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

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 total lifetime milk yield (TLM) 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 Partial Least Squares regression (PLS) procedure.

The Q^2 cumulated index (0.971) measures the global goodness of fit and the predictive quality of the TLM model. The variables important for the projection of the total lifetime milk yield (TLM) were thoses measured-in-time (day or month or lactation) e.g. LDIM, Plife , Cull , LI and NCL ,but the variables measured-in-kg , e.g. MDAY, TMYL, 305M1,305ML and TMY1, had low influence on the TLM model . The R^2 between the input variables (TLM and lifetime variables) and the PLS components was 0.97. Therefore the model is well fitted.

Keywords: Partial least squares regression, total lifetime milk yield, lifetime performance traits, Holstein cows

INTRODUCTION

Partial least squares analysis is a multivariate statistical technique that allows comparison between multiple response variables and multiple explanatory variables. Partial least squares is one of a number of covariance-based statistical methods which are often referred to as structural equation modeling. It was designed to deal with multiple regression when data has small sample, missing values. Partial least squares regression has been demonstrated on both real data and in simulations (Garthwaite, 1994 and Tennenhaus *et al.*, 2005).

The PLS overcomes the multicollinearity problems by combining features of principal components analysis (PCA) and multiple regression (Abdi, 2003).

This method is quick, efficient and optimal for a criterion based on covariances. It is recommended in cases where the number of variables is high, and where it is likely that the explanatory variables are correlated. (XLSTAT, 2009)

The objective of this work was to determine the extent to which total lifetime milk yield is influenced by lifetime variables.

MATERIALS 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 in 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 productivity.

Definitions and notations of the studied traits

- Total lifetime milk yield (TLM, Kg) = including production to the end of each lactation.
- Lifetime performance traits which are used in predicting TLM :
  1. Total Milk Yield at First Lactation (TMY1 ,Kg)
  2. Total Milk Yield at Last Lactation (TMYL ,Kg)
  3. 305 Milk Yield at First Lactation (305M1 ,Kg)
  4. 305 Milk Yield at Last Lactation (305ML ,Kg)
  5. Milk per Day at 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.

Raw means of the traits considered are reported in Table 1. The Correlation matrix between traits is presented in table 2.

Table 1. Descriptive statistics for total lifetime milk yield and lifetime performance

Variable	Measuring unit	Mean	SD
TLM	Kg	26935	11815
TMY1	Kg	9160	2675
TMYL	Kg	7365	3650
305M1	Kg	7175	1380
305ML	Kg	6200	2455
Plife	Month	48	20
MDAY	Kg	19	3.5
CULL	Month	75	19
NCL	Lactation	3.2	1.3
LDIM	Day	1220	480
LI	%	52	8

Table 2. Correlation matrix between lifetime performance traits and lifetime milk yield

Variables	TMYL	305M1	305ML	Plife	MDAY	CULL	NCL	LDIM	LI	TLM
TMY1	**0.230**	**0.720**	**0.238**	-0.020	**0.419**	-0.010	**-0.226**	0.037	**0.157**	**0.141**
TMYL		**0.198**	**0.928**	0.058	**0.394**	0.061	**-0.230**	**0.123**	**0.257**	**0.221**
305M1			**0.222**	**-0.077**	**0.590**	-0.058	**-0.152**	-0.040	0.015	**0.154**
305ML				0.018	**0.481**	0.022	**-0.221**	**0.073**	**0.199**	**0.214**
Plife					-0.057	**0.990**	**0.890**	**0.977**	**0.774**	**0.902**
MDAY						-0.053	-0.012	0.056	**0.216**	**0.345**
CULL							**0.880**	**0.966**	**0.732**	**0.894**
NCL								**0.865**	**0.653**	**0.819**
LDIM									**0.866**	**0.935**
LI										**0.821**

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

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 method used to build the components differs from Principal Components Analysis (PCA) , and presents the advantage of handling missing data. 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 was:

$$\mathbf{Y} = T_h C'_h + E_h$$
$$= X W^*_h C'_h + E_h$$
$$= X W_h (p'_h W_h)^{-1} C'_h + E_h$$

where Y is the matrix of the dependent variables, X is the matrix of the explanatory variables. T_h , C'_h , W^*_h , W_h and p'_h , are the matrices generated by the PLS algorithm, and Eh is the matrix of the residuals.

The matrix B of the regression coefficients of Y on X, with h components generated by the PLS regression algorithm is given by:

$$B = W_h (p'_h W_h)^{-1} C'_h$$

The predictor data matrix X is compressed into a set of h latent variables or factors $\mathbf{t}_h = Xw_h$ (h = 1, 2,. . .,H) with relative weights **w** determined such as to maximize the covariance between factor scores $\mathbf{t}_a$ and the corresponding factors of the dependent variables $u_h = Yc_h$, subject to some normalization and orthogonality conditions.

Each latent variable accounts for a certain amount of the variability of the **X** and **y**. According to the criterion of the covariance maximization, the first h latent variables explain the most of the variance of the **X** and **y**, while the remaining latent variables typically describe the random noise in the data. The optimal number of latent variables, h, is obtained considering the Residual Sum of Squares (RSS) and the PRediction Error Sum of Squares (PRESS)

Model quality indexes:

The Q² cumulated **(Q²cum)** index measures the global contribution of the h first components to the predictive quality of the model .The **R²Y** cumulated (**R²Ycum**) index is the sum of the coefficients of determination between the dependent variables and the h first components. The **R²X** cumulated (**R²Xcum**) index is the sum of the coefficients of determination between the explanatory variables and the h first components.

RESULTS AND DISCUSSION

Table3 and Fig.1 allow to visualize the quality of the PLS regression as a function of the number of components .

Table 3. Model quality of total lifetime milk yield trait

Index	Comp1	Comp2	Comp3	Comp4	Comp5
Q² cum	0.936	0.955	0.967	0.969	0.971
R²Y cum	0.937	0.956	0.968	0.970	0.972
R²X cum	0.441	0.709	0.812	0.928	0.957

The Q² cumulated index measures the global goodness of fit and the predictive quality of the TLM model. PLS has selected five components (Comp1, Comp2,…, Comp5). The values of Q² cum with the five components is very close to 1. The R²Y cum and R²X cum that correspond to the correlations between the lifetime performance traits (Xs) and TLM (Y) variable with the components are very close to 1 with last two components (Comp4 and Comp5). This indicates that the five components generated by the PLS regression summarize well both the Xs and the Y. As we noticed that the cumulated Q² corresponding to this model reaches its maximum value with 5 components**.**

Figure 1. Model quality by number of components

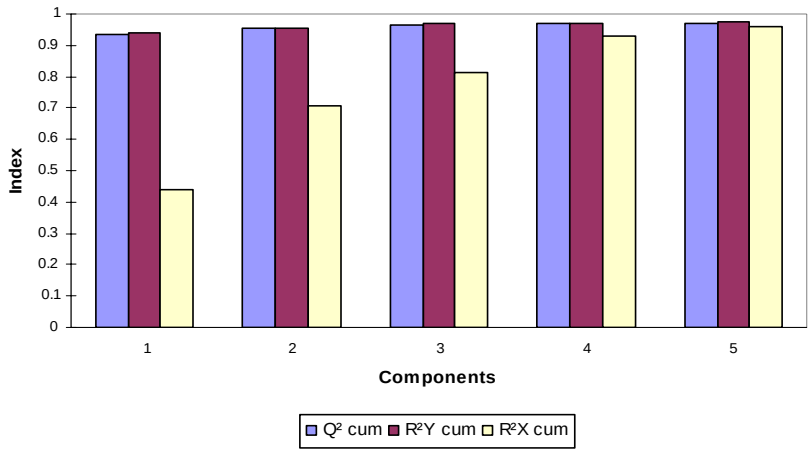


Table (4) presents the correlation matrix of the explanatory (Xs) and dependent (Y) variables with the t (t1, t2,…,t5) components. The correlations map (Fig.2) allows to visualize on the first two PLS components(t1, t2) the correlations between the Xs and the components, and the Y and the components.

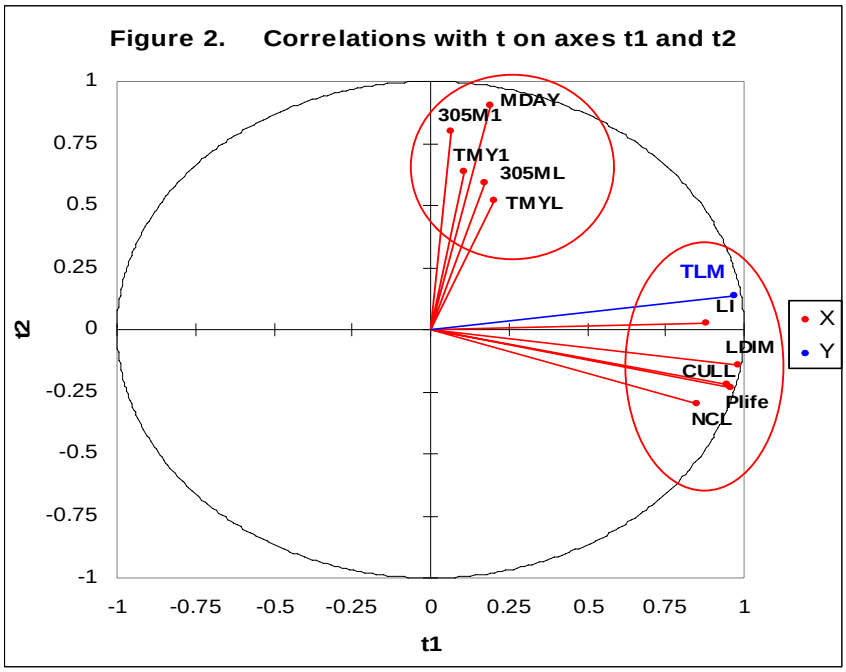
Table 4. Correlation matrix of the variables with the t components

Variable	t1	t2	t3	t4	T5
TMY1	0.106	**0.636**	-0.354	-0.570	0.300
TMYL	0.204	**0.518**	-0.615	0.522	-0.074
305M1	0.066	**0.796**	0.016	-0.470	0.057
305ML	0.175	**0.590**	-0.508	0.567	-0.046
Plife	**0.958**	-0.234	0.006	-0.016	0.124
MDAY	0.191	**0.899**	0.287	0.135	-0.097
CULL	**0.947**	-0.220	0.018	-0.003	0.190
NCL	**0.851**	-0.298	0.396	-0.016	0.000
LDIM	**0.985**	-0.142	-0.014	-0.024	0.021
LI	**0.880**	0.026	-0.169	-0.096	-0.357
TLM	**0.968**	0.138	0.109	0.044	0.047

High correlations are founded on the first two dimensions (t1,t2). Also strong correlations were found between **TLM and LDIM** and **Plife** and **CULL** and **LI** and **NCL**. By looking at the correlations map, we should also notice that the correlations are concentrated on the two positive parts of the correlations circle.
Table5 shows the Variable Importance for the Projection (VIPs) for each explanatory variable, at last component (Comp5), since their value of Q^2 cum is 0.971.

Table5. Variable importance in the projection (VIP Comp5)

Variable	VIP	Standard deviation	Lower bound(95%)	Upper bound(95%)
LDIM	**1.437**	0.030	1.377	1.496
Plife	**1.387**	0.050	1.289	1.486
CULL	**1.374**	0.048	1.280	1.469
LI	**1.271**	0.025	1.222	1.320
NCL	**1.265**	0.118	1.033	1.496
MDAY	0.650	0.090	0.474	0.826
TMYL	0.387	0.197	0.000	0.773
305ML	0.373	0.186	0.008	0.737
305M1	0.323	0.148	0.033	0.614
TMY1	0.304	0.244	-0.174	0.782



On the VIP charts (fig.3) one bar chart per component (variable), a border line is plotted to identify the VIPs that are greater than 0.8. These thresholds, suggested by Wold (1995) and Ericksson (2001), allow identifying the variables that are moderate (0.8<VIP<1) or highly influential (VIP>1).This allows to quickly identify which are the explanatory variables that are highly influential (LDIM, Plife , Cull , LI and NCL) on the TLM model. We can also see that the TMY1, 305M1,305ML , TMYL and MDAY have a low infludence on the TLM model**.**

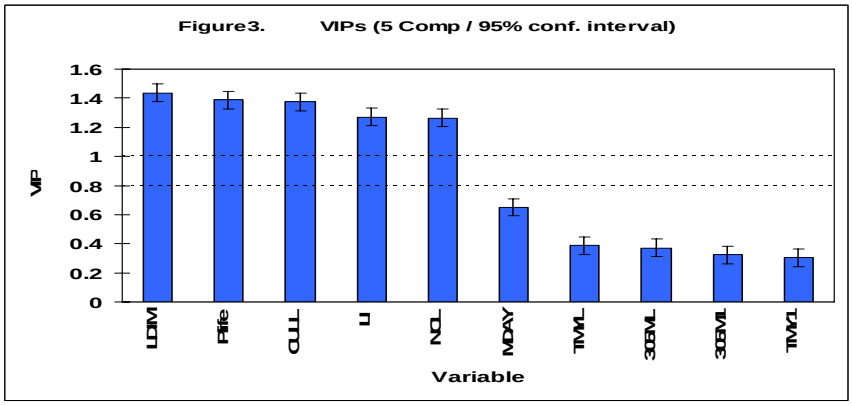


Table 6 displays the parameters (or coefficients) of the TLM dependent variable model. Table 7 shows the goodness of fit statistics of the PLS regression model for TLM dependent variable. The R^2 between the input variables (TLM and explanatory) and the components t allow evaluating the explanatory power 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 TLM allows to conclude that the model is well fitted (R^2 equals 0.97)

Table 6. Model parameters

Variable	TLM
Intercept	-29108.831
TMY1	0.098
TMYL	0.103
305M1	0.215
305ML	0.244
Plife	151.272
MDAY	995.844
CULL	147.385
NCL	1789.898
LDIM	5.750
LI	38.113

Table7. Goodness of fit statistics (Variable TLM)

Observations	850.000
Sum of weights	850.000
DF	844.000
R^2	0.972
Std. deviation	1971.942
MSE	3861106.224
RMSE	1964.970

CONCLUSIONS

Results of the present study highlight that the PLS method, developed to maintain a good predictive power of multivariate regression and to correct at the same time for high co linearity among predictors. The variables importance for the projection total lifetime milk yield (TLM) were those measured-in-time (day or month or lactation) e.g. LDIM, Plife , Cull , LI and NCL ,but the variables measured-in-kg , e.g. MDAY, TMYL, 305M1,305ML and TMY1, have a low influence on the TLM model . The R^2 between the input variables (TLM and lifetime variables) and the PLS components was 0.97. Therefore the results gave raise to the model is well fitted.

ACKNOWLEDGEMENTS

This work was funded by the cattle information systems/Egypt (CISE) and International Company for Animal Wealth.

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

Garthwaite, Paul H., 1994. An Interpretation of Partial Least Squares. *Journal of the American Statistical Association*, 89(425): 122.

Tenenhaus M., J. Pagès, L. Ambroisine and C. Guinot, 2005. PLS methodology for studying relationships between hedonic judgements and product characteristics. Food Quality and Preference. 16, 4, 315-325.

Wold S., 1995. PLS for multivariate linear modelling. 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

استخدام طريقة الانحدار الجزئي للحد الأدنى للمربعات لتحديد الأهمية النسبية لصفات الأداء طيلة العمر للتنبؤ بإنتاج اللبن الكلى خلال حياة الحيوان لأبقار الهولستين في مصر

محمد عبد العزيز محمد إبراهيم

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

استخدم في هذه الدراسة 2730 سجل لبن لـ 850 بقرة هولستين بنات 316 طلوقة تابعة لمزرعة الشركة العالمية للثروة الحيوانية بالجيزة . كان الهدف من الدراسة هو تحديد إلي أي مدى يتأثر إنتاج اللبن الكلى خلال حياة الحيوان 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 – دليل الحياتية وذلك باستخدام طريقة الانحدار الجزئي للحد الأدنى للمربعات .

وقد أظهرت النتائج أن الصفات الأكثر أهمية وذات التأثير المعنوي على إنتاج اللبن الكلى خلال حياة الحيوان أو التنبؤ به، هي على الترتيب عدد أيام الحلب خلال حياة الحيوان، العمر الإنتاجي، العمر عند الاستبعاد، دليل الحياتية و عدد مواسم الحلب والملاحظ أنها صفات ترتبط بعنصر الزمن. بينما كان تأثير الصفات الإنتاجية (إنتاج اللبن في الموسم الأول، إنتاج اللبن في الموسم الأخير، إنتاج 305 يوم للموسم الأول، إنتاج 305 يوم للموسم الأخير وإنتاج اللبن اليومي خلال الحياة الإنتاجية) غير معنوي. وقد وصلت درجة الدقة إلى 97% .