Partial Robust M-Regression Estimator in the Presence of Multicollinearity and Vertical Outliers

*Nur Hidayah Md Noh¹, Balkiah Moktar², Sarah Yusoff³, Mohd Noh Ab Majid⁴
¹,³Faculty of Computer and Mathematical Science, Universiti Teknologi MARA, 21080, Kuala Terengganu, Malaysia
²Faculty of Computer and Mathematical Science, Universiti Teknologi MARA, Perlis, Malaysia
⁴Faculty of Business Management, Universiti Teknologi MARA, 23000 Dungun Terengganu, Malaysia

*nurhidayah0738@uitm.edu.my

Abstract. The objective of using regression is to explain the variation in one or more response variables by associating the variation with proportional variation in one or more explanatory variables. However, if the number of independent variables is multiple, they tend to be highly collinear and this contributes to multicollinearity problem. For instance, Ridge Regression (RR), Principal Component Regression (PCR), and Partial Least Squares Regression (PLSR) are some of the prediction methods used to handle dataset with multicollinearity. In addition, another problem that arises is the existence of outlying objects in a dataset. The effect of outlying data points in the presence of multicollinearity problem could be reduced with the implementation of robust regression method. A recently studied robust PLSR, which is called Partial Robust M-Regression (PRM) is found to be able in dealing with multicollinearity and outliers simultaneously. This method was employed in this study. Throughout this study, five methods of regression were chosen; OLS, RR, PCR, PLSR, and PRM, to compare which is the best method in their predictive ability. To compare these five regression methods, a simulation study had been conducted, and the mean square error of \(\beta\) estimate (MSE(\(\beta\))) was calculated. The simulation results show that PRM outperforms other method in the presence of multicollinearity, outliers and leverage points.

1. Introduction

The Ordinary Least Squares (OLS) regression method is commonly used due to its simple way of computation but, their estimation of regression weights in multiple regression are affected by multicollinearity, outliers, non-normality and missing data [1]. Multicollinearity is a condition where two or more variables in the data are redundant and shares the same information [2]. It is a problem in regression analysis caused by high correlation among independent variables. High multicollinearity increases possibility of predictor variable to be rejected from the regression model [3] and may yield unstable results due to the increased in standard error of the estimated coefficients [4]. Few methods had been proposed to cope with multicollinearity problem, popularly called biased regression methods. Among such are Principle Component Regression (PCR) [5], Ridge Regression (RR) [6], and Partial Least Square Regression (PLSR) [7].
Al-hassan and Al-Kassab [8] compared Ridge Regression (RR) and Principal Components Regression (PCR) by using Monte Carlo simulation study in order to deal multicollinearity problem between explanatory variables and it was concluded that RR performed better than PCR due to smaller mean square error (MSE). Ordinary Least Squares (OLS) regression, Principle Component Regression (PCR), Partial Least Squares Regression (PLSR) and Variable Subset Selection (VSS) in dealing the problem in various degree of multicollinearity was compared and it was found that when multicollinearity increase, RR performs best closely followed by PLS and PCR whereas OLS performs worst [9].

Rougoor, Sundaram, and van Arendonk [10] performed a study on the advantages and disadvantages of PLSR and PCR in the use livestock management where the relation between breeding management and 305 days milk production was investigated. The data set was a high dimensional data (p>n) and presented multicollinearity. It was found that in a high dimensional data with multicollinearity problem, PLSR is a good alternative of PCR. The predictive ability of PLSR was compared by Yeniey and Goktas [11] among OLS regression, PCR and RR on the gross product per capita (GDDPC) in Turkey data set. The data set contains high collinearity. It was concluded that PLS yields better results in predicting the model with smallest number of factors when compared to other prediction methods. Ahmad, Adnan and Ahmad [2] compared the performances of RR, PCR and PLSR to deal multicollinearity problem through a simulation study. Mean Square Error (MSE) were compared and to sum up, PLSR and RR performed better than PCR with slight difference. Farahani, Rahiminiezhad and Same [12] compared OLS regression and PLSR in a study done to predict couple’s mental health and conclude that PLSR model gives stable results in dealing with small missing values, sample size and multicollinearity.

RR, PCR and PLSR was compared by Ramzan, Zahid and Ramzan [13] in a study of economic growth in Pakistan involving time series data with high level of multicollinearity. Root Mean Square Error (RMSE) and Root Mean Square Error of cross-validation (RMSECV) was measured in fitting regression model predict the model’s ability and it was concluded that PLSR is the best model in predicting the time series data that displays high multicollinearity.

Another factor that effects OLS regression and can change the magnitude of regression coefficients also even the coefficient signs is by the presence of outliers. In linear regression, an outlier is an observation with large residual. Wainer [14] defined outliers as contaminants. An outlier is considered a data point that is far beyond the norm for a variable or population [15], [16]. Outliers may increase error rates and significant errors of statistical estimator when employing either parametric or nonparametric tests [17]. There are three classes of outlier problem that have been identified relating to regression analysis which are (1) problems with outliers in the y-direction (response direction), (2) problems with multivariate outliers in the covariate space (i.e. outliers in the X-space, which are also referred to as leverage points) and (3) problems with outliers in both the y-direction and the X-space [18].

In the existence of multicollinearity and outliers, Partial Robust M-Regression (PRM) introduced by Serneels, Christoph, Filzmoser, and Espen [19] is an alternative to non-robust regression and will be employed in this study. PRM is a robustified version of the partial least square approach. Liebmann, Filzmoser and Varmuza [20] compared the performance of the SIMPLS approach with the partial robust M regression (PRM). Both methods were employed to three different data sets including outliers intentionally created. PRM outperforms the classical PLS models in the presence of outlying observations in the data. For clean data, the prediction performance of both the classical and the robust model are in the same range. Therefore, the goal of this paper is to compare the performance of Partial Robust M-Regression with Partial Least Square Regression, Principal Component Regression, Ridge Regression and Ordinary Least Square Regression in predicting data with multicollinearity and vertical outliers.
2. Methodology

Among regression methods that can deal multicollinearity problem are Ridge Regression, Principle Component Regression and Partial Least Squares Regression but the performance are affected by outliers. Hence, this paper focuses on the robust method which can deal with both multicollinearity and outliers in regression analysis, the robustified version of Partial Least Squares Regression namely Partial Robust M-Regression. OLS regression, PCR, RR and PLS are briefly outlined while PRM is presented in more detailed. The regression model used for the above methods is defined by the equation,

\[ Y = X\beta + \varepsilon \]  

Where \( Y \) is a \( n \times q \) matrix of observations on \( q \) dependent variables \( y_1, y_2, ..., y_q \), \( \varepsilon \) is a \( n \times q \) matrix of errors, whose rows are independently and identically distributed, and \( \beta \) is a \( p \times q \) matrix of parameters to be estimated.

2.1 Ordinary Least Squares Regression

Generally, the Ordinary Least Squares (OLS) regression is used to estimate the parameter values with the objective function to minimize the sum of squared residuals,

\[ \hat{e}^T \hat{e} = (y - X\hat{\beta})^T(y - X\hat{\beta}) \]  

Hence the estimation of \( \beta \) becomes

\[ B_{OLS} = (X^TX)^{-1}X^TY \]  

However, this method is not powerful when the problem of multicollinearity arises. When the independent variables are highly correlated, \( XX' \) have ill-conditioned and resulting in large value of variance of OLS estimator.

2.2 Principal Components Regression

Principal Component Regression is a biased estimation technique in handling multicollinearity and was developed by Massy [5] to deal the problem of multicollinearity by reducing the variances of the regression coefficients. It performs least squares estimation on a set of new variables called the Principal Components of the correlation matrix. The results in estimation and prediction are better than the ordinary least squares. PCR basically obtain the number of principle components providing the maximum variation of \( X \) which optimizes the predictive ability of the model. PCR consist the linear regression between the scores and the response matrix \( Y \), which is modeled by

\[ Y = TC + \varepsilon = XPC + \varepsilon = X\beta + \varepsilon \]  

\( T = XP \), where \( T \) is the score matrix and \( P \) is the loading matrix. The regression coefficient is given by

\[ \beta_{PCR} = P(T^T)^{-1}T^TY \varepsilon \]  

2.3 Ridge Regression

Ridge Regression which was introduced by Hoerl and Kennard [6] is a popular method in dealing with multicollinearity. It is a modification of ordinary least squares method that allows biased estimators of the regression coefficients and the estimator function is given by
\[ \beta_{\text{RIDGE}} = (X^T X + k I)^{-1} X^T Y + \varepsilon \]  

\[ \text{where } I \text{ is the pxp identity matrix and } k \text{ is biasing constant. When } k = 0, \beta_{\text{RIDGE}} = \beta_{\text{OLS}}, \text{ when } k > 0, \beta_{\text{RIDGE}} \text{ is more biased but more stable and precise than OLS and when } k \rightarrow 0, \beta_{\text{RIDGE}} \rightarrow 0 \text{ and there is always exist the value of } k > 0 \text{ which produce } \text{MSE}_{\beta_{\text{RIDGE}}} < \text{MSE}_{\beta_{\text{OLS}}}. \]

2.4 Partial Least Squares Regression

The purpose of Partial Least Squares Regression is to describe the relationship between response variable, Y, and a set of explanatory variables, X with the constraint that these components explain as much as possible of the covariance between covariates X and response variable, Y, by replacing the original variables with a few orthogonal latent factors, T, with maximal covariance with y [21]. The PLSR helps deal with multicollinearity problem [7]. The PLSR model uses a factor score matrix T = XP for an appropriate weight matrix P in the regression model

\[ Y = TQ + \varepsilon = X\beta + \varepsilon \]  

\[ \text{where } Q \text{ is a matrix of loadings of the regression coefficients and } \varepsilon \text{ is the error term. The regression coefficients, } \beta, \text{ are obtained as} \]

\[ \beta_{\text{PLS}} = PQ \]  

The algorithm of PCR and PLSR in detailed by Naes and Martens [22].

2.5 Partial Robust M-Regression

The algorithm used by Serneels et.al [19] to accomplish PRM regression uses robust starting values and making the weights depend not only on the residuals but also on the scores. There are two types of weights used in PRM regression which are the residual weight, \( w_i^r \) that deals with leverage points and the leverage weight \( w_i^L \) which deals with vertical outliers. Provided by the weight function \( f \) called the “Fair” function,

\[ f(z, c) = \frac{1}{1 - \left| \frac{z}{c} \right|} \]  

\[ \text{with } c = 4 \text{ is the tuning constant, the weights, } w_i^r \text{ have been computed as} \]

\[ w_i^r = f \left( \frac{r_i}{\hat{\sigma}} \cdot c \right) \]  

\[ \text{where } r_i = y_i - \text{median}(y_j) \text{ is the residuals and } \hat{\sigma} = \text{MAD}(r) = \text{median}_{j} \left| r_i - \text{median}(r_j) \right| \text{ be the mean absolute deviation of a residual vector } r. \text{ The leverage weight } w_i^L, \text{ is defined as} \]

\[ w_i^L = f \left( \frac{t_i - \text{med}_L(T)}{\text{median}_{i} \left| x_i - \text{med}_L(T) \right|} \right) \]
where $\| \cdot \|$ is the Euclidian norm, $t_i$ are the $f$ PLS scores of the $i$th object, $\text{medL}_1(T)$ denotes the L1-median calculated from score vectors. The PRM algorithm consists of the following steps as follow:

Step 1: Calculate the robust starting values for the weights, $w_i = w_i^r w_i^t$
Step 2: Perform classical PLS (SIMPLS) on weighted data, $w_i x_i$ and $w_i y_i$.
Step 3: Update the weights, $w_i$ by recomputing $w_i^r$ from PLS residuals and $w_i^t$ from PLS scores.
Step 4: Iterate step 2 and 3 until convergence of the estimated regression coefficient. Convergence is achieved if the relative difference in norm between estimated regression coefficients is smaller than certain threshold value.
Step 5: The estimated regression is obtained from the last weighted PLS.

3. A simulation study
The efficiency of OLS regression, PCR, RR, PLSR and PRM regression were compared by performing a simulation study on simulated dataset. The experiment described in this section considers only univariate responses represented by the following regression model.

$$Y_1 = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$ (12)

where $\beta_0 = \beta_0 = \ldots = \beta_p = 1$. Simulated data sets will be employed in this study with $p = 4$ and 10 predictor variables for sample sizes $n = 20$, 50, 100, 200 and 500. The goal is to develop a linear equation that relates all the predictor variables to a response variable. Data sets with multicollinearity problem will be simulated by using R with the following initial model;

$$x_1 = N(0,1)$$ (13)
$$x_{p-1} = N(0,0.1) + x_1$$ (14)
$$Y = x_1 + x_2 + \ldots + x_p + \epsilon_i$$ (15)

where $\epsilon_i$ is the error term following the standard normal distribution and $p = 4$ and 10. For each situation, $m = 1000$ dataset will be generated. To create vertical outliers in the simulated dataset, the error terms in which follows the normal distribution, N(0,\sigma) will be replaced with N(20,\sigma). The percentages of contamination in predictor variable used are 5, 10, 20 and 30 for each fixed value of number of predictors, $p$ and sample size, $n$ as done by González, Peña, and Romera [23]. In this simulation study, all the data generated are regressed on OLS regression, PCR, RR, PLSR and PRM regression.

4. Analysis and findings
This study is performed by investigating the effect of vertical outlier with severe multicollinearity. To detect multicollinearity, variance inflation factor (VIF) was calculated. VIF is defined as

$$\text{VIF} = \frac{1}{1 - R_j^2}$$ (16)

where $R_j^2$ is the coefficient of determination in the regression. According to Neter, Wassemen, and Kutner [24] if the value of VIF > 10, multicollinearity problem exists. In this study, data containing multicollinearity was simulated and VIF values for the simulated dataset were determined. Table 1 indicates the VIF values of $p=4$ and $p=10$ regressors for datasets with multicollinearity. From the VIF values, it is concluded that there exists multicollinearity among explanatory variables in the dataset simulated since all VIF>10.
Table 1. VIF values for $p=4$ and $10$ for $n=500$ datasets with multicollinearity

| Regressors | $p=4$   | $p=10$   |
|------------|---------|----------|
| x1         | 160.3162| 818.7042 |
| x2         | 53.1903 | 78.4229  |
| x3         | 51.9849 | 81.4417  |
| x4         | 99.6709 | 148.1541 |
| x5         |         | 50.9031  |
| x6         |         | 113.3492 |
| x7         |         | 152.8349 |
| x8         |         | 101.8985 |
| x9         |         | 78.2488  |
| x10        |         | 109.8038 |

VIF for $p=4$ and $p=10$ regressors for datasets containing multicollinearity problem and vertical outlier is presented in Table 2. It is observed that, VIF increases and becomes more severe on dataset with multicollinearity and vertical outlier.

Table 2. VIF values for $p=4$ and $10$ for $n=500$ datasets with multicollinearity and vertical outliers

| Regressors | $p=4$   | $p=10$   |
|------------|---------|----------|
| x1         | 34741.2060 | 76618.3380 |
| x2         | 7915.3320  | 8739.4970 |
| x3         | 11255.1450 | 9430.3100 |
| x4         | 11172.6200 | 8789.3080 |
| x5         |           | 9479.1640 |
| x6         |           | 8465.9540 |
| x7         |           | 9263.1780 |
| x8         |           | 10016.2310|
| x9         |           | 9313.0720 |
| x10        |           | 9226.8940 |

By comparing VIF values from Table 1 and Table 2, the VIF value becomes critical when vertical outliers are present in the dataset containing multicollinearity problem. This result suggested that vertical outliers increase the collinearity among predictor variables.

To see the efficiency of all regression methods involved in this study, the Mean Square Error (MSE) of estimated parameters are computed. The MSE($\beta$) of the estimator of parameter is defined as:

$$MSE(\beta) = \frac{\sum_{i=1}^{m}(\hat{\beta} - \beta)^2}{m}$$  \hspace{1cm} (17)

where $m$ is the number of simulation replications and $\beta$ is the true value of the simulated model. MSE($\beta$) is used as the efficiency test to find the best regression model that best fits the dataset with multicollinearity and vertical outlier. Since MSE($\beta$) measures to what extent the slope and intercept are correctly estimated, a value close to zero is the best.
Table 3 and Table 4 presents the MSE value of the estimated regression parameters, $\hat{\beta}$, from 1000 simulated dataset for $p = 4$ and $p = 10$ regressors, with different number of samples and in the existence of multicollinearity and vertical outlier. It can be observed that, MSE($\beta$) increase as the level of contamination increase in all cases (different observation, regressors and contamination level). MSE($\beta$) also increase as regressor increase but decrease as observation increase. RR has the highest MSE($\beta$) value followed by OLS regression for both 4 and 10 regressors in contamination up to 10% for all observation. But when contamination increase up to 30%, OLS regression has higher MSE($\beta$) value compared to RR. On the contrary, both PCR and PLSR have significantly lower MSE value for all cases in comparison to the OLS regression and RR. Whereas, PRM regression has the lowest MSE($\beta$) value which is closer to zero for all cases.

Table 3. $MSE(\beta)$ values from 1000 simulated samples for $p = 4$ using different number of contaminations and $n$ samples with multicollinearity and vertical outliers

| Outliers | $n$ | Method | OLS | PCR | RR   | PLS  | PRM |
|----------|-----|--------|-----|-----|------|------|-----|
| 5%       | 20  | 874.90 | 540.57 | 879.07 | 337.01 | 9.91 |
|          | 50  | 238.50 | 155.05 | 244.26 | 100.04 | 3.56 |
|          | 100 | 128.83 | 89.02  | 124.09 | 64.20  | 3.63 |
|          | 200 | 63.78  | 43.27  | 64.05  | 31.33  | 1.78 |
|          | 500 | 25.64  | 16.35  | 24.42  | 10.94  | 0.25 |
| 10%      | 20  | 1685.99 | 983.85 | 1620.61 | 675.74 | 12.27 |
|          | 50  | 545.11 | 321.55 | 501.92 | 245.18 | 4.56 |
|          | 100 | 270.72 | 170.77 | 248.41 | 123.20 | 3.92 |
|          | 200 | 122.39 | 76.79  | 117.10 | 54.79  | 2.00 |
|          | 500 | 49.62  | 24.66  | 44.39  | 24.66  | 0.30 |
| 20%      | 20  | 3223.57 | 1762.37 | 2837.83 | 1383.45 | 22.19 |
|          | 50  | 1115.19 | 563.63 | 915.77 | 491.61 | 8.00 |
|          | 100 | 496.69 | 287.45 | 412.23 | 197.37 | 5.80 |
|          | 200 | 239.98 | 135.77 | 204.12 | 86.81  | 2.71 |
|          | 500 | 100.73 | 56.30  | 79.30  | 36.42  | 0.59 |
| 30%      | 20  | 4522.17 | 2291.47 | 3619.16 | 2518.39 | 164.88 |
|          | 50  | 1680.09 | 756.45 | 1232.38 | 551.62 | 18.13 |
|          | 100 | 775.63 | 372.56 | 568.92 | 248.71 | 10.80 |
|          | 200 | 359.59 | 181.20 | 270.34 | 105.28 | 4.76 |
|          | 500 | 159.07 | 79.59  | 110.88 | 53.26  | 1.37 |

Table 4. $MSE(\beta)$ values from 1000 simulated samples for $p = 10$ using different number of contaminations and $n$ samples with multicollinearity and vertical outliers

| Outliers | $n$ | Method | OLS | PCR | RR   | PLS  | PRM |
|----------|-----|--------|-----|-----|------|------|-----|
| 5%       | 20  | 1378.37 | 158.25 | 1423.10 | 53.23 | 7.60 |
|          | 50  | 281.51  | 53.36  | 288.15 | 19.92  | 2.41 |
|          | 100 | 139.76  | 34.37  | 135.83 | 11.00  | 0.71 |
|          | 200 | 68.54   | 22.89  | 69.05  | 17.22  | 12.06 |
|          | 500 | 25.39   | 6.85   | 24.34  | 2.45   | 0.18 |
| 10%      | 20  | 2761.98 | 278.21 | 2783.99 | 89.63  | 8.44 |
|          | 50  | 647.73  | 119.96 | 602.34 | 44.62  | 3.01 |
|          | 100 | 288.70  | 63.18  | 266.88 | 19.99  | 0.98 |
|          | 200 | 130.89  | 37.97  | 125.04 | 21.14  | 12.21 |
PRM regression has lower MSE($\beta$) when number of regressors increase from 4 to 10. Compared to other regression methods, PRM regression is the best method in parameter estimation when a dataset contains multicollinearity and vertical outliers. Of the five methods compared, PRM regression outperforms other methods by the smallest value of MSE of the estimate in all contamination level. Hence, PRM regression is the most efficient and best precision method when multicollinearity and vertical outliers is present.

5. Application to a real dataset
Evaporated water data was used to illustrate the problem of multicollinearity problem and vertical outliers [25]. This data was edited to create outliers in Y-direction. For observation number 13, the value of 47 is replaced with 77 and observation 25, the value of 54 is replaced with 74 to create 20% vertical outliers. Evaporated data was used to illustrate the data that contains problems of multicollinearity problem and vertical outliers. VIF values larger than 10 proves that multicollinearity problem is present.

| Explanatory Variables | VIF |
|-----------------------|-----|
| X1                    | **20.3101** |
| X2                    | 8.6227 |
| X3                    | **27.2724** |
| X4                    | **14.3466** |
| X5                    | 9.3957 |
| X6                    | **16.9689** |

From Table 5, there exists serious multicollinearity among predictor variables by observing the value of VIF. It is proven that high collinearity exists among and explanatory variables X1, X3, X4 and X6.
Figure 1. Box plot of all variables for Evaporated Water Data.

Box plot is useful to illustrate the existence of outliers and from Figure 1, the box plot for all variables in Evaporated Water Data is presented. It is observed that there exist vertical outliers in the dependent variable.

Table 6. MSE(\(\hat{\beta}\)) for evaporated water dataset

|       | OLS  | PCR  | RR   | PLS  | PRM  |
|-------|------|------|------|------|------|
| RMSE  | 1.3947 | 1.5976 | 1.3750 | 1.5501 | 0.1723 |

Table 6 presents the MSE(\(\hat{\beta}\)) for Evaporated Water Data. The result proves that in the presence of multicollinearity and vertical outliers, PRM regression had the smallest MSE(\(\hat{\beta}\)). It can be concluded that PRM regression is the best method in determining the parameter estimate.

6. Conclusion

In this study, we discussed five different regression techniques in dealing data with multicollinearity and vertical outlier. The classical Ordinary Least Square (OLS) Regression was compared with three famous regression methods that can handle multicollinearity problem which are Principle Component Regression (PCR), Ridge Regression (RR) and Partial Least Square Regression (PLSR). In the presence of multicollinearity and outliers simultaneously, a new arising robust Partial Least Square Regression which is called Partial Robust M-Regression that was introduced by Serneels et al. [19] was applied and compared to the four previous methods. A simulation study was carried out and the comparison of performance for each method was made by means of the Mean Square Error of the estimate (MSE (\(\hat{\beta}\))). In the simulation study, it was shown that in the presence of multicollinearity and vertical outliers, PRM Regression has the lowest value of MSE of the estimate. Therefore, PRM is the most efficient regression method to determine parameter estimate in contrast to other methods, following the last numbered section of the paper.

References

[1] Ho K and Naugher JR 2000. Outlier Lies: An Illustrative Example of Identifying Outliers and Applying Robust Models, Mult. Linear Regres. Viewpoints, vol. 26, no. 2, pp. 2–6
[2] Adnan N, Ahmad MH, and Adnan RA 2006. comparative study on some methods for handling multicollinearity problems, Matematika, vol. 22, no. 2, pp. 109–119
[3] Lin JL and Tsay RS 2005. Comparisons of Forecasting Methods with Many Predictors, Work. Pap., pp. 1–23
[4] Field A 2009. Multilevel Linear Models, Discovering Statistics Using SPSS. pp. 725–781.
[5] Massy WF 1965. Principal Components Regression in Exploratory Research, J. Am. Stat. Assoc., vol. 60, no. 309, pp. 234–256
[6] Hoerl AE and Kennard RW 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, vol. 12, no. 1, pp. 55–67
[7] Wold S, Ruhe A, Wold H and Dunn WJ, 1984. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses, SIAM J. Sci. Stat. Comput., vol. 5, no. 3, pp. 735–743
[8] Al-hassan YM and Al-Kassab MM 2009. A Monte Carlo Comparison between Ridge and Principal Components Regression Methods, Appl. Math. Sci., vol. 3, no. 42, pp. 2085–2098.
[9] Frank LE and Friedman JH 1993. A Statistical View of Some Chemometrics Regression Tools, Technometrics, vol. 35, no. 2, pp. 109–135.
[10] Rougoor C, Sundaram R, and van Arendonk JA 2000. The relation between breeding management and 305-day milk production, determined via principal components regression and partial least squares, Elsevier, vol. 66, no. 1, pp. 71–83.
[11] Yeniey O and Goktas A 2002. A comparison of partial least squares regression with other prediction methods, Hacettepe J. Math., vol. 31, pp. 99–111.
[12] Farahani AH, Rahiminezhad A, Same L, and Immannezhad K 2010. A Comparison of Partial Least Squares (PLS) and Ordinary Least Squares (OLS) regressions in predicting of couples mental health based on their communicational patterns, Elsevier, vol. 5, pp. 1459–1463.
[13] Ramzan S, Zahid FM, and Ramzan S 2010. Prediction Methods for Time Series Regression Models with Multicolinearity, World Appl. Sci. J., vol. 11, no. 4, pp. 443–450.
[14] Wainer H 1976. Robust Statistics: A Survey and Some Prescriptions, J. Educ. Behav. Stat., vol. 1, no. 4, pp. 285–312
[15] Stevens JP 1984. Outliers and influential data points in regression analysis., Psychol. Bull., vol. 95, no. 2, pp. 334–344.
[16] Jarrel MG 1994. A comparison of two procedures, the Mahalanobis distance and the Andrews-Pregibon statistic, for identifying multivariate outliers, ERIC, vol. 1, pp. 49–58.
[17] Zimmerman DW 1994. A Note on the Influence of Outliers on Parametric and Nonparametric Tests, J. Gen. Psychol., vol. 121, no. 4, pp. 391–401.
[18] Chen C 2002. Statistics and Data Analysis Paper 265-27 Robust Regression and Outlier Detection with the ROBUSTREG Procedure,” Statistics (Ber).
[19] Serneels S, Christoph C, Filzmoser P, and Espen PJV 2005. Partial robust M-regression Chemom. Intell. Lab. Syst., vol. 79, no. 1–2, pp. 55–64.
[20] Liebmann B, Filzmoser P, and Varmuza K 2010. Robust and classical PLS regression compared,” J. Chemom., vol. 24, no. 3–4, pp. 111–120
[21] Abdi H 2010. Partial least squares regression and projection on latent structure regression (PLS Regression), Wiley Interdiscip. Rev. Comput. Stat., vol. 2, no. 1, pp. 97–106
[22] Naes T and Martens H 1985. Comparison of prediction methods for multicollinear data Commun. Stat. - Simul. Comput., vol. 14, no. 3, pp. 545–576
[23] González J, Peña D, and Romera R 2009. A robust partial least squares regression method with applications, J. Chemom., vol. 23, no. 2, pp. 78–90.
[24] Neter J, Wasserman W, and Kutner M 1983. Simultaneous inferences and other topics in regression analysis
[25] Späth H 2014. Mathematical algorithms for linear regression. Elsevier: United States of America