EMPIRICAL STUDY ON ROBUST REGRESSION ESTIMATORS AND THEIR PERFORMANCE

Article · August 2023

DOI: 10.24412/1932-2321-2023-273-466-478

2 authors:

Sajesh T.A
St. Thomas College, Thrissur Kerala
18 PUBLICATIONS **68** CITATIONS
SEE PROFILE

Lakshmi Raveendran
St. Thomas College, Thrissur
6 PUBLICATIONS **2** CITATIONS
SEE PROFILE

EMPIRICAL STUDY ON ROBUST REGRESSION ESTIMATORS AND THEIR PERFORMANCE

Lakshmi R[1], Dr.Sajesh T A [1]

•

Department of Statistics,St. Thomas College (Autonomous), Thrissur 680001, University of Calicut [1]
lakshmi.nss19@gmail.com, sajesh.t.abraham@gmail.com

Abstract

Regression Analysis is statistical technique to model data. But the presence of outliers and influential points affect data modelling and its interpretation. Robust regression analysis is an alternative choice to this. Here we made an attempt to study different robust estimators and propose a new robust reweighted S_n covariance based regression estimator. We have evaluated the performance empirically and the simulation study shows our proposed estimator is preferable to OLS and other robust regression estimators in terms of the MSE criteria. Also, proposed robust S_n covariance regression estimator produce outperforming results for regression equivaraince and breakdown criterion. Robustness of the proposed estimator is proved empirically. The proposed method is innovatively used to model fluid data. R software is used for simulation and study.

Keywords: robust S_n regression,influential observations, modelling,data analysis,

1. Introduction

One of the most essential statistical methods in data modelling is regression analysis. It helps in the prediction of a link between the predictors and the response variable. All academic disciplines, including social science, health science, engineering, physical science, and others, frequently use it. Regression Analysis mainly rely on ordinary least squares method, which is very vulnerable in the midst of the outliers. Informally outlier can be defined as those observations which lie out of the place with respect to other observations in the data set. Thus, when there are polluted points in the data set, robust regression was created as an improved and effective alternative to least squares. There are numerous robust regression techniques; among them some are resistant too. In this paper, we discuss about some of the mainstream and efficient robust regression techniques for contaminated data in multiple linear regression models. Apart from that, multiple regression can achieve efficiently through expressing classical normal equations in covariance matrix form. In this paper, apart from discussing robust regression estimators, we propose a robust reweighted regression based on S_n covariance matrix. The main inherent idea is to compare the techniques using simulated data set, and determine the properties of the proposed estimator through vast empirical simulations alone. Simulation is done and evaluated by using Monte Carlo technique. Section 2 briefly describes about the OLS method, necessity of robust techniques and important robust estimators developed over years and propose a new reweighted regression estimator based on robust covariance matrix technique. Section 3 of this paper presents different simulation methods for comparing proposed estimator and other robust regression estimators, along with that, properties of the proposed estimator studied through wide simulation. Section 4 provides the real life data application and conclusion of the paper.

2. Ordinary Least Squares

Linear regression model is about estimating the parameter $\beta \in R^p$

$$y_i = x_i\beta + \epsilon_i \quad i = 1, 2, 3..., n. \tag{1}$$

where $(x_i, y_i) \in (R^p, R)$ comprise the data and β is the p- dimensional unknown vector and ϵ_i are unknown errors. The best-known estimator of β is the least square estimators obtained by:

$$min \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

The least square estimators are very popular because of Gauss Markov theorem and very easy to use. These classical estimators are the best when their assumptions are met by the data. Whenever there are outliers in the data, OLS results in unstable estimate prediction and are renowned for misbehaving. The data may contain outliers for a number of reasons, including incorrect data entry, incorrect scoring, and unusual sample data. In regression, outliers can be classified according to their location and effect. Observations would be unusual with respect to y values or x values. They are categorised as outliers, leverages and influential points based on how they affect the model. The impact of these observations depends on the location where they occur. Extreme values in the predicted variables are called as leverages. Leverages measure how far an independent variable deviates from its mean. The direction of the distance between the remaining data points is not taken into account by leverages. Leverages do not affect the estimates of the regression coefficients. It affects the model summary statistics, standard errors of regression coefficient etc. Influential points are those points with unusual x coordinate and the unusual y value. The regression coefficients are noticeably affected by influential points. Influential points pull the regression model in its direction. Outliers in either the x or y directions constitute a significant hazard to least square estimators. Statistical or graphical methods can be used to identify outliers. Mahalanobis distance is a statistical procedure used to locate the outliers in the x direction. We cannot say Mahalanobis distance as a perfect method, as it fails to detect the outliers in y direction. Other statistical outlier diagnostics works on the idea of erasing one observation at a time and recalculates the regression coefficients; they are called as regression diagnostics, in which diagnostic quantities are obtained using the data with aim of identifying influential points. Following the identification, they are either eliminated or corrected, and then the least squares analysis is performed. As a result, such statistics estimate the change in regression coefficients that would occur if a single observation were removed following analysis. These statistics are also known as deletion statistics, useful for pinpointing influential points. Cook distance, Studentised residuals, DFFITS, DFBETAS and Jacknife residuals are some of such deletion statistics. Calculation of these diagnostic statistics become complicated when there are multiple unusual observations. Robust regression estimation is alternative strategy for handling outliers. Robust methods aim to create estimators that are immune to outliers. Diagnostic tools remove outliers before fitting the data using the least square approach, whereas, Robust regression, on the other hand, fits a regression model to the great majority of the data before identifying outliers as regions with substantial residuals. The breakdown point, concept of bounded influence and relative efficiency are ideas that are pertinent to the study of robust regression. The presence of single outlier can completely invalidate the OLS estimator. Contrast to it; we will see estimators that can handle certain percentage of outliers. This particular concept is called as breakdown point and [3] provided the first explanation of a breakdown point.It'd only evaluate location in a single dimension.Also,[5] provided broad description of braekdown, but it was highly mathematical in nature and asymptotic.It was [4], suggested a limited sample version of breakdown point.
For a sample $\mathbf{Z}$ of n observations,

$$Z = (x_{11}, ..., x_{1p}, y_1), ..., (x_{n1}, ...x_{np}, y_n)$$

Let T represents a regression estimator. When T is applied to such a sample, the result is a regression coefficient vector as $T(Z) = \hat{\beta}$. Let j of the sample data points swapped by arbitrary values and call them as corrupted sample $Z^{'}$. The maximal bias generated by such contamination is then calculated as:

$$bias(j; T, Z) = sup_{Z'} \|T(Z^{'}) - T(Z)\|,$$

Where the supremum is over all possible Z'. If the bias is infinite, j outliers have a significant impact on the estimator. Thus, breakdown of the estimator T at the sample Z is defined as

$$\epsilon_n^*(T, Z) = min(\frac{j}{n}; bias(j; T, Z) \quad is \quad infinite).$$

Or the least amount of contamination that an estimator can tolerate is known as the breakdown point. The breakdown point of ordinary least estimator is $\epsilon_n^*(T, Z) = \frac{1}{n}$. That is, even the presence of single outlier in the data set can affect least square estimators.

2.1. Proposed Method

OLS estimator can express as solution to 2 proposed by [8] in the following way. Let $z = (x, y)$ be the joint variable of independent and dependent variables. Letμ be the location and Σ be the scatter matrix of z. Partitioning μ and Σ yields the notation

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \text{ and } \quad \Sigma = \begin{pmatrix} \Sigma_{xx}\, \Sigma_{xy} \\ \Sigma_{xy}\, \Sigma_{yy} \end{pmatrix} \tag{2}$$

Generally the estimates $\hat{\mu}$ and $\hat{\Sigma}$ are estimated in empirical way.The least square estimates of β and α can be written as function of $\hat{\mu}$ and $\hat{\Sigma}$,namely,

$$\hat{\beta} = \hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}, \quad \hat{\alpha} = \hat{\mu}_y - \hat{\beta}^T\hat{\mu}_x. \tag{3}$$

Major drawback of the above mentioned estimators is, classical estimators of location and scatter are sensitive to the presence of the outliers. Robustification of the classical estimators of scatter and location improve the performance of the estimators and [9] in their paper proposed a robust method for detecting multiple outliers and thus robust covariance matrix estimation in multidimensional data set denoted as S_n method. Another objective of our paper is to propose a robust reweighted regression estimator based on S_n covariance estimator in equation (4). In this paper the performance of the proposed robust reweighted S_n regression estimator of the joint variable **z** is evaluated and studied through empirical simulations. The performance and properties of the estimator is investigated through wide range of simulations and Mean Squared error is used to compare the performance of proposed estimator with other estimators in different scenarios.
Let $X = X_1, X_2, X_3, \ldots, X_p$ be a $n \times p$ matrix of size n and p being the number of variables. The robust covariance matrix based on S_n method of the matrix **X** is defined as:

$$S_n(X_i, X_j) = med_i med_{j\neq i}[(x_i - x_j)(y_i - y_j)] \quad , i, j = 1, 2, 3, ...p, \tag{4}$$

where med is an abbrevation for low median ($[\frac{n+1}{2}]^{th}$ order statistic). Inner median will be taken up by $[n/2]^{th}$ order statistic for odd value of n. The corresponding correlation matrix of equation 4 is defined as:

$$\delta_{S_n}(X) = DCOV_{S_n}(X)D^t \tag{5}$$

where D is the diagonal matrix with diagonals $1/S_n(x_i), i = 1, 2, 3, .., p$. Here $S_n(x_i)$ is nothing but robust scale estimator of univariate random variable X and is defined as,

$$S_n(X) = 1.1926\ med_i\ med_j|x_i - x_j|$$

The covariance matrix mentioned in equation 4 is non-positive semi definite and [7] in their paper describe procedure to solve non positive semi definite and obtain positive semi definite and approximately affine equivariant estimators. The following steps provide us positive semi definite dispersion matrix and robust estimates:

- Let e_j be the eigen vector corresponding to the eigen value λ_j of correlation matrix δ_{S_n}. Let E be $p \times p$ matrix with columns e_j for $j = 1, 2, \ldots, p$.
- Let $R = D^{-1}E$ and $z_i = R^{-1}X_i$ and Z be an orthogonalised matrix with rows $z_i^T (i = 1, 2, 3 \ldots .n)$ and columns $Z_j (j = 1, 2, \ldots, p)$.

The resulting robust S_n estimate of location and scatter is defined as:

$$\hat{\mu}_{S_n} = Rv \quad \text{and} \quad \hat{\Sigma}_{S_n} = R\Gamma R^T \tag{6}$$

where $v = (med(Z_1), med(Z_2), ..., med(Z_p))^T$and $\Gamma = diag(S_n(Z_1)^2, ..., S_n(Z_p)^2)$. Here med stands for median and S_n is the univariate robust scale estimate. The process can be iterate to enhance the estimates by replacing covariance estimator used in equation 5 with above $\hat{\Sigma}_{S_n}$. Let us call the robust estimator 6 of location and scatter as initial S_n estimator of z. The associated robust squared Mahalanobis distance of each observation z_i is defined as The resulting robust S_n estimate of location and scatter is defined as:

$$RD(z_i) = (z_i - \hat{\mu}_{S_n})^t \hat{\Sigma}_{S_n}^{-1} (z_i - \hat{\mu}_{S_n})$$

Robust Mahalanobis distance is an efficient outlier detection method. Let w_i be a weighted function based on the above Mahalanobis distance defined as $w_i = w(RD(z_i))$.The reweighted estimators take over the robustness properties of initial estimators with increasing their efficiency ([6]). Therefore the reweighted S_n location and scatter matrix be obtained as:

$$\hat{\mu}_{wS_n} = \frac{\sum_{i=1}^n w_i z_i}{\sum_{i=1}^n w_i} \quad \text{and} \quad \hat{\Sigma}_{wS_n} = \frac{\sum_{i=1}^n w_i (z_i - \hat{\mu_{wS_n}})(z_i - \hat{\mu_{wS_n}})^T}{\sum_{i=1}^n w_i}. \tag{7}$$

The weights above are computed as $w_i = w(RD(z_i)) = I(RD(z_i) \leq c)$, which assign weight 1 to the z_i for $i = 1, 2, .., n$, where

$$c = \begin{cases} \chi^2_{0.95,p} & \text{if} \quad p < 15 \\ \dfrac{\chi^2_{0.95,p} med(rd_1, ...rd_n)}{\chi^2_{0.5,p}} & \text{if} \quad p \geq 15 \end{cases} \tag{8}$$

Based on $\hat{\mu}_{wS_n}$ and $\hat{\Sigma}_{wS_n}$ we obtain $\hat{\beta}_{wS_n}$ and $\hat{\alpha}_{wS_n}$the robust reweighted S_n regression estimator defined as:

$$\hat{\beta}_{S_n} = (\hat{\Sigma}_{wS_n})_{xx}^{-1} (\hat{\Sigma}_{wS_n})_{xy} \quad \text{and} \quad \hat{\alpha}_{S_n} = (\hat{\mu}_{wS_n})_y - (\hat{\beta}_{S_n})^T (\hat{\mu}_{wS_n})_x. \tag{9}$$

The efficiency, breakdown and affine equivariant property of the proposed estimator 9 is evaluated.

3. Simulation Study

Simulation study is done to evaluate the performance of the proposed robust reweighted S_n regression estimator. And the results are compared with ordinary least squares and some of the other robust regression estimators like: LTS, LMS, S, and MM estimator. The simulations are done in R and all the values are reported in tables at the end of the paper. Consider the linear regression model form:

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

where $\boldsymbol{X}$ is $n \times p$ matrix, $\boldsymbol{\beta} = (\beta_1, .., \beta_p)^T$ is the unknown regression coefficient vector of size $p \times 1$,α is the unknown intercept of the model and ϵ is the i.i.d error term and are independent from X. The X variables are distributed as $\boldsymbol{N(0_p, I_p)}$, where I_p is the p- dimensional identity matrix. Following sets of dimensions and sample sizes are considered in this study respectively: p=5, 10, and 20 with n=50,100,500. The simulations are repeated 1000 times and each time parameter estimates are noted.

Mainly three simulation scenarios, as that found in the literature [2] are considered here.

- The dependent variable is generated from standard normal distribution with corresponding regression coefficients including intercept equals zero, and standard normal errors are considered [NES].
- The dependent variable is simulated from t distribution with 3 degrees of freedom with corresponding regression coefficients including intercept equals zero and heavy tailed errors (t distribution with 3 degrees of freedom) [HTS].
- Regression with normal error, some percentage(δ) of randomly selected observation in independent variable replaced as $N(\lambda\sqrt{\chi^2_{(0.99,p)}}, 1)$ and the dependent variables were replaced as $N(k\sqrt{\chi^2_{(0.99,1)}}, 1)$where λ,k=0.5,1,1.5,2,3,5,7,8,10. The percentage of contamination considered in this scenario is 10% and 20%.

3.1. Efficiency

It is a well-known fact that ordinary least squares have maximum efficiency under normal errors. Thus under normal error case, the efficiency of each robust method is calculated relative to OLS. Let $\Phi = (\beta^T, \alpha)^T$be the joint vector of regression parameters, intercept and slope. Dimension of Φ is $(p+1) \times 1$. The finite efficiency for the joint estimator $\hat{\Phi}_{Re}$ of a robust method (Re) is defined as:

$$Eff = \frac{1/1000 \sum_{i=1}^{1000} \|\hat{\Phi}^i_{OLS} - \Phi\|_2^2}{1/1000 \sum_{i=1}^{1000} \|\hat{\Phi}^i_{Re} - \Phi\|_2^2}$$

Table1 exhibit the simulated efficiency of $\hat{\Phi}$, of proposed robust reweighted S_n estimator and other robust regression estimators, with respect to the classical least square estimator, under normal error scenario described above. In the table, bold letters in each row represents the highest efficiency and italic letters represent the lowest efficiency. Estimators with higher efficiencies are represented in bold and lowest efficiencies are represented in italics. Among the estimators under consideration, proposed reweighted S_n estimator exhibits highest efficiency throughout all the randomly chosen dimensions and sample size considered. MM estimator also possesses efficiency greater than 90% in some of the cases under consideration. Among the estimators LMS perform poorly. The proposed estimator has highest efficiency for all randomly chosen sample sizes and dimensions considered.
In the second scenario, we are considering heavy tailed error distribution. Thus least square estimators cannot be maximum efficient estimator. Hence we consider the Mean Squared Error of the estimators instead of Eff. The table 2 results shows that proposed estimator out perform in all scenarios with mean squared error lower than other estimators. Also, as the sample size increases the mean squared error decreases for all the robust methods.

3.2. Robustness

To study the robustness, simulations accordingly in third scenario [CS] defined above are carried out. Here we have randomly considered different dimensions (p=5 and 10) with sample size (50 and 500). The criteria used here to compare the different estimators is mean squared error of the estimated joint parameter vector $\hat{\Phi}$, averaged over 1000 simulation runs, similar criteria considered in [2].Tables 3 to 9 below shows the maximum (across λ and k) MSE for both estimated intercept and slope for different combination of dimension and sample size. We are considering the maximum value of mean squared error obtained over all considered k values, for each value of lambda.

$$i.e. MSE_\lambda(.) = max_{k\epsilon[0.5,1,1.5,2,3,5,7,8,10]} MSE_{\lambda,k}(.)$$

Lowest value of MSE in each row is notated in bold letter in the tables.Tables 3 to 9 gives $MSE_\lambda(.)$ of different robust estimators and proposed estimator. Among the values, proposed estimator shows minimum MSE in all cases considered. For higher dimensions, proposed estimator possess very low MSE than other estimators, even when percentage of contamination increases, proposed estimator shows low MSE and consistently maintain low error throughout different level of

contamination. Here we can see for same dimension, when number of observations increases, mean squared error of reweighted S_n estimator decreases. Thus the performance of proposed estimator increases with increase in the number of observations. Proposed estimator out perform in all the scenarios constantly. MM and S estimators mainly collapse for λ,k =0.5, 1, 1.5, 2, 3 of contamination. And for other values of (λ,k), MM and S estimators perform moderately with MSE values mostly greater than proposed estimator. LTS and LMS estimators performs consistently for all values of λ,k, but the mean squared error obtained is higher than proposed estimator in all scenarios. Among all the scenarios, OLS possess highest mean squared error than other estimators.

3.3. Breakdown Property

The breakdown point evaluates the maximum percentage of outliers an estimator can tolerate. 50% is the highest breakdown value an estimator can attain. Even though high contamination level occurs rarest of rare in general, here we propose to study the performance of the estimators in extreme contamination and evaluate the consistency in their performance. It has shown that repeated median regression estimator has 50% asymptotic breakdown point through simple mathematical induction by [11] . The same lemma quoted in [11] is applicable for reweighted S_n estimator, since the estimator is nothing but nested median of observations. For this, a criterion [CS] is used with percentage of contamination 30%, 35%, 40%. Dimensions considered here are 5 and 30. And we consider the maximum MSE across all combinations of λ,k. i.e.$MMSE(.) = max_{\lambda\epsilon[0.5,1,1.5,2,3,5,7,8,10]} MSE_\lambda(.)$. The results are shown in table 10.
In general, robust estimators like LTS, S, and MM have high breakdown point, but their computations are challenging. In all these mentioned methods regression estimators are obtained by resampling algorithm. Resampling algorithms are used to obtain number of subsamples, and then robust regression estimators are obtained by making use of an initial high breakdown estimator. Thus all these established methods depend purely on number of subsamples and initial estimates. Reweighted S_n estimator proposed here is not dependant on resampling and initial estimates. Also, proposed estimator possess high empirical breakdown even to large contamination and higher dimension.
Based on all simulation results, we can observe that all robust estimators, except our suggested estimator, have a constant increase in mean square error value, which reaches a maximum as the fraction of outliers in the vertical direction reaches a maximum for a certain lambda value. Although both MM and S estimators are stated to have a high breakdown, S estimator performs poorly as the dimension of the variable grows. MM estimator could be consider to be as reasonably good robust estimator, shows low MSE among other established robust estimators. However, the MSE of MM estimator is much higher than that of proposed reweighted S_n estimator. As the percentage of outliers in the data increases, LTS perform poorly. Even though LTS has a 50% breakdown point, the performance of LTS estimator depends merely on the correct choice of tuning constant h, here we used default value h=0.5. Throughout various levels of contamination, our suggested reweighted Sn estimator consistently maintains a low MSE. Also, even in higher dimensions, the proposed estimator has a lower MSE than other well-known robust estimators, indicating that proposed estimator is more resistant to large numbers of outliers, which can be termed as high empirical breakdown point.

3.4. Equivariance Property

Rather than theoretical goodness, practical usefulness of an estimator is determined by equivariance, breakdown and robustness properties. These three qualities are considerable properties of a regression estimator and discussed in our paper. Breakdown property is described in the above section. For regression estimators three types of equivariance are considered:

1. Regression equivariance is defined as: if we transform the dependent variable by adding a linear function of independent variables, is equal to adding the coefficients of this linear

function to the estimators.

2. y- equivariance is defined as, if the dependent variable is transformed linearly, then the estimators get transformed in the same manner.
Let $\hat{\Phi}(X,Y) = (\hat{\beta}^T, \hat{\alpha})^T$, where X is $n \times p$ matrix and Y is $n \times 1$ matrix. Then (1) and (2) can be combined to form: $\hat{\Phi}(X, Yb + Xg + u) = \hat{\Phi}(X,Y)b + (g^T, u)^T$, where $b \epsilon R$, a non-zero constant, g is $p \times 1$ vector and $u \epsilon R$ is any constant. Keeping X as same and transforming the dependent variable as $Yb + Xg + u$, then the resulting estimator would be: $\hat{\beta}_{new} = b(\hat{\beta}) + g$ and $\hat{\alpha}_{new} = b\hat{\alpha} + u$.

3. x- equivariance is defined as, if the independent variables are linearly transformed, then the equivalent transformed estimator is: $\hat{\Phi}(XA, Y) = (\hat{\beta}^T (A^{-1})^T, \hat{\alpha})^T$.

That is, if the independent variables are transformed as XA, with a non-singular $p \times p$ matrix A, the resulting new estimators are $\hat{\beta}_{new} = A^{-1}\hat{\beta}$and $\hat{\alpha}_{new} = \hat{\alpha}$.It is not possible to explore available transformations, so, [7] and [10] proposed in their papers, to generate A matrices randomly for the purpose of checking x-equivariance as $A = TD$, where T a random orthogonal matrix and D is a $p \times p$ diagonal matrix with diagonal entries are independently and uniformly distributed. The methods outlined above are employed in our paper to investigate the suggested estimator's equivariance property. When the above mentioned transformations are performed on simulated data sets, the MSE of the suggested estimator is examined. Here we consider two dimensions, p=5 and p=30. Also, we consider contaminated scenario [CS] with contamination of 10% and 20%. First the estimator is applied to untransformed data and the estimator obtained $\hat{\Phi}_{S_n}$ is recorded. The above mentioned transformed data is then used to estimate x-equivariance, y-equivariance, regression equivariance, and the resulting new estimator $\hat{\Phi}_{S_n new}$is stored. The MSE is calculated between $\hat{\Phi}_{S_n new}$ and estimator value which has to be obtained if the above properties hold. The estimator consistently performs and maintains low MSE. Even when the contamination increases with increase in dimension, the estimator exhibits low MSE. Low MSE indicates that the model can be predicted more accurately. The suggested estimator is essentially affine equivariant since the model's error is managed and kept low.
Table 11 shows MMSE results for x- equivariance. From the table, we can see that the error remains controlled and low for varied proportion of vertical outliers and leverage points. The error value increases as the dimension increases, but in a regulated manner. The suggested estimator is nearly x-equivariant since the errors are controlled.
Table 12shows the MMSE for y-equivariance and regression equivariance. The mean square error remains very low throughout for different proportion of vertical outliers and leverage points. Also, we can see mean square error shows decreasing pattern as dimension increases. Though mean square error increases with increase in the percentage of contamination level, the increments are very small and close to zero. As a result, we can state that the mean square error is well-controlled and kept to a minimum in all scenarios evaluated. Thus the proposed estimator is approximately y-equivariant and regression equivariant. We have empirically demonstrated three equivariance features using simulated samples with contamination at various degrees and dimensions.

4. Real life data application

4.1. Fluid Dynamics

A substance capable of flowing is termed as a fluid. Fluids are of two types, namely liquids and gases. The study of fluid's behaviour at rest (termed fluid statics) and in motion (termed fluid dynamics) is jointly known as fluid mechanics. Many real-world applications, including cancer treatment, car radiators, air conditioning, refrigeration, microwave ovens, blow moulding, and petrochemical processing, heavily rely on heat and mass transmission. Through extensive studies scientists have been successful in improving these transmission qualities. Choi and Eastman were the first to notice that the introduction of nano sized particles in a conventional fluid was able to bring a significant improvement in these transmission qualities. Internal heat sources play an

important role in various heat transfer applications.
For its vital role in photo thermal and photodynamic therapy, the dynamics of water-based TiO_2 nano liquid over an elongated nonlinear surface was elucidated by [1]. The flow problem was modeled using partial differential equations which were solved using finite-difference based bvp5c technique with the help of apposite similarity transformations. Further, the authors employed response surface methodology and sensitivity analysis to elucidate the heat transfer rate for the consequence of magnetic field $(0.5 \leq M \leq 1.5)$, thermal radiation $(0.5 \leq R_d \leq 1.5)$ and exponential heat source $(0.2 \leq Q_E \leq 0.4)$. The optimal heat transfer rate was observed when M=0.5,R_d =1.5, and Q_E =0.2.
Recently, researchers have examined the influence of effectual parameters on the engineering quantities using statistical methods like regression analysis, and response surface methodology. By establishing a quantitative relationship between the independent (relevant characteristics) and dependent (physical process of interest) variables, the inclusion of these statistical techniques tends to broaden perception. Typographical errors while handling such data is a possibility, owing to those outliers could occur in such datasets, which should be tackled scientifically.
In this paper, the fluid data for conducting multiple linear regression analysis has been derived from Areekara et al. The derived data has been analyzed using proposed regression estimator along with other robust regression estimators. The data consist of 20 observations with three independent variables and one independent variable. Initially we analyze the performance in original data without outlier and then we conduct the same procedure by replacing 10th and 19th observation into outliers. The results are reported in tables 13 and 14 below.Proposed estimator performs in normal scenario and outlier injected data as that of other existing robust estimator like MM, S estimators. We have also reported the OLS estimated values, from which it is clear that classical method fails in the performance of outliers. Also, LTS, LMS regression methods fail to perform in these data sets due to their computational complexity. Thus we have shown in our paper that, for conducting linear regression analysis in such fluid data sets, it is always better to use robust regression methods; even there are no outliers in the data set. Also, proposed estimator works well even without outliers in the datasets. In this paper, the fluid data for conducting multiple linear regression analysis has been derived from Areekara et al. The derived data has been analyzed using proposed regression estimator along with other robust regression estimators. The data consist of 20 observations with three independent variables and one independent variable. Initially we analyze the performance in original data without outlier and then we conduct the same procedure by replacing 10^{th} and 19^{th} observation into outliers. The results are reported in tables 13 and 14 below.Proposed estimator performs in normal scenario and outlier injected data as that of other existing robust estimator like MM, S estimators. We have also reported the OLS estimated values, from which it is clear that classical method fails in the performance of outliers. Also, LTS, LMS regression methods fail to perform in these data sets due to their computational complexity. Thus we have shown in paper that, for conducting linear regression analysis in such fluid data sets, it is always better to use robust regression methods; even there are no outliers in the data set. Also, proposed estimator works well even without outliers in the datasets.

4.2. Belgian Phone Call Data

Belgian phone calls data was published by Belgium Statistical Survey and [[6]] used the data in their work. The data consists of annual count of international calls from Belgium during the period 1950 to 1973. The data comprise of two variables, the year (X) and the number of call received (Y). The data contains six outliers in Y direction. From the table 9 below, it is clear that, OLS provides highly misleading estimates in the presence of anomalies. Also, M and LMS performances are not remarkable. MM, S and LTS don't exhibits remarkable performance in outlier detection and in providing estimates. Among them, S_n estimator detects the outlying observations as ($15^{th}, 16^{th}, 17^{th}, 18^{th}, 19^{th}, 20^{th}$, and 24^{th} observation) and MM estimator detects outlying observations as ($15^{th}, 16^{th}, 17^{th}, 18^{th}, 19^{th}, 20^{th}$, and 21^{st} observation). Thus, proposed

estimator performs well and detects outliers correctly as that of MM estimator and it can be considered as an add-on to the covariance based regression method.

5. Conclusion

Classical regression analysis is very sensitive to the presence of contaminated observations. Several robust alternative methods are available in the literature. In this paper, we propose an improved robust reweighted S_n regression estimator. Here we are proposing a new robust regression technique using alternative form of OLS. We propose a new robust reweighted robust S_n regression estimator. The properties and performance of our proposed estimator are inferred through wide range of empirical simulation methods. Also, the performance of proposed estimator in fluid dynamics data and Belgian phone call data is evaluated. Proposed estimator exhibit a consistent performance in all the cases considered. The robustness property, affine equivariance and breakdown property of the proposed estimator is compared with OLS, MMS, LTS, LMS, S estimators using simulation study. And in all scenarios considered, proposed estimator outperforms other existing robust estimators. The results are tabulated below. Although many robust regression estimators have already been proposed in the literature, we could add proposed estimator to the list of available regression estimators, since proposed estimator exhibit excellent performance than other estimators. A thorough comparison has done and we can conclude that proposed estimator possess high breakdown, robustness equivariance property. Also, the proposed estimator is suitable for multiple regression estimation and is a good alternative to the classical estimator. Developing theoretical properties of the proposed estimator is the future aim of our work.

Table 1: *Table showing efficiency in case of Normal error scenario*

p	n	S_n	MM	S	LTS	LMS
5	30	0.9684	0.9309	0.2813	0.6708	0.0873
	50	0.9286	0.9228	0.2814	0.4639	0.1797
	100	0.9235	0.8951	0.2312	0.5166	0.1299
	500	0.9857	0.9309	0.2728	0.7414	0.0509
10	50	0.9466	0.9832	0.2520	0.6917	0.0292
	100	0.9429	0.8461	0.2888	0.4819	0.1211
	500	0.9852	0.8706	0.1992	0.6314	0.0315
	1000	0.9283	0.9321	0.2123	0.7119	0.0163
30	100	0.8751	0.7409	0.2617	0.7157	0.0033
	500	0.9628	0.8706	0.1992	0.6314	0.0315
	1000	0.9335	0.9321	0.2123	0.7119	0.0163
	5000	0.9817	0.7409	0.2617	0.7157	0.0033

Table 2: *Table showing the MSE in case of t tailed error distribution*

p	n	S_n	MM	S	LTS	LMS
5	30	0.4530	0.5326	1.1059	0.9081	2.0202
	50	0.2429	0.2501	0.5465	0.3708	0.8954
	100	0.1083	0.1192	0.2059	0.1391	0.4232
	500	0.0196	0.0203	0.0349	0.0224	0.1449
10	50	0.5387	0.5307	2.3917	0.8783	2.4023
	80	0.2996	0.3984	0.6684	0.3959	1.3699
	100	0.2111	0.3667	0.4762	0.2979	0.9966
	500	0.0396	0.1467	0.0629	0.0421	0.5392
30	100	0.3448	0.8933	2.2357	1.2430	6.2807
	150	0.4834	0.5339	0.8678	0.6666	4.2174
	500	0.0254	0.0971	0.2479	0.1269	3.2732

Table 3: *Table 3(1) showing $MSE_\lambda(.)$ for n=500,p=5,δ=10%*

λ	S_n	MM	S	LTS	LMS	OLS
0.5	0.1595	2.7088	1.7421	0.4396	0.8785	91.915
1	0.1158	2.2949	1.8572	1.7479	1.0189	37.916
1.5	0.1166	1.7354	1.8309	1.7042	1.0428	18.686
2	0.1158	2.6151	2.6478	1.8624	1.0528	10.881
3	0.1114	2.7446	2.5763	1.8506	1.0556	4.9069
5	0.1139	1.7846	1.8011	2.6581	1.0500	1.7826
7	0.1122	0.9074	0.9328	2.4172	1.0560	0.9135
8	0.1142	0.6917	0.7117	2.6076	1.0563	0.6882
10	0.1162	0.4473	0.4600	1.8004	1.0519	0.4438

Table 4: *Table 3(2) showing $MSE_\lambda(.)$ for n=500,p=5,δ=20%*

λ	S_n	MM	S	LTS	LMS	OLS
0.5	1.6434	7.3536	7.4798	7.5241	1.0062	134.02
1	0.1488	10.669	9.9103	9.7256	1.0749	42.848
1.5	0.1216	10.811	9.7874	9.7707	1.0882	19.762
2	0.1217	7.2620	10.175	10.198	1.0854	11.196
3	0.1209	4.9965	5.0651	5.0491	1.0857	4.9839
5	0.1169	1.7835	1.8156	1.8243	1.0707	1.7938
7	0.1195	0.9116	0.9263	0.9398	1.0579	0.9109
8	0.1207	0.4474	0.7124	0.7150	1.0662	0.6961
10	0.1153	0.4455	0.3032	0.3072	1.0812	0.4463

Table 5: *Table 3(3) showing $MSE_\lambda(.)$ for n=50,p=5,δ=10%*

λ	S_n	MM	S	LTS	LMS	OLS
0.5	0.6232	3.0231	2.8642	3.2983	1.1107	99.802
1	0.4038	2.6021	2.8603	3.0759	1.1709	38.799
1.5	0.3998	2.6271	2.7735	3.4003	1.1968	18.827
2	0.3867	2.6629	2.8132	3.6539	1.1950	11.005
3	0.3904	2.8386	2.7222	3.3677	1.2375	4.9652
5	0.3834	1.6583	1.7868	1.6814	1.2058	1.8273
7	0.3945	0.9234	0.6863	0.9836	1.2035	0.9278
8	0.3804	0.7253	0.8329	0.7483	1.1877	0.7016
10	0.3995	0.4704	0.5515	0.5209	1.1998	0.4827

Table 6: *Table 3(4) showing $MSE_\lambda(.)$ for n=50,p=5,δ=20%*

λ	S_n	MM	S	LTS	LMS	OLS
0.5	1.1457	11.154	12.784	18.499	99.802	139.16
1	0.4178	12.066	12.407	18.179	38.799	42.948
1.5	0.4215	12.643	12.710	20.576	18.827	19.915
2	0.4124	10.342	10.441	11.408	11.005	11.268
3	0.4169	5.0657	5.2375	5.2001	4.9652	5.0114
5	0.4113	1.1892	1.8959	1.8915	1.8273	1.8074
7	0.4014	0.9862	1.0382	0.9977	1.5036	0.9311
8	0.4133	0.7476	0.8552	0.8458	1.1879	0.7362
10	0.4158	0.3324	0.6009	0.5703	1.1999	0.4736

Table 7: *Table 3(5) showing $MSE_\lambda(.)$ for n=50,p=10,δ=10%*

λ	S_n	MM	S	LTS	LMS	OLS
0.5	0.4056	6.8755	7.5017	8.4836	1.1244	1.0269
1	0.4073	7.8391	7.6248	9.9929	1.1958	1.1821
1.5	0.4216	7.8541	7.1342	9.2994	1.2212	12.714
2	0.3726	6.2529	6.1745	6.6639	1.2174	7.1902
3	0.3356	3.6121	3.2540	3.2603	1.2399	3.2836
5	0.3475	1.2000	1.2886	1.2639	1.2546	1.1877
7	0.3509	0.6345	0.4943	0.6841	1.2516	0.6207
8	0.3512	0.4756	0.5748	0.5516	1.2623	0.4852
10	0.3568	0.3197	0.2922	0.3760	1.2248	0.4229

Table 8: *Table 3(6) showing $MSE_\lambda(.)$ for n=50,p=10,δ=20%*

λ	S_n	MM	S	LTS	LMS	OLS
0.5	0.7139	42.630	37.898	7.5892	1.1537	108.34
1	0.4418	28.875	28.031	7.1634	1.1952	28.931
1.5	0.4239	13.086	13.160	7.2613	1.2349	12.923
2	0.4079	7.4023	7.5233	6.2072	1.2334	7.2643
3	0.4008	3.3492	2.2409	3.1659	1.2526	1.5981
5	0.4797	1.2284	1.3222	1.2963	1.2698	1.1756
7	0.4125	0.6363	0.7376	0.7444	1.2424	0.6244
8	0.4533	0.5037	0.4316	0.5726	1.2968	0.4769
10	0.4229	0.2499	0.4273	0.3948	1.2854	0.3298

Table 9: *Table 3(7) showing $MSE_\lambda(.)$ for n=500,p=10,δ=10%*

λ	S_n	MM	S	LTS	LMS	OLS
0.5	0.1655	4.0859	4.5578	0.2901	1.0055	90.827
1	0.1626	5.8239	4.5036	1.1616	1.0425	27.440
1.5	0.1598	5.9110	5.3589	2.6196	1.0638	12.582
2	0.1603	5.8077	4.6791	4.5252	1.0524	7.1467
3	0.1618	3.1899	3.2266	2.6401	1.0529	3.1912
5	0.1622	1.1576	1.1758	4.6335	1.0474	1.1535
7	0.1612	0.5905	0.4584	5.3122	1.0526	0.5885
8	0.1558	0.4490	0.3022	4.5747	1.0559	0.4499
10	0.1589	0.2865	0.3001	4.6989	1.0422	0.2879

Table 10: *Table.4 showing the MMSE(.)for checking breakdown property*

Method	5			30		
	δ=30 %	δ=35 %	δ=40 %	δ=30 %	δ=35 %	δ=40 %
OLS	12.438	12.721	12.934	8.6068	10.133	11.681
S_n	0.5108	1.5456	1.7754	0.9073	1.0807	2.1518
MM	6.3341	9.2566	6.1267	7.4952	9.2222	12.584
S	6.8982	8.6637	13.662	36.451	34.157	30.675
LTS	47.399	48.661	206.54	886.55	1055.6	1033.1

Table 11: *Table 5 showing the $MMSE_\lambda(\hat{\Phi}_{S_n new})$ for checking x-equivariance*

	p=5		p=30	
λ	δ=10%	δ=20%	δ=10%	δ=20%
0.5	0.01953	0.03739	0.10566	0.13301
1	0.05469	0.03865	0.29562	0.32564
1.5	0.03992	0.03566	0.11913	0.17027
2	0.03909	0.01530	0.13167	0.25173
3	0.03842	0.01900	0.16799	0.28351
5	0.03488	0.03279	0.16283	0.18638
7	0.03771	0.03911	0.19997	0.25546
8	0.01948	0.03119	0.10239	0.10734
10	0.02028	0.02552	0.12871	0.27826

Table 12: *Table.6 showing the $MMSE_\lambda(\hat{\Phi}_{S_n new})$for checking y-equivariance and regression equivariance*

	p=5		p=30	
λ	δ=10%	δ=20%	δ=10%	δ=20%
0.5	0.00319	0.00216	0.00067	0.00276
1	0.00145	0.00173	0.00013	0.01238
1.5	0.00102	0.00814	0.00039	0.01591
2	0.00056	0.00139	0.00039	0.00005
3	0.00109	0.00135	0.00050	0.00073
5	0.00178	0.00106	0.00141	0.00077
7	0.00021	0.00020	0.00024	0.00038
8	0.00012	0.00062	0.00064	0.00082
10	0.00011	0.00000	0.00438	0.00034

Table 13: *Table 7 showing output of fluid data without outlier*

	S_n	OLS	MM	S	LMS
β_0	1.1056	1.9304	1.8018	1.7995	2.0497
β_1	-0.1617	-0.1614	-0.1036	-0.0899	-0.2098
β_2	0.5713	0.5711	0.6274	0.6297	0.5077
β_3	-2.1968	-2.1954	-2.2016	-2.2424	-2.2676

Table 14: *Table 8 showing output of fluid data with outliers*

	S_n	OLS	MM	S	LMS
β_0	1.1109	0.6438	1.95204	2.0321	1.7494
β_1	-0.1547	0.0015	-0.1748	-0.2001	-0.1914
β_2	0.5688	0.9332	0.5556	0.5163	0.6315
β_3	-2.1839	0.2489	-2.2495	-2.2708	-1.7389

Table 15: *Table 9 showing results on Belgian Phone Call data*

Method	Intercept	Coefficient of X	MSE
OLS	58.566	0.587	78.1372
S_n	2.986	6.062	5001.829
MM	47.931	8.831	10041.74
S	48.060	8.894	10228.44
LTS	47.769	9.094	10717.14
LMS	48.439	8.658	9674.762
M	57.412	0.626	92.58984

References

[1] Areekara, S., Mackolil, J., Mahanthesh, B., Mathew, A., and Rana, P. (2022). A study on nanoliquid flow with irregular heat source and realistic boundary conditions: A modified Buongiorno model for biomedical applications.*ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift f?r Angewandte Mathematik und Mechanik*, 102, e202100167.

[2] Cabana Garceran del Vall, E., Lillo Rodriguez, R. E., and Laniado Rodas, H. (2019). Shrinkage reweighted regression.

[3] Hodges Jr, J. L. (1967, January). Efficiency in normal samples and tolerance of extreme values for some estimates of location.In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 163–186).

[4] Donoho, D. L., and Huber, P. J. (1983).The Notion of BreakdownPoint. *A festschrift for Erich L. Lehmann*157184.

[5] Hampel, F. R. (1971).A general qualitative definition of robustness. *The annals of mathematical statistics*, 42(6), 1887–1896.

[6] Rousseeuw, P. J., and Leroy, A. M. (2005).*Robust regression and outlier detection*. John wiley & sons.

[7] Maronna, R. A., and Zamar, R. H. (2002).Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*,44(4), 307–317.

[8] Maronna, R. and Morgenthaler, S. (1986). Robust regression through robust covariances.*Communications in Statistics-Theory and Methods*, 15(4), 1347–1365.

[9] Kunjunni, S. O., and Abraham, S. T. (2022).Multidimensional outlier detection and robust estimation using Sn covariance.*Communications in Statistics-Simulation and Computation*, 51(7), 3912–3922.

[10] Sajesh, T. A., and Srinivasan, M. R. (2012).Outlier detection for high dimensional data using the Comedian approach.*Journal of Statistical Computation and Simulation*, 82(5), 745–757.

[11] Siegel, A. F. (1982). Robust regression using repeated medians.*Biometrika, 69*(1), 242–244.

[12] Yohai, V. J. (1987).High breakdown-point and high efficiency robust estimates for regression.*The Annals of statistics*, 642–656.