TWO PENALIZED MIXED–INTEGER NONLINEAR PROGRAMMING APPROACHES TO TACKLE MULTICOLLINEARITY AND OUTLIERS EFFECTS IN LINEAR REGRESSION MODELS

MAHDI ROOZBEH*
SAMAN BABAIE–KAFAKI AND ZOHRE AMINIFARD

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

(Communicated by Harish Garg)

ABSTRACT. In classical regression analysis, the ordinary least–squares estimation is the best strategy when the essential assumptions such as normality and independency to the error terms as well as ignorable multicollinearity in the covariates are met. However, if one of these assumptions is violated, then the results may be misleading. Especially, outliers violate the assumption of normally distributed residuals in the least–squares regression. In this situation, robust estimators are widely used because of their lack of sensitivity to outlying data points. Multicollinearity is another common problem in multiple regression models with inappropriate effects on the least–squares estimators. So, it is of great importance to use the estimation methods provided to tackle the mentioned problems. As known, robust regressions are among the popular methods for analyzing the data that are contaminated with outliers. In this guideline, here we suggest two mixed–integer nonlinear optimization models which their solutions can be considered as appropriate estimators when the outliers and multicollinearity simultaneously appear in the data set. Capable to be effectively solved by metaheuristic algorithms, the models are designed based on penalization schemes with the ability of down–weighting or ignoring unusual data and multicollinearity effects. We establish that our models are computationally advantageous in the perspective of the flop count. We also deal with a robust ridge methodology. Finally, three real data sets are analyzed to examine performance of the proposed methods.

1. Introduction. In general, a data set can be modeled by parametric or non–parametric approaches [20]. Including linear, logistic and polynomial regression models, parametric schemes are applied when relationship between the input and output variables is known. However, when relationship between the variables is not clear, non–parametric curve estimation using kernels, splines and local polynomials can be helpful.

As known, parametric models are relatively simple to describe and implement, being also advantageous in contrast to the other approaches in terms of interpretation
and inference. But non-parametric approaches often suffer from the major disadvantage of requiring a huge number of observations to obtain an accurate estimate of the population model, because they do not reduce the problem of estimating the model to a small number of parameters. Also, since the relationship between each predictor and the corresponding response is modeled by a curve, non-parametric approaches are somewhat less interpretable than the linear regression. As a type of parametric regression, the most common functional form is the parametric linear model, frequently used to describe the relationship between a dependent variable and the explanatory variables. Regression analysis is a method for investigating the functional relationship among the variables. Since it commonly happens that more than one factor influence the outcome, fitting regression models to data involving two or more predictors is one of the most widely used statistical procedures.

Multiple regression model (MRM) is arguably the most widely used statistical tool in applied econometrics, mathematical finance, social sciences and other fields. MRM can be used to analyze the marketing effectiveness as well as pricing and promotions on sales of a product [26]. It can be also used to assess risk in financial services or insurance domain. General purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. The model is given by

\[ y = X\beta + \epsilon, \]

where \( y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n, X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times p} \) is a matrix of observations on the explanatory variables, \( \beta = (\beta_1, \ldots, \beta_p)^\top \) is a vector of unknown regression coefficients and \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top \) is a vector of error terms with \( E(\epsilon) = 0 \) and \( E(\epsilon \epsilon^\top) = \sigma^2 I_n \), where \( I_n \) is the unit matrix of order \( n \) and \( \sigma^2 \) is an unknown constant. The ordinary least-squares estimator (OLSE) of \( \beta \) is given by

\[ \hat{\beta} = \arg \min_{\beta} (y - X\beta)^\top (y - X\beta) = S^{-1}X^\top y, \]

where \( S = X^\top X \). In regression analysis, researchers often encounter with some problems such as collinearity between explanatory variables and outliers existence in the data set. In what follows, we review some common methods for solving the mentioned problems.

1.1. Review of the literature. Existence of outliers is a common problem in regression analysis. Generally, an outlier is an individual that lies outside the theme or the society which it is a piece of. In statistics, an outlier is a point that is well out of the expected scope of the amounts in a research, and in the regression modeling, an outlier is an observation point that fails to track the linear design of the data. Outliers corrupt the OLSE and this fact motivated the researchers to investigate robust estimation (see [34]). The robust regression estimation is one of the most popular interests in the statistical methods.

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. The basic measure of robustness of an estimator is its breakdown point; that is, the fraction (up to 50%) of outlying data points that can corrupt the estimator arbitrarily. The study of efficient algorithms for robust estimators has been an active area of research in
computational geometry. As examples, Holland and Welsch [19] used the iterative reweighed least-squares procedure in which the weights are set inversely proportional to the squared residuals. Rousseeuw [32] proposed least median of squares (LMS) regression which is defined as the hyperplane that minimizes the median squared residual. Although the vast majority of works on robust linear estimation in the field of computational geometry have been devoted to study the LMS estimator, it has been observed that LMS does not yield an appropriate estimator in the perspective of statistical properties. More precisely, Rousseeuw and Leroy [33] argued that a better choice is the least trimmed squares (LTS) which involves computing the hyperplane that minimizes the sum of the smallest $h$ squared residuals, where $h \leq n$ is the given integer trimming parameter. LTS is a robust estimator with a 50%-breakdown point, which means that the estimator is insensitive to corruption due to outliers, provided that the outliers constitute less than 50% of the set. Although both the LMS and LTS estimators have 50%-breakdown, LTS has a number of advantages in contrast to LMS. For example, the LTS objective function is smoother than that of LMS. Also, LTS has better statistical efficiency because it is asymptotically normal and converges faster.

Support vector machines (SVMs) are another modern day robust techniques to combat the outliers problem. SVM is a sparse kernel decision machine that avoids computing posterior probabilities when building its learning model (see [20, 6] for more details). It offers an essential approach for machine learning problems because of its mathematical foundation in statistical learning theory. SVM constructs its solution in terms of a subset of the training input. The fact that the support vector classifier’s decision rule is based only on a potentially small subset of the training observations (the support vectors) means that it is quite robust to the behavior of observations that are far away from the hyperplane. SVM has been extensively used for classification, regression, novelty detection tasks, and feature reduction. SVM concepts can be generalized to become applicable to regression problems. As in classification, support vector regression (SVR) is characterized by the use of kernels, sparse solution, and Vapnik-Chervonenkis control of the margin and the number of support vectors. Similar to SVM, SVR has been proved to be a robust tool in real-valued function estimation. SVR instead seeks coefficients that minimize a different type of loss, where only residuals larger (in absolute value) than some positive constant contribute to form the loss function. This is an extension of the margin used in support vector classifiers to the regression setting.

As known, a neural network (NN) is a two-stage regression or classification model. The central idea is to extract linear combinations of the inputs as derived features, and then model the target as a nonlinear function of these features. The result is a powerful learning method, with widespread applications in many fields. The NN approach also identifies cluster labels for each data record. The cluster label can often help to interpret the resulting outliers. For example, outliers are sometimes found to be concentrated in a single cluster or in a group of clusters with common characteristics [16]. The cluster label not only enables the individuals but also enables groups to be identified as outliers.

Besides the outliers problem in regression analysis, one may encounter with the problem of multicollinearity that is defined as the existence of nearly linear dependency among columns of the design matrix $X$. In this situation, the matrix $S = X^TX$ has one or more small eigenvalues which causes the estimates of the regression coefficients to be large in the absolute value. As known, condition number
is an effective tool to check the existence of multicollinearity [37]. The OLSE performs poorly in the presence of multicollinearity. Also, existence of multicollinearity may lead to wide confidence intervals for the individual parameters or their linear combinations, and can produce estimators with wrong signs.

The most popular approach to combat multicollinearity is the ridge regression estimator proposed by Hoerl and Kennard in the 1970s [18]. Several other methods for dealing with multicollinearity are the \(r-k\) class estimator which combines the techniques of ridge regression and principal component regression (PCR) while utilizing the ridge technique to reduce the mean square error of principal components estimators [10]; the biased estimator which combines the advantages of ridge and Stein estimators [23]; the \(r-d\) class estimator which includes OLS, PCR and Liu estimators, being superior to the OLSE, Liu estimator and PCR estimator in the scalar mean–squared error (mse) sense [21]; and a class of biased estimators based on the Cholesky and QR decompositions which make the data to be less distorted than the other biased methods [9, 28, 29]. As a brief review on applicability of these strategies, see [3, 4, 5, 15] in which the ridge methodology has been used.

The main part of this work is devoted to study two robust least–squares estimators under dependency among columns of the design matrix in an MRM. This work is organized as follows. In Section 2, we describe some preliminaries in the sense of reviewing the ridge LTS approach and a recent nonlinear integer programming model proposed by Roozbeh et al. [30], both being capable to dominate outliers and multicollinearity synchronously. Then, in Section 3 we propose two extended versions of the main model of [30] and discuss how the metaheuristic algorithms can be effectively used to compute the related estimators. To illustrate efficiency of the proposed estimators in real applications, in Section 4 we perform some numerical tests on a bridge construction data, an electricity consumption data and a larger data set of determinants of wages. Finally, concluding remarks are provided in Section 5.

2. Preliminaries. To reduce corruption of the least–squares fit occurs due to the outliers strong effect on the objective function, LTS approach minimizes the sum of smallest \(h\)-squared residuals rather than the complete sum of squares. Here, \(h\) is a threshold such that the ratio \(\alpha = h/(n-m)\) represents percentage of the good observations. If we let \(z_i\) to be the indicator demonstrating whether the observation \(i\) is good or not, the LTS problem can be formulated as follows:

\[
\begin{array}{ll}
\min & f(\beta, z) = (y - X\beta)^\top Z(y - X\beta) \\
\text{s.t.} & z^\top e = h, \\
& z_i \in \{0, 1\}, \ i = 1, 2, \ldots, n,
\end{array}
\]  

(3)

where \(Z\) is a diagonal matrix with the diagonal elements \(z = (z_1, z_2, \ldots, z_n)^\top\), and \(e = (1, \ldots, 1)^\top \in \mathbb{R}^n\). Solution of the recent optimization problem is the LTS estimator of the regression coefficients, obtained by

\[
\hat{\beta}_{LTS}(z) = S(z)^{-1} X^\top Z y,
\]  

(4)

where \(S(z) = X^\top Z X\). As seen, LTS approach deals with a quadratic integer programming problem with one linear equality constraint. So, although being NP–hard [12], simple binary structure of the problem makes it possible to effectively find a solution of (3) by metaheuristic algorithms, even in large scale cases. Defeating the outlier effects in the estimation process, the problem (3) is basis of our optimization.
models. Nevertheless, it should be noted that the model fails to extract those columns of the design matrix $X$ which are linearly independent in an acceptable level.

As is well known, covariance matrix of the OLSE is equal to $\sigma^2 S^{-1}$. As seen, both the OLSE and its covariance matrix are heavily dependent to the characteristics of the matrix $S$. Hence, OLSE may be affected by some computational errors when $S$ is ill-conditioned or equivalently, when multicollinearity appears in the data set. The multicollinearity effect can be reduced by providing more data, modifying the model and reselecting the variables [29].

Among the popular methods to combat the multicollinearity there is the ridge regression which yields the ridge least-squares estimator (RLSE), originally proposed by Hoerl and Kennard [18] as a solution of the following optimization problem:

$$\min_{\beta} f(\beta) = (y - X\beta)^\top (y - X\beta) + k(\beta^\top \beta - \phi^2), \quad (5)$$

in which $k \geq 0$ is the Lagrange multiplier and called the ridge or the shrinkage parameter. Solving problem (5), we get

$$\hat{\beta}(k) = S(k)^{-1} X^\top y, \quad S(k) = S + kI_p, \quad k > 0. \quad (6)$$

Under squared error loss function, the risk function of RLSE is

$$R(\hat{\beta}(k), \beta) = k^2 S(k)^{-1} \beta^\top \beta + \sigma^2 S(k)^{-1} SS(k)^{-1}. \quad (7)$$

For proper values of $k$, the risk function of RLSE is smaller than that of OLSE.

A review of the literature reveals an abundance of the studies made to improve the ridge strategy; see for example [29, 31] and the references therein. Although ridge methodology is the most popular approach to solve the multicollinearity problem, it has the following two main weaknesses: (i) the data are distorted by replacing the matrix $S$ with the matrix $S(k)$; (ii) depending on some unknown parameters, choosing an appropriate value for the shrinkage parameter $k$ is a complicated problem which has not yet been solved completely. In the procedure of choosing $k$, on one side we must control the condition number of $S$ to enhance stability of the estimation. Hence, we must do our best to let the parameter $k$ to be large. Furthermore, we know the bigger the $k$ is, the smaller the covariance of the estimator is. This implies that the estimator turns out to be more stable. On the other side, we know, in view of the biased estimator, when $k$ is smaller, the estimator is better. So, we must obey some regulations to determine a proper value for $k$.

2.1. Robust ridge methodology to combat outliers and multicollinearity problems. Here, we suggest a ridge LTS (RLTS) approach to simultaneously control presence of the outliers and the multicollinearity in the data set, combining (3) and (5) in the sense of

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

$$\text{s.t.} \quad z^\top e = h, \quad z_i \in \{0, 1\}, \quad i = 1, 2, \ldots, n. \quad (8)$$

Solution of the above optimization problem is called the RLTS estimator (RLTSE) of the parameter $\beta$, obtained in the MRM as

$$\hat{\beta}^{RLTS}(k, z) = S(k, z)^{-1} X^\top Zy, \quad (9)$$

where $S(k, z) = S(z) + kI_p$. 

2.1.1. **GCV criterion for selecting the ridge parameter.** As previously mentioned, it is difficult to give a satisfactory answer to the problem of selecting the ridge parameter. Because of good features of the generalized cross validation (GCV) and its simplicity (see [4]), here we use it for RLTSE to choose the optimum value of the ridge parameter \( k_{\text{opt}} \). The GCV criterion for RLTSE can be obtained by

\[
\text{GCV}(k, z) = \frac{\frac{1}{n} \left\| (I_n - H(k, z)) y \right\|^2}{\left( 1 - \frac{1}{n} \text{tr}(H(k, z)) \right)^2},
\]

where \( H(k, z) = X S(k, z)^{-1} X^\top Z \).

The proposed GCV function creates a balance between the precision of the estimators and the biasness caused by the ridge estimation. It behaves like a risk improvement estimator, but does not require an estimate of \( \sigma^2 \) and so, can be used when \( n - p \) is small or even \( n \leq p \) (as in high dimensional problems). According to [4], minimizer of the GCV expectation is essentially equivalent to the minimizer of risk expectation for the estimators.

2.2. **A heuristic optimization problem based on a penalization scheme to combat outliers and multicollinearity problems.** In a recent attempt to overcome the defects of the ridge approach, Roozbeh et al. [30] dealt with a modification on the model (3) by embedding a multiple of \( \kappa(X^\top ZX) \) on the cost function of (3) as follows:

\[
\min_{\beta, z} f(\beta, z) = (y - X\beta)^\top Z(y - X\beta) + \mu \kappa(X^\top ZX)
\]

s.t. \( z^\top e = h \),

\( z_i \in \{0, 1\}, \quad i = 1, 2, \ldots, n, \) \( (11) \)

in which \( \kappa(\cdot) \) stands for the spectral condition number and \( \mu > 0 \) is called the penalty parameter which should be appropriately selected based on the problem scale [35]. The corresponding estimator is called the modified LTS counter multicollinearity (MLTSCM) estimator. In contrast to (3), additional term of the cost function of (11) has been embedded to have control on \( \kappa(X^\top ZX) \). More precisely, if \( \kappa(X^\top ZX) \) is large (or equivalently, \( X^\top ZX \) is an ill-conditioned matrix), then the objective function value in (11) may be decreased by appropriately changing the values of \( z_1, \ldots, z_n \). That is, additional term of the model (11) (i.e. \( \mu \kappa(X^\top ZX) \) with \( \mu > 0 \)) has been considered as a penalty for inappropriate choices of \( z_1, \ldots, z_n \) (i.e. the choices which lead to multicollinearity). As seen, the model (11) presents a quadratic integer programming problem with one linear equality constraint. So, although being NP–hard [12], simple binary structure of the problem makes it possible to effectively find a solution by metaheuristic algorithms, even in large scale cases. Defeating simultaneously the outliers and multicollinearity effects, problem (11) is basis of our optimization models. Nevertheless, it should be noted that existence of the term \( \kappa(X^\top ZX) \) in the objective function makes the model to some extent complicated and increases the computational cost. It is worth noting that in contrast to the classical ridge approaches or the QR–based scheme of [29], the data are not distorted in the model (11).

3. **Two modified optimization models to simultaneously combat outliers and multicollinearity.** Here, we suggest two modified versions of the model (11) by changing the penalty term in the objective function. As an attractive feature,
in the suggested models it is not necessary to compute the condition number which is computationally expensive, while tending to move toward well-conditioning. In the models, effective upper bounds of the spectral condition number are employed.

One of the essential upper bounds of the condition number of a positive definite matrix is introduced by Byrd and Nocedal [14] (based on an analysis conducted on the convergence of the quasi–Newton methods [35]), i.e.

$$\psi(A) = \text{tr}(A) - \ln(\det(A)),$$

where $A \in \mathbb{R}^{n \times n}$ is a positive definite matrix with the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n > 0$. As a worth noting fact, $\psi(A) > 0$, because

$$\psi(A) = \sum_{i=1}^{n} (\lambda_i - \ln \lambda_i).$$

(13)

It can be seen that for all $t > 0$, the function $\omega(t) = t - \ln t$ is a strictly convex function with the minimum value 1 at $t = 1$. So, $\psi(A) \geq n$, and as mentioned in [14], $\psi(A)$ can be considered as a measure of closeness of the matrix $A$ to the identity matrix for which $\psi(I) = n$. Furthermore, since $\omega(t) > \ln t$, we have

$$\psi(A) \geq \ln \lambda_1 - \ln \lambda_n = \ln \frac{\lambda_1}{\lambda_n} = \ln \kappa(A),$$

(14)

which shows that $\psi(A)$ is large when $A$ is an ill-conditioned matrix. Hence, small values of $\psi(A)$ are favorable in practical computations. Based on these preliminaries, the optimization model (11) can be modified as follows:

$$\min_{\beta, z} f(\beta, z) = (y - X\beta)^\top Z(y - X\beta) + M\psi(X^\top ZX)$$

s.t. $z^\top e = h, \quad z_i \in \{0, 1\}, \ i = 1, 2, \ldots, n.$

(15)

in which $M > 0$ is the penalty parameter. The resulting estimator is called the first upper bound–based double modified LTS counter multicollinearity (UB-DMLTSCM1) estimator.

To present the other model, we use the upper bound of the spectral condition number of an arbitrary nonsingular matrix $A \in \mathbb{R}^{n \times n}$ established by Piazza and Politi [25], i.e.

$$\kappa(A) \leq \varphi(A) = \frac{2}{|\det(A)|} \left( \frac{||A||_F}{\sqrt{n}} \right)^n,$$

(16)

in which $||A||_F = \sqrt{\text{tr}(A^\top A)}$ stands for the Frobenius matrix norm [37]. So, we get the following modification of the model (11):

$$\min_{\beta, z} f(\beta, z) = (y - X\beta)^\top Z(y - X\beta) + M\varphi(X^\top ZX)$$

s.t. $z^\top e = h, \quad z_i \in \{0, 1\}, \ i = 1, 2, \ldots, n.$

(17)

The resulting estimator is called the second upper bound–based double modified LTS counter multicollinearity (UBDMLTSCM2) estimator.

To discuss a computational advantage of the models (15) and (17) in contrast to the model (11), we state the following result.

**Proposition 1.** Computational costs of the models (15) and (17) is less than computational costs of the model (11).
Proof. Assume that $A = [a_{ij}] \in \mathbb{R}^{n \times n}$ is a positive definite matrix with the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n > 0$. It is enough to show that the flop counts of $\text{tr}(A)$, $||A||_F$, and $\text{det}(A)$ is less than the flop count of $\kappa(A)$.

Note that $\kappa(A) = \frac{\lambda_1}{\lambda_n}$. As known, there is no general formula for the eigenvalues of an $n \times n$ matrix if $n > 4$ [37, page 310]. So, using the popular (iterative) QR algorithm on a tridiagonal reduction of $A$ (achieved by orthogonal similarity transformations) to estimate the eigenvalues, $\kappa(A)$ costs at least $\frac{4}{3}n^3 + O(n^2)$ flops [37, page 353]. On the other hand, the flop counts of $\text{tr}(A) = \sum_{i=1}^{n} a_{ii}$ and $||A||_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2}$ are respectively of the form $O(n)$ and $O(n^2)$. Also, to compute $\text{det}(A)$ we can perform $\frac{1}{3}n^3 + O(n^2)$ flops to get the Cholesky decomposition of $A$ as $A = R^T R$ [37, Page 40] where $R \in \mathbb{R}^{n \times n}$ is an upper triangular matrix whose determinant is equal to product of its diagonal elements. So, $\text{det}(A) = \text{det}(R)^2$ costs $\frac{1}{3}n^3 + O(n^2)$ flops. Hence, the proof is complete.

As seen, the optimization models (15) and (17) illustrate NP–hard mixed–integer nonlinear programming problems for which the classical methods are not practically efficient [12]. As known, metaheuristic algorithms have attracted special attention in developing efficiently robust computational procedures for solving a vast variety of such problems [38, 24]. These nature–inspired methods are so popular since their softwares can be flexibly reused and also, they can efficiently solve complicated problems even in large scale cases [30, 8, 38].

Among the earliest and most popular metaheuristic techniques of the optimization, there is the simulated annealing (SA) algorithm. The method origins from the successful annealing process of the materials which involves the cautious control of the cooling schedule [38]. SA is a local search algorithm capable of escaping from local optima by use of random hill–climbing moves in the search process [12, 17]. It is very efficient in practice [17, 27] and well–developed in theory [7, 1]. Motivated by these, here we use the SA algorithm to approximately compute UBDMLTSCM1 and UBDMLTSCM2.

4. Numerical experiments. In this section, we perform some numerical experiments to support our assertions, by investigating effectiveness of the proposed estimation methods into three real world data sets: a bridge construction data, an electricity consumption data and a larger data set of determinants of wages. Here, we implemented the SA algorithm in MATLAB software environment; the search procedure was stopped after 10000 iterations. Also, we set $\mu = 10$ in (11) and $M = 1$ in (15) and (17). Other computations were performed by the statistical software R.

4.1. Application to bridge construction data. Here, information from 45 bridge projects provided in [36, page 130] is analyzed. (The data can be also found in [34].) The dependent and independent variables are given as follows:

$Time$: Design time (person–days);
Based on the results of [30], there exists a linear relationship between the logarithmic transformation of response and the predictor variables and moreover, the predictors $\log(DArea)$ and $\log(Length)$ have statistically insignificant linear relationship with the response variable. So, specification of the final model is

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

(18)

Below, correlations between the predictors in model (18) is given, showing that all of them are considerable:

|          | log(CCost) | log(Dwgs) | log(Spans) |
|----------|------------|-----------|------------|
| log(CCost) | 1.0000     | 0.8315    | 0.7751     |
| log(Dwgs)  | 0.8315     | 1.0000    | 0.6297     |
| log(Spans) | 0.7751     | 0.6297    | 1.0000     |

For the model (18), condition number of the new design matrix is equal to 2258.34. This means that there exists a strong collinearity between the transformed predictors. Figure 1 gives the diagnostics plots to identify outliers for the model (18). We label the observations as outliers if the standardized residual for the point falls outside the interval $[-2, 2]$. The optimal value of $k (k_{opt})$ for the ridge estimators is obtained using the GCV score (see Figure 2).

Table 1 shows summary of the results. In this table, $SSE = \min_{\beta, \sigma} f(\beta, \sigma)$ and $R^2 = 1 - \frac{SSE}{S_{yy}}$ are respectively the residual sum of squares and the coefficient of determination of the model (as a measure for goodness of fit), i.e. $S_{yy} = \sum_{i=1}^{n} z_i (y_i - \bar{y})^2$ and $SSE = \sum_{i=1}^{n} z_i (y_i - \hat{y})^2$, with $\hat{y}_i = x_i^T \hat{\beta}$ where $x_i = (1, \log(CCost)_i, \log(Dwgs)_i, \log(Spans)_i)$. To compare the proposed methods with
the modern day non–parametric data science approaches, the linear SVR (LSVR) [22], the non–parametric SVR (NSVR) and NN regression (NNR) estimators are also reported in Table 1. As seen, UBDMLTSM1 is quite efficient in the sense that it has a significant value of goodness of fit. Also, the performance of UBDMLTSM2 is satisfactory.

4.2. Application to electricity consumption data. In order to more clearly compare performance of the proposed estimators, we use another real data set related to the determinants of electricity consumption in Germany, firstly considered by Akdeniz Duran et. al [2]. The data set are accessible at http://www.quantlet.org. Electricity consumption is known to be influenced negatively by the electricity price
and positively by the income of the consumers. The response variable and the predictors of the problem are given as follows:

- **LEC**: Log monthly electricity consumption per person;
- **LI**: Log income per person;
- **LREG**: Log rate of electricity price to the gas price;
- **Temp**: Cumulated average temperature index;

The data have been taken as the average of 20 German cities for 177 cases. So, the MRM can be considered as

\[
(LEC)_i = \beta_0 + \sum_{j=1}^{11} \beta_j x_{ij} + \beta_{12} (LI)_i + \beta_{13} (LREG)_i + \beta_{14} (Temp)_i + \epsilon_i, \tag{19}
\]

where \(x_1, \ldots, x_{11}\) are dummy variables for the monthly effects. Based on the adjusted R–squared and AIC criteria given in Table 2, the final MRM can be specified as

\[
(LEC)_i = \beta_0 + \beta_{12} (LI)_i + \beta_{13} (LREG)_i + \beta_{14} (Temp)_i + \epsilon_i. \tag{20}
\]

Condition number of the new design matrix in the model (20) is approximately \(\lambda_4/\lambda_1 = 18404279.00\), and so, there exists a very strong multicollinearity between the predictors. Figure 3 gives the diagnostics plots to identify outliers for the model (20). We calculated the optimal values of ridge parameter for RLTSE using GCV (see Figure 4). The results are summarized in Table 3. They show that UB-DMLTSCM1 is the most efficient estimator in contrast to the others. Performance of UB-DMLTSCM2 is also reasonable.

#### 4.3. Application to determinants of wages data.

To illustrate usefulness of the suggested strategies for a larger data in the regression model, we consider the determinants of wages data set (CPS 1985) which can be found in R package “AER” [11]. The Current Population Survey (CPS) is used to supplement census information between census years. These data consist of a random sample of 534 persons from the CPS, with information on wages and other characteristics of the workers,
TABLE 2. The most effective subgroup of predictor variables based on the $R_{adj}^2$ and AIC criteria for the electricity data set

| Subset size | Predictor variables                      | $R_{adj}^2$ | 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, $x_9$                  | 0.5891      | -1080.057    |
| 5           | Temp, LREG, LI, $x_{10}$               | 0.5882      | -1078.709    |
| 6           | Temp, LREG, LI, $x_9$, $x_{10}$, $x_{11}$ | 0.5875      | -1077.427    |
| 7           | Temp, LREG, LI, $x_9$, $x_{10}$, $x_{11}$, $x_1$ | 0.5858      | -1075.734    |
| 8           | Temp, LREG, LI, $x_9$, $x_{10}$, $x_{11}$, $x_1$, $x_3$ | 0.5873      | -1073.897    |
| 9           | Temp, LREG, LI, $x_9$, $x_{10}$, $x_{11}$, $x_1$, $x_3$, $x_5$ | 0.5812      | -1071.907    |
| 10          | Temp, LREG, LI, $x_9$, $x_{10}$, $x_{11}$, $x_1$, $x_3$, $x_5$, $x_4$ | 0.5789      | -1069.987    |
| 11          | Temp, LREG, LI, $x_9$, $x_{10}$, $x_{11}$, $x_1$, $x_3$, $x_5$, $x_4$, $x_7$, $x_2$ | 0.5764      | -1067.997    |
| 12          | Temp, LREG, LI, $x_9$, $x_{10}$, $x_{11}$, $x_1$, $x_3$, $x_5$, $x_4$, $x_7$, $x_2$, $x_6$ | 0.5740      | -1064.098    |
| 13          | Temp, LREG, LI, $x_9$, $x_{10}$, $x_{11}$, $x_1$, $x_3$, $x_5$, $x_4$, $x_7$, $x_2$, $x_6$, $x_8$ | 0.5709      | -1063.014    |

Figure 4. The diagram of GCV($k, z$) versus the ridge parameter for the electricity data

including sex, number of years of education, years of work experience, occupational status, region of residence and union membership. We wish to determine whether wages are related to these characteristics. A data frame containing 534 observations on 11 variables is described as follows:

- **wage**: Wage (in dollars per hour);
- **education**: Number of years of education;
- **south**: Indicator variable for Southern Region (1 = person lives in south and 0 = person lives elsewhere);
- **sex**: Indicator variable for sex (1 = female and 0 = male);
Table 3. Evaluation of the proposed estimators for the electricity data set

| Method   | OLS   | RLTS  | MLTSCM | UBDMLTSCM1 |
|----------|-------|-------|--------|-------------|
| Intercept| 4.4069| 5.1693| 4.9881 | 5.2039      |
| LI       | 0.1925| 0.0989| 0.1146 | 0.0956      |
| LREG     | -0.0778| -0.0939| -0.1054| -0.0956    |
| Temp     | -0.0002| -0.0002| -0.0003| -0.0003    |
| SSE      | 0.3765| 0.2637| 0.1982 | 0.1296      |
| $R^2$    | 0.5962| 0.6742| 0.7399 | 0.7559      |

| Method   | UBDMLTSCM2 | LSVR | NSVR | NNR |
|----------|------------|------|------|-----|
| Intercept| 4.0907     | 0.0881| -2.6215| - |
| LI       | 0.2225     | 0.1545| -1.2806| - |
| LREG     | -0.0940    | -0.1322| -3.7418| - |
| Temp     | -0.0003    | -0.7508| -0.8067| - |
| SSE      | 0.1413     | 0.3881| 0.2629| 0.4240 |
| $R^2$    | 0.7468     | 0.5838| 0.7181| 0.5452 |

Figure 5. The diagnostic plots for the model (21)

experience: Number of years of potential work experience;
union: Indicator variable for union membership (1 = union member and 0 = not union member);
age: Age in years;
race: Race (1 = other, 2 = hispanic and 3 = white);
occupation: Occupational category (1 = management, 2 = sales, 3 = clerical, 4 = service, 5 = professional and 6 = other);
sector: Sector (0 = other, 1 = manufacturing and 2 = construction);
marrried: Marital status (0 = unmarried and 1 = married).
Based on residual plots, wages were log–transformed to stabilize the variance.
For detecting the multicollinearity in design matrix, the “mctest” package in R is used which leads to the following output. It provides the Farrar-Glauber test and
FIGURE 6. The diagram of GCV$(k, z)$ versus the ridge parameter for the CPS data

other relevant tests for multicollinearity.

Overall Multicollinearity Diagnostics

| MC Results | detection |
|-------------|------------|
| Determinant $|X'X|$: | 0.0001 | 1 |
| Farrar Chi-Square: | 4818.3895 | 1 |
| Red Indicator: | 0.1983 | 0 |
| Sum of Lambda Inverse: | 10068.8439 | 1 |
| Theil's Method: | 1.2263 | 1 |
| Condition Number: | 739.7337 | 1 |

1 -- > COLLINEARITY is detected by the test
0 -- > COLLINEARITY is not detected by the test

As it can be seen, there exists a strong multicollinearity between the predictors. Figure 5 gives the diagnostics plots to identify outliers for the following model:

$$\log(WAGE) = \beta_0 + \beta_1 \text{education} + \beta_2 \text{south} + \beta_3 \text{sex} + \beta_4 \text{experience} + \beta_5 \text{union} + \beta_6 \text{age} + \beta_7 \text{race} + \beta_8 \text{occupation} + \beta_9 \text{sector} + \beta_{10} \text{married} + \epsilon.$$  \hspace{1cm} (21)

As a first step to data visualization, we employ diagnostic plots to identify data that are 'inconsistent' with the main bulk of the data sets. We label an observations as 'outlier' if the absolute standardized residual for the point is larger than 2; observations 8, 15, 105, 166, 247, 277, 339, 354, 362, 379, 410, 414, 416, 434 and 450 are outliers as shown in Figure 5. The presence of these outliers will inevitably affect the parameter estimation in the model. Thus, developing an efficient robust estimation strategy is necessary.

We calculated the optimal values of ridge parameter for RLTSE using GCV (see Figure 6). The results are summarized in Table 4. They show that UBDMLTSCM2 and NSVR are the most efficient estimators in contrast to the others based on the SSE and $R^2$ respectively.

5. Conclusions. In this study, we have proposed two mixed–integer nonlinear constrained optimization problems as multiple regression models to handle multicollinearity problem while making the estimators resistant against the outlying observations and non–normal error distributions. Stabilization of the data and freedom from choosing the shrinkage parameter by running a complicated algorithm are the invaluable merits of these techniques
in comparison with the classical methods such as ridge estimation. Because of some draw-
backs of the ridge approach, we have modified optimization problem of the robust ridge 
estimator using a penalization scheme based on upper bounds of the spectral condition 
number which are computationally advantageous in the perspective of the flop count. We 
have also studied three real–world applications related to the bridge projects data set, 
the electricity data and the wages data to compare performance of the proposed estima-
tors with some classical ones. The results demonstrate that both the proposed estimators 
(UBDMLTSCM1 and UBDMLTSCM2) have satisfactory performances in the existence of 
outliers and multicollinearity.

Acknowledgments. The authors thank the Research Council of Semnan University for 
it support. They are also grateful to the Associate Editor and the anonymous reviewers for 
their valuable comments and suggestions helped to improve the quality of this work.

REFERENCES

[1] E. H. L. Aarts, J. H. M. Korst and P. J. M. van Laarhoren, Simulated annealing, in Local
Search in Combinatorial Optimization, Wiley-Intersci. Ser. Discrete Math. Optim., Wiley-
Intersci. Publ., Wiley, Chichester, 1997, 91–121.
[2] E. Akdeniz Duran, W. K. Härdle and M. Osipenko, Difference based ridge and Liu type estimators in semiparametric regression models, *J. Multivariate Anal.*, 105 (2012), 164–175.

[3] F. Akdeniz and M. Roozbeh, Generalized difference-based weighted mixed almost unbiased ridge estimator in partially linear models, *Statist. Papers*, 60 (2019), 1717–1739.

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

[5] M. Arashi and T. Valizadeh, Performance of Kibria’s methods in partial linear ridge regression model, *Statist. Pap.*, 56 (2015), 231–246.

[6] M. Awad and R. Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, Apress, Berkeley, CA, 2015.

[7] S. Babaie-Kafaki, R. Ghanbari and N. Mahdavi-Amiri, An efficient and practically robust hybrid metaheuristic algorithm for solving fuzzy bus terminal location problems, *Asia–Pac. J. Oper. Res.*, 29 (2012), 1–25.

[8] S. Babaie-Kafaki, R. Ghanbari and N. Mahdavi-Amiri, Hybridizations of genetic algorithms and neighborhood search metaheuristics for fuzzy bus terminal location problems, *Appl. Soft Comput.*, 46 (2016), 220–229.

[9] S. Babaie-Kafaki and M. Roozbeh, A revised Cholesky decomposition to combat multicollinearity in multiple regression models, *J. Stat. Comput. Simul.*, 87 (2017), 2298–2308.

[10] M. R. Baye and D. F. Parker, Combining ridge and principal component regression: A money demand illustration, *Comm. Statist. A—Theory Methods*, 13 (1984), 197–205.

[11] E. R. Berndt, *The Practice of Econometrics*, New York, Addison-Wesley, 1991.

[12] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, Athena Scientific, Massachusetts, 1997.

[13] P. Bühlmann, M. Kalisch and L. Meier, High–dimensional statistics with a view towards applications in biology, *Ann. Rev. Stat. Appl.*, 1 (2014), 205–278.

[14] R. H. Byrd and J. Nocedal, A tool for the analysis of quasi–Newton methods with application to unconstrained minimization, *SIAM J. Numer. Anal.*, 26 (1989), 727–739.

[15] M. Hassanzadeh Bashtian, M. Arashi and S. M. M. Tabatabaey, Using improved estimation strategies to combat multicollinearity, *Comm. Statist. Theory Methods*, 30 (2001), 2699–2705.

[16] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, Berlin, Heidelberg, 2013.

[17] D. Henderson, S. H. Jacobson and A. W. Johnson, The theory and practice of simulated annealing, in *Handbook of Metaheuristics*, Kluwer Academic Publishers, Boston, MA, (2003), 287–319.

[18] A. E. Hoerl and R. W. Kennard, Ridge regression: Biased estimation for non–orthogonal problems, *Technometrics*, 12 (1970), 55–67.

[19] W. M. Pride and O. C. Ferrel, *Marketing*, 15th edition, South-Western, Cengage Learning, International Edition, 2010.

[20] C. R. Reeves, Modern heuristic techniques, in *Modern Heuristic Search Methods*, John Wiley and Sons, Chichester, (1996), 1–24.

[21] M. Roozbeh, Optimal QR-based estimation in partially linear regression models with correlated errors using GCV criterion, *Computational Statistics & Data Analysis*, 117 (2018), 45–61.
[29] M. Roozbeh, S. Babaie-Kafaki and M. Arashi, A class of biased estimators based on QR decomposition, *Linear Algebra Appl.*, **508** (2016), 190–205.

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

[31] M. Roozbeh, Robust ridge estimator in restricted semiparametric regression models, *J. Multivariate Anal.*, **147** (2016), 127–144.

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

[33] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, John Wiley and Sons, New York, 1987.

[34] S. J. Sheather, *A Modern Approach to Regression with R*, Springer, New York, 2009.

[35] W. Sun and Y. X. Yuan, *Optimization Theory and Methods: Nonlinear Programming*, Springer, New York, 2006.

[36] P. Tryfos, *Methods for Business Analysis and Forecasting: Text & Cases*, John Wiley and Sons, New York, 1998.

[37] D. S. Watkins, *Fundamentals of Matrix Computations*, 2nd edition, John Wiley and Sons, New York, 2002.

[38] X. S. Yang, *Nature–Inspired Optimization Algorithms*, Elsevier, Amsterdam, 2014.

Received September 2019; 1st revision May 2020.

E-mail address: mahdi.roozbeh@semnan.ac.ir
E-mail address: sbk@semnan.ac.ir
E-mail address: aminisor@semnan.ac.ir