc Reg* procedure, SAS 9.2).

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 (SAS Inc 2009).

Curve estimation analysis was undertaken to know the appropriateness of the models (lifetime yields with first principal as predictor). Eight types of models, viz. linear, quadratic, cubic, logarithmic, inverse, power, S-curve, and exponential were fitted for prediction of LTMY4 and LTMY5. The adequacy of the best fitted model was adjudged based on adjusted R^2-value and significance of coefficients (Draper and Smith 1986, SPSS Inc. 2003).

RESULTS AND DISCUSSION

Principal component analysis could retain following first 3 components explained 93.08% variation of the original variables. The first principal component showed 61.86% variation followed by second principal component (26.14%) and third principal component (5.09%) (Table 1; Fig. 1). Here first 2 PCs had eigen values more than one, but scree plot showed the appropriateness of first 3 components (clearly as curve bend at PC_3). So, first 3 PCs were retained to be used as predictors.

Table 1. 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	5.5674	3.2149	0.6186	0.6186
2	2.3525	1.8949	0.2614	0.8800
3	0.4577	0.1502	0.0509	**0.9308**
4	0.3074	0.1673	0.0342	0.9650
5	0.1401	0.0439	0.0156	0.9806
6	0.0962	0.0544	0.0107	0.9913
7	0.0419	0.0051	0.0047	0.9959
8	0.0368	0.0368	0.0041	1.0000

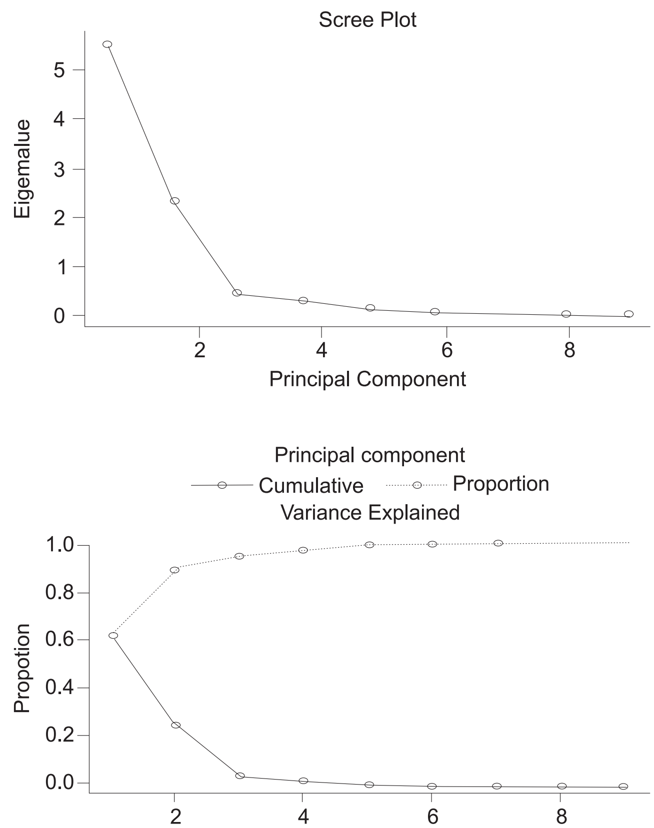

Fig. 1. Scree plots of Principal components formulated in Vrindavani cattle part milk yield data

The following 3 principal components (PCs) were retained:

PC_1 = 0.316 MY100_1 + 0.331 MY_170_1 + 0.331 MY240_1 +0.313TLMY1+ 0.309 MY100_2 + 0.336 MY170_2+ 0.346 MY240_2+ 0.312 TLMY2+ 0.398 LTMY2

PC_2 = 0.329 MY100_1 + 0.368 MY_170_1 + 0.381 MY240_1 + 0.350 TLMY1 -.339 MY100_2 -.357 MY170_2 -.341 MY240_2 -.360 TLMY2 -.016 LTMY2

PC_3 = 0.430 MY100_1 + 0.245 MY_170_1 + 0.046 MY240_1 -.442 TLMY1+ 0.398 MY100_2 + 0.199 MY170_2+ 0.051 MY240_2 -.336 TLMY2 -.494 LTMY2

The PCRA for lifetime milk yield (LTMY4) evolved equation, $LTMY4=4410.305^{**}+0.596^{*}PC\_1-1.171^{*}PC\_3$ and could explained 54.46% variation in the estimated values with adjusted R^2= 53.20% (*-Significant $p<0.05$, **-significant $P<0.01$).

In case of lifetime milk yield prediction up to 5 parities (LTMY5) the following evolved equation, LTMY5 = 7987.560**- 2.301**PC_3 could explained 39.74% variation in the estimated values of lifetime milk yield with adjusted R^2= 38.92%.

Fit diagnostics for LTMY4 and LTMY5 showed goodness of fit of the models. The quantile plot and histogram of residuals indicated observance of normality assumption. Residual plots indicated absence of heteroscedasticity (among the PCs). The observed values verses the predicted values plot indicated a good fit for the model. The fit-mean residual plot indicated that the fitted model accounted for a good deal of variability in both LTMY4 and LTMY5.

Curve estimation: The regression coefficients along with standard errors of the fitted models (lifetime yields-Y with first principal component-PC_1, as predictor) are presented in Table 2. The value of adjusted R^2, by the models for LTMY4 showed maximum variation (53.50%) by power function with minimum in inverse function (47.00%). Since in power function, regression coefficient was highly significant (P<0.01) and had maximum adjusted R^2, the function was decided to be more appropriate for estimating the LTMY4, though intercept was not significant (Fig. 2). Same trend was observed for LTMY5 and values of adjusted R^2, by the fitted models showed major variation (39.30%) explained by power function and minimum by inverse function (34.93%). So in both the lifetime traits (LTMY4 and LTMY5) the power function was more appropriate. The power function explained 38.40% variation in estimated values of LTMY5, which was found more appropriate (Fig. 3).

The study was to show the use of principal components

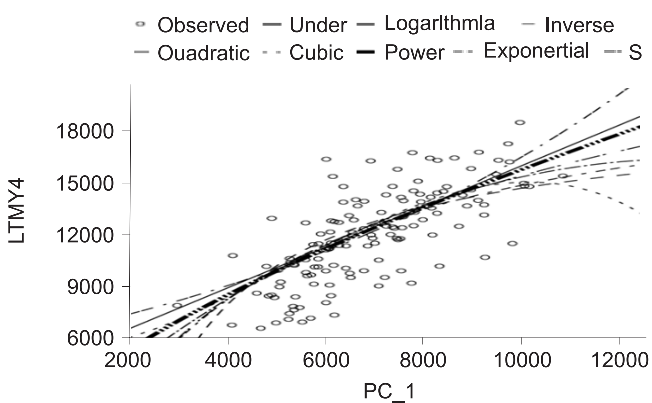

Fig. 2. Appropriateness of power function in estimation of LTMY4 (kg).

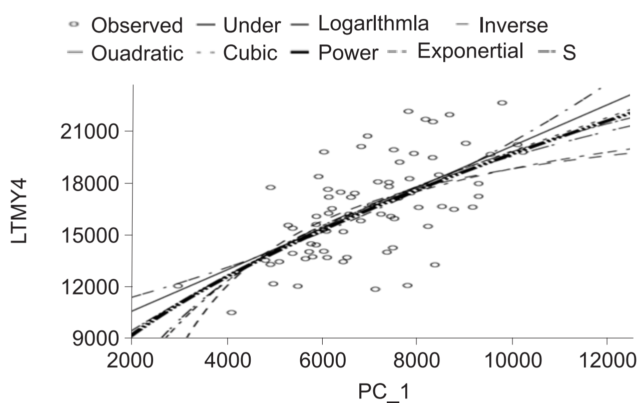

Fig. 3. Appropriateness of power function in estimation of LMMYS (kg).

Table 2. Significance of coefficients of fitted models and respective adjusted R^2- values

Model	LTMY4					LTMY5				
	b_0	b_1	b_2	b_3	Adj. R^2	b_0	b_1	b_2	b_3	Adj. R^2
Linear	4250.0** (973.143)	1.178** (0.136)	-	-	50.00	6609.843** (1352.684)	1.295** (0.189)	-	-	38.30
Quadratic	–127.389^{NS} (3472.125)	2.476* (0.997)	–9.2E–05^{NS} (7E–05)	-	50.50	4106.786^{NS} (4874.072)	2.037^{NS} (1.400)	–5E–05^{NS} (1E–04	-	37.30
Logarithmic	–56941.458** (7912.128)	7862.887** (895.449)	-	-	50.70	–60017.832** (11084.718)	8570.997** (1254.505)	-	-	38.20
Cubic	5425.791^{NS} (9216.147)	–0.217^{NS} (4.257)	3E–04^{NS} (0.001)	–2E–08^{NS} (3E–08)	50.10	3143.927^{NS} (12975.327)	2.505^{NS} (5.993)	–1E–04^{NS} (0.001)	4E–09^{NS} (4E–08)	36.80
Inverse	19333.760** (859.087)	–4562588.172** (558658.324)	-	-	47.00	23080.531** (1191.144)	–4942712.185** (7745930.581)	-	-	34.90
Power	32.609^{NS} (20.860)	0.671** (0.072)	-	-	53.50	103.769NS (74.238)	0.566** (0.081)	-	-	39.30
S-curve	10.007** (0.069)	–3959.246** (445.942)	-	-	51.30	10.138** (0.076)	–3313.178** (496.106)	-	-	37.10
Exponent	6122.418** (489.643)	9.9E–05** (1.1E–05)	-	-	51.30	8519.867** (750.036)	8E–05** (1E–05)	-	-	38.40

Figures in parenthesis are standard error of coefficient. *,>significant (P<0.05); **, > significant (P<0.01). NS, > nonsignificant.

regression analysis as PCs are orthogonal contrasts, free from problem of multicolliniarity. 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 (Hotelling 1933, Chapman *et al.* 2001, Rugoor *et al.* 2000).

PCRA with step-wise procedure, showed fitted models for LTMY4 and LTMY5 with the 3 predictors (PC_1, PC_2 and PC_3): LTMY4=4410.305**+0.596*PC_1-1.171*PC_3 and LTMY5 = 7987.560**- 2.301**PC_3 and could explain 54.46% and 39.74% variation in the estimated values of LTMY4 and LTMY5, respectively. Curve estimation analysis showed that power function was the most appropriate model for both the lifetime traits. The model LTMY4=32.609 $(PC\_1)^{0.671}$ was found best fit and explained 53.50% variation in estimation. The relationships LTMY5==103.769 $(PC\_1)^{0.566}$, explained 38.40% variation in estimated values of LTMY5. Since every biological phenomenon is curvilinear in nature and the variation explained in estimation of LTMY4 and LTMY5 by step-wise procedure and curve estimation are at par, these models are recommended as best fitted models. In both the traits, the regressions were highly significant (P<0.01) whereas intercepts were not significant (as the curve is crossing y-axis nearby the origin as observed in Figs 2, 3). Khan *et al*. (2012), found 40.32% variation in estimated life time yields (total of first 4 lactations- LTMY4) with initial growth, reproduction, part lactation records with step-wise procedure of regression analysis in Vrindavani cattle. Whereas principal components based on only part lactation records could estimate 54.46% variation in estimation in the same crossbred strain. Bhatacharya and Gandhi, (2005) have compared multiple regression analysis and principal components analysis 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. This showed the importance of principal component regression analysis (PCRA) in estimation of lifetime production traits. So, it is concluded that fitted prediction model LTMY4=32.609 $(PC\_1)^{0.671}$ and LTMY5==103.769 $(PC\_1)^{0.566}$, may be helpful in early selection of Vrindavani cattle based on initial part lactation records.

ACKNOWLEDGEMENT

Authors are thankful to Director and Incharge Livestock Production and Management Section, IVRI, for constant encouragement and providing facilities

REFERENCES

Basic Animal Husbandry Statistics. 2012. Department of Animal Husbandry Dairying and Fisheries. Ministry of Agriculture, Government of India. Krishi Bhavan, New Delhi.

Bhatacharya T K and 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): 1317–20.

Chapman K W, Lawless H T and Boor K J. 2001. Quantitative descriptive analysis and principal component analysis for sensory characterization of ultrapasteurized milk. *Journal of Dairy Science* **84**: 12–20.

Draper N R and Smith H. 1966. *Applied Regression Analysis*. Pp. 407. Wiley Press, New York, USA.

Hotelling H. 1933. Analysis of the complex of statistical variables into principal components. *Journal of Educational Psychology* **24**: 417–41, 498–520.

Khan T A, Tomar A K S and Dutt Triveni. 2012. Prediction of lifetime milk production in synthetic crossbred cattle strain *Vrindavani* of North India. *Indian Journal of Animal Sciences* **82**: 1367–71.

Lukibisi F B, Muhuyi W B, Muia J M K, Ole Sinkeet S N and Wekesa W F. 2008a. Statistical use and Interpretation of Principal Component Analysis in Applied Research. *Egerton University's 3rd Annual Research Week and International Conference*. 16–18 September, 2008.

Malhotra P K and Singh R P. 1980. Estimation of life-time production in Red Sindhi cattle using ridge-trace criterion. *Indian Journal of Animal Sciences* **50**(3): 215–18.

Puri T R and Sharma K N S. 1965. Prediction of lifetime production on basis of first lactation yield and age at first calving for selection of dairy cattle. *Journal of Dairy Science* **48**(4): 462–67.

Panda K K, Meheta R K and Das B C. 2006. Prediction of total lactation milk yield based on most frequent daily milk yield and highest daily milk yield of a month in Sahiwal cows. *Indian Journal of Animal Sciences* **76** (10): 851–52.

Pearson K. 1901. On lines and planes of closet fit to a system of points in space. *Philosophical Magazine* **2**: 557–72.

Rao C R. 1964. The use and interpretation of principal component analysis in applied research. *Sankhya* **A 26**: 329–58.

Rougoor C W, Sundaram R, van Arendonk J A M. 2000. The relation between breeding management and 305–day milk production determined via principal components regression and partial least squares. *Livestock Production Science* **66**: 71–83.

Singh R R, Dutt Triveni, Kumar Amit, Tomar A K S and Singh Mukesh. 2011. On-farm characterization of Vrindavani cattle in India. *Indian Journal of Animal Sciences* **81**(3): 267–71.

Shinde N V, Mote M G, Khutal B B and Jagtap D Z. 2010. Prediction of lifetime milk production on the basis of lactation traits in Phule Triveni crossbred cattle. *Indian Journal of Animal Sciences* **80**(10): 968–88.

SPSS Inc. 2003 User's Guide. SPSS Inc., Chicago.

SAS Institute Inc. 2009. *SAS/STAT ® 9.2User's Guide.* 2nd edn. Cary, NC: SAS Institute Inc.