Original Article

Robust High-Dimensional Semiparametric Regression using Optimized Differencing Method Applied to Vitamin B2 Production Data

Mahdi Roozbeh[1*] Monireh Maanavi[2] Saman Babaie-Kafaki[3]

1. Associate Professor, Faculty of Mathematics, Statistics & Computer Science, Semnan University, Semnan, Iran
2. MSc, Faculty of Mathematics, Statistics and Computer Science, Semnan University, Semnan, Iran
3. Professor, Faculty of Mathematics, Statistics & Computer Science, Semnan University, Semnan, Iran

*__Correspondence to__: Mahdi Roozbeh
mahdi.roozbeh@semnan.ac.ir

(Received: 2 Jan. 2020; Revised: 6 Apr. 2020; Accepted: 19 May. 2020)

Abstract

***Background and Purpose*:** By evolving science, knowledge, and technology, we deal with high-dimensional data in which the number of predictors may considerably exceed the sample size. The main problems with high-dimensional data are the estimation of the coefficients and interpretation. For high-dimension problems, classical methods are not reliable because of a large number of predictor variables. In addition, classical methods are affected by the presence of outliers and collinearity.

Methods: Nowadays, many real-world data sets carry structures of high-dimensional and outliers problems. In the regression concept, an outlier is a point that fails to follow the main linear pattern of the data. The ordinary least-squares estimator is potentially sensitive to the outliers; this fact provided necessary motivations to investigate robust estimations. To handle these problems, we combined the least absolute shrinkage and selection operator (LASSO) with the least trimmed squares (LTS) estimation.

Results: Due to the flexibility and applicability of the semiparametric model in medical data, a penalized optimization approach for semiparametric regression models to simultaneously combat high-dimension and outliers in the data set. Based on the numerical study, it was deduced that the proposed model is quite efficient in the sense that it has a significant value of goodness of fit (MSE=1.3807).

Conclusion: We have proposed an optimization approach for semiparametric models to combat outliers in the data set. Especially, based on a penalization LASSO scheme, we have suggested a nonlinear integer programming problem as the semiparametric model which can be effectively solved by any evolutionary algorithm. We have also studied a real-world application related to the riboflavin production.

Keywords: High-Dimensional Data Set; Ordinary Least Square Method; Outliers; Robust Regression

Citation: **Roozbeh M*,** Maanavi M, Babaie-Kafaki S. Robust High-Dimensional Semiparametric Regression using Optimized Differencing Method Applied to Vitamin B2 Production Data. Iran J Health Sci. 2020; 8 (2): 9-22.

1. Introduction

The linear regression model can be shown as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \qquad (1)$$

Where $\boldsymbol{y} = (y_1, \dots, y_n)^{\mathrm{T}}$ is an $n \times 1$ vector of observations on the dependent variable, $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_p)$ is an $n \times p$ matrix of observations on the explanatory variables such that $\boldsymbol{x}_i = (x_{1i}, \dots, x_{ni})^{\mathrm{T}}$, $i = 1, \dots, p$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^{\mathrm{T}}$ is a $p \times 1$ vector of unknown regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^{\mathrm{T}}$ is an $n \times 1$ vector of the error terms with $E(\boldsymbol{\varepsilon}) = 0$ and $E(\boldsymbol{\varepsilon\varepsilon}^{\mathrm{T}}) = \sigma^2 \boldsymbol{I}_p$.

Outliers (i.e. points that fail to follow a partial linear pattern of the majority of the points) are a common problem in using the ordinary least squares (OLS) method. In such situations, robust regression methods are used to overcome undesirable effects of the outliers (inflated sum of squares, bias or distortion of estimation, distortion of p-values, etc.). Outliers may be observed because of a recording error, a disruption in production processes, human errors, or may be formed differently from the large portion of the data. They may cause the wrong model formations, wrong parameter estimations or erroneous analysis results (1). The importance of these points were to the extent that the researchers made various definitions, which will be discussed in some of these terms:

- An outlier is an observation that deviates so much from other observations as to arouse suspicions that is generated by a different mechanism (2).
- An outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of the data set (3).
- The outlier is viewed as an observation whose value is in the pattern generated by the other data (4).
- Outliers are observations that do not follow the pattern of the majority of the data (5).
- An outlier is an observation that lies outside the overall pattern of a distribution (6).
- To be an influential one, a point should cause a dramatic change in the model after its deletion (7).

A point $(\boldsymbol{x_i}, y_i)$ which does not follow the linear pattern of the majority of the data but whose $\boldsymbol{x}_i$ is not outlying is called a vertical outlier. A point $(\boldsymbol{x_i}, y_i)$ whose $\boldsymbol{x}_i$ is outlying is called a leverage point. We say that it is a good leverage point when $(\boldsymbol{x_i}, y_i)$ follows the pattern of the majority, and a bad leverage point, otherwise. After summarizing, a data set can contain four types of points: regular observations, vertical outliers, good leverage points, and bad leverage points. Of course, most data sets do not have all the four types (8). Figure 1 shows the types of points.

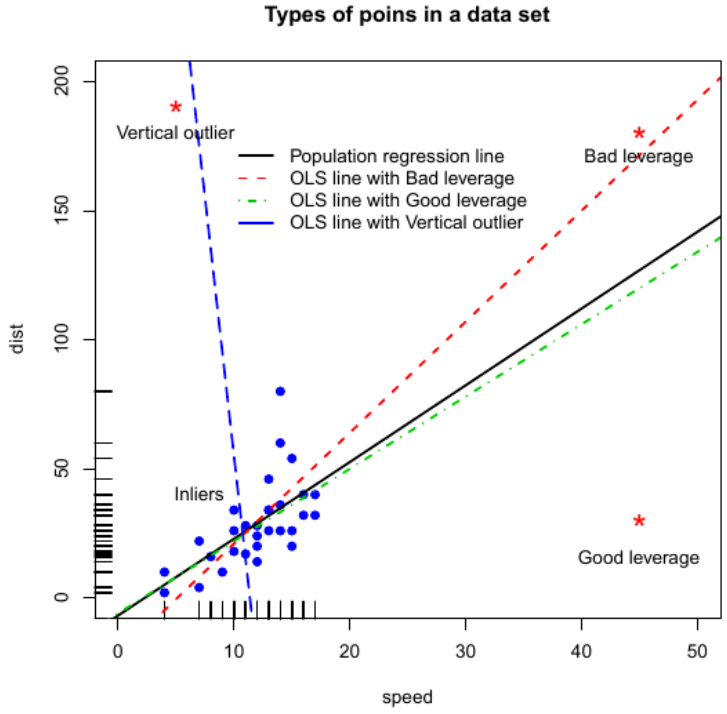

Figure 1. Types of points

Semiparametric models have many applications. Engle et al. (1986) were among the first researchers who considered the semiparametric model to analyze the relationship between temperature and electricity usage, and found them to have highly nonlinear relationship. Yatchew (1997) estimated the relationship between variable costs of distributing electricity per customer as a nonlinear function of the scale of operation as measured by the number of customers.

The course of high-dimensionality is another common problem in the modern statistical methods. In statistical theory, the field of high-dimensional statistics studies data whose dimension is larger than dimensions considered in classical multivariate analysis. As mentioned before, big data has been one of the hottest topics in computer science, data mining, engineering, and applied mathematics. In fact, various research activities surrounding the big data are so vast that they form a new discipline, namely, data science.

There are many challenging issues associated with big data (12,13), and among them, the most important issue is the high-dimensional data analysis. Even with some moderate size data, high-dimensionality can pose extra challenges. High-dimensional data are relevant to a wide range of fields, such as biometric, medicine, e-commerce, network security, and industrial applications. In order to use data characteristics, proper techniques and methods are needed to handle such high-dimensional data (14).

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 (1996), 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. LASSO regression is a simple technique to reduce model complexity and prevent over-fitting which may result from simple linear regression. Ordinary least squares regression chooses the coefficients by minimizing the residual sum of squares (RSS), which is the difference between the observed and the estimated values, as follows:

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

LASSO is an extension of OLS which adds a penalty to the RSS equal to the sum of the absolute values of the non-intercept beta coefficients multiplied by parameter λ that slows or accelerates the penalty. That is, if λ is less than 1, then it slows the penalty, while if it is more than 1, it accelerates the penalty. Therefore, the following optimization problem should be solved:

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

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 deals with the bias-variance tradeoff. Bias refers to how correct (or incorrect) the model is.

There are several ways to choose the optimal λ, such as AIC, BIC, C_p and so on. For this purpose, one of the most popular methods is the cross-validation (CV) method.

In order to find the optimal value of λ, a range of λ values are tested and the optimal value is chosen using cross-validation. Cross-validation 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 MSE in the test set.

Rousseeuw (1984) introduced several robust regression estimators including least median of squares (LMS) and least trimmed squares (LTS) (see also the monograph (17,18)). LTS converges at rate $n^{\frac{1}{2}}$ with the same assymptotic efficiency under normality as Huber's skip estimator. The LMS convergence rate is $n^{\frac{1}{3}}$ and its objective function is less smooth than LTS. As a consequence, as argued in (8), LTS is now preferred over LMS.

LTS estimator is the solution of the following optimization problem:

$$\min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{h} e_{(i)}^2\right\}, \qquad s.t. \qquad e_i = y_i - \hat{y}_i, \qquad (2)$$

Based on the ordered absolute residuals $|e_{(1)}| \leq \cdots \leq |e_{(n)}|$, where h is the trimmed parameter. When $h = [\frac{n}{2}]$, the LTS estimator locates that half of the observations which has the smallest estimated variance. In that case, the breakdown point is 50%. When h is set to the sample size, LTS and OLS coincide, i.e. $\widehat{\boldsymbol{\beta}}_{OLS} = \widehat{\boldsymbol{\beta}}_{LTS}$.

As discussed earlier, LASSO offers interpretable models, but it is not robust with respect to the outliers. The breakdown point of the LASSO is $\frac{1}{n}$, that is, only one

single outlier can make the LASSO estimator completely unreliable. Therefore, robust alternatives are needed. In this situation, Alfons et al. (2013) suggested the sparse LTS estimator as follows:

$$\min_{\boldsymbol{\beta},\boldsymbol{z}}\{\phi(\boldsymbol{\beta},\boldsymbol{z})\} = \min_{\boldsymbol{\beta},\boldsymbol{z}}\{(\boldsymbol{y}-\boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X\beta}) + h\lambda p(\boldsymbol{\beta})\},$$

$$s.t. \quad \boldsymbol{e}^{\mathrm{T}}z = h,$$

$$z_i \in \{0,1\}, \quad i = 1,\dots,n,$$

where

$$\boldsymbol{Z} = \begin{bmatrix} z_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & z_n \end{bmatrix}, \boldsymbol{z} = \begin{bmatrix} z_1 \\ \vdots \\ \boldsymbol{z_n} \end{bmatrix}, \boldsymbol{e} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},$$

and λ is a penalty parameter. Alfons et al. (2013) show that the breakdown point of this estimator is $\frac{n-h+1}{n}$.

Semiparametric regression models are appropriate models when a suitable link function of the mean response is assumed to have a linear parametric relationship to some explanatory variables, while its relationship to the other variables has an unknown form. Let $(y_1, \boldsymbol{x}_1^{\mathrm{T}}, t_1), \dots, (y_n, \boldsymbol{x}_n^{\mathrm{T}}, t_n)$ be the observations that follow the semiparametric regression model, that is:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + f(\boldsymbol{t}) + \varepsilon_i, \qquad i = 1,\dots,n, \qquad (3)$$

where $\boldsymbol{x}_i = (x_{1i},\dots,x_{pi})^{\mathrm{T}}$ is a vector of explanatory variables, $\boldsymbol{\beta} = (\beta_1,\dots,\beta_p)^{\mathrm{T}}$ is an unknown p-dimensional vector parameter, t_i's are design points which belong to some bounded domain $D \in \mathbb{R}$, $f(\boldsymbol{t})$ is an unknown smooth function and ε_i's are random errors, assumed to be independent of $(\boldsymbol{x}_i, t_i)$. Surveys regarding the estimation and application of the model (3) can be found in the monograph of Härdle et al. (2000). Speckman (1988) studied partial residual estimation of $\boldsymbol{\beta}$ and $f(\boldsymbol{t})$ in (3), and obtained asymptotic bias and variance of the estimators. Roozbeh (2016) developed robust statistical inference for the model (3) for both heteroscedastic and correlated errors under general assumption $E(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{V}$. For bandwidth selection in the context of kernel-based estimation in model (3), generalized cross-validation criterion has been used for optimal bandwidth selection (23-25). The difference-based estimation approach is optimal in the sense that the estimator of the linear component is asymptotically efficient and the estimator of the nonparametric component is asymptotically minimax rate optimal (26). Hall et al. (1990) extended the idea to higher-order differences for efficient estimation of the variance in such a setting.

In fact, in the differencing method, the differencing matrix $\boldsymbol{D}_{(n-m)\times n}$ is multiplied to both sides of the model (3). That means:

$$\boldsymbol{Dy} = \boldsymbol{DX\beta} + \boldsymbol{D}f(\boldsymbol{t}) + \boldsymbol{D\varepsilon},$$

So that

$$\boldsymbol{D} = \begin{bmatrix} d_0 & \cdots & d_m & 0 & \cdots & 0 & 0 \\ 0 & d_0 & \cdots & d_m & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & d_0 & d_1 & \cdots & d_m \end{bmatrix} \text{ with } \sum_{j=0}^{m} d_j = 0 \text{ and } \sum_{j=0}^{m} d_j^2 = 1.$$

Hall et al. (27) computed the optimal differencing coefficients numerically for $m \leq 10$, as reported in Table 1.

Table 1. Optimal differencing coefficients numerically up to 10

Order / Weight	1	2	3	4	5	6	7	8	9	10
d_0	0.7071	0.8090	0.8090	0.8873	0.9200	0.9200	0.9302	0.9380	0.9443	**0.9494**
d_1	-0.7071	-0.500	-0.3832	-0.3099	-0.2600	-0.2238	-0.1965	-0.1751	-0.1587	**-0.1437**
d_2		-0.309	-0.2809	-0.2464	-0.2197	-0.1925	-0.1728	-0.1389	-0.1439	**-0.1314**
d_3			-0.1942	-0.1901	-0.1774	-0.1635	-0.1506	-0.1224	-0.1287	**-0.1197**
d_4				0.1409	-0.1420	-0.1369	-0.1299	-0.1069	-0.1152	**-0.1085**
d_5					-0.1103	-0.0112	-0.1107	-0.0925	-0.1025	**-00978**
d_6						-0.0906	-0.0930	-0.0791	-0.0905	**-0.0877**
d_7							-0.0765	-0.0666	-0.0792	**-0.0782**
d_8									-0.0687	**-0.0691**
d_9									-0.0588	**-0.0606**
d_{10}										**-0.0527**

2. Methods and Result

To illustrate the usefulness of the suggested strategies for high dimensional data in the semiparametric regression model, we consider the data set of riboflavin (also known as vitamin B2) production in Bacillus subtilis which can be found in R package "hdi" (28). 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). These coenzymes play major roles in energy production, cellular function, growth, and development, and metabolism of fats, drugs, and steroids. The conversion of the amino acid tryptophan to niacin (sometimes referred to as vitamin B3) requires FAD. Similarly, the conversion of vitamin B6 to the coenzyme pyridoxal 5-phosphate needs FMN. In addition, riboflavin helps to maintain normal levels of homocysteine, an amino acid in the blood. More than 90% of dietary riboflavin is in the form of FAD or FMN; the remaining 10% is comprised of the free form and glycosides or esters. Most riboflavin is absorbed in the proximal small intestine.

The body absorbs little riboflavin from single doses beyond 27 mg and stores only small amounts of riboflavin in the liver, heart, and kidneys. When excess amounts are consumed, they are either not absorbed or the small amount, that is absorbed, is excreted in the urine. Bacteria in the large intestine produce free riboflavin that can be absorbed by the large intestine in amounts that depend on the diet. More riboflavin is produced after ingestion of vegetable-based than meat-based foods. Riboflavin status is not routinely measured in healthy people. The current EARs for riboflavin for women and men aged 14 and older are 0.9 mg/day and 1.1 mg/day, respectively, and the RDAs are 1.1 and 1.3 mg/day, respectively. RDAs are higher than EARs so as to identify amounts that will cover people with higher than average requirements. RDA for pregnancy is 1.4 mg/day and for lactation is 1.6 mg/day. For infants up to 12 months, the Adequate Intake (AI) is 0.3-0.4 mg/day. A stable and sensitive measure of riboflavin deficiency is the erythrocyte glutathione reductase activity coefficient (EGRAC) which is based on the ratio between this

enzyme's in-vitro activities in the presence of FAD to that without added FAD.

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 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. To analyze this data, we first used the LASSO Method to specify a sparse model. To determine the penalty parameter, a 10-fold cross-validation method was used with 70% data as a training set and 30% as a test set.

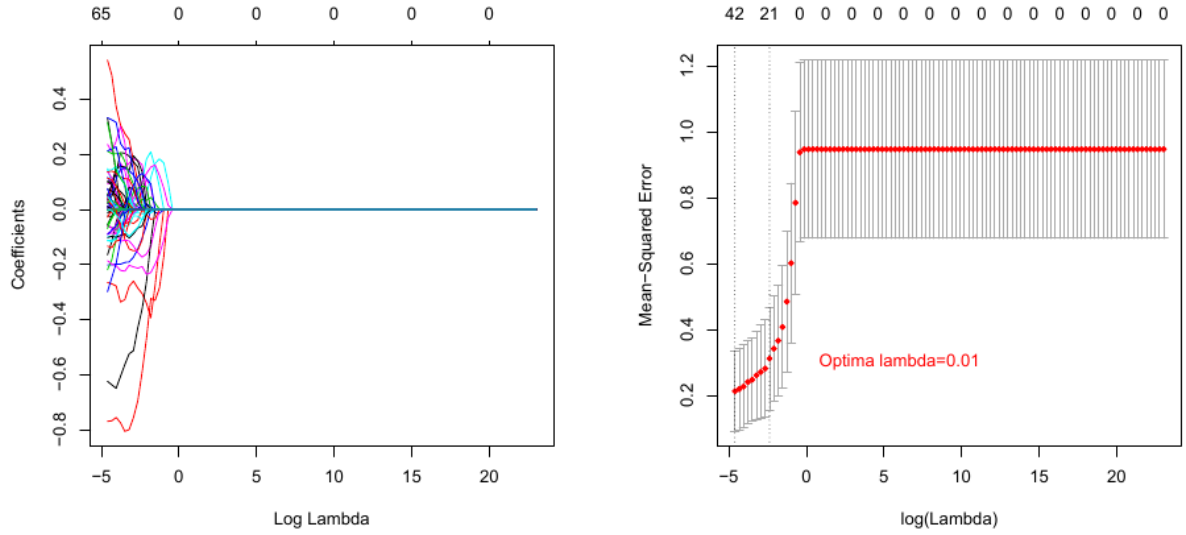

Figure 2. The LASSO Plots

In Figure 2, the 10-fold cross-validation and the coefficients estimation diagrams for different values of the penalty parameter are depicted. The LASSO Method selected 64 variables. To detect the nonparametric part of the model we calculated:

$$s_i^2 = \frac{1}{n - p_1 - 1}(\boldsymbol{y} - \boldsymbol{X}_{new}[,-i]\ \widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}_{new}[,-i]\ \widehat{\boldsymbol{\beta}}), \quad i = 1, \ldots, 64, \quad (4)$$

where $\boldsymbol{X}_{new}[,-i]$ was 6 obtained by deleting the i^{th} column of the matrix $\boldsymbol{X}$ (29,30). Among all 64 remained genes, "PURR-at" had minimum s^2 value, and so, it could be considered as a nonparametric part. We also used the added-variable plots to identify the parametric and nonparametric components of the model. Added-variable plots enabled us to visually assess the effect of each predictor, having adjusted for the effects of the other predictors. By looking at the added-variable plot shown in Figure 3, we considered "PURR-at" as a nonparametric part. The regression model is:

$$\boldsymbol{y} = \boldsymbol{X}_{new}\boldsymbol{\beta} + \mathrm{f}(\mathrm{PURR} - \mathrm{at}) + \boldsymbol{\varepsilon}, \quad (4)$$

where $\boldsymbol{X}_{new}$ is an $n \times (64 - 1)$ design matrix without "PURR-at" variable, with the explanatory variables, such that $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{ni})^{\mathrm{T}}$, $i = 1, \ldots, p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{(64-1)})^{\mathrm{T}}$ is a $p \times 1$ vector of unknown regression coefficients, and

$\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^{\mathrm{T}}$ is an $n \times 1$ vector of error terms with $E(\boldsymbol{\varepsilon}) = 0$ and $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}) = \sigma^2 \boldsymbol{I}_p$.

The differencing method was then used to separate the identified nonparametric part. In this example, the third-order differencing matrix was used. So, the specification of the new model is as follows:

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}, \qquad (5)$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{D}\boldsymbol{y}$, $\tilde{\boldsymbol{X}} = \boldsymbol{D}\boldsymbol{X}$ and $\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{D}\boldsymbol{\varepsilon}$.

In this section, we identified outliers using the diagrams in Figures 4 and 5. It was clear that there were outliers in the data set. Therefore, a robust method should be used to estimate the coefficients. Table 2 reports the estimation of the coefficients by the LTS Method. The mean squared error (MSE) value for this model was documented to be equal to 112.1112.

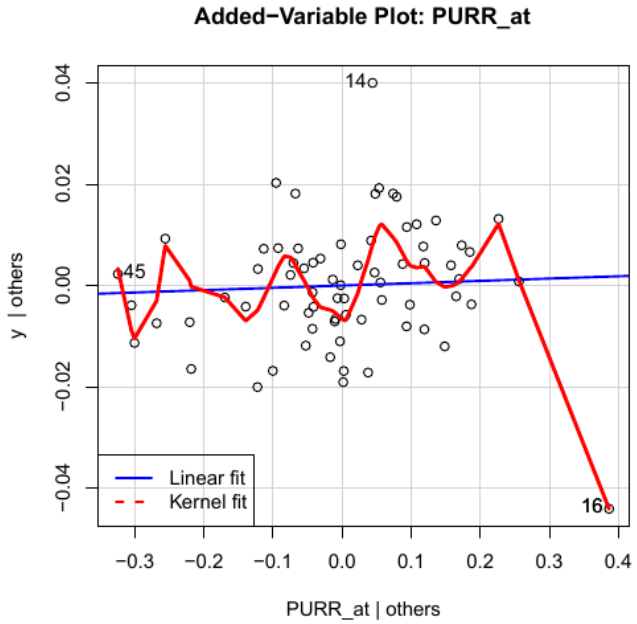

Figure 3. The added-variable plot for "PURR-at"

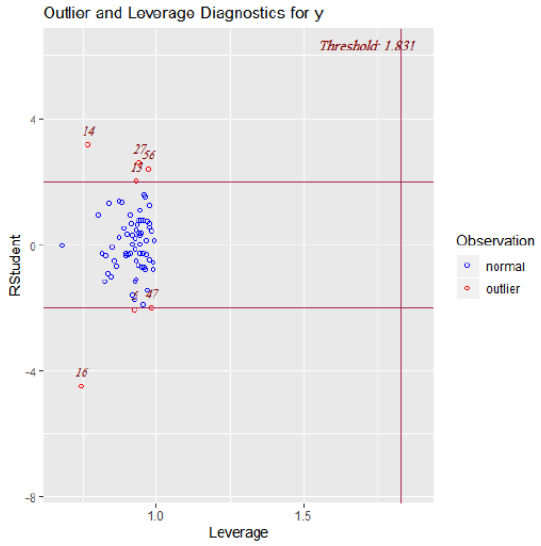

Figure 4. Types of the points in data set

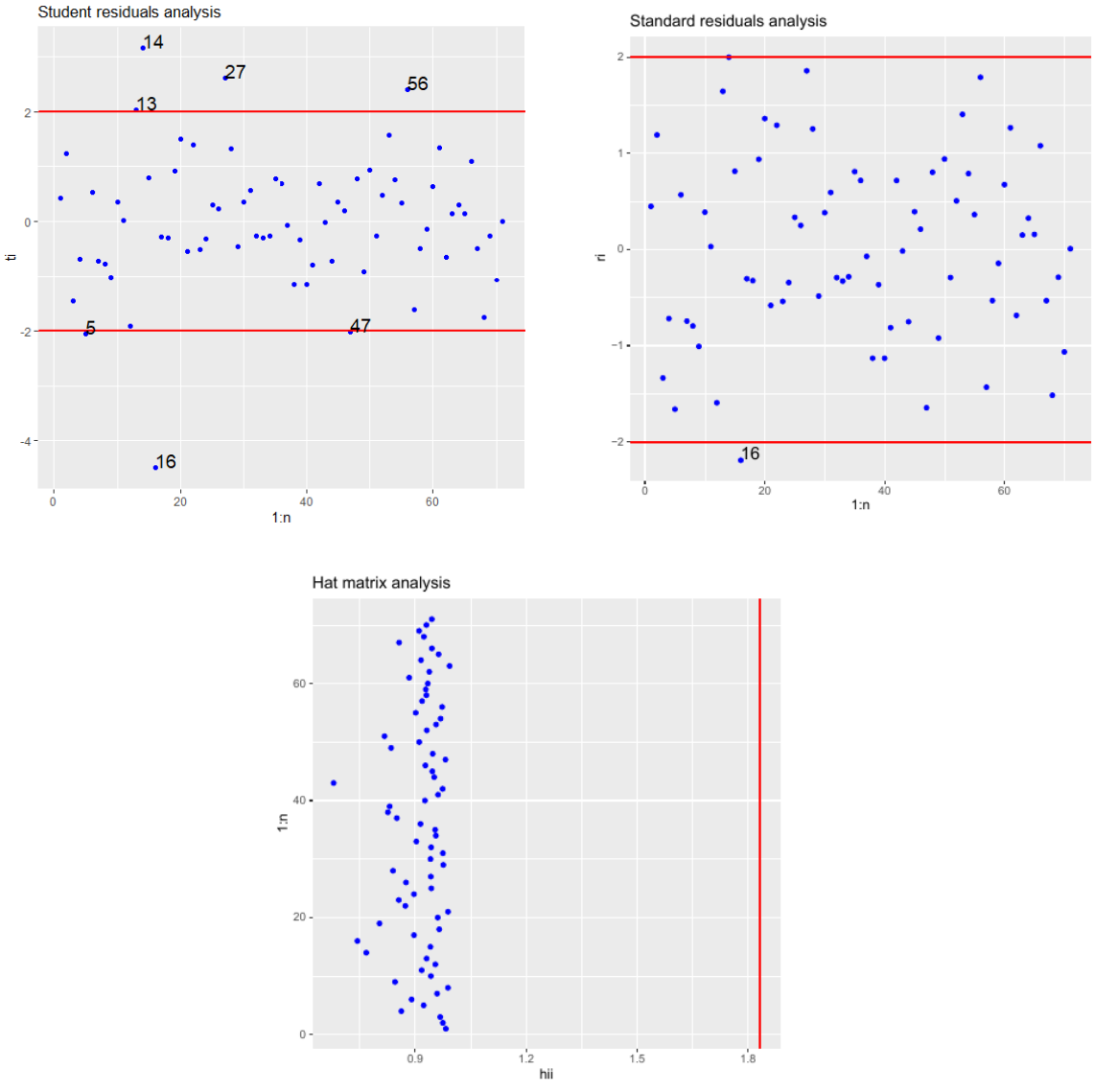

Figure 5. Standard and student residual plot with leverage plot

Table 2. Estimation of the coefficients based on the LTS Method

Parameter	Estimation	Parameter	Estimation
Intercept	0.0021	YHDS-r-at	0.4286
ALD-at	0.0533	YIST-at	-0.2288
ARAM-at	0.4057	YISU-at	0.2021
ARAN-at	-0.1519	YKBA-at	0.1739
ARGF-at	-0.0703	YKNV-at	0.2523
ARGH-at	-0.1453	YKVJ-at	0.2999
DEGA-at	0.5056	YLXW-at	0.1485
ECSB-at	0.1056	YMAH-i-at	-0.0971
GAPB-at	-0.0342	YMFE-at	0.0022
GUTR-at	0.2502	YOAB-at	-0.8605
LYSC-at	-0.2045	YOMT-at	0.4535
METK-at	-0.2365	YOSU-at	-0.0629
PHOA-at	0.0795	YPGA-at	-0.2818
PRIA-a	-0.1879	YPUI-at	0.4879
PYRAA-at	0.0358	YQED-at	-0.1725
sigM-at	0.0259	YQGJ-at	0.2432
SPOIVA-at	-0.6759	YQJT-at	0.4727
SPOVAA-at	0.4881	YQJU-at	0.4641
XHLB-at	0.0718	YRVJ-at	0.1293
XKDB-at	0.0553	YTGB-at	-0.0635
XLYA-at	0.0401	YTSA-at	-0.3836
YACN-at	0.1966	YUID-at	-0.0611
YBFI-at	0.1348	YULC-at	-0.1546
YBXA-at	0.1362	YVFM-at	0.0305
YCLB-at	-0.0044	YWBI-at	0.0612
YDAO-at	-0.0938	YWRO-at	-0.3544
YDDH-at	-0.3378	YXAF-at	-0.1120
YDDK-at	0.0055	YXIB-at	0.0122
YEBC-at	-0.9018	YXLD-at	-0.3641
YETH-at	-0.2432	YXLE-at	0.0669
YFHE-r-at	0.0735	YYBI-at	-0.1588
YFIO-at	0.6113	YYCO-at	-0.4923

In this phase, the researchers estimated the parameters using the sparse LTS Method. The "YCIC-at" variable was selected as a nonparametric part. The regression model is:

$$\boldsymbol{y} = \boldsymbol{X}_{new}\boldsymbol{\beta} + \mathrm{f}(\mathrm{YCIC} - \mathrm{at}) + \boldsymbol{\varepsilon}, \qquad (5)$$

where $\boldsymbol{X}_{new}$ is an $n \times (\boldsymbol{p} - \boldsymbol{1})$ design matrix without "YCIC-at" variable, with the explanatory variables such that $\boldsymbol{x}_i = (x_{1i}, \dots, x_{ni})^{T}, i = 1, \dots, p,$ $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{(p-1)})^{T}$ is a $p \times 1$ vector of unknown regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^{\mathrm{T}}$ is an $n \times 1$ vector of error terms with $E(\boldsymbol{\varepsilon}) = 0$ and $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{T}) = \sigma^2 \boldsymbol{I}_p$.

Figure 6. Estimation of nonparametric part of model (5)

Figure 6 shows the estimation of the nonparametric part. The optimal value of the penalty parameter was equal to 0.0040871, and 63 explanatory variables remained in the model.

The outliers can be seen in Figure 7, and it was clear that there were some outliers in the data set. Therefore, a robust method should be used to estimate the coefficients.

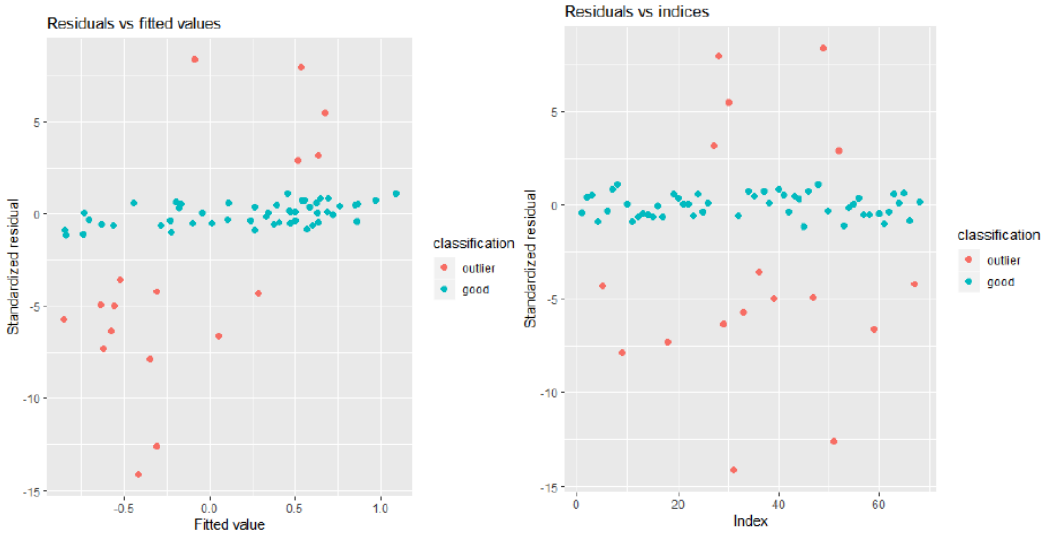

Figure 7. Outlier detection plots

Table3. Estimation of the coefficients based on the sparse LTS Method

Parameter	Estimation	Parameter	Estimation
Intercept	0.1487	YDHB-at	0.1027
ALAS-at	-0.1492	YDJL-at	0.0686
ALST-at	-0.1066	YFMF-at	0.1115
CGEC-at	-0.1173	YHJO-at	-0.1865
CITC-at	-0.1469	YISH-at	0.1143
CITR-at	-0.0644	YKUH-at	0.4568
LEUA-at	-0.0181	YNGE-at	0.2526
MENE-at	-0.0200	YPBH-at	0.2395
mrpF-at	0.2649	YQAD-r-at	-0.4528
MUTS-at	-0.5433	YQDB-at	0.0022
OPUBA-at	0.7186	YQHL-at	0.5141
PADC-at	-0.2245	YQJY-at	-0.0908
PRKA-at	-0.0318	YQKD-at	1.2713
PYRE-at	-0.0355	YTPS-at	0.3151
SDHA-at	-0.2160	YTQB-at	0.3070
SPOIIIAF-at	-0.0244	YUSB-at	-0.02733
SRFAC-at	0.0914	YVDH-at	-0.0033
TAGA-at	0.0232	YVEA-at	-0.2210
TRER-at	0.0883	YWTB-at	-0.4932
YBBK-at	0.1109	YXKI-at	-0.4932

Table 3 reports the estimation of the coefficients by the LTS Method. As is shown in the table, the MSE for this model was equal to 1.3807. It should be noted that the MSE was equal to 148.9527 based on the nonrobust LASSO Method, and so, it was quite clear that the robust method performed better than nonrobust type.

3. Discussion

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 be exactly solved. In this regard, we have proposed a penalized optimization approach for semiparametric regression models to simultaneously combat high-dimension and outliers in the data set. Especially, based on a difference-based scheme, we have suggested a robust programming problem using LTS regression estimation. The results of this work can potentially be extended to the case of heteroscedastic or correlated errors. However, the findings of the present study may be extended by revised Cholesky decomposition, QR decomposition, and extended least trimmed squares estimation to combat multicollinearity due of high-dimension and outliers problem in the data sets (31-33).

Acknowledgments

The authors thank the anonymous referees and associate editor for their useful comments and suggestions on an earlier version of the manuscript which resulted in this improved version. This paper is derived from Master's thesis (Mathematical Statistics) of M. Maanavi in Semnan University with code 2599975.

Conflicts of Interest

The authors declare that there are no conflicts of interests.

References

1. Liu H, Sirish S, Wei J. On-line outlier detection and data cleaning. Computers and Chemical Engineering. 2004;28(9):1635-1647.
2. Hawkins D. Identification of outliers. London: Chapman and Hall; 1980.
3. Barnett V, Lewis T. Outliers in statistical data. 3rd ed. Chichester: John Wiley and Sons; 1994.
4. Beckman RJ, Cook RD. Outlier ... ... s. Technometrics. 1983;25(2):119-149.
5. Sheather SJ. A modern approach to regression with R. New York: Springer; 2009.
6. Moore DS, Mccabe GP, Criag BA. Introduction to the practice of statistics. 9th ed. New York: WH Freeman and Company; 2017.
7. Das MK, Gogoi B. Influential observations and cutoffs of different influence measures in multiple linear regression. International Journal of Computational and Theoretical Statistics. 2015;2 (2):79-85.
8. Rousseeuw PJ, Van Driessen K. Computing LTS regression for large data sets. Data mining and Knowledge Discovery. 2006;12(1):29-45.
9. Engle RF, Granger CWJ, Rice J, WEISS A. Semiparametric estimation of the relation between weather and electricity sale. Journal of the American Statistical Association. 1986;81(394):310-320.
10. Yatchew A. An elementary estimator of the partial linear model. Economics Letters. 1997;57(2):135-143.
11. Yatchew A. Nonparametric regression techniques in economics. Journal of Economic Literature. 1998;36(2):669-721.
12. Leskovec J, Rajaraman A, Ullman JD. Mining of massive datasets. 2nd ed. Cambridge: Cambridge University Press; 2014.
13. Samet H. Foundations of multidimensional and metric data structures. 1st ed. Burlington: Morgan Kaufmann; 2006.
14. Tan N, Steinbach M, Kumar V. Introduction to data mining. Son Francisco: Pearson Addison Wesley, 2006.
15. Tibshirani R. Regression shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Ser. B. 1996;58(1):267-288.
16. Rousseeuw PJ. Least median of squares regression. Journal of the American Statistical Association. 1984;79(388):871-880.
17. Visek JA. Regression with high breakdown point. Proceedings of the 11-th summer school JCMF; 2000 Sep 11-15; Nectiny, Czech. 2001. p. 324-356.
18. Rousseeuw PJ, Leroy AM. Robust regression and Outlier Detection. New York: John Wiley and Sons: 1987.
19. Alfons A, Croux C, Gelper S. Sparse least trimmed squares regression for analyzing high dimesional large data set. The Annals of Applied Statistics. 2013;7(1): 226-248.
20. Härdle WK, Liang H, Gao J. Partially linear models. Heidelberg: Physika Verlag; 2000.
21. Speckman P. Kernel smoothing in partial linear models. Journal of the Royal Statistical Society, Ser B. 1988;50(3):413-436.
22. Roozbeh M. Robust ridge estimator in restricted semiparametric regression models. Journal of Multivariate Analysis. 2016;147:127-144.
23. Amini M, Roozbeh M. Optimal partial ridge estimation in restricted semiparametric regression models. Journal of Multivariate Analysis. 2015;136:26-40.
24. Roozbeh M. Optimal QR-based estimation in partially linear regression models with correlated errors using GCV criterion. Computational Statistics & Data Analysis. 2018;117:45-61.
25. Roozbeh M, Babaie-Kafaki S, Naeimi Sadigh A. A heuristic approach to combat multicollinearity in least trimmed squares regression analysis. Applied Mathematical Modeling. 2018;57:105-120.
26. Akdeniz F, Roozbeh M. Generalized difference-based weighted mixed almost unbiased ridge estimator in partially linear models. Statistical Papers. 2019;60:1717-1739.
27. Hall P, Kay J, Titterington DM. Asymptotically optimal difference-based estimation of variance in nonparametric regression. Biometrika. 1990;77(3):521-528.
28. Buhlmann P, Kalisch M, Meier L. High dimensional statistics with a view towards applications in biology. Annual Review of

Statistics and its Applications. 2014;1(1):255-278.
29. Arashi M, Roozbeh M. Some improved estimation strategies in high dimensional semiparametric regression models with application to riboflavin production data. Statistical Papers. 2019;60:667-686.
30. Amini M, Roozbeh M. Improving the prediction performance of the LASSO by subtracting the additive structural noises. Computational Statistics. 2019;34:415–432.
31. Babaie-Kafaki S, Roozbeh M. A revised Cholesky decomposition to combat multicollinearity in multiple regression models. Journal of Statistical Computation & Simulation. 2017;87(12):2298--2308.
32. Roozbeh M, Babaie-Kafaki S, Arashi M. A class of biased estimators based on QR decomposition. Linear Algebra and its Applications. 2016;508:190-205.
33. Roozbeh M, Babaie-Kafaki S. Extended least trimmed squares estimator in semiparametric regression models with correlated errors. Journal of Statistical Computation & Simulation. 2016;86(2):357-372.