Applied Mathematical Modelling 57 (2018) 105–120

Contents lists available at ScienceDirect

Applied Mathematical Modelling

journal homepage: www.elsevier.com/locate/apm

A heuristic approach to combat multicollinearity in least trimmed squares regression analysis

Mahdi Roozbeh*, Saman Babaie-Kafaki, Alireza Naeimi Sadigh

Faculty of Mathematics, Statistics and Computer Science, Semnan University, P.O. Box 35195-363, Semnan, Iran

ARTICLE INFO

Article history:
Received 29 September 2016
Revised 11 November 2017
Accepted 21 November 2017
Available online 24 December 2017

Keywords:
Breakdown point
Constrained optimization
Heuristic algorithm
Least trimmed squares estimator
Multicollinearity
Ridge estimation

ABSTRACT

In order to down-weight or ignore unusual data and multicollinearity effects, some alternative robust estimators are introduced. Firstly, a ridge least trimmed squares approach is discussed. Then, based on a penalization scheme, a nonlinear integer programming problem is suggested. Because of complexity and difficulty, the proposed optimization problem is solved by a tabu search heuristic algorithm. Also, the robust generalized cross validation criterion is employed for selecting the optimal ridge parameter. Finally, a simulation case and two real-world data sets are computationally studied to support our theoretical discussions.

1. Introduction

Regression model (RM) has attracted especial attentions in statistics, econometrics, engineering, psychology and other areas. Generally, the model can be given by

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

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times p}$ is a non-stochastic design matrix of full column rank, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ is often assumed to be a vector of disturbances (errors) distributed with $\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \sigma^2\boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the identity matrix of order n and σ^2 is an unknown parameter. The ordinary least-squares estimator (OLSE) of the parameter $\boldsymbol{\beta}$ is obtained in the RM as

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

where $\boldsymbol{S} = \boldsymbol{X}^\top\boldsymbol{X}$.

The RM may encounter with some difficulties such as existence of the outliers and the multicollinearity in a data set. In the following literature review, we discuss several essential approaches to overcome such problems.

* Corresponding author.
E-mail address: mahdi.roozbeh@semnan.ac.ir (M. Roozbeh).

https://doi.org/10.1016/j.apm.2017.11.011
0307-904X/

1.1. Review of literature

Among the common problems in regression analysis there is the problem of outliers. In general, an outlier is an observation point that lies outside the main body or the group that it is a part of. In statistics, an outlier is an observation that is well outside of the expected range of the values in a study or an experiment, and in the regression concept, an outlier is a point that fails to follow the main linear pattern of the data. The OLSE is potentially sensitive to the outliers; this fact provided necessary motivations to investigate robust (resistance) estimations. Generally, the robust estimation and the regression are among the most popular problems in the statistics community. In the statistical literature, robust regression estimators are labeled as "robust" because they are relatively insensitive to extreme observations. Robust regression estimators may still retain the mentioned desirable characteristic even when the assumptions under which they are derived do not hold.

As an intuitive scheme to detect the outliers, OLSE can be used on the data set and then, the observations with unusually large squared residuals can be ignored. However, since the outliers have already corrupted the estimation, this strategy may not be effective in practice.

Multicollinearity is another difficulty in addition to the outliers problem in the RMs. As known, multicollinearity appears when the columns of X are to some extent linearly dependent. So, some of the eigenvalues of the matrix S tend to zero, yielding some large regression coefficients estimations (in the absolute value). The multicollinearity problem can be effectively discovered using the condition number [1]. More exactly, under multicollinearity the condition number of $\boldsymbol{S}$ is large and hence, it is an ill-conditioned matrix. Multicollinearity may considerably decrease efficiency of the ordinary least-squares (OLS) method in the sense of leading to wide confidence intervals as well as wrong signs for the parameter $\boldsymbol{\beta}$.

The most effective strategy to overcome the multicollinearity is the ridge estimation suggested by Hoerl and Kennard [2]. Examples of alternative approaches to reduce the effect of multicollinearity include the $r-k$ class estimator proposed by Baye and Parker [3], the biased estimation suggested by Liu [4], the $r-d$ class estimator proposed by Kaçiranlar and Sakallioğlu [5], and the biased estimation suggested by Roozbeh et al. [6] in which the QR decomposition has been employed.

The main part of this work is related to investigate some robust estimation approaches when multicollinearity appears in the data set. This work is organized as follows: a metaheuristic algorithm is briefly discussed for the least trimmed squares (LTS) estimation in Section 2. In Section 3, the ridge LTS methodology and a constrained optimization approach are described, both being designed to decrease the negative effects of the outliers and multicollinearity simultaneously. Section 4 includes a simulation study (to bring together the high percentage of the contaminated data with outliers and strong multicollinearity), and two real applications in bridge construction and electricity consumption data sets, to illustrate effectiveness of the proposed estimators. Finally, conclusions are drawn in Section 5.

2. A tabu search algorithm for the least trimmed squares estimation

Here, at first we present the constrained optimization model of LTS. Then, we briefly describe a tabu search algorithm to solve the problem.

2.1. LTS approach to combat outliers

As mentioned before, outliers essentially damage the results of the OLS regression fit. As a practical tool to decrease the outliers effect, LTS attempts to minimize the smallest h-squared residuals summation instead of whole squared residuals. Here, h is the number of good observations, called the threshold parameter, and the ratio $\alpha = \frac{h}{n} \times 100$ stands for the percentage of the good observations.

Let z_i be the indicator demonstrating whether the $i$the observation is good or not. Consider the following nonlinear binary optimization problem used in the LTS method:

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{z}} \quad & f(\boldsymbol{\beta},\boldsymbol{z}) = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\top}\boldsymbol{Z}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\ s.t. \quad & \boldsymbol{z}^{\top}\boldsymbol{e} = h, \\ & z_i \in \{0,1\},\ i = 1,2,\ldots,n, \end{aligned} \tag{2.1}$$

in which the matrix $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$ is diagonal such that the vector $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)^{\top}$ contains its diagonal elements and $\boldsymbol{e} = (1,\ldots,1)^{\top} \in \mathbb{R}^n$. Solution of (2.1) is called the LTS estimator of the parameter $\boldsymbol{\beta}$, computed in the RM as follows:

$$\hat{\boldsymbol{\beta}}^{LTS}(\boldsymbol{z}) = \boldsymbol{S}(\boldsymbol{z})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Z}\boldsymbol{y},$$

where $\boldsymbol{S}(\boldsymbol{z}) = \boldsymbol{X}^{\top}\boldsymbol{Z}\boldsymbol{X}$.

The LTS problem (2.1) is a nonlinear integer programming problem which contains the binary variables $z_i \in \{0, 1\}$, $i = 1, 2, ..., n$. So, finding its optimal solution in such discrete feasible region by classical algorithms is often expensive or even impractical, especially in large-scale cases. As known, there are many studies reporting promising results of the heuristic algorithms for solving such problems (see [7] and the references therein). The interest in these strategies remains particularly

vivid for several motivations: the high flexibility that make it possible to re-use the softwares, and the good performances that allow to efficiently address some large-scale and complicated problems. Hence, we apply a heuristic algorithm for solving (2.1).

2.2. A tabu search algorithm for LTS

Although it seems that a relaxed version of LTS problem (2.1) can be easily solved by the classical methods, it should be noted that the solution of the relaxed problem may be infeasible for the original problem [8]. It is also notable that the LTS problem has only one simple linear constraint and n binary variables. Hence, every evolutionary metaheuristic algorithm can be effectively employed to solve it.

Among the well-known and efficient heuristic algorithms there is the tabu search (TS) technique [8–12]. TS is a local search evolutionary algorithm that can be used for solving optimization problems in the general form $\min_{x\in\Omega} f(x)$, where Ω is the set of feasible solutions and f is the objective function. Proposed by Glover in 1986, TS has emerged as one of the leading techniques for handling optimization problems that are difficult or impossible to solve with classical procedures [9]. It is an iterative algorithm based on an adaptive memory programming scheme which provides necessary tools to overcome local optima and generates solutions that are very close to optimality. So, here we employ the TS algorithm to find an approximate solution of (2.1).

3. An optimization approach to simultaneously combat outliers and multicollinearity

Here, we describe the biased estimation strategy under multicollinearity in the RMs. In this context, we need the following preliminaries.

As known, the OLSE covariance matrix is equal to $\sigma^2\boldsymbol{S}^{-1}$. So, similar to OLSE, its covariance matrix is strongly dependent to the characteristics of the matrix $\boldsymbol{S}$. Hence, OLSE can be affected by some computational errors when $\boldsymbol{S}$ is an ill-conditioned matrix. The multicollinearity problem may be solved by providing more information, changing the model and selecting the variables again [6]. There exist two popular methods to solve the multicollinearity problem: the 'principal components regression' and the 'ridge regression' approaches which here we focus on the second one.

In recent decades, researchers paid especial attentions to the ridge strategy. Initiated by Hoerl and Kennard [2], they suggested to enlarge the diagonal elements of $\boldsymbol{S}$ by adding some small positive scalars. Indeed, the ridge estimator can be determined by minimizing the squared residuals summation under the constraint $\boldsymbol{\beta}^\top\boldsymbol{\beta}\le\phi^2$. More exactly, the ridge estimator is a solution of the following optimization problem:

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + k(\boldsymbol{\beta}^\top\boldsymbol{\beta}-\phi^2), \tag{3.1}$$

in which $k\ge 0$ is the Lagrange multiplier. Solving problem (3.1), we get

$$\hat{\boldsymbol{\beta}}(k) = \boldsymbol{S}(k)^{-1}\boldsymbol{X}^T\boldsymbol{Y},\ \ \boldsymbol{S}(k) = \boldsymbol{S}+k\boldsymbol{I}_p, \tag{3.2}$$

where is called the ridge least-squares estimator (RLSE). By direct computation, the risk function of $\hat{\boldsymbol{\beta}}(k)$ is given by

$$\mathrm{R}\big(\hat{\boldsymbol{\beta}}(k),\boldsymbol{\beta}\big) = k^2\boldsymbol{S}(k)^{-1}\boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{S}(k)^{-1} + \sigma^2\boldsymbol{S}(k)^{-1}\boldsymbol{S}\boldsymbol{S}(k)^{-1}.$$

For some appropriate values of $k>0$, called the ridge or the shrinkage parameter, the mean squared error of RLSE is smaller than that of OLSE. The constant k is often evaluated based on the data set. Using the vector of correlation coefficient between $\boldsymbol{X}$ and $\boldsymbol{y}$, recently Dorugade [13] proposed the adjusted ridge approach by avoiding the complicated computations of determining the optimal ridge parameter. Roozbeh [14] considered robust ridge estimation of the semiparametric RM under a multicollinearity setting, using the LTS method. Although the ridge strategy is frequently used to dominate the multicollinearity effect and it has been shown that these robust ridge estimators are more effective than the ordinary ridge estimator, it has some defects. For example, the data are distorted by changing the diagonal entries of $\boldsymbol{S}$. In the ridge regression methodology, choosing the ridge parameter plays an important role in the estimation procedure. It is difficult to propose a satisfactory strategy for selecting k, because in practice the best choice of k always depends on the unknown parameters, making the problem to be more difficult. To determine k, on one side we should maintain the condition number of $\boldsymbol{S}$ in a low level to prevent the instability of the evaluated coefficients. Hence, we should select some large values for the parameter k, yielding a small covariance for the evaluated coefficients as well as a more stable estimator. On the other side, small values of k lead to less biased estimators. So, we should be careful to select a proper value for k. Although many researchers have proposed various methods for determining the ridge parameter, this problem has not yet been solved completely. In a recent attempt, Roozbeh et al. [6] introduced a QR-based approach to solve multicollinearity problem in the RM by increasing the numerical rank of the design matrix $\boldsymbol{X}$, yielding appropriate practical results and causing less data distortion in contrast to the classical ridge strategy. Same as the ridge method, choosing the optimal value of biasing parameter is so complicated such that it has not still been solved.

In the presence of outliers in the data besides multicollinearity, combining two optimization problems (2.1) and (3.1), we propose ridge LTS (RLTS) problem in the RM as follows:

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{z}_1} \quad & f(\boldsymbol{\beta},\boldsymbol{z}_1) = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{Z}_1(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + k(\boldsymbol{\beta}^\top\boldsymbol{\beta}-\phi^2) \\ s.t. \quad & \boldsymbol{z}_1^\top \boldsymbol{e} = h, \\ & z_{1_i} \in \{0,1\},\ i=1,2,\ldots,n, \end{aligned} \tag{3.3}$$

where its solution is called the RLTS estimator (RLTSE) of the parameter $\boldsymbol{\beta}$, calculated by

$$\hat{\boldsymbol{\beta}}^{LTS}(k,\boldsymbol{z}_1) = \boldsymbol{S}(k,\boldsymbol{z}_1)^{-1}\boldsymbol{X}^\top\boldsymbol{Z}_1\boldsymbol{y}, \tag{3.4}$$

where $\boldsymbol{S}(k,\boldsymbol{z}_1) = \boldsymbol{S}(\boldsymbol{z}_1) + k\boldsymbol{I}_p$. Based on the approach of [15], we can consider the relaxed version of (3.3) by using the constraint $0 \le z_{2_i} \le 1$ instead of $z_{1_i} \in \{0,1\}$. The corresponding optimization problem is called the ridge relaxed LTS (RRLTS) problem, yielding the ridge relaxed LTS estimator (RRLTSE) of $\boldsymbol{\beta}$. Moreover, employing the extension scheme of Roozbeh and Babaie-Kafaki [16], we can develop an extension of the RRLTS problem by replacing the constraints of (3.3) with $h_1 \le \boldsymbol{z}_3^\top\boldsymbol{e} \le h_2$ and $0 \le z_{3_i} \le 1,\ i=1,2,\ldots,n$, where the positive integers h_1 and h_2 are determined such that $h \in [h_1, h_2]$. The solution of this optimization problem is called the ridge extended relaxed LTS estimator (RERLTSE) of $\boldsymbol{\beta}$. Here, we use robust generalized cross validation (RGCV) to select the optimal value of the ridge parameter (k_{opt}) for evaluation of RLTSE. According to the achievements of Amini and Roozbeh [17], the RGCV score functions can be procured by

$$\mathrm{RGCV}(k,\boldsymbol{z}_j) = \frac{\frac{1}{n}\left\|\left(\boldsymbol{I}_n - L(k,\boldsymbol{z}_j)\right)\boldsymbol{y}\right\|_2^2}{\left(1-\frac{1}{n}\mathrm{tr}\left(L(k,\boldsymbol{z}_j)\right)\right)^2}, \quad j=1,2,3,$$

where $L(k,\boldsymbol{z}_j) = \boldsymbol{X}\boldsymbol{S}(k,\boldsymbol{z}_j)^{-1}\boldsymbol{X}^\top\boldsymbol{Z}_j$.

Although frequently used for overcoming multicollinearity, the ridge estimator has some defects, as mentioned before. Especially, selection of the ridge parameter k plays an essential role in the estimation result in the sense that large values of k lead to considerable biasness while small values of k lead to remarkable instability. The optimal value of k which yields the most efficient estimator is an open problem. As seen in (3.2) and (3.4), the estimator $\hat{\boldsymbol{\beta}}(k)$ is not a simple function of k.

The mentioned defects of the ridge approach which can be transmitted to RLTS motivated us to investigate another strategy to overcome the multicollinearity and outliers effects simultaneously. In this context, we suggest a modified form of the optimization problem (2.1) based on a penalization scheme. More exactly, we add a positive multiple of the spectral condition number of the matrix $\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{X}$, i.e. $\kappa(\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{X})$, to the objective function of (2.1) as a penalty term. This yields the following constrained optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{z}_1} \quad & f(\boldsymbol{\beta},\boldsymbol{z}) = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{Z}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + M\kappa(\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{X}) \\ s.t. \quad & \boldsymbol{z}^\top \boldsymbol{e} = h, \\ & z_i \in \{0,1\},\ i=1,2,\ldots,n, \end{aligned} \tag{3.5}$$

in which M is a positive constant, namely the penalty parameter, which should be chosen appropriately (see [18] for more details). The resulting estimator is called the modified LTS counter multicollinearity (MLTSCM) estimator. As seen, we add the last term of the objective function of (3.5) to have control on the spectral condition number of the matrix $\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{X}$. More exactly, if $\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{X}$ is ill-conditioned, then $\kappa(\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{X})$ is large and so, the current choice of the binary variables $z_1,\ldots,z_n$ is not appropriate in the perspective of the objective function value. In such situation, it is better to change the values of $z_1,\ldots,z_n$ to achieve a smaller value of $\kappa(\mathbf{X}^\top\mathbf{Z}\mathbf{X})$. In other words, last term of the objective function of (3.5) is added as a penalty for choosing inappropriate values for $z_1,\ldots,z_n$, or equivalently, as a penalty for leading to an ill-conditioned matrix $\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{X}$. As known, well-conditioned matrices enhance the numerical stability [1] and consequently, the obtained results are more accurate. The worthwhile advantage of the extended RM (3.5) over the ridge and QR-based approaches is that it does not distort the data at all.

In addition to the difficulties mentioned in Section 2.2 which may appear by solving a relaxed version of (3.5), computing the objective function derivative in (3.5) is practically expensive; especially, we may encounter with considerable numerical errors in large scale (ill-conditioned) cases. Thus, considering the argument of Section 2, here we solve the problem (3.5) by the well-known TS algorithm, although it can be effectively solved by every evolutionary algorithm.

4. Numerical experiments

In this section, based on the solution approach flowchart illustrated by Fig. 1, we make some numerical experiments to support our assertions. At first, we consider a simulation scheme to investigate efficiency of the suggested estimators and next, we consider two real-world examples related to the bridge construction and the electricity consumption. Here, we stopped the TS algorithm by reaching the maximum of 1000 iterations. Also, we set $M = 10$ in (3.5).

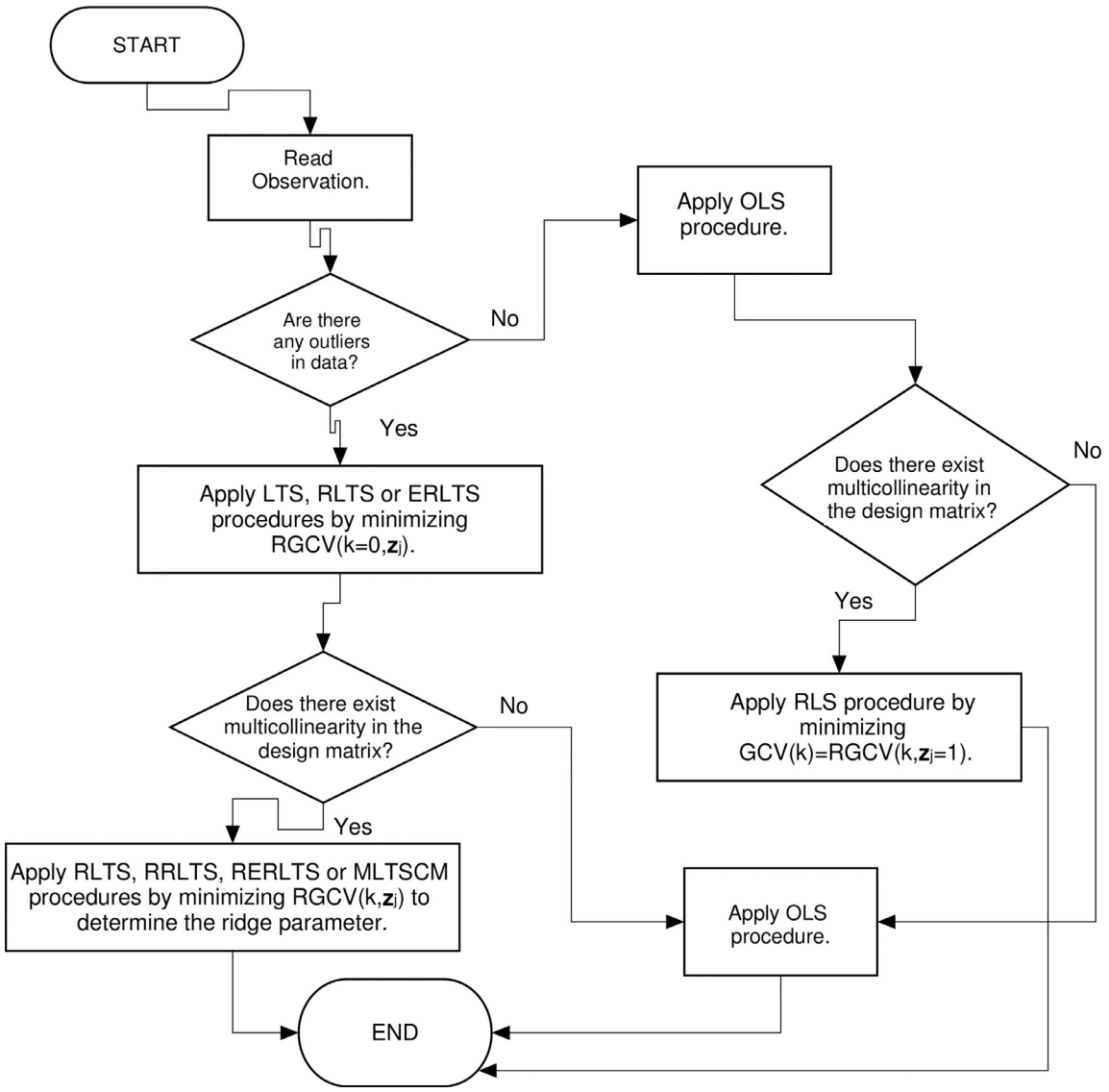

Fig. 1. General flowchart of the solution approach.

4.1. A simulated data example

Here, we computationally evaluate accuracy of the suggested estimators when multicollinearity appears in the matrix $\boldsymbol{X}$ with contaminated data. To achieve high degree of collinearity, following the approach of [19,20] the explanatory variables were generated for $n = 250,\ p = 5$ and $\gamma = 0.90$ using the following model:

$$x_{ij} = (1-\gamma^2)^{\frac{1}{2}} z_{ij} + \gamma z_{ip},\ \ i = 1, \ldots, n,\ \ j = 1, \ldots, p,$$

where z_{ij} are independent standard normal pseudo-random numbers and γ is specified so that the correlation between any two explanatory variables to be equal to γ^2. These variables are then standardized so that $\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{X}^\top\boldsymbol{y}$ are in correlation forms. Afterwards, n observations for the dependent variable are determined by

$$y_i = \beta_0 + \sum_{j=1}^{5} \beta_j x_{ij} + \epsilon_i,\ \ i = 1, \ldots, n, \tag{4.1}$$

where

$$\boldsymbol{\beta} = (4, -1, 2, -5, -3, 5)^\top,$$

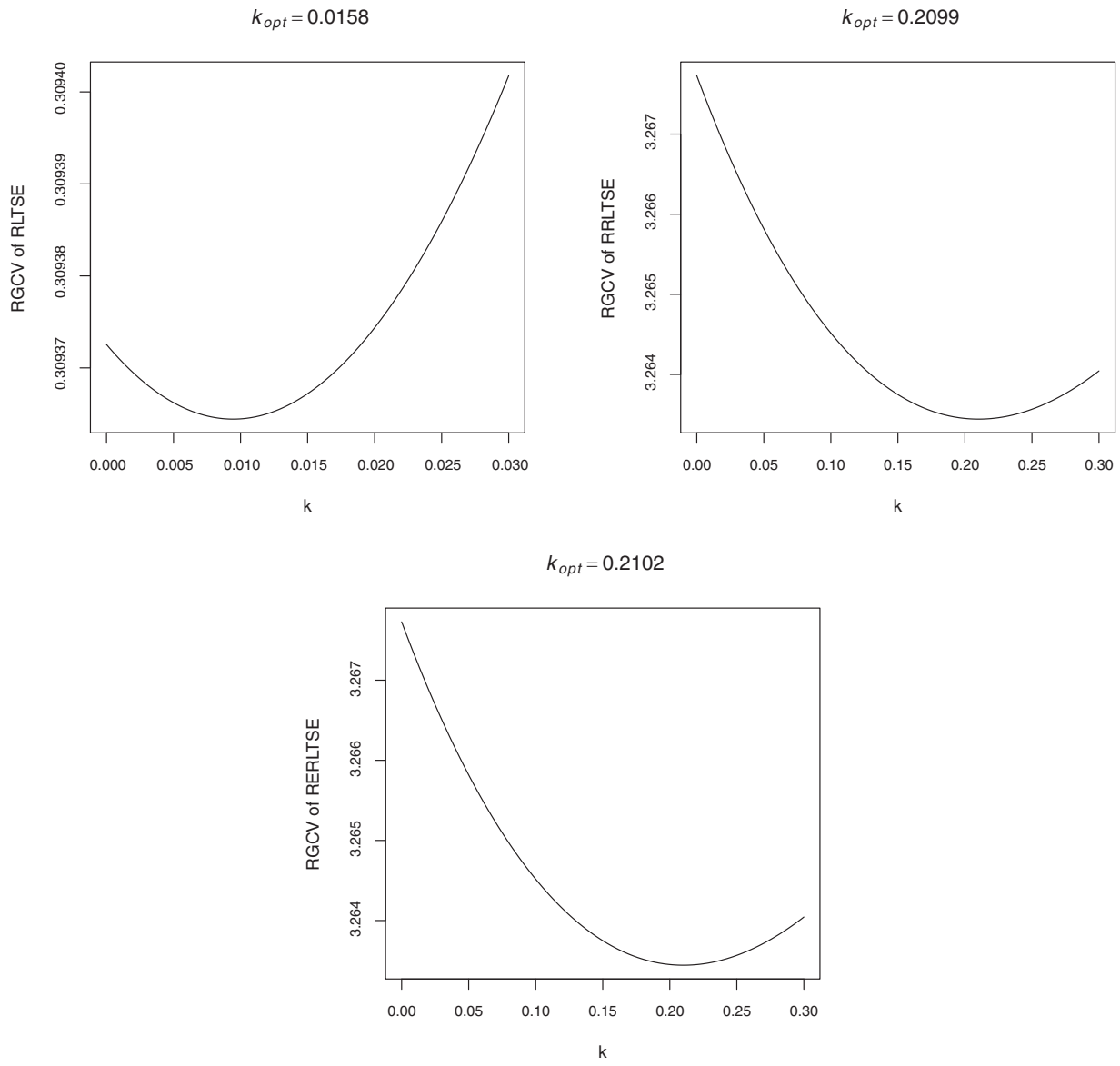

Fig. 2. The diagram of RGCV(k, $\boldsymbol{z}_j$) versus ridge parameter for $\gamma = 0.90$.

and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^\top, \boldsymbol{\epsilon}_2^\top)^\top$, with

$$\boldsymbol{\epsilon}_{1\ (h\times 1)} \sim \mathcal{N}_h(\mathbf{0}, \sigma^2 \boldsymbol{I}_n),\ \ h = 0.60n,\ \ \sigma^2 = 0.44,\ \ \boldsymbol{\epsilon}_{2\ ((n-h)\times 1)} \overset{i.i.d.}{\sim} \chi_1^2(15),$$

in which $\chi_\nu^2(\delta)$ stands for the noncentral Chi-squared distribution where δ and ν are the noncentrality parameter and degrees of freedom, respectively. Note that using such distribution for error terms corrupts the data which here is necessary for checking robustness of the estimators. We produce the first h and the last $n-h$ error terms as random variables from dependent normal and independent noncentral Chi-squared distributions, 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.

All computations were done using the statistical package R. The spectral condition numbers of the design matrices in model (4.1) are approximately equal to $\kappa_2(\boldsymbol{S}) = \lambda_1/\lambda_6 = 13659.97$ for $\gamma = 0.90$. The diagrams of RGCV versus ridge parameter are plotted for RLTSE, RRLTSE and RERLTSE in Fig. 1. In Table 1, we have illustrated the proposed estimators, respectively. We numerically calculated the SSE $= f(\hat{\boldsymbol{\beta}}, \boldsymbol{z})$ and $R^2 = 1 - SSE/S_{yy}$ for the proposed estimator. They are measures for the error and goodness of prediction, respectively. More exactly, $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$ and $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, with $\hat{y}_i = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$ where $\boldsymbol{x}_i^\top = (1, x_{i1}, \ldots, x_{ip})$. As seen in Table 1, the MLTSCM method fits the simulated data better than the others.

Table 1
Evaluated parameters for different estimators with $\gamma = 0.90$.

Coefficients/Method	OLS	RLTS	RRLTS	RERLTS	MLTSCM
$\hat{\beta}_1$	16.4668	4.2309	4.8789	4.8310	4.3115
$\hat{\beta}_2$	−2.2759	−1.0983	−1.1786	−1.1726	−1.0596
$\hat{\beta}_3$	1.6309	1.8442	1.8241	1.8262	1.7871
$\hat{\beta}_4$	−5.9407	−5.0901	−5.0957	−5.0923	−5.1168
$\hat{\beta}_5$	−5.3705	−2.9104	−3.0662	−3.0556	−2.9794
$\hat{\beta}_6$	8.3224	5.1606	5.3490	5.3341	5.2518
R^2	0.1508	0.9426	0.9484	0.9509	**0.9536**

Table 2
Bridge construction data.

Observation	Time	CCost	Dwgs	Length	Spans	DArea
1	78.8	82.4	6	90	1	3.60
2	309.5	422.3	12	126	2	5.33
3	184.5	179.8	9	78	1	6.29
4	69.6	100.0	5	60	1	2.20
5	68.8	103.0	5	60	1	1.44
6	95.7	134.4	5	60	1	5.40
7	112.3	173.2	5	180	3	6.60
8	171.9	207.9	7	188	2	7.90
9	177.8	327.7	9	336	2	4.88
10	65.8	56.2	3	25	1	0.85
11	51.2	46.8	3	50	1	1.45
12	59.6	118.9	6	114	2	4.10
13	85.4	113.3	6	108	2	3.89
14	138.1	309.0	6	128	2	5.48
15	125.7	309.0	6	128	2	5.48
16	70.1	106.1	5	90	1	3.78
17	342.9	374.5	7	430	6	19.58
18	70.3	98.3	5	50	1	1.50
19	124.9	99.3	6	68	1	4.83
20	90.0	60.0	5	64	1	1.89
21	111.7	67.4	5	70	1	3.43
22	104.7	1123.6	9	273	2	33.18
23	82.1	123.6	6	95	1	4.22
24	199.8	222.9	9	73	1	7.43
25	256.9	498.0	15	185	3	9.98
26	179.7	563.3	9	395	6	22.53
27	296.6	749.1	9	450	5	45.00
28	159.6	187.3	10	140	3	10.36
29	182.1	187.3	10	140	3	10.36
30	173.0	500.5	9	308	3	22.50
31	216.0	701.7	12	438	4	23.07
32	329.8	1066.8	12	405	4	42.67
33	325.3	766.7	15	902	7	30.67
34	47.8	30.0	4	35	1	1.01
35	57.0	65.5	5	70	1	2.52
36	59.4	63.7	5	60	1	2.16
37	59.2	68.4	4	70	1	2.14
38	84.4	187.3	5	242	5	7.74
39	202.1	421.4	9	369	1	13.65
40	46.6	107.7	6	124	1	4.22
41	257.5	1264.1	11	860	5	20.73
42	167.1	220.1	7	220	3	7.48
43	418.1	336.2	11	285	3	7.21
44	302.2	641.4	8	560	5	10.89
45	87.2	70.2	6	90	1	3.24

4.2. Application to bridge construction data

To more reasonably show achievements of our study, a real-world data adopted from [21] (pp. 130) is also analyzed in which information from 45 projects of bridge construction was gathered. They are illustrated in Table 2. Also, the variables are given as follows:

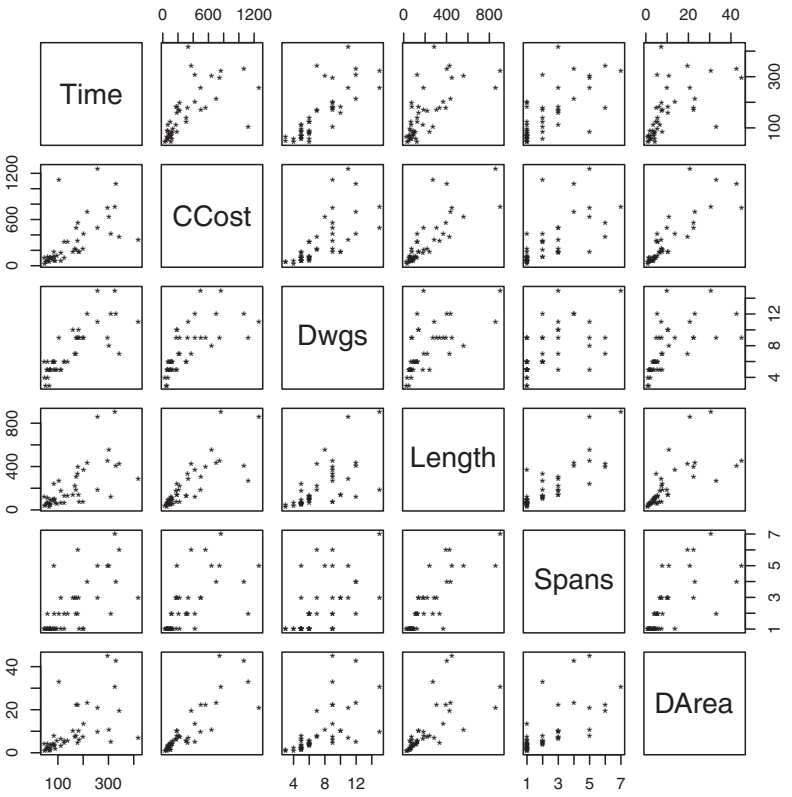

Fig. 3. Matrix of scatter plot for the response variable/predictors.

Time: Time of construction (person-days);
CCost: Cost of construction ($000);
Dwgs: Structural drawings number;
Length: Bridge length (ft);
Spans: Spans number;
DArea: Deck area of bridge (000 sq ft).
At first, let us consider the following model:

$$Time = \beta_0 + \beta_1 CCost + \beta_2 Dwgs + \beta_3 Length + \beta_4 Spans + \beta_5 DArea + \epsilon. \tag{4.2}$$

To investigate relations of the variables, we illustrate the scatter plot of the data by Fig. 3. As seen, there exists a striking nonlinear pattern among the response and some of the predictor variables and also, some of the variables seem to be skewed. Moreover, there is an evidence of nonconstant variance according to Fig. 4. So, it is necessary to transform the variables using the effective multivariate Box-Cox approach (see [22] for more details). Below, the output of R using the powerTransform command from alr3 is given:

```
bcPower Transformations to Multinormality
          Est.Power   Std.Err.   Wald Lower Bound   Wald Upper Bound
 Time       -0.1795     0.2001      -0.5716           0.2127
 DArea      -0.1346     0.0893      -0.3096           0.0404
 CCost      -0.1762     0.0942      -0.3609           0.0085
 Dwgs       -0.2507     0.2402      -0.7215           0.2200
 Length     -0.1975     0.1073      -0.4078           0.0127
 Spans      -0.3744     0.2594      -0.8828           0.1340
Likelihood ratio tests about transformation parameters
                                    LRT      df      pval
 LR test, lambda = (0 0 0 0 0 0)   8.121991   6    0.2293015
 LR test, lambda = (1 1 1 1 1 1)   283.1840   6    0.0000000
```

Based on the output detailed above, the logarithmic transformation is suitable. The matrix of scatter plot of the log-transformed variables is shown in Fig. 5. In contrast to Fig. 3, Fig. 5 shows that the transformation improved the linear

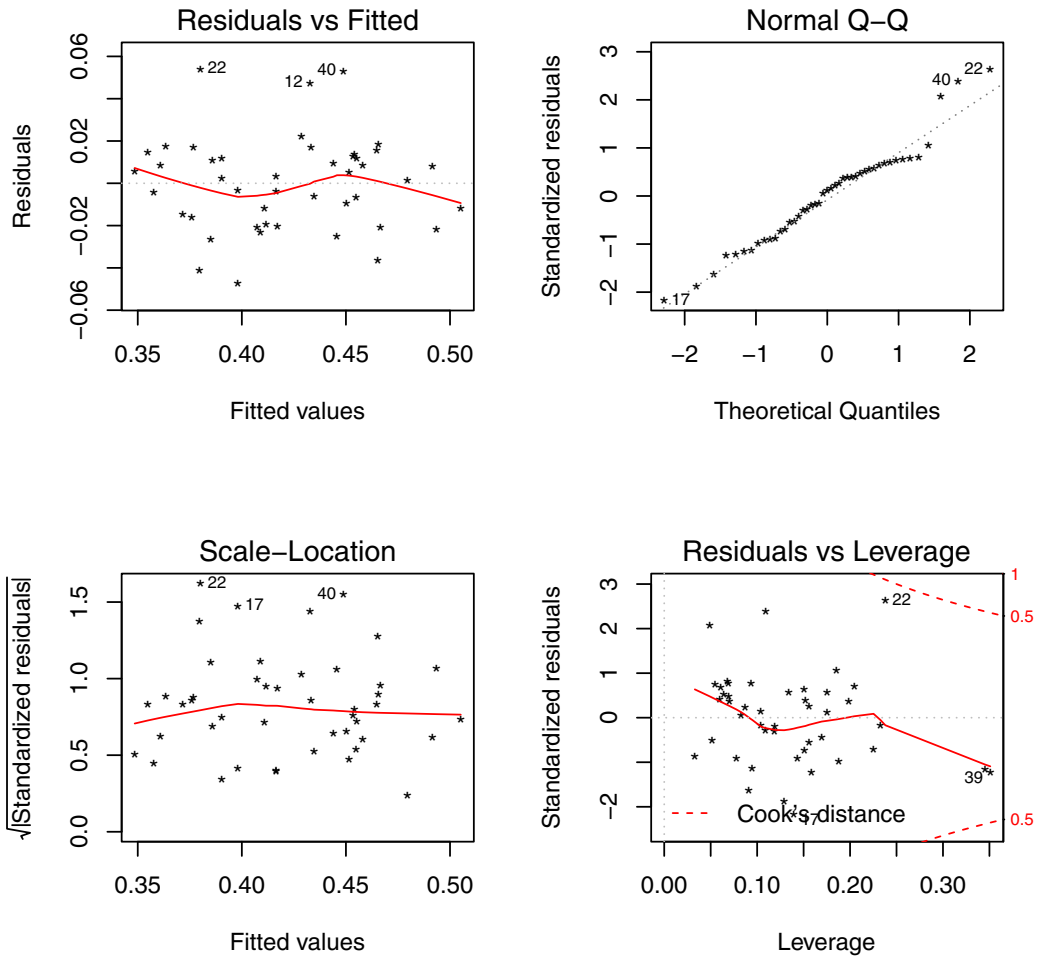

Fig. 4. Diagnostic plots for the model (4.2).

relationships between the response and predictor variables. Now, the transformed RM is considered as follows:

$$\log(Time) = \beta_0 + \beta_1 \log(CCost) + \beta_2 \log(Dwgs) + \beta_3 \log(Length) + \beta_4 \log(Spans) + \beta_5 \log(DArea) + \epsilon, \tag{4.3}$$

which leads to the following output:

```
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
 Intercept    2.28590     0.61926     3.691    0.000681
log(CCost)    0.19609     0.14445     1.358    0.182426
log(Dwgs)     0.85879     0.22362     3.840    0.000440
log(Length)  -0.03844     0.15487    -0.248    0.805296
log(Spans)    0.23119     0.14068     1.643    0.108349
log(DArea)   -0.04564     0.12675    -0.360    0.720705
Residual standard error: 0.3139 on 39 degrees of freedom
Multiple R-squared: 0.7762,   Adjusted R-squared: 0.7475
F-statistic: 27.05 on 5 and 39 DF,  p-value: 1.043e-1.
```

As seen, while the overall F-test for the model (4.3) is statistically significant in a high level, only one of the evaluated coefficients (β_2) is significant. As a more troubling result, the estimations of β_3 and β_5 are not positive while it is trivial that bridges with longer length or larger area need more time to be constructed. These wrong signs for β_3 and β_5 are the result of multicollinearity between the predictors.

Below, the correlation matrix of the transformed predictors is given, demonstrating that some of the correlations are considerable:

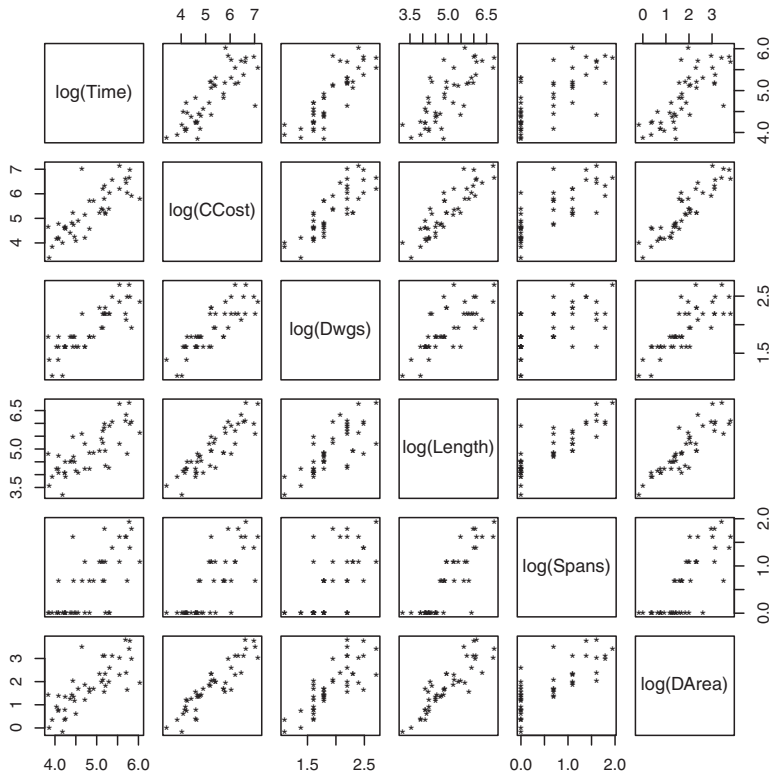

Fig. 5. Matrix of scatter plot of the transformed variables.

Table 3
The best subset of predictors based on the values of R^2_{adj} and AIC for the bridge construction data set.

Subset size	Predictor variables	R^2_{adj}	AIC
1	log(*Dwgs*)	0.702	−94.90
2	log(*Dwgs*), log(*Spans*)	0.753	−102.37
3	**log(Dwgs), log(Spans), log(CCost)**	**0.758**	**−102.41**
4	log(*Dwgs*), log(*Spans*), log(*CCost*), log(*DArea*)	0.753	−100.64
5	log(*Dwgs*), log(*Spans*), log(*CCost*), log(*DArea*), log(*Length*)	0.748	−98.71

	TCCost	TDwgs	TLength	TSpans	TDArea
TCCost	1.0000	0.8315	0.8905	0.7751	0.9092
TDwgs	0.8315	1.0000	0.7523	0.6297	0.8012
TLength	0.8905	0.7523	1.0000	0.8585	0.8842
TSpans	0.7751	0.6297	0.8585	1.0000	0.7815
TDArea	0.9092	0.8012	0.8842	0.7815	1.0000

To detect the linear relation between the response and the predictor variables of the model, we used the Added-variable plots [22] which provides a visual tool to investigate the effect of each predictor, separately. According to the plot shown by Fig. 6, the predictors log(*DArea*) and log(*Length*) have statistically insignificant linear relationship with the response variable. As seen in the figure, the slopes of the least-squares lines are approximately equal to zero for these variables. The adjusted R-squared and AIC criteria approved this result which reported in Table 3 (Table 1, Section 2, Ch. 7, p.235 in [22]). So, specification of the final RM is

$$\log(Time) = \beta_0 + \beta_1 \log(CCost) + \beta_2 \log(Dwgs) + \beta_4 \log(Spans) + \epsilon. \tag{4.4}$$

For the above model, the condition number is equal to 2258.34. This means that there exists a strong multicollinearity in the new design matrix. We obtained the optimal values of k (k_{opt}) for the ridge estimators using the RGCV (see Fig. 7). Table 4 shows summary of the results. As seen, MLTSCM turns out to be practically promising in contrast to the others. Also, the results of RERLTS and MLTSCM are reasonable.

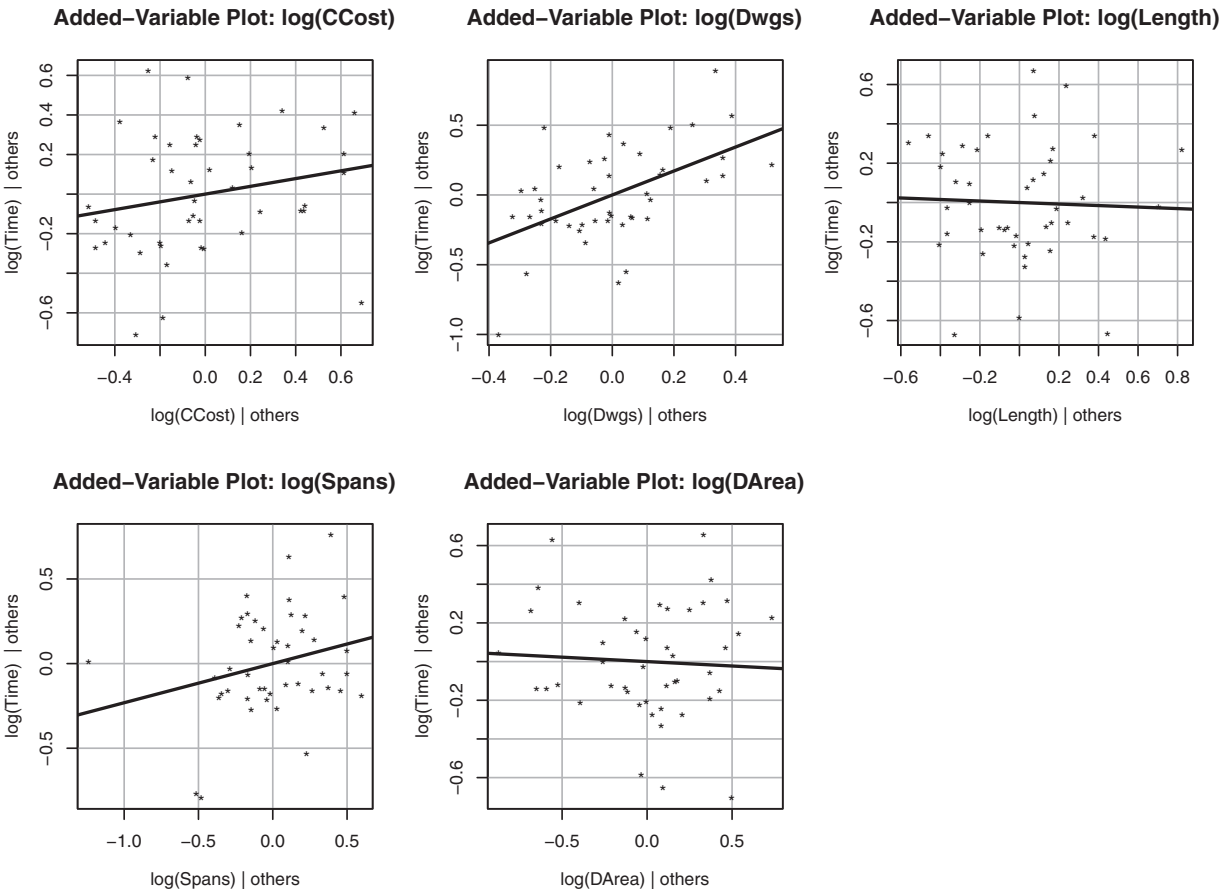

Fig. 6. Added-variable plots of individual explanatory variables versus dependent variable for the the bridge construction data set.

Table 4
Estimated regression coefficients of the model (4.4) for the bridge construction data set.

Coefficients/Method	OLS	RLTS	RRLTS	RERLTS	MLTSCM
Intercept	2.3317	1.91363	1.7999	1.8923	2.0304
log(*CCost*)	0.1483	0.33718	0.3464	0.3320	0.3056
log(*Dwgs*)	0.8356	0.58002	0.6423	0.6272	0.6210
log(*Spans*)	0.1963	0.06662	0.0012	0.0315	0.0657
SSE	3.8692	2.3018	2.0034	1.9788	**1.9778**
R^2	0.7747	0.8480	0.8544	0.8579	**0.8600**

4.3. Application to electricity consumption data set

In order to provide another practical example to examine the efficiency of our estimators, we use a real-world data related to the electricity consumption, firstly considered by Akdenïz Duran et al. [23]. The data are available at http://www.quantlet.org. Since electricity is a non-storable good, electricity providers are interested in understanding and hedging demand fluctuations. Electricity consumption is known to be affected by the temperature, the price of electricity and the income of the consumers.

The variables of the problem comprise the log monthly electricity consumption per person (*LEC*) as the response variable and the predictors log income per person (*LI*), log rate of electricity price to the gas price (*LREG*) and cumulated average temperature index (*Temp*) for the corresponding month taken as average of 20 German cities computed from the data of German weather service for 177 cases. So, the model can be specified as

$$(LEC)_i = \beta_0 + \beta_1(LI)_i + \beta_2(LREG)_i + \beta_3(Temp)_i + \sum_{j=4}^{14} \beta_j x_{ij} + \epsilon_i, \tag{4.5}$$

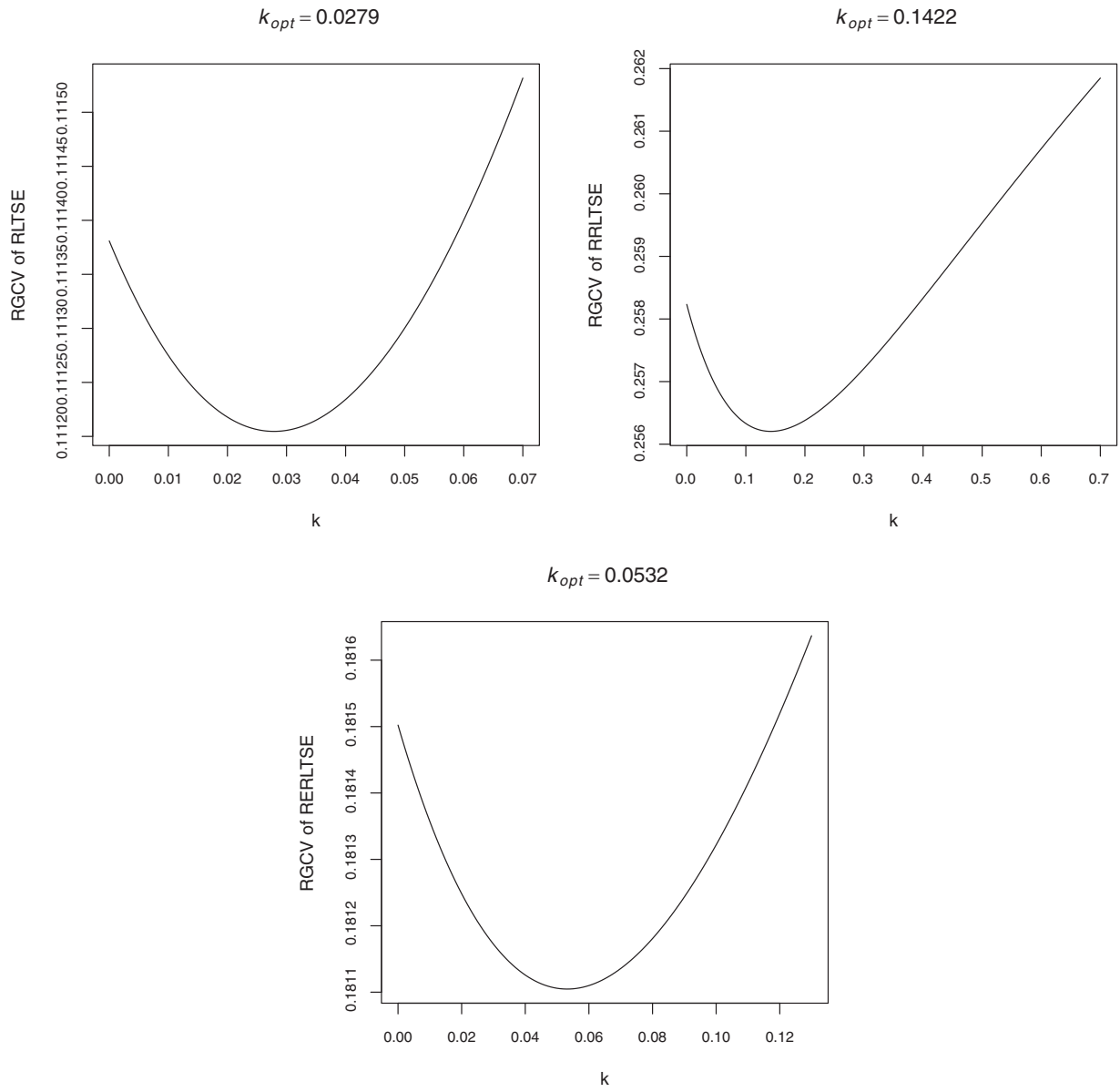

Fig. 7. The diagram of RGCV(k, $\boldsymbol{z}_j$) versus the ridge parameter for the bridge construction data set.

where x_j are dummy variables for the monthly effects. Fig. 8 shows the added-variable plot of model (4.5). As seen, the lack of statistical significance of the predictor variables x_j, $j = 4, 5, \ldots, 14$, is evident. This result is approved by the adjusted R-squared and AIC criterions (see Table 5). So, the final RM can be written as

$$(LEC)_i = \beta_0 + \beta_1 (LI)_i + \beta_2 (LREG)_i + \beta_3 (Temp)_i + \epsilon_i. \tag{4.6}$$

The condition number of the new design matrix in the model (4.6) is approximately $\lambda_4/\lambda_1 = 18404279.00$, and so, there exists a very strong multicollinearity among the columns of the design matrix. We obtained the optimal values of k (k_{opt}) for the ridge estimators using RGCV (see Fig. 9). Table 6 shows summery of the results. As seen, MLTSCM turns out to be practically promising in contrast to the others. Also, we achieved $R^2 = 0.739$ using the MLTSCM method for a subset of three predictor variables while Akdenïz Duran et al. [23] achieved $R^2 = 0.749$ using all of the predictor variables. Note that R^2 has insensible growth when the number of predictors increase from 3 to 14, and so the three-predictor subset is preferable because of the easier interpretation. This implies that the predictors x_j, $j = 4, 5, \ldots, 14$, have insignificant linear relationship with the response variable. The results of Table 5 and Fig. 8 also confirm this achievement.

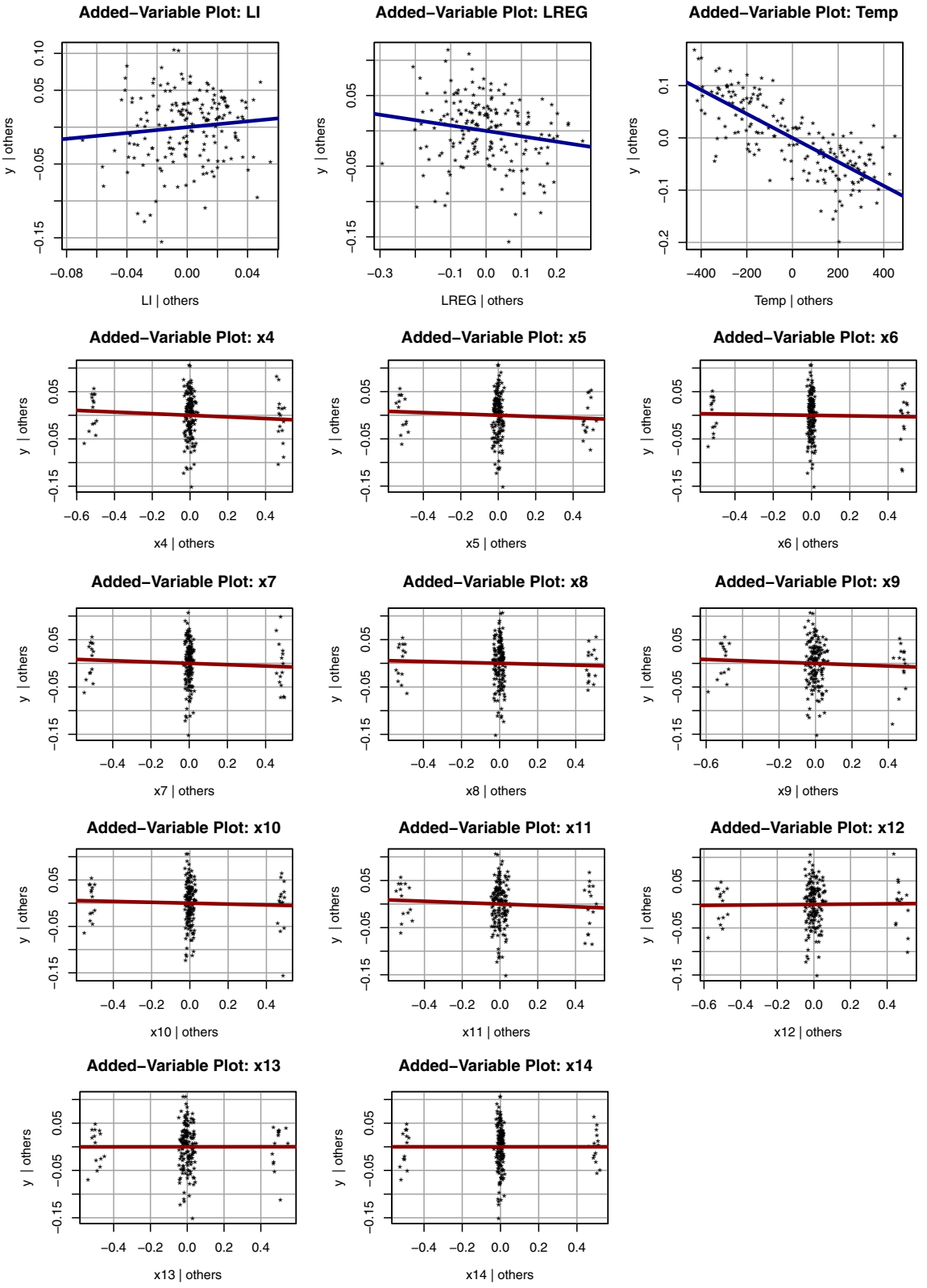

Fig. 8. Added-variable plots of individual explanatory variables versus dependent variable for the electricity consumption data set.

Table 5
The best subset of predictors based on the values of R^2_{adj} and AIC for the electricity consumption data set.

Subset size	Predictor variables	R^2_{adj}	AIC
1	*Temp*	0.5523	−1067.814
2	*Temp*, *LREG*	0.5781	−1077.339
3	**Temp, LREG, LI**	**0.5892**	**−1081.063**
4	*Temp*, *LREG*, *LI*, x_{12}	0.5891	−1080.057
5	*Temp*, *LREG*, *LI*, x_{12}, x_{13}	0.5882	−1078.709
6	*Temp*, *LREG*, *LI*, x_{12}, x_{13}, x_{14}	0.5875	−1077.427
7	*Temp*, *LREG*, *LI*, x_{12}, x_{13}, x_{14}, x_4	0.5858	−1075.734
8	*Temp*, *LREG*, *LI*, x_{12}, x_{13}, x_{14}, x_4, x_6	0.5837	−1073.897
9	*Temp*, *LREG*, *LI*, x_{12}, x_{13}, x_{14}, x_4, x_6, x_8	0.5812	−1071.907
10	*Temp*, *LREG*, *LI*, x_{12}, x_{13}, x_{14}, x_4, x_6, x_8, x_7	0.5789	−1069.987
11	*Temp*, *LREG*, *LI*, x_{12}, x_{13}, x_{14}, x_4, x_6, x_8, x_7, x_{10}	0.5764	−1067.997
12	*Temp*, *LREG*, *LI*, x_{12}, x_{13}, x_{14}, x_4, x_6, x_8, x_7, x_{10}, x_5	0.5740	−1064.098
13	*Temp*, *LREG*, *LI*, x_{12}, x_{13}, x_{14}, x_4, x_6, x_8, x_7, x_{10}, x_5, x_9	0.5718	−1064.281
14	*Temp*, *LREG*, *LI*, x_{12}, x_{13}, x_{14}, x_4, x_6, x_8, x_7, x_{10}, x_5, x_9, x_{11}	0.5709	−1063.014

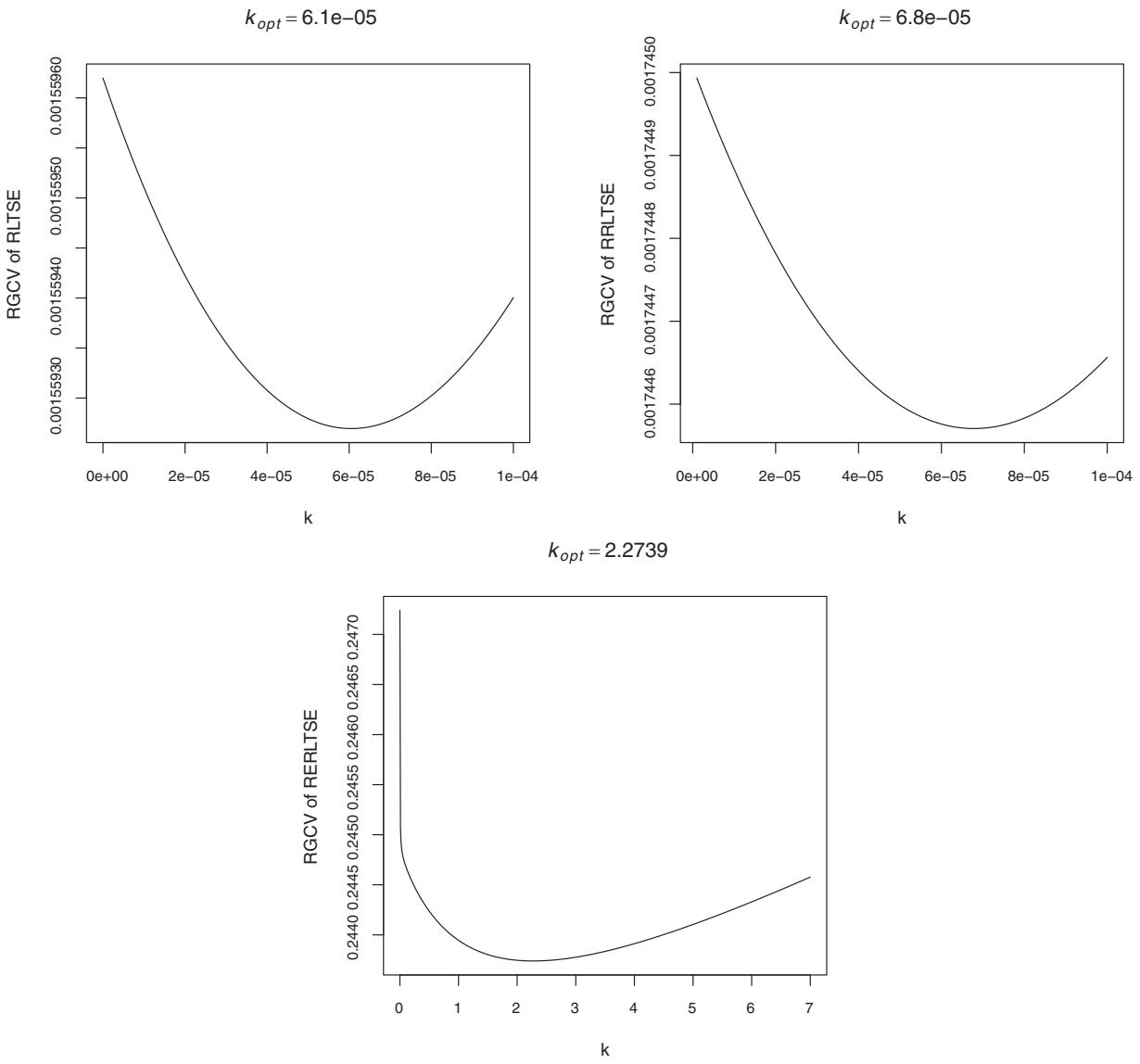

Fig. 9. The diagram of RGCV(k, $\boldsymbol{z}_j$) versus the ridge parameter for the electricity consumption data set.

Table 6
Estimated regression coefficients of the model (4.6) for the electricity consumption data set.

Coefficients/Method	OLS	RLTS	RRLTS	RERLTS	MLTSCM
Intercept	4.40692	5.16925	4.85097	5.08809	4.98810
LI	0.19253	0.09885	0.13730	0.10721	0.11461
LREG	−0.07775	−0.09387	−0.08908	−0.09579	−0.10541
Temp	−0.00023	−0.00023	−0.00023	−0.00023	−0.00025
SSE	0.37649	0.26372	0.26386	0.23526	**0.19815**
R^2	0.59620	0.67418	0.67535	0.69854	**0.73990**

5. Conclusions

A range of procedures in robust techniques require optimization of an objective function over all the subsamples of the given size. Such combinatorial problems are often extremely difficult to solve exactly. In this regards, we have proposed several optimization approaches for RMs to simultaneously combat multicollinearity and outliers in the data set. Especially, based on a penalization scheme we have suggested a nonlinear integer programming problem as the RM which can be effectively solved by any evolutionary algorithm. Among them, here we employed the well-known tabu search algorithm for the LTS regression estimation. The worthwhile advantages of our method (MLTSCM) in contrast to the rival methods in similar situations such as ridge and QR-based approaches are (i) the data are not distorted at all; (ii) there is no need to solve a complicated problem for choosing the biasing parameter. The results of this work can potentially be extended to the case of heteroscedastic or correlated errors. The GCV is a popular parameter selection criterion, but similar to the OLSE, it yields a severely undersmoothed estimate in the presence of outliers and multicollinearity. For this reason, the stable robust GCV (RGCV) criterion was introduced for obtaining the optimal value of the ridge parameter. To numerically compare performance of the proposed estimators, we have also studied a simulation case and two real-world applications related to the bridge construction data and the electricity consumption data. The results showed that the proposed methods RERLTS and MLTSCM are reasonably efficient in contrast to the OLS method.

Acknowledgments

This research was supported by Research Council of Semnan University. The first author's research is also supported in part by a grant No. 96001062 from the Iran National Science Foundation (INSF). The authors are grateful to the anonymous reviewers, associate editor and editor-in-chief for their valuable comments and suggestions helped to improve the quality of this work.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.apm.2017.11.011.

References

[1] D.S. Watkins, Fundamentals of Matrix Computations, John Wiley and Sons, New York, 2002.
[2] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for non-orthogonal problems, Technometrics 12 (1970) 55–67.
[3] M.R. Baye, D.F. Parker, Combining ridge and principal component regression: a money demand illustration, Commun. Stat. Theory Methods 13 (1984) 197–205.
[4] K.J. Liu, A new class of biased estimate in linear regression, Commun. Stat. Theory Methods 22 (1993) 393–402.
[5] S. Kaçiranlar, S. Sakallioğlu, Combining the Liu estimator and the principal component regression estimator, Commun. Stat. Theory Methods 30 (2001) 2699–2705.
[6] M. Roozbeh, S. Babaie-Kafaki, M. Arashi, A class of biased estimators based on QR decomposition, Linear Algebra Appl. 508 (2016) 190–205.
[7] S. Babaie-Kafaki, R. Ghanbari, N. Mahdavi-Amiri, Hybridizations of genetic algorithms and neighborhood search metaheuristics for fuzzy bus terminal location problems, Appl. Soft Comput. 46 (2016) 220–229.
[8] R.J. Vanderbei, Linear Programming: Foundations and Extensions, Springer, New York, 2008.
[9] F. Glover, Future paths for integer programming and links to artificial intelligence, Comput. Oper. Res. 13 (1986) 533–549.
[10] A.R. Hedar, M. Fukushima, Tabu search directed by direct search methods for nonlinear global optimization, Eur. J. Oper. Res. 170 (2006) 329–349.
[11] Z. Lü, J.K. Hao, Adaptive tabu search for course timetabling, Eur. J. Oper. Res. 200 (2010) 235–244.
[12] L. Chen, A. Langevin, D. Riopel, A tabu search algorithm for the relocation problem in a warehousing system, Int. J. Prod. Econ. 129 (2011) 147–156.
[13] A.V. Dorugade, Adjusted ridge estimator and comparison with Kibria's method in linear regression, J. Assoc. Arab Univ. Basic Appl. Sci. 21 (2016) 96–102.
[14] M. Roozbeh, Robust ridge estimator in restricted semiparametric regression models, J. Multivar. Anal. 147 (2016) 127–144.
[15] T.D. Nguyen, R. Welsch, Outlier detection and least trimmed squares approximation using semi-definite programming, Comput. Stat. Data Anal. 54 (2010) 3212–3226.
[16] M. Roozbeh, S. Babaie-Kafaki, Extended least trimmed squares estimator in semiparametric regression models with correlated errors, J. Stat. Comput. Simul. 86 (2016) 357–372.
[17] M. Amini, M. Roozbeh, Optimal partial ridge estimation in restricted semiparametric regression models, J. Multivar. Anal. 136 (2015) 26–40.
[18] W. Sun, Y.X. Yuan, Optimization Theory and Methods: Nonlinear Programming, Springer, New York, 2006.
[19] G.C. McDonald, D.I. Galarneau, A Monté–Carlo evaluation of some ridge-type estimators, J. Am. Stat. Assoc. 70 (1975) 407–416.

[20] D.G. Gibbons, A simulation study of some ridge estimators, J. Am. Stat. Assoc. 76 (1981) 131–139.
[21] P. Tryfos, Methods for Business Analysis and Forecasting: Text & Cases, John Wiley and Sons, New York, 1998.
[22] S.J. Sheather, A Modern Approach to Regression with R, Springer, New York, 2009.
[23] E.A. Duran, W.K. Härdle, M. Osipenko, Difference-based ridge and Liu type estimators in semiparametric regression models, J. Multivar. Anal. 105 (2012) 164–175.