Iraqi Journal of Statistical Sciences

http://stats.uomosul.edu.iq

Estimating Outliers Using the Iterative Method in Partial Least Squares Regression Analysis for Linear Models

Mahammad Mahmoud Bazid[1] Taha Hussein Ali[2]

[1,2]Department of Statistics and Informatics, College of Administration & Economics, Salahaddin University, Erbil, Iraq

Article information

Article history:
Received October 28,2024
Revised: April 25,2025
Accepted: May 8,2025
Available June 1, 2025

Keywords:
Partial Least Squares Regression, Linear Model, Outliers, Residuals, and Iterative Method.

Correspondence:
Prof. Dr Taha Hussein Ali
taha.ali@su.edu.krd

Abstract

Outliers affect the accuracy of the estimated parameters of the partial least squares regression model and give unacceptably large residual values. Traditional robust methods (used in ordinary least squares) cannot be used to treat outliers in estimating partial least squares regression model, due to the number of independent variables greater than the sample size, therefore, it was proposed to use an iterative method to treat outliers and estimation of partial least squares regression model parameters. The iterative method relies on identifying outliers and then estimating them using the initial estimated values and the residual and determining the optimal value that gives the least sum of squares error for the partial least square regression model. To illustrate the proposed method, simulated and real data were used based on a program MATLAB designed for this purpose. The proposed method provided accurate results for the partial squares regression model parameters depending on MSE criteria and addressed the problem of outliers.

DOI 10.33899/iqjoss.2025.187757 , ©Authors, 2025, College of Computer Science and Mathematics University of Mosul.

1. Introduction

Partial least squares analysis (PLS) is a statistical method used in multivariate analysis to compare many response variables with the corresponding explanatory factors. PLS is a representative statistical technique referred to as structural equation modelling. Its purpose was to address the challenges of multiple regression in cases where the data has a limited sample size, missing values, or multicollinearity (Kalivas, 1997). This method has gained immense popularity in hard science fields, particularly chemistry and chemometrics, where there is a significant issue of many correlated variables and a restricted number of observations. Wright pioneered path analysis and causal modelling in the 1920s. Partial least squares regression was first designed for econometrics but later adopted by the chemistry sector for analytical, physical, and clinical chemistry research (Pirouz, 2006). PLS is an abbreviation which initially stood for partial least squares regression, however, recently, some writers have decided to develop this term as a projection of latent structures. PLS regression integrates information from and generalizes principal component analysis (PCA) and multiple linear prediction in every scenario. Its purpose is to assess or predict a collection of dependent variables from a set of independent variables or predictors. This determination is made possible by identifying from the predictors a group of orthogonal factors termed hidden variables that have the best predictive ability (Abdi, H., 2010). PLS is still a very active study subject from a theoretical point of view; see for instance for new advances on the links of PLS with subspaces and conjugate gradients. PLS began to catch the attention of statisticians only approximately 28 years ago. This was mainly owing to the capacity of PLS to operate extremely well for data with relatively small sample sizes and many factors. Thus, it is only natural that in the past several years this paradigm has been successfully applied to difficulties in genomics and proteomics. PLS approaches are in general characterized by strong computational and statistical efficiency. They also provide significant flexibility and adaptability in terms of the analytical issues that may be handled. However, the scientific discussion of PLS is quite different because of the presence of a huge number of algorithmic versions of PLS, which made it exceedingly difficult to comprehend the principles behind PLS. This paper aims to address this hole by re-estimating and getting SSE for a new PLS model and so on until the residual is less than (0.001), giving a comprehensive review of the various PLS techniques (Boulesteix and Strimmer, 2007).

2. Partial Least Squares Regression

Partial Least Squares is a broad category of techniques used to represent relationships between sets of observable data indirectly via hidden variables. The methodology includes regression and classification tasks, along with decreasing dimension methods and modelling tools.

The fundamental premise of all Partial Least Squares (PLS) approaches is that the observable data is produced by a system or process that is influenced by a limited number of latent variables. The PLS method may be inherently expanded to regression issues. Each of the predictor and forecast (response) variables is regarded as a constituent block of variables. PLS therefore separates the score vectors that act as a newer predictor representation and goes back to the answer variables on these new predictors. The inherent imbalance among predictor and responder variables is mirrored in the method in which score vectors are produced. This variety is known under the designations of PLS1 (one response variable) and PLS2 (a minimum of two response variables). Previously disregarded by statisticians, PLS regression is still seen more as an algorithm than a serious statistical model. However, there has been a surge in interest in PLS's statistical characteristics in recent years. PLS has been linked to other regression techniques such as Ridge Regression (RR) and Principal Component Regression (PCR), and these techniques may all be grouped under a common concept known as continuum regression (Rosipal and Krämer, 2005). The partial least-squares regression approach (PLS) is gaining relevance in many sectors of chemistry; analytical, physical, clinical chemistry and industrial process control may benefit from the usage of the method. The pioneering work in PLS was done in the final years of the sixties by Wold in the discipline of econometrics. The employment of the PLS approach for chemical applications was pioneered by the groups of S. Wold and H. Martens in the late seventies following an initial application. The nonlinear iterative partial least squares (NIPALS) algorithm’s characteristics serve as the foundation for the PLS model. The data matrix may be represented by the scoring matrix. A regression between the scores for the X and Y blocks would make up a simpler model. An inner relation (connecting both blocks) and outside relations (X and Y blocks separately) can make up the PLS model. The fundamental underlying structure of multivariate PLS with l component is:

$$X = TP^T + E \tag{1}$$

$$Y = UQ^T + F \tag{2}$$

Where:

- X is a n x m predictor matrix.
- Y is a n x p response matrix.
- T and U are n x l matrices that are, as well, projectors of X (the X score, component or factor matrix) and projectors of Y (the Y scores).
- P and Q are, accordingly, m x l and p x l loading matrices.
- matrices E and F are the error terms, supposed to be independent and symmetrically distributed random normal variables.

The breakdown of Y is done to optimize the covariance between T and U.

The covariance of column *i* of T (length n) with 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 one to the 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 minimizingPFP. 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:

$$\hat{u}_h = b_h k_h \tag{3}$$

Where $b_h = u'_h t_h / t'_h t_h$. In the MLR and PCR models, theb_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 (Geladi and Kowalski, 1986).

The procedure of separating each factor of the matrix X. determines the initial vector it is multiplied by the matrix X to discover the linear structures t, that is:

$$t = XW \tag{4}$$

Where W is a vector of random values or it is the first eigenvector equal to the first eigenvalue for the matrix ($X'YY'X$), in the same manner, the linear combination of the matrix Y, which is with vector U as follows:

$$U = YC \tag{5}$$

You may determine that ($X'Y$) is the covariance matrix between X and Y if C is a vector of random values or the first eigenvector corresponding to the first eigenvalue for the matrix ($X'YY'X$). This approach looks for a collection of elements known as latent vectors, which attempt to explain as much of the covariance between X and Y as they can. The following equation may be used to analyze the independent variables:

$$X = TP^t \tag{6}$$

Whereas (P) is a loaded vector and is a linear combination between the orthogonal factors t and the original matrix of the independent variables, that is: where (T) is a linear combination of predictive variables, but in the form of orthogonal factors, that is, each column contains all the independent variables present in (X), but in the form of a linear combination of weights.

$$P = X't \tag{7}$$

In the matrix (T), are represented by the columns (t). $T'T = I \quad and \quad P'P = 1$

The process is repeated after determining the first eigenvector, which is then subtracted from both X and Y. Continue until X turns into a zero matrix (Omer et al. 2024).

3. Outliers

In the context of model construction, outliers refer to data points (vectors) that deviate significantly from the rest of the data, therefore causing substantial distortion in the obtained finding. Every extensive dataset usually has outliers that must be detected and removed from the training set before the modelling process. The basic techniques of trimming the data eliminate most or all the major outliers and should consequently be an essential step in the preparation of large data sets. The techniques require the ranking of each variable and removing it or changing a tiny fraction of the extreme values of the variable. Normally this proportion is round one and five. Note that only the extreme parts of one variable have been changed at each step the complete observation is never eliminated. With trimming, the most extreme components are set to missing, and thus 2 percentage and 10 percentage missing data are interested in the data, which typically has no detrimental effect on future data analysis. the most extreme components are instead assigned a value nearer to the mean, frequently 3% points (calculated in a rigorous method), or the last acceptable value in data processing (Kettaneh and Wold, 2005).

3.1. Outliers' Effect on Parameter Estimation

Outliers may have a significant influence on statistical studies, particularly those that depend on the assumption of normally distributed data or that are sensitive to high values (Ali, 2017).

- **Bias parameter estimates:** Outliers may affect the mean, inflating or deflating the estimate of central tendency. For example, in regression models, outliers may significantly alter both the slope and intercept.
- **Increase variability:** Outliers may artificially inflate standard deviation and variance estimates, creating the mistaken impression that the data is more widely distributed.
- **Influence hypothesis testing:** Because skewed test results may result in incorrect conclusions, such as the inaccurate rejection or acceptance of null hypotheses, outliers can have this effect.
- **Distort model predictions:** Outliers have a disproportionately large impact on predictions in machine learning and regression models, which makes it harder for the models to generalize to fresh data.

3.2. Finding and Classifying Outliers:

Finding observations that differ considerably from the rest of the data requires using statistical techniques to identify outliers. Several methods may be applied (Ali et al. 2024):

3.3. Methods of Statistics and Visualization:

- **Z-scores**, also known as standard scores, are used to express how much a data point deviates from the mean in standard deviations. Values that deviate more than three standard deviations from the mean are usually regarded as outliers.

$$Z = \frac{X-\mu}{\sigma} \tag{8}$$

Where the data point is X. μ represents the mean and σ represents the standard deviation.

- **Boxplots:** The interquartile range (IQR) is shown as a boxplot, with outliers being identified as points that are more than 1.5 times the IQR from the quartiles.
- **Scatterplots:** Outliers in multivariate data may be visually identified via scatterplots, sometimes called pair plots, as points that are distant from data clusters.
- **Histograms:** Outliers may be identified as solitary bars at extreme values by using the frequency distribution of the data to visualize them.
- **Cook's Distance:** This regression analysis metric gauges a data point's impact on the regression parameters. Large Cook's distance points represent significant observations or possible outliers (Kareem et al. 2019).

3.4. Strategies for Handling Outliers:

Outliers are treated differently depending on their cause and context once they have been recognized. Typical methods include:

I. Eliminating Outliers:

- **Case by case:** Removal of outliers on a case-by-case basis is possible if they result from measurement mistakes, data entry problems, or abnormalities unrelated to the research.
- **Automated removal:** Methodically eliminating outliers using statistical criteria like IQR, Z-scores, or visual examination (like boxplots). Removing data should be done carefully, however, to prevent losing crucial information (.

II. Data transformation:

- **Logarithmic transformation:** By condensing the range of extreme values, log transformation may lessen the effect of significant outliers (Omer et al. 2024).
- **Square root transformation:** This method is effective with somewhat skewed data and is comparable to log transformation.

- **Winsorizing:** This technique reduces the impact of outliers without eliminating them by substituting them with the closest non-outlier value or a predetermined threshold.

III. robust Estimation Techniques: These techniques exhibit reduced sensitivity to outliers in contrast to conventional parametric techniques:

- **Robust regression:** reduces the impact of outliers on parameter estimations by using techniques like Huber regression or Least Absolute Deviations (Ali et al. 2022).
- **Median-based estimators:** The median may be substituted for the mean in robust parameter estimation since it is less impacted by extreme values than the mean.

V. Cutting or Capping: In cases where outliers are expected (such as in finance with extreme returns), they can be capped at a certain threshold. For example, setting extreme values to the 1st and 99th percentiles ensures that no value lies beyond these points (Ali. 2018).

VI. Making Use of Models with Inbuilt Outlier Resistance: Certain algorithms are more resilient to outliers by nature, like:

- **Random Forests:** Individual outliers have less of an impact on decision trees in random forests since the trees are constructed using subsets of the data.
- **Ridge and Lasso Regression:** By decreasing coefficients, regularization approaches like Lasso (L_1) and Ridge (L_2) regression may lessen the impact of outliers (Cousineau and Chartier, 2010).

4. Mean square error and Coefficient of Determination:

Outliers may be limited to a certain level in situations where they are anticipated, such as in finance with extraordinary returns. To guarantee that no value is beyond the first and 99th percentiles, for instance, extreme values should be assigned to these locations (Ali et al. 2023). When the MSE is lower, the model's predictions are more accurate than the real data. MSE is always positive; therefore, a value near zero is preferable. larger mistakes are penalized more than smaller ones since they square the discrepancies.

$$\mathrm{MSE} = \frac{1}{m}\sum_{j=1}^{m}\left(y_j - \hat{y}_j\right)^2 \qquad \mathbf{(9)}$$

- y_j : The real data.
- $\hat{y}_j$: The estimated data.
- m : Number of observations.

Coefficient of Determination R^2 quantifies the percentage of the dependent variable's variation that can be predicted based on the independent variables. It provides insight into how well the regression model matches the available data. R^2 has a range of 0 to 1. A model that fully explains the variance in the data and accurately predicts it has an R^2 of 1. When a model's R^2 is zero, it is as poor at explaining variance in data as it is at forecasting the data mean. When the model performs worse than just forecasting the mean, negative R^2 values might appear (Ali et al. 2023).

$$R^2 = 1 - \frac{SS_{residual}}{SS_{general}} \qquad (10)$$

$SS_{residual}$The total error may be defined as the sum of squared discrepancies between the expected and actual values.

$SS_{general}$the total squared discrepancies between the mean of the real values and their actual values.

5. Proposed Method

The proposed method for treating outliers is summarized in the following steps:

- Identifying outliers y(o) from the standard residuals of a partial least squares regression model that are outside an interval (∓2.5).
- Calculate the initial Sum of Squares Error (ISSE) of the model from the following formula:

$$ISSE = SSE = Residuals^{T} \times Residuals \tag{11}$$

The residuals from the classical PLS model.

- Estimate outliers using the following equation:

$$Y(o) = \hat{y}(o) - Residual(o) \tag{12}$$

Residual (o) is the outlier residual, using Y(o) to estimate a PLS model and compute SSE.

- Calculate the absolute difference E between the SSE of the previous and current.
- If the residual value is greater than (0.001), then the outlier in equation (12) will be re-estimated and get SSE for a new PLS model and so on until the residual is less than (0.001).
- Finally, the estimated values of the outliers that give the least sum of squares error are used to get the PLS model.

6. Simulation Study

To measure the efficiency of the proposed method, and its handling of outliers in partial least squares regression analysis, the simulation study. The spectral and octane data of the gasoline experiment (Kalivas, 1997, p. 256) were used in the simulation by using the estimated parameters as the assumed parameter values (402 parameters for 401 independent variables), and (60) observations. This experiment was simulated with a random error having a standard normal distribution for several different sample sizes (10, 20, …, 60) and different numbers of independent variables (15, 20, 30, …, 180). Also, two outliers were added to the dependent variable and the experiment was repeated a thousand times. The MSE average of the PLSR models for classical and proposed methods with the average number of times the outlier is estimated is summarized in Tables 1 and 2 based on five and ten factors.

Table 1. The MSE average for 5 Factors

Method	Sample Size	Number of Independent Variables	N	MSE
Classical	10	15	-----	90.6833
Proposed			86	**1.6526**
Classical		20	-----	86.1192
Proposed			62	**1.6522**
Classical	20	30	-----	617.9439
Proposed			37	**9.9567**
Classical		40	-----	153.1791
Proposed			40	10.2957
Classical	30	50	-----	268.0350
Proposed			21	19.3367
Classical		60	-----	290.7048
Proposed			21	18.9583
Classical	40	70	-----	43.0301
Proposed			13	28.8203
Classical		80	-----	42.7187
Proposed			13	28.5590
Classical	50	90	-----	185.0618
Proposed			353	37.2553
Classical		100	-----	127.9305
Proposed			33	36.3577

Table 2. The MSE average for 10 Factors

Method	Sample Size	Number of Independent Variables	N	MSE
Classical	20	40	-----	2.8818
Proposed			100	**1.6498**
Classical		60	-----	3.5523

Proposed			91	**1.3549**
Classical	30	70	-----	10.2435
Proposed			78	**5.2328**
Classical		90	-----	14.5009
Proposed			123	**5.0256**
Classical	40	80	-----	70.6633
Proposed			294	**9.9960**
Classical		120	-----	33.5776
Proposed			34	**8.9680**
Classical	50	100	-----	559.7075
Proposed			49	**14.8587**
Classical		150	-----	57.0230
Proposed			63	**22.5774**
Classical	60	120	-----	142.7310
Proposed			17	**19.8851**
Classical		180	-----	50.2524
Proposed			127	**34.5525**

7. Simulation Results Discussing

For all simulation cases, the proposed method was highly efficient in handling the outlier problem, estimating the parameters of the PLSR model, and minimum the model residuals based on the efficiency criterion of MSE (it has the lowest MSE compared to the classical method).

The number of times the outliers were estimated ranged between (13-353) times and was lowest at sample size (40) and number of independent variables (70 and 80) and highest at sample size (50) and number of independent variables (90), Sometimes, but not always, the repeat number of outlier estimates decreases as the number of observations and independent variables in the model increases. The MSE increases with the number of observations and independent variables. The MSE values decrease as the number of factors increases in the PLS model.

The traditional method was greatly affected by the presence of outliers and presented a large MSE for the PLS model which cannot be handled using robust methods due to the presence of a large number of independent variables greater than the number of observations, the proposed method presented minimum MSE, the estimated values of the outliers were very close to their true values, fewer important variables were included in the model, and there was a significant increase in the coefficient of determination, which will be discussed in detail in the real data.

8. Real Data

The real data represents the spectral and octane data of gasoline (Kalivas, 1997, p. 256) so the predictor X is a numeric matrix that contains the near-infrared (NIR) spectral intensities of 60 samples of gasoline at 401 wavelengths (number of independent variables). The response y is a numeric vector that contains the corresponding octane ratings (with X data).

Two outliers were added to the dependent variable data, and the traditional method (without processing outliers) was applied. Ten factors were used to obtain the largest cumulative percentage of explanation of the total variance as in Figure 1:

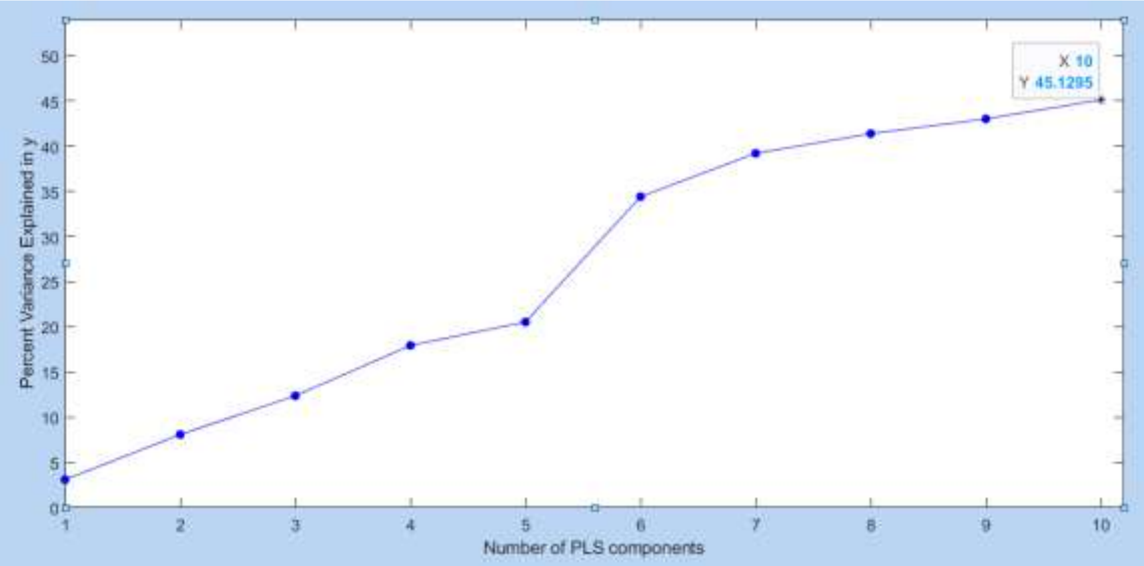

Figure 1. Cumulative Percentage of explanation of the total variance for Classical Method

Figure 1 shows that ten factors explain only 45.1295% of the total variance in the response variable (octane ratings) and represent determination coefficients R^2 in the PLSR model, which is an unacceptably low percentage due to outliers. Compute the fitted response variable and display the residuals in Figures 2 and 3, respectively.

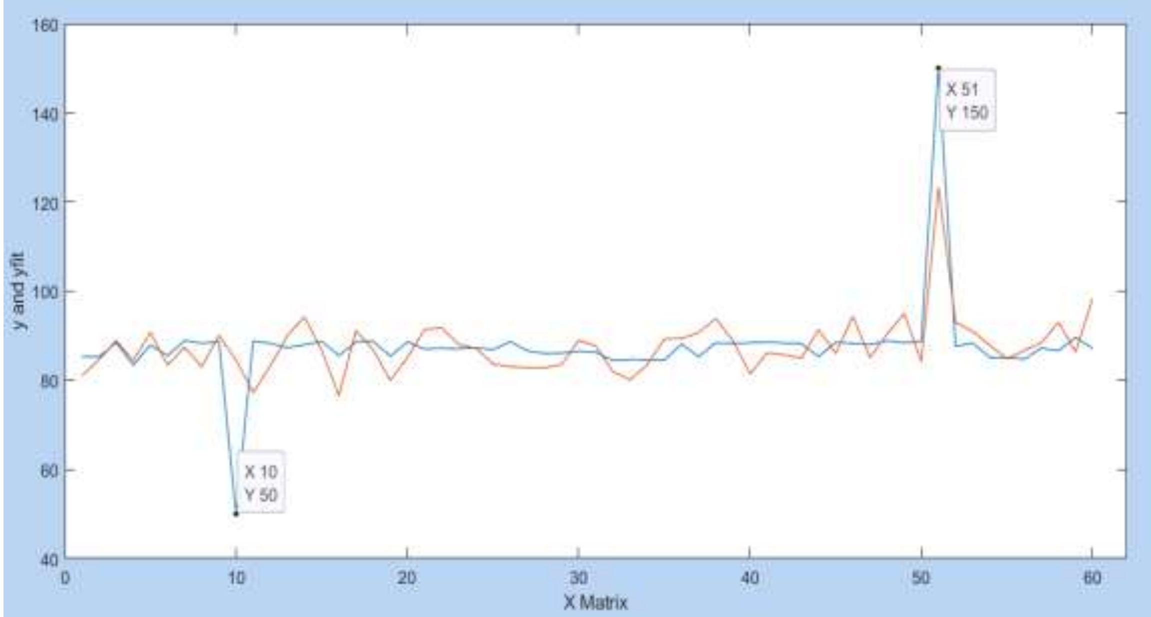

Figure 2. Fitted Response Variable for Classical Method

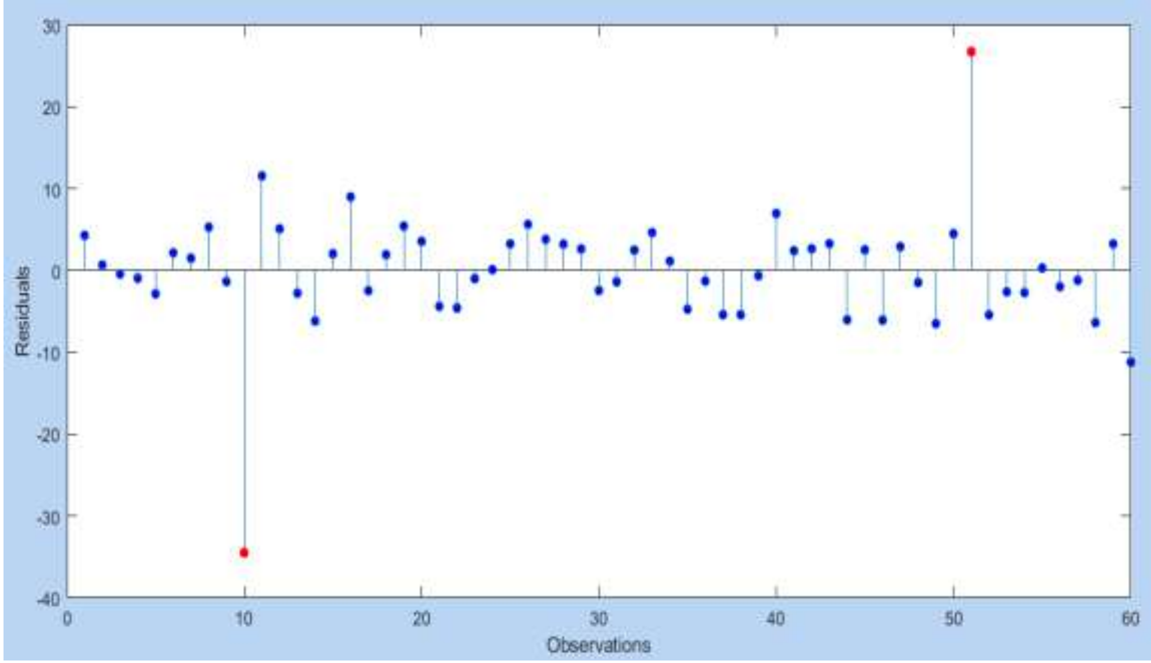

Figure 3. Residuals PLS Model for Classical Method

Figure 2 shows the actual response values (blue line) and the estimated values (red line) from the PLSR model. The estimated values were affected by outliers (y_{10} = 50 and y_{51}= 150) and were unacceptably bad. Figure 3 shows the large and unacceptable residual values, especially Residual_{10} and Residual_{51} marked in red, with the sum of squared errors equal to (2993.4).

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 4 which shows that there are only (40) significant independent variables (red points are VIP) out of a total of (401).

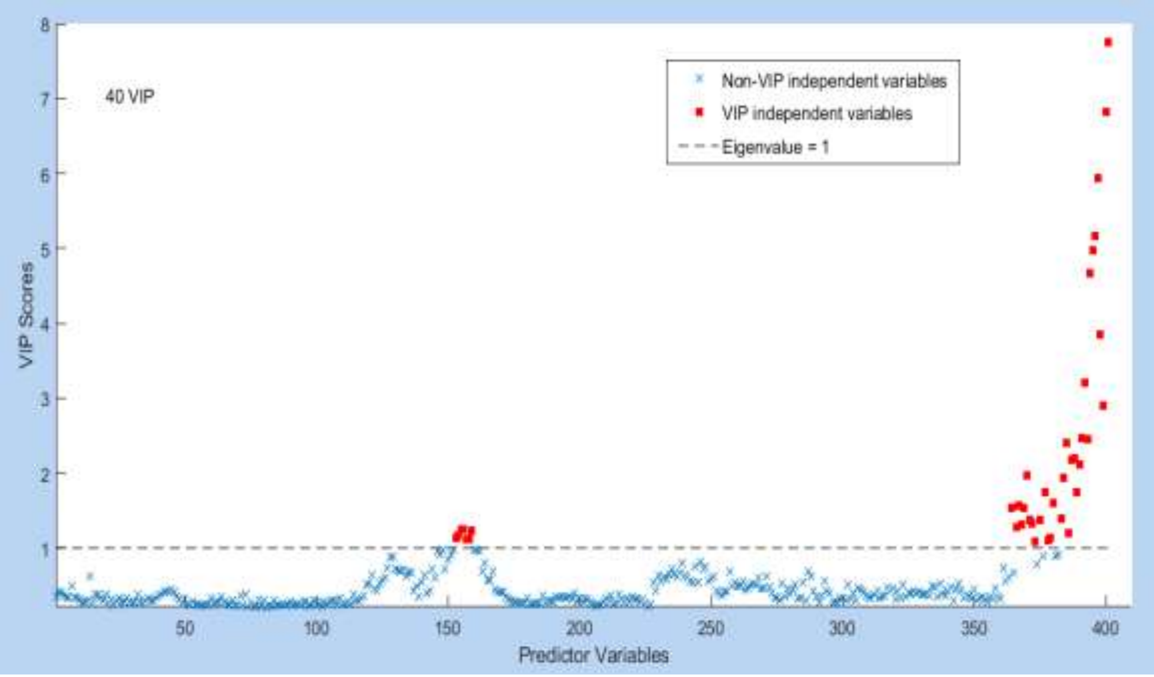

Figure 4. VIP Independent Variables Classical Method

The proposed iterative method depends on estimating outliers. This is done first by identifying outliers based on the residuals of the PLS model as in Figure 3. The two values (y_{10} and y_{51}) are considered outliers thus they will be estimated using the proposed method. When the estimation of outliers was repeated (17) times, the iterative method provided the lowest sum of squared errors (1.0121) for the PLS model when y_{10} = 88.6852 and y_{51} = 88.0356.

Figure 5 shows the actual response values (blue line) and the estimated values (red line) from the PLS model. The estimated values were unaffected by outliers and were acceptably good. Figure 6 shows the small and acceptable residual values (-0.3-0.3), compared with the classical method (-38-28).

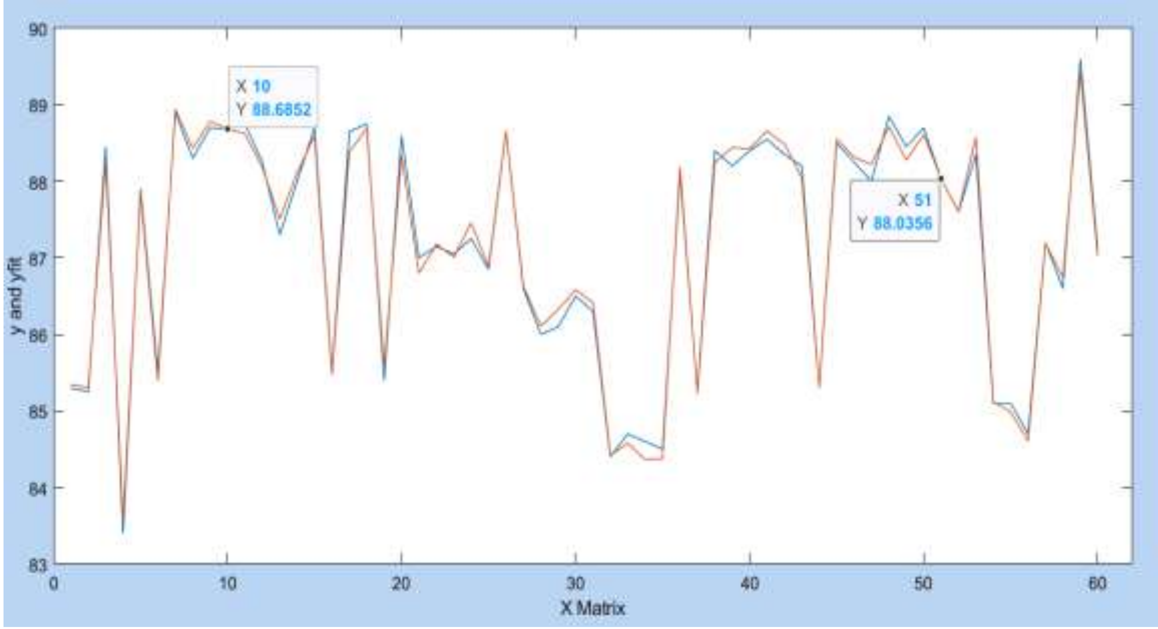

Figure 5. Fitted Response Variable for Proposed Method

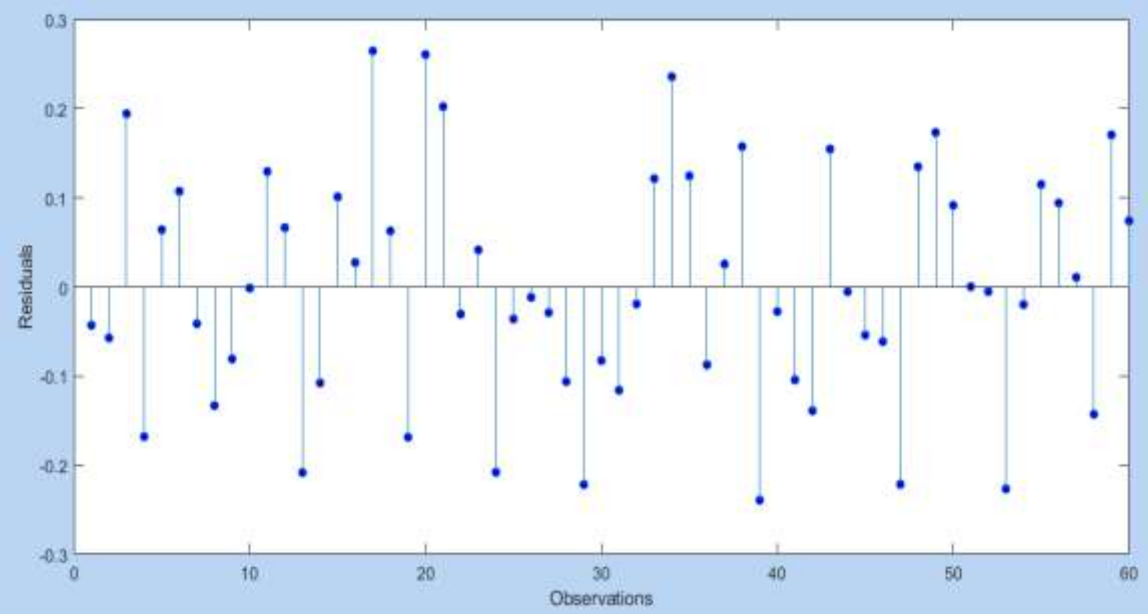

Figure 6. Residuals PLS Model for Proposed Method

Figure 7 shows that ten factors explain 99.2701% of the total variance in the response variable (octane ratings) and represent determination coefficients R^2 in the PLS model, which is an acceptably high percentage compared with the classical method of 45.1295%.

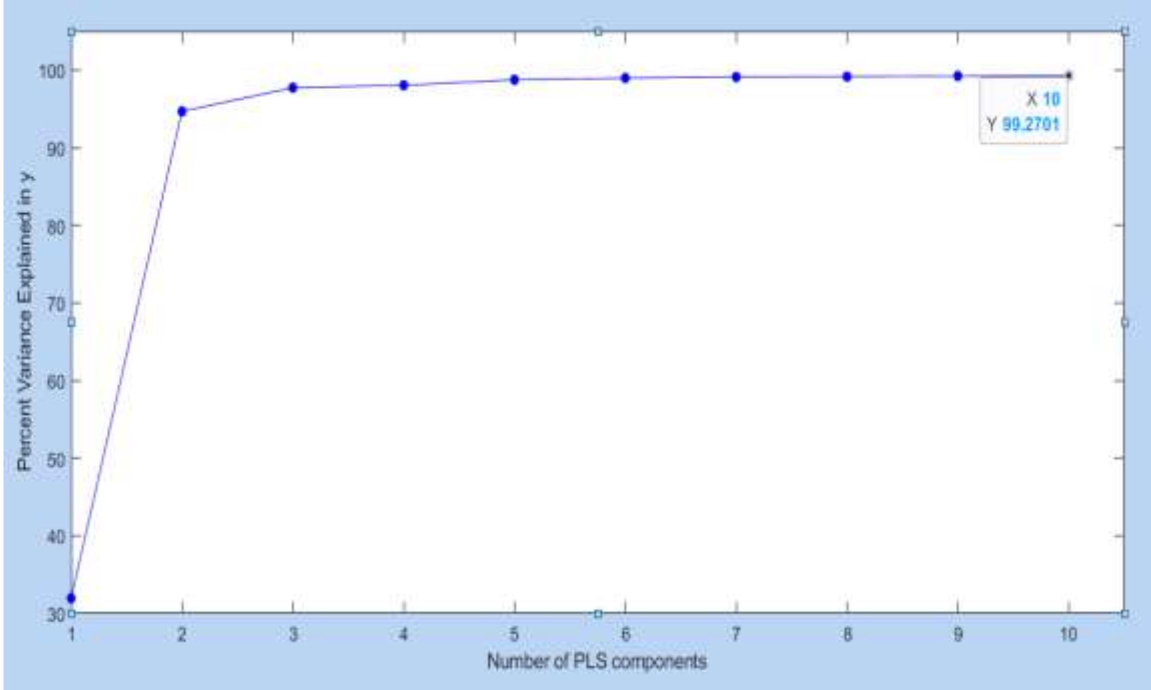

Figure 7. Cumulative Percentage of explanation of the total variance for the Proposed Method

Calculate variable importance in projection (VIP) scores for a PLS model as in Figure 8 which shows that there are (91) significant independent variables (red points are VIP) out of a total of (401). It is completely identical to the results of the traditional method before adding outliers.

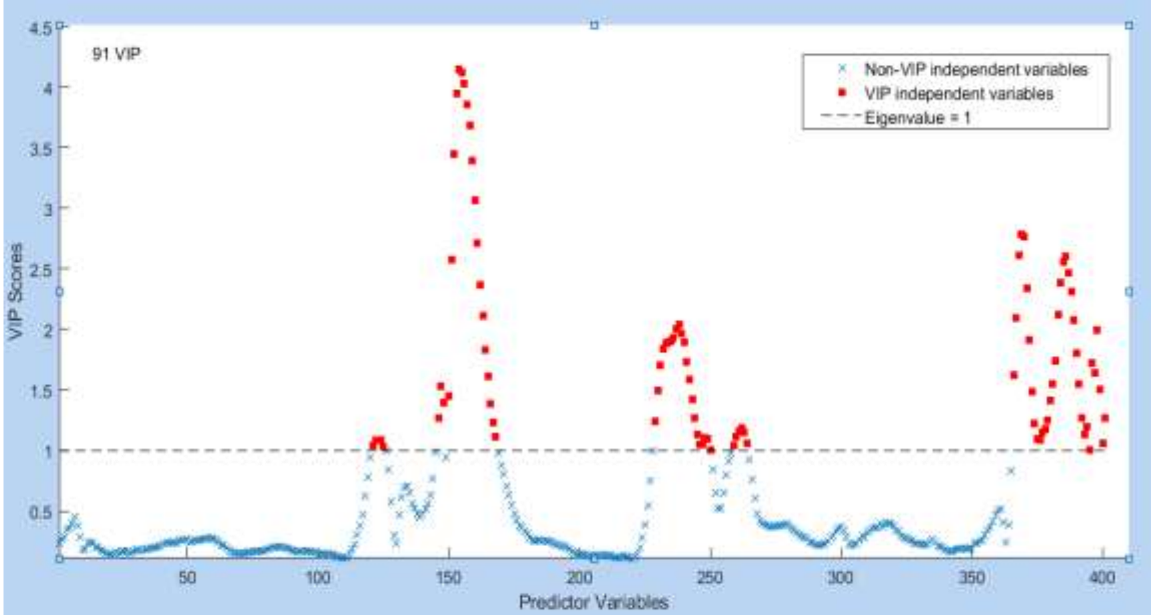

Figure 8. VIP Independent Variables Classical Method

9. Conclusions

1. For all simulation cases, the iterative method provided highly efficient outlier handling with the lowest sum of squares of the residuals of the PLSR model.
2. The proposed method was more efficient than the traditional method depending on the MSE.
3. The MSE increases with the number of observations and independent variables.

4. The MSE values decrease as the number of factors increases in the PLSR model.
5. For real data, the proposed method provided a lower MSE, max VIP, and max explanation ratio (coefficient of determination), compared with the classical method.

10. Recommendations

1. Using the proposed method to address the problem of outliers when estimating a PLSR model.
2. Comparison of the proposed method with wavelet shrinkage methods for handling data noise and outliers.
3. Conduct a future study on handling outliers using modified robust methods consistent with the PLSR model's assumptions.

References

1. Kalivas, John H., "Two Data Sets of Near Infrared Spectra," Chemometrics and Intelligent Laboratory Systems, v.37 (1997) pp.255–259.
2. Pirouz, D. M. (2006). An Overview of Partial Least Squares. Irvine: Business Publications.
3. Abdi, H., (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). Wiley Interdisciplinary Reviews: computational statistics, 2(1), pp.97-106.
4. Boulesteix, A.L. and Strimmer, K., (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in bioinformatics, 8(1), pp.32-44.
5. Rosipal, R. and Krämer, N., (2005), February. Overview and recent advances in partial least squares. In International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection" (pp. 34-51). Berlin, Heidelberg: Springer Berlin Heidelberg.
6. Geladi, P. and Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Analytica chimica acta, 185, pp.1-17.
7. Omer, A. W., Sedeeq, B. S., & Ali, T. H. (2024). A proposed hybrid method for Multivariate Linear Regression Model and Multivariate Wavelets (Simulation study). Polytechnic Journal of Humanities and Social Sciences, 5(1), 112-124. https://doi.org/10.25156/ptjhss.v5n1y2024.pp112-124
8. Kettaneh, N., Berglund, A. and Wold, S., (2005). PCA and PLS with very large data sets. *Computational Statistics & Data Analysis*, *48*(1), pp.69-85.
9. Ali, Taha Hussein & Awaz Shahab M. "Uses of Waveshrink in Detection and Treatment of Outlier Values in Linear Regression Analysis and Comparison with Some Robust Methods", Journal of Humanity Sciences 21.5 (2017): 38-61.
10. Ali, T. H., Sedeeq, B. S., Saleh, D. M., & Rahim, A. G. (2024). Robust multivariate quality control charts for enhanced variability monitoring. Quality and Reliability Engineering International, 40(3), 1369-1381. https://doi.org/10.1002/qre.3472
11. Kareem, Nazeera Sedeek and Mohammad, Awaz Shahab, and Ali, Taha Hussein, "Construction robust simple linear regression profile Monitoring" Journal of Kirkuk University for Administrative and Economic Sciences, 9.1. (2019): 242-257.
12. Omer, A. W., Sedeeq, B. S., & Ali, T. H. (2024). A proposed hybrid method for Multivariate Linear Regression Model and Multivariate Wavelets (Simulation study). Polytechnic Journal of Humanities and Social Sciences, 5(1), 112-124. https://doi.org/10.25156/ptjhss.v5n1y2024.pp112-124
13. 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.
14. Ali, Taha Hussein, 2018, Solving Multi-collinearity Problem by Ridge and Eigenvalue Regression with Simulation, Journal of Humanity Sciences, 22.5: 262-276.
15. Cousineau D., Chartier, S. (2010). Outliers' detection and treatment: a review. International Journal of Psychological Research,3(1), 58-67.58
16. 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.
17. Ali, Taha Hussein, Nasradeen Haj Salih Albarwari, and Diyar Lazgeen Ramadhan. "Using the hybrid proposed method for Quantile Regression and Multivariate Wavelet in estimating the linear model parameters." Iraqi Journal of Statistical Sciences 20.1 (2023): 9-24.

تقدير القيم الشاذة باستخدام الطريقة التكرارية في تحليل الانحدار الجزئي للمربعات الصغرى للنماذج الخطية

محمد محمود بازيد[1] و طه حسين علي[2]

[2,1]قسم الإحصاء والمعلوماتية، كلية الإدارة والاقتصاد، جامعة صلاح الدين، أربيل، العراق.

الخلاصة: تؤثر القيم الشاذة على دقة المعلمات المقدرة لنموذج الانحدار الجزئي للمربعات الصغرى وتعطي قيم بولقي كبيرة بشكل غير مقبول. لا يمكن استخدام الطرق التقليدية الحصينة (المستخدمة في المربعات الصغرى العادية) لمعالجة القيم الشاذة في تقدير نموذج الانحدار الجزئي للمربعات الصغرى، وذلك بسبب عدد المتغيرات المستقلة الأكبر من حجم العينة، لذلك، تم اقتراح استخدام طريقة تكرارية لمعالجة القيم الشاذة وتقدير معلمات نموذج الانحدار الجزئي للمربعات الصغرى. تعتمد الطريقة التكرارية على تحديد القيم الشاذة ومن ثم تقديرها باستخدام القيم المقدرة الأولية والبواقي وتحديد القيمة المثلى التي تعطي أقل خطأ لمجموع المربعات لنموذج الانحدار الجزئي للمربعات الصغرى. لتوضيح الطريقة المقترحة، تم استخدام بيانات محاكاة وحقيقية بناءً على برنامج MATLAB المصمم لهذا الغرض. قدمت الطريقة المقترحة نتائج دقيقة لمعلمات نموذج الانحدار الجزئي للمربعات الصغرى اعتمادًا على معايير MSE وعالجت مشكلة القيم الشاذة.

الكلمات المفتاحية: الانحدار الجزئي للمربعات الصغرى، النموذج الخطي، القيم الشاذة، البواقي، والطريقة التكرارية.