JIRBDAI Journal of Institutional Research, Big Data Analytics and Innovation

Solving Multicolinearity Problem in a Linear Regression: A Comparative Study of Ordinary Least Squares and Partial Least Squares Regression

Babafemi D. Ogunbona, Folorunsho O. Balogun, Kayode S. Famuagun
Department of Mathematics, Adeyemi Federal University of Education, Ondo
Correspondence Author email: ogunbonabd@aceondo.edu.ng

ABSTRACT Ordinary Least Squares (OLS) estimator usually yields inefficient estimates when multicollinearity is present in a Linear Regression Model. The inefficiency of OLS can be mitigated by Partial Least Squares Regression (PLSR). However, this method requires selecting latent variables in order to yield efficient estimates of regression parameters. This paper proposes using weighted standard errors and ranking standard errors of regression coefficients for latent variable extraction, alongside model selection methods such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Adjusted R Squared (ARS), and Standard Error of Regression Coefficient (SER). Monte-Carlo experiments on a Linear Regression Model with six explanatory variables were conducted one thousand (1000) times across eleven levels of multicollinearity [low (0.0, 0.2 and 0.4), moderate (0.6 and 0.8), high (0.95, 0.96, 0.97, 0.98,) and very high (0.99 and 0.999)] and five levels of sample size [small (20), medium (30 and 50) and large (100 and 250)]. Results indicated that while OLS estimates are preferred overall, Partial Least Squares with BIC becomes preferable at high multicollinearity levels. Additionally, the efficiency of OLS improves with larger sample sizes despite multicollinearity.

Keywords: Multicollinearity, OLS, PLS, Linear Regression, Latent Variables

I. INTRODUCTION

Multicollinearity refers to the presence of high intercorrelation among two or more predictor variables in a regression model. It is a common issue in various fields, including economics, social sciences, and machine learning, where multiple factors influence a dependent variable [1, 2]. While multicollinearity does not affect the predictive power of the model, it poses challenges in interpreting the individual coefficient estimates. This problem arises because the regression coefficients become sensitive to small changes in the data, making it difficult to disentangle the unique impact of each predictor variable on the dependent variable [3, 4]. To address multicollinearity, researchers and practitioners have proposed two main approaches: variable selection and dimensionality reduction. The variable selection approach involves choosing a subset of the original predictors that have lower correlations with each other [5]. Popular methods within this category include forward selection, backward elimination, and stepwise regression. For instance, a study by [6] demonstrated the application of stepwise regression in selecting a subset of variables to improve the prediction of water quality parameters. While variable selection can mitigate multicollinearity, it may result in the loss of important information if variables with significant relationships to the dependent variable are discarded. On the other hand, dimensionality reduction techniques aim to transform the original set of predictors into a new set of variables with lower correlations [7]. This can be achieved through factor analysis, Principal Component Analysis (PCA), or Partial Least Squares (PLS) regression [8, 9]. These methods create new, uncorrelated variables that capture the maximum amount of variance in the data, thereby reducing the multicollinearity problem.

Ordinary Least Squares (OLS) regression is a fundamental technique for linear regression analysis. It estimates the unknown parameters of a linear model by minimizing the sum of squared differences between observed and predicted values of the dependent variable. However, OLS regression assumes the absence of multicollinearity among the predictor variables. In the presence of high multicollinearity, OLS

estimates can become biased and inconsistent, leading to unreliable conclusions [10]. Partial Least Squares (PLS) regression, a dimensionality reduction technique, addresses multicollinearity by constructing new latent variables called components. These components are linear combinations of the original predictor variables, designed to maximize the covariance with the dependent variable while being uncorrelated with each other [11]. PLS regression performs regression analysis on these components instead of the original variables, reducing the impact of multicollinearity. PLS has gained popularity due to its ability to handle multicollinear data and provide accurate predictions, particularly when the number of predictor variables is large [12]. The comparative study presented in this study aims to provide a comprehensive evaluation of OLS and PLS regression in the presence of multicollinearity. By simulating datasets with varying levels of multicollinearity and sample size and applying both methods, we compare their performance in terms of coefficient estimation and prediction accuracy.

II. RELATED WORKS

Multicollinearity is a common issue in linear regression analysis, and it occurs when two or more independent variables in a model are highly correlated. This can lead to unstable coefficient estimates, inflated standard errors, and difficulty interpreting the individual effects of the correlated variables. The presence of multicollinearity can affect the accuracy and reliability of the regression model's predictions and inferences. Ordinary Least Square (OLS) regression is a widely used statistical technique for estimating the relationships between dependent and independent variables. It minimizes the sum of the squared differences between the observed and predicted values of the dependent variable. While OLS regression is simple and effective in many cases, it can be sensitive to multicollinearity.

Studies have shown that the presence of multicollinearity can lead to biased and inefficient estimates in OLS regression. For instance, [13] stated in their study that when multicollinearity exists among the independent variables, the OLS estimates can be highly variable and unstable, leading to unreliable predictions. Similarly, [14, 15] emphasized that multicollinearity can cause standard errors to be inflated, making it challenging to determine the significance of individual coefficients accurately. The main challenge with OLS regression in the presence of multicollinearity is the interpretation of individual variable effects. In the study OF [16], they stated that when two or more independent variables are highly correlated, their estimated coefficients in the OLS model may change significantly with small changes in the sample data. This instability makes it difficult to draw meaningful conclusions about the unique contribution of each predictor variable to the dependent variable.

To address these issues, researchers have proposed various diagnostics and remedies for multicollinearity in OLS regression. For example, [17] in their article, suggested several diagnostics, such as tolerance, variance inflation factor (VIF), and condition index, to detect multicollinearity. They also discussed common remedies, including variable selection techniques (e.g., backward elimination, forward selection), ridge regression, and principal component regression. These techniques aim to reduce the dimensionality of the data or modify the estimation process to mitigate the impact of multicollinearity.

Partial Least Squares (PLS) regression is a multivariate analysis technique that has gained popularity as an alternative to OLS regression when multicollinearity is present. Unlike OLS, which focuses on finding the best fit for the dependent variable, PLS aims to maximize the covariance between the predicted values of the dependent variable and the actual values. By doing so, PLS can handle multicollinearity more effectively. The advantage of PLS regression also lies in its ability to handle complex relationships between variables.

Comparative studies have been conducted to evaluate the performance of OLS and PLS regression in the presence of multicollinearity. These studies offer insights into the strengths and weaknesses of each method. For instance, [18] compared the two models (PLS and OLS) in order to determine which one is more robust in a study of predicting couples mental health and found that PLS provided more stable and

reliable coefficient estimates compared to OLS. Similarly, [19] compared the predictive performance of PLS, OLS, RR and PCR in a real-life data environment with multicollinearity. They concluded that PLS regression yielded more accurate predictions and was less affected by the multicollinearity issue.

The comparative studies also highlight the importance of considering the specific characteristics of the data and research objectives when choosing between OLS and PLS regression. PLS regression is particularly useful when the focus is on prediction and understanding the underlying latent structure. In contrast, OLS may still be preferred when the primary interest lies in estimating the unique effects of individual variables, even in the presence of moderate multicollinearity.

III. MATERIALS AND METHODS

A. THE PROPOSED METHODS OF WEIGHT ALLOCATION

In this study, standard errors of the estimate of the regression coefficients of simple regression model of each explanatory variable was used as the weighting scheme. Weight was allocated to each explanatory variable using:

i. Their standard errors (w_{ij} α $SE(\beta_j)$). Here, weight is defined as

$$w_{ij} = \frac{SE(\beta_j)}{\sum SE(\beta_j)} \tag{1}$$

where $SE(\beta_j)$ is the standard error of the jth the regression coefficient.

ii. The ranking of their standard errors (w_{ij} α rank of $SE(\beta_j)$). Weight is defined here as

$$w_{ij} = \frac{SE_R(\beta_j)}{\sum SE_R(\beta_j)} \tag{2}$$

where $SE_R(\beta_j)$ is the ranking of the standard error of the jth regression coefficient.

B. MONTE CARLO EXPERIMENTS AND DATA

The Linear Regression Model of single response Y, being a vector, is predicted by a set of predictive variables X_{ij}. The linear model is defined as:

$$Y_t = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_p X_{pi} \tag{3}$$

where, β_0 is the intercept term, β_j is the regression coefficients and $\varepsilon_j \sim N(0, \sigma^2)$

Monte Carlo experiments were conducted at different degree of multicollinearity and different sample sizes in order to generate data. For the Monte Carlo simulation study, the parameters of equation (3.3) were specified and fixed as β_0=0.5, β_1=1.2, β_2=0.9, β_3=0.8, β_4=2.1, β_5=1.8, β_6=3.2. The explanatory variables X_1, X_2, X_3, X_4, X_5 and X_6 are assumed to be normally distributed $X_i \sim N(0,1)$ and non-stochastic and are generated by adopting the equation given by [20, 21]:

$$X_{ij} = (1 - GM^2)^{\frac{1}{2}} E_{ij} + GM * E_{ip};\ i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, 6 \tag{4}$$

where, GM = Multicollinearity level and E_{ij} are independent standard normal pseudo-random numbers.

The error term was generated to be normally distributed with mean 0 and variance 1. The regress and, Y, was calculated using equation (3). A Monte Carlo experiment of 1000 trials was carried out with 20, 30, 50, 100 and 250, five (5) sample sizes at 0.0, 0.2, 0.4, 0.6, 0.8, 0.95, 0.96, 0.97, 0.98, 0.99, 0.999, multicollinearity levels.

C. DIFFERENT COMBINATIONS OF LATENT VARIABLES

In general, a model with p number of regressors will yield $2^p - 1$ possible number of models. In this case, in which $p=6$, the possible number of models is 63 possible models. The different combinations of latent variables were used as regressors on the dependent variable. The different combinations of latent variables are summarized in the Table 1.

Table 1: Different Combinations of the Latent Variables

Combination	Number of ways	All the Combinations
$^6C_1$	6	1, 2, 3, 4, 5,6
$^6C_2$	15	12, 13, 14, 15, 16, 23, 24, 25, 26, 34, 35, 36, 45, 46, 56
$^6C_3$	20	123, 124, 125, 126, 134, 135, 136, 145, 146, 156, 234, 235, 236, 245, 246, 256, 345, 346, 356, 456
$^6C_4$	15	1234, 1235, 1236, 1245, 1246, 1256, 1345, 1346, 1356, 1456, 2345, 2346, 2356, 2456, 3456
$^6C_5$	6	12345, 12346, 12356, 12456, 13456, 23456
$^6C_6$	1	123456
Total	63	

The different combinations of latent variables were extracted and the Monte Carlo experiment conducted.

D. EXTRACTION OF THE LATENT VARIABLES AND ESTIMATION OF THE COEFFICIENTS OF THE REGRESSION MODEL

In univariate Partial Least Squares, Y, a vector, and a set of X variables are considered. This is the typical multiple regression setting. Attempt is made to find model of the form

$$y = \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_3 + \ldots + \beta_p T_p \quad (5)$$

where T_p, the pth latent variable, is a linear combination of the $X's$.

All variables are assumed to be centred to have mean 0. This means that the intercept terms will always be zero. Here is the algorithm for determining the $T's$:

1. Regress $(Y - \overline{Y})$ on each $(X_i - \overline{X})$ in turn to get b_{1i}.
2. Form each latent variable T_i, using the weights, the coefficients and the explanatory variables:

$$T_t = \sum_{i=1} w_{1i} b_{1i} X_{1i} \quad (6)$$

3. Regress $(Y - \overline{Y})$ on T_i and each $(X_i - \overline{X})$ on T_i. The residuals from these regressions have the effect of T_i removed.
4. Replace $(Y - \overline{Y})$ and each $(X_i - \overline{X})$ by the residuals of each corresponding regression.
5. Go back to step one updating the index.
6. Continue in this manner till all the latent variables are derived.

The dependent variable was then regressed on the extracted latent variables and only those which satisfied each model selection criterion were used to compute the dependent variable (y). Thus, for the 63 regression models (Y on the different combinations of latent variables), the one which satisfied each model selection criterion was chosen to estimate the model parameter. This was done for the 1000 replications. This was also incorporated into the program written using TSP software. The PLS regression coefficients corresponding to that of OLS were obtained using:

$$\beta_{PLS} = (X'X)^{-1} X'y \quad (7)$$

E. PROCEDURES FOR GETTING THE PERFORMANCE METRIC

Mean Square Error (MSE) was used as criterion for comparison:

$$MSE(\beta) = \frac{1}{R} \sum_{i=1}^{6} \sum_{j=1}^{R} (\beta_{ij} - \beta_i)^2 \quad (8)$$

The following steps were followed in order to get the Mean Square Errors (MSEs) for PLS the methods and OLS estimators:

a) Regression analysis was run on all the 63 different combinations of latent variables.
b) Model with the minimum AIC, BIC or SER or the one with the maximum ARS was identified.
c) ŷ was predicted using the above identified model.
d) 4. PLS estimate was calculated using $\hat{\beta}_{PLS} = (X'X)^{-1}X'\hat{y}$ to get $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6$
e) The MSE was calculated using MSE = $\sum_{i=1}^{6} (\hat{\beta}_{ij} - \beta_i)^2$
f) This was repeated 999 times to make 1000 repeatitions.
g) The average was taken to get MSE = $\frac{1}{R} \sum_{i=1}^{6} \sum_{j=1}^{R} (\hat{\beta}_{ij} - \beta_i)^2$

The Mean Square Errors (MSEs) for the methods and OLS estimator were compared to identify the minimum MSEs.

IV. RESULTS AND DISCUSSION

A. WEIGHT PROPORTION TO STANDARD ERROR OF REGRESSION COEFFICIENTS

The Mean Square Errors based on PLS using different methods of model selection and that of OLS estimator at different levels of multicollinearity and sample size using weight proportion to standard errors of regression coefficients are presented in Table 2 below:

Table 2: **Results of PLS and OLS Estimators Using Weight Proportional to Standard Errors of Regression Coefficients**

	ESTIMATOR		MULTICOLLINEARITY										
N			0.0	0.2	0.4	0.6	0.8	0.95	0.96	0.97	0.98	0.99	0.999
	OLS		0.541	0.569	0.67	0.918	1.75	7.344	9.298	12.606	19.344	40.065	430.826
		AIC	0.452	0.644	0.687	0.95	2.276	8.112	9.163	10.82	14.28	25.095	250.196
20	PLS	BIC	0.44	0.785	0.703	1.022	2.855	8.344	9.097	10.202	12.77	22.026	202.65
		ARS	0.471	0.614	0.684	0.921	2.021	7.856	9.239	11.624	16.28	30.568	308.416
		SER	0.471	0.614	0.684	0.921	2.021	7.856	9.239	11.624	16.28	30.568	308.416
	OLS		0.333	0.342	0.4	0.55	1.062	4.58	5.82	7.925	12.229	25.523	278.378
		AIC	0.384	0.382	0.422	0.595	1.442	4.805	6.949	7.944	9.952	20.401	154.515
30	PLS	BIC	0.389	0.474	0.45	0.661	1.921	6.931	7.356	7.976	9.06	18.524	120.172
		ARS	0.375	0.376	0.422	0.567	1.232	4.528	6.419	8.083	10.865	21.845	193.436
		SER	0.375	0.376	0.422	0.567	1.232	4.528	6.419	8.083	10.865	21.845	193.436
	OLS		0.155	0.155	0.178	0.238	0.447	1.877	2.38	3.233	4.976	10.357	112.541
		AIC	0.157	0.191	0.195	0.255	0.523	2.34	3.702	4.61	6.068	10.706	77.786
50	PLS	BIC	0.231	0.199	0.212	0.297	0.635	2.856	4.681	5.547	6.539	9.909	66.538
		ARS	0.155	0.164	0.18	0.247	0.486	2.08	3.304	4.232	5.789	11.166	85.582
		SER	0.155	0.164	0.18	0.247	0.486	2.08	3.304	4.232	5.789	11.166	85.582
	OLS		0.071	0.072	0.083	0.109	0.202	0.828	1.047	1.418	2.174	4.502	48.484
		AIC	0.074	0.074	0.09	0.125	0.219	1.013	1.45	2.05	3.285	3.508	33.327
100	PLS	BIC	0.109	0.102	0.1	0.125	0.255	1.45	2.032	3.005	4.081	3.941	24.846
		ARS	0.073	0.073	0.085	0.125	0.212	0.901	1.255	1.743	2.704	3.698	38.113
		SER	0.073	0.073	0.085	0.125	0.212	0.901	1.255	1.743	2.704	3.698	38.113
	OLS		0.029	0.029	0.033	0.044	0.081	0.332	0.42	0.568	0.871	1.802	19.365
250		AIC	0.032	0.035	0.036	0.047	0.09	0.399	0.514	0.643	0.933	2.189	13.993
	PLS	BIC	0.036	0.036	0.036	0.047	0.095	0.408	0.619	0.835	1.723	2.088	12.564
		ARS	0.029	0.028	0.035	0.045	0.085	0.384	0.459	0.618	0.848	2.041	14.904
		SER	0.029	0.028	0.035	0.045	0.085	0.384	0.459	0.618	0.848	2.041	14.904

In Table 2, weights are assigned based on the standard errors of regression coefficients across different levels of multicollinearity and sample sizes. The metrics, AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), ARS (Adjusted R-Squared) and SER (Standard Error of Regression), when analyzed across different multicollinearity levels, demonstrate the stability and performance of PLS relative to OLS. The consistent reduction in AIC, BIC, and SER as sample size increases suggests that PLS can better manage the bias-variance trade-off under multicollinearity compared to OLS. This detailed analysis highlights the suitability of PLS in addressing multicollinearity.

Table 3: Frequency of PLS and OLS Estimators Using Weight Proportional to Standard Errors of Regression Coefficients

MULTICOLLINEARITY	SAMPLE SIZE					
	SMALL		MEDIUM		LARGE	
	Method	Frequency	Method	Frequency	Method	Frequency
LOW	BIC	1	OLS	5	OLS	5
	OLS	2	SER/ARS	1	SER/ARS	1
MODERATE	OLS	2	OLS	4	OLS	4
	OLS	1	SER/ARS	1	OLS	7
HIGH	BIC	3	OLS	6	SER/ARS	1
			BIC	1		
	BIC	2	BIC	4	AIC	1
VERY HIGH					OLS	1
					BIC	2

From Table 3, it can be seen that when sample size is small, OLS is preferred at low and moderate multicollinearity while PLS with BIC is preferred at high and very high multicollinearity. Also, at medium and large sample sizes, the OLS estimator is preferred except at very high multicollinearity. At these instances, PLS with BIC provides minimum MSE.

B. WEIGHT PROPORTION TO RANKING OF STANDARD ERROR OF REGRESSION COEFFICIENT

The Mean Square Errors based on PLS using different methods of model selection and that of OLS estimator at different levels of multicollinearity and sample size using weight proportion to the ranking of the standard errors of regression coefficients are presented in Table 3 below:

Table 4: Results of PLS and OLS Estimators Using Weight Proportional to Ranking of Standard Error of Regression Coefficient

n	ESTIMATOR		MULTICOLLINEARITY										
			0.000	0.200	0.400	0.600	0.800	0.950	0.960	0.970	0.980	0.990	0.999
	OLS		0.541	0.569	0.670	0.918	1.750	7.344	9.298	12.606	19.344	40.065	430.826
		AIC	0.506	0.447	0.634	0.692	1.435	5.512	6.394	8.927	13.141	26.239	273.564
20	PLS	BIC	0.509	0.435	0.675	0.628	1.337	5.265	6.394	8.114	11.371	21.637	211.120
		ARS	0.500	0.483	0.625	0.744	1.523	5.948	7.483	10.012	15.334	30.634	327.710
		SER	0.500	0.483	0.625	0.744	1.523	5.948	7.483	10.012	15.334	30.634	327.710
	OLS		0.333	0.342	0.400	0.550	1.062	4.580	5.820	7.925	12.229	25.523	278.378
		AIC	0.337	0.352	0.415	0.650	1.227	5.943	7.588	10.240	15.046	28.153	241.767
30	PLS	BIC	0.334	0.410	0.422	0.723	1.317	6.786	8.411	11.049	15.649	27.363	197.600
		ARS	0.329	0.342	0.414	0.611	1.181	5.438	6.975	9.525	14.422	27.816	257.918
		SER	0.329	0.342	0.414	0.611	1.181	5.438	6.975	9.525	14.422	27.816	257.918
	OLS		0.155	0.155	0.178	0.238	0.447	1.877	2.380	3.233	4.976	10.357	112.541
		AIC	0.157	0.167	0.184	0.261	0.489	2.331	2.984	4.000	5.928	11.161	89.213
50	PLS	BIC	0.165	0.185	0.265	0.275	0.605	3.207	3.990	5.204	7.004	11.540	71.448
		ARS	0.156	0.162	0.179	0.241	0.472	2.108	2.644	3.536	5.347	10.432	97.997
		SER	0.156	0.162	0.179	0.241	0.472	2.108	2.644	3.536	5.347	10.432	97.997
	OLS		0.071	0.072	0.082	0.109	0.202	0.828	1.047	1.418	2.174	4.502	48.484

		AIC	0.077	0.072	0.086	0.110	0.198	0.970	1.295	1.840	2.905	5.762	39.114
100	PLS	BIC	0.078	0.090	0.100	0.148	0.220	1.519	1.987	2.739	3.997	6.799	26.028
		ARS	0.076	0.072	0.082	0.109	0.198	0.875	1.131	1.577	2.504	5.072	45.012
		SER	0.076	0.072	0.082	0.109	0.198	0.875	1.131	1.577	2.504	5.072	45.012
	OLS		0.029	0.029	0.033	0.044	0.081	0.332	0.420	0.568	0.871	1.802	19.365
		AIC	0.033	0.033	0.035	0.044	0.083	0.379	0.475	0.645	1.015	2.212	18.161
250	PLS	BIC	0.038	0.034	0.034	0.044	0.104	0.462	0.558	0.743	1.234	3.039	13.910
		ARS	0.029	0.031	0.036	0.045	0.080	0.390	0.482	0.624	0.946	2.061	19.183
		SER	0.029	0.031	0.036	0.045	0.080	0.390	0.482	0.624	0.946	2.061	19.183

Table 4 displays the performance of OLS and PLS estimators when weights are assigned based on the ranking of the standard errors of regression coefficients across different levels of multicollinearity and sample sizes. As shown, the Mean Squared Error (MSE) generally decreases with larger sample sizes and higher multicollinearity levels, indicating improved estimation stability. In addition, PLS consistently yields lower AIC, BIC, and SER values across most multicollinearity levels compared to OLS, particularly in small to medium sample sizes (n=20 and n=30). This suggests that weighting by standard error ranking benefits PLS more than OLS, as it allows PLS to maintain a more parsimonious model and a tighter error distribution under high multicollinearity conditions.

At larger sample sizes (n=100 and n=250), the differences in performance between OLS and PLS estimators diminish, with both achieving comparably low MSE and stable AIC, BIC, and SER values. This observation further supports the asymptotic equivalence between OLS and PLS, as previously noted, but also highlights that PLS remains more robust to multicollinearity at smaller sample sizes when weights are applied in proportion to the ranking of standard errors.

Table 5: Frequency of PLS and OLS Estimators Using Weight Proportional to Ranking of Standard Errors of Regression Coefficients

MULTICOLLINEARITY	**SAMPLE SIZE**					
	SMALL		MEDIUM		LARGE	
	M	F	M	F	M	F
	SER/ARS	2	SER/ARS	2	OLS	6
LOW	BIC	1	OLS	4		
	BIC	2	OLS	4	SER/ARS	3
MODERATE					BIC	1
HIGH	BIC	4	OLS	8	BIC	8
	BIC	2	BIC	3	OLS	2
VERY HIGH			OLS	1	BIC	2

From Table 5, it can be observed that PLS with BIC is preferred when sample size is small. At medium and large sample sizes, OLS and PLS with BIC are preferred even though PLD with BIC performs better when multicollinearity is high and very high.

Table 6: Best Estimator using the Proposed Methods of Weight Allocation

		SAMPLE SIZE									
		20		30		50		100		250	
ρ	**W**	BEST	VALUE	BEST	VALUE	BEST	VALUE	BEST	VALUE	BEST	VALUE
	W1	BIC	**0.43953**	OLS	0.33313	SER	**0.15451**	OLS	**0.071069**	OLS	**0.02852**
0.0	W2	SER	0.50004	SER	**0.32865**	OLS	0.155	OLS	0.071069	OLS	0.02852
	W1	OLS	0.56873	OLS	0.34173	OLS	**0.15539**	OLS	**0.071973**	SER	**0.028375**
0.2	W2	BIC	**0.43548**	SER	**0.34172**	OLS	0.15539	OLS	0.071973	OLS	0.028893
	W1	OLS	0.66993	OLS	**0.39984**	OLS	**0.17751**	OLS	0.08251	OLS	**0.032988**

0.4	W2	SER	**0.62542**	OLS	0.39984	OLS	0.17751	OLS	0.08215	OLS	0.032988
	W1	OLS	0.9179	OLS	**0.5496**	OLS	**0.23787**	OLS	0.10919	OLS	0.043868
0.6	W2	BIC	**0.62809**	OLS	0.5496	OLS	0.23787	SER	**0.10862**	BIC	**0.04382**
	W1	OLS	1.74976	OLS	**1.06182**	OLS	**0.44659**	OLS	0.20174	OLS	0.081069
0.8	W2	BIC	**1.33694**	OLS	1.06182	OLS	0.44659	SER	**0.19795**	SER	**0.080218**
	W1	OLS	7.34367	SER	**4.52769**	OLS	**1.87655**	OLS	**0.8279**	OLS	**0.33209**
0.95	W2	BIC	**5.26502**	OLS	4.5799	OLS	1.87655	OLS	0.8279	OLS	0.33209
	W1	BIC	9.09655	OLS	**5.82021**	OLS	**2.37969**	OLS	**1.04703**	OLS	**0.41983**
0.96	W2	BIC	**6.39407**	OLS	5.82021	OLS	2.37969	OLS	1.04703	OLS	0.41983
	W1	BIC	10.20177	OLS	**7.92516**	OLS	**3.23304**	OLS	**1.41803**	OLS	**0.56832**
0.97	W2	BIC	**8.11378**	OLS	7.92516	OLS	3.23304	OLS	1.41803	OLS	0.56832
	W1	BIC	12.79697	BIC	**9.05996**	OLS	**4.97637**	OLS	**2.17434**	SER	**0.84803**
0.98	W2	BIC	**11.37137**	OLS	12.22852	OLS	4.97637	OLS	2.17434	OLS	0.87088
	W1	BIC	22.02566	BIC	**18.5239**	BIC	**9.90897**	AIC	**3.50824**	OLS	**1.80157**
0.99	W2	BIC	**21.63709**	BIC	27.3632	OLS	10.3572	OLS	4.50225	OLS	1.80157
	W1	BIC	**202.6498**	BIC	**120.1719**	BIC	**66.5382**	BIC	**24.84613**	BIC	**12.56399**
0.999	W2	BIC	211.1197	BIC	197.6003	BIC	71.4477	BIC	26.0279	BIC	13.91008

Table 6 summarizes the best-performing estimator (either OLS or PLS) under different weighting methods (W1: Weight Proportional to Standard Error of Regression Coefficient; W2: Weight Proportional to Ranking of Standard Error of Regression Coefficient) across various sample sizes and levels of multicollinearity (ρ). The table reveals that as the sample size increases, the differences between the weighting methods diminish, with both methods frequently favoring OLS as the best estimator, especially at low to moderate multicollinearity levels (e.g., $P \leq 0.8$).

At smaller sample sizes (20 and 30), the results show a noticeable preference for PLS under the W2 scheme at higher multicollinearity levels ($P \geq 0.6$), as indicated by the best values for BIC and SER. This highlights the effectiveness of the W2 weighting approach in enhancing the stability and accuracy of PLS when the data suffers from multicollinearity and limited sample size. In contrast, at high sample sizes (100 and 250), OLS is consistently identified as the best estimator under both weighting methods across all multicollinearity levels, suggesting that OLS is more robust and efficient when the sample size is large, regardless of multicollinearity.

This comparison underscores that PLS with W2 weighting may offer advantages in smaller datasets with high multicollinearity, while OLS performs well across most conditions as sample sizes increase, supporting its asymptotic efficiency.

V. CONCLUSION

In this study, OLS and PLS (using methods of model selection and proposed methods of weight allocation) have been employed to solving the problem of multicollinearity. The Mean Square Error has been used as criterion for comparison of performances of the two estimators. It was observed that OLS estimator is preferred when sample sizes are medium and large at virtually all levels of multicollinearity except when multicollinearity is very high. At these instances, PLS with BIC model selection provides minimum MSE. When the sample size is small at virtually all levels of multicollinearity, PLS with BIC model selection provides most efficient estimates of the regression coefficients. Moreover, it was also observed that out of the two weight allocation schemes proposed, weight based on standard error of regression coefficient gave lower MSE except when sample size is small. At these instances, weight based on ranking of standard error of regression coefficient provided minimum MSE.

References

[1] Etaga, H. O., Roseline, C. N. and Ngonadi, L. O. Effect of Multicollinearity on Variable Selection in Multiple Regression. *Science Journal of Applied Mathematics and Statistics,* 9(6), 141-153, 2021. doi:10.11648/j.sjams.20210906.12.

[2] Noora, S. Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics,* 8(2), 39-42, 2020.

[3] Vatcheva, K. P., MinJae L., Joseph B. M. and Mohammad H. R.. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale)*. 6(2), 1-20, 2016.

[4] Kyriazos, T. and Poga, M. Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions. *Scientific Research Publishing,* 13, 404-424, 2023.

[5] Chan, J.Y.-L., Leow, S.M.H., Bea, K.T., Cheng, W.K, Phoong, S.W., Hong, Z.-W. and Chen, Y.-L. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics,* 10, 1-17, 2022. https://doi.org/10.3390/math10081283

[6] Sharma, M. J. and Jin, Y. S. Stepwise regression data envelopment analysis for variable reduction. *Applied Mathematics and Computation,* 253, 126-134, 2015.

[7] Cheng, J., Jun S., Kunshan Y., Min X. and Yan C. A variable selection method based on mutual information and variance inflation factor, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 268, 2022, 120652, https://doi.org/10.1016/j.saa.2021.120652.

[8] Liu, C., Zhang, X., Nguyen, T. T, Jinyuan, L., Tsungchin, W., Ellen, L. and Xin, M. T. Partial least squares regression and principal component analysis: similarity and differences between two popular variable reduction approaches. *General Psychiatry*, 35, 1-5, 2022: e100662. doi:10.1136/gpsych-2021-100662.

[9] Johnson, R. A. and Wichern, D. W. Applied Multivariate Statistical Analysis, 7th ed., Pearson,2014.

[10] Gogoi, C. J. and Gogoi, B. OLS, M, S and Adaptive Estimation in Linear Regression. *International Journal of Management (IJM),* 11(5), 602-608, 2020.

[11] Korkmazoglu, O. B. and Kemalbay, G. Econometrics Application of Partial Least Squares Regression: An Endogeneous Growth Model for Turkey. *Procedia - Social and Behavioral Sciences,* 62, 906-910, 2012

[12] Rosipal, R., and Krämer, N. Overview and Recent Advances in Partial Least Squares. 34 – 51, 2006.

[13] Adewoye, K. B, Ayinla, B. R., Aminu, T. F. and Onikola, I. O. Investigating the Impact of Multicollinearity on Linear Regression Estimates. *Malaysian Journal of Computing,* 6 (1), 698 -714, 2021.

[14] K. Ayinde, Lukman, A. F. and Arowolo, O. T. Combined Parameters Estimation Methods of Linear Regression Model with Multicollinearity and Autocorrelation. *Journal of Asian Scientific* Research, 5(5): 243-250, 2015.

[15] Adeboye, N. O, Fagoyinbo, I. S. and Olatayo, T. O.. Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients. *Journal of Mathematics (IOSR-JM),* 10(3), 1-5, 2014.

[16] Gregorich, M.; Strohmaier, S.; Dunkler, D.; Heinze, G. Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution. *Int. J. Environ. Res. Public Health*, 18, 1-12,2021. https://doi.org/10.3390/ijerph18084259

[17] Dormànn, C. F. Collinearity: A Review of Methods to deal with it and a Simulation Study Evaluating Their Performance. *Ecography,* 36, 27–46, 2013.

[18] Farahania, H. A., Rahiminezhad, A,. Laleh, S. and Kobra, I. A Comparison of Partial Least Squares (PLS) and Ordinary Least Squares (OLS) regressions in predicting of couples mental health based on their communicational patterns. *Procedia Social and Behavioral Sciences*, 5, 1459–1463, 2010.

[19] Yeniay, O. and Atilla, G. A Comparison of Partial Least Squares Regression with Other Prediction Methods. *Hacettepe Journal of Mathematics and Statistics,* 31, 99–111, 2002.

[20] Gibbons, D. G. A Simulation Study of Some Ridge Estimators. *Journal of the American Statistical Association* 76, 131–139, 1981.

[21] Muniz, G. and Kibria, B. M. G. On Some Ridge Regression Estimators: An Empirical Comparisons. *Communications in Statistics - Simulation and Computation,* 38, 621–630, 2009.