On the Partial Least Square Regression Modeling to Collinear Regressors: Contribution of Transportation Sector to Nigeria Economic Growth

Ogunnusi, O. Nurudeen[1*], **Ojo, G. Olugbenga**[2] **and Sikiru, O. Abdulwasiu**[3]

[1,2,3] Department of Mathematics & Statistics, Federal Polytechnic, Ilaro, Nigeria. P.M.B 50

E-mail: ogunnusioluwatobI@gmail.com; gabriel.ojo@federalpolyilaro.edu.ng
olusegun.sikiru@federalpolyilaro.edu.ng

Corresponding Author: **Ogunnusi, O. Nurudeen**[1*]

Phone Numbers: +2348134340254, +2348037458377, +2348032544458

Abstract

This study dwelt on the application of partial least square regression (plsr) modeling using the contribution of transportation sector to economic growth data. Twenty-nine (29) years data covering 1981 to 2019, extracted from Central Bank of Nigeria statistical bulletin which consist the contribution of road, air, water, rail, transport service, post and couriers' services was proxied with economic growth to confirm its pattern of contribution of the predictor variables. Ordinary Least Square Regression model was first fitted to the data to confirm if there exists multicollinearity in the set of predictors. Result of the Variance Inflation Factors (VIF) technique adopted indicated severe multicollinearity among the set of predictors of road, rail, transport service, post and courier services with associated coefficients of 33.0751, 7.3543, 21.2836 and 62.0324, violating the assumption of predictors independence. The partial least square model fitted using kernel function of the cross-validation technique indicated maximum of three components explaining 99.99% variance of the predictors and 99.24% variance of economic growth. Comparative analysis of the out of sample predictions in R for data science was carried out using RMSE evaluation technique as this indicated that PLSR is better fitted to data with multicollinearity effect as recorded from the lower RMSE of predictions.

***Keywords**: Transportation, PLSR, VIF, Multicollinearity, Regression, Kernel Algorithm*

1.0 Introduction

In recent past, transportation as one of the most important mode of systems in Nigerian sectors has derived an exponential pattern in moving of goods and services within and outside where most researchers have denoted as one of the highly pervasive factors in any economy. Hence, since the transportation industries plays a credible role in the aspect of work and leisure to people globally especially in Nigeria, majority comprehends the quality-of-life improvements, standard of living of the populace. Thence, these all helps to generate employments opportunities, increasing revenues from taxes, and in high-esteem create developmental growth of the nation economy tremendously (Nwaogbe, 2013). Therefore, transportation in any nation has been previewed such immensely of great interest as the major factor that support the process of growth in development of the economy. Aghadiaye and Adebayo (2013), denotes economic growth as increase in the quantity of goods and services produced in a nation which raises her national income. They further deduced that the process occurs whenever there is a quantitative increase in a country's input and output over a specific period of time. However, there is positive relationship between transportation and economic development which thereby links people with jobs, delivers products to markets, underpins supply chains and logistics networks which in turn key to foreign and domestic trades [Pteg (2014), Adeyi (2018), Adeniyi et al (2018)]

Multicollinearity is a threatening phenomenon to linear model formulation and its negative impact on prediction and inference is so enormous that its problems require serious attention (Ayinde et al(2015), Tyagi and Chandra(2017)). The problem of multicollinearity as experienced in modeling real life data originated from the existence of strong correlation between two or more explanatory variables thereby violating the independency assumption of explanatory variables (Gujarati and Porter(2009), Ayinde et al (2012), Daoud (2017)).

Linear model formulation in the presence of Multicollinearity is characterized by ill-conditioned $(X'X)$ where its inverse may not exist as a result of zero value determinant. Where the inverse exists, the diagonal elements of $(X'X)^{-1}$ may be so high thereby making the variance of the estimated parameters too high even when the $R^2$is large. Therefore, in the presence of multicollinearity, the variances and covariances of the estimated coefficients are large (Field; 2000).

Partial Least Squares regression analysis is a multivariate predictive modeling technique developed by Herman Wold in 1960 for use in the area of econometrics but its use was extended to chemistry and chemometrics (Geladi and Kowalski, 1986). Wold developed Partial Least Square to address the problem of weak theory and weak data (Wold,1983).

Partial Least Squares involves predicting Y based on X with the aim of describing or identifying the structures underlying the two variable (Abdi, 2003).

Lphan (2016) investigated the kind of lifestyle affecting the interests in river transport in Banjannasin. In their research, two models such as Partial Least Squares (PLS) approach (Second Order Confirmatory Factor Analysis) and Full Latent Variable Model were used. Since the objects of the study are the existing river transport and planned river transport whereby suited to the will of the users. Hence, based on the results of the model, it is deduced that there is a positive influence of the lifestyle with dimension cognition to the use of river transport.

Bjørn-Helge and Ron (2019) affirmed that Partial least square regression has been widely applied to variety of fields, including classifying wastewater pollution, to distinguish coffee beans, classify soy sauce, tumor classification for breast cancer and distinguishing between diagnoses of mental disorder.

Ozlem and Gulder (2012) analyzed the use of PLS in economic growth for Turkish economy and compared two algorithms of PLS and discovered that results of NIPALS and Kernel are not different but the kernel algorithm is being faster than the NIPALS algorithm for most problems.

Hence, Partial least squares as a multivariate estimation technique were adopted to withstand problems in data specifically, small datasets, missing values and multicollinearity. Therefore, aim of this study is to fit a Partial Least Square Regression (PLSR) model on contributory latent variables of transportation sector to Nigeria economic growth in the presence of multicollinearity and thereby provide relevant and reliable information about the past, present and likely future of Nigeria economic growth in the hand of transportation sector.

2.0 Methodology

The real-life data used for this study comprises of the Real Gross Domestic Product (RGDP), contribution of Road, Air, Rail/Pipeline transport, transport services, and Post and Courier Services spanning from 1981-2019 which are mainly sourced from the Annual Statistical report of the Central Bank of Nigeria (CBN) data site. The method of estimation adopted in this study was carried out using the Generalized Least Square, taking into account Partial Least Square Regression (PLSR) modeling approach thereby, testing each of the set of predictors for collinearity using Variance Inflation Factor approach. **pls** package of the R software developed by Bjorn-Helge et al., (2020) was adopted in fitting the specified models.

2.1 Model Specification

In this research, RGDP(Y) is a function of Road (X_1), Rail (X_2), Water (X_3), Air (X_4), Transport Services (X_5) and Courier/Postal services (X_6). The functional relationship is specified thus:

$$Y = f(X_{1i}, X_{2i}, \dots, X_{6i}) \qquad (2.1)$$

The econometric model in matrix form of this functional relationship is given as:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21}X_{31} & \dots. & X_{k1} \\ 1 & X_{12} & X_{22}X_{32} & \dots. & X_{k2} \\ \vdots & \vdots & \vdots \;\; \vdots & \dots. & \vdots \\ \vdots & \vdots & \vdots \;\; \vdots & \dots. & \vdots \\ 1 & X_{1k} & X_{2k}X_{3k} & \dots. & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ \vdots \\ e_k \end{bmatrix} \qquad (2.2)$$

$$Y = X\beta + e_i \qquad (2.3)$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}; \; X = \begin{bmatrix} 1 & X_{11} & X_{21}X_{31} & \dots. & X_{k1} \\ 1 & X_{12} & X_{22}X_{32} & \dots. & X_{k2} \\ \vdots & \vdots & \vdots \;\; \vdots & \dots. & \vdots \\ \vdots & \vdots & \vdots \;\; \vdots & \dots. & \vdots \\ 1 & X_{1k} & X_{2k}X_{3k} & \dots. & X_{kn} \end{bmatrix}; \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}; \; e_i = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ \vdots \\ e_k \end{bmatrix} \qquad (2.4)$$

Where,

2.2 Test for Multicollinearity

Implications in the presence of multicollinearity is derived such in the matrix equations below;

If $|X'X| \to 0$, then, $(X'X)^{-1} \to \infty$ (2.5)

Therefore, Var(b) = $S^2(X'X)^{-1} \to \infty$ (2.6)

Where the test statistic $t = \frac{b_i - B_i}{S(b_i)}$ → has small value (2.7)

2.3 Collinearity check using Variance Inflation Factors

Consider the following linear model with k independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon. \quad (2.8)$$

The standard error of the estimate of β_j is the square root of the j+1, j+1 element of

$$\sigma_\varepsilon = s^2 (X'X)^{-1} \quad (2.9)$$

By definitions, $\boldsymbol{s}$ is deduced as the Root mean Squared Error (RMSE), Root Mean Square Error Squared $(\text{RMSE})^2$ whereof, is an unbiased estimator of the true variance of the error term $\boldsymbol{\sigma^2}$. $\boldsymbol{X}$ is the regression design matrix, a matrix such that $X_{i,\ j+1}$ is the value of the j^{th} independent variable for the i^{th} case or observation, such that $X_{i,1}$ equal 1 for all i.

Hence, it turns out that the square of the standard error in equation (2.9), the estimated variance of the estimate of β_j, can be equivalently expressed as

$$Var\left(\hat{\beta}_j\right) = \frac{S^2}{(n-1)} \cdot \frac{1}{1-R_j^2} \quad (2.10)$$

From equation (2.10), $R_j^{\,2}$ is denoted as the multiple R^2 for the regression of X_j on the other covariates; that is, a regression that does not involve the response variable Y in which its identity separate the influences of several distinct factors on the variance of the coefficient estimate such as $\boldsymbol{s^2}$ shows a greater scatter in the data around the regression surface leads to proportionately more variance in the coefficient estimates, n is a greater sample size results in proportionately less variance in the coefficient estimates, $Var\left(\hat{\beta}_j\right)$ deduced a greater variability in a particular covariate leads to proportionately less variance in the corresponding coefficient estimate.

However, the Variance Inflation Factors (VIF) such as $\frac{1}{1-R_j^2}$ reflects all other factors that influence the uncertainty in the coefficient estimates which is always equals to one (1) when the vector X_j is orthogonal to each column of the design matrix for the regression of X_j on the other covariates. By contrast, the VIF is greater than 1 when the vector X_j is not orthogonal to all columns of the design matrix for the regression of X_j on the other covariates. Therefore, in lieu of the above facts, the VIF is invariant to the scaling of the variables.

2.4 Mathematical Estimation of Variance Inflation Factors

Step one: Calculate k different VIFs, one for each X_i by first running an ordinary least square regression that has X_i as a function of all the other explanatory variables in the first equation. If i = 1, for example, the equation would be

$$X_1 = \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_k X_k + C_0 + e \quad (2.11)$$

Where C_0 is a constant and e is the error term

Step two: Then, calculate the VIF factor for $\hat{\beta}_i$ with the following formula:

$$VIF = \frac{1}{1-R_i^2} \quad (2.12)$$

Where R_i^2 is the coefficient of determination of the regression equation in step one, but with X_i on the left-hand side, and all other predictor variables on the right-hand side.

Step three: Analyze the magnitude of multicollinearity by considering the size of $VIF\ \hat{\beta}_i'$. A common rule of thumb is that if $VIF\ \hat{\beta}_i' > 10$ then multicollinearity is high. The square root of the variance inflation factor tells how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the model.

For simplicity, we focus on the simple regression problem of predicting a single response variable. In the usual multiple linear regression (MLR) we are interested in equation 2.3 above. The problem often arises if X is likely to be singular, either because of the number of variables exceeds the number of observations or because of the multicollinearity. The scores and loadings are chosen in such a way to maximize the covariance between X and Y.
However, the least square solution of equation 3.3 is given by

$$\hat{B} = (X'X)^{-1} X'Y \quad (2.13)$$

The problem often is that $X'X$ is singular, either because the number of variables (columns) in X exceeds the number of objects (rows), or because of collinearities. PLSR circumvent this by decomposing X into orthogonal scores T and loadings P

$$X = TP \quad (2.14)$$

and regressing Y not on X itself but on the first a columns of the scores T.

2.5 Kernel Algorithm of Estimating PLSR

According to Bjon-Helge and Ron (2007), Latent Variables (LVs) as a partial least square regression components are iteratively obtained whereby the step process starts with the Singular Value Decomposition (SVD) of the cross-product matrix in equation (2.15).

$$A = X'X \quad (2.15)$$

This inclusively shows variation in both X and Y, and as well correlation between them. Hence, Left and Right first singular vectors such as w and q, are used as weight vectors for X and Y, respectively, to obtain scores t and u;

$$t = Xw = Ew \quad (2.16)$$

$$u = Yq = Fq \quad (2.17)$$

where E and F are initialized as X and Y, respectively. The X scores t is often normalized:

$$t = t/\sqrt{t't} \quad (2.18)$$

By regressing, X and Y are loaded against the vector t in the previous equations;

$$p = E't \quad (2.19)$$

$$q = F't \quad (2.20)$$

Finally, the data matrices are “deflated”: the information related to this latent variable in the form of the outer products tp' and q' , is subtracted from the (current) data matrices E and F.

$$E_{n+1} = E_n - tp' \quad (2.21)$$

$$F_{n+1} = F_n - tq' \quad (2.22)$$

The estimation of the next component then can start from the SVD of the cross-product matrix $E_{n+1}F_{n+1}$. After every iteration, vectors w, t, p and q are saved as columns in matrices W, T , P and Q, respectively. Another way of representing the weights such that all columns are related to the original matrix X is given by

$$R = W(P^T W)^{-1} \tag{2.23}$$

As in Principal Component Regression (PCR) where Y is regressed on X, scores T is used to calculate the regression coefficients and thereby convert T to real of the original variables by pre-multiplying with matrix R ($since\ T\ =\ XR$), then

$$B_{PLS} = R(T^T T)^{-1} T^T Y = RQ^T \tag{2.24}$$

The number of optimal components is determined by cross-validation. Hence, the response variable (GDP) can then be predicted using multivariate regression formula thus

$$\hat{Y} = XB_{PLS} \tag{2.25}$$

3.0 Results and Discussion

Table 3.1: Descriptive Statistics of Analyzed Variables

Variables	Mean	SD	Min	Max
Real GDP (Y)	38043	29557.26	13779	144210
Road Transport (rt) X_1	438.3	508.0629	127.5	2727.5
Rail Transport (rtp) X_2	1.913	3.167089	0.040	10.160
Water Transport (wt) X_3	3.920	1.731346	2.540	10.125
Air Transport (at) X_4	34.55	36.87315	13.260	198.62
Transport Services (ts) X_5	19.28	21.80278	3.330	93.530
Post and Courier Services (pcs) X_6	8.976	6.790526	2.680	22.620

Values expressed in Billion Naira

Source: Extracted from R-Studio Output

Table 3.1 depicts the descriptive statistics of the analyzed variables of transportation sector such as contribution of road transport, rail transport, water transport, air transport, transport services and post and courier services sub sectors and Nigeria economic growth. Analysis indicates that there is an average 38,043 GDP recorded between 1981year 2000 on an average spread of ±29,557.26. Also, minimum and maximum GDP was found to be 13779 and 144210 for the response variable under study. This implies that since the existence of economic growth records of Nigeria ab-initio, highest contribution of all sectors of the economic were found to reach its peak on a maximum value of 144210. Summary of predictor variables of contributory road transport, rail transport, water transport, air transport, transport services and post and courier services are were also captured in the table with their average, minimum and maximum measurements recorded. In addition, the maximum value for the set of the six predictors indicated the point where the sub-sectors can be adjudged to have the highest predicted values over the years under study.

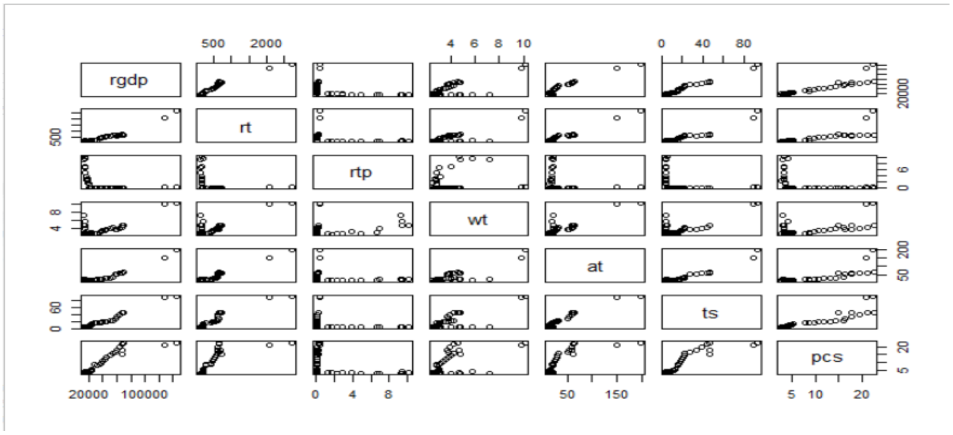

Figure 3.1: Paired variables assessment

It can be seen from the paired variables assessment in figure 4.1 that there is relationship existing between economic growth and the selected predictor variables. However, it can also be assessed that the set of predictors were positively related and can adjudged to be collinear, which can thereby be confirmed using Variance Inflation factors of the classical linear regression model as shown in table 3.2.

In view of this, we split the dataset into train and test where the trained dataset represented 77% of the entire data and the test dataset takes the remaining 23%. These splits were used to fit the classical linear model and the partial least square model as a set of latent variables.

Table 3.2: OLSR Result on Trained Data (Response Variable *Y* = Real Gdp)

Variables	Parameter	Coefficient	t-Value	Pr(>|t|)	VIF
Constant (Intercept)	α	7878.256	9.074	0.0000	-
X_1: Road Transport (rt)	β_1	24.415	4.074	0.0000	33.0751
X_2: Rail Transport (rtp)	β_2	-771.702	-5.785	0.0000	7.3543
X_3: Water Transport (wt)	β_3	184.990	0.595	0.5576	3.5602
X_4: Air Transport (at)	β_4	188.722	3.810	0.0009	5.0281
X_5: Transport Services (ts)	β_5	35.650	0.329	0.7448	21.2836
X_6: Post and Courier Services (pcs)	β_6	1408.417	4.542	0.0001	62.0324

R^2 = 0.9958; ***Adj. R^2*** = 0.9948; ***MSE_{OLS}*** = 923; ***F-statistic*** on 6 and 24 DF = 33.72;
p-value (F-statistic) = 0.00000
Source: *Extracted from R-Studio Output*

The model specification of table 4.1 is written as:

$$\hat{Y} = XB_{OLS} = 7878.256 + 24.415X_1 - 771.702X_2 + 184.990X_3 + 188.722X_4 + 35.650X_5 + 1408.417X_6 \quad (2.26)$$

Table 3.2 indicates estimates of linear relationship existing between Nigeria economic growth and the duo of predictor variables of the trained dataset under study. The R-square of 0.9958 indicates the existence of high positive relationship between the set of predictors and response variable which also implies that about 99.6% variation in real Gross Domestic Product (GDP Y) could be largely explained by variation in Road, Rail, Water, and Air transport with inclusion of transport services, post and courier services respectively. This also indicates that adding other independent variables to the model gave an adjusted R-squared of 99.5% as it reveals the validity of the coefficient of multiple determinations R^2 of 0.9958.

The intercept of 7878.256 shows the autonomous Real GDP when the predictor variables are held constant. This value implies that Real GDP would be positively inclined without the influence of the considered predictors, and has been in existence ab-initio. Also, a unit increase in Road Transport, Water Transport, Air Transport, Transport Services and Post and Courier Services will give an incremental rate of 24.415, 184.990, 188.722, 35.650 and 1408.417 incremental rate in economic growth under study. Only Rail Transport sub-sector was found to have reduced economic growth as the coefficient of determination was found to be negatively inclined. The positively inclined variables were within apriori expectation as they are to have positive contribution to the response variable. The negative influenced experienced in the rail transport sub-section may be due to external forces rallying round the sub-sector in terms of development in the past.

Individual contribution of the aforementioned predictors' variables to the model can be evidenced from the t-statistic column of the table 3.2. Test revealed that Road Transport, Rail Transport, Air Transport and Post and Courier Services are statistically significant in the contribution of predicting economic growth at 1% significance level (p-value < 0.01). It can be deduced from the test of significance that variables of water transport and transport services do not contribute significantly to the model (P-value > 0.10) but can be adjudged to be a variable of importance.

The presented F- statistic value of 33.72 (df 6 and 24) with P-value of 0.000 < 0.01 can be adjudged that all the predictors variables jointly contributed significantly to the measured rates of economic growth observed. Although, the OLS model may be unbiased as it requires no tuning parameter but may not be efficient due to the collinearity in the set of predictors as these could increase variance of evaluated parameters. For diagnosing collinearity problems, Variance-inflation factor (VIF) was focused on as shown in Table 3.2. Typically, 5 was used as the threshold at which it is considered it to be a problem, simply a rule of thumb. To deal with this problem in our data, we apply PLS decomposing **X** and **Y** into components. To determine the optimal number of components to take into account cross-validation method was mainly used by minimizing the root mean squared error of prediction (RMSEP). The cross-validation is shown in Table 3.3.

Table 3.3: Cross Validation Result for Real GDP

Validation: RMSEP	**1 component**	**2 component**	**3 component**	**4 component**	**5 component**	**6 component**
Cross Validation	2593	2141	1426	1711	1384	1357
% Variance *(X)*	99.85	99.93	99.99	100.00	100.00	100.00
% Variance Explained *(Real GDP)*	96.18	98.37	99.24	99.34	99.58	99.58
Intercept				12968		

Source: *Extracted from R-Studio Output*

From Table 3.3, it can be seen that three components explain 99.99% of the variance of **X** and 99.24% of variance of **Y**. It is often simpler to judge the RMSEP by plotting them against the number of components as shown in figure 3.2 and 3.3 respectively.

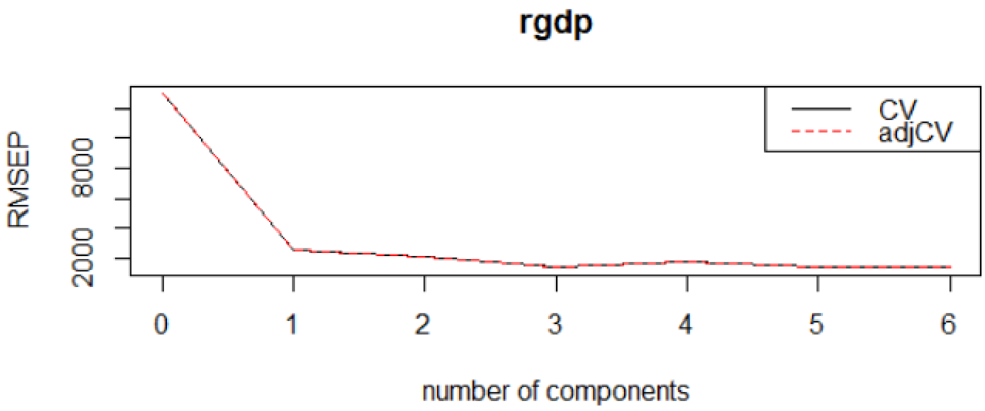

Figure 3.2: Cross-validated RMSEP curves for Real GDP data

The number of selected components is three since the cross validation error RMSEP (1426) does not show a significant decrease after three components.

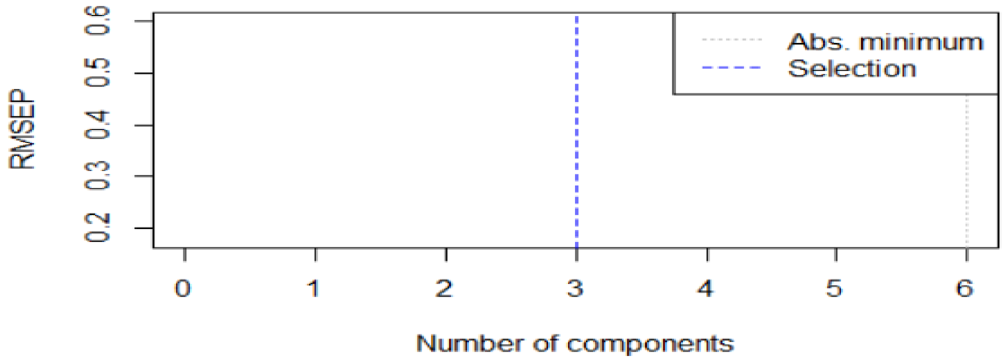

Figure 3.3: Cross-validated RMSEP curves for Real GDP data

The cross-validated RMSEP curves for Real GDP data in figure 4.3 also confirmed an optimal three components in the data.

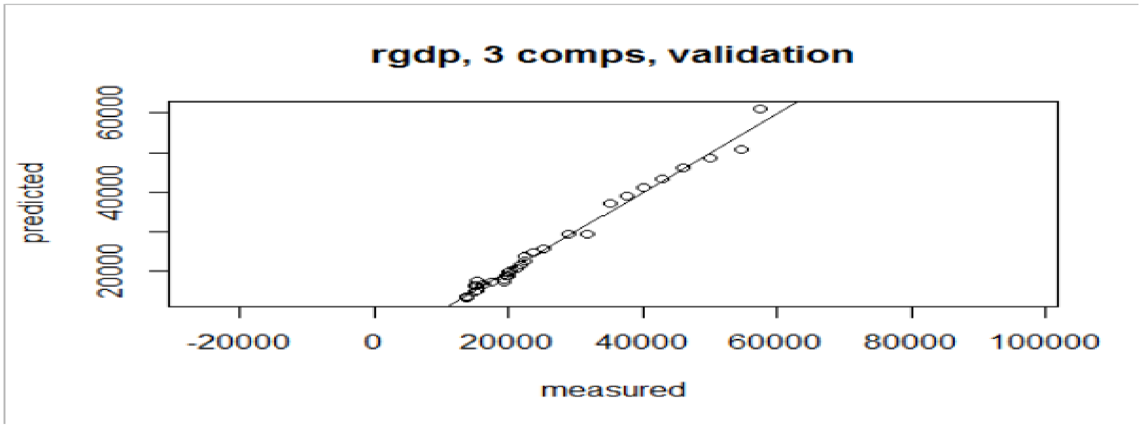

Figure 3.4: Cross-validated predictions for the Nigerian Transport data

After choosing the number of components, different aspect of the fit was inspected by plotting the cross-validated predictions as given in figure 3.4 which shows the cross-validated predictions with three components versus measured values. There is no curvature or other anomalies as indication of grouping or outliers. The scores for three components obtained by the algorithms Kernel is as obtained in table 3.4

Table 3.4. Scores for X and Y by Kernel Algorithm

s/n	*Scores for X (T)*			*Scores for Y (U)*		
	t_1	t_2	t_3	u_1	u_2	u_3
1	-48.869	-7.723	-6.029	-143.384	-14.256	-11.125
2	-102.693	-6.730	-0.486	-146.921	-6.671	0.101
3	-143.193	-5.617	2.724	-161.633	-3.497	3.788
4	-143.193	-5.617	2.724	-162.547	-2.919	4.594
5	-107.247	-4.830	2.198	-147.324	-6.045	-2.069
6	-126.078	-3.351	0.603	-143.643	-2.649	1.194
7	-124.310	-2.027	1.600	-143.307	-2.865	-1.427
8	-122.575	-1.196	1.343	-130.977	-1.267	-0.121
9	-119.638	-0.265	0.690	-116.991	0.399	1.131
10	-116.607	-0.402	0.812	-90.931	3.873	7.280
11	-110.446	-0.231	0.808	-92.312	2.735	5.051
12	-100.811	0.307	0.535	-86.855	2.105	3.062
13	-92.409	0.791	-2.279	-82.866	1.439	1.104
14	-89.746	2.217	-2.799	-82.204	1.138	-1.838
15	-87.896	2.596	-2.321	-77.356	1.590	-1.713
16	-84.187	2.925	-1.726	-66.669	2.642	-0.482
17	-78.545	3.318	-1.411	-58.748	2.986	-0.566
18	-71.745	3.976	-0.749	-51.702	3.023	-1.6234
19	-65.686	4.837	0.202	-50.192	2.337	-4.256
20	-59.281	5.108	0.843	-34.138	3.792	-2.241
21	-50.125	5.476	1.496	-13.672	5.498	0.038
22	-8.838	5.679	3.530	34.148	6.484	1.371

23	-5.722	5.313	4.380	69.807	11.392	10.352
24	145.383	0.989	-4.056	112.715	-4.928	-10.075
25	172.060	0.676	-3.991	144.521	-4.154	-8.225
26	203.002	0.234	-3.722	177.184	-3.894	-7.029
27	236.260	-0.235	-3.421	215.113	-3.190	-5.032
28	271.966	-0.743	-3.080	255.157	-2.535	-3.052
29	309.535	-1.353	-2.650	304.966	-0.689	1.131
30	349.085	-1.987	-2.308	366.600	2.642	7.881
31	367.809	-2.031	16.257	404.164	5.484	12.795

Source: *Extracted from R-Studio Output*

Table 3.4 represents the score matrix for X indicating the linear combination of weight matrix W and X which is expressed as $T = XW$. The t-columns are referred to as the latent vectors used in fitting the Partial Least Square models. The Y scores also showed the latent vector of the dependent variables due to the introduction of normalization in the data.

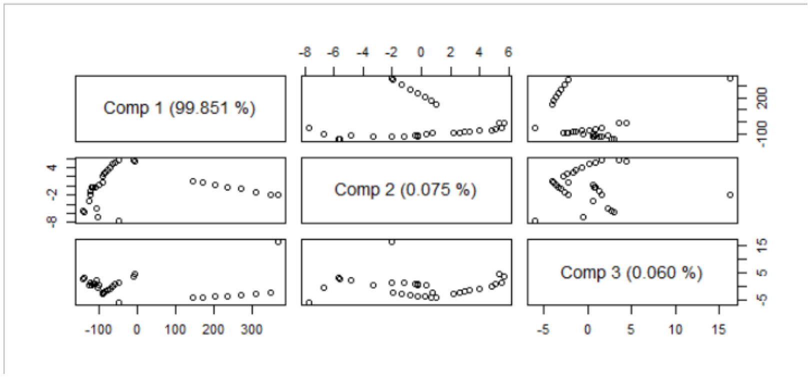

Figure 3.5: Score plot for the Nigerian Transport Data

The score plot of figure 3.5 showed the component that loaded more heavily on the data as 99.85% of the entire predictors' variables explains the response. Other two components contributed to about 0.135%

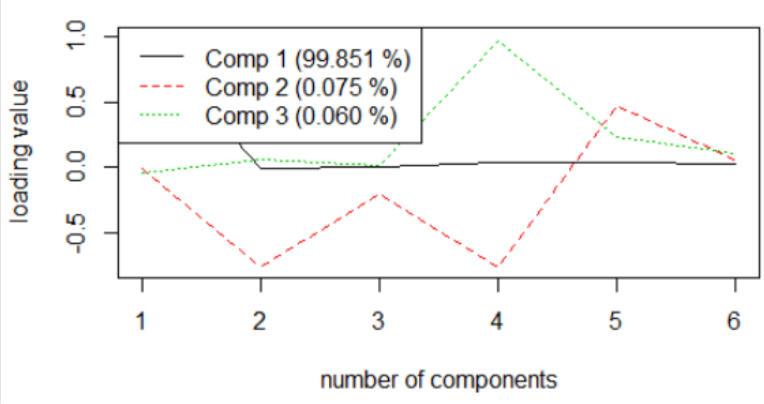

Figure 3.6: Loading plot for the Nigerian Transport Data

Table 3.5 Weights for Score, Loading Matrix of X and regression Coefficients by Kernel Algorithm

Variables	*X Loadings*			*Beta Coefficients*
	p₁	***p₂***	***p₃***	Constant =3.45767
Road Transport (X_1)	0.998			2.570
Rail Transport (X_2)		-0.773		34.785
Water Transport (X_3)		0.201		-9.188
Air Transport (X_4)		-0.762	0.971	13.836
Transport Services (X_5)		0.477	0.234	28.664
Post and Courier Services X_6			0.109	10.957

Source: *Extracted from R-Studio Output*

The response variable is predicted using multivariate regression formula $\hat{Y} = XB_{PLS}$

$$\hat{Y}_{PLS} = 3.45767 + 2.570rt + 34.785rtp - 9.188wt + 13.836at + 28.664ts + 10.957pcs$$

The loadings value p_1, p_2 and p_3 in table 3.5 indicates the amount of the explanation of components to the explanatory variables. Furthermore, showed that component p_1 loaded more on road transport, while$p_2$loaded more on rail, water, air, and transport services with p_3 having to load more on-air transport, transport services and post and courier services. As deduced from the plsr coefficients, only contribution of water transport system was found to be negatively related with economic growth as this do not conform with apriori expectation.

Table 3.6: Out of sample Predictions Result

Observed RGDP	**Predicted RGDP $\hat{Y} = XB_{OLS}$**	**Predicted RGDP $\hat{Y} = XB_{PLS}$**
59929.89	59662.61	33638.77
63218.72	63684.03	35757.19
67152.79	67350.32	36762.03
69023.93	70348.51	37261.37
67931.24	63241.31	35703.50
68490.98	60225.57	35110.67
127736.8	120452.83	38659.58
144210.5	148442.67	39308.26
RMSEP	**4518.874**	**1426.000**
R^2	**0.983**	**0.934**

Source: *Extracted from R-Studio Output*

From the out of sample predictions result, comparative analysis of the OLSRM and PLSRM indicated that PLSR predicted the observed RGDP well compared to the OLSRM since the Root means square (RMSE) of predictions of the PLSRM is lower than its counterpart of OLSRM. It can also be seen from the predicted GDP that large variances existed between the observed and predicted RGDP taking the partial least square regression model into consideration as compared to the predicted OLSR values which are closer than the observed values. Although, higher R-square value of the OLSRM does not make it better in terms of prediction as compared with the RMSEP

4.1 Conclusion

Based on the empirical analysis of result findings, it was deduced that: Ordinary least square model was found to be unbiased but inefficient in predicting the influence of transportation sector on economic growth, the set of predictors were found to be significantly related with each other, as a result of the collinearity existing between the variables; Partial Least square regression model was fitted to solve the problem, cross validated method of selecting numbers of components were adopted, contributory sub-sectors of road transportation, rail, air, post and courier services were found to significantly contribute to economic growth of Nigeria, water transportation and transport services do not have significant influence on economic growth, transportation sector was found to significantly predict economic growth in the absence of collinearity effect, partial least square regression modeling technique predicts well the trained data as evidenced from the RMSE. From the findings emanating from this study, we hereby recommend that:

i. The developed model (PLSR) could be used by policy makers to forecast economic growth taking into consideration, contributors of transportation sector of the economy.
ii. Transportation sector of the economy should be improved upon especially in the area of constructing goods roads as the road transport variable possessed the largest contribution.
iii. Other algorithms such as NIPALS should also be adopted by researchers in solving collinearity problem using PLSR approach.
iv. Policy makers should focus more on how this sector can further improve the Nigeria economic growth through construction of more rail lines, rehabilitation of water transport system, encouraging foreign direct investment in the air transport subsector which will further create employment opportunities for the citizenry.

Reference

Adeniyi, J.O., Akinrinmade, Y., and Abiodun, A. L. (2018) Analysis of Road Transport Impact on Rural Development in Nigeria. A Study on Akure North Local Government Area, Ondo State. *International Journal of New Technology and Research. ISSN:2454-4116, Volume-4, Issue-3, PP 102-110.*

Adetose, E. O. and Oluwatosin, O.O. (2020) Seaport development as an agent for economic growth and international transportation. *European Journal of Logistics, Purchasing and Supply Chain Management. ISSN 2054-0949, Vol.8 No.1, pp.19-34.*

Adeyi, E.O. (2018) The Impact of Transportation on Economic Development in Nigeria. *International Journal of Contemporary Applied Researches. ISSN: 2308-1365, Vol. 5, No. 8, www.ijcar.net.*

Ayinde K, Lukman, A., and Arowolo, O (2015). Combined parameters estimation methods of linear regression model with multicollinearity and autocorrelation. *Journal of Asian Scientific Research. DOI: 10.18488/JOURNAL.2/2015.5.5/2.5.243.250*

Ayinde, K., Alao, R.F., and Ayoola, F.J. (2012). Effect of multicolinearity and autocorrelation on predictive ability of some estimators of linear regression model. *Mathematical Theory and Modeling 2* (11), 41 – 52.

Bjørn-Helge, M. and Ron, W. (2007) The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software. Volume 18, Issue 2, PP1–24.* http://www.jstatsoft.org/

Bjørn-Helge, M. and Ron, W. (2019) Introduction to the pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software.*

Bjorn-Helge, M., Ron, W., Kristian, H. and Paul, H. (2020) Partial Least Square and Principal Components Rgression. *CRAN. URL* *https://mevik.net/work/software/pls.html*

Daoud J.I (2017). Multicollinearity and regression analysis. *J. Phys.: Conf. Ser. 949 012009*

Edith, A.O. and Adebayo, A.E. (2013). The Role of Road Transportation in Local Economic Development: A Focus on Nigeria Transportation System. *Developing Country Studies. ISSN: 2225-0565 Vol.3, No.6.* *http://www.jstatsoft.org/.*

Geladi, P; Kowalski, B (1986). Partial least squares regression: a tutorial. *Analytica Chimica Acta 185*, 1 – 17.

Gujarati, DN; Porter, DC (2009). *Basic Econometrics*, 5th edition., New York, Mc Graw-Hill.

Lphan, F. R., Agus, T.M., and Bagus, H.S. (2016) The Analysis of Lifestyle Affecting the Choice on River Transport in Banjannasin. *International Business Management*. ISSN: 1993-5250: 10 (19), PP 4090-4698.

Nwaogbe, O. R., Wokili, H.O., and Asiegbu, B. (2013) An Analysis of the Impact of Air Transport Sector to Economic Development in Nigeria. *IOSR Journal of Business and Management (IOSR-JBM). e-ISSN: 2278-487X, p-ISSN: 2319-7668. Volume 14, Issue 5, PP 41-48.*

Okorie, S. (2020) Transportation and Economic Development Nexus in Nigerian Economy. *World Journal of Innovative Research. ISSN: 2454-8236, Vol. 8, Issue 4, PP 59-66*

Ozlem B.K, Gulder K (2012). Econometrics application of partial least squares regression: an endogeneous growth model for Turkey. Journal of Procedia-Social and Behavioral Sciences 62(2012) 906-910. DOI: 10.1016/j.sbspro.2012.09.153

Tyagi, G; Chandra, S (2017). Note on the performance of biased estimators with autocorrelated errors. *International Journal of Mathematics and Mathematical Sciences*. Volume *2017*, 12 Pages, Article ID 2045653.

Pteg, (2014) Transport Works for Growth and Jobs. *The voice of Urban transport*. www.pteg.net.

Wold, H. (1966). Estimation of principal components and related models by iterative least square. In P.R. Krishnaiah, editor, Multivariate Analysis, 391 420. *Academic Press, New-York.*

Wold, S; Martens, H; Wold, HOA (1983). The multivariate calibration method in chemistry solved by PLS method In: *Proceedings of Conference Matrix Pencils Lecture Notes in Mathematics*, Ruhe A. and Kagstrom B., eds., Springer-Verlag. Heidelberg, 286 – 293.

Zizi, H. Z. and Avanenge, F. (2016) An analysis of the issues and challenges of transportation in Nigeria and Egypt. *The Business and Management Review*: *Vol.7 No. 2*.