Iranian Journal of Numerical Analysis and Optimization
Vol. 12, No. 1, (2022), pp 173-186
DOI:10.22067/ijnao.2021.68160.1008
https://ijnao.um.ac.ir/

How to cite this article
Research Article

A heuristic algorithm to combat outliers and multicollinearity in regression model analysis

M. Roozbeh*, S. Babaie-Kafaki and M. Manavi

Abstract

As known, outliers and multicollinearity in the data set are among the important difficulties in regression models, which badly affect the least-squares estimators. Under multicollinearity and outliers' existence in the data set, the prediction performance of the least-squares regression method is decreased dramatically. Here, proposing an approximation for the condition number, we suggest a nonlinear mixed-integer programming model to simultaneously control inappropriate effects of the mentioned problems. The model can be effectively solved by popular metaheuristic algorithms. To shed light on importance of our optimization approach, we make some numerical experiments on a classic real data set as well as a simulated data set.

AMS subject classifications (2020): 62J05, 90C11, 90C59.

Keywords: Condition number; Linear regression; Penalty method; Metaheuristic algorithm; Nonlinear mixed-integer programming.

*Corresponding author

Received 7 January 2021; revised 27 July 2021; accepted 2 October 2021

Mahdi Roozbeh
Faculty of Mathematics, Statistics and Computer Science, P.O. Box: 35195-363, Semnan University, Semnan, Iran. e-mail: mahdi.roozbeh@semnan.ac.ir

Saman Babaie-Kafaki
Faculty of Mathematics, Statistics and Computer Science, P.O. Box: 35195-363, Semnan University, Semnan, Iran. e-mail: sbk@semnan.ac.ir

Monireh Manavi
Faculty of Mathematics, Statistics and Computer Science, P.O. Box: 35195-363, Semnan University, Semnan, Iran. e-mail: m.maanavi95@gmail.com

1 Introduction

As an effective statistical tool, multiple regression is widely and increasingly used in econometrics, engineering, social sciences, and so on. Generally, in a multiple regression model, the relationship between several independent (predictor) variables with a dependent (response) variable is investigated. The model is formulated by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ is a vector of observations on the response variable, $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times p}$ is a matrix of observations on the predictor variables, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ is a vector of unknown regression coefficients, and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\top$ is a vector of error terms with $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \sigma^2 \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the unit matrix of order n and σ^2 is an unknown constant. The ordinary least-squares estimator (OLSE) of $\boldsymbol{\beta}$ is given by

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \; (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{S}^{-1}\boldsymbol{X}^\top \boldsymbol{y},\end{aligned}$$

where $\boldsymbol{S} = \boldsymbol{X}^\top \boldsymbol{X}$. In regression models with intercept β_0, we can simply rewrite (1) by considering $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ and $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{X})$.

Some difficulties often appear in the regression analysis, such as collinearity between explanatory variables as well as outliers' existence in the data set. Generally, in the regression modeling an outlier is an observation point that fails to track the linear pattern of the data [10]. Outliers corrupt the OLSE; this fact motivated the researchers to investigate robust estimations [9]. As another problem in regression analysis, one may encounter with multicollinearity in the data set that is defined as the existence of nearly linear dependency among columns of the design matrix $\boldsymbol{X}$ [7]. In this situation, the matrix $\boldsymbol{S} = \boldsymbol{X}^\top \boldsymbol{X}$ has one or more small eigenvalues which causes the OLSE to perform poorly [8]. One effective approach to detect the outliers in a data set is the least trimmed squares (LTS) [9] in which the sum of smallest h-squared residuals is minimized rather than the complete sum of squares. Here, h is a prespecified threshold value and denotes the number of normal or good observations that are not outliers. If $z_i \in \{0, 1\}$ is the indicator demonstrating whether the ith observation is ordinary (nonoutlier) or not, then the model can be formulated by

$$\begin{aligned}&\min_{\boldsymbol{\beta},\boldsymbol{z}} \psi(\boldsymbol{\beta}, \boldsymbol{z}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{X} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &s.t. \;\; \boldsymbol{z}^\top \boldsymbol{e} = h, \\ &\qquad z_i \in \{0, 1\}, \; i = 1, 2, \dots, n,\end{aligned} \tag{2}$$

where $\boldsymbol{Z}$ is a diagonal matrix with the diagonal elements $\boldsymbol{z} = (z_1, z_2, \dots, z_n)^\top$, and $\boldsymbol{e} = (1, \dots, 1)^\top \in \mathbb{R}^n$.

Let A be an arbitrary nonsingular matrix. Denoted by $K_p(A)$, the p-norm condition number of A defined by $\kappa_p(A) = \|A\|_p \|A^{-1}\|_p$ [12]. When A is positive definite and $p = 2$, then we get the spectral condition number, which can be determined as the ratio of the largest eigenvalue to the smallest eigenvalue of the matrix A [12]. As known, the condition number is an important factor to check the existence of the multicollinearity [12]. Based on this fact, Roozbeh, Babaie-Kafaki, and Naeimi Sadigh [8] developed an extension of the optimization model (2) to simultaneously control presence of the outliers and the multicollinearity in the data set; that is,

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{z}} \; & \psi(\boldsymbol{\beta}, \boldsymbol{z}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{Z} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \mu\kappa(\boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{X}) \\ s.t. \; & \boldsymbol{z}^\top \boldsymbol{e} = h, \\ & z_i \in \{0, 1\}, \; i = 1, 2, \dots, n, \end{aligned} \tag{3}$$

in which $\kappa(\cdot)$ stands for the spectral condition number and $\mu > 0$ is called the penalty parameter. Here, the corresponding estimator is called the modified LTS counter multicollinearity (MLTSCM) estimator. In model (3), the additional term $\kappa(\boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{X}) > 0$ has been embedded as a penalty for generating inappropriate values for $z_1, z_2, \dots, z_n$ and decreases the condition number of the final model to combat multicollinearity problem.

Here, we deal with a modified version of the regression model (3) with less computational cost in the objective function. This work is organized as follows. In Section 2, we introduce a penalized regression method. The method is then combined with the LTS method in Section 3, to combat the sparsity of the model. In Section 4, we deal with our approximate version of the mixed-integer nonlinear programming model (6) using a simple estimation of the spectral condition number. In Section 5, we provide some numerical experiments to show effectiveness of our model. Finally, concluding remarks are presented in Section 6.

2 Least absolute shrinkage and selection operator methodology

The amount of data we are faced with keeps growing. From around the late 1990s, wide data sets emerged, in which the number of variables far exceeds the number of observations. This was mainly due to our increasing ability to measure a large amount of information automatically [6].

Penalized regression can perform variable selection and prediction in a "Big Data" environment more effectively and efficiently in contrast to the other methods. Initially proposed by Tibshirani [11], the LASSO (least absolute shrinkage and selection operator) is based on minimizing mean squared error, which is based on balancing the opposing factors of bias and variance to build the most predictive model. In fact, LASSO shrinks the regression coefficients toward zero by penalizing the regression model with an ℓ_1-norm penalty term, that is, the sum of the absolute value of the coefficients. LASSO regression is a simple technique to reduce model complexity and usually prevent over-fitting which may result from simple linear regression.

In the case of LASSO regression, the penalty term is embedded to force the coefficient estimates with minor contributions to the model to be exactly set to zero. This means that, LASSO can be also seen as an alternative to the subset selection methods for performing variable selection in order to reduce complexity of the model.

Ordinary least squares (OLS) regression chooses the coefficients by minimizing the residual sum of squares (RSS) as follows:

$$\min_{\boldsymbol{\beta}} RSS = \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top (\boldsymbol{y} - \hat{\boldsymbol{y}}) = \min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\},$$

where $(x_{i1}, \dots, x_{ip})$ can be called $\boldsymbol{x}_i^\top$. LASSO is an extension of the OLS, which adds a penalty to the RSS, being equal to sum of the absolute values of the nonintercept beta coefficients multiplied by the parameter λ that slows (when $\lambda < 1$) or accelerates (when $\lambda > 1$) the penalty. Therefore, the following optimization problem should be solved in LASSO problem:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Figure 1 shows the constraint area of the LASSO method for $p = 2$, in which elliptical contours of the function are shown by the full. They are centered at the OLSE. The constraint region is the rotated square. LASSO solution is the first place that the contours touch the square, and this will sometimes occur at a corner, corresponding to a zero coefficient. LASSO is frequently used in practice since the ℓ_1 penalty allows us to shrink some coefficients to zero, that is, to produce sparse estimation models that are highly interpretable.

It is notable that increasing λ will increase bias and decrease variance. Likewise, decreasing λ reduces bias and increases variance. A big part of the building, the best models in LASSO deal with the bias-variance tradeoff. Bias refers to how correct (or incorrect) the model is. A very simple model that makes a lot of mistakes is said to have a high bias. A very complicated model that does well on its training data is said to have a low bias. There are several ways to choose the optimal λ, such as AIC, BIC, $\mathrm{C_p}$, and so on [10]. For this purpose, one of the most popular methods is the cross validation (CV) method.

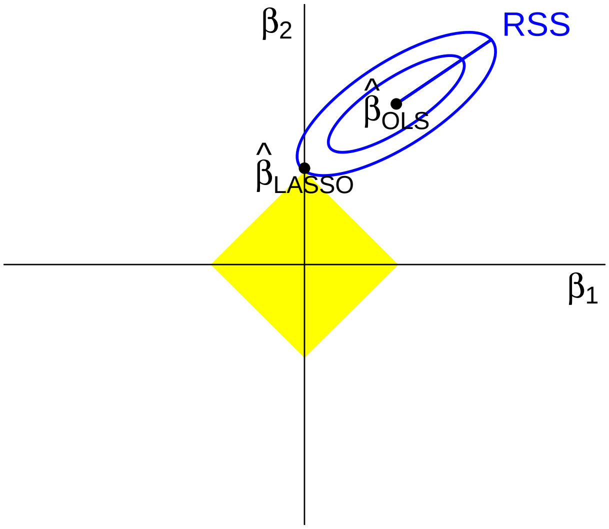

Figure 1: Constraint region of LASSO.

In order to find the optimal value of λ, a range of λ values is tested and the optimal value is chosen using CV, which involves

- separating the data into a training set and a test set;
- building the model in the training set;
- estimating the outcome in the test set using the model from the training set;
- calculating mean squared error (MSE) in the test set.

There are different types of cross validation method like Leave-P-Out, Leave-one-out, k-fold, standard k-fold, Monte Carlo, and so on [2]. One of the most important methods is the k-fold CV, which is one way to improve over the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other $k-1$ subsets are put together to form a training set. Then, the average error across all k trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set $k-1$ times. The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into the test and the training set k different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.

3 Sparse least trimmed squares method

As discussed, LASSO offers interpretable models but is not robust with respect to the outliers. The breakdown point of LASSO is $\frac{1}{n}$ (see [1] for more details); that is, only one single outlier can make the LASSO estimator completely unreliable. Therefore, robust alternatives are needed. In this situation, Alfons, Croux, and Gelper [1] suggested the sparse LTS estimator as follows:

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{z}} \psi(\boldsymbol{\beta},\boldsymbol{z}) = \left\{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{Z}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + h\lambda \textstyle\sum_{j=1}^p |\beta_j|\right\} \\ &s.t. \;\; \boldsymbol{z}^\top \boldsymbol{e} = h, \\ &\qquad z_i \in \{0,1\}, \; i = 1,2,\ldots,n, \end{aligned}$$

where λ is a penalty parameter. They showed that the breakdown point of this estimator is $\frac{n-h+1}{n}$.

4 A penalized nonlinear mixed-integer programming model in linear regression

As known, the condition number is an effective tool to check the existence of multicollinearity [12]. The OLSE performs poorly in the presence of multicollinearity. Also, the existence of multicollinearity may lead to wide confidence intervals for the individual parameters or their linear combinations, and can produce estimators with wrong signs.

As known, if $a_i, i = 1,\ldots,n$ denotes the ith column of A, then, for any i and j, it can be seen that

$$\kappa_p(A) \geq \frac{\|a_i\|_p}{\|a_j\|_p} \tag{4}$$

(see [12, Theorem 2.2.25]). Hence, we can write

$$\kappa_p(A) \geq \frac{\max\limits_{i=1,\ldots,n} \|a_i\|_p}{\min\limits_{j=1,\ldots,n} \|a_j\|_p} \overset{def}{=} \Psi_p(A). \tag{5}$$

Now, based on the above inequality, it is reasonable to select a set of h appropriate observations of the data that makes the computationally low cost term $\Psi_p(\boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{X})$ as small as possible. Hence, we propose the computationally low cost approximation $\kappa_p(A) \approx \Psi_p(A)$ and then, propose the following approximation of model (3):

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{z}} \psi(\boldsymbol{\beta},\boldsymbol{z}) = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{Z}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + \mu\Psi(\boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{X}) \\ &s.t. \ \ \boldsymbol{z}^\top \boldsymbol{e} = h, \\ &\qquad z_i \in \{0,1\}, \ i = 1,2,\ldots,n. \end{aligned} \tag{6}$$

Here, the corresponding estimator is called the approximate LTS counter multicollinearity (ALTSCM) estimator. As seen, the given optimization model illustrates an NP-hard mixed-integer (having both continuous ($\boldsymbol{\beta}$) and (discrete) integer ($\boldsymbol{z}$) variables) nonlinear programming problem for which the classical methods are not practically efficient [4]. Note that in complexity theory, NP-hardness is viewed as strong evidence that a problem is not polynomially solvable [4]. As known, metaheuristic algorithms have attracted special attention in developing efficiently robust computational procedures for solving a vast variety of such problems. Among them there is the simulated annealing (SA) algorithm [4]. SA is a local search algorithm capable of escaping from local optima by use of random hill-climbing moves in the search process. It is very efficient in practice and well-developed in theory. Motivated by these, here we use the SA algorithm to approximately compute ALTSCM.

5 Numerical experiments

In this section, we investigate computational efficiency of the given estimator firstly on a real data set and then, on a simulated data set.

5.1 A real data set related to the riboflavin vitamin B2 production in Bacillus subtilis

To illustrate usefulness of the suggested strategies, we consider the data set about riboflavin vitamin B2 production in Bacillus subtilis, which can be found in R package “hdi” [5]. Riboflavin is one of the B vitamins, which are all water soluble. Riboflavin is naturally present in some foods, added to some food products, and available as a dietary supplement. This vitamin is an essential component of two major coenzymes, flavin mononucleotide (FMN; also known as riboflavin-5’-phosphate), and flavin adenine dinucleotide (FAD).

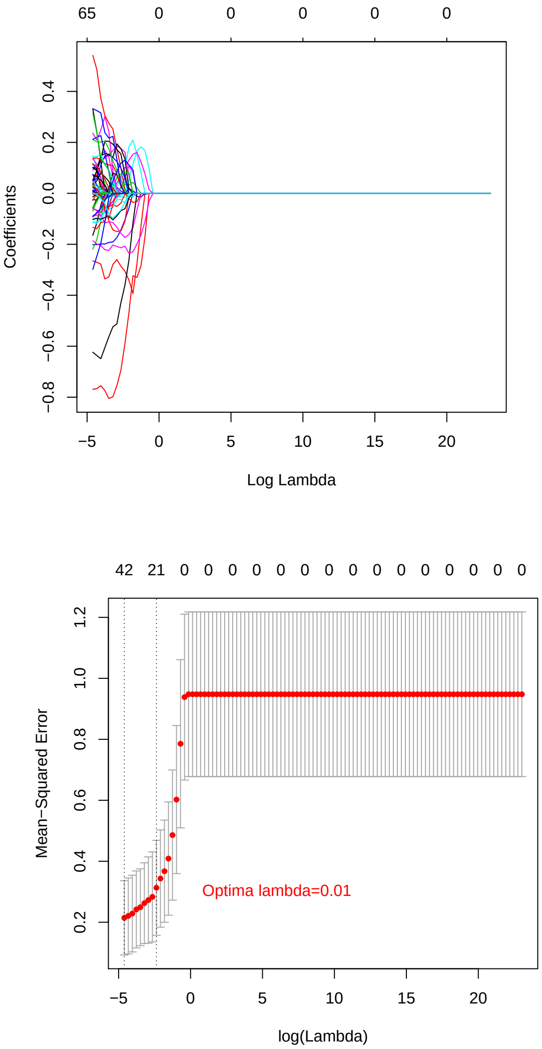

Figure 2: LASSO plots for the riboflavin data set.

There is a single real valued response variable, which is the logarithm of the riboflavin production rate. Furthermore, there are $p = 4088$ explanatory variables measuring the logarithm of the expression level of 4088 genes. There is also one rather homogeneous data set from $n = 7$ samples that were hybridized repeatedly during a fed batch fermentation process, where different engineered strains and strains grown under different fermentation conditions were analyzed. There is one rather homogeneous data set from $n = 71$ samples that were hybridized repeatedly during a fed batch fermentation process, where different engineered strains and strains grown under different fermentation conditions were analyzed. In Figure 2, the 10-fold cross-validation and

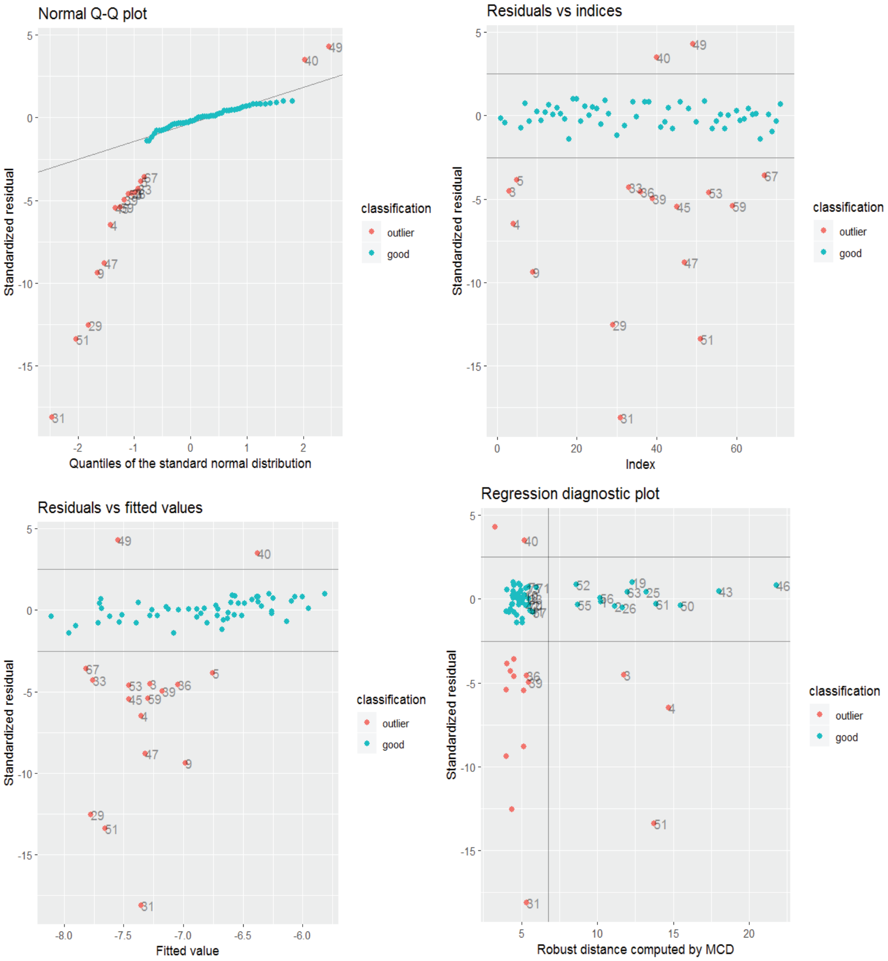

Figure 3: Outlier detection plots for the riboflavin data set.

the coefficients estimation diagrams for different values of the penalty parameter are depicted. We plotted Figure 2 to find the best value of the LASSO parameter (λ_n), which minimizes the CV criterion. As seen in Figure 2, the minimal mean squared error, estimated by CV, is achieved at $\lambda_n = 0.01$. The LASSO method selects 53 variables. Figure 3 produces the diagnostic plots for a sequence of regression models, such as submodels along a robust sparse least trimmed squares regression models for a grid of values for the penalty parameter. In the normal Q-Q plot of the standardized residuals, a reference line is drawn through the first and third quartiles. The index number of observations with the largest distances from that line are identified by a label (the observation number). In the plots of the standardized residuals versus their index or the fitted values, horizontal reference lines are drawn at 0 and ±2.5. The index number of observations with the largest absolute values of the standardized residuals is identified by a label (the observation number). For the regression diagnostic plot, the robust Mahalanobis distances of the predictor variables are computed via the minimum covariance determinant (MCD) based on only those predictors with nonzero coefficients. Horizontal reference lines are drawn at ±2.5 and a vertical reference line is drawn at the upper 97.5% quantile of the chi-squared distribution with p degrees of freedom, where p denotes the number of predictors with nonzero coefficients. The index number of observations with the largest absolute values of the standardized residuals and/or largest robust Mahalanobis distances are identified by a label (the observation number). According to Figure 3, it is clear that there exist some outliers in the data and so, it is necessary to use the robust methods same as the sparse LTS and ALTSCM methods for modeling the data.

We reported the results for the three methods listed in Table 1, in which we numerically calculated $\mathrm{RSS} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$ and $\mathrm{R}^2 = 1 - \mathrm{SSE}/\mathrm{S_{yy}}$ with $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\mathrm{S_{yy}} = \sum_{i=1}^{n} z_i(y_i - \bar{y})^2$. They are measures for the error and goodness of prediction, respectively. It is clear that the penalized mixed-integer method performs better than the other methods according to the goodness of fit criteria.

5.2 A simulated data analysis

To examine the performance of the proposed estimators, we perform a simulation data study. For this purpose, the following model is considered with $n = 55$, $p = 450$, and $h = 41$ (see [3] for more details):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^\top,\ \boldsymbol{\beta}_1 = (-1.5, 2, 2.5, 4, -3, 5)^\top,\ \boldsymbol{\beta}_{2\ (444\times 1)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1),$$

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{450} \sim \mathcal{N}_{450}(\mathbf{1}_{450}, \boldsymbol{I}_{450}),$$

$$\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2)^\top,\ \boldsymbol{\epsilon}_{1\ (h\times 1)} \sim \mathcal{N}(0,1),\ \boldsymbol{\epsilon}_{2\ ((n-h)\times 1)} \overset{i.i.d.}{\sim} t_2(8),$$

where $t_m(\delta)$ is the noncentral t-student distribution with m degrees of freedom and noncentrality parameter δ.

We produced the first h and the last $n-h$ error terms as random variables from dependent normal and independent noncentral t-student distributions,

Method	LASSO		Sparse LTS		ALTSCM	
	Effective genes	Estimation	Effective genes	Estimation	Effective genes	Estimation
	Intercept	0.6712	*Intercept*	0.3093	*Intercept*	–
	ARGF_at	-0.1911	CHED_at	-0.0895	ABH_at	0.0124
	DNAJ_at	-0.1364	CSBA_at	0.0546	ALD_at	0.0124
	GAPB_at	0.0380	CTAA_at	0.0336	AMYC_at	0.0205
	LYSC_at	-0.2977	DEAD_at	-0.1260	ARGB_at	-0.0132
	PKSA_at	0.0236	MREBH_at	-0.0819	ARGF_at	-0.0177
	PRIA_at	0.0982	SPOIISB_at	0.0248	ARGG_at	-0.0133
	SPOIIAA_at	0.0224	THIK_at	-0.2799	ARGH_at	-0.0160
	SPOVAA_at	0.2652	XKDI_at	0.0001	BIOB_at	-0.0133
	THIK_at	-0.0051	XKDS_at	0.0988	CARA_at	-0.0140
	XHLB_at	0.1608	YCDH_at	-0.0018	CARB_at	-0.0145
	YACN_at	-0.0404	YCGM_at	-0.0257	GAPB_at	0.0130
	YBFI_at	0.1329	YCGP_at	-0.0168	GSIB_at	-0.0129
	YCDH_at	-0.0067	YCKJ_at	-0.0761	NDK_at	-0.0129
	YCGO_at	-0.0057	YDAR_at	-0.0175	PAND_at	-0.0128
	YCKE_at	0.0092	YDBM_at	-0.3054	PBUX_at	0.0140
	YCLB_at	0.1994	YFLL_at	-0.0542	PHRI_r_at	-0.0128
	YCLF_at	-0.0533	YHCB_at	-0.1021	PTSG_at	-0.0120
	YDDH_at	-0.0176	YHCL_at	-0.0850	SIGY_at	-0.0155
	YDDK_at	-0.1142	YHDO_at	-0.0762	SSPA_at	-0.0122
	YEBC_at	-0.5347	YJBJ_at	-0.1005	YBGB_at	-0.0139
	YFHE_r_at	0.1451	YJCJ_at	0.2446	YCDH_at	-0.0307
	YFII_at	0.0100	YKUH_at	0.2855	YCGM_at	-0.0183
	YFIO_at	0.1588	YORB_i_at	0.0519	YCGN_at	-0.0206
	YFIR_at	0.0441	YQET_at	-0.0217	YCGO_at	-0.0207
	YHDS_r_at	0.1452	YRZI_r_at	0.0109	YCIC_at	-0.0229
	YKBA_at	0.1108	YTGB_at	-0.0203	YDDH_at	-0.0131
	YKVJ_at	0.0237	YUIA_at	0.0113	YFMH_r_at	-0.0164
	YLXW_at	0.0731	YVBY_at	0.0420	YHDS_r_at	0.0166
	YMFE_at	0.0183	YWQD_at	-0.0098	YHFH_r_at	0.0178
	YOAB_at	-0.8123	YXEL_at	-0.0442	YHZA_at	-0.0396
	YPGA_at	-0.0102	YXLD_at	-0.0389	$YOEB_at$	-0.0122
	YQJT_at	0.0415	YXLE_at	-0.1714	$YOPS_at$	0.0149
	YQJU_at	0.2320			YOQX_i_at	0.0155
	YRVJ_at	-0.0547			YPUD_at	0.0349
	YTGB_at	-0.0390			YPUF_at	0.0248
	YUID_at	0.0134			YPUG_at	0.0208
	YURQ_at	0.0245			YRPE_at	-0.0150
	YXLD_at	-0.2005			YRZI_r_at	0.0173
	YXLE_at	-0.1068			YTGA_at	-0.0168
	YYBG_at	-0.0781			YTGB_at	-0.0191
	YYDA_at	-0.1042			YTGC_at	-0.0129
					YTGD_at	-0.0198
					YTIA_at	-0.0290
					YUSA_at	-0.0126
					YXLC_at	-0.0204
					YXLD_at	-0.0270
					YXLE_at	-0.0270
					YXLF_at	-0.0194
					YXLG_at	-0.0216
RSS	594.0313		25.6582		**0.1487**	
R^2	0.6519		0.9093		**0.9458**	

Table 1: Results of the proposed estimators for the riboflavin data set

respectively. Such data generating method makes the outliers to appear on one side of the true regression hyperplane, corrupting the nonrobust estimators by tending them to the outliers.

In Table 2, we have reported the results. To save space, the estimations of nonzero coefficients are just reported in Table 2. Also, the outliers are shown in Figure 4. As seen, ALTSCM is again preferable to the other methods.

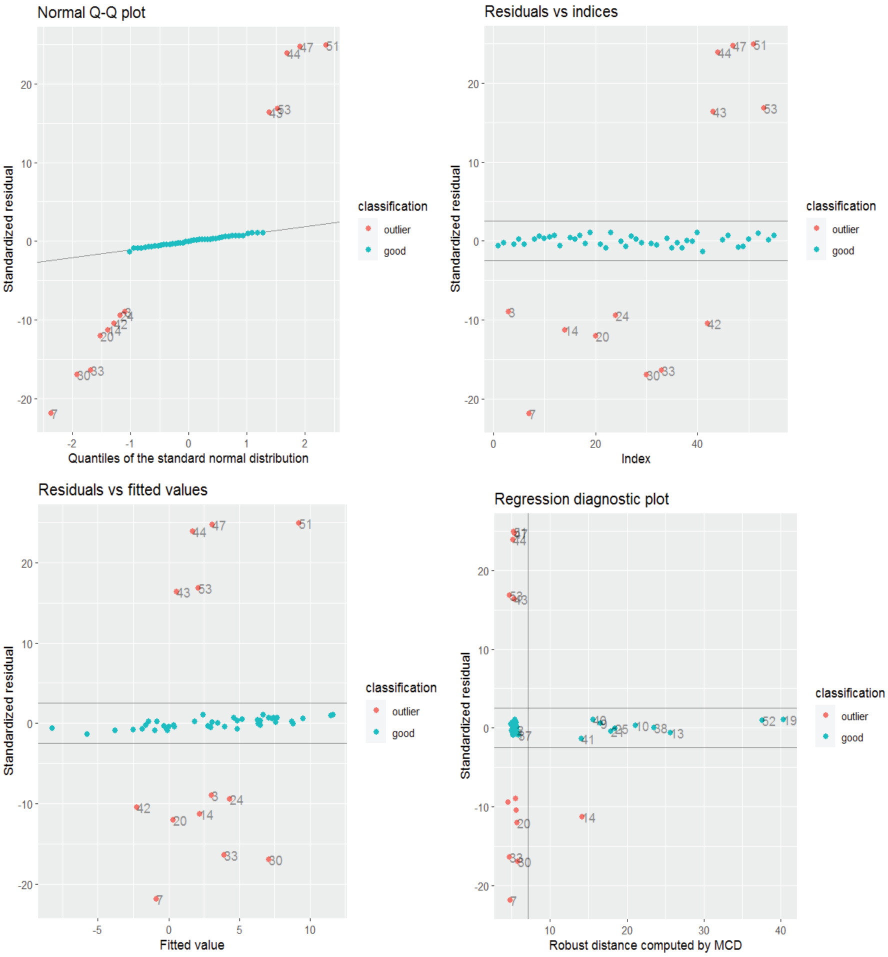

Figure 4: Outlier detection plots for the simulated data set.

Method Parameters	LASSO	Sparse LTS	ALTSCM
β_1	0.00000	0.00000	-1.44790
β_2	0.00000	0.00000	1.98652
β_3	2.08386	2.69022	2.48500
β_4	3.83184	3.47647	4.01102
β_5	-3.01472	-2.03737	-2.99656
β_6	2.27159	2.87485	4.94529
RSS	4203.47	2850.05	**4.71450**
R^2	0.14287	0.41885	**0.93325**

Table 2: Results of the proposed estimators for the simulated data set

6 Conclusions

We dealt with a computationally low cost nonlinear mixed-integer optimization model to combat both the outliers and multicollinearity effects in the high dimensional regression. Suggesting an approximation for the condition number, our model also has a simple (LTS-based) structure. We used the simulated annealing algorithm to solve the model effectively. Computational tests showed that the given estimator (ALTSCM) is practically promising.

References

1. Alfons, A., Croux, C. and Gelper, S., *Sparse least trimmed squares regression for analyzing high dimensional large data set,* Ann. Appl. Stat. 7 (2013), 226–248.

2. Amini M. and Roozbeh, M., *Optimal partial ridge estimation in restricted semiparametric regression models,* J. Multivar. Anal. 136 (2015), 26–40.

3. Arashi M., Roozbeh, M., Hamzah, N.A. and Gasparini, M., *Ridge regression and its applications in genetic studies,* PLoS ONE 16(4) (2021), e0245376.

4. Bertsimas, D. and Tsitsiklis, J.N., *Introduction to linear optimization,* Athena Scientific, Massachusetts, 1997.

5. Buhlmann, P., Kalisch, M. and Meier, L., *High-dimensional statistics with a view towards applications in biology,* Annu. Rev. Stat. Appl. 1 (2014), 255–278.

6. Efron, B. and Hastie, T., *Computer age statistical inference,* Cambridge University Press, Cambridge, 2017.

7. Roozbeh, M., Babaie-Kafaki, S. and Arashi, M., *A class of biased estimators based on QR decomposition*, Linear Algebra Appl. 508 (2016), 190–205.

8. Roozbeh, M., Babaie-Kafaki, S. and Naeimi Sadigh, A., *A heuristic approach to combat multicollinearity in least trimmed squares regression analysis,* Appl. Math. Model. 57 (2018), 105–120.

9. Rousseeuw, P.J. and Leroy, A.M., *Robust regression and outlier detection,* John Wiley and Sons, New York, 1987.

10. Sheather, S.J., *A modern approach to regression with R,* Springer, New York, 2009.

11. Tibshirani, R., *Regression shrinkage and selection via the LASSO,* J. R. Stat. Soc. Ser. B, 58 (1996), 267–288.

12. Watkins, D.S., *Fundamentals of matrix computations,* John Wiley and Sons, New York, 2002.

How to cite this article

Roozbeh, M., Babaie-Kafaki, S., Maanavi, M. A heuristic algorithm to combat outliers and multicollinearity in regression model analysis. *Iranian Journal of Numerical Analysis and Optimization*, 2022; 12(1): 173-186. doi: 10.22067/ijnao.2021.68160.1008