A PARTIAL LEAST SQUARES REGRESSION MODEL OF RESPONSES AND CORRELATED PREDICTORS

Olasupo Ahmed O, A.I. Taiwo Peter I Ogunyinka, B.T. Efuwape, Abdullah K.-K.A

Olabisi Onabanjo University, Ago Iwoye, Ogun State, Faculty of Science, Department of Mathematical Sciences

Corresponding author: Ahmed Olalekan Olasupo, ahmed.olasupo@oouagoiwoye.edu.ng

ABSTRACT: *This work presents the partial least squares (PLS) regression model as applicable to a wide range of fields especially where there is a number of correlated explanatory variables. The strategy is to decompose variables into components. This can be contrasted with other methods such as principal component regression, ridge regression where vital variable may be accidentally deleted. It is also not dependent on any unknown factor upon formulating a model of two dependent variable and four independent variables. The model is subsequently applied to central bank of Nigeria data. The* R^2 *value of 0.9608 and 0.8607 were obtained for external reserves and surplus respectively.*

The work shows that partial least squares regression is a more robust and creditable model for multivariate regression model.

KEYWORDS: *Partial least Squares, Responses, Predictor External reserve, Surplus*

1. INTRODUCTION

The absence of multicollinearity is vital to a multiple regression model. In regression when several independent variables are highly correlated, this problem is called multicollinearity. When predictors suffer from multicollinearity, using OLS might lead to inflated regression coefficients. These coefficients could fluctuate in sign and magnitude as a result of a small variation in the dependent or independent variables [2]. Collinearity makes it more difficult to achieve significance of the collinear parameters. But if such estimates are statistically significant, they are as reliable as any other variables in a model and even if they are not significant, the sum of the coefficient is likely to be reliable. In this case, increasing the sample size is a viable remedy for collinearity when prediction instead of explanation is the goal [5].

There are many solutions to this problem, such as centered-score regression, Orthogonalization, partial least squares (PLS), ridge regression, and principal component regression (PCR).

In this research work, a partial least square regression model is formulated and analyzed for dependent variables and the correlated independent variables. PLS Regression has provided a solution to the problem of multicollinearity in regression models, due to the problem of small sample size, correlated and high number of predictors, and high noise to signal relationships, ecologist have opted to use PLS-R as an alternative to current regression methods used in ecology [4]. [1] introduced PLS method as a new tool for functional neuro-image analysis because of its exclusive way of generating spatial patterns of brain activity therefore explaining the relationship between image pixels and task or behavior. Many Econometric models which include time series data have a multicollinearity problem because economic variables usually have correlations with each other. In such situations, decomposition by PLS-R is ideal tool since it is designed to deal with this condition [6].

The purpose of this work is to formulate, estimate and evaluate the performance of the PLS model on the time series data collected from the website of the central bank of Nigeria.

The data consists of four explanatory and two response variations which includes External reserve, surplus available to FGN, investment, Share capital, Deposit and Operating expenses. However, the External reserves, surplus available to FGN serve as the response variable which covers a period of forty one (41) years within the period of 1976 and 2016 inclusively.

2. PARTIAL LEAST SQUARES

Multicollinearity refers to a situation, in which one or more predictor variables in a multiple regression model are highly correlated, if collinearity is perfect (EXACT), the regression coefficients are indeterminate and their standard errors are indefinite, if it is less than perfect, the regression coefficients although determinate but possess large standard errors, which means that the coefficients cannot be estimated with great accuracy ([GUJ95]).

In the modeling of PLS, the underlying assumption is that there exist some causal variables known as latent variables (LV's) that actually influence the system that is being investigated. The causal variables are unknown, so PLSR process is used to estimate the

exact number of these variables. The variance inflation factor (VIF) provides a measure of how the variance of the parameter estimate changes relative to a model in which all predictor variables are uncorrelated.

Nipal Algorithm

The PLS method, which in its classical form is based on the nonlinear iterative partial least squares (NIPALS) algorithm has been developed by Wold [7] with following algorithms:

Step 1: Compute the X – weight as $w = X^T y$, $w = |{}^{W}/_{W}|$, the X- scores $t = Xw$,the X-loadings as $p = X^T t/(t^T t)$

Step 2: Compute the Y- loadings as $q = Y^T t/(t^T t)$ and Y – scores as $u = Yq$.

Step 3: Deflate X and Y by subtracting the computed latent vectors from them as $X_{new} = X - tp^T$ and $Y_{new} = Y - tq$

Step 4: Store w,t,p,q,u in W,T,P,Q,U respectively and calculate final regression vector as $B_{PLS} = W^T(PW^T)^{-1}Q$.

3. APPLICATION OF PLS ON THE CBN DATA

The database of the central bank of Nigeria consists of time series data on all the sectors of the economy. It covers a period of forty one (41) years within the period of 1976 and 2016 inclusively. External reserve (Y_1) and Surplus available to FGN (Y_2) was regressed on the following determinants: Investment (X_1), Share capital (X_2), Deposits (X_3) and Operating Expenses (X_4). Multicollinearity could increase the variance of evaluated parameters. For diagnosing focus on the variance inflation factor (VIF) is shown in Table 1.

Table1. Variance Inflation Factor of Explanatory variable

X_1	X_2	X_3	X_4
7.6	5185.1	5106.2	4.9

Typically, we often use 10 as our 'threshold' at which we consider it to be a problem, but this simply a rule of thumb, To deal with this problem in our data, we apply PLS decomposing **X** and **Y** into components.

Table 2: Cross validation results for External reserve and Surplus

Y1	Y2	
Intercept	17131	516
X_1	-4	-1
X_2	-2	1
X_3	3	-1
X_4	1	1

The percentage of variance explained in y and x by the component are

X	0.9991	0.0005	0.0004	0.0000
Y	0.9071	0.0537	0.0011	0.0001

By plotting the percent variance explained in Y against the number of components as follows:

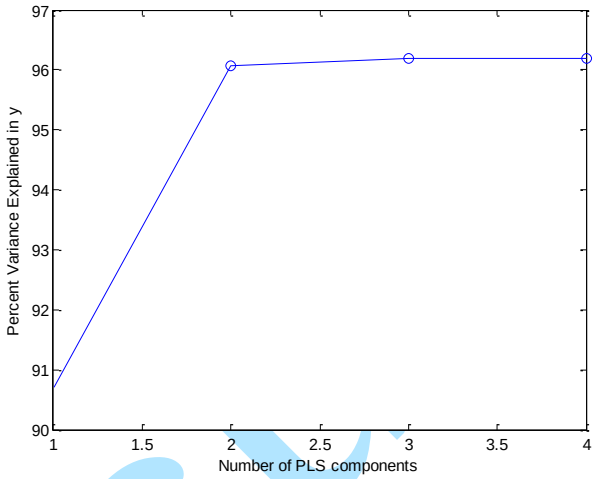

Figure 1. Curve for variance in Y in against number of components

Choosing the number of components in a PLS model is a critical step. The plot above gives a rough indication, showing that nearly 91% of the variance in `Y is` explained by the first component making a significant contribution.

The curve R^2 can be shown as follows:

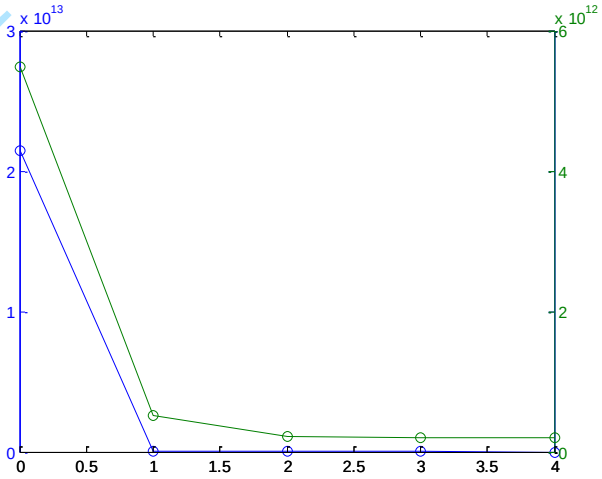

Figure 2. Curve of R2

$$R^2 \; of \; Y_1 = 0.9608$$
$$R^2 \; of \; Y_2 = 0.8607$$

The R^2 value of 0.9608 indicates that 96.1% of the total variation in Y_1 is explained by $X_1, X_2, X_3 and \; X_4$.

The R^2 value of 0.8607 indicates that 86.1% of the total variation in Y_2 is explained by $X_1, X_2, X_3 \; and \; X_4$.

CONCLUSION

In this study, PLS regression is analyzed and the objectives of the research were achieved and the findings are outlined below.

$$External\ Reserves = 17131 - 4(Investment) - 2(Share\ Capital) + 3(Deposit) + (Operating\ Expenses)$$

External Reserve = 17131 - 4(Investment) - 2(Share Capital) + 3(Deposit) + (Operating Expenses).

The coefficient of Investment i.e., 4 indicates that if X_1 is increased and holding X_2,X_3 and X_4 constant, then the External reserve will be decreased.
The coefficient of Share capital i.e., 2 indicates that if X_2 is increased and holdingX_1,X_3 and X_4 constant, then External reserve will be decreased.
The coefficient of deposit i.e., 3 indicates that if X_3 is increased and holding X_1,X_2 and X_4 constant, then External reserve will be increased.
The regression coefficient of operating expenses, i.e., 1indicates that if X_4 is increased and holding $X_1, X_2\ and\ X_3$ constant, then External reserve will be increased.
Also, Surplus available to

FGN = 516 - 4(Investment) + 2(Share Capital) - 1(Deposit) + (Operating Expenses).

The coefficient of Investment i.e., 4 indicates that if X_1 is increased and holding X_2,X_3 and X_4 constant, then surplus available to FGN will decrease.
The coefficient of Share capital i.e., 2 indicates that if X_2 is increased and holding X_1,X_3 and X_4 constant, then surplus available to FGN will increase.
The coefficient of deposit i.e., 1indicates that if X_3 is increased and holding X_1,X_2 and X_4 constant, then surplus available to FGN will decrease.
The regression coefficient of operating expenses, i.e., 1 indicates that if X_4 is increased and holding $X_1, X_2\ and\ X_3$ constant, then surplus available to FGN will increase.
Percentage of variance explained in y and x by the components were obtained and the graph was plotted, the result shows nearly 91% of the variance in y is explained by just only one of the components.
Also, the weight of the 4 predictors in each of the four components shows that the last components explains majority of the data, so, this suggests keeping one dimensional components as the final solution.
The result of the model of one dimensional component shows that

External Reserve = 7721.9 – 4.1(Investment) – 1.9(Share Capital) + 2.6(Deposit) + 2.0(Operating Expenses).

Surplus available to

FGN = 595.2+0.1(Investment) + 0.023(Share Capital) – 0.054(Deposit) -0.03(Operating Expenses).

The R^2 value of 0.9608 indicates that 96.1% of the total variation in Y_1 is explained by X_1,X_2,$X_3 and\ X_4$.
The R^2 value of 0.8607 indicates that 86.1% of the total variation in Y_2 is explained by X_1, X_2,$X_3\ and\ X_4$.

REFERENCES

[1] **A.R. Mclntosh, F.L Bookstein, J.V. Haxby, G.L. Grady** (1995), Spatial Pattern Analysis of Functional Brain Images Using Partial Least Squares, Neuroimage Journal, vol.3, 143-157, 1996

[2] **Fekedulegn, B. D; Colbert, J.J; Hicks, R.R; Schuckers, Michael E.** (2002). Coping with multicollinearity: An example on application of principal component analysis in Dendroecology. Newton Square, PA: U.S. Department of Agriculture, Forest Service, North eastern Research Station. 43p.

[3] **Gujarati D.N** (1995) Basic Econometrics. 4th Edition, United State.

[4] **Luis, M. Carrascal, Ismael Galvan and Oscar Gordo**, (2009) Partial least squares regression as an alternative to current regression methods used in ecology. ES28006 Madrid, Spain.

[5] **Leahy, R.L.** (2001) Overcoming resistance in cognitive therapy. Newyork, NY, US Guilford Press.

[6] **Ozlem Berak Korkmazoglu, Gulder Kemalbay** (2012), Academic research paper on Economics and business.

[7] **Wold, H**. (1966) Estimated of Principal Component and related models by iterative least squares. In P. R. Krishnajah (Ed.), Multivariate analysis (pp. 391-420). NewYork: Academic Press.