Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 52 (2) (2023), 459 – 486
DOI : 10.15672/hujms.1122113

Research Article

Robust correlation scaled principal component regression

Aiman Tahir*, Maryam Ilyas

College of Statistical and Actuarial Sciences, University of the Punjab, Lahore, Pakistan

Abstract

In multiple regression, different techniques are available to deal with the situation where the predictors are large in number, and multicollinearity exists among them. Some of these approaches rely on correlation and others depend on principal components. To cope with the influential observations (outliers, leverage, or both) in the data matrix for regression purposes, two techniques are proposed in this paper. These are Robust Correlation Based Regression (RCBR) and Robust Correlation Scaled Principal Component Regression (RCSPCR). These proposed methods are compared with the existing methods, i.e., traditional Principal Component Regression (PCR), Correlation Scaled Principal Component Regression (CSPCR), and Correlation Based Regression (CBR). Also, Macro (Missingness and cellwise and row-wise outliers) RCSPCR is proposed to cope with the problem of multicollinearity, the high dimensionality of the dataset, outliers, and missing observations simultaneously. The proposed techniques are assessed by considering several simulated scenarios with appropriate levels of contamination. The results indicate that the suggested techniques seem to be more reliable for analyzing the data with missingness and outlyingness. Additionally, real-life data applications are also used to illustrate the performance of the proposed methods.

Mathematics Subject Classification (2020). 62H25, 62D10, 62J20

Keywords. Principal component regression, influential observations, robust regression

1. Introduction

Linear regression analysis is used to estimate the linear relationship between the response variable and one or more independent variables for prediction purposes (e.g., [15, 47]). In multiple linear regression, explanatory variables may be linearly related to each other which causes the problem of multicollinearity (e.g., [5, 48, 68, 79]). The predictive power of the model and model fitting procedure is not much affected by the presence of multicollinearity (e.g., [55, 73]). However, the estimates of the regression coefficients are highly affected by multicollinearity as the sampling variances of ordinary least square estimates inflate in the presence of multicollinearity (e.g., [39]). Additionally, with a large number of predictor variables, the multiple regression becomes complex (e.g., [42]). Many methods are available in the literature that can cope with the problems of multicollinearity

*Corresponding Author.
Email addresses: aimantahir78@gmail.com (A. Tahir), maryam.stat@pu.edu.pk (M. Ilyas)
Received: 28.05.2022; Accepted: 10.09.2022

and dimension reduction (e.g., [32,39,70]). However, the focus of this paper is Principal Component Regression (PCR) [39] and Correlation Based Regression (CBR) [70].

Consider a multiple linear regression model in Eq. (1.1) based on n-observations and p-predictors. Here, $\mathbf{y}$ is a $(n \times 1)$ vector of the dependent variable, $\mathbf{X}$ is a $(n \times p)$ matrix of known standardized predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients and $\boldsymbol{\epsilon}$ is a $(n \times 1)$ vector of random errors that is normally distributed with zero mean and unknown variance σ^2.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.1}$$

The Ordinary Least Square (OLS) estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ (e.g., [47]) and vector of fitted values is $\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$, where,

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{1.2}$$

Principal component regression [39] is closely related to multiple linear regression. In principal component regression, the subset of Principal Components (PCs) (e.g., [1,40]) of independent variables are used as regressors instead of original predictors. Often the PCs with large variances are chosen as regressors although the PCs with low variances may also be important in prediction purposes (e.g., [38]). Rest of the regression fitting is performed just like the classical least square method. More precisely, in principal component regression, the explanatory variables ($\mathbf{X}$) are transformed in terms of its p-eigenvalues and eigenvectors. Mathematically, it can be written as

$$\mathbf{X^T X} = \mathbf{P D P^T} = \mathbf{Z^T Z}, \tag{1.3}$$

where $\mathbf{D}$ is a diagonal matrix containing the eigenvalues of $\mathbf{X}^T\mathbf{X}$ on the diagonal, $\mathbf{P}$ is the matrix containing the eigenvectors of $\mathbf{X}^T\mathbf{X}$. Also, $\mathbf{P}$ is an orthogonal matrix such that $\mathbf{P}^T\mathbf{P} = \mathbf{P}\mathbf{P}^T = \mathbf{I}$ and $\mathbf{Z}$ is a new orthogonal data matrix having same order as $\mathbf{X}$ which is composed of the principal components.

The PCR overcomes the problem of multicollinearity and reduces the dimensions [39]. However, it is important to note that it does not account for the correlation between the predictor variables and the response variable. To fill this gap, Correlation Scaled Principal Component Regression (CSPCR) has been introduced [70]. It utilizes the correlation between dependent and independent variables. In CSPCR, each independent variable is multiplied by the corresponding correlation of this variable with the response variable. Then PCR is applied over this correlation scaled predictors. CBR (e.g., [70]) is another technique that is used to reduce the dimensions in multiple linear regression and is based on the correlation between response variable and predictors. Unlike CSPCR, it only considers the variables that are highly correlated with response variable and then the regression analysis is performed by utilizing the subset of these highly correlated variables as regressors.

In regression analysis, a common problem occurs when the outliers, influential observations or leverage points exist in the data set (e.g., [11,12,14,64]). This problem may have a significant impact on the estimation of the parameters (e.g., [14]). These contaminated values typically pull the regression line in their direction. Therefore, the estimates, their standard errors, the coefficient of determination, and the residual sum of squares become sensitive due to this contamination problem (e.g., [11,27,28]).

The presence of outliers in the data limits the use of traditional PCR. Walczak and Massart [76] proposed a robust PCR approach. In this approach, robust PCA is utilized instead of classical PCA. Robust PCA relies on robust covariance matrix that is estimated by using the least median of squares regression [62]. Pell [56] introduced a method of detecting outliers in response matrix by using the technique Resampling by HalfMeans (RHM) [23]. This approach removes the samples contaminated with outliers before computing the principal components. Filzmoser [25] proposed robust PCR method which uses the idea of projection pursuit introduced by [43]. This method computes the

robust principal components and incorporates these robust PCs in LTS regression [62] for prediction purposes. Hubert and Verboven [36] introduced another robust approach for PCR. This method comes with two variations. For the low-dimensional data ($p < n$), covariance matrix is estimated by Minimum Covariance Determinant (MCD) estimator [60]. For the high dimensional data ($p > n$), the ROBPCA method [35] was proposed to be implemented followed by a robust regression approach. There is an empirical approach of robust PCR [78] as well. This is based on regression diagnostics so-called principal sensitive vectors [57]. These diagnostics are used to detect the outliers before carrying out the classical least square PCR approach. Engelen et al. [24] conducted a comparative study in which two approaches (robust PCR and robust partial least square) were compared on the basis of their efficiency, goodness of fit, predictive power and robustness. Denhere and Billor [19] proposed the estimator of parameter function for a functional logistic regression model which dealt with functional outliers. Gagnon [26] introduced the Bayesian approach that resists outliers in both the dependent variable and independent variables. This approach penalizes the unusual observations and gives predictions which are consistent with the rest of the data matrix. It also discovers significant outliers by using the geometry of PCs.

Missing observations in the data matrix also limit the use of PCA. There exist some robust covariance estimators in the literature that jointly cope with the outliers and missing values (e.g., [67, 71]). Missingness and outlyingness can also be handled using imputation procedures on the data matrix. MacroPCA [34] is one such technique that is based on imputation techniques. It generates fully imputed matrices for missingness and outliers (both cellwise and row-wise). Principal component scores are then generated using the imputed data matrix. Thus, MacroPCA jointly deals with multicollinearity, high dimensions, missingness, and outliers (row-wise and cellwise).

In this paper, Robust Correlation based Regression (RCBR) and Robust Correlation Scaled Principal Component Regression (RCSPCR) which are the robust version of two existing methods CBR and CSPCR [70] are proposed. Additionally, MacroRCSPCR is introduced. It combines the features of RCSPCR and MacroPCA [34]. Therefore, the proposed MacroRCSPCR is a form of PCR that can simultaneously deal with missing values, cellwise outliers, row-wise outliers, high dimensions and multicollinearity.

This paper is distributed in different sections. Subsections of Section 1 give overview of outliers, influential observations and leverage points. Their detection tests and robust regression are also briefly discussed in the subsections. In Section 2, details of the developed methods are presented. Section 3 provides details of data generation procedure for simulation study and different simulation scenarios. The performance of proposed methods using simulated and real-life data sets is evaluated in Section 4.

1.1. Outliers and leverage points

Outliers are the observations that are inconsistent with the rest of the data set [10]. It is generally an observation with large residual. Outliers give outlying values for corresponding error term (in the y-direction) and are not typically outlying in the space of predictors (e.g., [12, 64]). On the other hand, leverage measures how far an independent variable deviate from its mean. These points are typically far from the cloud of the rest of the data. Leverage points exist in the space of independent variables and not in the response variable (e.g., [64]). The outlying points in the predictors that are close to the regression line and do not influence the model fit are considered as good leverage points. Bad leverage points are the ones that are inconsistent with the rest of the data in the space of predictors and also have strong impact on regression line (e.g., [12, 60]). The outliers or leverage points that influence the slope of regression line, estimates of regression coefficients and their standard errors are considered as influential observations (e.g., [11]).

In multivariate analysis, data are presented in a rectangular matrix. Cases and variables are arranged in rows and columns, respectively (e.g., [37]). Sometimes certain cases deviate from the majority of the cases. These are referred as row-wise outliers [34]. Cellwise outliers are the unusual entries that exist anywhere in the dataset (e.g., [2, 7]).

1.2. Diagnostic test

Different methods have been introduced in the literature to detect the outlier, leverage points and influential observations in linear regression analysis (e.g., [14, 16, 17, 62, 75]). Graphical diagnostic tools can also be used to identify the outliers (e.g., [15, 18, 21, 66]).

The standard diagnostic approaches involve the least square projection matrix (**H**). This **H** matrix in Eq. (1.2) is known as the hat matrix. The diagonal entries h_{ii} of hat matrix **H** are the classical diagnostic tools for detecting leverage and influential points (e.g., [14,62]). The points for which $h_{ii} > 2p/n$ are considered problematic (e.g., [17, 29–31, 54, 72]). It is important to note that h_{ii} just detects outliers in x-space. It does not take into account the outliers present in the y-direction.

The outliers in the y-direction can be detected by making use of least square residuals (e_i) (e.g., [63]). These residuals are often scaled to scrutinize the outliers (e.g., [9, 16, 17, 30, 31, 54, 72]). For example, studentized residuals [9] are scaled version of least square residuals.

$$studentized\ residuals : t_i = \frac{e_i}{s\sqrt{(1-h_{ii})}} \tag{1.4}$$

Here, $e_i = y_i - \hat{y}_i$ and s is the estimate of σ. The large value of t_i gives the indication of outliers in the response variable [14].

There are three forms of diagnostic tools. These are single case diagnostics, multiple case diagnostics and high breakdown diagnostics [62]. The classical single case diagnostics of influential observations are Cooks squared distance (CD^2) [16] and Difference of Fits (DFFITS) [11].

$$CD^2(i) = \frac{t_i^2}{p}\frac{h_{ii}}{1-h_{ii}} \tag{1.5}$$

$$DFFITS(i) = \frac{e_i}{s(i)}\frac{(h_{ii})^{1/2}}{1-h_{ii}} \tag{1.6}$$

Here, $s(i)$ is the estimate of σ that is obtained after deleting the i^{th} observation. The observation for which $CD^2(i)$ is equal to 1 is referred to as influential observation (e.g., [17, 44]). If the DFFITS exceeds $2(p/n)^{1/2}$, the observation may be considered as unusual point [11].

The generalization of single-case diagnostics is known as multiple-case diagnostics. For example, the generalization of Cooks squared distance has been proposed by [17] that checks the joint effect of removing more than one case. High breakdown diagnostics are able to deal with the multiple outliers without taking the masking effect [62]. The observations for which the diagnostic increase its threshold value are firstly deleted. Then, the least square regression is performed on the rest of the dataset. There are other more recent techniques that are able to detect cellwise outliers [61] and good and bad leverage points [34].

1.3. Robust regression

In regression analysis, the existence of outliers in the data matrix often inflate the standard errors of the OLS regression estimators [6]. To fix this, robust regression is a suitable alternative to OLS (e.g., [8, 74]). The classical robust regression approach is least absolute values regression [22]. This approach is also known as L_1 regression. The

regression coefficients are estimated by minimizing the sum of absolute residuals. It is robust to outliers and non-robust for the leverage points. Another traditional robust regression approach is the M-estimate [33]. It is based on maximum likelihood. M-estimator minimizes the sum of the weight function of the residuals. It resists the outliers in response variable only. If the unusual observations exist in explanatory variables (leverage points), M-estimate and OLS give similar results [20].

The classical method to deal with both outliers in the response variable and the explanatory variable is Least Trimmed Squares (LTS) regression [60]. The objective function of the LTS regression is

$$\underset{\hat{\boldsymbol{\beta}}}{minimize} \sum_{i=1}^{h} (e^2)_{i:n}, \tag{1.7}$$

where $(e^2)_{1:n} \leq , , \leq (e^2)_{n:n}$ are the ordered statistics of the squared residuals and h is the trimming constant. Note that, when $h = n$, this is equivalent to the standard least square estimator. Typically, h is set to some constant fraction of n based on the expected number of outliers. To guarantee a unique solution it is often assumed that h is at least $n/2$. Some other robust estimators existing in literature are S-estimate [65], GM-estimators [60], least winsorised squares [77] and repeated median estimate [69].

2. Statistical methodology

This section describes the three proposed methods. These are RCSPCR, RCBR and MacroRCSPCR. RCSPCR and RCBR accounts for the influential observations in the data. However, MacroRCSPCR is proposed to deal with the influential and missing observations simultaneously.

2.1. Robust correlation scaled principal component regression (RCSPCR)

The CSPCR [70] is based on the steps below.

(i) Obtain correlation between the response variable and each of the independent variables.
(ii) Multiply each explanatory variable by the corresponding correlation value obtained in step (i). This gives the independent variables scaled by the corresponding correlation with the dependent variable.
(iii) Use the correlation scaled explanatory variables obtained in step (ii) and apply Principal Component Analysis (PCA).
(iv) Use the first k-PCs (where $k < p$) that explain most of the variation of data obtained in step (iii) as regressors and carry out classical OLS regression.

We proposed RCSPCR that accounts for the influential observations in the data set. Classical least square gives misleading results in the presence of influential observations. Therefore, step (iv) in CSPCR is replaced with a robust regression approach and the resulting methodology is termed RCSPCR. We used Least Trimmed Square Estimate (LTS) as an alternative robust regression estimator to the least square method in RCSPCR.

2.2. Robust correlation based regression (RCBR)

To perform CBR [70], the following steps are carried out.

(i) Compute the correlation between the dependent variable and each of the independent variables.
(ii) Sort out the predictors concerning the non-increasing order of correlation with the response variable.
(iii) Select the first 1/4 of the predictors with high correlations as regressors and perform regression.

The RCBR is proposed in this paper that is based on CBR [70]. To account for the influential observations, OLS regression in step (iii) in CBR is replaced with the least trimmed squares regression in RCBR.

2.3. Macro correlation scaled principal component regression (MacroRCSPCR)

Traditional PCR cannot be applied to missing values because the classical method which is used to compute PCs does not corporate with NAs, row-wise and cellwise outliers. MacroPCA [34] can cope with the outliers (both cellwise and row-wise) and NAs simultaneously. It starts by applying DetectDeviatingCells (DDC) algorithm [61] that gives initial imputation to NAs, cellwise and row-wise outliers. It imputes the missing entries and obtains the estimate of the expected values of outlying cells simultaneously. DDC algorithm is used as initial step in estimating the multivariate location and scatter matrices. Further, MacroPCA collaborate Iterative Classical Principal Component Analysis (ICPCA) [41,46] and Robust Principal Component Analysis (ROBPCA) [35] for improved imputation of NAs, row-wise and cellwise outliers. These imputed predictors scaled by corresponding correlation with response variable are used in computing PCs.

The proposed MacroRCSPCR changes two steps in CSPCR [70] (see Subsection 2.1). Instead of using classical PCA, MacroPCA is performed in step (iii). Also, the OLS regression in step (iv) is replaced with least trimmed squares regression in MacroRCSPCR.

3. Simulation study

3.1. Data generation

In order to examine the finite sample performance of the proposed approaches (RCSPCR, RCBR and MacroRCSPCR), a simulation study is conducted. Following the ethos of [3], the predictors are simulated using the normal distribution.

$$x_1 = N(0, 1) \tag{3.1}$$

$$x_{p-1} = N(0, 1.5) + x_1 \tag{3.2}$$

These predictors are set in the columns of a $(n \times p)$ matrix of predictors $\mathbf{X}$. Furthermore, a response variable is simulated as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta}. \tag{3.3}$$

Here, the vector of regression coefficients ($\boldsymbol{\beta}$) is generated from a uniform distribution in the interval [-1, 1]. The error term $\boldsymbol{\delta}$ is obtained from N (0, 1). In the next Subsections (3.2)–(3.5), different simulation scenarios are discussed. These scenarios are designed to test the performance of the proposed methods (see Section 2).

3.2. Scenario-I: Outliers in response variable

In this case, different levels of contamination, i.e., 0%, 2%, and 10% in the response variable are considered without modifying the predictors. For example, to infuse 2% outliers in the data, 2% observations of the response variable are randomly replaced by the values from the normal distribution with $\mu = 25$ and $\sigma = 1.5$. Other levels of contamination are processed in a similar way.

3.3. Scenario-II: Leverage points in the predictors

This scenario studies leverage points. Different levels of contamination, i.e., 0%, 2%, and 10% are infused in the predictors without contaminating the dependent variable. These leverage points are found as good leverage points according to the criteria proposed in MacroPCA for detecting the good and bad leverage points [34]. For example, to infuse 2% leverage points in independent variables, 2% values of 11 to 13 predictors at the same position are randomly replaced by the observations from $N(25, 2.25)$.

3.4. Scenario-III: Data with both outliers and leverage points

In this scenario, we considered both the outliers and leverage points. 2% values of response are randomly replaced by the observations normally distributed with $\mu = 25$ and $\sigma = 1.5$ and 10% observations of 11 to 13 predictors at the same positions are randomly replaced with the values from normal distribution $N(25, 2.25)$. 0%, 2%, and 10% outliers in response variable and 10% leverage points are taken in this case.

3.5. Scenario-IV: Data matrix with "Macro"

This scenario considers cellwise and row-wise outliers and missing observations. 2% values of response variable are replaced by the observations generated from N(50, 2.25) that introduces outliers. In order to produce missing observations in predictors, a random subset (0%, 5%, and 15%) of $n \times p$ cells are replaced with NAs. Also, different percentages (0%, 15%, 30%) of cellwise outliers are randomly infused by replacing original cells with the values generated from $N(250, 25)$. Furthermore, the predictors are modified by replacing 0% and 20% of random rows with the rows generated from $N(240, 9)$ to introduce the row-wise outliers. Several combinations of contamination levels with missing values, cellwise outliers and row-wise outliers were considered and some of these are presented in this paper.

3.6. Performance evaluation criteria

In order to compare the proposed methods with the existing ones, different statistical tools are used. For example, the coefficient of determination R^2 is used by several researchers (e.g., [70]) for similar comparisons. The classical (R^2) and robust (R^2_w) coefficient of determination [59] used in this research are as follows:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.4}$$

$$R^2_w = 1 - \frac{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} w_i(y_i - \bar{y}_w)^2}. \tag{3.5}$$

Here, $\bar{y}$ is the mean of the dependent variable, $\bar{y}_w = (1/\sum_{i=1}^{n} w_i)\sum_{i=1}^{n} w_i y_i$ and w_i are the weights.

For assessing the predictive accuracy of the proposed methods, cross-validated Average Prediction Error (APE) is calculated.

$$APE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{3.6}$$

Further, the efficiency of the proposed methods is assessed using bootstrap estimates of the variances of regression coefficients. As an example, the detailed steps of the calculation of R^2, APE and variance-covariance matrix of regression coefficients for RCBR are discussed in Appendix A.

4. Results

4.1. Simulated data

Simulated data sets were generated under various settings discussed in Subsections 3.2 to 3.5. The tables regarding Scenarios I-III were generated by considering the sample sizes as 50 and 500. The number of predictors is considered 20. For Scenario-IV, the sample size is reported as 500 for 20 predictors. The proposed methods were implemented on the simulated data sets. The results of the proposed methods resulting from these simulated data sets are presented in Tables 1-4 and in Appendix B. The values of R^2, APE and variance-covariance matrices of the regression coefficients are presented for different contamination levels and the different number of retained regressors / PCs.

In Scenario-I, it can be observed from Table 1 that our proposed methods (RCBR and RCSPCR) outperform other existing methods (CBR, PCR and CSPCR). The presence of outliers markedly reduces the values of R^2 in the case of existing methods (CBR, PCR and CSPCR). The predictive accuracy assessed by APE of our proposed techniques is slightly better than the existing methods. The regression coefficients of our proposed methods are observed marginally more efficient than the existing ones. Therefore, the proposed methods (RCBR and RCSPCR) are not influenced by the presence of these outliers. This is due to the fact that outliers affect the OLS estimates and slope of the regression line. As CBR, PCR and CSPCR are based on OLS estimates that are highly sensitive to outliers but our proposed methods (RCBR and RCSPCR) give better performance because of their robust aspects. Another aspect that needs to be noted is the levels of contamination. The values of R^2 decrease as the contamination levels increase for the existing methods but in the case of our proposed methods, the increase in contamination fraction does not influence the values of R^2. This is because the proposed methods have been introduced in a robust context to overcome the extraordinary impact of contaminated entries in the data set. However, for increasing contamination levels the APE and variances of the estimates inflate for RCBR and RCSPCR.

Table 1. Scenario-I: The results of simulation study in the presence of outliers for n=500 and p=20 at different contamination levels.

	No. of retained regressors/PCs		Contamination levels		
Method	M		0%	2%	10%
CBR	2	R^2	0.395	0.182	0.060
		APE	0.611	13.278	63.856
	5	R^2	0.493	0.223	0.076
		APE	0.562	13.203	64.033
RCBR	2	R^2_w	0.432	0.419	0.383
		APE	0.612	13.194	63.294
	5	R^2_w	0.527	0.530	0.488
		APE	0.564	13.036	63.016
PCR	2	R^2	0.325	0.157	0.048
		APE	1.408	13.939	64.257
	5	R^2	0.433	0.195	0.062
		APE	1.461	14.094	64.257
CSPCR	2	R^2	0.403	0.188	0.059
		APE	1.539	17.337	69.994
	5	R^2	0.447	0.207	0.069
		APE	1.521	18.722	69.994
RCSPCR	2	R^2_w	0.436	0.443	0.411
		APE	1.541	16.710	65.793
	5	R^2_w	0.480	0.494	0.478
		APE	1.526	17.068	65.793

The Scenario-II is related to leverage points where outlyingness is introduced only in predictors without contaminating the response variable. It can be seen from Table 2 that leverage points and contamination fraction do not highly influence the values of R^2 in this scenario. In this case, the proposed methods (RCBR, RCSPCR) marginally give better performance in comparison with the existing ones (CBR, PCR, CSPCR) in terms of the coefficient of determination R^2. However, the existing methods (PCR, CBR, CSPCR) have smaller APE and variances of the estimates as compared to those of the proposed methods (RCBR, RCSPCR). It is important to note that CBR has the smallest APE and PCR has smallest variances of estimates (Table 2).

Table 2. Scenario-II: The results of simulation study in the presence of leverage points for n=500 and p=20 at different contamination levels.

	No. of retained regressors/PCs		Contamination levels		
Method	M		0%	2%	10%
CBR	2	R^2	0.395	0.369	0.402
		APE	0.611	0.662	0.629
	5	R^2	0.493	0.456	0.500
		APE	0.562	0.550	0.551
RCBR	2	R^2_w	0.432	0.403	0.435
		APE	0.612	0.663	0.630
	5	R^2_w	0.527	0.490	0.535
		APE	0.564	0.552	0.554
PCR	2	R^2	0.325	0.262	0.308
		APE	1.408	1.182	1.260
	5	R^2	0.433	0.386	0.399
		APE	1.461	1.388	1.388
CSPCR	2	R^2	0.403	0.366	0.401
		APE	1.539	1.581	2.110
	5	R^2	0.447	0.407	0.450
		APE	1.520	1.685	2.294
RCSPCR	2	R^2_w	0.436	0.412	0.445
		APE	1.541	2.755	19.678
	5	R^2_w	0.480	0.462	0.505
		APE	1.526	3.230	10.261

In Scenario-III, both the outliers and leverage points are infused in the response variable and the explanatory variables. In this contamination context, the values of coecient of determination R^2 for existing methods (PCR, CSPCR and CBR) are highly aected due to their sensitivity to the outliers. However, the proposed methods (RCBR, RCSPCR) successfully overcome the impact of these contaminated values in terms of R^2. The values of R^2 decrease by increasing the contamination fraction in existing methods but our proposed methods are independent of these contaminated entries due to less sensitivity of outlyingness (Table 3). The APE associated with the proposed RCBR is marginally better than the existing methods (CBR, PCR, CSPCR). RCBR is more efficient as compared to those of CBR, CSPCR (Table 3). Although, PCR has smallest variances of regression coefficients but it does not perform well in terms of coefficient of determination R^2 and APE being compared with CBR and RCBR. In this scenario, CSPCR and RCSPCR give poor performance in terms of APE. In order to see the impact of small sample size ($n = 50$), the proposed and existing methods were compared for Scenario-I to Scenario-III. Reducing the sample size does not have significant impact on the results stated above (Appendix B). Additionally, error distribution was varied for Scenario-I to Scenario-IV. These results are presented in Appendix B as well. It can be observed that change of error distribution (Cauchy distribution (0, 1)) does not change the pattern of results of the proposed and existing methods (Appendix B).

Table 3. Scenario-III: The results of simulation study regarding the Scenario-III for n=500 and p= 20 at different contamination levels.

	No. of retained regressors/PCs		Contamination levels		
Method	M		0%	2%	10%
CBR	2	R^2	0.395	0.146	0.063
		APE	0.611	13.414	63.98
	5	R^2	0.493	0.179	0.078
		APE	0.562	13.344	64.238
RCBR	2	R^2_w	0.432	0.378	0.383
		APE	0.612	13.327	63.596
	5	R^2_w	0.527	0.468	0.486
		APE	0.564	13.376	63.987
PCR	2	R^2	0.325	0.106	0.047
		APE	1.408	13.959	64.649
	5	R^2	0.433	0.149	0.060
		APE	1.461	14.076	65.916
CSPCR	2	R^2	0.403	0.144	0.062
		APE	1.539	23.503	99.652
	5	R^2	0.447	0.160	0.072
		APE	1.520	23.927	112.29
RCSPCR	2	R^2_w	0.436	0.387	0.415
		APE	1.541	42.701	126.637
	5	R^2_w	0.480	0.425	0.461
		APE	1.526	87.742	145.059

In Scenario-IV, several contamination models were taken with NAs, row-wise outliers, cellwise outliers and combinations of them. For brevity, a few of them are presented here. The proposed method MacroRCSPCR successfully handles the several contamination levels of row-wise and cellwise outliers and the fraction of missingness. It can be observed from Table 4, the values of R^2 are not affected by these contaminated values for most of the contamination models. However, when the fraction of the row-wise and cellwise outliers exceed 30% the values of R^2 are highly influenced. The proposed method performs well when the percentage of row-wise and cellwise outliers is less than 30% and the fraction of missingness is less than 20%. APE and variances increase for increasing contamination levels (row and column wise both) (Table 4).

Table 4. Scenario-IV: The results of simulation study in the presence of Macro for n = 500 and p = 20 at different contamination levels

Method	No. of retained regressors/PCs		Cellwise contamination		
MacroRCSPCR	M		0%	15%	30%
Rowwise outliers=0%	2	R^2_w	0.412	0.402	0.377
NAs=0%		APE	0.653	51.099	53.477
	5	R^2_w	0.475	0.459	0.416
		APE	0.780	53.619	123.374
Rowwise outliers=20%	2	R^2_w	0.317	0.318	0.252
NAs=5%		APE	54.438	52.407	1721.450
	5	R^2_w	0.375	0.366	0.300
		APE	55.359	189.013	1275.218
Rowwise outliers=20%	2	R^2_w	0.316	0.325	0.274
NAs=15%		APE	50.345	82.384	2494.971
	5	R^2_w	0.361	0.367	0.319
		APE	53.053	109.374	1888.149

4.2. Application

Our study shows interesting results of multiple regression techniques, whose results may be a novelty in the environmental field, specifically in tropical soils, as shown by scientific studies with multiple regression techniques (e.g., [50,59]). However, beyond its application in the specific case of the mathematical-statistical field and a delimitation to a type of problem (levels of contamination), this work develops a scientific logic that we believe is easily exportable to other areas, not only the environmental but agriculture in general (e.g., [13, 49, 51, 52]), social sciences (e.g., [45, 53]), economics [58] which use multivariate techniques and machine learning for prediction purposes.

We evaluate the performance of the proposed method RCSPCR and RCBR on the Cancer data set taken from https://data.world/exercises/linear-regression-exercise-1/workspace/data-dictionary (see, Table 5). American community survey (census.gov, clinicaltrials.gov and cancer.gov) are the sources that were being used to generate this data set. The details of the data preparation procedure can be found at https://data.world/nrippner/cancer-trials. The dataset aims to explore the incidence and mortality due to cancer across the world with respect to various demographic variables. This dataset contains a sample of 3047 patients on 34 variables. 27 of these variables are taken into account to perform regression. Target_deathrate (mean per capita (100,000) cancer mortalities) is taken as the response variable. It is predicted by 27 explanatory variables that are incidenceRate (mean per capita (100,000) cancer diagnoses), medIncome (median income per county), povertyPercent (percent of populace in poverty), MedianAge (median age of county residents), MedianAgeMale (median age of male county residents) , MedianAgeFemale (median age of female county residents), AvgHouseholdSize (average household size of occupied housing units), PercentMarried (percent of county residents who are married), PctNoHS18_24 (percent of county residents aged 18-24 highest education attained: less than high school), PctHS18_24 (percent of county residents aged 18-24 highest education attained: high school diploma), PctBachDeg18_24 (percent of county residents aged 18-24 highest education attained: bachelors degree), PctHS25_Over (percent of county residents aged 25 and over highest education attained: high school diploma), PctBachDeg25_Over (percent of county residents aged 25 and over highest education attained: bachelors degree), PctUnemployed16_Over (percent of county residents aged 16 and over unemployed), PctPrivateCoverage (percent of county residents with private health coverage), PctEmpPrivCoverage (percent of county residents with employee-provided private health coverage), PctPublicCoverage (percent of county residents with government-provided health coverage), PctPublicCoverageAlone (percent of county residents with government-provided health coverage alone), PctWhite (percent of county residents who identify as white), PctBlack (percent of county residents who identify as black), PctAsian (percent of county residents who identify as Asian), PctOtherRace (percent of county residents who identify in a category which is not White, Black, or Asian), PctMarriedHouseholds (percent of married households), BirthRate (number of live births relative to number of women in county), avgAnnCount (mean number of reported cases of cancer diagnosed annually), avgDeathsPerYear (mean number of reported mortalities due to cancer), popEst2015 (population of county). The response variable has almost 6% outliers which are detected by studentized residuals [9] and explanatory variables have 9% leverage points which are diagnosed by Cooks squared distance [16]. Seven PCs are retained and the values of R^2, APE and variance-covariance matrix of regression coefficients are tabulated in Table 5. The results indicate that proposed methods RCBR and RCSPCR outperform other competing methods (CBR, CSPCR and PCR) in terms of R^2. The proposed methods (RCBR and RCSPCR) give large values of the APE and variances of the estimates as compared to those of CBR, CSPCR and PCR (Table 5).

Table 5. The results from the cancer dataset.

Scenario-III Method	R^2	APE	$\text{Cov}(\hat{\beta})$
CBR	0.479	0.5246	$\begin{bmatrix} 0.041 & – & – & – & – & – & – \\ -0.024 & 0.038 & – & – & – & – & – \\ -0.003 & -0.003 & 0.020 & – & – & – & – \\ -0.001 & -0.003 & -0.004 & 0.018 & – & – & – \\ -0.002 & 0.001 & -0.003 & -0.005 & 0.018 & – & – \\ 0.001 & 0.005 & -0.002 & -0.001 & -0.003 & 0.020 & – \\ 0.001 & -0.004 & -0.0022 & -0.003 & -0.002 & -0.011 & 0.023 \end{bmatrix}$
RCBR	0.593	0.5275	$\begin{bmatrix} 0.041 & – & – & – & – & – & – \\ -0.025 & 0.038 & – & – & – & – & – \\ -0.004 & -0.004 & 0.021 & – & – & – & – \\ -0.002 & -0.002 & -0.003 & 0.023 & – & – & – \\ 0.000 & 0.002 & -0.002 & -0.003 & 0.027 & – & – \\ 0.000 & 0.007 & -0.003 & -0.001 & -0.003 & 0.039 & – \\ 0.001 & -0.003 & -0.001 & -0.004 & -0.003 & -0.021 & 0.038 \end{bmatrix}$
PCR	0.466	1.3341	$\begin{bmatrix} 0.032 & – & – & – & – & – & – \\ -0.000 & 0.000 & – & – & – & – & – \\ -0.004 & 0.000 & 0.004 & – & – & – & – \\ 0.017 & -0.006 & -0.006 & 0.029 & – & – & – \\ 0.006 & 0.001 & -0.001 & 0.009 & 0.018 & – & – \\ 0.000 & 0.000 & -0.000 & 0.002 & 0.003 & 0.007 & – \\ 0.002 & 0.000 & 0.000 & -0.003 & -0.005 & 0.000 & 0.020 \end{bmatrix}$
CSPCR	0.465	1.0669	$\begin{bmatrix} 0.248 & – & – & – & – & – & – \\ -0.003 & 0.028 & – & – & – & – & – \\ -0.012 & 0.148 & 0.900 & – & – & – & – \\ -0.003 & 0.001 & 0.009 & 0.015 & – & – & – \\ 0.000 & -0.033 & -0.209 & -0.001 & 0.057 & – & – \\ -0.002 & 0.013 & 0.083 & 0.000 & -0.018 & 0.012 & – \\ 0.002 & -0.010 & -0.061 & 0.000 & 0.013 & -0.005 & 0.013 \end{bmatrix}$
RCSPCR	0.574	1.0816	$\begin{bmatrix} 0.236 & – & – & – & – & – & – \\ -0.002 & 0.025 & – & – & – & – & – \\ -0.010 & 0.134 & 0.902 & – & – & – & – \\ -0.005 & -0.017 & -0.006 & 0.082 & – & – & – \\ 0.003 & -0.044 & -0.301 & 0.006 & 0.110 & – & – \\ 0.003 & -0.000 & -0.001 & -0.000 & -0.000 & 0.006 & – \\ 0.001 & 0.007 & 0.053 & 0.001 & -0.018 & -0.009 & 0.010 \end{bmatrix}$

To examine the method MacroRCSPCR on a real dataset, top gear data [4] were taken. This data set is available in R-package, i.e., robustHD. It is based on the BBC television show Top Gear and consists of variables that specify features of cars. In this dataset, 297 cars with 32 variables are given. Only 11 continuous variables are price (in UK pounds), displacement (the displacement of the engine in cc), BHP (the power of the engine in bhp), torque (the torque of the engine in lb/ft), acceleration (the time it takes the car to get from 0 to 62 mph in seconds), topSpeed (the car's top speed in mph), MPG (the combined fuel consumption (urban + extra urban; in miles per gallon)), weight (the cars curb weight in kg), length (the cars length in mm), width (the cars width in mm), height (the cars height in mm) are used here in this application. Five variables price, displacement, BHP, torque, topspeed are highly skewed. Therefore, these highly skewed variables are logarithmically transformed. This dataset consists of 95 missing cells which is 2.9% of total cells and also

there are some cellwise outliers whose number is very small as compared to the dataset. These cellwise outliers are detected through DetectDeviatingCells (DDC) algorithm [61]. The variable price is used as the dependent variable and all other variables are used as independent variables here. With two retained PCs, the calculated values of coefficient of determination R^2 and APE are 0.90031 and 107.380 respectively. The values in a variance-covariance matrix of regression coefficients are very close to zero.

5. Conclusion and discussion

In this paper, three approaches are proposed that can deal with multicollinearity, high dimensionality and missingness and outlyingness. The methods developed to handle these features are Robust Correlation Scaled Principal Component Regression (RCSPCR) (Subsection 2.1, Robust Correlation Based Regression (RCBR) (Subsection 2.2) and MacroRCSPCR (Subsection 2.3). RCBR and RCSPCR which are the robust version of two existing methods CBR and CSPCR are proposed. The properties of the proposed methods were observed by varying the form and level of contamination, sample size, retained regressors and error distribution. In the presence of outliers only, RCSPCR outperforms all the other methods being compared. It is found to be the most accurate and efficient method of PCR in the presence of outliers. Leverage points and influential observations are found to be the best handled by CBR and RCBR respectively. MacroRCSPCR is just a variant of RCSPCR under missing observations and two-dimensional outliers. Therefore, there is potential to improve or modify the proposed methods to cope with leverage points as well.

References

[1] H. Abdi and L.J. Williams, *Principal component analysis*, Wiley Interdiscip. Rev. Comput. Stat. **2** (4), 433-459, 2010.

[2] C. Agostinelli, A. Leung, V.J. Yohai and R.H. Zamar, *Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination*, Test **24** (3), 441-461, 2015.

[3] M.H. Ahmad, R. Adnan and N. Adnan, *A comparative study on some methods for handling multicollinearity problems*, Matematika (Johor) **22** (2), 109-119, 2006.

[4] A. Alfons, Package "robustHD: Robust methods for high-dimensional data", R package version: 0.5.1, 2016.

[5] A. Alin, *Multicollinearity*, Wiley Interdiscip. Rev. Comput. Stat. **2** (3), 370-374, 2010.

[6] O.G. Alma, *Comparison of robust regression methods in linear regression*, Int. J. Contemp. Math. Sciences **6** (9), 409-421, 2011.

[7] F. Alqallaf, S. Van Aelst, V.J. Yohai and R.H. Zamar, *Propagation of outliers in multivariate data*, Ann. Statist. **37** (1), 311-31, 2009.

[8] R. Andersen, *Modern Methods for Robust Regression*, Sage, 2008.

[9] A.C. Atkinson, *Regression diagnostics, transformations and constructed variables*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **44** (1), 1-22, 1982.

[10] V. Barnett and T. Lewis, *Outliers in Statistical Data*, John Wiley & Sons, New York, 1994.

[11] D.A. Belsley, E. Kuh and R.E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley & Sons, 2005.

[12] D. Blatná, *Outliers in regression*, Trutnov **30**, 1-6, 2006.

[13] B. Campos, F. Paredes, J. Rey, D. Lobo and S. Galvis-Causil, *The relationship between the normalized difference vegetation index, rainfall, and potential evapotranspiration in a banana plantation of Venezuela*, Sains Tanah **18** (1), 58-64, 2021.

[14] S. Chatterjee and A.S. Hadi, *Influential observations, high leverage points, and outliers in linear regression*, Statist. Sci. **1** (3), 379-393, 1986.

[15] S. Chatterjee and A. S. Hadi, *Regression Analysis by Example*, John Wiley & Sons, 2015.
[16] R.D. Cook, *Detection of influential observation in linear regression*, Technometrics **19** (1), 15-18, 1977.
[17] R.D. Cook and S. Weisberg, *Residuals and Influence in Regression*, Chapman and Hall, New York, 1982.
[18] C. Daniel and F.S. Wood, *Fitting Equations to Data: Computer Analysis of Multifactor Data*, John Wiley & Sons, 1980.
[19] M. Denhere and N. Billor, *Robust principal component functional logistic regression*, Comm. Statist. Simulation Comput. **45** (1), 264-281, 2016.
[20] A.K. Dey, M.A. Hossain and K.P. Das, *Regression analysis for data containing outliers and high leverage points*, Ala. J. Math **39**, 1-6, 2015.
[21] N.R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley & Sons, 1998.
[22] F.Y. Edgeworth, *On observations relating to several quantities*, Hermathena **6** (13), 279285, 1887.
[23] W.J. Egan and S.L. Morgan, *Outlier detection in multivariate analytical chemical data*, Anal. Chem. **70** (11), 2372-2379, 1998.
[24] S. Engelen, M. Hubert, K.V. Branden and S. Verboven, *Robust PCR and Robust PLSR: a comparative study*, in: Theory and Applications of Recent Robust Methods, 105-117, Birkhäuser, Basel, 2004.
[25] P. Filzmoser, *Robust principal component regression*, in: Proceedings of the Sixth International Conference on Computer Data Analysis and Modeling, 132137, Minsk, Belarus, 2001.
[26] P. Gagnon, M. Bédard and A. Desgagné, *An automatic robust Bayesian approach to principal component regression*, J. Appl. Stat. **48** (1), 84-104, 2021.
[27] D.N. Gujarati, *Basic Econometrics*, Tata McGraw-Hill Education, New York, 2009.
[28] L.C. Hamilton, *Statistics with Stata: Version 12*, Cengage Learning, 2012.
[29] H.V. Henderson and P.F. Velleman, *Building multiple regression models interactively*, Biometrics **37** (2), 391-411, 1981.
[30] D.C. Hoaglin and R.E. Welsch, *The hat matrix in regression and ANOVA*, Amer. Statist. **32** (1), 17-22, 1987.
[31] R.R. Hocking, *Developments in linear regression methodology: 19591982*, Technometrics **25** (3), 219230, 1983.
[32] A.E. Hoerl and R.W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics **12** (1), 5567, 1970.
[33] P.J. Huber, *Robust estimation of a location parameter*, Ann. Math. Statist. **35**, 73-101, 1964.
[34] M. Hubert, P.J. Rousseeuw and W. Van den Bossche, *MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers*, Technometrics **61** (4), 459-473, 2019.
[35] M. Hubert, P.J. Rousseeuw and K.V. Branden, *ROBPCA: a new approach to robust principal component analysis*, Technometrics **47** (1), 64-79, 2005.
[36] M. Hubert and S. Verboven, *A robust PCR method for highdimensional regressors*, J. Chemom. **17** (89) 438-452, 2003.
[37] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, Pearson, London, 2014.
[38] I.T. Jolliffe, *A note on the use of principal components in regression*, J. R. Stat. Soc. Ser. C. Appl. Stat. **31** (3), 300-303, 1982.
[39] I.T. Jolliffe, *Principal components in regression analysis*, in: Principal Component Analysis, 129-155, Springer, New York, 1986.
[40] I.T. Jolliffe, *Principal component analysis*, in: Encyclopedia of Statistics in Behavioural Science, John Wiley & Sons, 2005.

[41] H.A. Kiers, *Weighted least squares fitting using ordinary least squares algorithms*, Psychometrika **62** (2), 251-266, 1997.

[42] M.R. Lavery, P. Acharya, S.A. Sivo and L. Xu, *Number of predictors and multicollinearity: What are their effects on error and bias in regression?*, Comm. Statist. Simulation Comput. **48** (1), 27-38, 2019.

[43] G. Li and Z. Chen, *Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo*, J. Amer. Statist. Assoc. **80** (391), 759-66, 1985.

[44] D.C. Montgomery and A.E. Peck, *Introduction to Linear Regression Analysis*, John Wiley & Sons, New York, 1982.

[45] E. Montenegro, J. Pitti and O.B. Olivares, *Identification of the main subsistence crops of Teribe: a case study based on multivariate techniques*, Idesia **39** (3), 83-94, 2021.

[46] P.R. Nelson, P.A. Taylor and J.F. MacGregor, *Missing data methods in PCA and PLS: Score calculations with incomplete observations*, Chemom. Intell. Lab. Syst. **35** (1), 45-65, 1996.

[47] J. Neter, M.H. Kutner, C.J. Nachtsheim and W. Wasserman, *Applied Linear Statistical Models*, McGraw-Hill, New York, 1996.

[48] K. Ntotsis and A. Karagrigoriou, *The impact of multicollinearity on big data multivariate analysis modeling*, in: I. Dimotikalis, A. Karagrigoriou, C. Parpoula and C. Skiadas (ed.) Applied Modeling Techniques and Data Analysis 1: Computational Data Analysis methods and Tools, ISTE, 2021.

[49] O.B. Olivares, *Determination of the Potential Influence of Soil in the Differentiation of Productivity and in the Classification of Susceptible Areas to Banana Wilt in Venezuela*, 89-111, UCOPress, Spain, 2022.

[50] O.B. Olivares, M. Araya-Alman and C. Acevedo-Opazo, *Relationship between soil properties and banana productivity in the two main cultivation areas in Venezuela*, Soil Sci. Plant Nutr. **20** (3), 2512-2524, 2020.

[51] O.B. Olivares, J. Calero, J.C. Rey, D. Lobo, B.B. Landa and J.A. Gómez, *Correlation of banana productivity levels and soil morphological properties using regularized optimal scaling regression*, Catena **208**, 105718, 2022.

[52] O.B. Olivares and R. Hernández, *Application of multivariate techniques in the agricultural lands aptitude in Carabobo, Venezuela*, Trop. Subtrop. Agroecosystems **23** (2), 1-12, 2020.

[53] O.B. Olivares, J. Pitti and E. Montenegro, *Socioeconomic characterization of Bocas del Toro in Panama: an application of multivariate techniques*, Rev. Bras. de Gestao e Desenvolv. Reg. **16** (3), 59-71, 2020.

[54] S. Paul, *Sequential detection of unusual points in regression*, J. R. Stat. Soc. Ser. D. (The Statistician) **32** (4), 417-424, 1983.

[55] R.K. Paul, *Multicollinearity: Causes, effects and remedies*, IASRI, New Delhi, 2006.

[56] R.J. Pell, *Multiple outlier detection for multivariate calibration using robust statistical techniques*, Chemom. Intell. Lab. Syst. **52** (1), 87-104, 2000.

[57] D. Pena and V. Yohai, *A fast procedure for outlier diagnostics in large regression problems*, J. Amer. Statist. Assoc. **94** (446), 434-445, 1999.

[58] J. Pitti, O.B. Olivares and E. Montenegro, *The role of agriculture in the Changuinola District: a case of applied economics in Panama*, Trop. Subtrop. Agroecosystems **25** (1), 1-11, 2021.

[59] O. Renaud and M.P. Victoria-Feser, *A robust coefficient of determination for regression*, J. Statist. Plann. Inference **140** (7), 1852-1862, 2010.

[60] P.J. Rousseeuw, *Least median of squares regression*, J. Amer. Statist. Assoc. **79** (388), 871-880, 1984.

[61] P.J. Rousseeuw and W.V.D. Bossche, *Detecting deviating data cells*, Technometrics **60** (2), 135-145, 2018.

[62] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, 1987.
[63] P.J. Rousseeuw and A.M. Leroy, *A robust scale estimator based on the shortest half*, Stat. Neerl. Statistica **42** (2), 103-116, 1988.
[64] P.J. Rousseeuw and B.C. Van Zomeren, *Unmasking multivariate outliers and leverage points*, J. Amer. Statist. Assoc. **85** (411), 633-639, 1990.
[65] P.J. Rousseeuw and V. Yohai, *Robust regression by means of S-estimators*, in: Robust and Nonlinear Time Series Analysis, 256-272, Springer, New York, 1984.
[66] G.A. Seber and A.J. Lee, *Linear Regression Analysis*, John Wiley & Sons, 2012.
[67] S. Serneels and T. Verdonck, *Principal component analysis for data containing outliers and missing elements*, Comput. Statist. Data Anal. **52** (3), 1712-1727, 2008.
[68] N. Shrestha, *Detecting multicollinearity in regression analysis*, Am. J. Appl. Math. Stat. **8** (2), 39-42, 2020.
[69] A.F. Siegel and R.H. Benson, *A robust comparison of biological shapes*, Biometrics **38** (2), 341-350, 1982.
[70] K.K. Singh, A. Patel and C. Sadu, *Correlation scaled principal component regression*, in: Intelligent Systems Design and Applications, 17th International Conference on Intelligent Systems Design and Applications, 350-356, Springer, 2017.
[71] I. Stanimirova, M. Daszykowski and B. Walczak, *Dealing with missing values and outliers in principal component analysis*, Talanta **72** (1), 172-178, 2007.
[72] J.P. Stevens, *Outliers and influential data points in regression analysis*, Psychol. Bull. **95** (2), 334, 1984.
[73] M. Suhail, S. Chand and B.G. Kibria, *Quantile based estimation of biasing parameters in ridge regression model*, Comm. Statist. Simulation Comput. **49** (10), 2732-2744, 2020.
[74] Y. Susanti and H. Pratiwi, *M estimation, S estimation, and MM estimation in robust regression*, Int. J. Pure Appl. Math. **91** (3), 349-360, 2014.
[75] M.A. Ullah and G.R. Pasha, *The origin and developments of influence measures in regression*, Pakistan J. Statist. **25** (3), 2009.
[76] B. Walczak and D.L. Massart, *Robust principal components regression as a detection tool for outliers*, Chemom. Intell. Lab. Syst. **27** (1), 41-54, 1995.
[77] C. Yale and A.B. Forsythe, *Winsorized regression*, Technometrics **18** (3), 291-300, 1976.
[78] M.H. Zhang, Q.S. Xu and D.L. Massart, *Robust principal components regression based on principal sensitivity vectors*, Chemom. Intell. Lab. Syst. **67** (2), 175-185, 2003.
[79] N. Zhao, Q. Xu, M.L. Tang, B. Jiang, Z. Chen and H. Wang, *Highdimensional variable screening under multicollinearity*, Chemom. Intell. Lab. Syst. **9** (1), 1-11, 2020.

Appendix A. Computational steps involved in R^2, APE and variance-covariance matrix of regression co-efficients

A.1. The R^2 under RCBR is computed as,

(1) Generate the standardized data matrix $\mathbf{X}$ using Eq. (3.1, 3.2).
(2) Generate the response variable $\mathbf{y}$ using Eq. (3.3).
(3) Infuse the outliers, leverage points or influential observations according to Scenario I-III respectively.
(4) Compute the correlation of each predictor $X_1, X_2, , X_p$ with the response variable $\mathbf{y}$.
(5) Sort the above correlation vector in order of decreasing magnitude and then predictors are sorted out according to their correlations.
(6) Take ij of the variables with the highest correlations with the response variable.
(7) Obtain the least trimmed squares estimates using Eq. (1.7).
(8) Obtain the coefficient of determination R^2 and R_w^2 of the multiple regression model obtained in step 7 using Eq. (3.4) and Eq. (3.5) respectively.
(9) Iterate steps 1-7 for 100 Monte Carlo runs. The coefficient of determination is averaged value over 100 replications.

A.2. The APE under RCBR is computed as,

(1) Generate the standardized data matrix $\mathbf{X}$ using the Eq. (3.1, 3.2).
(2) Generate the response variable $\mathbf{y}$ using Eq. (3.3).
(3) Infuse the outliers, leverage points or influential observations according to Scenario I-III respectively.
(4) Define k-folds over the sample size. Leave the first fold. On the reduced data consisting of (k-1) folds, perform CBR (steps 4-7 mentioned above in the calculation of R^2). Then, estimate $\mathbf{y}$ for the observations in the first fold. This process continues until we have estimated for each of the k-dropped folds.
(5) Iterate the steps 1-4 for 100 Monte Carlo runs. APE is averaged value over 100 replications.

A.3. The variance-covariance matrix of regression coefficients under RCBR is computed as,

(1) Generate the standardized data matrix $\mathbf{X}$ using Eq. (3.1, 3.2).
(2) Generate the response variable $\mathbf{y}$ using Eq. (3.3).
(3) Infuse the outliers, leverage points or influential observations according to Scenario I-III respectively.
(4) Take a with replacement sample of observations from the data generated above. Apply CBR (steps 4-7 mentioned above in the calculation of R^2) to estimate regression coefficients. This process continues until we have estimates for each of the k-bootstrap samples. Calculate the variance-covariance matrix over k-bootstrap estimates of regression coefficients.
(5) Iterate steps 1-4 for 100 Monte Carlo runs. Bootstrap variance-covariance matrix is averaged over 100 replications.

Appendix B. Tables

Table B.1. Scenario-I: The variance covariance matrix of regression coefficients regarding simulation study for $n = 500$ and $p = 20$ at different contamination levels.

Method	No. of retained regressors/PCs M	Contamination levels 0%	2%	10%
CBR	2	$\begin{bmatrix} 0.001 & - \\ -0.001 & 0.005 \end{bmatrix}$	$\begin{bmatrix} 0.207 & - \\ -0.059 & 0.149 \end{bmatrix}$	$\begin{bmatrix} 0.199 & - \\ -0.015 & 0.182 \end{bmatrix}$
	5	$\begin{bmatrix} 0.007 & - & - & - & - \\ 0.002 & 0.004 & - & - & - \\ 0.003 & 0.001 & 0.015 & - & - \\ -0.002 & 0.001 & -0.001 & 0.040 & - \\ -0.009 & -0.011 & -0.015 & -0.028 & 0.068 \end{bmatrix}$	$\begin{bmatrix} 0.229 & - & - & - & - \\ 0.033 & 0.198 & - & - & - \\ -0.004 & -0.025 & 0.124 & - & - \\ -0.042 & -0.050 & -0.007 & 0.132 & - \\ -0.061 & -0.056 & -0.039 & -0.023 & 0.141 \end{bmatrix}$	$\begin{bmatrix} 0.266 & - & - & - & - \\ 0.034 & 0.213 & - & - & - \\ 0.034 & 0.010 & 0.240 & - & - \\ -0.049 & -0.064 & -0.067 & 0.383 & - \\ -0.074 & -0.024 & -0.097 & -0.137 & 0.348 \end{bmatrix}$
RCBR	2	$\begin{bmatrix} 0.003 & - \\ -0.002 & 0.007 \end{bmatrix}$	$\begin{bmatrix} 0.040 & - \\ -0.017 & 0.035 \end{bmatrix}$	$\begin{bmatrix} 0.023 & - \\ -0.014 & 0.020 \end{bmatrix}$
	5	$\begin{bmatrix} 0.009 & - & - & - & - \\ 0.001 & 0.006 & - & - & - \\ 0.003 & 0.001 & 0.020 & - & - \\ -0.000 & 0.000 & -0.002 & 0.047 & - \\ -0.011 & -0.011 & -0.019 & -0.034 & 0.078 \end{bmatrix}$	$\begin{bmatrix} 0.052 & - & - & - & - \\ 0.006 & 0.059 & - & - & - \\ -0.002 & -0.010 & 0.059 & - & - \\ -0.012 & -0.016 & -0.011 & 0.062 & - \\ -0.022 & -0.023 & -0.019 & -0.023 & 0.076 \end{bmatrix}$	$\begin{bmatrix} 0.042 & - & - & - & - \\ -0.006 & 0.048 & - & - & - \\ -0.007 & -0.007 & 0.053 & - & - \\ -0.013 & -0.016 & -0.015 & 0.074 & - \\ -0.013 & -0.009 & -0.014 & -0.031 & 0.070 \end{bmatrix}$
PCR	2	$\begin{bmatrix} 0.001 & - \\ 0.000 & 0.022 \end{bmatrix}$	$\begin{bmatrix} 0.007 & - \\ 0.003 & 0.094 \end{bmatrix}$	$\begin{bmatrix} 0.060 & - \\ 0.005 & 0.198 \end{bmatrix}$
	5	$\begin{bmatrix} 0.001 & - & - & - & - \\ 0.000 & 0.026 & - & - & - \\ -0.000 & -0.000 & 0.022 & - & - \\ -0.000 & -0.000 & 0.001 & 0.037 & - \\ -0.000 & -0.003 & 0.001 & -0.002 & 0.044 \end{bmatrix}$	$\begin{bmatrix} 0.008 & - & - & - & - \\ 0.002 & 0.077 & - & - & - \\ -0.000 & -0.003 & 0.090 & - & - \\ 0.000 & 0.008 & 0.012 & 0.089 & - \\ 0.000 & 0.005 & 0.006 & 0.001 & 0.103 \end{bmatrix}$	$\begin{bmatrix} 0.069 & - & - & - & - \\ 0.003 & 0.236 & - & - & - \\ -0.004 & 0.019 & 0.226 & - & - \\ -0.003 & 0.015 & 0.011 & 0.243 & - \\ 0.009 & -0.002 & -0.023 & 0.023 & 0.265 \end{bmatrix}$
CSPCR	2	$\begin{bmatrix} 0.417 & - \\ 0.074 & 0.056 \end{bmatrix}$	$\begin{bmatrix} 6.133 & - \\ 5.340 & 6.116 \end{bmatrix}$	$\begin{bmatrix} 17.975 & - \\ 8.379 & 18.473 \end{bmatrix}$
	5	$\begin{bmatrix} 0.423 & - & - & - & - \\ 0.103 & 0.077 & - & - & - \\ -0.300 & -0.076 & 0.277 & - & - \\ 0.385 & 0.091 & -0.276 & 0.508 & - \\ -0.872 & -0.209 & 0.611 & -0.902 & 2.701 \end{bmatrix}$	$\begin{bmatrix} 6.291 & - & - & - & - \\ 5.263 & 5.837 & - & - & - \\ -1.189 & -0.834 & 8.393 & - & - \\ 3.807 & 4.194 & 0.668 & 9.537 & - \\ -2.504 & -2.675 & -0.046 & -0.801 & 9.144 \end{bmatrix}$	$\begin{bmatrix} 19.744 & - & - & - & - \\ 9.814 & 24.386 & - & - & - \\ -2.547 & -2.588 & 22.479 & - & - \\ 1.432 & 0.090 & -0.284 & 25.245 & - \\ -1.426 & 0.208 & -0.635 & 2.145 & 31.087 \end{bmatrix}$
RCSPCR	2	$\begin{bmatrix} 0.412 & - \\ 0.102 & 0.129 \end{bmatrix}$	$\begin{bmatrix} 0.217 & - \\ 0.496 & 2.532 \end{bmatrix}$	$\begin{bmatrix} 0.225 & - \\ 0.069 & 1.093 \end{bmatrix}$
	5	$\begin{bmatrix} 0.437 & - & - & - & - \\ 0.118 & 0.124 & - & - & - \\ -0.295 & -0.079 & 0.284 & - & - \\ 0.255 & 0.066 & -0.179 & 0.309 & - \\ -0.832 & -0.200 & 0.552 & -0.548 & 2.661 \end{bmatrix}$	$\begin{bmatrix} 0.189 & - & - & - & - \\ 0.479 & 3.144 & - & - & - \\ 0.030 & 0.133 & 0.940 & - & - \\ 0.683 & 4.143 & 0.191 & 9.429 & - \\ -0.382 & -2.722 & -0.269 & -3.643 & 6.438 \end{bmatrix}$	$\begin{bmatrix} 0.193 & - & - & - & - \\ 0.097 & 0.843 & - & - & - \\ 0.075 & 0.106 & 1.821 & - & - \\ -0.006 & 0.102 & -0.153 & 3.123 & - \\ 0.026 & 0.037 & -0.164 & -0.439 & 3.884 \end{bmatrix}$

Table 2. Scenario-II: The variance covariance matrix of regression coefficients regarding simulation study for $n = 500$ and $p = 20$ at different contamination levels.

Method	No. of retained regressors/PCs M	Contamination levels 0%	2%	10%
CBR	2	$\begin{bmatrix} 0.001 & - \\ -0.001 & 0.005 \end{bmatrix}$	$\begin{bmatrix} 0.003 & - \\ -0.003 & 0.012 \end{bmatrix}$	$\begin{bmatrix} 0.003 & - \\ -0.002 & 0.008 \end{bmatrix}$
	5	$\begin{bmatrix} 0.007 & - & - & - & - \\ 0.002 & 0.004 & - & - & - \\ 0.003 & 0.001 & 0.015 & - & - \\ -0.002 & 0.001 & -0.001 & 0.040 & - \\ -0.009 & -0.011 & -0.015 & -0.028 & 0.068 \end{bmatrix}$	$\begin{bmatrix} 0.003 & - & - & - & - \\ 0.001 & 0.020 & - & - & - \\ 0.001 & -0.007 & 0.065 & - & - \\ -0.001 & -0.001 & -0.025 & 0.069 & - \\ -0.001 & -0.005 & -0.017 & -0.039 & 0.061 \end{bmatrix}$	$\begin{bmatrix} 0.008 & - & - & - & - \\ 0.000 & 0.020 & - & - & - \\ 0.000 & -0.009 & 0.071 & - & - \\ -0.003 & -0.004 & -0.018 & 0.069 & - \\ 0.002 & -0.002 & -0.026 & -0.037 & 0.062 \end{bmatrix}$
RCBR	2	$\begin{bmatrix} 0.003 & - \\ -0.002 & 0.007 \end{bmatrix}$	$\begin{bmatrix} 0.007 & - \\ -0.006 & 0.014 \end{bmatrix}$	$\begin{bmatrix} 0.007 & - \\ -0.004 & 0.008 \end{bmatrix}$
	5	$\begin{bmatrix} 0.009 & - & - & - & - \\ 0.001 & 0.006 & - & - & - \\ 0.003 & 0.001 & 0.020 & - & - \\ -0.000 & 0.000 & -0.002 & 0.047 & - \\ -0.011 & -0.011 & -0.019 & -0.034 & 0.078 \end{bmatrix}$	$\begin{bmatrix} 0.005 & - & - & - & - \\ 0.000 & 0.022 & - & - & - \\ 0.001 & -0.006 & 0.067 & - & - \\ -0.001 & -0.002 & -0.025 & 0.075 & - \\ -0.001 & -0.007 & -0.018 & -0.042 & 0.069 \end{bmatrix}$	$\begin{bmatrix} 0.008 & - & - & - & - \\ -0.001 & 0.023 & - & - & - \\ 0.001 & -0.010 & 0.080 & - & - \\ -0.000 & -0.004 & -0.020 & 0.072 & - \\ -0.003 & -0.004 & -0.027 & -0.038 & 0.064 \end{bmatrix}$
PCR	2	$\begin{bmatrix} 0.001 & - \\ 0.000 & 0.022 \end{bmatrix}$	$\begin{bmatrix} 0.000 & - \\ 0.000 & 0.011 \end{bmatrix}$	$\begin{bmatrix} 0.000 & - \\ 0.000 & 0.001 \end{bmatrix}$
	5	$\begin{bmatrix} 0.001 & - & - & - & - \\ 0.000 & 0.026 & - & - & - \\ -0.000 & -0.000 & 0.022 & - & - \\ -0.000 & -0.000 & 0.001 & 0.037 & - \\ -0.000 & -0.003 & 0.001 & -0.002 & 0.004 \end{bmatrix}$	$\begin{bmatrix} 0.000 & - & - & - & - \\ 0.000 & 0.001 & - & - & - \\ 0.000 & 0.000 & 0.023 & - & - \\ 0.000 & -0.000 & -0.001 & 0.030 & - \\ 0.000 & -0.000 & 0.000 & -0.002 & 0.041 \end{bmatrix}$	$\begin{bmatrix} 0.000 & - & - & - & - \\ 0.000 & 0.001 & - & - & - \\ 0.000 & 0.000 & 0.015 & - & - \\ 0.000 & 0.000 & 0.001 & 0.023 & - \\ 0.000 & -0.000 & 0.001 & -0.002 & 0.030 \end{bmatrix}$
CSPCR	2	$\begin{bmatrix} 0.417 & - \\ 0.074 & 0.056 \end{bmatrix}$	$\begin{bmatrix} 0.246 & - \\ 0.155 & 0.281 \end{bmatrix}$	$\begin{bmatrix} 0.058 & - \\ -0.080 & 0.466 \end{bmatrix}$
	5	$\begin{bmatrix} 0.423 & - & - & - & - \\ 0.103 & 0.077 & - & - & - \\ -0.300 & -0.076 & 0.277 & - & - \\ 0.385 & 0.091 & -0.276 & 0.508 & - \\ -0.872 & -0.209 & 0.611 & -0.902 & 2.701 \end{bmatrix}$	$\begin{bmatrix} 0.232 & - & - & - & - \\ -0.146 & 0.284 & - & - & - \\ -0.125 & 0.129 & 0.129 & - & - \\ 0.476 & -0.521 & -0.338 & 1.461 & - \\ -0.360 & 0.369 & 0.281 & -0.913 & 2.876 \end{bmatrix}$	$\begin{bmatrix} 0.065 & - & - & - & - \\ -0.091 & 0.442 & - & - & - \\ -0.048 & 0.205 & 0.142 & - & - \\ 0.191 & -0.753 & -0.361 & 1.453 & - \\ -0.173 & 0.606 & 0.303 & -1.083 & 3.005 \end{bmatrix}$
RCSPCR	2	$\begin{bmatrix} 0.412 & - \\ 0.102 & 0.129 \end{bmatrix}$	$\begin{bmatrix} 0.649 & - \\ 0.068 & 0.516 \end{bmatrix}$	$\begin{bmatrix} 0.544 & - \\ 0.668 & 0.689 \end{bmatrix}$
	5	$\begin{bmatrix} 0.437 & - & - & - & - \\ 0.118 & 0.124 & - & - & - \\ -0.295 & -0.079 & 0.284 & - & - \\ 0.255 & 0.066 & -0.179 & 0.309 & - \\ -0.832 & -0.200 & 0.552 & -0.548 & 2.661 \end{bmatrix}$	$\begin{bmatrix} 6.264 & - & - & - & - \\ 3.228 & 4.374 & - & - & - \\ -0.948 & -0.772 & 0.425 & - & - \\ 2.388 & 1.992 & -0.735 & 1.908 & - \\ -1.618 & -1.387 & 0.648 & -1.204 & 4.013 \end{bmatrix}$	$\begin{bmatrix} 3.143 & - & - & - & - \\ 0.188 & 0.867 & - & - & - \\ -0.234 & 0.293 & 0.268 & - & - \\ 0.154 & -0.497 & -0.344 & 0.859 & - \\ -0.219 & 0.462 & 0.422 & -0.820 & 2.895 \end{bmatrix}$

Table 3. Scenario-III: The variance covariance matrix of regression coefficients regarding simulation study for $n = 500$ and $p = 20$ at different contamination levels.

Method	No. of retained regressors/PCs M	Contamination levels 0%	2%	10%
CBR	2	$\begin{bmatrix} 0.001 & - \\ -0.001 & 0.005 \end{bmatrix}$	$\begin{bmatrix} 0.206 & - \\ -0.035 & 0.090 \end{bmatrix}$	$\begin{bmatrix} 0.190 & - \\ -0.023 & 0.161 \end{bmatrix}$
	5	$\begin{bmatrix} 0.007 & - & - & - & - \\ 0.002 & 0.004 & - & - & - \\ 0.003 & 0.001 & -0.015 & - & - \\ -0.002 & 0.001 & -0.001 & 0.040 & - \\ -0.009 & -0.011 & -0.015 & -0.028 & 0.068 \end{bmatrix}$	$\begin{bmatrix} 0.275 & - & - & - & - \\ -0.016 & 0.151 & - & - & - \\ -0.034 & -0.017 & 0.101 & - & - \\ -0.020 & -0.033 & -0.009 & 0.134 & - \\ -0.061 & -0.037 & -0.028 & -0.044 & 0.147 \end{bmatrix}$	$\begin{bmatrix} 0.272 & - & - & - & - \\ 0.066 & 0.253 & - & - & - \\ 0.026 & -0.010 & 0.266 & - & - \\ -0.053 & -0.087 & -0.047 & 0.334 & - \\ -0.099 & -0.0104 & -0.136 & -0.131 & 0.426 \end{bmatrix}$
RCBR	2	$\begin{bmatrix} 0.003 & - \\ -0.002 & 0.007 \end{bmatrix}$	$\begin{bmatrix} 0.048 & - \\ -0.019 & 0.034 \end{bmatrix}$	$\begin{bmatrix} 0.019 & - \\ -0.014 & 0.022 \end{bmatrix}$
	5	$\begin{bmatrix} 0.009 & - & - & - & - \\ 0.001 & 0.006 & - & - & - \\ 0.003 & 0.001 & 0.020 & - & - \\ -0.000 & 0.000 & -0.002 & 0.047 & - \\ -0.011 & -0.011 & -0.019 & -0.034 & 0.078 \end{bmatrix}$	$\begin{bmatrix} 0.074 & - & - & - & - \\ -0.013 & 0.080 & - & - & - \\ -0.016 & -0.025 & 0.068 & - & - \\ -0.014 & -0.005 & -0.015 & 0.077 & - \\ -0.012 & -0.026 & -0.011 & -0.031 & 0.075 \end{bmatrix}$	$\begin{bmatrix} 0.025 & - & - & - & - \\ -0.006 & 0.047 & - & - & - \\ -0.003 & -0.009 & 0.048 & - & - \\ -0.005 & -0.016 & -0.007 & 0.049 & - \\ -0.006 & -0.010 & -0.024 & -0.014 & 0.052 \end{bmatrix}$
PCR	2	$\begin{bmatrix} 0.001 & - \\ -0.000 & 0.022 \end{bmatrix}$	$\begin{bmatrix} 0.001 & - \\ 0.000 & 0.014 \end{bmatrix}$	$\begin{bmatrix} 0.004 & - \\ 0.003 & 0.098 \end{bmatrix}$
	5	$\begin{bmatrix} 0.001 & - & - & - & - \\ 0.000 & 0.026 & - & - & - \\ -0.000 & -0.000 & 0.022 & - & - \\ -0.000 & -0.000 & 0.001 & -0.037 & - \\ -0.000 & -0.003 & 0.001 & -0.002 & -0.044 \end{bmatrix}$	$\begin{bmatrix} 0.001 & - & - & - & - \\ 0.000 & 0.013 & - & - & - \\ -0.000 & -0.001 & 0.108 & - & - \\ 0.000 & -0.004 & -0.012 & 0.100 & - \\ -0.001 & -0.000 & -0.001 & -0.007 & 0.0889 \end{bmatrix}$	$\begin{bmatrix} 0.004 & - & - & - & - \\ -0.001 & 0.104 & - & - & - \\ 0.001 & 0.000 & 0.234 & - & - \\ 0.001 & 0.034 & -0.002 & 0.326 & - \\ -0.002 & 0.025 & -0.027 & 0.037 & 0.269 \end{bmatrix}$
CSPCR	2	$\begin{bmatrix} 0.417 & - \\ 0.074 & 0.056 \end{bmatrix}$	$\begin{bmatrix} 0.210 & - \\ 0.210 & 8.551 \end{bmatrix}$	$\begin{bmatrix} 7.519 & - \\ 2.917 & 47.678 \end{bmatrix}$
	5	$\begin{bmatrix} 0.0423 & - & - & - & - \\ 0.103 & 0.077 & - & - & - \\ -0.3000 & -0.076 & 0.277 & - & - \\ 0.385 & 0.091 & -0.276 & 0.508 & - \\ -0.872 & -0.209 & 0.611 & -0.902 & 2.701 \end{bmatrix}$	$\begin{bmatrix} 0.207 & - & - & - & - \\ 0.035 & 8.617 & - & - & - \\ 0.080 & 6.162 & 6.880 & - & - \\ -0.104 & -4.612 & -2.674 & 13.689 & - \\ -0.017 & 2.498 & 1.152 & 1.509 & 12.004 \end{bmatrix}$	$\begin{bmatrix} 6.552 & - & - & - & - \\ 2.120 & 51.585 & - & - & - \\ 0.653 & 3.658 & 15.715 & - & - \\ -0.218 & -9.866 & -4.121 & 23.102 & - \\ 0.991 & -1.280 & -2.68 & -2.866 & 29.520 \end{bmatrix}$
RCSPCR	2	$\begin{bmatrix} 0.437 & - \\ 0.102 & 0.129 \end{bmatrix}$	$\begin{bmatrix} 12.379 & - \\ 0.840 & 3.772 \end{bmatrix}$	$\begin{bmatrix} 0.231 & - \\ 0.149 & 0.738 \end{bmatrix}$
	5	$\begin{bmatrix} 0.437 & - & - & - & - \\ 0.118 & 0.124 & - & - & - \\ -0.295 & -0.079 & 0.284 & - & - \\ 0.255 & 0.066 & -0.179 & 0.309 & - \\ -0.832 & -0.200 & 0.552 & -0.548 & 2.661 \end{bmatrix}$	$\begin{bmatrix} 8.137 & - & - & - & - \\ 0.227 & 2.573 & - & - & - \\ 0.119 & 0.584 & 3.522 & - & - \\ 0.058 & -0.071 & -0.673 & .891 & - \\ -0.534 & -0.008 & 0.422 & 2.544 & 8.684 \end{bmatrix}$	$\begin{bmatrix} 0.143 & - & - & - & - \\ -0.069 & 0.234 & - & - & - \\ 0.029 & -0.041 & 0.804 & - & - \\ -0.084 & 0.153 & 0.042 & 2.338 & - \\ -0.023 & 0.029 & -0.026 & 0.057 & 1.464 \end{bmatrix}$

Table 4. Scenario-IV: The variance covariance matrix of regression coefficients regarding simulation study for $n = 500$ and $p = 20$ at different contamination levels.

Method	No. of retained regressors/PCs M	Contamination levels 0%	2%	10%
CBR	2	$\begin{bmatrix} 0.001 & - \\ -0.001 & 0.005 \end{bmatrix}$	$\begin{bmatrix} 0.207 & - \\ -0.059 & 0.149 \end{bmatrix}$	$\begin{bmatrix} 0.199 & - \\ -0.015 & 0.182 \end{bmatrix}$
	5	$\begin{bmatrix} 0.007 & - & - & - & - \\ 0.002 & 0.004 & - & - & - \\ 0.003 & 0.001 & 0.015 & - & - \\ -0.002 & 0.001 & -0.001 & 0.040 & - \\ -0.009 & -0.011 & -0.015 & -0.028 & 0.068 \end{bmatrix}$	$\begin{bmatrix} 0.229 & - & - & - & - \\ 0.033 & 0.198 & - & - & - \\ -0.004 & -0.025 & 0.124 & - & - \\ -0.042 & -0.050 & -0.007 & 0.132 & - \\ -0.061 & -0.056 & -0.039 & -0.023 & 0.141 \end{bmatrix}$	$\begin{bmatrix} 0.266 & - & - & - & - \\ 0.034 & 0.213 & - & - & - \\ 0.034 & 0.010 & 0.240 & - & - \\ -0.049 & -0.064 & -0.067 & 0.383 & - \\ -0.074 & -0.024 & -0.097 & 0.137 & 0.348 \end{bmatrix}$
RCBR	2	$\begin{bmatrix} 0.003 & - \\ -0.002 & 0.007 \end{bmatrix}$	$\begin{bmatrix} 0.040 & - \\ -0.017 & 0.035 \end{bmatrix}$	$\begin{bmatrix} 0.023 & - \\ -0.014 & 0.020 \end{bmatrix}$
	5	$\begin{bmatrix} 0.009 & - & - & - & - \\ 0.001 & 0.006 & - & - & - \\ 0.003 & 0.001 & 0.020 & - & - \\ -0.000 & 0.000 & -0.002 & 0.047 & - \\ -0.011 & -0.011 & -0.019 & -0.034 & 0.078 \end{bmatrix}$	$\begin{bmatrix} 0.052 & - & - & - & - \\ 0.006 & 0.059 & - & - & - \\ -0.002 & -0.010 & 0.059 & - & - \\ -0.012 & -0.016 & -0.011 & 0.062 & - \\ -0.022 & -0.023 & -0.019 & -0.023 & 0.076 \end{bmatrix}$	$\begin{bmatrix} 0.042 & - & - & - & - \\ -0.006 & 0.048 & - & - & - \\ -0.007 & -0.007 & 0.053 & - & - \\ -0.013 & -0.016 & -0.015 & 0.074 & - \\ -0.013 & -0.009 & -0.014 & -0.031 & 0.070 \end{bmatrix}$
PCR	2	$\begin{bmatrix} 0.001 & - \\ 0.000 & 0.022 \end{bmatrix}$	$\begin{bmatrix} 0.007 & - \\ 0.003 & 0.094 \end{bmatrix}$	$\begin{bmatrix} 0.060 & - \\ 0.005 & 0.198 \end{bmatrix}$
	5	$\begin{bmatrix} 0.001 & - & - & - & - \\ 0.000 & 0.026 & - & - & - \\ -0.000 & -0.000 & 0.022 & - & - \\ -0.000 & -0.000 & 0.001 & 0.037 & - \\ -0.000 & -0.003 & 0.001 & -0.002 & 0.044 \end{bmatrix}$	$\begin{bmatrix} 0.008 & - & - & - & - \\ 0.002 & 0.077 & - & - & - \\ -0.000 & -0.003 & 0.090 & - & - \\ 0.000 & 0.008 & 0.012 & 0.089 & - \\ 0.000 & 0.005 & 0.006 & 0.001 & 0.103 \end{bmatrix}$	$\begin{bmatrix} 0.069 & - & - & - & - \\ 0.003 & 0.236 & - & - & - \\ -0.004 & 0.019 & 0.226 & - & - \\ -0.003 & 0.015 & 0.011 & 0.243 & - \\ 0.009 & -0.002 & -0.023 & 0.023 & 0.265 \end{bmatrix}$
CSPCR	2	$\begin{bmatrix} 0.417 & - \\ 0.074 & 0.056 \end{bmatrix}$	$\begin{bmatrix} 6.133 & - \\ 5.340 & 6.116 \end{bmatrix}$	$\begin{bmatrix} 17.975 & - \\ 8.379 & 18.473 \end{bmatrix}$
	5	$\begin{bmatrix} 0.423 & - & - & - & - \\ 0.103 & 0.077 & - & - & - \\ -0.300 & -0.076 & 0.277 & - & - \\ 0.385 & 0.091 & -0.276 & 0.508 & - \\ -0.8720 & -0.209 & 0.611 & -0.902 & 2.701 \end{bmatrix}$	$\begin{bmatrix} 6.291 & - & - & - & - \\ 5.263 & 5.837 & - & - & - \\ -1.189 & -0.834 & 8.393 & - & - \\ 3.807 & 4.194 & 0.668 & 9.537 & - \\ -2.504 & -2.675 & -0.046 & -0.801 & 9.144 \end{bmatrix}$	$\begin{bmatrix} 19.744 & - & - & - & - \\ 9.814 & 24.386 & - & - & - \\ -2.547 & -2.588 & 22.479 & - & - \\ 1.432 & 0.090 & -0.284 & 25.245 & - \\ -1.426 & 0.208 & -0.635 & 2.145 & 31.087 \end{bmatrix}$
RCSPCR	2	$\begin{bmatrix} 0.412 & - \\ 0.102 & 0.129 \end{bmatrix}$	$\begin{bmatrix} 0.217 & - \\ 0.496 & 2.532 \end{bmatrix}$	$\begin{bmatrix} 0.225 & - \\ 0.069 & 1.093 \end{bmatrix}$
	5	$\begin{bmatrix} 0.437 & - & - & - & - \\ 0.118 & 0.124 & - & - & - \\ -0.295 & -0.079 & 0.284 & - & - \\ 0.255 & 0.066 & -0.179 & 0.309 & - \\ -0.832 & -0.200 & 0.552 & -0.548 & 2.661 \end{bmatrix}$	$\begin{bmatrix} 0.189 & - & - & - & - \\ 0.479 & 3.144 & - & - & - \\ 0.030 & 0.133 & 0.940 & - & - \\ 0.683 & 4.143 & 0.191 & 9.429 & - \\ -0.382 & -2.722 & -0.269 & -3.643 & 6.438 \end{bmatrix}$	$\begin{bmatrix} 0.193 & - & - & - & - \\ 0.097 & 0.843 & - & - & - \\ 0.075 & 0.106 & 1.821 & - & - \\ -0.006 & 0.102 & -0.153 & 3.123 & - \\ 0.026 & 0.047 & -0.164 & -0.439 & 3.884 \end{bmatrix}$

Table 5. The results of simulation study regarding the Scenario-I for $n = 50$, $p = 20$ and no. of retained regressors / PCs $(M) = 5$ at different contamination levels.

		Contamination levels														
Method		0%					2%					10%				
CBR	R^2	0.539					0.328					0.229				
	APE	0.601					15.106					76.541				
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	0.014	–	–	–	–	1.249	–	–	–	–	3.079	–	–	–	–
		−0.000	0.029	–	–	–	0.464	0.727	–	–	–	0.133	2.376	–	–	–
		−0.002	−0.002	0.035	–	–	0.063	−0.003	0.848	–	–	0.090	−0.260	2.743	–	–
		−0.004	−0.011	−0.014	0.059	–	−0.283	−0.282	−0.294	0.721	–	−0.113	−0.592	−0.311	2.637	–
		−0.005	−0.013	−0.011	−0.026	0.060	−0.101	−0.122	−0.208	−0.158	0.722	−0.322	−0.438	−0.465	−0.471	2.395
RCBR	R^2_w	0.539					0.580					0.521				
	APE	0.655					12.998					63.505				
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	0.075	–	–	–	–	0.241	–	–	–	–	0.302	–	–	–	–
		−0.020	0.095	–	–	–	−0.026	0.233	–	–	–	−0.022	0.307	–	–	–
		−0.017	−0.011	0.128	–	–	−0.025	−0.044	0.202	–	–	−0.094	−0.052	0.401	–	–
		−0.012	−0.025	−0.044	0.166	–	−0.093	−0.021	−0.035	0.214	–	−0.070	−0.063	−0.112	0.277	–
		−0.014	−0.033	−0.045	−0.050	0.152	−0.041	−0.055	−0.070	−0.044	0.220	−0.041	−0.071	−0.102	−0.024	0.285
PCR	R^2	0.497					0.270					0.521				
	APE	1.729					16.533					74.562				
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	0.042	–	–	–	–	0.037	–	–	–	–	0.328	–	–	–	–
		0.005	0.034	–	–	–	−0.036	0.198	–	–	–	−0.104	0.568	–	–	–
		−0.005	−0.001	0.053	–	–	0.013	−0.017	0.148	–	–	0.166	−0.057	0.918	–	–
		0.006	−0.002	−0.001	0.069	–	−0.003	0.004	0.001	0.206	–	−0.060	0.090	−0.061	1.363	–
		0.000	0.002	−0.001	0.000	0.065	0.016	−0.064	0.039	−0.009	0.406	−0.054	−0.016	0.097	0.035	1.819
CSPCR	R^2	0.521					0.305					0.212				
	APE	2.710					56.415					269.472				
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	0.394	–	–	–	–	3.647	–	–	–	–	38.743	–	–	–	–
		0.016	0.116	–	–	–	−0.431	50.199	–	–	–	−0.847	13.891	–	–	–
		−0.032	−0.016	0.102	–	–	0.928	6.328	34.383	–	–	0.709	−1.858	29.893	–	–
		0.019	0.033	−0.048	0.195	–	−0.345	−8.781	−1.449	23.480	–	4.622	1.498	−4.823	47.392	–
		−0.004	−0.030	0.047	−0.070	0.205	0.969	7.666	2.024	−1.025	23.914	4.724	−4.237	9.417	−15.993	74.948
RCSPCR	R^2_w	0.582					0.586					0.560				
	APE	2.799					22.711					73.783				
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	0.377	–	–	–	–	3.913	–	–	–	–	1.262	–	–	–	–
		0.025	0.184	–	–	–	0.220	8.141	–	–	–	−0.091	2.472	–	–	–
		−0.022	−0.036	0.287	–	–	−0.295	0.220	8.250	–	–	−0.167	0.212	4.364	–	–
		0.056	0.019	−0.049	0.395	–	−0.798	−1.959	0.621	8.407	–	0.019	−0.368	−0.113	6.858	–
		0.031	−0.012	0.094	−0.122	0.597	0.062	1.654	−0.054	−0.077	12.605	−0.339	−0.621	0.769	−0.984	10.340

Table 6. The results of simulation study regarding the Scenario-II for $n = 50$, $p = 20$ and $M = 5$ at different contamination levels.

Method		Contamination levels: 0%	2%	10%
CBR	R^2	0.539	0.521	0.558
	APE	0.601	1.241	0.596
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.014 & - & - & - & - \\ -0.000 & 0.029 & - & - & - \\ -0.002 & -0.002 & 0.035 & - & - \\ -0.004 & -0.011 & -0.014 & 0.059 & - \\ -0.005 & -0.013 & -0.011 & -0.026 & 0.060 \end{bmatrix}$	$\begin{bmatrix} 0.020 & - & - & - & - \\ -0.000 & 0.019 & - & - & - \\ -0.003 & -0.000 & 0.047 & - & - \\ -0.009 & -0.011 & -0.019 & 0.062 & - \\ -0.007 & -0.005 & -0.021 & -0.018 & 0.057 \end{bmatrix}$	$\begin{bmatrix} 0.021 & - & - & - & - \\ -0.000 & 0.029 & - & - & - \\ -0.005 & -0.008 & 0.053 & - & - \\ -0.007 & -0.009 & -0.019 & 0.066 & - \\ -0.002 & -0.006 & -0.015 & -0.030 & 0.059 \end{bmatrix}$
RCBR	R^2_w	0.595	0.585	0.608
	APE	0.655	1.349	0.765
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.075 & - & - & - & - \\ -0.020 & 0.095 & - & - & - \\ -0.017 & -0.011 & 0.128 & - & - \\ -0.012 & -0.025 & -0.044 & 0.166 & - \\ -0.014 & -0.033 & -0.045 & -0.050 & 0.152 \end{bmatrix}$	$\begin{bmatrix} 0.058 & - & - & - & - \\ -0.014 & 0.073 & - & - & - \\ -0.017 & -0.005 & 0.135 & - & - \\ -0.017 & -0.029 & -0.044 & 0.153 & - \\ -0.000 & -0.0156 & -0.051 & -0.048 & 0.142 \end{bmatrix}$	$\begin{bmatrix} 0.086 & - & - & - & - \\ -0.014 & 0.086 & - & - & - \\ -0.020 & -0.024 & 0.155 & - & - \\ -0.015 & -0.026 & -0.060 & 0.185 & - \\ -0.015 & -0.011 & -0.039 & -0.067 & 0.151 \end{bmatrix}$
PCR	R^2	0.497	0.449	0.494
	APE	1.729	2.479	1.484
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.042 & - & - & - & - \\ 0.005 & 0.034 & - & - & - \\ -0.005 & -0.001 & 0.053 & - & - \\ 0.006 & -0.002 & -0.001 & 0.069 & - \\ 0.000 & 0.002 & -0.001 & 0.000 & 0.065 \end{bmatrix}$	$\begin{bmatrix} 0.014 & - & - & - & - \\ 0.001 & 0.038 & - & - & - \\ -0.002 & 0.006 & 0.043 & - & - \\ 0.001 & 0.003 & -0.001 & 0.067 & - \\ 0.000 & -0.001 & -0.001 & 0.005 & 0.055 \end{bmatrix}$	$\begin{bmatrix} 0.000 & - & - & - & - \\ 0.000 & 0.045 & - & - & - \\ 0.000 & 0.002 & 0.038 & - & - \\ -0.000 & -0.002 & 0.00 & 0.058 & - \\ -0.000 & 0.005 & -0.002 & -0.001 & 0.063 \end{bmatrix}$
CSPCR	R^2	0.521	0.505	0.526
	APE	2.710	4.375	9.473
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.394 & - & - & - & - \\ 0.016 & 0.116 & - & - & - \\ -0.032 & -0.016 & 0.102 & - & - \\ 0.019 & 0.033 & -0.048 & 0.195 & - \\ -0.004 & -0.030 & 0.047 & -0.070 & 0.205 \end{bmatrix}$	$\begin{bmatrix} 0.384 & - & - & - & - \\ 0.013 & 0.075 & - & - & - \\ -0.029 & 0.005 & 0.114 & - & - \\ 0.000 & 0.011 & -0.024 & 0.132 & - \\ -0.012 & -0.005 & -0.040 & -0.020 & 0.217 \end{bmatrix}$	$\begin{bmatrix} 0.395 & - & - & - & - \\ 0.029 & 0.099 & - & - & - \\ -0.014 & -0.017 & 0.092 & - & - \\ 0.023 & 0.015 & -0.021 & 0.187 & - \\ 0.014 & -0.018 & 0.033 & -0.042 & 0.208 \end{bmatrix}$
RCSPCR	R^2_w	0.582	0.579	0.600
	APE	2.799	7.703	59.228
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.377 & - & - & - & - \\ 0.025 & 0.184 & - & - & - \\ -0.022 & -0.036 & 0.287 & - & - \\ 0.056 & 0.019 & -0.049 & 0.395 & - \\ 0.031 & -0.012 & 0.094 & -0.122 & 0.597 \end{bmatrix}$	$\begin{bmatrix} 0.422 & - & - & - & - \\ -0.118 & 1.369 & - & - & - \\ -0.073 & -0.072 & 0.429 & - & - \\ -0.007 & 0.102 & -0.066 & 0.414 & - \\ -0.007 & -0.199 & 0.089 & -0.087 & 0.482 \end{bmatrix}$	$\begin{bmatrix} 0.377 & - & - & - & - \\ 0.022 & 0.242 & - & - & - \\ -0.007 & -0.027 & 0.246 & - & - \\ -0.009 & -0.012 & -0.034 & 0.609 & - \\ 0.025 & -0.037 & -0.066 & 0.009 & 0.474 \end{bmatrix}$

Table 7. The results of simulation study regarding the Scenario-III for $n = 50$, $p = 20$ and $M = 5$ at different contamination levels.

Method		Contamination levels: 0%	2%	10%
CBR	R^2	0.539	0.319	0.217
	APE	0.601	15.802	76.363
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.014 & – & – & – & – \\ -0.000 & 0.029 & – & – & – \\ -0.002 & -0.002 & 0.035 & – & – \\ -0.004 & -0.011 & -0.014 & 0.059 & – \\ -0.005 & -0.013 & -0.011 & -0.026 & 0.060 \end{bmatrix}$	$\begin{bmatrix} 0.184 & – & – & – & – \\ 0.010 & 0.217 & – & – & – \\ -0.268 & -0.258 & 0.814 & – & – \\ -0.061 & -0.128 & 0.200 & 0.610 & – \\ -0.009 & -0.008 & -0.036 & -0.191 & 0.477 \end{bmatrix}$	$\begin{bmatrix} 2.801 & – & – & – & – \\ 0.190 & 2.701 & – & – & – \\ -0.118 & -0.447 & 3.122 & – & – \\ -0.255 & -0.291 & -0.885 & 2.683 & – \\ 0.102 & -0.411 & -0.718 & -0.648 & 2.488 \end{bmatrix}$
RCBR	R^2_w	0.595	0.571	0.544
	APE	0.655	14.234	66.296
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.075 & – & – & – & – \\ -0.020 & 0.095 & – & – & – \\ -0.017 & -0.011 & 0.128 & – & – \\ -0.012 & -0.025 & -0.044 & 0.166 & – \\ -0.014 & -0.033 & -0.045 & -0.050 & 0.152 \end{bmatrix}$	$\begin{bmatrix} 0.101 & – & – & – & – \\ -0.012 & 0.113 & – & – & – \\ -0.014 & -0.014 & 0.153 & – & – \\ -0.014 & -0.048 & 0.007 & 0.267 & – \\ -0.010 & -0.022 & -0.080 & -0.074 & 0.198 \end{bmatrix}$	$\begin{bmatrix} 0.250 & – & – & – & – \\ -0.003 & 0.306 & – & – & – \\ -0.056 & -0.080 & 0.324 & – & – \\ -0.076 & -0.067 & -0.052 & 0.301 & – \\ -0.023 & -0.101 & -0.095 & -0.073 & 0.310 \end{bmatrix}$
PCR	R^2	0.497	0.248	0.149
	APE	1.757	15.865	73.041
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.042 & – & – & – & – \\ 0.005 & 0.034 & – & – & – \\ -0.005 & -0.001 & 0.053 & – & – \\ -0.006 & -0.002 & -0.001 & 0.069 & – \\ -0.000 & 0.002 & -0.001 & 0.000 & 0.065 \end{bmatrix}$	$\begin{bmatrix} 0.019 & – & – & – & – \\ 0.000 & 0.041 & – & – & – \\ 0.005 & -0.022 & 0.308 & – & – \\ -0.001 & 0.011 & -0.039 & 0.266 & – \\ -0.000 & -0.006 & 0.012 & 0.004 & 0.342 \end{bmatrix}$	$\begin{bmatrix} 0.002 & – & – & – & – \\ -0.002 & 0.367 & – & – & – \\ 0.005 & -0.140 & 0.767 & – & – \\ 0.001 & 0.105 & 0.009 & 1.249 & – \\ -0.008 & 0.010 & -0.050 & -0.064 & 1.248 \end{bmatrix}$
CSPCR	R^2	0.521	0.291	0.208
	APE	2.937	114.744	225.19
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.394 & – & – & – & – \\ 0.016 & 0.116 & – & – & – \\ -0.032 & -0.016 & 0.102 & – & – \\ -0.019 & 0.033 & -0.048 & 0.195 & – \\ -0.004 & -0.030 & 0.047 & -0.070 & 0.205 \end{bmatrix}$	$\begin{bmatrix} 0.135 & – & – & – & – \\ -0.012 & 3.483 & – & – & – \\ -0.015 & -1.436 & 10.169 & – & – \\ -0.056 & -1.590 & 4.991 & 9.707 & – \\ 0.053 & 1.714 & -3.944 & -2.447 & 33.725 \end{bmatrix}$	$\begin{bmatrix} 0.335 & – & – & – & – \\ -0.351 & 39.425 & – & – & – \\ -0.055 & 4.320 & 14.234 & – & – \\ 0.612 & -2.343 & -2.504 & 27.987 & – \\ -0.368 & -0.050 & 2.823 & -1.680 & 46.899 \end{bmatrix}$
RCSPCR	R^2_w	0.582	0.548	0.561
	APE	3.129	259.880	178.307
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.377 & – & – & – & – \\ 0.025 & 0.184 & – & – & – \\ -0.022 & -0.036 & 0.287 & – & – \\ -0.056 & 0.019 & -0.049 & 0.395 & – \\ -0.031 & -0.012 & 0.094 & -0.122 & 0.597 \end{bmatrix}$	$\begin{bmatrix} 0.979 & – & – & – & – \\ -0.085 & 0.717 & – & – & – \\ 0.044 & -0.287 & 2.376 & – & – \\ -0.045 & -0.357 & -1.372 & 3.664 & – \\ 0.144 & -0.177 & 0.693 & -0.312 & 7.348 \end{bmatrix}$	$\begin{bmatrix} 5.440 & – & – & – & – \\ -0.259 & 1.995 & – & – & – \\ 0.240 & -0.196 & 5.624 & – & – \\ -0.123 & 0.313 & -1.260 & 7.579 & – \\ – & – & 0.254 & -0.171 & -0.344 & 0.711 & 8.702 \end{bmatrix}$

Table 8. The results of simulation study regarding the Scenario-I for $n = 500$, $p = 20$ and $M = 5$ at different contamination levels with error term from Cauchy distribution $(0, 1)$.

Method		Contamination levels: 0%	Contamination levels: 2%	Contamination levels: 10%
CBR	R^2	0.029	0.029	0.027
	APE	1.003	13.714	64.862
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.021 & - & - & - & - \\ 0.000 & 0.013 & - & - & - \\ -0.013 & -0.015 & 0.029 & - & - \\ 0.000 & 0.003 & -0.006 & 0.005 & - \\ -0.001 & 0.001 & -0.003 & -0.001 & 0.004 \end{bmatrix}$	$\begin{bmatrix} 0.138 & - & - & - & - \\ 0.018 & 0.091 & - & - & - \\ 0.003 & -0.001 & 0.069 & - & - \\ -0.033 & -0.020 & -0.019 & 0.094 & - \\ -0.025 & -0.027 & -0.021 & -0.022 & 0.079 \end{bmatrix}$	$\begin{bmatrix} 0.546 & - & - & - & - \\ 0.040 & 0.410 & - & - & - \\ -0.034 & -0.024 & 0.356 & - & - \\ -0.061 & -0.119 & -0.073 & 0.320 & - \\ -0.112 & -0.107 & -0.103 & 0.031 & 0.329 \end{bmatrix}$
RCBR	R_w^2	0.331	0.309	0.316
	APE	0.986	13.481	63.579
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.000 & - & - & - & - \\ 0.000 & 0.000 & - & - & - \\ 0.000 & 0.000 & 0.000 & - & - \\ 0.000 & 0.000 & 0.000 & 0.000 & - \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{bmatrix}$	$\begin{bmatrix} 0.000 & - & - & - & - \\ 0.000 & 0.000 & - & - & - \\ 0.000 & 0.000 & 0.000 & - & - \\ 0.000 & 0.000 & 0.000 & 0.000 & - \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{bmatrix}$	$\begin{bmatrix} 0.000 & - & - & - & - \\ 0.000 & 0.000 & - & - & - \\ 0.000 & 0.000 & 0.000 & - & - \\ 0.000 & 0.000 & 0.000 & 0.000 & - \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{bmatrix}$
PCR	R^2	0.020	0.020	0.019
	APE	1.025	13.737	64.619
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.000 & - & - & - & - \\ 0.000 & 0.002 & - & - & - \\ -0.000 & -0.000 & 0.002 & - & - \\ -0.000 & -0.000 & -0.000 & 0.000 & 0.003 \\ 0.000 & 0.000 & -0.000 & -0.000 & 0.003 \end{bmatrix}$	$\begin{bmatrix} 0.003 & - & - & - & - \\ -0.000 & 0.060 & - & - & - \\ -0.000 & -0.004 & 0.062 & - & - \\ 0.002 & 0.000 & -0.005 & -0.054 & - \\ 0.000 & 0.003 & 0.004 & -0.001 & 0.072 \end{bmatrix}$	$\begin{bmatrix} 0.030 & - & - & - & - \\ -0.004 & 0.286 & - & - & - \\ 0.001 & 0.025 & 0.282 & - & - \\ -0.002 & -0.004 & -0.016 & 0.301 & - \\ -0.007 & 0.035 & 0.016 & 0.025 & 0.358 \end{bmatrix}$
CSPCR	R^2	0.027	0.027	0.0255
	APE	0.562	0.550	0.551
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.242 & - & - & - & - \\ 1.182 & 6.419 & - & - & - \\ 1.000 & 5.274 & 8.471 & - & - \\ -0.447 & -2.119 & -2.918 & 3.882 & - \\ -0.019 & -0.052 & -0.216 & -0.355 & 2.118 \end{bmatrix}$	$\begin{bmatrix} 6.746 & - & - & - & - \\ 0.932 & 2.711 & - & - & - \\ -5.146 & -1.321 & 15.199 & - & - \\ 5.452 & 0.371 & -5.518 & 28.984 & - \\ -3.885 & -1.113 & 6.066 & -6.659 & 26.095 \end{bmatrix}$	$\begin{bmatrix} 29.161 & - & - & - & - \\ 6.290 & 28.939 & - & - & - \\ -5.504 & -2.780 & 25.028 & - & - \\ 11.244 & 0.512 & -7.505 & 43.784 & - \\ -5.866 & -2.970 & -3.941 & -0.390 & 42.848 \end{bmatrix}$
RCSPCR	R_w^2	0.341	0.316	0.331
	APE	1.079	13.746	63.674
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.002 & - & - & - & - \\ 0.003 & 0.025 & - & - & - \\ -0.000 & -0.000 & 0.008 & - & - \\ -0.001 & 0.014 & -0.002 & 3.882 & - \\ -0.001 & -0.006 & 0.001 & -0.355 & 2.118 \end{bmatrix}$	$\begin{bmatrix} 0.001 & - & - & - & - \\ -0.001 & 0.028 & - & - & - \\ 0.000 & -0.003 & 0.008 & - & - \\ -0.000 & 0.011 & -0.001 & 00.024 & - \\ -0.000 & -0.004 & 0.000 & -0.000 & 0.020 \end{bmatrix}$	$\begin{bmatrix} 0.001 & - & - & - & - \\ -0.000 & 0.007 & - & - & - \\ 0.000 & -0.001 & 0.010 & - & - \\ 0.001 & -0.001 & 0.001 & 0.009 & - \\ -0.001 & -0.000 & 0.000 & -0.001 & 0.020 \end{bmatrix}$

Table 9. The results of simulation study regarding the Scenario-II for $n = 500$, $p = 20$ and $M = 5$ at different contamination levels with error term from Cauchy distribution $(0, 1)$.

Method		contamination levels 0%	2%	10%
CBR	R^2	0.029	0.029	0.034
	APE	1.003	1.008	1.012
	$\text{Cov}(\hat{\boldsymbol{\beta}})$	0.021 − − − −	0.014 − − − −	0.004 − − − −
		0.000 0.013 − − −	0.005 0.003 − − −	−0.000 0.000 − − −
		−0.013 −0.015 0.029 − −	−0.007 −0.003 0.005 − −	−0.001 0.000 0.000 − −
		−0.000 0.003 −0.006 0.005 −	0.008 0.003 −0.005 0.008 −	0.007 −0.000 −0.002 0.019 −
		−0.001 0.001 −0.003 −0.001 0.004	−0.014 −0.005 0.007 −0.011 0.017	−0.007 0.000 0.003 0.019 0.020
RCBR	R^2_w	0.331	0.315	0.309
	APE	0.986	0.988	0.991
	$\text{Cov}(\hat{\boldsymbol{\beta}})$	0.000 − − − −	0.000 − − − −	0.000 − − − −
		0.000 0.000 − − −	0.000 0.000 − − −	0.000 0.000 − − −
		0.000 0.000 0.000 − −	0.000 0.000 0.000 − −	0.000 0.000 0.000 − −
		0.000 0.000 0.000 0.000 −	0.000 0.000 0.000 0.000 −	0.000 0.000 0.000 0.000 −
		0.000 0.000 0.000 0.000 0.000	0.000 0.000 0.000 0.000 0.000	0.000 0.000 0.000 0.000 0.000
PCR	R^2	0.020	0.019	0.021
	APE	1.025	1.019	1.022
	$\text{Cov}(\hat{\boldsymbol{\beta}})$	0.000 − − − −	0.004 − − − −	0.000 − − − −
		0.000 0.002 − − −	−0.001 0.001 − − −	0.000 0.000 − − −
		−0.000 −0.000 0.002 − −	0.000 −0.000 0.001 − −	0.000 0.000 0.002 − −
		−0.000 −0.000 0.000 0.003 −	0.000 −0.000 0.000 0.004 −	0.000 0.000 0.000 0.003 −
		0.000 0.000 −0.000 −0.000 0.003	−0.000 0.000 0.000 −0.000 0.005	0.000 0.000 0.001 0.000 0.005
CSPCR	R^2	0.027	0.026	0.030
	APE	1.366	1.437	1.700
	$\text{Cov}(\hat{\boldsymbol{\beta}})$	0.242 − − − −	0.059 − − − −	0.013 − − − −
		1.182 6.419 − − −	0.028 0.116 − − −	0.048 0.272 − − −
		1.000 5.274 8.471 − −	0.003 −0.000 0.309 − −	−0.001 −0.001 0.157 − −
		−0.447 −2.119 −2.918 3.882 −	−0.110 −0.079 0.020 0.376 −	−0.006 −0.023 0.031 0.133 −
		−0.019 −0.052 −0.2160 −0.355 2.118	0.052 0.163 −0.563 −0.153 10.708	−0.092 −0.137 0.108 0.479 12.821
RCSPCR	R^2_w	0.341	0.329	0.309
	APE	1.079	1.274	1.361
	$\text{Cov}(\hat{\boldsymbol{\beta}})$	0.002 − − − −	0.000 − − − −	0.000 − − − −
		0.003 0.025 − − −	0.000 0.001 − − −	−0.000 0.005 − − −
		−0.000 −0.000 0.008 − −	0.000 −0.000 0.002 − −	0.000 0.000 0.003 − −
		0.001 0.014 −0.002 3.882 −	−0.000 −0.000 0.000 0.003 −	−0.000 −0.000 −0.000 0.003 −
		−0.001 −0.006 0.001 −0.355 2.118	0.000 0.000 0.000 0.000 0.171	0.000 −0.000 0.000 −0.000 0.014

Table 10. The results of simulation study regarding the Scenario-III for $n = 500$, $p = 20$ and $M = 5$ at different contamination levels with error term from Cauchy distribution $(0, 1)$.

		Contamination levels														
Method		0%					2%					10%				
CBR	R^2	0.029					0.031					0.028				
	APE	1.003					13.838					64.772				
	$\text{Cov}(\hat{\boldsymbol{\beta}})$	0.021	–	–	–	–	0.131	–	–	–	–	0.488	–	–	–	–
		0.000	0.013	–	–	–	0.027	0.093	–	–	–	−0.011	0.269	–	–	–
		−0.013	−0.015	0.029	–	–	−0.004	−0.016	0.076	–	–	0.043	0.003	0.255	–	–
		0.000	0.003	−0.006	0.005	–	−0.028	−0.022	−0.013	0.078	–	−0.047	−0.042	−0.068	0.272	–
		−0.001	−0.001	−0.003	−0.001	0.004	−0.027	−0.021	−0.016	−0.027	0.076	−0.114	−0.075	−0.066	−0.062	0.350
RCBR	R^2_w	0.331					0.313					0.284				
	APE	0.986					13.552					63.636				
	$\text{Cov}(\hat{\boldsymbol{\beta}})$	0.000	–	–	–	–	0.000	–	–	–	–	0.000	–	–	–	–
		0.000	0.000	–	–	–	0.000	0.000	–	–	–	0.000	0.000	–	–	–
		0.000	0.000	0.000	–	–	0.000	0.000	0.000	–	–	0.000	0.000	0.000	–	–
		0.000	0.000	0.000	0.000	–	0.000	0.000	0.000	0.000	–	0.000	0.000	0.000	0.000	–
		−0.000	0.000	0.000	0.000	0.000	−0.000	0.000	0.000	0.000	0.000	−0.000	0.000	0.000	0.000	0.000
PCR	R^2	0.020					0.019					0.017				
	APE	1.025					13.736					65.236				
	$\text{Cov}(\hat{\boldsymbol{\beta}})$	0.000	–	–	–	–	0.000	–	–	–	–	0.007	–	–	–	–
		0.000	0.002	–	–	–	0.000	0.003	–	–	–	0.001	0.031	–	–	–
		−0.000	−0.000	0.000	–	–	−0.000	0.000	0.070	–	–	0.002	0.002	0.263	–	–
		−0.000	−0.000	0.000	0.003	–	−0.000	−0.001	−0.003	0.044	–	−0.001	0.010	−0.017	0.290	–
		0.000	0.000	−0.000	−0.000	0.003	−0.000	0.000	0.003	−0.002	0.061	−0.002	−0.000	−0.010	−0.017	0.248
CSPCR	R^2	0.027					0.028					0.025				
	APE	1.366					21.178					108.618				
	$\text{Cov}(\hat{\boldsymbol{\beta}})$	0.242	–	–	–	–	3.786	–	–	–	–	3.376	–	–	–	–
		1.182	6.419	–	–	–	−2.592	5.660	–	–	–	0.524	30.780	–	–	–
		1.000	5.274	8.471	–	–	0.999	3.622	21.717	–	–	0.156	5.760	27.370	–	–
		−0.447	−2.119	−02.918	3.882	–	−0.293	−1.766	−2.984	15.940	–	0.538	1.874	2.476	22.398	–
		−0.019	−0.052	−0.216	−0.355	2.118	−2.045	3.013	−4.740	1.009	21.920	−1.094	10.091	−0.847	−1.423	35.124
RCSPCR	R^2_w	0.341					0.308					0.288				
	APE	1.079					14.150					64.457				
	$\text{Cov}(\hat{\boldsymbol{\beta}})$	0.002	–	–	–	–	0.018	–	–	–	–	0.000	–	–	–	–
		0.003	0.025	–	–	–	0.009	0.008	–	–	–	0.000	0.001	–	–	–
		−0.000	−0.000	0.008	–	–	−0.009	−0.004	0.021	–	–	0.000	−0.000	0.005	–	–
		0.001	0.014	−0.002	3.882	–	0.000	−0.000	0.005	0.022	–	0.000	0.000	−0.001	0.007	–
		−0.001	−0.006	0.001	−0.335	2.118	0.006	0.004	−0.005	−0.003	0.021	−0.000	0.000	−0.001	0.001	0.007

Table 11. The results of simulation study in the presence of Macro for $n = 500$, $p = 20$ and $M = 5$ at different contamination levels with error term from Cauchy distribution $(0, 1)$.

Method		Cellwise contamination		
MacroRCSPCR		0%	15%	30%
Rowwise outliers=0%	R^2_w	0.307	0.289	0.275
NAs=0%	APE	51.424	50.691	57.785
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.002 & – & – & – & – \\ 0.000 & 0.018 & – & – & – \\ 0.000 & 0.002 & 0.027 & – & – \\ 0.000 & -0.002 & 0.001 & 0.028 & – \\ 0.001 & -0.001 & -0.003 & 0.002 & 0.035 \end{bmatrix}$	$\begin{bmatrix} 0.001 & – & – & – & – \\ 0.001 & 0.027 & – & – & – \\ 0.000 & 0.002 & 0.045 & – & – \\ 0.000 & -0.001 & -0.002 & 0.021 & – \\ -0.001 & -0.001 & 0.002 & -0.001 & 0.030 \end{bmatrix}$	$\begin{bmatrix} 0.002 & – & – & – & – \\ 0.000 & 0.018 & – & – & – \\ 0.000 & 0.000 & 0.027 & – & – \\ 0.000 & -0.001 & -0.004 & 0.017 & – \\ 0.000 & -0.002 & 0.002 & -0.001 & 0.018 \end{bmatrix}$
Rowwise outliers=20%	R^2_w	0.232	0.235	0.184
NAs=5%	APE	50.245	53.328	155.954
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.002 & – & – & – & – \\ 0.001 & 0.019 & – & – & – \\ 0.000 & 0.002 & 0.021 & – & – \\ 0.000 & -0.001 & 0.003 & 0.025 & – \\ 0.000 & 0.001 & 0.000 & -0.001 & 0.032 \end{bmatrix}$	$\begin{bmatrix} 0.003 & – & – & – & – \\ 0.000 & 0.025 & – & – & – \\ 0.000 & -0.001 & 0.062 & – & – \\ -0.000 & -0.005 & 0.010 & 0.075 & – \\ 0.000 & -0.001 & 0.003 & 0.000 & 0.043 \end{bmatrix}$	$\begin{bmatrix} 0.001 & – & – & – & – \\ 0.000 & 0.009 & – & – & – \\ 0.000 & 0.001 & 0.061 & – & – \\ 0.000 & 0.000 & -0.002 & 0.047 & – \\ 0.000 & -0.001 & -0.002 & -0.001 & 0.050 \end{bmatrix}$
Rowwise outliers=20%	R^2_w	0.253	0.223	0.189
NAs=15%	APE	51.364	57.762	120.641
	$\mathrm{Cov}(\hat{\boldsymbol{\beta}})$	$\begin{bmatrix} 0.001 & – & – & – & – \\ 0.000 & 0.012 & – & – & – \\ 0.000 & 0.000 & 0.009 & – & – \\ -0.001 & 0.001 & 0.001 & 0.030 & – \\ 0.000 & 0.000 & -0.001 & 0.000 & 0.031 \end{bmatrix}$	$\begin{bmatrix} 0.002 & – & – & – & – \\ -0.001 & 0.024 & – & – & – \\ 0.000 & 0.002 & 0.025 & – & – \\ -0.001 & -0.001 & 0.001 & 0.046 & – \\ -0.001 & 0.001 & -0.002 & -0.005 & 0.030 \end{bmatrix}$	$\begin{bmatrix} 0.001 & – & – & – & – \\ 0.000 & 0.015 & – & – & – \\ 0.000 & 0.002 & 0.083 & – & – \\ 0.001 & -0.001 & 0.006 & 0.063 & – \\ 0.000 & 0.000 & -0.001 & 0.006 & 0.037 \end{bmatrix}$