COMPARISON BETWEEN DIFFERENT PROCEDURES TO DETERMINE THE RELATIVE IMPORTANCE OF THE LIFETIME PERFORMANCE TRAITS IN PREDICTING BREEDING VALUES OF HOLSTEIN COWS

M. A. M. Ibrahim

Animal Production Department, Faculty of Agriculture, Cairo University, Giza, Egypt

SUMMARY

Data used in this study comprised 2730 lactation records of 850 Holstein cows sired by 316 sires. The Holstein cows belong to a commercial farm. The objective of this work was to determine to what extent breeding value (BV) is influenced by lifetime variables (Total milk yield at first lactation (TMY1),Total milk yield at last lactation(TMYL), 305 milk yield at first lactation (305M1), 305 milk yield at last lactation (305ML), milk per day at productive life (Mday),number of complete lactations (NCL), lifetime days in milk (LDIM),productive life (Plife), age at disposal (CULL) and longevity index (LI, %)) using the stepwise multiple linear regression analysis (SMLR) and partial least squares regression (PLS) procedures.

The variables are important for the projection of breeding values (BV) were the variables measured-in-weight (TMY1, MDAY, TMYL, 305M1 and 305ML).The variables measured-in-duration (day or month or lactation) (NCL, LI , Cull , Plife and LDIM) have low influence on the BV estimations model . PLS resulted in R^2 of 0.64 and root mean squares for error) of 223. SMLR results were slightly lower R^2 of 0.61 with root mean squares for error of 231 using the same lifetime variables.

Keywords: Stepwise regression, partial least squares regression, breeding value, lifetime performance traits, Holstein cows

INTRODUCTION

Stepwise regression is a procedure for automatically selecting, in a stepwise manner variables that are close to optimal in the sense of maximizing the squared multiple correlation coefficients (R^2) of the dependent variable with a set of selected independent variables.

The main goal of PLS procedure is to minimize the sample prediction error, seeking linear functions of the predictors that explain as much as possible variation in each response. In addition, it has the additional goal of accounting for variation in the predictors, under the assumption that directions in the predictor space which are well sampled should provide better prediction for new observations when the predictors are highly correlated. All of the techniques implemented in the PLS procedure, which are derived from principal components regression (PCR), are to reduce errors in rank regression and PLS regression. The procedure works by extracting successive linear combinations of the predictors called either components or latent vectors to explain predictor variation (Abdi, 2003).

The main objective of this work was to determine to what extent estimates of breeding values are influenced by lifetime variables, and the variables importance for the projection of breeding values (BV) by SMLR and PLS using the same data.

MATERIAL AND METHODS

Data used in this study included 2730 lactation records for 850 Holstein cows sired by 316 sires. Data were collected from a commercial farm (International Company for Animal Wealth), located at Giza Governorate, Egypt. The records covered the period from 1991 to 2006.

Cows were imported as pregnant heifers from USA. Cows were artificially inseminated using frozen semen imported from USA and Canada. All cows were machine milked according to their level of production.

Definitions and abbreviations of the studied traits

- Breeding value (BV).
- Lifetime performance traits which are used in predicting BV :
  1. Total milk yield in first lactation (TMY1 ,Kg)
  2. Total milk yield in last lactation (TMYL ,Kg)
  3. 305 milk yield in first lactation (305M1 ,Kg)
  4. 305 milk yield in last lactation (305ML ,Kg)
  5. Milk per day in productive life (Mday, Kg)
  6. Number of complete lactations (NCL, lactation)
  7. Lifetime days in milk (LDIM ,day)
  8. Productive life or longevity (Plife, months) = the period from first calving to disposal form the farm,
  9. Age at disposal or lifetime (Cull, months) = the time between birth date and disposal date.
  10. Longevity index (LI, %) = lifetime days in milk divided by its longevity, expressed as a percentage. It measures the cow's lifetime efficiency, because it represents the days spent producing milk.

Means of the considered traits are given in Table 1. The correlation matrix between traits is presented in table 2.

Table 1. Descriptive statistics for breeding value and lifetime performance

Variable	Measuring unit	Minimum	Maximum	Mean	SD
BV	Kg	-766	1057	15	370
TMY1	Kg	1556	11010	9160	2675
TMYL	Kg	970	11335	7365	3650
305M1	Kg	1556	10875	7175	1380
305ML	Kg	970	11216	6200	2455
Plife	Month	15	142	48	20
MDAY	Kg	8	28	19	3.5
CULL	Month	40	169	75	19
NCL	Lactation	2	7	3.2	1.3
LDIM	Day	394	3828	1220	480
LI	%	26	81	52	8

Table 2. Correlation matrix between lifetime performance traits and breeding value

Variables	TMYL	305M1	305ML	Plife	MDAY	CULL	NCL	LDIM	LI	BV
TMY1	**0.230***	**0.720***	**0.238***	-0.020	**0.419***	-0.010	**-0.226***	0.037	**0.157***	**0.634***
TMYL		**0.198***	**0.928***	0.058	**0.394***	0.061	**-0.230***	**0.123***	**0.257***	**0.481***
305M1			**0.222***	**-0.077***	**0.590***	-0.058	**-0.152***	-0.040	0.015*	**0.390***
305ML				0.018	**0.481***	0.022	**-0.221***	**0.073***	**0.199***	**0.398***
Plife					-0.057	**0.990***	**0.890***	**0.977***	**0.774***	-0.053
MDAY						-0.053	-0.012	0.056	**0.216***	**0.453***
CULL							**0.880***	**0.966***	**0.732***	-0.048
NCL								**0.865***	**0.653***	**-0.271***
LDIM									**0.866***	0.019
LI										**0.154***

* Values in bold typing are different from 0 with a significance level alpha=0.05

Statistical analysis:

The breeding values of milk yield were estimated using VCE 6.0.2 software (Groeneveld *et al.,* 2008). The following statistical model was used:

$$Y_{ijklm} = \mu + YC_i + SC_j + bx_{ijk} + A_l + e_{ijklm},$$

Where,

Y_{ijklm} = Observation on the total milk yield trait,

μ = The overall mean,

YC_i = The fixed effect of the i^{th} year of calving (i=1, 2, 3, 4), years were divided into four intervals, 1= (1991-1994), 2= (1995-1998), 3= (1999-2002) and 4= (2003- 2006).

SC_j = The fixed effect of the j^{th} season of calving (j=1, 2, 3, 4), where ,1- winter (December- February) ,2- Spring (March- May), 3- Summer (June- August) and 4- Autumn (September – November)

b = The regression coefficient of total milk yield trait on age at first calving,

x_{ijk} = Age at first calving, as a co-variable,

A_l = Animal's random additive genetic effect, and

e_{ijklm} = random error.

A partial least squares (PLS) analysis (PLS regression, XLSTAT, 2009.) and stepwise multiple linear regression analysis (SMLR), and maximum R-square (MaxRsquare) methods (Proc stepwise, SAS/STAT, SAS Inc, 2004) were used to determine the relationship between breeding values and lifetime performance traits and the accuracy of predicting breeding values using different methods.

The PLS model:

The idea of PLS regression is to create, starting from a table with n observations described by p variables, a set of h components with h<p. The determination of the number of components to keep is usually based on a criterion that involves a cross-validation. (XLSTAT, 2009). The equation of the PLS regression model and the model quality indexes were described in detail by Ibrahim (2009).

RESULTS AND DISCUSSION

Table 3 shows the traits used in predicting breeding values by using stepwise method (SMLR). Predictions with SMLR produced Maximum R^2 of 0.61 with root mean square error (RMSE) of 231 when the model used the following five traits: TMY1, TMYL , 305ML, 305M1 and MDAY (step 5).

Table 3. The SMLR models for prediction of breeding values from lifetime performance traits and there R^2 values, and mean squares for error (MSE) and its root (RMES)

Step	Intercept	TMY1	TMYL	305 ML	305 M1	MDAY	R^2	MSE	RMSE
1	-787.2	0.088					0.402	81858	286
2	-948.2	0.076	0.036				0.520	65719	256
3	-852.6	0.078	0.077	-0.067			0.548	62033	249
4	-1079.9	0.069	0.086	-0.090		21.352	0.574	58569	242
5	-914.2	0.095	0.087	-0.099	-0.088	35.549	0.613	53160	231

Table 4 and Fig.1 allow to visualize the quality of the PLS regression as a function of the number of components .

Table 4. Model quality of breeding value (BV)

Index	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6
Q^2 cum	0.435	0.482	0.515	0.544	0.581	0.601
R^2Y cum	0.449	0.519	0.552	0.581	0.615	0.636
R^2X cum	0.288	0.472	0.788	0.900	0.952	0.980

The Q^2 cumulated index measures the global goodness of fit and the predictive quality of the BV model. PLS has selected six components (Comp1, Comp2,…, Comp6). The values of Q^2 cum and the R^2Y cum with the six components are 0.6. The R^2X cum that corresponds to the correlations between the lifetime performance traits (Xs) and BV (Y) variable with the components are very close to 1 with last two components (Comp5 and Comp6). This indicates that the six components generated by the PLS regression summarize well both the Xs and the Y. The cumulated Q^2 corresponding to this model reaches its maximum value with 6 component**.**

Table 5 presents the correlation matrix of the explanatory (Xs) and dependent (Y) variables with the t (t1, t2,…,t6) components. The correlations map (Fig. 2) allows to visualize for the first two PLS components(t1, t2) the correlations between the Xs and the components, and Y and the components.

High correlations are found on the first two dimensions (t1, t2). Also strong correlations were found among all possible combinations of BV, TMY1, TMYL, 305 M1, 305 ML and MDAY. By looking at the correlations map, we should also notice that the correlations are concentrated on the three parts of the correlations circle and they are far from the center (the correlations around the center is low).

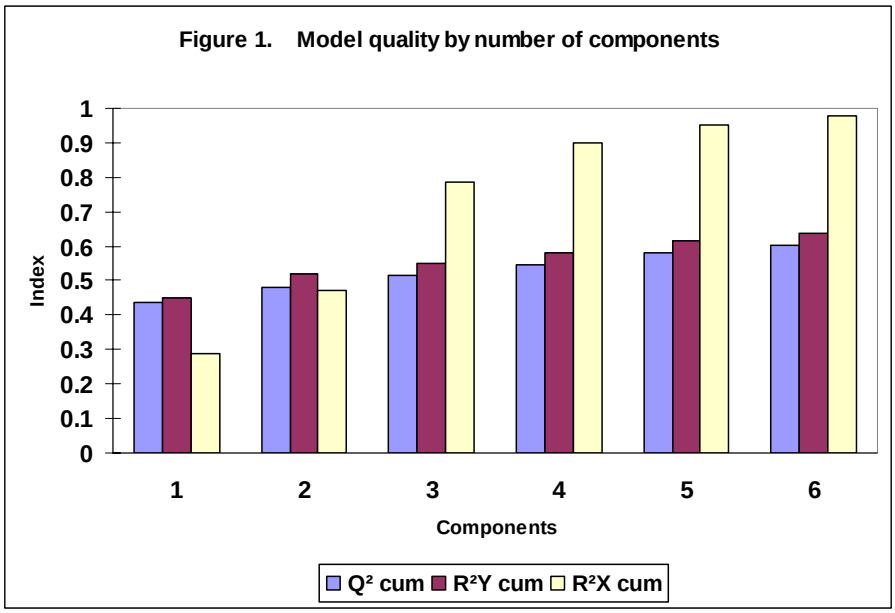
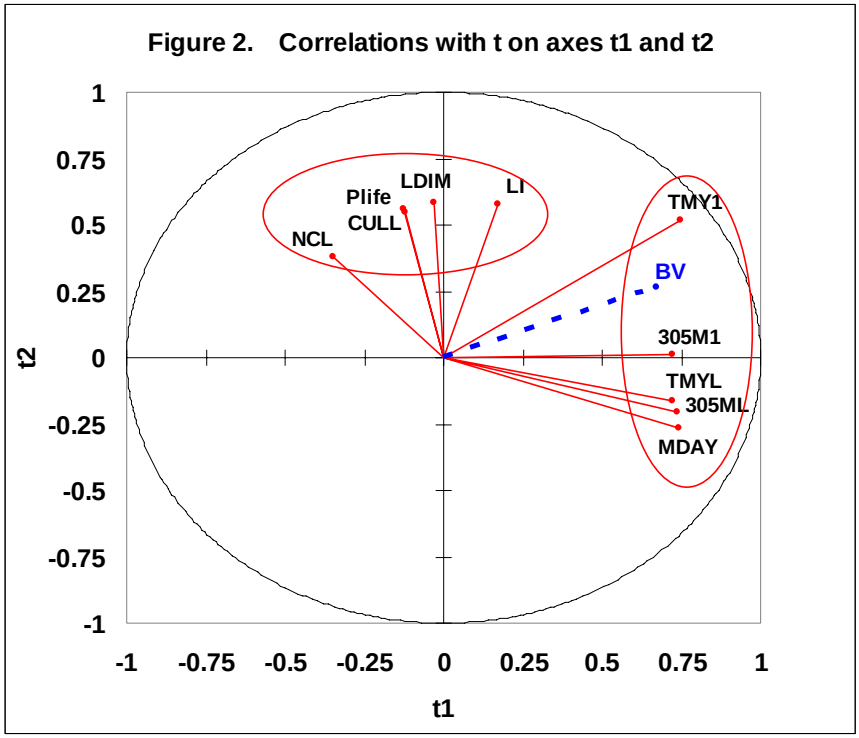


Table 5. Correlation matrix of the variables with the t components

Variable	t1	t2	t3	t4	t5	t6
TMY1	0.749	0.515	0.233	-0.306	0.070	0.022
TMYL	0.723	-0.160	-0.179	0.523	-0.299	0.200
305M1	0.722	0.014	-0.105	-0.666	0.074	0.100
305ML	0.744	-0.262	-0.219	0.445	-0.309	0.021
Plife	-0.127	0.563	-0.789	0.117	-0.067	0.128
MDAY	0.735	-0.204	-0.276	-0.024	0.536	-0.228
CULL	-0.119	0.550	-0.786	0.105	-0.046	0.194
NCL	-0.348	0.377	-0.825	-0.016	0.159	-0.052
LDIM	-0.031	0.586	-0.787	0.169	-0.018	0.025
LI	0.174	0.576	-0.627	0.253	-0.074	-0.341
BV	0.670	0.264	0.181	0.171	0.186	0.144

Table (6) shows the variable importance for the projection (VIPs) for each explanatory variable, at the last component (Comp6). The first five VIP values showed the most influential traites on the prediction of BV.



On the VIP chart (Fig.3), a border line is plotted to identify the VIPs that are greater than 0.8. This threshold, suggested by Wold (1995) and Ericksson (2001), allow for identifying the variables that are moderately (0.8<VIP<1) or highly influential (VIP>1).This allows to quickly identify which are the explanatory variables that are highly influential (namely TMY1, TMYL, 305M1, 305ML and MDAY) on the BV model. We can also see that the NCL, LI, CULL, Plife and LDIM have a low infludence on the BV model**.** The software in both proceedures (PLS & SMLR) selected the same traits (TMY1, TMYL, 305M1, 305ML and MDAY) as a principal components.

Table 7 displays the parameters (or coefficients) of the BV dependent variable model. Table 8 shows the goodness of fit statistics of the PLS regression model for BV (dependent) variable. The R^2 between the input variables (BV and explanatory) and the components t allow evaluating the explanatory power of the t. We define it as the mean of the squares of the correlation coefficients between the variables and the component.The analysis of the model corresponding to BV allows to conclude that the model is fit (R^2 equals 0.64)

Table 6. Variable Importance for the projection (VIP Comp6)

Variable	VIP*	Standard deviation	Lower bound(95%)	Upper bound(95%)
TMY1	**1.659**	0.272	1.127	2.191
MDAY	**1.245**	0.012	1.220	1.269
TMYL	**1.239**	0.099	1.044	1.433
305M1	**1.166**	0.079	1.011	1.320
305ML	**1.163**	0.027	1.110	1.217
NCL	0.768	0.180	0.416	1.120
LI	0.625	0.275	0.086	1.163
CULL	0.409	0.221	-0.024	0.842
Plife	0.406	0.229	-0.043	0.856
LDIM	0.374	0.424	-0.458	1.206

* Listed in a dessending order

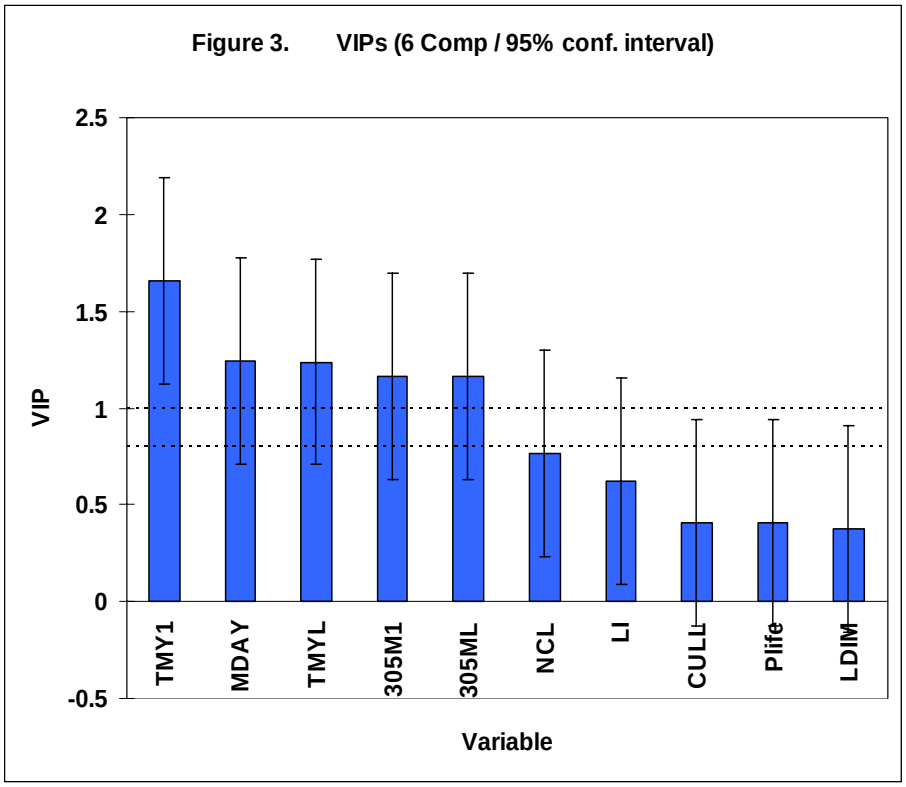


Table 7. Model parameters

Observations	850
Sum of weights	850
DF	843
R^2	0.64
Std. deviation	224
MSE	49692
RMSE	223

Table 8. Goodness of fit statistics (Variable BV)

Variable	BV
Intercept	-582.695
TMY1	0.092
TMYL	0.066
305M1	-0.115
305ML	-0.085
Plife	2.988
MDAY	53.122
CULL	3.895
NCL	-117.780
LDIM	0.104
LI	-10.639

CONCLUSIONS

The variables important for the projection of breeding value (BV) were those measured-in-kg , e.g. TMY1, MDAY**,** TMYL, 305M1 and 305ML . They have high influence on the BV model .The variables measured-in-time **(**day or month or lactation) e.g. NCL, LI , Cull , Plife and LDIM have low influence on the BV model . There is little difference between P L S results and those of RMSE.

ACKNOWLEDGEMENTS

The author would like to express his high appreciation to the International Company for Animal Wealth for providing the data. Also, the author is grateful to the Cattle Information Systems/Egypt (CISE) for the availability of the facilities for data analysis.

REFERENCES

Abdi, H., 2003. Partial least squares (PLS) regression. In Encyclopedia of social sciences research methods (ed. M. Lewis–Beck, A. Bryman and T. Futing), pp. 1-7. Sage Publication, Thousand Oaks, CA.

Eriksson L., Johansson E., Kettaneh-Wold N. and Wold S. 2001. Multi- and Megavariate Data Analysis. Principles and Applications, Umetrics Academy, Umeå.

Groeneveld, E.; M. Kovac and N. Mielenz, 2008. VCE 6.0.2, Co-variance components estimation package, Institute of Farm Animal Genetics, Mariensee, Germany.

Ibrahim, M.A.M., 2009. The use of Partial Least Squares regression procedure to determine the relative importance of the lifetime performance traits in predicting total lifetime milk yield of Holstein cows in Egypt. Egyptian J. Anim. Prod., (2009) 46(1):1-9

SAS, 2004. SAS Statistics. Guide Release Edition. SAS. Inst., Inc., Cary, NC.

Wold S. 1995. PLS for multivariate linear modeling. In: van de Waterbeemd H. (ed.), QSAR: Chemometric Methods in Molecular Design. Vol. 2. Wiley-VCH, Weinheim, Germany. 195-218.

XLSTAT 2009 , Statistical software for MS Excel- Statistics and data analysis with MS

Excel Addinsoft , 224 Centre Street, 3rd Floor New York, NY 10013 USA.

المقارنة بين طرق مختلفة لتحديد الأهمية النسبية لصفات الآداء طيلة العمر للتنبؤ بالقيم التربوية لأبقار الهولستين

محمد عبد العزيز محمد إبراهيم

قسم الإنتاج الحيواني، كلية الزراعة، جامعة القاهرة، الجيزة، ج .م.ع

استخدم في هذه الدراسة 2730 سجلا لإنتاج اللبن لعدد 850 بقرة هولستين بنات 316 طلوقة تابعة لمزرعة الشركة العالمية للثروة الحيوانية بالجيزة . كان الهدف من الدراسة هو تحديد إلي أي مدى تتأثر القيمة التربوية للحيوان لصفة إنتاج اللبن بكل من الصفات التالية:- إنتاج اللبن في الموسم الأول، إنتاج اللبن في الموسم الأخير، إنتاج 305 يوم للموسم الأول، إنتاج 305 يوم للموسم الأخير، إنتاج اللبن اليومي خلال الحياة الإنتاجية، عدد أيام الحلب خلال حياة الحيوان، عدد مواسم الحلب الكاملة، العمر الإنتاجي، العمر عند الاستبعاد ودليل الحياتية وذلك باستخدام طريقة الانحدار الجزئي للحد الأدنى للمربعات وكذلك الانحدار بطريقة الخطوة خطوة (الانحدار التدريجي).

وقد أظهرت النتائج أن الصفات الأكثر أهمية وذات التأثير المعنوي على القيمة التربوية للحيوان لصفة إنتاج اللبن أو التنبؤ به، هي على الترتيب صفات إنتاج اللبن في الموسم الأول، إنتاج اللبن اليومي خلال الحياة الإنتاجية، إنتاج اللبن في الموسم الأخير، إنتاج 305 يوم للموسم الأول و إنتاج 305 يوم للموسم الأخير على الترتيب والملاحظ إنها صفات تقاس بوحدة الوزن . في حين جاءت صفات عدد أيام الحلب خلال حياة الحيوان، العمر الإنتاجي، العمر عند الاستبعاد، دليل الحياتية و عدد مواسم الحلب (الملاحظ أنها صفات ترتبط بعنصر الزمن) في مرتبة الصفات الأقل أهمية . وقد ظهر هذا واضحا من النتائج المتحصل عليها بكلا الطريقتين اللتين استخدمتا في التحليل .

وقد إظهر التحليل بطريقة الإنحدار الجزئي للحد الأدنى للمربعات دقة أعلى نسبياً (0.64 مقابل 0.61) مع إنخفاض في الجذر التربيعي لتباين الخطأ (223 مقابل 231) عند المقارنة مع الإنحدار بطريقة الخطوة خطوة.