Tech-Sphere Journal of Pure and Applied Sciences (TSJPAS)
A Subsidiary of Tech-Sphere Multidisciplinary International Journal (TSMIJ)
Aladesuyi et al. Vol 2, Issue 1, 2025 Publication Edition

Assessing The Role of Significant Roots in Parameter Estimation of Linear Regression Models Under Multicollinearity

Research Article https://stem.techspherejournal.com

Author(s) Details

A. Aladesuyi*[1], K. Ayinde[2] and T. S. Fayose[3]

[1, 3]Department of Statistics, Federal Polytechnic, Ado – Ekiti, 360101, Nigeria
[2]Department of Mathematics and Statistics, Northwest Missouri State University, Maryville, Missouri, USA.

https://doi.org/10.5281/zenodo.15470100

****Corresponding author's email**: aladesuyialademomi@gmail.com

ABSTRACT

Linear regression is a widely used method for modelling relationships between an effect variable and one or more concomitant variables. However, multicollinearity (high correlation among concomitant variables) undermines the stability and accuracy of parameter estimates, often reflected in the small eigenvalues (significant roots) of the design matrix. This study investigates the impact of significant roots on parameter estimation and explores Partial Least Squares (PLS) regression as a remedy. PLS addresses multicollinearity by constructing latent variables that maximize covariance with the response offering improved prediction and stability. Through simulation study across various sample sizes (20 – 250) and eleven multicollinearity levels (0 – 0.999), the performance of PLS and Ordinary Least Squares (OLS) estimators is compared using Mean Squared Error (MSE). Results show that OLS performs best at low multicollinearity, while PLS, particularly at 1% significance level, excels when multicollinearity is high. Findings highlight the diagnostic role of eigenvalues and confirm PLS as a robust alternative for achieving efficient regression estimates under multicollinearity thereby validating that PLS regression provides more stable and accurate parameter estimates than OLS in the presence of high multicollinearity.

Keywords: Linear Regression, Partial Least Squares, Latent Variables, Multicollinearity.

1 Introduction

Linear regression remains one of the most widely used statistical methods to model relationship between an effect variable and one or more concomitant variables. Its simplicity, interpretability and versatility make it essential in numerous scientific disciplines including economics, engineering, social sciences and biological sciences. However, the presence of multicollinearity poses significant challenges in linear regression modelling. Multicollinearity is a situation where the concomitant variables are highly linearly correlated (Fayose and Ayinde, 2019). The presence of multicollinearity causes the regression parameters estimates to be misleading (Fayose *et al.,* 2023a; Fayose *et al.,* 2023b). One of the less explored yet critical aspect in this context is the effect of significant roots on the stability and accuracy of parameter estimates. Significant roots are the eigenvalues of the design matrix or its correlation matrix. Multicollinearity leads to near – linear dependencies among concomitant variables inflating the variance of estimated coefficients and making them sensitive to small changes in the data.

This instability is often characterized through the condition number or by examining the roots of the matrix involved in the normal equations of linear regression. Significant or small roots (eigenvalues close to zero) indicate ill – conditioning which can render least squares estimates unreliable. Consequently, understanding how these roots affect parameter estimation is essential for developing robust regression models and mitigation techniques such as Ridge Regression (RR), Principal Component Regression (PCR) and Partial Least Squares (PLS).

Partial least squares method is a dimension – reduction method that constructs latent variables (components) from the original concomitant variables in such a way that these components maximize the covariance between concomitant variables and the effect variable. PLS integrates response information into component construction making it particularly effective in addressing multicollinearity and improving prediction accuracy.

This paper explores the dual focus of understanding how significant roots affect parameter estimates under multicollinearity and how PLS was employed as a robust remedy.

Belsley *et al.,* (1980) provided one of the earliest frameworks for diagnosing multicollinearity using condition indices which are directly related to the ratio of the largest to smallest eigenvalues (roots) of the concomitant variable's matrix. They linked multicollinearity to the eigen – structure of the concomitant variable's matrix using condition number as a practical diagnostic. A small eigenvalue typically implies that concomitant variables lie nearly in a lower – dimensional subspace undermining parameter stability.

Alqaralleh and Abdullah (2021) conducted a simulation – based study demonstrating how small eigenvalues directly leads to inflated variance in OLS estimates even when the overall fit of the model remains high. Khan *et al.,* (2022) utilized eigenvalues analysis to classify the severity of multicollinearity suggesting that root magnitudes offer better diagnostic precision than traditional variance inflation factors (VIF). Liu *et al.,* (2020) compared and explored the performance of ridge regression, principal component regression and partial least squares methods, compared these methods and highlighted the limitations of PCR in capturing response – oriented structure, arguing for the use of techniques like PLS that better preserve model predictiveness under multicollinearity. Chen *et al.,* (2020) applied PLS to forecast macroeconomic variables using highly correlated indicators and concluded that PLS provided better forecasting accuracy than OLS and PCR particularly when multicollinearity was severe. Meanwhile, Duran and Mert (2021) provided an evaluation of multicollinearity diagnostics and proposed a root – based index that correlates well with PLS performance suggesting that eigenvalue analysis can guide the optimal number of PLS components to retain.

Zhang and Sun (2022) analytically showed that PLS components can stabilize estimation by avoiding predictor directions associated with near – zero eigenvalues. This aligns with the findings of Wang *et al.,* (2023) that showed through simulation that PLS substantially reduces parameter variance in the presence of multicollinearity without sacrificing interpretability.

Finally, Nadarajah and Krishna (2023) explored adaptive versions of PLS and demonstrated that weighting concomitant variables by root magnitude further improved prediction performance while Alshamrani *et al.,* (2024) applied PLS in a high – dimensional setting and demonstrated its superiority in cases where small roots dominate the concomitant variables space confirming its robustness even in modern data environments. The findings of these various authors collectively demonstrate that significant roots not only serve as indicators of multicollinearity but also guide the implementation of effective remedies. It is therefore noted that this paper quantifies eigenvalue impact on regression stability and introduces weighted PLS using standard error to enhance multicollinearity resilience in a linear regression model.

Section 2 contains the methodology. Section 3 contains simulation studies and design; Section 4 contains simulation results and Section 5 contains conclusion.

2 Methodology

2.1 The NIPALS Algorithm and Bilinear Decomposition

The NIPALS algorithm first employed by Wold (1966) is an iterative method used to compute the latent variables otherwise called scores and loadings in PLS especially when missing values occurred in a dataset or is too large to handle in a single matrix decomposition. Expanded work was done by Wold (1975).

Consider the standard regression model:

$$Y = X\beta + U, \qquad (1)$$

where is an $n \times p$ matrix with full rank, is a $n \times 1$ vector of effect variable,
The general NIPALS algorithm for Partial Least Squares is by bilinear decomposition as follows:
Standardize the predictor matrix X and effect variable matrix Y to have zero mean and unit variance. The PLS model is defined as:

$$Y = TQ + E \qquad (2)$$

$$X = t_1 p_1' + \ldots + t_k p_k' + F \qquad (3)$$

$$Y = u_1 q_1' + \ldots + u_k q_k' + E \qquad (4)$$

With E and F corresponding to residuals, p and q are loading matrices, t and u are latent components of x and y respectively. The decomposition in (3) and (4) is extensively analyzed by Martens and Naes (1989). Indeed the bilinear decomposition links matrix Y to matrix X using k latent vectors $t_1,\ldots,t_k$. Expressions (3) and (4) are successively used together with scores (t_k and u_k) and weight vectors (w_k and q_k) as defined in Algorithm below. The use of symbol in NIPALS algorithm are as follows.

Input: ($X_0 \leftarrow X; Y_0 \leftarrow Y$)

For k = 1,..., p, and uk a column of Y

Until convergence is met do:

Step1. $w_k \propto X_{k-1}' u_k$ C(X)

Step2. $t_k \propto X_{k-1} w_k$ R(X)

Step3. $q_k \propto Y_{k-1}' t_k$ C(Y)

Step4. $u_k \propto Y_{k-1} q_k$ R(Y)

If convergence is met, compute loading vector:

$$p_k = X_{k-1}' t_k / t_k' t_k \qquad (5)$$

Store: $T[k] \leftarrow t_k; U[k] \leftarrow u_k; P[k] \leftarrow p_k; Q[k] \leftarrow q_k$

Take as new data

$X_k = X_{k-1} - t_k p_k'$ and $Y_k = Y_{k-1} - t_k q_k'$

and go to Step 1, with X = Xk and Y = Yk

The fact that different normalizations can be used is emphasized. Normalizing for example the extracted vectors, by using the normalizing constants $w_k' w_k, t_k' t_k$ etc, allows forming the whole procedure as a sequence of simple regressions. The derived vectors then correspond to the slopes in simple univariate regressions (Bastien *et al.,* 2005). This approach permits to effectively handle missing data in a simple and direct way. The notation C and R indicate the column and the row space of matrices X and Y. Indeed, the score vectors t_k and u_k lie in R, while w_k and q_k lie in C of X and Y, respectively. This is also true for all $2 \leq k \leq p$ with data being deflated according to the above expression. The deflated data can be expressed in terms of the original data X_0 and Y_0 at any k, according to the equations below

$$X_k = (1 - \mathrm{P}_{Tk}) X_O; \tag{6}$$

$$Y_k = (1 - \mathrm{P}_{Tk}) Y_O; \tag{7}$$

With P_{Tk} denoting the projection or the hat score matrix Tk($T_k' T_K$)-1 T_k'

2.2 Method of Allocating Weights

In this paper, the standard error (SE) of the estimate of Simple Linear Regression model of each concomitant variable was used as the weighting scheme. The weights allocated to the concomitant variables using *Weight Proportional to Standard error* of regression coefficient and the weight is defined as:

$$w_{ij} = \frac{SE(\beta_j)}{\sum SE(\beta_j)} \tag{8}$$

2.3 The Linear Regression Model for Simulation Study

The linear regression model having six concomitant variables is defined as:

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon_i, \ i = 1,2,\ldots 6; \tag{9}$$

Where y_i is the response variable,
β_0, is the intercept,
β_j, j= 1...6 are the regression coefficients,
ε_i is the error term.

The concomitant variables are assumed to be fixed with the following degrees of multicollinearity (0, 0.2, 0.4, 0.6, 0.8, 0.9, 0.95, 0.97, 0.98, 0.99 and 0.999) respectively.
In this article, NIPALS iteratively extracts latent components maximizing covariance between concomitant variables and effect variable enabling efficient PLS regression under multicollinearity and missing data.

3 Simulation Studies and Design

3.1 Monte Carlo Simulation Method

Monte Carlo simulation is a computational technique that uses repeated random sampling to model and analyze the behaviour of complex systems or processes (Metropolis and Ulam, 1949). The method involves generating a large number of possible outcomes using stochastic inputs then statistically analyzing the results to approximate probabilities i.e. means, variances or mean squared errors respectively (Kroese et al., 2014). In this paper, we used Monte Carlo simulation process to generate multiple random subsets and applied PLS algorithm to each subset and analyzed the regression parameters. We evaluated the robustness of PLS under different levels of multicollinearity and sample sizes and the optimal number of latent components was selected. The approach was previously used by (Mevik and Cederkvist, 2004; Chong and Jun, 2005).

3.2 Generation of the Concomitant Variables

The potential concomitant variables X1, X2, X3, X4, X5 and X6 are normally distributed Xi ~ N(0,1) and non – stochastic and are generated following approached used by Wichern and Churchill (1978) and Gibbons (1981).

$$\left.\begin{aligned}
X_1 &= ((1-M^2)^{0.5})\,E_1+M\,(E_6)\\
X_2 &= ((1-M^2)^{0.5})\,E_2+M\,(E_6)\\
X_3 &= ((1-M^2)^{0.5})\,E_3+M\,(E_6)\\
X_4 &= ((1-M^2)^{0.5})\,E_4+M\,(E_6)\\
X_5 &= ((1-M^2)^{0.5})\,E_5+M\,(E_6)\\
X_6 &= ((1-M^2)^{0.5})\,E_6+M\,(E_6)
\end{aligned}\right\} \tag{10}$$

Where, M is Multicollinearity, $E_{1}{}_{-}E_{6}$ Error term

The error term (ε_i) was generated to be normally distributed with mean zero and variance one. Concomitant variables are generated by combining independent errors with a shared component to simulate controlled multicollinearity across varying correlation levels. This is the rationale behind equation (10).

3.3 Generation of Effect Variable

The model parameters are = 0.5, = 0.8. The sample sizes are 20, 30, 50, 100 and 250 and eleven levels of multicollinearity were used (0, 0.2, 0.4, 0.6, 0.8, 0.9, 0.95, 0.97, 0.98, 0.99 and 0.999) at each of the level of multicollinearity and sample sizes, the experiment was repeated 1000 times.

3.4 Criterion for Comparison

Mean squared error was used for comparison. The mean squared error is defined as:

$$MSE(\beta) = \frac{1}{R}\sum_{i=1}^{6}\sum_{j=1}^{R}(\hat{\beta}_{ij} - \beta_i)^2 \tag{11}$$

.

3.5 Procedure for Selection of Best Significant Levels and OLS Estimator

In order to clearly see the performances of the techniques, the multicollinearity levels were classified as low (M = 0, 0.2 and 0.4), moderate (M = 0.6 and 0.8), high (M = 0.95, 0.96 and 0.97), very high (0.98, 0.99 and 0.999); and sample size as small (20), medium (30 and 50) and large (100 and 250). The number of times each technique has the smallest MSE was counted in these classifications. Estimator with the highest count was considered best. In other words, in this paper, we categorized multicollinearity levels and sample sizes into distinct groups to assess estimator performance under varying conditions. For each combination, the number of times each estimator achieved the lowest mean squared error (MSE) was recorded. The estimator with the highest frequency of lowest MSE across all classifications was selected as the best.

3.6 Latent Variables Extraction and Estimation of Regression Coefficient of the Model

Considering univariate Partial Least Squares (PLS), that is to say l =1 so that Y is scalar. This is the typical multiple regression setting (Faraway, 1994).

Model of the form is defined: $\hat{y} = \beta_1 T_1 + \ldots + \beta_P T_P$ (12)

Where T_p is the linear combination of the X'S and β_p is the parameter.
Variables are assumed to be centered to have mean zero, this means the intercept terms will always be zero. Here is the algorithm for determining the T's.

1. Regress (Y- $\bar{Y}$)on each (Xi - $\bar{X}$) in turn to get b_{1i}.
2. Form $$T_1 = \sum_{i=1}^{m} w_{1i} b_{1i} X_{1i} \tag{13}$$

Where b_{1i} is the parameter, X_{1i} Explanatory variables and weights w_{1i} *sum* to one.

3. Regress (Y- $\bar{Y}$) on T_1 and each (Xi - $\bar{X}$) on T_1. The residuals from these regressions have the effect of T_i removed.
4. Replace (Y- $\bar{Y}$) and each (Xi - $\bar{X}$) by the residuals of each corresponding regression.
5. Go back to step one updating the index.
6. Continue in this manner till all the components or latent variables are extracted.

The effect variable was then regressed on the extracted Latent variables and those that were insignificant were omitted. Time Series Program (TSP) software was used to write the program. Only those significant were used to predict the effect variable ($\hat{y}$). The Partial Least Squares regression coefficient corresponding to that of OLS was obtained using:

$$\hat{\beta}_{PLS} = (X^1 X)^{-1} X^1 \hat{y} \tag{14}$$

The results are presented in section 4.

4 Results and Discussion

We represented the simulation results visually and in Table. The results using weight proportional to the standard error of regression coefficient were presented in the following figures:

4.1 Results Using Weight Proportional to Standard Error of Regression Coefficient

Table 1: MSE values at various levels of multicollinearity, significant levels and sample sizes when Weight was attached as Proportional to Standard error of regression coefficients

								MULTICOLLINEARITY LEVEL					
SS	**EST.**	**SL**	**0**	**0.2**	**0.4**	**0.6**	**0.8**	**0.95**	**0.96**	**0.97**	**0.98**	**0.99**	**0.999**
20	PLS	**0.1**	**0.43743**	0.98651	1.58937	5.1546	5.9022	10.40383	**8.8683**	9.80288	11.956	12.293	163.4435
		0.05	0.47701	1.33519	1.98731	6.43635	7.03931	11.05315	9.0723	9.77331	10.8101	**11.7842**	107.645
		0.01	0.73809	2.14788	3.03804	8.88254	8.97997	11.71713	9.79611	**9.76314**	**9.91102**	12.2136	**45.19548**
	OLS	**OLS**	0.54126	**0.56873**	**0.66993**	**0.9179**	**1.74976**	**7.34367**	9.29823	12.6058	19.344	40.0652	430.8261
30	PLS	**0.1**	0.62674	0.48172	1.1619	1.8451	3.31868	8.59024	7.3122	**7.91422**	9.02641	9.30529	118.6723
		0.05	0.52704	0.63179	1.35938	2.24157	4.00234	9.08943	7.62274	8.01757	8.48609	8.5094	87.10795
		0.01	0.62672	0.79004	1.9281	3.25469	5.60473	9.68669	8.39568	8.32621	**8.29985**	**8.36487**	**41.13732**
	OLS	**OLS**	**0.33313**	**0.34173**	**0.39984**	**0.5496**	**1.06182**	**4.5799**	**5.82021**	7.92516	12.2285	25.5233	278.3782
50	PLS	**0.1**	0.20159	0.49586	0.33902	0.3954	1.36225	7.93205	4.23885	5.18081	6.42801	8.80679	72.32237
		0.05	0.23848	0.50697	0.37405	0.44863	1.65546	8.70821	4.98639	5.77026	6.59101	**8.62183**	62.79803
		0.01	0.24237	0.58736	0.49743	0.57442	2.24641	10.05512	6.33425	6.83133	7.21629	8.84817	**37.4955**
	OLS	**OLS**	**0.155**	**0.15539**	**0.17751**	**0.23787**	**0.44659**	**1.87655**	**2.37969**	**3.23304**	**4.97637**	10.3573	112.5406
100	PLS	**0.1**	0.10087	0.08725	0.11793	0.13272	0.2494	6.67914	1.65655	2.39793	3.60864	7.84357	29.97051
		0.05	0.12385	0.10132	0.12317	0.14169	0.26803	7.32045	1.91052	2.81888	3.99483	7.93888	26.19008
		0.01	0.13354	0.10272	0.12218	0.17589	0.34927	8.30579	2.82915	3.66814	4.59273	7.92498	**17.48316**
	OLS	**OLS**	**0.071069**	**0.07197**	**0.08215**	**0.10919**	**0.20174**	**0.8279**	**1.04703**	**1.41803**	**2.17434**	**4.50225**	48.48393
250	PLS	**0.1**	0.076133	0.03607	0.03642	0.07124	0.58803	5.3339	0.56191	0.64705	1.05741	8.93712	13.866
		0.05	0.077653	0.03607	0.03645	0.07967	0.68682	6.61374	0.60474	0.6725	1.34704	9.05116	13.56257
		0.01	0.096788	0.03607	0.03645	0.0799	0.83485	8.59677	0.63772	1.01636	1.94199	9.29765	**11.52105**
	OLS	**OLS**	**0.028521**	**0.02889**	**0.03299**	**0.04387**	**0.081069**	**0.33209**	**0.41983**	**0.56832**	**0.87088**	**1.80157**	19.36509

SS –Sample Space, SL – Significant level, EST – Estimators

From Table 1: It can be observed that the OLS estimator still produces the minimum MSE in most cases at different levels of sample sizes and low to medium levels of multicollinearity but when the multicollinearity levels tend to unity ($\rho \to 1$) i.e. as multicollinearity level tends towards 1 while PLS method with significant latent variables provides better efficient estimates and most efficient. This further described that the results show that OLS consistently achieves the lowest Mean Squared Error (MSE) at low and moderate multicollinearity levels across all sample sizes. However, as multicollinearity increases toward unity i.e., $\rho \cong 1$, OLS performance deteriorates significantly. In high multicollinearity settings, PLS (Partial Least Squares) with lower significance levels (e.g. $\alpha = 0.01$) produces smaller MSEs especially for larger sample sizes (n = 100, 250). Thus, PLS becomes a more reliable and efficient estimator than OLS under severe or very high multicollinearity scenario.

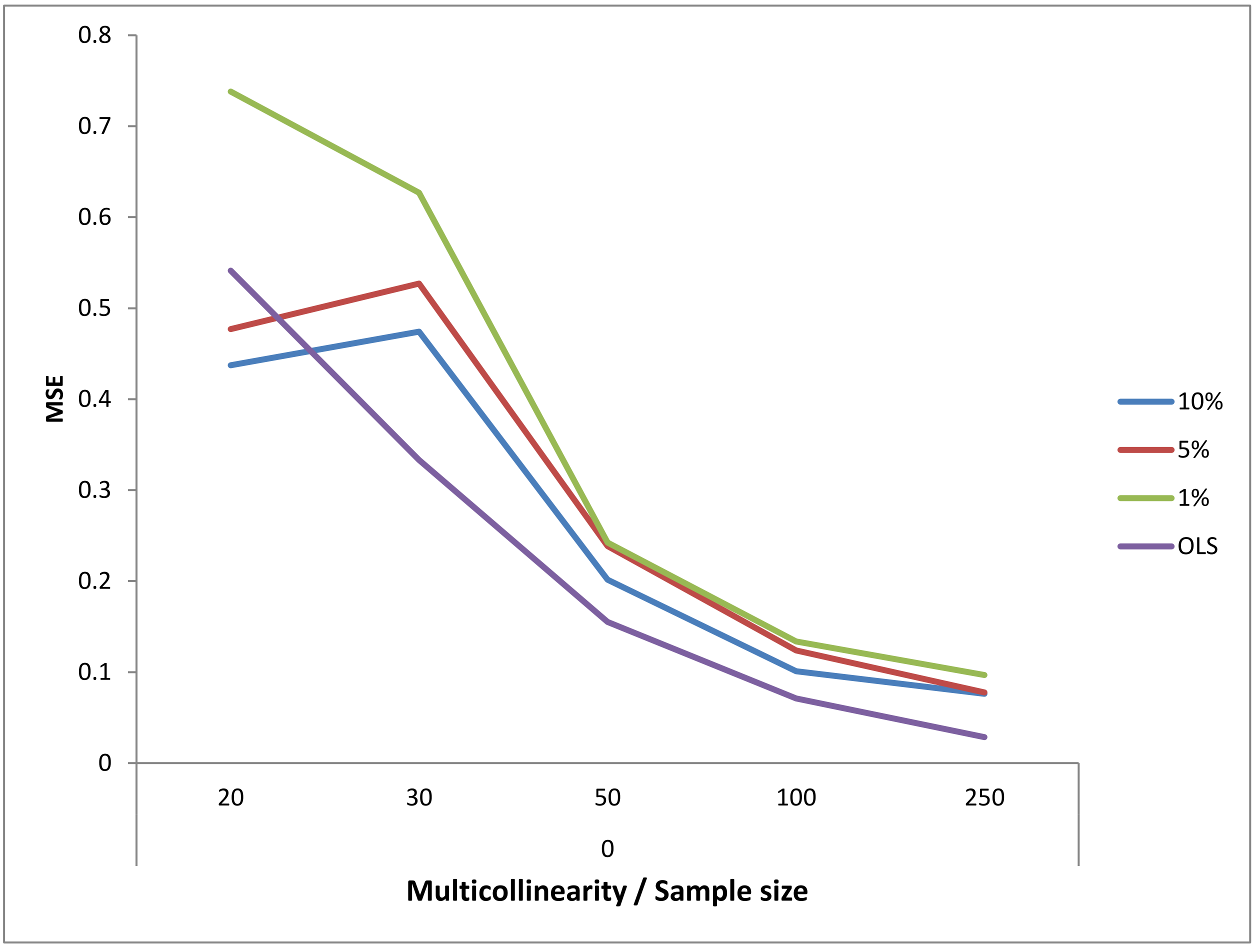

Figure 1: Line Graph showing MSE of OLS estimator and PLS with significant Latent variables at zero multicollinearity level using weight proportional to standard error of regression coefficient.

It is observed that at multicollinearity = 0, PLS with significant latent variables shows comparable or slightly higher MSE than OLS when using weights proportional to regression coefficient standard errors

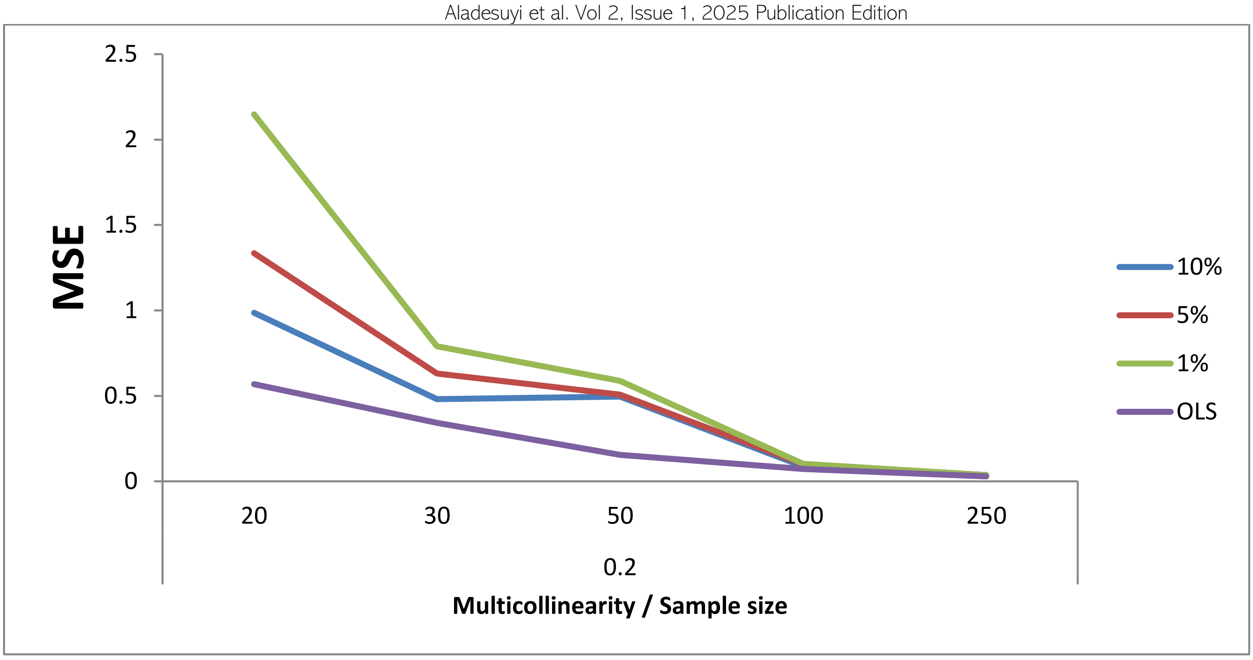

Figure 2: Line Graph showing MSE of OLS estimator and PLS with significant Latent variables at Multicollinearity = 0.2 using weight proportional to standard error of regression coefficient.

It is observed that PLS begins to show marginal MSE improvement over OLS with significant latent variables under weighted regression.

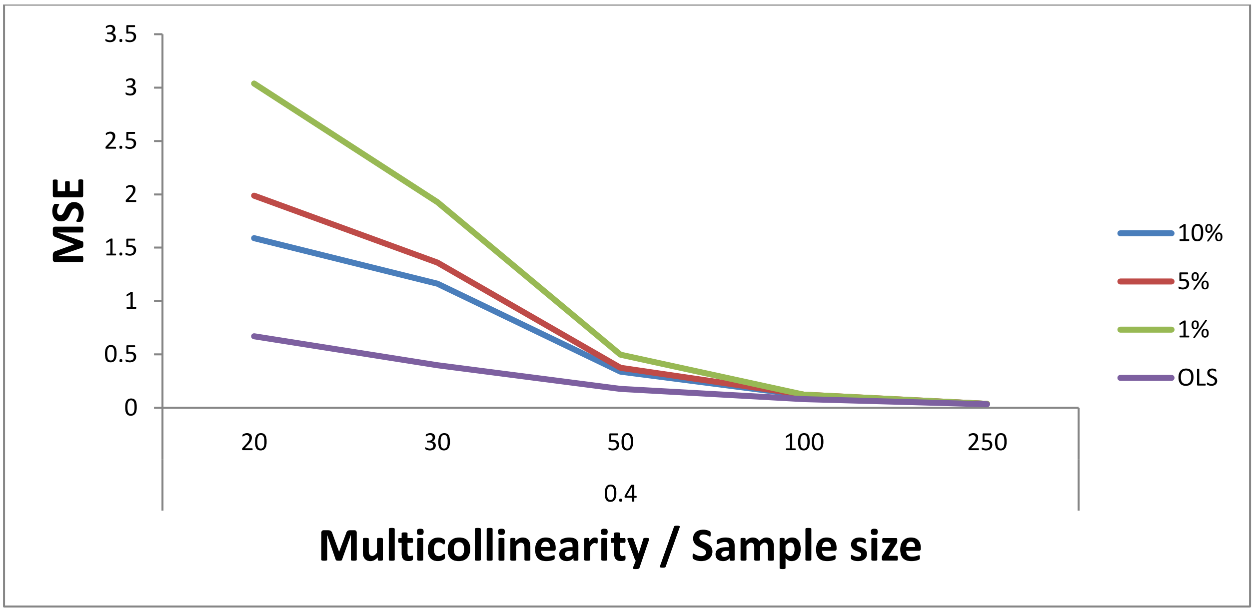

Figure 3: Line Graph showing MSE of OLS estimator and PLS with significant Latent variables at Multicollinearity = 0.4 using weight proportional to standard error of regression coefficient.

It is observed that PLS demonstrates a clearer reduction in MSE compared to OLS when incorporating significant latent variables.

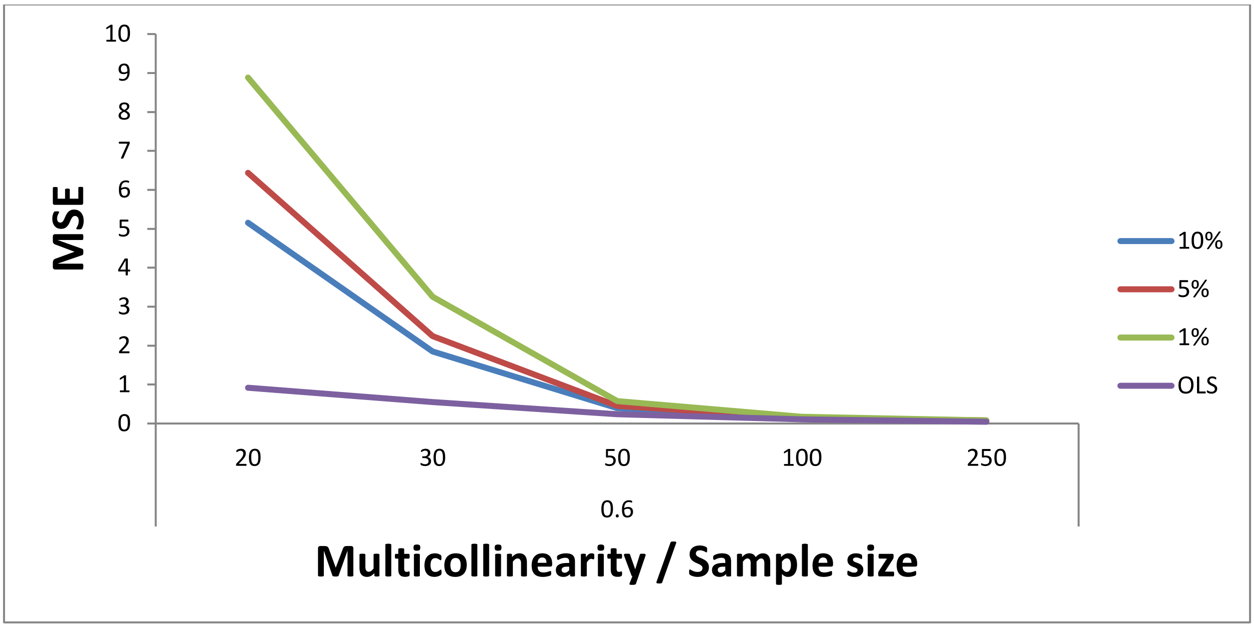

Figure 4: Line Graph showing MSE of OLS estimator and PLS with significant Latent variables at Multicollinearity = 0.6 using weight proportional to standard error of regression coefficient.

It is observed that PLS with significant latent variables yields a lower MSE than OLS, indicating better performance under moderate multicollinearity.

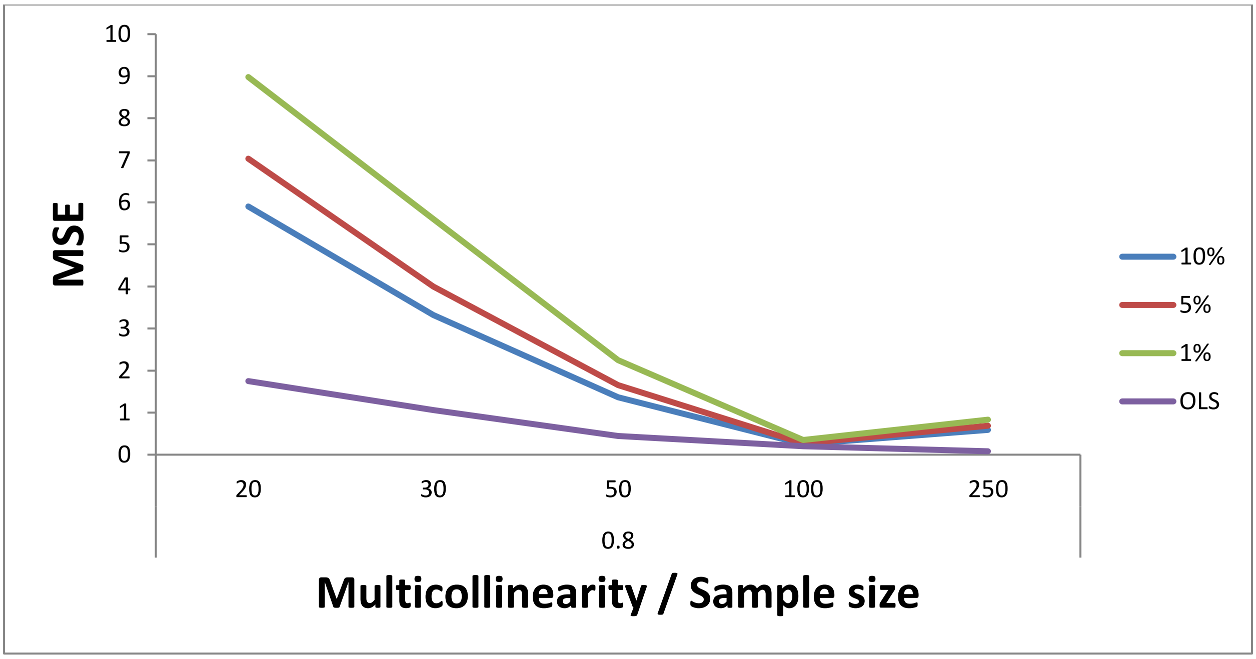

Figure 5: Line Graph showing MSE of OLS estimator and PLS with significant Latent variables at Multicollinearity = 0.8 using weight proportional to standard error of regression coefficient.

It is discovered that PLS at 1% significance level begins to produce lower MSE than OLS for larger sample sizes, indicating its growing advantage.

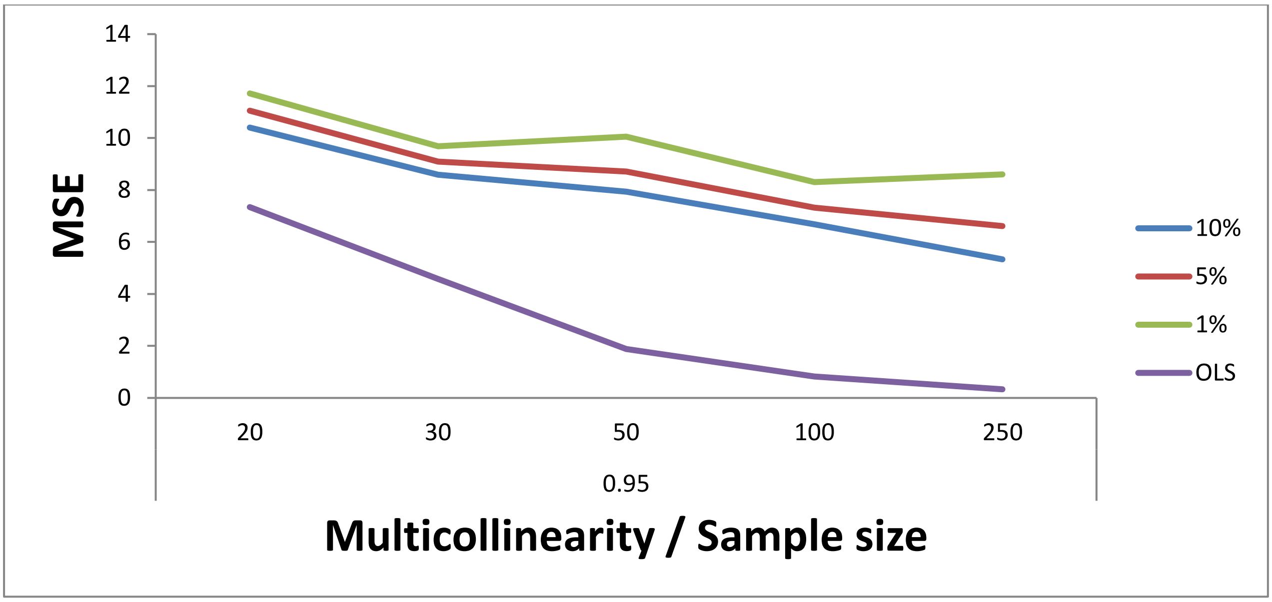

Figure 6: Line Graph showing MSE of OLS estimator and PLS with significant Latent variables at Multicollinearity = 0.95 using weight proportional to standard error of regression coefficient.

Inference indicates that PLS maintains a strong MSE advantage over OLS highlighting its robustness in high multicollinearity settings.

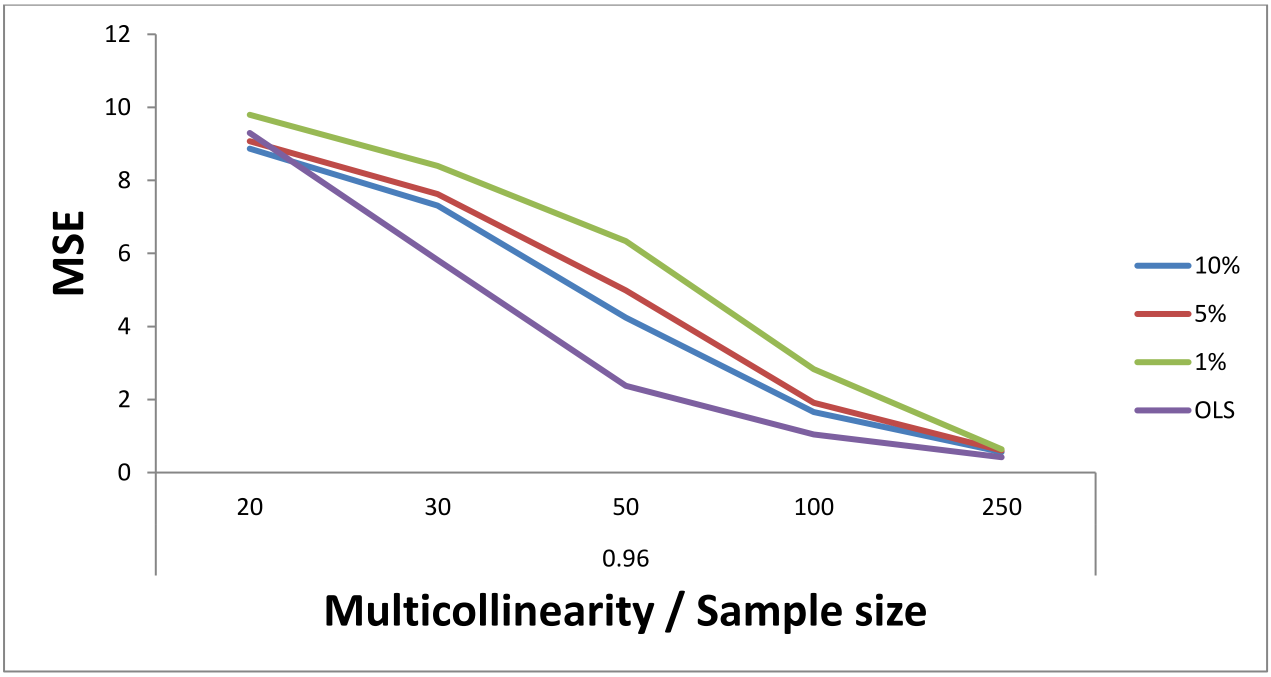

Figure 7: Line Graph showing MSE of OLS estimator and PLS with significant Latent variables at Multicollinearity = 0.96 using weight proportional to standard error of regression coefficient.

Inference shows that the MSE of OLS increases sharply, while PLS with 1% significance shows robust efficiency and lower MSE.

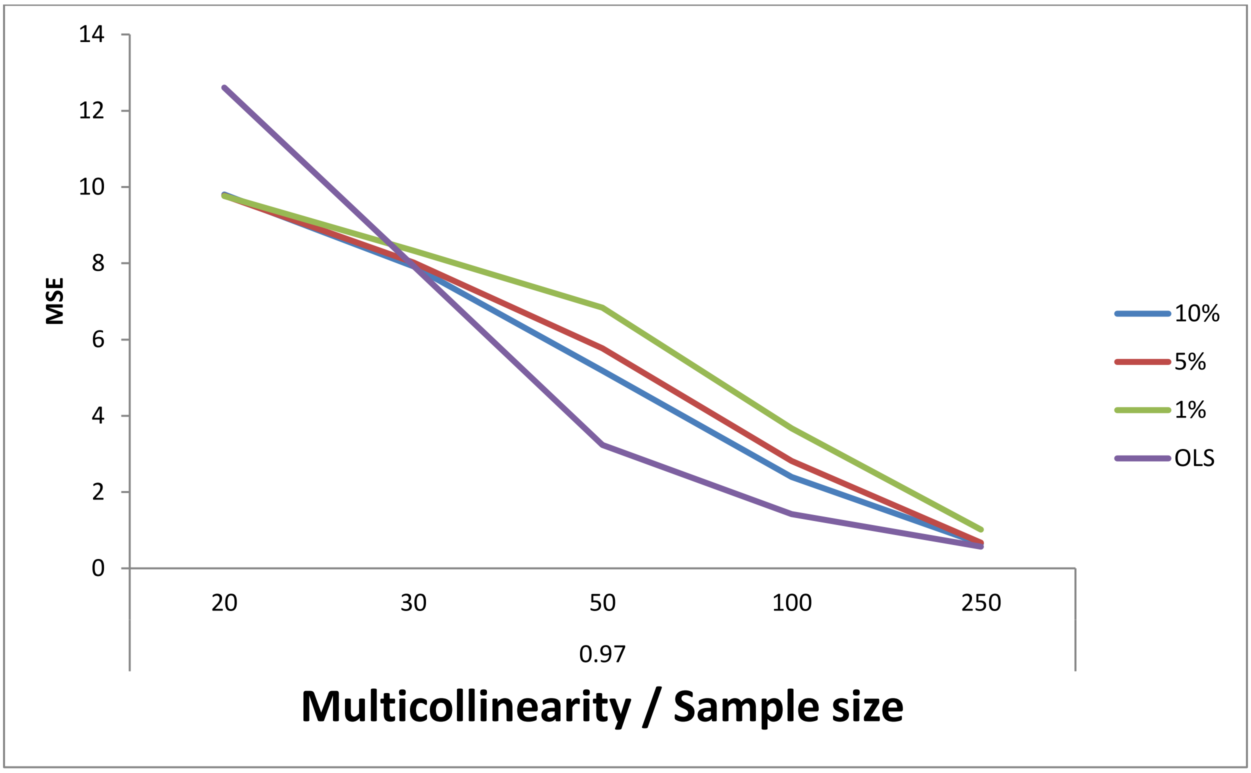

Figure 8: Line Graph showing MSE of OLS estimator and PLS with significant Latent variables at Multicollinearity = 0.97 using weight proportional to standard error of regression coefficient.

It is discovered that PLS further extends its MSE improvement over OLS, reinforcing its effectiveness in near – extreme multicollinearity.

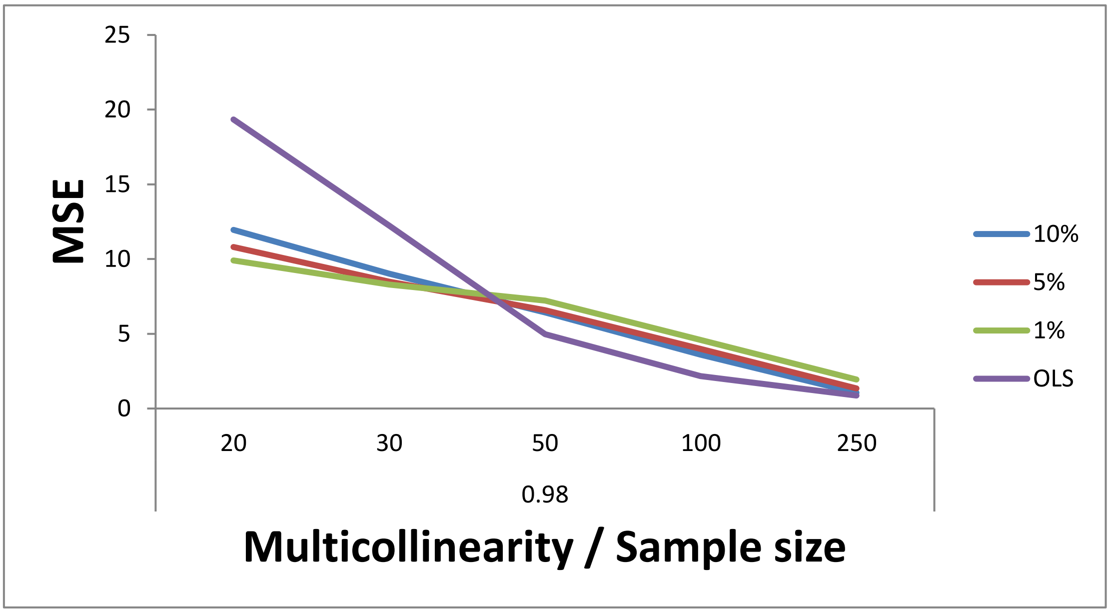

Figure 9: Line Graph showing MSE of OLS estimator and PLS with significant Latent variables at Multicollinearity = 0.98 using weight proportional to standard error of regression coefficient.

Inference suggests that PLS sustains a significantly lower MSE than OLS demonstrating resilience against very high multicollinearity.

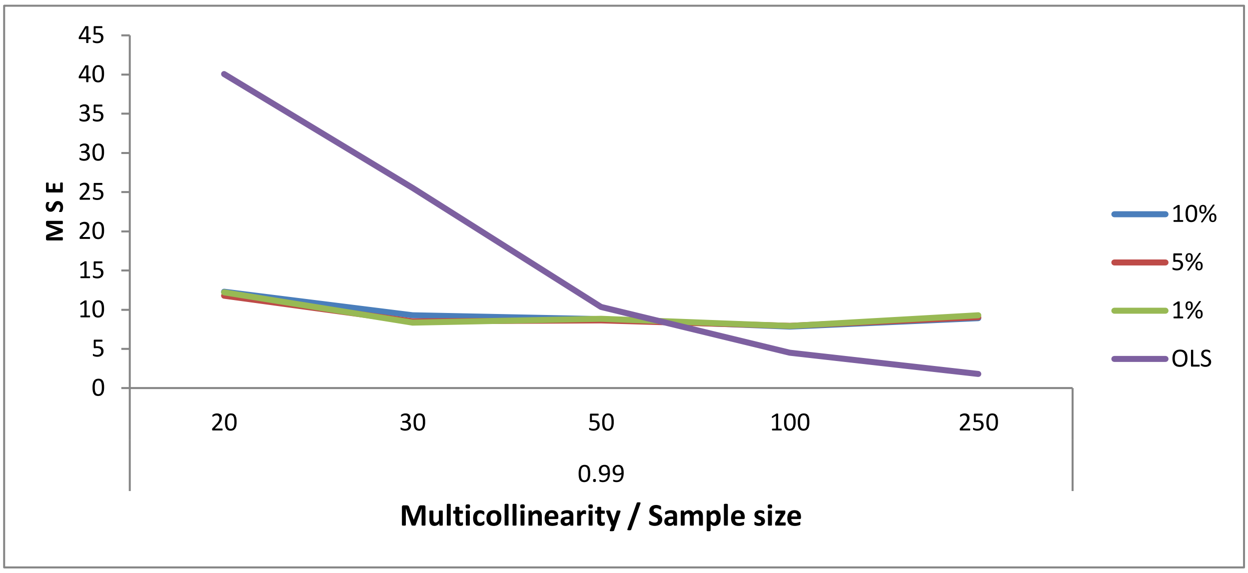

Figure 10: Line Graph showing MSE of OLS estimator and PLS with significant Latent variables at Multicollinearity = 0.99 using weight proportional to standard error of regression coefficient

It is discovered PLS becomes decisively more efficient than OLS which shows large MSE values making PLS the preferred method.

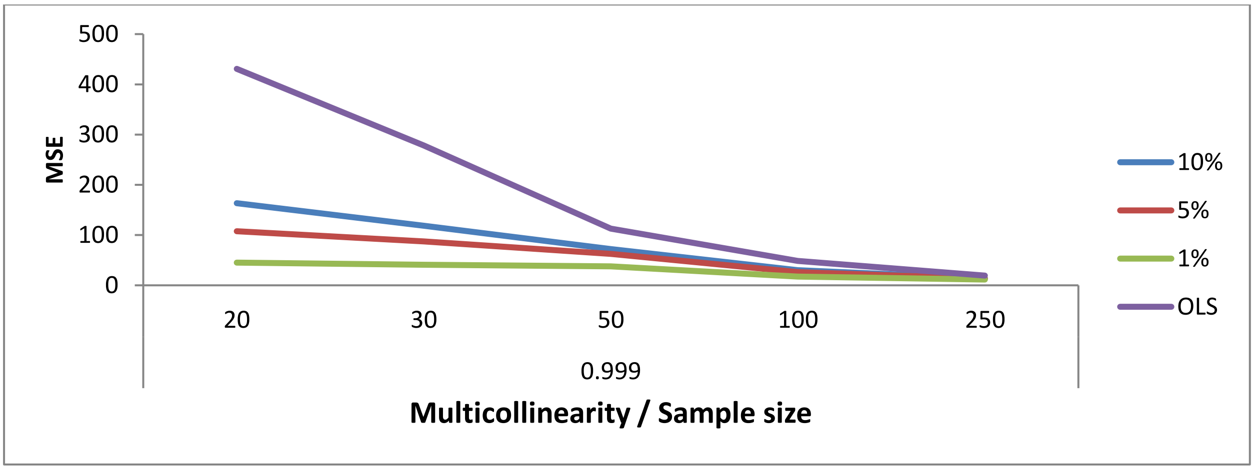

Figure 11: Line Graph showing MSE of OLS estimator and PLS with significant Latent variables at Multicollinearity = 0.999 using weight proportional to standard error of regression coefficient.

Inference suggests MSE values from OLS become extremely large and unstable while PLS at low significance levels yields the most efficient and reliable estimates.

5 Conclusion

The effect of significant latent variables on having efficient estimates of parameters of linear regression models has been examined in this paper. The PLS Estimator was used to find ways to overcome the problem of multicollinearity. The MSE has been a criterion for comparison of performances of various different significant levels and OLS Estimator. Partial least squares regression has been proven to be a very powerful versatile data analytical tool applied in many areas of research and industrial applications. One of the approach's strong points is its computational and implementation simplicity, which favours the technique as a first step to understand the existing relationships and to analyse real – world data. The practical implications of the findings are that while OLS is effective and yields the lowest MSE in low to moderate multicollinearity scenarios, it becomes unreliable as multicollinearity increases. In such high multicollinearity conditions, PLS with lower significance levels (particularly 1% or $\alpha = 0.01$) provides more accurate and efficient parameter estimates, making it a better alternative to OLS for large datasets with highly correlated concomitant variables. The following are takeaways from Figure (1) to Figure (11):

1. It is evident that the OLS estimator and various significant levels are employed when allocating weight to concomitant variables in proportion to the standard error of the regression coefficient.
2. When multicollinearity is minimal, the OLS Estimator yields the most efficient estimates across all sample size categorization. OLS performs better at low levels of multicollinearity because under such conditions, the concomitant variables are relatively uncorrelated allowing the OLS method to effectively estimate the regression coefficients. Also, at low multicollinearity, the variance of the coefficient estimates is smaller leading to more accurate and stable predictions. In these scenarios, OLS can efficiently minimize the Mean Squared Error (MSE) and provide reliable results, as it assumes that concomitant variables are not strongly interrelated, which aligns well with the assumptions of the OLS method.
3. At moderate and high multicollinearity level, OLS estimator produces best efficient across all sample sizes while 1% and 10% significant levels perform best when sample size is small.
4. The most efficient estimate was obtained across all sample sizes at 1% significance level when multicollinearity is very high or when multicollinearity tends towards perfect. PLS (Partial Least Squares) performs better at high levels of multicollinearity because it reduces dimensionality by extracting latent variables that capture the underlying structure of the concomitant variables even when they are highly correlated. This allows PLS to avoid issues of inflated variance and instability in the coefficient estimates which often occur in OLS under high multicollinearity.

Finally, the paper focuses on controlled simulation on multicollinearity and three significance levels which may not fully capture the complexity and variability of real – life datasets and the use of weights proportional to the standard error of regression coefficients assumes that this approach is optimal across all scenarios, which may limit the generalizability of the findings to cases where alternative weighting or regularization strategies could perform better. These are supposed limitations to the paper. Other criteria for comparison which may be considered for future study are bias, RMSE among others.

Acknowledgements

The authors acknowledge the contribution of FUTA – LISA LAB in the Department of Statistics, Federal University of Technology, Akure for allowing us to use the Lab for our simulation and research work.

Author Contributions

A. Aladesuyi: Original Draft Preparation, Conceptualization, Methodology, Results Discussion and Review and Editing. All authors have read and agreed to the published version of the manuscript
K. Ayinde: Original Draft Preparation, Conceptualization, Methodology, Review.
T. S. Fayose: Methodology, Review, Editing and Result Discussion.

Declaration of Competing Interest

The authors assert that they had no known conflicting financial interests or personal ties that may have seemed to have influenced the research that was published in the paper. Lastly, the authors concurred that all of the materials presented here are authentic and unique.

Funding

No financial assistance was received for this study.

ORCID

Taiwo Stephen Fayose: https://orcid.org/0000-0002-7457-9575

References

Alqaralleh, B. and Abdullah, S. (2021). Influence of Multicollinearity on Linear Regression Coefficients: A Simulation Approach. Communications in Statistics – Simulation and Computation, 50(3), 812 – 830.
Alshamrani, A.; Zhou, L. and Malik, K. (2024). Partial Least Squares Regression in High – Dimensional Multicollinear Settings. Computational Statistics and Data Analysis, 185, 107803.
Bastien, P.; Esposito – Vinzi, V. and Tenenhaus, M. (2005): PLS Generalized Linear Regression. Computational Statistics and Data Analysis, 48(1), 17 – 46.
Belsley, D. A.; Kuh, E. and Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley and Sons.
Chen, Y., Liu, W. and Li, G. (2020). Forecasting Economic Indicators Using PLS Under Multicollinearity. Applied Economics Letters, 27(19), 1573 – 1577.
Chong, I. G. and Jun, C. H. (2005). Performance of Some Variable Selection Methods when Multicollinearity is Present. Chemometrics and Intelligent Laboratory Systems, 78 (1 – 2), 103 – 112.
Duran, A. and Mert, A. (2021). A New Multicollinearity Diagnostic Based on Eigenvalue Analysis. Statistica Neerlandica, 75(4), 405 – 423.
Faraway, J. J. (1994): Order of Variables in Regression Analysis. Journal of the Royal Statistical Society: Series B (Methodological), 56(1), 115 – 117.
Fayose, T. S. and Ayinde, K. (2019). Different Forms Biasing Parameter for Generalized Ridge Regression Estimator. International Journal of Computer Applications, 181, 21 – 29.
Fayose, T. S.; Ayinde K. and Alabi, O. O. (2023a). M Robust Weighted Ridge Estimator in Linear Regression Model. African Scientific Reports, 2(123), 1 – 28.
Fayose, T. S.; Ayinde K.; Alabi, O. O. and Bello, A. H. (2023b). Robust Weighted Ridge Regression Based on S – Estimator. African Scientific Reports, 2(126), 1 – 28.
Gibbons, D. G. (1981): A Simulation Study of Some Ridge Estimators. Journal of the American Statistical Association, 76(373), 131 – 139.
Khan, M. A.; Rehman, R. and Jamil, M. (2022). Spectral Analysis of Multicollinearity in Regression: A New Perspective. Journal of Statistical Computation and Simulation, 92(4), 703 – 720.
Kroese, D. P.; Brereton, T.; Taimre, T. and Botev, Z. I. (2014). Why the Monte Carlo Method is so Important Today? Wiley Interdisciplinary Reviews: Computational Statistics, 6(6), 386 – 392.
Liu, Y.; Gao, H. and Wang, L. (2020). A Comparative Study of Ridge, Principal Component and Partial Least Squares

Regression under Multicollinearity. Statistical Methods and Applications, 647, 241 – 260.
Martens, H. and Naes, T. (1989): Multivariate Calibration. Chichester. John Wiley and Sons.
Metropolis, N. and Ulam, S. (1949). The Monte Carlo Method. Journal of the American Statistical Association, 44(247), 335 – 341.
Mevik, B. H. and Cederkvist, H. R. (2004). Mean Squared Error of Prediction (MSEP) estimates for Principal Component Regression and Partial Least Squares Regression. Journal of Chemometrics, 18(9), 422 – 429.
Nadarajah, S. and Krishna, R. (2023). Adaptive PLS Regression Under Multicollinearity: A Spectral Approach. Statistical Modelling, 23(1), 15 – 31.
Wang, T.; Lin, C. and Zhang, Y. (2023). Improving Regression Stability Under Multicollinearity Using Partial Least Squares. Journal of Multivariate Analysis, 195, 105049.
Wichern, D. and Churchill, G. A. (1978): A Comparison of Ridge Estimators. Technometrics, 20, 301 – 311.
Wold, H. (1966): Estimation of Principal Components and Related Models by Iterative Least Squares. In P. R. Krishnaiah, (ed.): Multivariate Analysis, 391 – 420. New York: Academic Press.
Wold, H. (1975): Path Models with Latent Variables: The NIPALS Approach. In H. M. Blalock, Aganbegian, F. M. Borodkin, R. Boudon, and V. Capecchi (Eds.), Quantitative Sociology: International Perspectives on Mathematical and Statistical Modeling, 307 – 357. New York: Academic Press.
Zhang, H. and Sun, D. (2022). Partial Least Squares and Eigenvalue Stabilization in Multicollinear Regression. Linear Algebra and Its Applications, 647, 241 – 260.