Research on Several Problems in Partial Least Squares Regression Analysis

Ju Wu*

College of Mathematics and Information Science, Neijiang Normal University, Neijiang 641000, China

Abstract: Purpose: preliminary discussion on model prediction precision in the partial least squares regression analysis method; Method: introduce current development conditions of partial least squares regression analysis, analyze problems of traditional regression analysis method such as multiple linear regression analysis, introduce the mathematic principle and modeling method of the partial least squares regression analysis method, and conduct detailed analysis on the partial least squares regression analysis modeling and prediction by using the classical Linnerud data. Result: The partial least squares regression analysis has the basic features of the multiple linear regression analysis and principal component analysis, can precisely predict multiple data and establish a precise mathematical model; Conclusion: The partial least squares regression analysis can provide precise mathematical model and can reserve the explaining variants remarkably associated to explained variants to most extent, so it is feasible to some extent and can meet the general requirements of engineering, economy, biology and medical statistical analysis.

Keywords: Partial least squares regression analysis, model prediction, multiple linear regression analysis, principal component analysis, medical statistics.

1. INTRODUCTION

The partial least squares regression (PLSR) analysis is a new and very advanced multiple statistics method, which was first proposed by Herman Wold as an economist in Sweden in 1964. Later PLSR was studied by Wold and Abano, so the theory, methods and application quickly develop. Like the classic multiple linear regression analysis (MLR) and logistics multiple regression analysis, the PLSR mainly solves the actual problems in the MLR analysis such as variant multiple correlations or variants more than sample points. The PLSR analysis integrates basic functions of multiple linear regression analysis, principal component analysis and classical association analysis, so this method is extensively applied in the biomedicine, economy, chemistry measurement and industry design. This method is called as the 2G multiple statistics analysis method in the world [1, 2].

PLSR can better solve many problems, which cannot be solved or are difficult to solve by using a common multiple regression method. We are frequently restricted by objective conditions in traditional multiple statistics analysis. E.g. if a common least squares method is used in research on multiple correlations problem among independent variants, such variant multiple correlations will severely endanger the parameter estimate, expand model error and damage model stability. Multiple correlations problem is very complicated. No satisfactory answer is given in theory and method for a long period, so it troubles working staff, who are engaged in comprehensive variants in the dependent variants, recognizes information and noises in the system, and better overcomes actual system analysis [3]. A valid technical means is explored in the PLSR, which decomposes and screens data information in the system, extracts the strongly explaining the unfavourable role of multiple variant correlation in system modelling [4].

The PLSR provides one multiple-to-multiple regression modeling method, especially two variant groups include many variants with multiple correlations. Experiences indicate that the model based on PLSR method has unique strengths which are different from them of a traditional and classical regression analysis method (1G multiple statistics analysis method) when the observed data (samples) is less. The PLSR integrates strengths of principal component analysis, classic association analysis and linear regression analysis method in modeling. It not only provides one reasonable regression model in the analysis results, but also completes research contents such as principal component analysis and typical correlation analysis and provides some rich and deep information [5, 6].

Model selection problem is very important inmultiple regression analysis. The order identification of the PLSR is one hot subject studied by the statistics experts all the time. This paper introduces the modeling method of the PLSR analysis and principal component analysis and compares and analyzes the model from prediction view by combing individual cases in order to conduct preliminary discussion on mathematical model problems of PLSR.

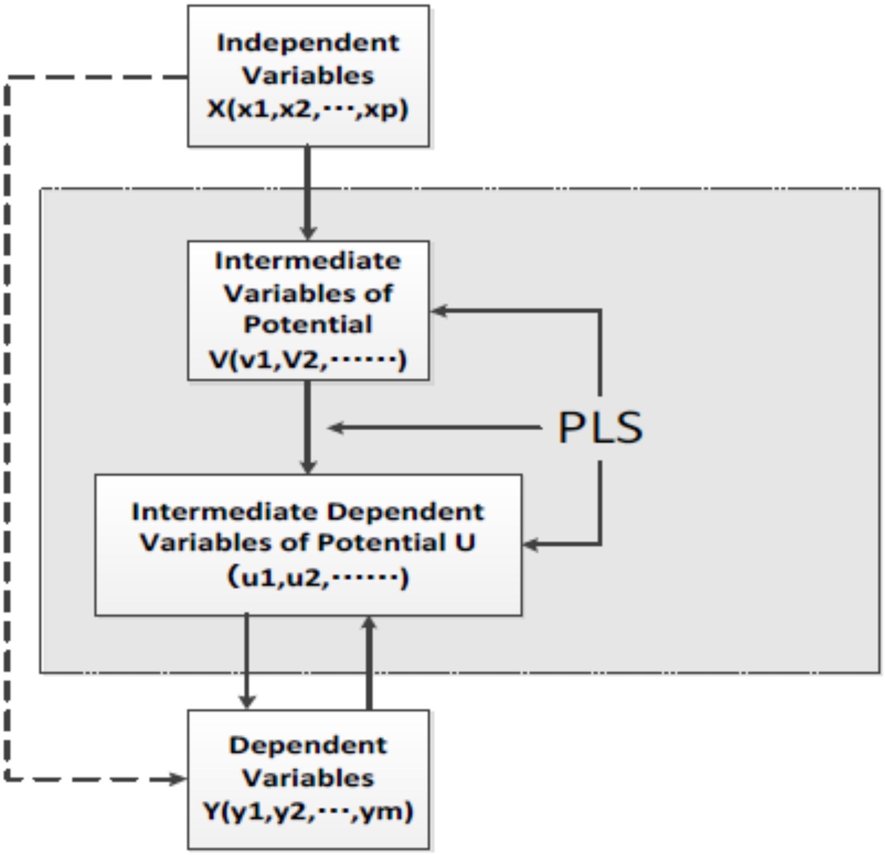
Independent Variables X(x1,x2,···,xp)
Intermediate Variables of Potential V(v1,V2,······)
PLS
Intermediate Dependent Variables of Potential U (u1,u2,······)
Dependent Variables Y(y1,y2,···,ym)

Fig. (1). Mathematic principle framework of PLSR.

2. PLSR ANALYSIS

The PLSR is an advanced multiple statistics analysis method, which studies relation between multiple dependent variants and multiple independent variants. This principle was first proposed by economist Herman Wold. The original assumption is that all independent variants are associated with the dependent variants and aims to find some linear combinations in the independent variant space and better explain mutation information of the dependent variants [7].

The basic idea of PLSR is shown as the Fig. (**1**). Generally the PLSR aims to establish the linear relation of the potential independent variants to the potential dependent variants. It can indirectly reflect the relation between the dependent variants and independent variants (For relation establish, refer to the solid line in the Fig. (**1**)). The potential independent variants and potential dependent variants reflect linear combination of the variants in the PLSR principle and should meet two assumptions:

1) Two groups of potential variants bear mutation information of the independent variants or dependent variants;

2) The correlation between the corresponding potential independent variants and potential dependent variants is maximized. Based on the above discussion, the establishment idea of the PLSR model is briefly described as follows [8, 9].

The modeling problem of p dependent variants Y (y_1, y_2, y_3, …, y_p) and m independent variant X(x_1, x_2, x_3, …, x_m) is considered. The basic method of the PLSR is described as follows: first extract the first component u1 (u1 is linear combination of X (x1, x2, x3, …, xm) and mutation information in the original independent variant set should be extracted as much as possible, extract the first component v1 from the dependent. If the regression equation is satisfactory, the algorithm will terminate.

The above steps describe the PLSR modeling process. Specific computing method deduction is not emphasized in this paper, so it will not be further described.

3. ALGORITHM ANALYSIS

To verify excellence of the PLSR analysis, this paper refers to a large number of related references for model verification. This paper uses physical training data of the male users of the health club given by the classical Linnerud as the PLSR analysis verification data. For data, refer to the Table **1**.

From the Table **1**, we know that the observation matrix of the independent variants and dependent variants are 20*3 data matrix. Different parameter variants have different dimensions, so the data dimensions (standardization) should be normalized prior to data processing. This paper correlates all

Table 1. Linnerud physical training data.

No	x_1	x_2	x_3	y_1	y_2	y_3
1	191	36	50	5	162	60
2	189	37	52	2	110	60
3	192	38	58	12	101	101
4	162	35	62	12	105	37
5	189	35	46	13	155	58
6	182	36	56	4	101	42
7	211	38	56	8	101	38
8	167	34	60	6	125	40
9	176	31	74	15	200	40
10	154	33	56	17	251	250
11	169	34	50	17	120	38
12	166	33	52	13	210	115
13	154	34	64	14	215	105
14	247	46	50	1	50	50
15	193	36	46	6	70	31
16	202	37	62	12	210	120
17	176	37	54	4	60	25
18	157	32	52	11	230	80
19	156	33	54	15	225	73
20	138	33	68	2	110	43

Table 2. Correlation coefficient matrix.

	x_1	x_2	x_3	y_1	y_2	y_3
x_1	1	0.7821	-0.3142	-0.3568	-0.5014	-0.2322
$x2$		1	-0.3634	-0.5122	-0.7011	-0.1980
x_3			1	0.6723	0.2541	0.0297
y_1				1	0.7012	0.3687
y_2					1	0.7412

standardized data by using the simple standardization processing process. The processing results are shown as the Table **2**.

From the above Table, we can know that single pole, bending and high jump variants are positively correlated with the pulse and are negatively correlated with the waistline and body weight, namely two variants are correlated. The variant groups include correlation. E.g. the body weight variant is highly correlated with the pulse variant, r=0.7821. Matlab 7.0 is used to process the standardized data, namely the function plsregress (x, y, n), is used. n indicates component number and the independent variant maximum is used, namely n=3. The variants are few in this paper, so the component number can be computed via 1-2-3 iteration. When n is 1, the ratio of the explaining variants is 61.2%. When n is 2, the ratio of the explaining variant is 89.7%. When n is 3, the ratio of the explaining variants is 93.5%. The results for n as 3 are computed.

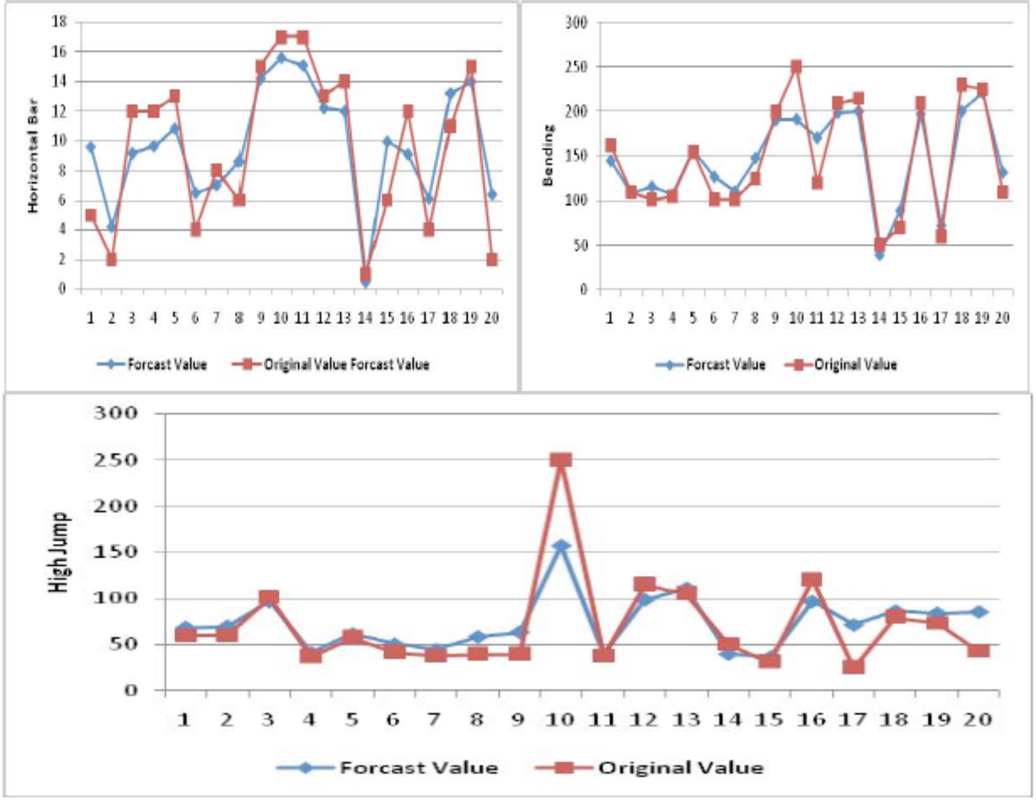

Fig. (2). Model prediction results.

To further investigate the fitting degree of the model, we should draw the actual values and prediction graph for comparison and analysis. The prediction results are shown as the Fig. (**2**).

The model prediction and computing results indicate that the PLSR analysis method can better predict the data and reserve the explaining variants remarkably correlated to the explained variants to most extent, so it is feasible.

CONCLUSION

The methods for processing relation between the independent variants and dependent variants are plentiful in research on modern science. The PLSR method is used in this paper, which is one multiple regression algorithm and quickly develops in recent years. With learning in this paper, we can know that the PLSR analysis have the basic features of the multiple linear regression analysis method and principal component analysis method and can quickly find the relation between the independent variants and dependent variants. Now the PLSR has been extensively applied in the engineering, economy and medical science. Now research on the PLSR analysis algorithm theory has been formed and is gradually improved. E.g. kernel partial least squares regression theory. Some scholars point out that the PLSR can provide precise mathematical model. The improved algorithms such as kernelpartial least squares regression have also some strengths, but they are on the preliminary research phase and should be discussed and studied later.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] B. Li, “Research on dam security monitoring statistics model based on PLSR”, M.S. thesis, *Xi’an: Xi’an University of Technology*, 2007.

[2] P. Wang, *Application of PLSR and kernel PLSR---with commercial house sale price as one example*, M.S. thesis, Kunming: Kunming University of Technology, 2012.

[3] P. Wang, G. H. Fu, and L. Dai, “Application of PLSR in commercial house sale price analysis”, *Statistics and Decision,* vol. 403, no. 7, pp. 84-86, 2014.

[4] R. X. Suo and F. L. Wang, “Prediction of per capital net income rural residents based on PLSR”, *Science and Comprehensive Research of Mathematical and Agricultural System,* vol. 27, no. 2, pp. 142-147, 2011.

[5] G. H. Chen and X. J. Wen, “Non-linear hidden kernel PLSR algorithm and its application”, *Journal of Wuhan University of Technology,* vol. 30, no. 12, pp. 114-116, 2008.

[6] Z. S. Wang and K. Z. Deng, “Multi-dimension kernel PLSR method of parameter recognition based on probability integral method”, *Journal of Rock Dynamics and Engineering,* vol. 30, no. 5, pp. 3864-3870, 2011.

[7] H. Y. Song, W. H. Gui, and C. H. Yang, “Application of dynamic prediction mode based on kernel PLSR method in steel converter blowing and melting”, *Journal of China Non-Ferrous Metals,* vol. 17, no. 7, pp. 1201-1206, 2007.

[8] H. W. Wang, *PLSR method and its application*. Defense Industry Press, 1999.

[9] G. B. Chen, "Application of PLSR in classroom teaching quality evaluation", *Technology Information,* vol. 15, no. 17-18, 2009.

Received: February 18, 2015 Revised: March 22, 2015 Accepted: March 31, 2015