Animal Science 2006, **82**: 463–468

ISSN 1357-7298
DOI: 10.1079/ASC200659

Use of a partial least-squares regression model to predict test day of milk, fat and protein yields in dairy goats

N. P. P. Macciotta[1†], C. Dimauro[1], N. Bacciu[1], P. Fresi[2] and A. Cappio-Borlino[1]

[1]*Dipartimento di Scienze Zootecniche, Università di Sassari, Via De Nicola 9, 07100 Sassari, Italy*
[2]*Associazione Nazionale della Pastorizia, Via Togliatti 1587, 00155 Rome, Italy*

[†] E-mail: macciott@uniss.it

Abstract

A model able to predict missing test day data for milk, fat and protein yields on the basis of few recorded tests was proposed, based on the partial least squares (PLS) regression technique, a multivariate method that is able to solve problems related to high collinearity among predictors. A data set of 1731 lactations of Sarda breed dairy Goats was split into two data sets, one for model estimation and the other for the evaluation of PLS prediction capability. Eight scenarios of simplified recording schemes for fat and protein yields were simulated. Correlations among predicted and observed test day yields were quite high (from 0·50 to 0·88 and from 0·53 to 0·96 for fat and protein yields, respectively, in the different scenarios). Results highlight great flexibility and accuracy of this multivariate technique.

Keywords: *goats, milk fat, milk protein, prediction, regression analysis.*

Introduction

The future development of supervised recording plans for milk production traits of small ruminants farmed in semi extensive conditions has to cope with two opposite requirements. First, an increase of the number of recorded animals is needed to enhance the impact of breeding programs for dairy sheep and goats. Such an increase necessarily implies a reduction of the number of recorded tests per lactation in order to contain selection costs (Bouloc *et al.*, 1991; Giaccone *et al.*, 1996; Gonzalo *et al.*, 2003). On the other hand, the availability of a minimum number of tests along lactation still remains a fundamental requisite both to guarantee the efficiency of selection schemes and to direct management decisions. Moreover, being almost all of sheep and goat milk destined to cheese processing, it is of great importance the knowledge of the evolution along the lactation of fat and protein yields.

These two opposite requirements can be reconciled by a suitable mathematical model able to predict with a sufficient accuracy missing tests on the basis of a few tests actually recorded. Several methods able to predict test day (TD) yields with reasonable accuracies have been proposed for dairy cattle, mainly based on multiple-trait and test day models (Schaeffer and Jamrozik, 1996; Pool and Meuwissen 1999; Mayeres *et al.*, 2004; Vasconcelos *et al.*, 2004). Time series analysis and neural network approaches have been specifically tested for dairy sheep (Kominakis *et al.*, 2002; Macciotta *et al.*, 2002). However, the economic relevance of official milk recording in dairy species justifies further research on predictive methods.

In this paper, the capability of the partial least-squares regression (PLS) to predict missing tests in simplified recording schemes for small ruminants is checked. The PLS, originally developed in the computational chemistry context (Hoeskuldsson, 1988), has become an established tool for modelling linear relations between multivariate measurements. It is particularly useful when a set dependent variables has to be predicted from a set of independent variables highly correlated. The PLS overcomes the multicollinearity problems by combining features of principal components analysis (PCA) and multiple regression (Abdi, 2003).

Material and methods

Data

Data were test day records of milk production traits of 1731 Sarda goats, recorded by the Italian Association of Animal Breeders in the period 1989 to 2001. Each animal had five records for milk (MILK1 to MILK5), fat (FAT1 to FAT5) and protein (PROT1 to PROT5) yields arranged in a multivariate setting. Raw means of the traits considered for each test are reported in Table 1. The data set was split into two sub sets: an estimation data set (EDS) that consisted of 1000 goats, which was used for model estimation; a validation data set (VDS), made by the remaining 731 goats, which

Table 1 *Means and standard deviation for milk, fat and protein test day yields (g/day)*

	MILK 1	MILK 2	MILK 3	MILK 4	MILK 5
Mean	1640	1540	1350	1270	1050
s.d.	830	810	670	590	570
	FAT 1	FAT 2	FAT 3	FAT 4	FAT 5
Mean	80	80	70	60	50
s.d.	40	40	30	30	30
	PROT 1	PROT 2	PROT 3	PROT 4	PROT 5
Mean	70	60	60	50	40
s.d.	30	30	20	20	20

was used for making predictions by using the model estimated with the EDS records.

The PLS model

The most simple and intuitive method to predict values of *m* dependent variables $\mathbf{Y} = (\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_m})$ on the basis of values of *p* independent variables (predictors) $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p})$ is the multivariate multiple regression of **Y** on **X**. However, in cases of multicollinearity, i.e. when he independent variables are highly correlated among them (Draper and Smith, 1981), the resulting inflation of parameter variance compromises the predictive capability of the model. This is just the case of this study, where both **Y** and **X** are TD yields of milk, fat and protein recorded at different time distance among lactation: it is well known that both correlations between yields of milk, fat and protein on a given test day and correlations of yields on consecutive days are generally high (Schaeffer and Jamrozik, 1996).

PLS is a quite recent statistical tool able to handle multivariate regression models characterized by high collinearity among predictors (Geladi and Kowlaski, 1986). It develops a biased regression in order to stabilize the parameter estimates that lead to a more reliable prediction. At this aim the predictor data matrix **X** is compressed into a set of *A* latent variables or factors $\mathbf{t_a} = \mathbf{Xr_a}$ (a = 1, 2,...,A) with relative weights $\mathbf{r_a}$ determined such as to maximize the covariance between factor scores $\mathbf{t_a}$ and the corresponding factors of the dependent variables $\mathbf{u_a} = \mathbf{Yq_a}$, subject to some normalization and orthogonality conditions. Specifically, the following conditions control the PLS solutions: (1) maximization of covariance of score vectors $\mathbf{t_a}$ and $\mathbf{u_a}$; (2) normalization of weights $\mathbf{r_a}$, i.e. $\mathbf{r_a},\mathbf{r_a} = 1$; (3) normalization of weights $\mathbf{q_a}$, i.e. $\mathbf{q_a},\mathbf{q_a} = 1$; (4) orthogonality of **t** scores, i.e. $\mathbf{t_i},\mathbf{t_j} = 0$ for $i \neq j$.

The matrix of latent factors extracted from **X**, $\mathbf{T} = (\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_A})$ is then used for predicting values of latent factors extracted from Y, $\mathbf{U} = (\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_A})$

$$\mathbf{U} = \mathbf{BT}$$

Finally, the values of **Y** are calculated by back transforming the latent factors of **Y**, i.e.

$$\mathbf{Y} = \mathbf{BTQ}'$$

where **Q** is the matrix of the loadings of factors extracted from **Y**.

In brief, PLS extracts from the independent (**X**) and dependent (**Y**) variables a number of orthogonal latent variables that account for most of the variation of original variables. Therefore, PLS is similar to the regression on principal components (PC), that is the more commonly used method to overcome multicollinearity problems. However, unlike regression on PC, where the problem of choosing an optimum subset of predictor remains, PLS finds latent variables able to maximize the goodness of fit of the regression of factor scores of **Y** on factor scores of **X** (De Jong, 1993).

In the present study, the predictive capability of the PLS method has been tested in several scenarios that mimic simplified recording plans that differ from one another because of the number of available and missing tests (Table 2). Moreover, a comparison between predictions obtained by PLS and by regression on PC has been performed.

Tests selected as available are included in the **X** matrix of predictor variables, whereas missing tests are included in the **Y** matrix of variables to be predicted.

The adequacy of the PLS and of the regression of PC models for all the plans considered has been assessed by examining Pearson's correlations between predicted (Y_p) and observed (Y_o) values of dependent variables. Moreover, accuracy and precision of PLS predictions have been separately evaluated. Accuracy measures how closely predicted values are to the observed values whereas precision measures how closely individual model predicted values are within each others (Tedeschi, 2006). Consequently, inaccuracy (or bias) refers to the systematic

Table 2 *Scenarios of missing test day considered in the present study (asterisks indicate tests available, blanks tests to be predicted)*

Test day	Plan 1			Plan 2			Plan 3			Plan 4		
	Milk	Fat	Protein	Milk	Fat	Protein	Milk	Fat	Protein	Milk	Fat	Protein
1	*	*	*	*	*	*	*	*	*	*		
2	*	*	*	*			*			*	*	*
3	*			*			*	*	*	*		
4	*			*			*			*	*	*
5	*			*	*	*	*	*	*	*		
	Plan 5			Plan 6			Plan 7			Plan 8		
1	*	*	*	*	*	*	*	*	*			
2	*	*	*							*	*	*
3							*	*	*			
4										*	*	*
5				*	*	*	*	*	*			

deviation from the truth, whereas imprecision (or uncertainty) indicates a magnitude of the scatter about the predicted means. Accuracy can be distinguished from precision by partitioning the different sources of variation of the mean square errors of predictions (MSEP) (Theil, 1961). Two different partitions are possible (see **Appendix**). The first one consists of three terms that may be interpreted as error in central tendency (UM), error due to unequal variation (US) and error due to incomplete covariation (UC), whereas the three terms of the second partition represent error in central tendency (UM), error due to the linear regression of predicted on observed values (UR), and random errors, i.e. unexplained variance that cannot be accounted for by the linear regression (UD). The different terms are usually express as a percentage of the total MSEP.

Results and discussion

Results reported in Tables 3 and 4 are examples that illustrates the rationale of the PLS method and the essential steps of model estimation. They refer to the plans 3 and 7 of Table 2, that differ from one another because of the number of available milk tests (all for plan 3 and only MILK2 and MILK5 for plan 7). Latent factors are extracted in succession both from dependent and independent variables. The optimum number of factors is determined with a cross validation method by comparing the root of the predictive residual sum of squares (PRESS), measured in standard deviation units and averaged for all variables to be

Table 3 *Calculation steps of the partial least squares (PLS) method with cross-validation for plan 3 of Table 2*

	Cross-validation	
No. of PLS factors	PRESS	Comparison significance
0	1·612	$<0·0001$
1	0·906	$<0·0001$
2	0·831	$<0·0001$
3	0·702	$<0·0001$
4	0·617	0·0700
5	0·618	$<0·0001$
6	0·610	$<0·0001$
7	0·605	$<0·0001$
8	0·595	0·3800
9	0·594	1·0000
Minimum root mean PRESS		0·594
Minimizing number of factor		9
Smallest number of factors with $P > 0·1$		8

	Percent variation accounted for			
	Independent variables		Dependent variables	
Retained factors	Current	Total	Current	Total
1	63·074	63·074	59·165	59·165
2	13·160	76·234	7·937	66·203
3	5·032	81·267	8·428	74·631
4	3·710	84·976	6·511	81·141
5	2·330	87·305	2·840	83·981
6	1·420	88·725	1·562	85·543
7	8·486	97·211	0·164	85·707
8	1·202	98·413	0·304	86·012

Table 4 *Calculation steps of the partial least squares (PLS) method with cross-validation for plan 7 of Table 2*

	Cross-validation	
No. of PLS factors	PRESS	Comparison significance
0	1·544	$<0·0001$
1	0·947	$<0·0001$
2	0·911	0·662
3	0·896	0·640
4	0·897	0·025
5	0·892	1·000
Minimum root mean PRESS		0·892
Minimizing number of factor		5
Smallest number of factors with $P > 0·1$		2

	Percent variation accounted for			
	Factors		Responses	
Retained factors	Current	Total	Current	Total
1	63·906	63·906	52·690	52·690
2	17·034	80·940	2·125	54·815

Table 5 *Pearson correlations between observed and predicted values for fat, protein and milk yield in all the considered plans obtained with the partial least squares (PLS) and regression on principal components (PC) (below in italics) methods*

	Plan							
Test	1	2	3	4	5	6	7	8
				FAT				
1				0·86				0·61
				0·73				*0·62*
2		0·86	0·86			0·64	0·75	
		0·81	*0·79*			*0·53*	*0·73*	
3	0·84	0·84		0·84	0·75	0·69		0·79
	0·76	*0·79*		*0·77*	*0·76*	*0·66*		*0·77*
4	0·88	0·89	0·88		0·69	0·71	0·78	
	0·68	*0·86*	*0·84*		*0·69*	*0·70*	*0·79*	
5	0·84			0·86	0·50			0·61
	0·63			*0·84*	*0·48*			*0·76*
				PROTEIN				
1				0·98				0·65
				0·82				*0·65*
2		0·96	0·98			0·67	0·80	
		0·90	*0·85*			*0·60*	*0·79*	
3	0·98	0·97		0·96	0·80	0·69		0·84
	0·85	*0·86*		*0·73*	*0·80*	*0·67*		*0·85*
4	0·96	0·96	0·96		0·65	0·73	0·78	
	0·56	*0·92*	*0·89*		*0·67*	*0·66*	*0·78*	
5	0·95			0·98	0·53			0·69
	0·69			*0·92*	*0·52*			*0·76*
				MILK				
1								0·70
								0·66
2						0·66	0·79	
						0·57	*0·78*	
3					0·80	0·67		0·87
					0·82	*0·65*		*0·83*
4					0·71	0·72	0·80	
					0·74	*0·70*	*0·79*	
5					0·58			0·72
					0·57			*0·73*

Table 6 *Sources of variation of the mean square errors of predictions (MSEP) of the partial least squares (PLS) and regression on principal component (PC) (below in italics) in different plans for FAT3, PROTEIN3 and MILK3*

	Plan					
	1	2	4	5	6	8
Source	FAT 3					
Mean bias (U_M) (%)	0·39	2·40	3·81	2·15	0·02	1·64
	26·99	*28·65*	*27·42*	*25·15*	*22·33*	*26·98*
Unequal variances (U_S) (%)	0·22	2·68	0·08	34·88	18·12	18·67
	35·19	*33·11*	*35·02*	*44·99*	*39·81*	*36·32*
Incomplete (co) variation (U_C) (%)	97·38	94·92	96·10	62·97	81·86	79·70
	37·82	*38·24*	*37·56*	*29·86*	*37·86*	*37·70*
Systematic or slope bias (U_R) (%)	5·39	1·50	5·51	5·06	0·04	0·90
	9·04	*9·18*	*9·37*	*14·96*	*6·23*	*9·89*
Random errors (U_D) (%)	92·22	96·10	90·68	92·79	99·84	97·47
	63·97	*62·17*	*63·21*	*59·88*	*71·44*	*63·13*
	PROTEIN 3					
Mean bias (U_M) (%)	0·98	2·02	0·00	11·01	3·40	11·97
	41·98	*43·72*	*34·98*	*34·93*	*30·76*	*38·44*
Unequal variances (U_S) (%)	1·15	6·80	1·54	42·47	26·75	32·41
	33·18	*30·87*	*27·70*	*45·22*	*39·39*	*42·20*
Incomplete (co) variation (U_C) (%)	97·87	91·18	98·45	46·51	69·85	55·62
	24·84	*25·36*	*37·32*	*19·85*	*29·85*	*19·37*
Systematic or slope bias (U_R) (%)	0·01	1·90	0·11	12·91	0·97	9·72
	15·05	*14·16*	*4·68*	*20·66*	*7·91*	*21·91*
Random errors (U_D) (%)	99·01	96·08	99·89	76·08	95·63	78·31
	42·97	*42·07*	*60·33*	*44·41*	*61·33*	*39·65*
	MILK 3					
Mean bias (U_M) (%)				7·01	0·95	14·00
				34·95	*28·58*	*37·41*
Unequal variances (U_S) (%)				36·23	18·61	35·26
				42·55	*34·97*	*37·04*
Incomplete (co) variation (U_C) (%)				56·76	80·44	50·73
				22·51	*36·44*	*25·54*
Systematic or slope bias (U_R) (%)				9·09	0·00	13·62
				19·44	*4·49*	*13·14*
Random errors (U_D) (%)				83·90	99·04	72·37
				45·61	*66·93*	*46·45*

Note that $U_M + U_S + U_C = U_M + U_R + U_D = 100\%$.

predicted. The factor extraction process stops when a minimum value of PRESS is reached. However, the model with the minimum value of the PRESS statistic may not be significantly better than a previous one with fewer factors. Consequently, for each model, the significance level of a test of whether it is different from the one with the lowest PRESS is given. Finally, is retained the model that minimizes both the PRESS and the number of factors. Thus, for plan 3 (Table 3), although the minimum value of the PRESS is reached when nine factors are extracted, this value is not statistically different from the one with eight factors, and the latter is therefore retained. The variance explained increases together with the number of extracted factors, both for dependent and independent variables: with eight factors, the PLS model is able to explain about the 98% and 86% of the variance of independent and dependent variables, respectively.

Table 4 reports PLS model statistics for the plan 7: five factors are extracted but only two are retained with a 81% and 55% of variance explained for independent and dependent variables, respectively. In this case, fewer predictors available results in a reduction of both the number of factors retained and the magnitude of explained variances.

Table 5 reports Pearson correlations between observed and predicted values in all the considered plans obtained by the PLS and PC (in italics) models. The criteria used to choose the number of PCs to retain for the prediction step was a sum of their eigen values higher than 80% of the variance of original variables. Results highlight a good predictive ability of the PLS method, in most of cases higher than that of the regression on PC. The difference between the two methods is particularly evident for plans where the number of predictors is higher (i.e. plans 1–5). Actually, this is an expected result because PLS is a prediction method that is particularly suitable for cases in which predictors are many and highly collinear (Naes and Martens, 1985). For all the three traits, best results are obtained when available tests are regularly spread across lactation stages. Correlations are of the same order of those obtained in dairy cattle with other prediction methods (Schaeffer and Jamrozik, 1996; Macciotta *et al*., 2002; Mayeres *et al*., 2004; Vasconcelos *et al*., 2004).

From a technical point of view, particularly relevant are results concerning the two milk components. As expected, values are quite high for the first four plans (about 0·86 and 0·97 for fat and protein yields, respectively), where milk test day records were available for all the tests. Moreover, it is worth noticing that in all plans correlations tend to remain the same even when the number of available tests of fat and protein yields decreases (from 3 to 2). Finally, better results have been obtained for protein yield then for fat yield in all plans, thus confirming the higher predictability of this trait (Macciotta *et al.*, 2002; Vasconcelos *et al.*, 2004).

The analysis of the prediction error, obtained by partitioning the MSEP in different sources of variation, gives further indications on the predictive capability of the two models considered. As an example, Table 6 reports results of MSEP decomposition for MILK3, FAT3 and PROTEIN3 in all plans where these variables are predicted.

By and large, all methods aimed at solving the problem of multicollinearity among predictors result in biased regression. However, in the PLS method component of MSEP related to the inaccuracy (the mean bias, the inequality of variances and the systematic or slope bias) are of minimum importance (0 to 5%) whereas a marked predominance of the random component (in most cases larger than 90%), measured by the random error (UD) and by the incomplete covariance (UC), can be observed for all the three traits. Similar results have been obtained also in the other plans (not reported for brevity).

On the other hand, results of the MSEP analysis for the regression on PC highlight a relevant quota of the components related to prediction inaccuracy, such as means bias or slope bias (Table 6).

The incidence of random variation in the MSEP of PLS predictions could be due, apart from measurement error, to sources of variation of TD records that have not been properly accounted for in this work. Actually, in order to build a data set as larger as possible, goats of different parities, flocks, months of lambing, number of kids have been considered. A stratification of data according to these factors would probably result in more precise estimates.

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 collinearity among predictors, can be usefully applied to predict missing tests for milk production traits in individual lactations on the basis of a few tests recorded. The application of PLS models in some scenarios extracted from an archive of 1731 Sarda goats and characterised by different number and distributions of missing tests revealed the great flexibility and accuracy of this multivariate technique. The PLS technique seems therefore suitable for dairy small ruminants, where lactations with few and irregularly located TD records, especially for fat and protein contents, are often found. Such a situation is likely to become more widespread in the future for the diffusion of simplified milk recording schemes.

Acknowledgements

Work funded by the Italian Ministry of Education, University and Research (grant FAR and PRIN).

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.

Bouloc, N., Barillet, F., Boichard, D., Sigwald, J. P. and Bridoux, J. 1991. Etudes des possibilities d'allegement des controle laitier official chez les caprins. *Annales de Zootechnie* **40:** 125-139.

De Jong, S. 1993. SIMPLS: an alternative approach to Partial Least Squares regression. *Chemometrics and Intelligent Laboratory Systems* **18:** 251-263.

Draper, N. R. and Smith, H. 1981. *Applied regression analysis*. John Wiley and Sons, New York.

Geladi, P. and Kowlaski, B. 1986. Partial least squares regression: a tutorial. *Analytica Chimica Acta* **35:** 1-17.

Giaccone, P., Portolano, B., Todaro, M. and Leto, G. 1996. Semplificazione dei metodi di controllo della produzione del latte nella specie caprina. *Zootecnica e Nutrizione Animale* **22:** 139-148.

Gonzalo, C., Othmane, M. H., Angel Fuertes, J., De La Fuente, L. F. and San Primitivo, F. 2003. Losses of precision associated with simplified designs of milk recording in dairy ewes. *Journal of Dairy Research* **70:** 441-444.

Hoeskuldsson, A. 1988. Partial least squares PLS methods. *Journal of Chemometrics* **88:** 211-228.

Kominakis, A. P., Abas, Z., Maltaris, I. and Rogdakis, E. 2002. A preliminary study of the application of artificial neural networks to prediction of milk yield in dairy sheep. *Computers and Eletronics in Agriculture* **35:** 35-48.

Macciotta, N. P. P., Vicario, D., Pulina, G. and Cappio−Borlino, A. 2002. Test day and lactation yield predictions in Italian Simmental cows by ARMA methods. *Journal of Dairy Science* **85:** 3107-3114.

Mayeres, P., Stoll, J., Bormann, J., Reents, R. and Gengler, N. 2004. Prediction of daily milk, fat and protein production by a random regression test day model. *Journal of Dairy Science* **87:** 1925-1933.

Naes, T. and Martens, H. 1985. Comparison of prediction methods for multicollinear data. *Communications in Statistics, Simulation and Computation* **14:** 545-576.

Pool, M. H. and Meuwissen, T. H. E. 1999. Prediction of daily milk yields from a limited number of test days using test day model. *Journal of Dairy Science* **82:** 1555-1564.

Schaeffer, L. R. and Jamrozik, J. 1996. Multiple−trait prediction of lactation yields for dairy cows. *Journal of Dairy Science* **79:** 2044-2055.

Tedeschi, L. O. 2006. Assessment of adequacy of mathematical models. *Agricultural Systems* **89:** 225−247.

Theil, H. 1961. Economic forecasts and policy. In *Contribution of economic analysis* (ed. R. Strotz, J. Timbergen, P. J. Verdoorn and H. J. Witteveen), pp. 6-48. North−Holland Publishing Company, Amsterdam.

Vasconcelos, J., Martins, A., Petim−Batista, M. F., Colaco, J., Blake, R. W. and Carvalheira, J. 2004. Prediction of daily and lactation yields of milk, fat, and protein using an autoregressive repeatability test day model. *Journal of Dairy Science* **87:** 2591-2598.

(*Received* 20 *April* 2005—*Accepted* 26 *March* 2006)

Appendix

The mean square error of prediction (MSEP) is the most common and reliable measure of adequacy of mathematical models used for the prediction of a variable *Y*. It is defined as:

$$MSEP = \frac{\sum_{i=1}^{n} (Y_{pi} - Y_{oi})^2}{n}$$

where Y_{pi} and Y_{oi} are the *i*th predicted and observed value of *Y*, Theil (1961) has introduced methods of analysis of the different sources of variation of MSEP in order to distinguish accuracy and precision of a model of a given MSEP. Two readily interpretable partitions are possible.

$$\text{MSEP} = (\overline{Y}_P - \overline{Y}_o)^2 + (s_P - s_o)^2 + 2(1 - r)s_o s_P \quad (1)$$

$$\text{MSEP} = (\overline{Y}_P - \overline{Y}_o)^2 + s_o^2(1 - b)^2 + (1 - r^2)s_P^2 \quad (2)$$

where S_o^2 and S_P^2 (S_o and S_P) indicate the variances (standard deviations) associated with observed and model predicted values of *Y* respectively; *r* is the coefficient of correlations; *b* is the slope of the linear regression of Y_p on Y_o.

The terms of equations [1] and [2] are usually divided by the total MSEP to obtain the five proportions of MSEP (Tedeschi, 2006):

Proportion	Formula	Description
U_M	$\frac{(\overline{Y}_p - \overline{Y}_o)^2}{MSEP}$	Mean bias
U_S	$\frac{(s_p - s_o)^2}{MSEP}$	Unequal variances
U_C	$\frac{2(1-r)s_o s_P}{MSEP}$	Incomplete (co)variation
U_R	$\frac{s_o^2(1-b)^2}{MSEP}$	Systematic or slope bias
U_D	$\frac{(1-r^2)s_p^2}{MSEP}$	Random error

Note that $U_M + U_S + U_C = U_M + U_R + U_D = 1$.