Al-Rafidain Journal of Computer Sciences and Mathematics (RJCM)

www.csmj.uomosul.edu.iq

Haemoglobin Levels Analysis Using Robust Partial Least Square Regression Models

Taha H Ali[1] and Mahammad Mahmoud Bazid[2]

[1,2] Department of Statistics and Informatics, College of Administration and Economics, Salahaddin University-Erbil, Iraq
Email: taha.ali@su.edu.krd[1] and mahammad.bazid@su.edu.krd[2]

Article information

Article history:
Received 17 January ,2025
Revised 22 February ,2025
Accepted 09 March ,2025
Published 26 June ,2025

Keywords:
Haemoglobin Levels,
Partial Least Squares Regression,
Robust Partial Least Squares Regression,
Outliers, noise data

Correspondence:
Taha H Ali
Email: taha.ali@su.edu.krd

Abstract

The Robust Partial Least Square Regression method is used to handle outliers and increase the explanation proportion, but it does not reduce the average of the mean square error. In this article, three methods are proposed to handle the problem of outliers, reduce the average of the mean square error, and increase the explanation proportion of the predictor and dependent variables. The first proposed method (Iteration) depends on identifying outliers by estimating the initial Partial Least Square Regression and then estimating outliers based on the residuals of those values to obtain the lowest mean square error, while the second and third proposed methods depend on a hybrid process between iteration and robust Partial Least Square Regression. The proposed and conventional methods were applied to estimate PLSR models on data Datasets for various ordinary patients in Iraq. The Dataset provides the patients' Cell Blood Count test information that can be used to create a Hematology diagnosis/prediction system. Also, this Data was collected in 2022 from Al-Zahraa Al-Ahly Hospital. The proposed iterative method with higher efficiency provided 5 variables (importance in the projection score that explain changes in HGB levels. The proposed methods gave better results than the robust Partial Least Square Regression method.

DOI: 10.33899/csmj.2025.156730.1163,

1. Introduction

Partial Least Squares Regression (PLSR) is a statistical method used for the analysis of data that has the same advantage as principal component regression in terms of reducing the number of explanatory variables before conducting a regression on the response variables in a multivariate dataset. PLSR will extract a set of underlying factors, or principal components, from the data that contains the basic information about the relationship between the response variables and the explanatory variables. PLSR also has another advantage over principal component regression, in that it can take the relationship between the response and explanatory variables into account when fitting or estimating those factors. This basic approach often makes PLSR an excellent choice for the analysis of datasets that comprise many explanatory variables, are 'wide' and have a limited number of observations, or are 'short'. The appeal of PLSR in analyzing a wide dataset is that it can model complex relationships between response and explanatory variables without including all the corresponding operational effects needed in, for example, analysis of variance or covariance applied to each explanatory variable separately. However, employing PLSR requires an awareness of the theoretical properties of the model and the necessity to test model assumptions (Zeng et al. 2021; Burnett et al. 2021; Hair & Alamer, 2022). This paper will present robust models of Partial Least Squares regression to analyze Haemoglobin data in the presence of outliers. The data we consider contains Haemoglobin levels for n = 100 patients taken from the hospital in Baghdad. Clinical measurements in medicine, Haemoglobin, help in the diagnosis of blood diseases such as anemia. This is a heme protein contained in red blood cells; its major function in the body is molecular oxygen. Its absence weakens other cells, which reduces the ability of capillaries to effectively supply oxygen, leading to problems and can lead to death. The dataset considered is of interest as it contains several outliers, which make almost every statistical model considered here fail. The robust PLS models proposed here

can handle these problems and, in addition, reveal useful structure in the data (Prager et al. 2023). This paper is organized into sections where the Methodology in section 2, the outline of Partial Least Squares and robust PLS are given in sections 3, 4, and 5. include outliers, results for real data analysis are discussed in section 6. with the proposed method, and finally section 7. concludes the paper.

2. Methodology

2.1. Multiple Linear Regression (MLR):

The problem of multiple linear regression, or MLR, can be expressed as follows (Ali and Saleh, 2022). The objective is to determine a linear connection between the variables, $x_j = (j = 1 - m)$, and a variable, y, by feature measurement. This has a mathematical representation of

$$\boldsymbol{y = b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_m x_m + e} \quad (1)$$

$$\boldsymbol{y = x' b + e} \quad (2)$$

In equation (1), the x_m are referred to as independent variables, while the dependent variable is y. the b_m are the sensitivities, while $\boldsymbol{e}$ is the residual or error. In equation (2), x' is a row vector, b is a column vector, and y is a scalar. Multilinear dependencies are described for a single sample in Equation 1. When n samples are obtained, the column vector $y_i (i = 1 - n)$ may be expressed as follows: b stays constant, and the rows of a matrix X are formed by the vectors x'_j.

$$\boldsymbol{y = Xb + e} \quad (3)$$

The "least squares method" is the most often used technique for this. The solution using the least squares is:

$$\boldsymbol{b = (X'X)^{-1} X'y} \quad (4)$$

The most common issue with MLR is hinted at in Equation (4): the inverse of $\boldsymbol{X'X}$ could not exist. The same problem goes by the titles of singularity, zero determinant, and collinearity.

Although it would seem obvious at this point that there must always be an equal number of samples and variables, there are many ways to formulate this problem. Eliminating a few variables ($\boldsymbol{m}$) in the scenario when $\boldsymbol{m > n}$ ($\boldsymbol{n}$ is the number of observations) is one of them. There are several techniques for selecting which variables to remove, one of these techniques is PLS.

2.2. Model construction

The NIPALS algorithm's characteristics provide the foundation of the PLS model (Ali et al. 2023). The data matrix may be represented by the score matrix. A regression between the scores for the X and Y blocks would make up a basic model. The outside relations (X and Y blocks separately) and the inside relation (connecting both blocks) make up the PLS model, the X block's outer relation is:

$$\mathbf{X = TP' + E} \quad (5)$$

$$\mathbf{Y = UQ' + F} \quad (6)$$

Where:

- **X** is a $n \times m$ predictor matrix.
- **Y** is a $n \times p$ response matrix.
- **T** and **U** are $n \times 1$ matrices that are, as well, projectors of **X** (the **X** score, component or factor matrix) and projectors of **Y** (Ali et al. 2023).
- **P** and **Q** are, accordingly, $m \times 1$ and $p \times 1$ loading matrices.
- Matrices **E** and **F** are the error terms, supposed to be independent and symmetrically distributed random normal variables.

The breakdown of $\boldsymbol{Y}$ are done to optimize the covariance between **T** and **U** (Shahla et al. 2023).

The covariance of column i of T (length n) with the column i of U (length n) is maximized. Take note that this covariance is determined pair by pair. Furthermore, there is zero covariance between column i of T and column j of U (with $i \neq j$).

For PLSR, the scores constitute an orthogonal basis, so the loadings are selected accordingly. When orthogonality is applied upon loadings (and not the scores) in PCA, there is a significant difference.

The sums range from 1 to **a**. It is possible to define every component and determine whether E = F = 0. We go into how and why this is done below. The goal is to achieve as helpful a relationship between **X** and **Y** as feasible while also describing **Y** as well as practical and minimizing $\|F\|$. A graph of the Y block score, u, versus the **X** block score, t, for each component may be used to determine the inner relation. A linear model is the most basic for this relation:

$$\boldsymbol{\hat{u}_h = b_h k_h} \quad (7)$$

Where $b_h = u_h' t_h / t_h' t_h$. In the MLR and PCR models, the b_h function as the regression coefficients b. This model is not optimal. The principal components are estimated for each block independently, resulting in a weak relationship between them, which explains the rationale. It would be preferable if they knew more about one another, resulting in components that are slightly rotated and closer to the regression line. To produce slightly rotated components that are closer to the regression line in Figure 3, it would be preferable to provide them with information about one another. Model oversimplification: 2 × PCA The NIPALS section is an example of an algorithmic representation of an oversimplified model.

For the X block:

1- take $\mathrm{t_{start}} = \text{some } \mathrm{x_j}$
2- $\mathrm{p}' = \mathrm{t}'\mathbf{X}/\mathrm{t}'\mathrm{t}\ (= \mathrm{u}'\mathbf{X}/\mathrm{u}'\mathrm{u})$
3- $\mathrm{p}'_{new} = \frac{\mathrm{p}'_{old}}{\|\mathrm{p}'_{old}\|}$
4- $t = \frac{\mathbf{X}\mathrm{p}}{p'p}$
5- If t in steps 2 and 4 are equal, stop; if not, go on to step 2.

For the Y block:

1- take $\mathrm{u_{start}} = \text{some } \mathrm{y_j}$
2- $\mathrm{q}' = \mathrm{u}'\mathbf{X}/\mathrm{u}'\mathrm{u}\ (= \mathrm{t}'\mathbf{X}/\mathrm{t}'\mathrm{t})$
3- $\mathrm{q}'_{new} = \frac{\mathrm{q}'_{old}}{\|\mathrm{q}'_{old}\|}$
4- $u = \frac{\mathbf{Y}\mathrm{q}}{q'q}$
5- If u in steps 2 and 4 are equal, quit; if not, carry on stepping 2.

Exchanging scores improves the internal relationship. The relationships are expressed as entirely distinct algorithms. By allowing t and u to switch positions in step 2, one may learn more about the other. In this stage, take note of the sections included in parenthesis. As a result, the two algorithms may be expressed sequentially:

1- Take $\mathrm{u_{start}} = \text{some } \mathrm{y_j}$
2- $\mathrm{p}' = \mathrm{u}'\mathbf{X}/\mathrm{u}'\mathrm{u}\ (\mathrm{w}' = \mathrm{u}'\mathbf{X}/\mathrm{u}'\mathrm{u})$
3- $\mathrm{p}'_{new} = \frac{\mathrm{p}'_{old}}{\|\mathrm{p}'_{old}\|}\ (\mathrm{w}'_{new} = \frac{\mathrm{w}'_{old}}{\|\mathrm{w}'_{old}\|})$
4- $t = \frac{\mathbf{X}\mathrm{p}}{p'p}\ (\ t = \frac{\mathbf{X}\mathrm{w}}{w'w})$
5- $\mathrm{q}' = \mathrm{t}'\mathbf{Y}/\mathrm{t}'\mathrm{t}\ (= \mathrm{t}'\mathbf{X}/\mathrm{t}'\mathrm{t})$
6- $\mathrm{q}'_{new} = \frac{\mathrm{q}'_{old}}{\|\mathrm{q}'_{old}\|}$
7- $u = \frac{\mathbf{Y}\mathrm{q}}{q'q}$
8- Compare the t in step 4 with the 1 in the prior iteration step. If they're equal (within a given rounding error), stop; otherwise move to 2 (In the scenario for which the Y block contains just one variable, steps 5-8 may be avoided by writing Q = 1).

This algorithm generally converges extremely rapidly to yield rotated components for X and Y blocks. acquiring scores for orthogonal X blocks. The algorithm's failure to provide orthogonal t values remains an issue. The reason for this is that the PCA's computation sequence was changed. As a result, weights w' are used instead of the p' (refer to the formulae in parenthesis in the preceding paragraph). After convergence, an additional loop may be added to get orthogonal t values:

$$\mathbf{p}' = \mathbf{t}'\mathbf{X}/\mathbf{t}'\mathbf{t} \tag{8}$$

With $\mathbf{p}'_{new} = \frac{\mathbf{p}'_{old}}{\|\mathbf{p}'_{old}\|}$ It is now feasible to compute the new $\boldsymbol{t}$: $\boldsymbol{t} = \frac{\mathbf{X}\mathbf{p}}{\boldsymbol{p'p}}$ However, this ultimately constitutes only scalar multiplication with the norm of p' in Eqn.8: $\boldsymbol{t_{new}} = \boldsymbol{t_{old}}\|\mathbf{p}'_{\boldsymbol{old}}\|$. While orthogonal t values are not strictly required, their use facilitates a more straightforward comparison with PCR. To ensure accurate predictions without error, it is essential to apply the same resealing to the weights $\boldsymbol{w}'$: $\boldsymbol{w}'_{\boldsymbol{new}} = \boldsymbol{w}'_{\boldsymbol{old}}\|\mathbf{p}'_{\boldsymbol{old}}\|$. Subsequently, t may be used for the inner relation as delineated in Equation 19, and the residuals may be computed from: $\boldsymbol{E_1 = X - t_1 p_1'}$ $\boldsymbol{and}$ $\boldsymbol{F_1 = Y - u_1 q_1'}$. In general,

$$\boldsymbol{E_h = E_{h-1} - t_h p_h';}\ \mathbf{X = E_0} \tag{9}$$

$$\boldsymbol{F_h = F_{h-1} - u_h q_h';}\ \mathbf{Y = F_0} \tag{10}$$

However, u_h is substituted with its estimator, $u_h = b_h t_h$, in the outer relation for the Y block, yielding a mixed relation:

$$\boldsymbol{F_h = F_{h-1} - b_h t_h q_h'} \tag{11}$$

(Remember that the goal is to reduce $\|\boldsymbol{F_h}\|$.) The ability to utilize the model parameters for prediction from a test set is ensured by this mixed connection. Moreover, one may continue until the rank of the **X** block is depleted since the rank of **Y** is not reduced by **1** for every component. The Appendix contains the whole procedure as well as a matrix and vector illustration (Geladi and Kowalski, 1986).

2.3. Partial Least Squares Regression

Partial least squares regression is a popular multivariate method. It is used when creating a regression model for data that is multicollinear or in which there are more explanatory variables than samples. SIMPLS is one of the primary methods used in PLSR, which aims to repeatedly extract uncorrelated latent variables (Alin and Agostinelli). Partial least squares methods are useful as exploration tools for the analysis of large data blocks of high dimension, especially for data involving a great number of highly collinear and low-level intensity predictors. The central theme of robust partial least squares regression concerns the high sensitivity of robust methods to the leverage values of the predictors. Collinearity and predictor contamination in the form of low-level intensity measurements are common challenges in analyzing biomedical spectral data with PLS regression methods. Traditional PLS methods are highly sensitive to the leverage of the input variables, and this high leverage may result in overfitting. Outliers, if present, will distort the results, and thus the use of traditional PLS regression in the presence of outlying observations is not advisable. In particular, the inclusion of outlying observations in PLS regression models developed for the analysis of microarray data is highly suspect (Chen et al.2021). The ability of different data analysis approaches to supply the user with statistical and cognitive tools differs. The soft multivariate bilinear modelling approach of PLS Regression enables cognitive access to important and trustworthy information in data when linked with suitable interactive computer graphics. Additionally, cross-validation enables a statistical assessment of the results' reliability. As far as I'm aware, no other statistical approach has comparable versatility. The PLS Regression appears to regularly rank among the top regression procedures in terms of statistical prediction ability when compared to competing approaches that have all been adequately calibrated. Therefore, it is particularly well-suited for non-statisticians, or researchers who are unable to commit the required time to learning complex, abstract statistical approaches and who wish to apply their vital contextual knowledge when evaluating data. The PLS Regression was created in response to traditional statistics' inadequacies and unmet data analysis objectives. It developed from the close collaboration of chemist Svante Wold and his father, statistician Herman Wold. Two very different but occasionally equally oppressive scientific cultures the parameter estimation of traditional statistical modelling, which focused on distribution theory and hypothesis testing, and the mathematical modelling in traditional chemistry and physics, which focused on hard causal models developed the PLS Regression at the start of the 1980s (Martens, 2001).

2.4. Robust Partial Least Squares Regression

The traditional Partial Least Squares Regression methods are sensitive to various problems such as outliers, and noise, which prevail in real datasets. Such problems deteriorate their modelling performance, and in turn, practitioners may fail to obtain useful information from the constructed models after further drawing decisions or making predictions. Robust regression that mitigates such problems becomes an important research line (robust versus the outliers and proving accurate coefficient estimators). By investigating some cases, we discuss the pivotal role of error detection and the incompleteness of its list. Different iterations of when to use robust PLSR can have different levels of robustness. We provide case studies considering practical operations (Chan et al., 2022). Robustness improves the generalization of PLSR, as sufficiently small errors for iterations can occasionally provide less robust results with the traditional methods. The improved robustness of modelling can reveal a very strong ideal to unrealistic relation between the predictors and regressors. With different robust techniques, these cases can be separated and returned to let alone. Robust PLSR can require shorter dimensions to be used. Subsequently, the number of increments can be longer, but also the fitting and standard simplicity values can lead to clearer and more usable results than those for the traditional PLSR. The practical application of robust PLSR brings some advantages. From a theoretical perspective, it can be semi-automatically quantified, i.e., employed while the splits work under the general mathematical background and are implemented carefully (Hair and Alamer, 2022). Robust PLS has specialized in using criteria for the robustification of PLSR against both outliers and leverage points.

Three well-known methods for down-weighting leverage points and large residuals have been proposed. Furthermore, robust PLS estimation known as robust Iterative Robust PLS estimation has been implemented to the power weighting coefficient of the method to detect highly influential observations. These and other studies have addressed the robustness of PLS to leverage points in multivariate and high-dimensional settings with some similarity to the well-known algorithm of PLSR, but they involve significantly different estimation procedures due to their choice of robust criteria and respective characteristics (Ali et al. 2025). The effectiveness of these methods under different conditions or with diverse applications shows that the criterion of robustness plays a crucial role in the performance of the introduced methods. Also, Meetings and variations of the related literature demonstrate that the justifications of the

PLS method have a promising performance by modelling the influence of both leverage and low-leverage points (Kılıç et al., 2021).

2.5. Iterative Methods in Robust PLSR

Iterative methods refine statistical models sequentially. They first establish a tentative solution and then adjust the model to efficiently approximate some targets. Such methods act as chains of feedback loops. For PLSR, iterative methods enable obtaining optimal (or locally optimal) models according to a chosen criterion. Iteration is essential in predictive modelling when we build a final model based on available data that can be leveraged continuously. Such methods apply to any problem given only the ability to calculate volatility, gradient, or a function adjoint. The iterative strategy in predictive modelling involves defining an initial linear combination, improving the fitness of the obtained latent variables, reducing their cross-correlation, or concentrating on a high-leverage subgroup of data to enhance interpretability. Non-linearity concerning the directions of maximal covariance can be efficiently approached by alternating between space construction and label estimation (Ali, 2018).

Iterative methods provide robust solutions when non-normality is considered. Step-by-step improvement can result in models resistant to the presence of strong outliers. However, a major drawback of such an approach can be slow convergence, even close to the optimum of a stopping criterion. Thus, this iterative strategy is widely used when seeking robust solutions in PLS2. Iterative methods can also simply improve a model to yield greater interpretability. These methods permit performing feature engineering to extract information that may be important for understanding the observations. Because they do not yield the most optimal predictive models, decreasing the model complexity may be related to a decrease in their predictive power. Overall, it seems there is a trade-off between convergence and model accuracy (Knief and Forstmeier, 2021).

2.6. Outliers

Outliers are a common occurrence for any applied statistician who has examined real data sets. An observation is considered an outlier if it differs significantly from other observations, raising questions about whether it was caused by a different factor. When analyzing a sample that contains outliers, the notable differences between outlying and inlying observations, as well as the extent of deviation between the outliers and the inlier group, are evaluated using an appropriately standardized scale (Omar et al. 2020). Outliers are the extreme values of variables in each dataset. Outliers are the values that diverge from the overall pattern of data. The presence of outliers in the data often results in misleading interpretations, which could lead to incorrect decisions. Outliers can skew predictions and misrepresent results, resulting in wrong conclusions. Additionally, as most multivariate techniques assume normally distributed data in consecutive steps, the inclusion of outliers can seriously distort results and conduct tests of significance. If adequate procedures are not considered, outliers can negatively affect external validity and generalizability. However, not all the points that appear extreme are necessarily 'bad.' Some extreme values might have interesting information. A thorough investigation of the outlying points is necessary. It is important to remember that they have tails, but they are few. However, they do have an impact and should not be ignored (Sullivan et al., 2021). Based on the nature of outliers, they can be classified into various categories: univariate, multivariate, and contextual outliers. It is of utmost importance to understand the cause, i.e., why the outlying value exists before choosing a methodology to manage the outliers in the data. Outliers can occur for various reasons, and those reasons are divided into three main categories: data entry errors, measurement errors, and natural variations (Omar and Ali, 2025). Data entry errors occur due to human intervention or technological malfunction, or due to poor procedures. Measurement errors due to technological problems and data collection variability are made while measuring each value. Finally, natural variations occur due to differences in measurements because many samples are being taken. We have a broad phase variation in the measurements. Occasionally, these extreme values could be actual data, but humans are so biased and unconvinced that these outliers could be removed to achieve them. Even unusual cases lead users to remove these outliers and not draw an appropriate end from them. It is also important to remember that people have a specific idea about the significance of data. The higher or lower values of some variables deviate from these beliefs and are marked as outliers. Therefore, we can assume that they are of particular interest in our study. There is a risk of making mistakes if we omit outliers just because they appear to be different. We, therefore, need to understand the reason why outliers arise or exist before considering an adequate procedure for managing the outliers in our data (Smiti, 2020).

2.7. Proposed Methods

The three proposed methods for treating outliers are summarized in the following:

First Proposed:

- Estimate a partial least squares regression model that maximizes the covariance matrix between the Predictor and dependent variables after choosing many suitable components to obtain predictions of the initial values of the dependent variable and the residual.
- Identifying outliers y(o) from the standard residuals of a partial least squares regression model that are outside an interval (∓ 2.5) or the largest residual value.
- Calculate the initial average of mean Squares Error (AMSE) of the model from the following formula:

$$\text{AMSE} = \sum_{k=1}^{2} \sum_{j=1}^{p+1} (\text{MSE}(k,j)/2(p+1) \qquad (12)$$

The number of principal components is p. MSE includes two parts, the mean square error of X (MSEx) which measures how the model explained the variation in the Predictor variables, and the mean square error of Y (MSEy) which measures the accuracy of the model:

$$\text{MSEx} = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{ij} - \hat{x}_{ij}\right)^2 \qquad (13)$$

MSEx quantifies the error between the original x and the reconstructed x from the model.

$$\text{MSEy} = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(y_{ij} - \hat{y}_{ij}\right)^2 \qquad (14)$$

MSEy quantifies the error between the actual value y and the predicted by the PLSR model.

- Estimate outliers using the following equation:

$$Y(o) = \hat{y}(o) - Residual(o) \qquad (15)$$

Residual (o) is the outlier residual, using Y(o) instead of $\hat{y}(o)$ with Y to estimate a PLS model and compute AMSE.

- If the AMSE value is greater than (0.001), then the outlier in equation (14) will be re-estimated and get AMSE for a new PLS model and so on until the AMSE is less than (0.001).
- Finally, the estimated values of the outliers with the least AMSE are used to create the PLS model.

Second Proposed:

The second proposed method is based on the hybrid method (Robust-Iteration) which uses a robust estimator (Savitzky-Golay filter using iterative reweighting in combination) to handle outliers and noise in data based on maximizing the explanation ratio of the Predictor and dependent variables as inputs to the iterative method that minimizes the AMSE as in the first proposal.

Third Proposed:

The third proposed method is based on the hybrid method (Iteration-Robust) which uses the iterative process that minimizes the AMSE as in the first proposal as inputs a robust estimator to handle outliers and noise in data based on maximizing the explanation ratio of the Predictor and dependent variables.

3. Application Aspect

The proposed and conventional methods were applied to estimate PLSR models on data Datasets for various ordinary patients in Iraq. The Dataset provides the patients' Cell Blood Count test information that can be used to create a Hematology diagnosis/prediction system. Also, this Data was collected in 2022 from Al-Zahraa Al-Ahly Hospital. The dependent variable represents Hemoglobin (HGB), Normal Ranges: 11.0 to 16.0, Unit: g/Dl, while the Predictor variables represent 19 tests, in **Table 1**:

Table 1. Cell Blood Count test

No.	Symbol	Predictor Variable
1	WBC	White Blood Cell, Normal Ranges: 4.0 to 10.0, Unit: 10^9/L.
2	LYMp	Lymphocyte percentage, which is a type of white blood cell, Normal Ranges: 20.0 to 40.0, Unit: %
3	MIDp	Indicates the percentage combined value of the other types of white blood cells not classified as lymphocytes or granulocytes, Normal Ranges: 1.0 to 15.0, Unit: %
4	NEUTp	Neutrophils are a type of white blood cell (leukocytes); neutrophils percentage, Normal Ranges: 50.0 to 70.0, Unit: %
5	LYMn	Lymphocyte numbers are a type of white blood cell, Normal Ranges: 0.6 to 4.1, Unit: 10^9/L.
6	MIDn	Indicates the combined number of other white blood cells not classified as lymphocytes or granulocytes: 0.1 to 1.8, Unit: 10^9/L.
7	NEUTn	Neutrophils are a type of white blood cell (leukocytes); neutrophils Number, Normal Ranges: 2.0 to 7.8, Unit: 10^9/L.
8	RBC	Red Blood Cell, Normal Ranges: 3.50 to 5.50, Unit: 10^12/L
9	HCT	Hematocrit is the proportion, by volume, of the Blood that consists of red blood cells, Normal Ranges: 36.0 to 48.0, Unit: %
10	MCV	Mean Corpuscular Volume, Normal Ranges: 80.0 to 99.0, Unit: FL
11	MCH	Mean Corpuscular Hemoglobin is the average amount of Haemoglobin in the average red cell, Normal Ranges: 26.0 to 32.0, Unit: pg.
12	MCHC	Mean Corpuscular Hemoglobin Concentration, Normal Ranges: 32.0 to 36.0, Unit: g/dL
13	RDWSD	Red Blood Cell Distribution Width, Normal Ranges: 37.0 to 54.0, Unit: fL
14	RDWCV	Red blood cell distribution width, Normal Ranges: 11.5 to 14.5, Unit: %
15	PLT	Platelet Count, Normal Ranges: 100 to 400, Unit: 10^9/L
16	MPV	Mean Platelet Volume, Normal Ranges: 7.4 to 10.4, Unit: fL
17	PDW	Red Cell Distribution Width, Normal Ranges: 10.0 to 17.0, Unit: %

18	PCT	The level of Procalcitonin in the Blood, Normal Ranges: 0.10 to 0.28, Unit: %
19	PLCR	Platelet Large Cell Ratio, Normal Ranges: 13.0 to 43.0, Unit: %

A random sample of 100 observations was taken from these examinations, and the statistical description is in **Table 2**. The mean level of the dependent variable HGB was (11.4610), and it is within the normal period (11-16) with a standard deviation (2.74921). **Figure 1** shows that several observations are outside this normal period and that some values are much less than the minimum (11). All means of the Predictor variables for blood tests were within the normal range.

Table 2. Descriptive Statistics

Predictor Variable	Mean	Normal Range	Std. Deviation
HGB	11.4610	11-16	2.74921
WBC	7.0520	4-10	3.59249
LYMp	26.2470	20-40	11.38818
MIDp	8.6840	1-15	7.12226
NEUTp	65.6570	50-70	11.02517
LYMn	1.6900	0.6-4.1	0.85298
MIDn	0.5970	0.1-1.8	0.40613
MEUTn	4.7660	2-7.8	2.90023
RBC	4.6032	3.5-5.5	0.66308
HCT	40.3190	36-48	28.40610
MCV	81.8640	80-99	7.87709
MCH	25.4800	26-32	3.39233
MCHC	32.0740	32-36	6.72509
RDWSD	37.4610	37-54	4.71923
RDWCV	13.1010	11.5-14.5	1.40133
PLT	164.7600	100-400	48.85272
MPV	9.8490	7.4-10.4	1.23219
PDW	13.6420	10-17	2.03068
PCT	0.1548	0.10-0.28	0.04804
PLCR	27.0220	13-43	7.26473

PLSR analysis is used to measure the effect of Predictor variables on the dependent variable when the number of observations minus one is less than the number of Predictor variables, which is not available in this data, and when there is a multicollinearity problem between the Predictor variables, so multiple linear regression (MLR) analysis was used to verify it as in **Table 3**.

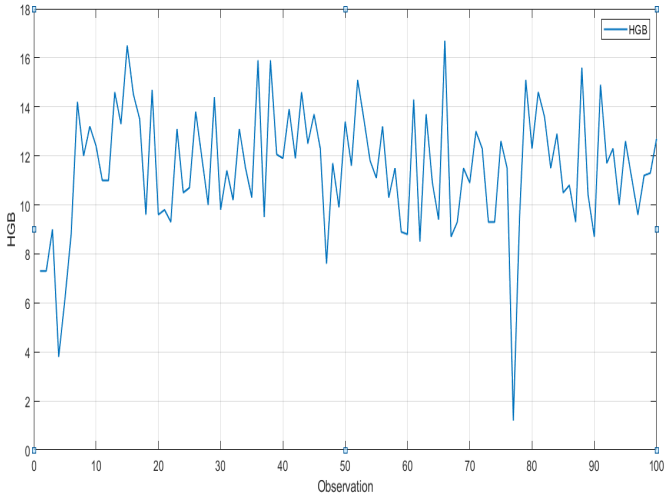

Figure 1. Scatter plot of the dependent variable (HGB)

Table 3. Multiple Linear Regression Model

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	-6.19	20.920		-.296	.768		
LYMp	-.052	.188	-.217	-.280	.780	.003	298.29
MIDp	-.008	.019	-.020	-.416	.679	.840	1.190
NEUTp	-.050	.202	-.199	-.246	.806	.003	322.70
LYMn	.111	.289	.034	.385	.702	.252	3.962
MIDn	.269	1.940	.040	.139	.890	.025	40.577
MEUTn	-.111	.262	-.117	-.424	.673	.026	37.770
RBC	2.915	.386	.703	7.543	.000	.233	4.293
HCT	-.002	.034	-.019	-.054	.957	.017	60.551
MCV	.092	.089	.263	1.039	.302	.031	31.779
MCH	.378	.173	.466	2.186	.032	.044	22.496
MCHC	.009	.020	.021	.444	.658	.879	1.138
RDWSD	-.069	.107	-.119	-.650	.517	.060	16.568
RDWCV	.037	.328	.019	.114	.910	.073	13.792
PLT	-.012	.031	-.220	-.403	.688	.007	147.55
MPV	-.790	.951	-.354	-.831	.408	.011	89.799
PDW	-.040	.109	-.030	-.369	.713	.315	3.179
PCT	14.80	30.499	.259	.485	.629	.007	140.36
PLCR	.072	.083	.190	.871	.386	.043	23.523

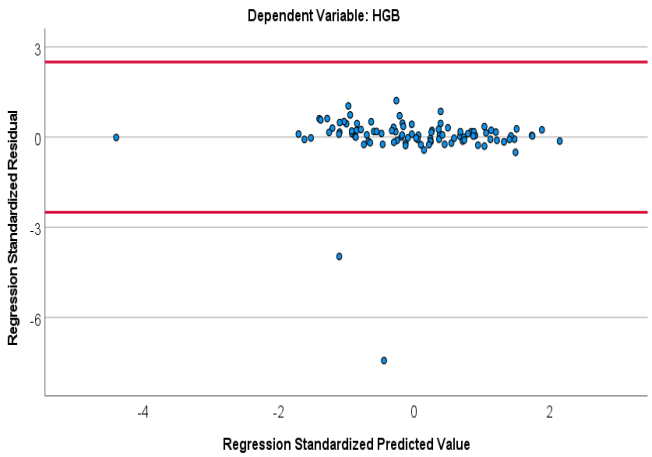

Figure 2. Regression Standardized Residuals for MLR

Table 3 shows the existence of the problem of multicollinearity between most of the Predictor variables (except for variables MIDp, LYMn, RBC, MCHC, and PDW) because the values of variance inflation factor (VIF) were greater than 5 and that the Predictor variables had no effect on the dependent variable because the values of p-values were greater than the significance level (0.01) except for variable RBC. Because of the problem of multicollinearity, using PLSR is more appropriate for analyzing this data, and with outliers in the model, As can be seen from the plot of the residuals of the MLR model (there are two outliers 4 and 77 were outside the interval ± 2.5) in **Figure 2** and Table 4 thus robust PLSR will be more appropriate than traditional PLSR. **Table 4** shows that there were two outliers (4 and 77) with very low HGB levels (3.80 and 1.20) while the predictive values (8.6867 and 10.3460) and standard residuals (-3.971 and -7.432), and residuals (-4.88667 and -9.14598) values were unacceptably large.

Table 4. Outliers Diagnostics

Case Number	Std. Residual	HGB	Predicted Value	Residual
4	-3.971	3.80	8.6867	-4.88667
77	-7.432	1.20	10.3460	-9.14598

The proposed methods depend on estimating outliers. This is done first by identifying outliers based on the standard residuals of the PLS model as in **Figure 3**. The two values (y_4 and y_{77}) are considered outliers thus they will be estimated using the robust PLSR and proposed methods. When the estimation of outliers was repeated (22) times, the iterative method provided the lowest sum of squared errors. To demonstrate the efficiency of the proposed methods (Iteration, Robust-Iteration, and Iteration-Robust) and compare them with the robust method in handling noise and outliers in the PLSR model, principal components (1-8) were used for five methods, and the comparison criteria were calculated as in **Table 5** which shows the results of the efficiency criteria for the five methods, where the first method represents PLSR (without filter), the second method represents robust PLSR (Robust), the third method is the proposed (Iteration), the fourth method (Robust-Iteration) and the fifth method (Iteration-Robust). Eight principal components were identified that were appropriate for this data and had an explanation proportion R^2X greater than 90%, R^2Y greater than 50%, and minimum MSE (366.8932, 152.1978, 0.0942, 0.0039, and 85.9875, respectively) for all methods used (the residuals are shown in **Figures 3-7**). The robust PLSR method was robust against outliers and provided an increase in the explanation proportions for the Predictor variable (from 98.3902 to 99.1087) and a decrease in the explanation proportions for the dependent variable (from 71.9128 to 51.0907) while decreasing the value of MSE (from 366.8932 to 152.1978). The result is logical because the robust PLSR method focuses on increasing the explanation ratio and reducing the MSE. The first proposed method (Iteration) is also strong against outliers and provided an increase in the explanation proportions for the Predictor variable (from 98.3902 to 98.4067) and dependent variable (from 71.9128 to 71.9805) while reducing the value of MSE (from 366.8932 to 0.0942). The increase in the proportion of explanation of the Predictor variables was limited. Still, the decrease was large in MSE, and this is logical in the mechanism of the iterative method in minimizing MSE and does not focus on maximizing the proportion of explanation. Also, note the big difference in reducing the value of MSE compared to the robust method (from 152.1978 to 0.0942).

Table 5. PLSR Model Results

Method	Number of principal components	R^2X	R^2Y	MSE
Without Filter	1	61.7492	9.2495	1276.300
Robust		69.6564	22.1909	481.9676
Iteration		61.7614	9.3976	0.1779
Robust-Iteration		50.7881	3.2079	0.0382
Iteration-Robust		70.7870	21.0093	458.8203
Without Filter	2	85.2478	13.9313	942.4720
Robust		83.2461	43.3924	362.2439
Iteration		85.2817	14.0427	0.1655
Robust-Iteration		83.0960	15.3060	0.0240
Iteration-Robust		83.4525	39.5373	295.9905
Without Filter	3	90.3799	23.0456	751.8545
Robust		91.8080	39.4192	298.9431
Iteration		90.3255	23.4796	0.1423
Robust-Iteration		87.6032	11.5868	0.0151
Iteration-Robust		93.3105	40.8180	289.5752
Without Filter	4	94.9597	27.3721	620.5871
Robust		95.2583	38.4271	252.6010
Iteration		95.0061	26.6463	0.1217

Robust-Iteration		95.2336	15.1892	0.0108
Iteration-Robust		95.9177	41.8393	220.4305
Without Filter		96.3657	33.0465	528.7259
Robust		93.6559	39.7218	215.0365
Iteration	5	96.3762	32.7080	0.1076
Robust-Iteration		95.9402	15.8539	0.0077
Iteration-Robust		97.5739	43.3489	171.5248
Without Filter		97.2024	39.2208	460.8768
Robust		97.5959	39.5672	199.1697
Iteration	6	97.1986	39.4864	0.0976
Robust-Iteration		97.8790	25.9688	0.0062
Iteration-Robust		98.4002	45.3286	135.3833
Without Filter		97.6233	61.2320	408.9185
Robust		98.4031	45.0333	165.6799
Iteration	7	97.6266	60.2657	0.0987
Robust-Iteration		98.3945	21.5845	0.0047
Iteration-Robust		98.7675	48.0387	107.7974
Without Filter		**98.3902**	**71.9128**	**366.8932**
Robust		**99.1087**	**51.0907**	**152.1978**
Iteration	8	**98.4067**	**71.9805**	**0.0942**
Robust-Iteration		**99.0860**	**50.2236**	**0.0039**
Iteration-Robust		**99.0623**	**52.7409**	**85.9875**

The second proposed method (Robust-Iteration) is also robust against outliers and provided an increase in the explanation proportions for the Predictor variable (from 98.3902 to 99.0860) and a decrease in the explanation proportions for the dependent variable (from 71.9128 to 50.2236) while decreasing the value of MSE (from 366.8932 to 0.0039), noting the big difference in reducing the value of MSE compared to the robust method (from 152.1978 to 0.0039). The third proposed method (Iteration-Robust) is also strong against noise and provided an increase in the explanation proportions for the Predictor variable (from 98.3902 to 99.0623) and a decrease in the explanation proportions for the dependent variable (from 71.9128 to 52.7409) while decreasing the value of MSE (from 366.8932 to 85.9875), noting the big difference in reducing the value of MSE compared to the robust method (from 152.1978 to 85.9875). General the proposed method (Iteration) gave the best results compared to other proposed methods and PLSR and robust PLSR method because it has the lowest MSE with the highest explanation ratio. **Figures 3-7** show the plot of the residuals of the five models and the proposed methods obtained the lowest standard residual values compared to the PLSR and robust PLSR models.

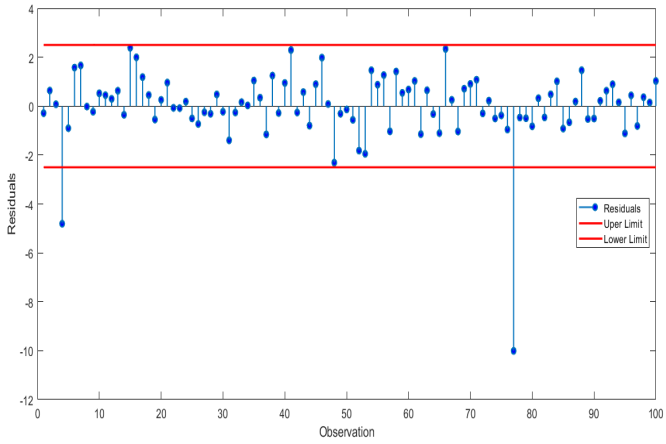

Figure 3. Residuals of the PLSR Model

Figure 3 shows that there were two outliers (4 and 77) for the PLSR model that were outside the interval (±2.5) Thus robust PLSR will be more appropriate than traditional PLSR as in **Figure 4**.

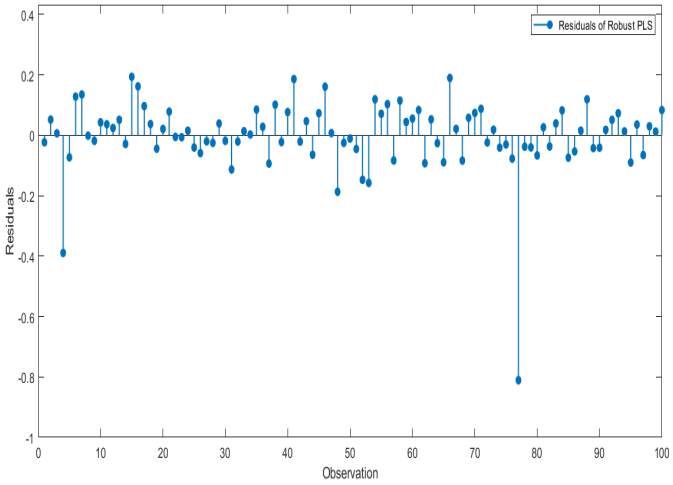

Figure 4. Residuals of the robust PLSR Model

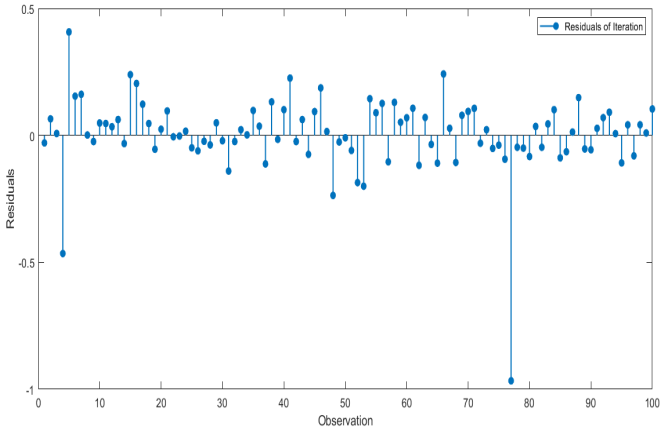

Figure 5. Residuals of the Iteration PLSR Model

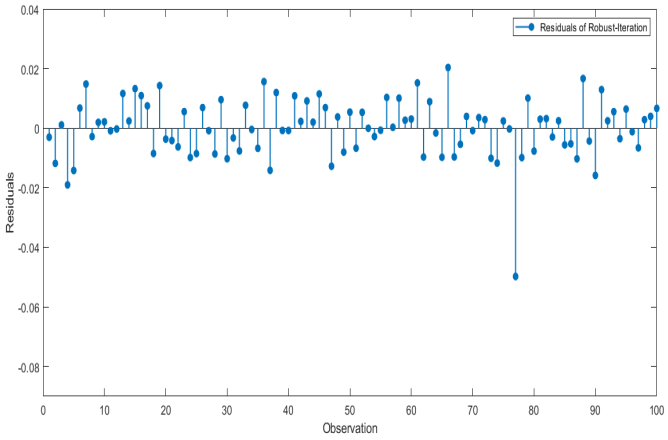

Figure 6. Residuals of the Robust-Iteration PLSR Model

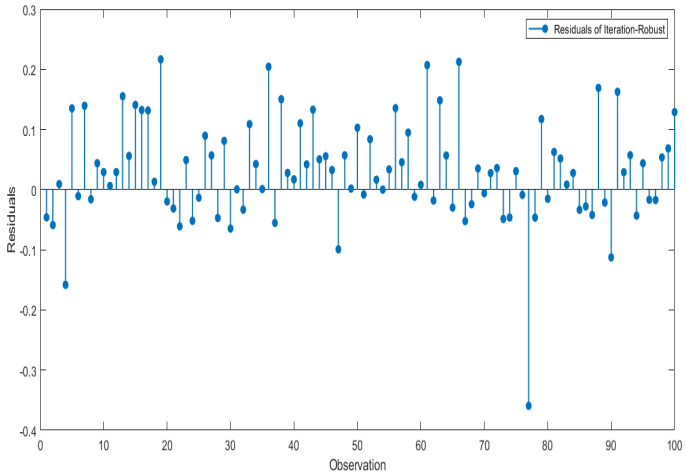

Figure 7. Residuals of the Iteration-Robust PLSR Model

Figures 4-7 show the small and acceptable residual values (-1-0.5), compared with the classical method (-38-28). The four methods with outliers provided results of different efficiency depending on the number of principal components used in the analysis. Depending on the 8 principal components, the robust method and the proposed methods address the problem of outliers and provide highly efficient estimators sorted by order of least MSE (Iteration PLSR, Robust-Iteration, Iteration-Robust, and Robust PLSR). **Figure 8** shows the actual and estimated values for the HGB levels from the five models and shows the large variation in estimated values depending on the method used to calculate the PLSR model parameters (using eight principal components).

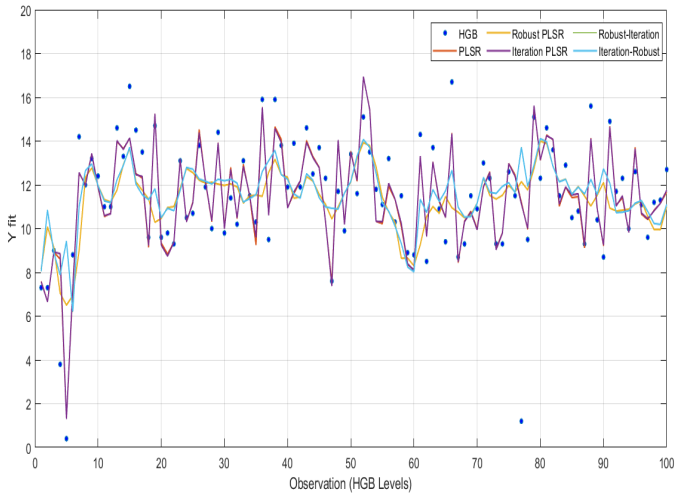

Figure 8. Estimated values for the HGB levels

Calculate variable importance in projection (VIP) scores for a PLS model. Use VIP to select predictor variables when multicollinearity exists among variables. Variables with a VIP score greater than 1 are considered important for the projection of the PLSR as in **Figure 9** which shows that there are only (6) significant Predictor variables (red points are VIP) out of a total of (19).

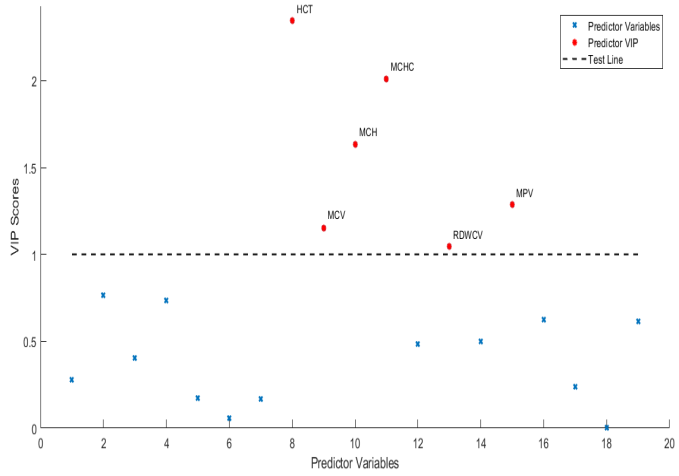

Figure 9. VIP Score for PLSR Method

The robust PLSR and proposed methods gave different results for the VIP Score as shown in the figures (10-12):

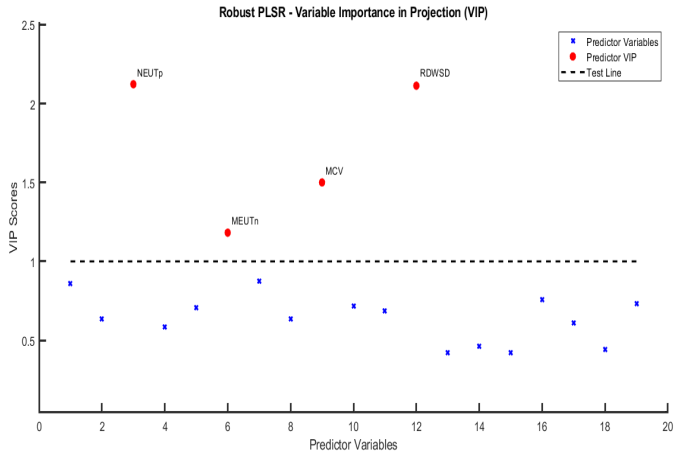

Figure 10. VIP Score for Robust PLSR Method

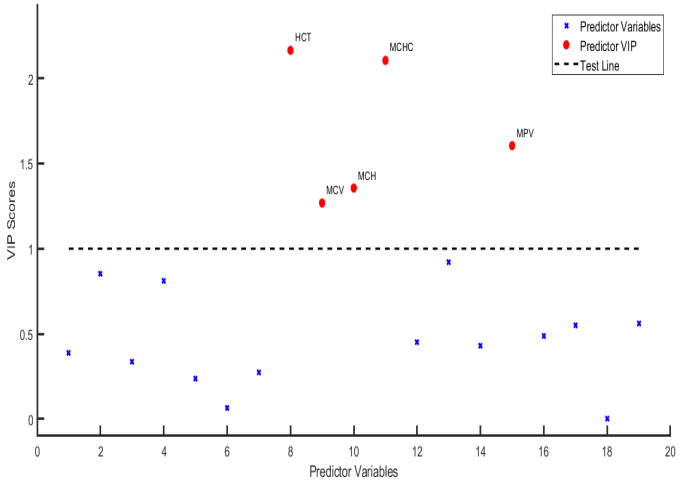

Figure 11. VIP Score for Iteration PLSR Method

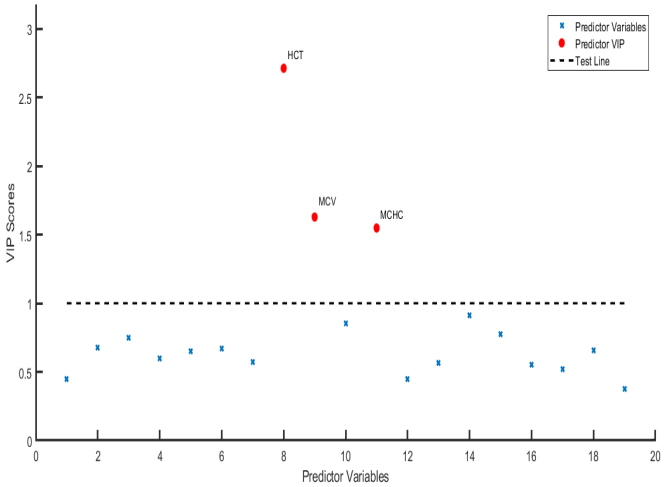

Figure 12. VIP Score for Robust-Iteration and Iteration-Robust Method

The robust PLSR method provided 4 VIPs that explain the changes in HGB levels. The proposed iterative method with higher efficiency provided 5 VIPs that explain the changes in HGB levels. The proposed methods (Robust-Iteration) and (Iteration-Robust) provided the same number of VIP (three variables) that explain the changes in HGB levels. **Table 6** shows the VIP for the five methods. Finally, the iterative method and the five tests (HTC, MCV, MCH, MCHC, and MPV) that affect the HGB levels can be relied upon.

Table 6. VIP Score for Five Methods

Method	VIP	Predictors
Classical PLSR	6	HCT, MCV, MCH, MCHC, RDWCV, and MPV
Robust PLSR	4	NEUTp, MEUTn, MCV, and RDWSD
Iteration	5	HTC, MCV, MCH, MCHC, and MPV
Robust-Iteration	3	HCT, MCV, and MCHC
Iteration-Robust	3	HCT, MCV, and MCHC

Conclusion

1. The three proposed methods address the problem of outliers in PLSR model data.
2. The proposed methods gave better results than the robust PLSR method.
3. The proposed methods provide highly efficient estimators sorted (Iteration PLSR, Robust-Iteration, and Iteration-Robust).
4. Increasing the number of principal components resulted in lower values of the MSE and increases for R^2X and R^2Y of all methods used for this data.
6. The proposed iterative method with higher efficiency provided 5 VIPs (HTC, MCV, MCH, MCHC, and MPV) that explain the changes in HGB levels.

Acknowledgement

The authors would express they're thanks to the College of Administration and Economics, University of Salahaddin University for supporting this report.

Conflict of interest

None.

References

[1] Ali, Taha Hussein, 2018, Solving Multi-collinearity Problem by Ridge and Eigenvalue Regression with Simulation, Journal of Humanity Sciences, 22.5: 262-276.

[2] Ali, Taha Hussein, Heyam Abd Al-Majeed Hayawi, and Delshad Shaker Ismael Botani. "Estimation of the bandwidth parameter in Nadaraya-Watson kernel non-parametric regression based on universal threshold level." Communications in Statistics-Simulation and Computation 52.4 (2023): 1476-1489.

[3] Ali, Taha Hussein, Avan Al-Saffar, and Sarbast Saeed Ismael. "Using Bayes weights to estimate parameters of a Gamma Regression model." Iraqi Journal of Statistical Sciences 20.1 (2023): 43-54.

[4] Ali, Taha Hussein, Heyam Abd Al-Majeed Hayawi, and Delshad Shaker Ismael Botani. "Estimation of the bandwidth parameter in Nadaraya-Watson kernel non-parametric regression based on universal threshold level." Communications in Statistics-Simulation and Computation 52.4 (2023): 1476-1489.

[5] Ali, Taha Hussein, and Saleh, Dlshad Mahmood, "Proposed Hybrid Method for Wavelet Shrinkage with Robust Multiple Linear Regression Model: With Simulation Study" QALAAI ZANIST JOURNAL 7.1 (2022): 920-937.

[6] Chen, Y., Meng, L., Zhou, H. and Xue, G., 2021. A Blockchain-Based Medical Data Sharing Mechanism with Attribute-Based Access Control and Privacy Protection. Wireless Communications and Mobile Computing, 2021(1), p.6685762. wiley.com

[7] Sullivan, J. H., Warkentin, M., & Wallace, L., 2021. So many ways to assess outliers: What works and does it matter? Journal of Business Research. [HTML]

[8] Smiti, A., 2020. A critical overview of outlier detection methods. Computer Science Review. [HTML]

[9] Zeng, N., Liu, Y., Gong, P., Hertogh, M. and König, M., 2021. Do right PLS and do PLS right: A critical review of the application of PLS-SEM in construction management research. Frontiers of Engineering Management, 8, pp.356-369. springer.com

[10] Burnett, A.C., Anderson, J., Davidson, K.J., Ely, K.S., Lamour, J., Li, Q., Morrison, B.D., Yang, D., Rogers, A. and Serbin, S.P., 2021. A best-practice guide to predicting plant traits from leaf-level hyperspectral data using partial least squares regression. Journal of Experimental Botany, 72(18), pp.6175-6189. oup.com

[11] Hair, J. & Alamer, A., 2022. Partial Least Squares Structural Equation Modeling (PLS-SEM) in second language and education research: Guidelines using an applied example. Research Methods in Applied Linguistics. researchgate.net

[12] Prager, G.W., Taieb, J., Fakih, M., Ciardiello, F., Van Cutsem, E., Elez, E., Cruz, F.M., Wyrwicz, L., Stroyakovskiy, D., Pápai, Z. and Poureau, P.G., 2023. Trifluridine–tipiracil and bevacizumab in refractory metastatic colorectal cancer. New England Journal of Medicine, 388(18), pp.1657-1667. nejm.org

[13] ALIN, A. & AGOSTINELLI, C. Robust Outlier Detection in Partial Least Squares Regression.

[14] CHAN, J. Y.-L., LEOW, S. M. H., BEA, K. T., CHENG, W. K., PHOONG, S. W., HONG, Z.-W. & CHEN, Y.-L. 2022. Mitigating the multicollinearity problem and its machine learning approach: a review. Mathematics, 10, 1283.

[15] GELADI, P. & KOWALSKI, B. R. 1986. Partial least-squares regression: a tutorial. Analytica chimica acta, 185, 1-17.

[16] HAIR, J. & ALAMER, A. 2022. Partial Least Squares Structural Equation Modeling (PLS-SEM) in second language and education research: Guidelines using an applied example. Research Methods in Applied Linguistics, 1, 100027.

[17] KıLıÇ, M., UYAR, A., KUZEY, C. & KARAMAN, A. S. 2021. Drivers and consequences of sustainability committee existence? Evidence from the hospitality and tourism industry. International Journal of Hospitality Management, 92, 102753.

[18] KNIEF, U. & FORSTMEIER, W. 2021. Violating the normality assumption may be the lesser of two evils. Behaviour Research Methods, 53, 2576-2590.

[19] MARTENS, H. 2001. Reliable and relevant modelling of real-world data: a personal account of the development of PLS regression. Chemometrics and intelligent laboratory systems, 58, 85-95.

[20] Omar, Cheman, Taha Hussien Ali, and Kameran Hassn, Using Bayes weights to remedy the heterogeneity problem of random error variance in linear models, IRAQI JOURNAL OF STATISTICAL SCIENCES 17.2 (2020): 58-67.

[21] Omer, A. W., & Ali, T. H. (2025). Dealing with the Outlier Problem in Multivariate Linear Regression Analysis Using the Hampel Filter. Kurdistan Journal of Applied Research, 10(1). https://doi.org/10.24017/science.2025.1.1.

[22] Shahla Hani Ali, Heyam A.A.Hayawi, Nazeera Sedeek K., and Taha Hussein Ali, (2023) "Predicting the Consumer price index and inflation average for the Kurdistan Region of Iraq using a dynamic model of neural networks with time series", The 7th International Conference of Union if Arab Statistician-Cairo, Egypt 8-9/3/2023:137-147.

[23] Ali, T. H., Saleh, D., Mustafa Abdulqader, Q., & Omer Ahmed, A. (2025). Comparing Methods for Estimating Gamma Distribution Parameters with Outliers Observation. Journal of Economics and Administrative Sciences, 31(145), 163-174. https://doi.org/10.33095/cc5b9h49.