




A hard re-descending hybrid robust regression estimation technique using direct weights

Greeshmagiri , Thangavel Palanisamy* 

Department of Mathematics, Amrita School of Physical Sciences, Coimbatore, Amrita Vishwa Vidyapeetham, India

Abstract

A hybrid approach of M and R estimators using an iterative procedure is proposed to detect outliers and estimation of regression parameters for linear models. We consider the deviation of each residual from its median to measure the likelihood of the corresponding data point to be an outlier. Also, the proposed work develops a reliable algorithm to estimate parameters of regression model that is unaffected by outliers. The significance of the proposed work is a novel hybrid approach of weighing the observations based on the order of residuals and is computationally simpler. Our proposal is illustrated using Monte Carlo simulation and analysed for few empirical benchmark data sets.

Mathematics Subject Classification (2020). 62G32, 62M10, 62J05

Keywords. Robust regression, hybrid estimator, M-estimator, R-estimator, outlier detection, Monte Carlo simulation

1. Introduction

The classical Ordinary Least Square (OLS) method which minimizes the sum of squares of residuals for the estimation of regression coefficients, has many drawbacks in case of a contaminated dataset. Though it is been used in various real-time applications [2, 22], one flaw that should be noted in the context of outlier detection is that each observation in the data is given equal weight. The literature on the estimation of the parameters in linear regression model reveals that efforts were made to consider observations depending on the magnitude of the residuals. By replacing square of errors with absolute errors, F. Y Edgeworth [6, 23] proposed an estimation method in which the outliers were weighed much lesser than that of OLS estimator. In addition, based on the performance indicators of the estimator, such as breakdown point, efficiency, and computational simplicity, a large number of approaches have emerged which are M, R, MM, Generalized M (GM), S, Least Median Square (LMS) and Least Trimmed Square (LTS) estimators. However, these efforts can be majorly classified either as M-estimation technique by weighing the square of residuals or as R-estimation technique [14, 15] by minimizing the dispersion of residuals using a suitable linear combination. The weights in M-estimation technique are obtained

*Corresponding Author.

Email addresses: g_greshma@cb.students.amrita.edu (Greeshmagiri), t_palanisamy@cb.amrita.edu (T. Palanisamy)

Received: 31.10.2023; Accepted: 05.09.2024

by minimizing an arbitrary function of residuals. The coefficients of the linear combination in R-estimation, is given by a non-decreasing score function on ordered residuals. Rather than the estimation of regression coefficients, anomaly detection is also given importance in these methods, as the presence of unusual observations may badly affect the actual fit. In fact, these robust regression techniques were established under this circumstance to reduce the impact of influential observations to an extent and to reveal the model of the clean data. A review and comparison of these methods was done by Chun Yu et al. in the year 2015 [26]. Robust regression procedures for outlier detection in both low and high dimensional datasets were broadly analysed in the works of Rousseeuw et al. [19–21]. In 2020, a review of outlier detection by robust regression was done by Getnet et al. [4] and reviews confined only to the M-estimators were done by Menezes et al. [5] and Khan et al. in 2021 [17].

The improved computational facility blended with the rich statistical tools facilitated the research community to address the surge in the problem of outliers in various domains. This inspired the authors to suggest an estimation method that incorporates outlier detection using a hybrid strategy by taking into account the benefits of R and M estimators. The proposed work implements a down weighing algorithm as in M-estimator but by ordering the absolute deviation of residuals from its median similar to R-estimator.

In Section 2, the proposed weight function along with the estimation procedure is broadly discussed. Section 3 discusses the optimization of threshold used in the proposed weight function. In Section 4, the proposed method is analyzed and evaluated using Monte Carlo simulation and few significant empirical data. The article concludes the study in Section 5.

2. Proposed estimator

The proposed work aims at the estimation of parameters in linear regression model which are not affected by the outlying observations in the data set and the detection of outliers as well. For this purpose, the authors propose a new weight function, which is hard re-descendant in nature motivated by the performance of Andrew's wave function. The proposed approach is hybrid combining M and R estimation techniques in the sense that it is accomplished by down weighing after ordering the absolute deviation of the residuals from its median. However, the very purpose of robust estimation either M or R, is to lessen the influence of the observations corresponding to the residuals of extreme magnitudes over the estimation of the parameters of the model. Therefore, the normal density curve of the residual argument is a natural choice to generate weights for the observations. Of course, any existing M-estimator can be observed to have weight function that is either a fine or coarse approximation of normal curve as depicted in Figure 1. Although it might slightly vary depending on the datasets selected, the weight function shown in Figure 1 actually refers to a specific dataset. Also, the weight functions given by Andrew's and Tukey's are serving the purpose better among the other M-estimators for robustness and closer geometrically to normal as well. Therefore, the authors attempt a weight function for the proposed robust estimation which is also an approximation of normal density.

The procedure begins with the initial parameter estimates obtained from OLS. The residual, r resulted by this set of parameters which is then scaled and acts as the argument to the proposed weight function. The median by its definition is robust against the presence of outliers in any set of observations. Therefore, we order the observations based on the distance of the residuals from its median $\tilde{\mu}_z$ which when exceeds a fixed threshold value are identified as outliers. This threshold value t , depends on the data [7, 24] under consideration which leads to higher efficiency. The weight function, w is defined as,

$$w(z) = \begin{cases} \frac{(t-z)^2}{\sigma_z \sqrt{2\pi}} e^{-\frac{(z-\tilde{\mu}_z)^2}{2\sigma_z^2}} & \text{for } |z - \tilde{\mu}_z| \leq t \\ 0 & \text{for } |z - \tilde{\mu}_z| > t \end{cases} \quad (2.1)$$

where $z = \frac{r}{s}$ in which r is the residual and $s = \frac{\tilde{\mu}_{|r-\tilde{\mu}_r|}}{0.6745}$, t is defined by $t = \mu_{|d|} * \frac{\pi}{2}$ with $\mu_{|d|}$, the mean absolute deviation d of z .

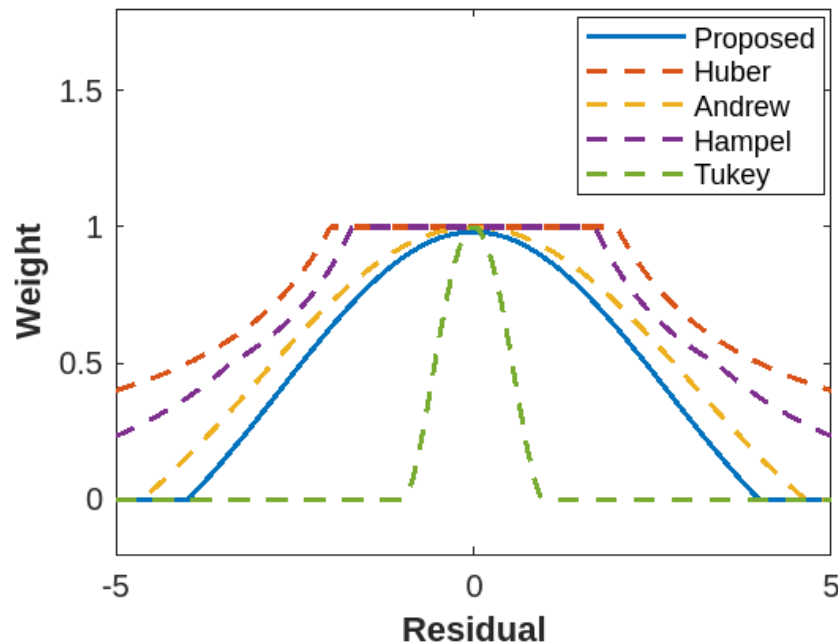


Figure 1. Weight functions graphs.

The new set of parameters are obtained using Iterative Re-weighted Least Square method (IRLS) [11] subject to a tolerance criterion. In fact, IRLS determine weights for the square of residuals iteratively by ordering the absolute deviation of residuals from its median in every iteration as used in R estimator. The proposed method is expected to give a relatively simpler weight function to identify the outliers in fewer number of iterations when compared to the conventional M-estimators. Moreover, the resulting model should be robust against the presence of outliers. The performance about the robustness of the proposed estimator is analyzed considering breakdown point and efficiency. The analysis reveals that the breakdown point is found to be more than 50% and the efficiency is as good as the well-known M-estimators. A comparative study has been done between the proposed hybrid estimator and existing estimators such as Huber t function [12], Andrews wave function [1], Hampels 17A function [8], Tukey's Bi-weight function [3], Least Weighted Square (LWS) [16] and L1-norm [6]. Huber - t function is a soft redescender which weigh the outlying observations lesser but the latter mentioned M-estimators are hard redescenders in the sense that the weight equals zero for sufficiently large absolute scaled residuals, $|z|$. Thus the proposed estimation technique will also fall in the category of hard re-descending estimators.

3. Threshold optimization

In this section, the justification for the choice of multiple $\frac{\pi}{2} \approx 1.57$ of mean absolute deviation of residuals, involved in the threshold t , is been provided. The authors feel it is essential to dispense this before the simulation study. We present an analysis by choosing various values, m instead of $\frac{\pi}{2}$ by brute force method. The probability distribution for the

choice of t is also picturized as a histogram of t versus number of data sets. It is clearly visible from the Table 1 that the perfect detection and swamping probability is attained when $t = \mu_{|d|} * \frac{\pi}{2}$. It is also revealed from Figure 2 that the probability distribution of t is approximately normal.

Table 1. Threshold optimization.

m	Detection probability	Swamping probability
1	1	0.222
1.1	1	0.178
1.2	1	0.154
1.3	1	0.114
1.4	1	0.106
1.5	1	0.048
1.51	1	0.046
1.52	1	0.046
1.53	1	0.046
1.54	1	0.044
1.55	1	0.044
1.56	1	0.044
1.57	1	0.034
1.58	1	0.044
1.59	1	0.052
1.6	1	0.052
1.7	1	0.049
1.8	0.998	0.046
1.9	0.996	0.046
2	0.988	0.044
3	0.978	0
4	0.168	0

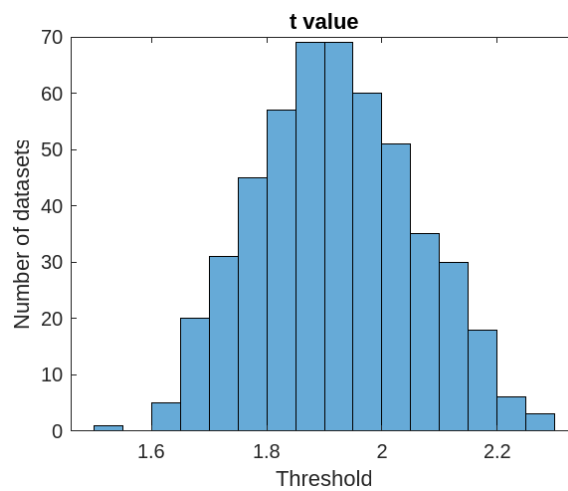


Figure 2. Threshold distribution.

4. Simulation study

To evaluate the performance of the proposed estimator, the algorithm has been tested using Monte Carlo simulation and also on few empirical benchmark data sets such as delivery time data, stack loss data and Hawkins Bradu and Kass (HBK) data [9, 18, 19].

The detection and swamping probabilities of the proposed estimator are investigated and compared with the competent estimation methods throughout our Monte Carlo simulation. A study is also carried out on the breakdown point of the estimator in our simulation and efficiency on the empirical benchmark data sets. We have also investigated the sensitivity due to hyper-parameters of our study. However, the results of the research analysis are found to be consistent with all the considered data sets.

We initially estimated the parameters of the model by OLS method which provides the residual for further iterative procedure. The residuals thus obtained (\mathbf{r}) are scaled using \mathbf{s} which results in $\mathbf{z} = \frac{\mathbf{r}}{\mathbf{s}}$. These scaled residuals are arranged appropriately to obtain its median. A threshold, \mathbf{t} is computed as discussed in Section 3, to identify the residuals having higher deviation from its median. Then, the scaled residuals are used to compute weights as per definition of the weight function given in (2.1). The initial estimation of the parameters is improved using these weights by weighted least square method. We iterate the algorithm until the largest change in any coefficient of estimation is less than 0.1%. It is clear to see that by implicitly ranking and re-weighting the residuals, the aforementioned estimate algorithm combine elements of the M and R techniques. The above sequential algorithm of IRLS is implemented using MATLAB 2023b. Those observations with residuals to median distance higher than the prescribed threshold are considered to be the outliers and their low weights play a key role in the robustness of estimates of the proposed method.

This section also includes the presentation of various results of the simulation study. Table 2 to Table 6 display the results about the Monte Carlo study. The estimated parameter values of the empirical benchmark data sets by the proposed hybrid method and other different M estimators in addition to OLS estimator excluding the outliers are provided in Table 7, Table 8 and Table 9. Additionally, a performance indicator known as the statistical efficiency, defined as the ratio of the residual mean square (RMS) of OLS for the data without outliers and that of the suggested method for the entire data is calculated. The efficiency and the RMS of the proposed method for all the empirical data sets in the simulation study is depicted in Table 10. The comparison of the efficiency of the proposed method with the other methods is given in Table 11. The empirical study also includes the plot of observations vs residual to median distance, say, \mathbf{d}_M in Figure 3, Figure 4 and Figure 5 to visualize clearly the outlying points. The results of the study on the breakdown point and sensitivity analysis is depicted in Figure 6 and Figure 7.

4.1. Monte Carlo simulation

In this section, Monte Carlo simulation is used to assess the performance of the proposed method. Also, a comparison on the performance of the proposed method with that of the existing better performing estimators is done for which results are presented as in [25]. Therefore, it has become essential to proceed with specifications described in [25]. In addition, we also consider Bi-square, L1-norm and a recent algorithm Least Weighted Square (LWS) [16] for comparison. For this purpose we examine linear regression models with number of regressors, k as 2 and 6 and with number of observations n as 40 and 60 respectively. For this purpose, 500 data sets are generated for each case randomly. The simulation study in the case of clean observations uses regressors following multivariate normal distribution with a mean vector of suitable dimension having 7.5 as each component and a suitable order scalar matrix for covariance with the diagonal entries 16. The corresponding responses for the above scenarios are obtained using the multiple linear regression model with constant term as zero and other coefficients as five, added to the error following standard normal distribution. It is obvious to understand the necessity of implanting outliers of various kind, the details of which will be discussed in each scenario.

The proposed study using Monte Carlo simulation is classified based on the outlier location in interior X space and exterior X space. The performance of the proposed method is also studied by varying the amount of outliers say, 10% and 20%. In all such above cases, we evaluate and compare the proportion of outlier detection and swamping. The discussion is made on all considered cases of interior X space and exterior X space and the results are presented in tables in the following sections.

We first consider outlying observations whose regressors are falling well within the bandwidth of the rest whereas the responses alone are deviating away. This is implemented by classifying the interior X space regression outlier cases into three sub cases: 1. Isolated outliers at random positions, 2. Outliers as cloud near to the centroid of X space and 3. Outliers as cloud away from the centroid of X space. Further, the deviations of the responses are executed by adding a multiple of the standard deviation σ of the error say $c\sigma$ denoted as δ_R , where $c = 3, 4, 5$ to the response generated by the model. The results of these three cases on interior X space are presented in Table 2, Table 3 and Table 4 respectively.

In the study of isolated outliers at random places, the simulation reveals the detection probability of the proposed method is nearly perfect for higher deviations. The swamping probability for most of the higher contaminated cases are found to be better in comparison with the considered estimators. Further the average swamping probability by considering all the different combinations is better except MM estimator.

Table 2. Random outliers.

n	k	contamination %	δ_R	Huber		Tukey's		LWS		MM estimator		Proposed Estimator	
				Detection	Swamping	Detection	Swamping	Detection	Swamping	Detection	Swamping	Detection	Swamping
40	2	10	3σ	0.978	0.061	0.982	0.302	0.984	0.34	0.989	0.060	0.972	0.124
60	6	10	3σ	0.910	0.058	0.964	0.032	0.95	0.338	0.918	0.057	0.87	0.176
40	2	20	3σ	0.890	0.090	0.956	0.296	0.996	0.262	0.938	0.085	0.812	0.064
0	6	20	3σ	0.740	0.092	0.864	0.268	0.852	0.338	0.770	0.088	0.478	0.104
40	2	10	4σ	1	0.058	1	0.328	0.998	0.292	1	0.058	0.996	0.072
0	6	10	4σ	0.997	0.054	0.988	0.338	0.99	0.308	0.998	0.054	0.988	0.118
0	2	20	4σ	0.987	0.096	0.996	0.322	0.998	0.298	0.993	0.086	0.978	0.03
60	6	20	4σ	0.952	0.109	0.984	0.34	0.976	0.272	0.961	0.099	0.94	0.044
40	2	10	5σ	1	0.055	1	0.336	1	0.326	1	0.053	0.998	0.062
60	6	10	5σ	1	0.050	0.998	0.374	0.998	0.334	1	0.049	1	0.078
40	2	20	5σ	0.998	0.088	0.998	0.364	1	0.312	1	0.074	1	0.012
60	6	20	5σ	0.985	0.113	1	0.302	0.998	0.348	0.993	0.092	0.994	0.018
Average probability				0.953	0.077	0.978	0.324	0.978	0.314	0.963	0.071	0.919	0.075

In the case of outliers as cloud near to the centroid of the X space and away from it, we consider outliers as a single cloud and two clouds separately. The outliers for single cloud are implanted as three different cases by adding a multiple of the standard deviation σ of the error say $c\sigma$ denoted as δ_R , where $c = 3, 4, 5$ to the response generated by the model whereas for two clouded outliers, it is done by adding and subtracting δ_R .

The simulation study for outliers near to the centroid of X space, reveals that the average detection probability of the proposed method is perfect and same as the estimators under consideration. However, the swamping probability is the same as MM estimator and higher than the other estimators. It is observed from the Table 3 that the detection and swamping probability for highly contaminated two clouded cases are found to be the same as Tukey's estimator and better than the other estimators considered.

Table 3. Outliers near to centroid of X space.

n	k	contamination %	δ_R	Cloud	Huber		Tukey's		LWS		MM estimator		Proposed Estimator	
					Detection	Swamping	Detection	Swamping	Detection	Swamping	Detection	Swamping	Detection	Swamping
40	2	10	3σ	1	1	0.06	1	0.352	1	0.292	1	0.059	1	0.12
60	6	10	3σ	1	1	0.055	1	0.36	1	0.282	1	0.056	1	0.158
40	2	20	3σ	1	1	0.087	1	0.326	1	0.286	1	0.083	1	0.074
60	6	20	3σ	1	1	0.082	1	0	1	0	1	0.079	1	0
40	2	10	4σ	1	1	0.09	1	0.352	1	0.3	1	0.055	1	0.076
60	6	10	4σ	1	1	0.042	1	0.35	1	0.296	1	0.052	1	0.096
40	2	20	4σ	1	1	0.05	1	0.33	1	0.296	1	0.083	1	0.03
60	6	20	4σ	1	1	0.058	1	0	1	0	1	0.081	1	0
40	2	10	3σ	2	1	0.052	1	0.308	0	0.302	1	0.052	1	0.148
60	6	10	3σ	2	1	0.051	1	0.33	0	0.248	1	0.050	1	0.16
40	2	20	3σ	2	1	0.052	1	0.332	0	0.288	1	0.049	1	0.056
60	6	20	3σ	2	1	0.09	1	0	0	0	1	0.045	1	0
40	2	10	4σ	2	1	0.052	1	0.348	0	0.302	1	0.052	1	0.09
60	6	10	4σ	2	1	0.050	1	0.35	0	0.278	1	0.049	1	0.124
40	2	20	4σ	2	1	0.051	1	0.308	0	0.306	1	0.050	1	0.034
60	6	20	4σ	2	1	0.045	1	0	0	0	1	0.044	1	0
Average probability					1	0.058	1	0.252	0.478	0.219	1	0.059	1	0.059

The simulation study for outliers away from the centroid of X space, reveals that the average detection and swamping probabilities are not significantly different from Tukey's and MM estimators respectively. However, results in Table 4 shows that the proposed method yields better results for the other estimators considered.

Table 4. Outliers away from centroid.

n	k	contamination %	δ_R	Cloud	Huber		Tukey's		LWS		MM estimator		Proposed Estimator	
					Detection	Swamping	Detection	Swamping	Detection	Swamping	Detection	Swamping	Detection	Swamping
40	2	10	3σ	1	0.770	0.063	1	0.288	1	0.338	0.780	0.061	0.998	0.136
60	6	10	3σ	1	0.705	0.060	0.988	0.322	1	0.234	0.724	0.058	0.998	0.132
40	2	20	3σ	1	0.588	0.095	0.998	0.322	1	0.28	0.646	0.085	0.99	0.076
60	6	20	3σ	1	0.424	0.108	0.952	0.256	1	0.294	0.456	0.107	0.906	0.108
40	2	10	4σ	1	0.951	0.060	1	0.332	1	0.33	0.951	0.062	1	0.104
60	6	10	4σ	1	0.943	0.059	1	0.372	1	0.29	0.958	0.055	1	0.114
0	2	20	4σ	1	0.843	0.116	1	0.304	1	0.3	0.898	0.093	1	0.042
60	6	20	4σ	1	0.628	0.150	0.996	0.318	1	0.264	0.695	0.132	0.996	0.058
40	2	10	5σ	1	0.946	0.058	1	0.31	1	0.358	0.996	0.057	1	0.068
60	6	10	5σ	1	0.994	0.057	1	0.288	1	0.276	0.994	0.050	1	0.086
40	2	20	5σ	1	0.953	0.133	1	0.324	1	0.296	0.982	0.083	1	0.016
60	6	20	5σ	1	0.803	0.189	0.988	0.342	1	0.274	0.875	0.138	1	0.028
40	2	10	3σ	2	0.790	0.061	1	0.344	0	0.284	0.797	0.060	1	0.136
60	6	10	3σ	2	0.773	0.056	0.996	0.32	0	0.33	0.776	0.055	0.992	0.172
40	2	20	3σ	2	0.661	0.085	0.998	0.366	0	0.4	0.709	0.080	0.978	0.076
60	6	20	3σ	2	0.561	0.095	0.968	0.326	0	0.314	0.598	0.089	0.878	0.086
40	2	10	4σ	2	0.960	0.060	0.988	0.346	0	0.302	0.960	0.060	1	0.082
60	6	10	4σ	2	0.961	0.053	1	0.34	0	0.272	0.960	0.053	1	0.134
40	2	20	4σ	2	0.899	0.097	1	0.318	0	0.298	0.933	0.083	0.998	0.044
60	6	20	4σ	2	0.811	0.121	0.996	0.348	0	0.288	0.856	0.100	0.98	0.062
40	2	10	5σ	2	0.998	0.057	1	0.316	0	0.324	0.998	0.056	1	0.094
60	6	10	5σ	2	0.997	0.049	1	0.37	0	0.302	0.999	0.049	1	0.1
40	2	20	5σ	2	0.984	0.094	1	0.334	0	0.322	0.994	0.073	1	0.01
60	6	20	5σ	2	0.941	0.137	1	0.314	0	0.306	0.965	0.092	0.994	0.008
Average probability					0.829	0.088	0.994	0.326	0.478	0.303	0.854	0.076	0.987	0.082

Next, we consider outlying observations where both regressors are deviating away with no restrictions on the responses. This is implemented by classifying the exterior X space regression outlier cases into two sub cases: 1. Exterior X space and exterior Y space, 2. Exterior X space and interior Y space. In fact, the analysis made in the case of exterior X space by earlier research studies have used only GM estimators to compare their proposed estimate due to the vulnerability of M and MM estimators. The deviations of the responses are executed similar to the earlier study of interior X space. Moreover, the deviation of the regressors is carried out by adding a multiple of the standard deviation σ of the error say $c\sigma$ denoted as δ_L , where $c = 3, 5$ to the regressors generated by the model. The results of these two cases on exterior X space are presented in Table 5 and Table 6 respectively.

The results of the simulation study for outliers in the external X and Y spaces show that none of the approaches under consideration have very good average detection rates. Further, the swamping probability is reasonably good in case on GM and the proposed estimator. The results of various combinations are presented in Table 5.

Table 5. Exterior X space and exterior Y space.

n	k	contamination %	δ_R	δ_L	cloud	GM		Tukey's		LWS		Proposed Estimator	
						Detection	Swamping	Detection	Swamping	Detection	Swamping	Detection	Swamping
40	2	10	2σ	3σ	1	0.253	0.071	0.874	0.372	0.992	0.288	0.9	0.124
60	6	10	2σ	3σ	1	0	0.088	0.014	0.338	0.014	0.296	0	0.292
40	2	20	2σ	3σ	1	0	0.144	0	0.468	0.12	0.346	0	0.284
60	6	20	2σ	3σ	1	0	0.087	0	0.378	0	0.332	0	0.082
40	2	10	5σ	3σ	1	0	0.063	0	0.36	0	0.37	0.002	0.254
60	6	10	5σ	3σ	1	0	0.061	0	0.358	0	0.308	0	0.316
40	2	20	5σ	3σ	1	0	0.063	0	0.4	0	0.314	0	0.31
60	6	20	5σ	3σ	1	0	0.058	0	0.35	0	0.314	0	0.066
40	2	10	2σ	5σ	1	0.961	0.074	0.992	0.296	1	0.28	0.998	0.068
60	6	10	2σ	5σ	1	0.080	0.130	0.006	0.452	1	0.314	0.664	0.154
40	2	20	2σ	5σ	1	0.071	0.258	0.692	0.382	0.992	0.328	1	0.012
60	6	20	2σ	5σ	1	0	0.145	0	0.436	0	0.37	0	0.054
40	2	10	5σ	5σ	1	0	0.098	0	0.41	0.006	0.346	0.028	0.248
60	6	10	5σ	5σ	1	0	0.071	0	0.372	0	0.302	0	0.27
40	2	20	5σ	5σ	1	0	0.098	0	0.42	0	0.362	0	0.336
60	6	20	5σ	5σ	1	0	0.065	0	0.404	0	0.318	0	0.078
40	2	10	2σ	3σ	2	0.620	0.071	0.948	0.37	0.992	0.27	0.916	0.136
60	6	10	2σ	3σ	2	0.478	0.089	0.652	0.368	0.898	0.26	0.538	0.174
40	2	20	2σ	3σ	2	0	0.144	0.748	0.314	0.922	0.324	0.492	0.132
60	6	20	2σ	3σ	2	0.249	0.088	0.006	0.488	0.382	0.288	0	0.03
40	2	10	5σ	3σ	2	0.365	0.063	0.106	0.38	0.298	0.318	0.104	0.18
60	6	10	5σ	3σ	2	0.060	0.058	0	0.392	0	0.286	0	0.254
40	2	20	5σ	3σ	2	0.101	0.063	0	0.376	0	0.348	0	0.196
60	6	20	5σ	3σ	2	0.005	0.057	0	0.376	0	0.35	0	0.03
40	2	10	2σ	5σ	2	0.961	0.074	1	0.326	1	0.308	1	0.042
60	6	10	2σ	5σ	2	0.080	0.130	0.984	0.364	0.992	0.308	0.99	0.066
40	2	20	2σ	5σ	2	0.071	0.258	0.994	0.326	0.996	0.308	0.996	0.008
60	6	20	2σ	5σ	2	0	0.145	0.872	0.322	0.986	0.29	0	0.03
40	2	10	5σ	5σ	2	0.470	0.098	0.644	0.396	0.91	0.324	0.722	0.084
60	6	10	5σ	5σ	2	0.133	0.072	0	0.41	0	0.266	0	0.23
40	2	20	5σ	5σ	2	0.123	0.099	0	0.4	0	0.354	0	0.232
60	6	20	5σ	5σ	2	0.011	0.067	0	0.392	0	0.33	0	0.03
Average probability						0.200	0.099	0.304	0.378	0.391	0.317	0.292	0.154

From the simulation study for outliers in the exterior X space and interior Y space, we observe that the proposed method detects moderately. In fact, it is similar to Tukey's and LWS and is 90% more than GM. Despite the unimpressive swamping probabilities of Tukey's and LWS, the proposed method is reasonably good as GM. The results of various combinations are presented in Table 6.

Table 6. Exterior X space and interior Y space.

n	k	contamination %	δ_R	δ_L	cloud	GM		Tukey's		LWS		Proposed Estimator	
						Detection	Swamping	Detection	Swamping	Detection	Swamping	Detection	Swamping
40	2	10	2σ	5σ	2	0.980	0.074	0.998	0.308	1	0.28	1	0.08
60	6	10	2σ	5σ	1	0.998	0.308	1	0.332	1	0.294	0.976	0.09
40	2	20	5σ	5σ	1	0	0.100	0	0.374	0	0.314	0	0.308
60	6	20	5σ	5σ	2	0.011	0.068	0	0.344	0.006	0.282	0	0.23
Average probability						0.261	0.096	0.499	0.339	0.502	0.293	0.494	0.177

Summarizing the performance, we observe that none of the existing methods emerges as the best in all the possible scenarios considered. It is significant to observe that LWS works extremely poor in many cases though it is reliable in few cases. The proposed method is detecting outliers as good as Tukey's algorithm in most of the cases. But, when both swamping and detection probabilities are taken into account, the suggested approach appears to be better, if previous knowledge about the data is not accessible. Further, the conventional detection metrics is used in L1-norm as it is not involving weights to obtain the detection and swamping. Since the results are so dismal, we have excluded L1-norm from the comparison tables, Table 2 to Table 11.

4.2. Empirical study

Our study investigates the most significant empirical data sets that are often used by research fraternity in the evaluation of robustness of regression models. They are Hawkins Bradu and Kass (HBK) data, stack loss data and delivery time data [9, 18, 19].

HBK data set is an artificially constructed data set with 14 outliers. This data set contains both good and bad leverage points where the existing regression methods commonly identify either of them or both types. The variable (y) depends on three independent variables x_1 , x_2 and x_3 . The parameter estimates obtained from OLS method for data without outliers and those obtained from Huber, Hampel, Andrew, Tukey and the new estimator for observations with outliers are listed in Table 7.

Table 7. Estimated coefficients - HBK data.

Coefficients	OLS	Huber	Hampel	Andrew	Tukey	Proposed Estimator
β_0	-0.1805	-0.6933	-0.1806	-0.1846	-0.5342	-0.1584
β_1	0.0814	0.1869	0.1815	0.0831	0.2329	0.0878
β_2	0.0399	-0.0988	0.0398	0.0404	0.0495	0.0405
β_3	-0.0517	0.3265	-0.0515	-0.0999	-0.0526	-0.0549

Figure 3 gives the observation against d_M plot where the horizontal line represents the threshold, t . It is interesting to note that the ten observations corresponding to the residuals falling above the threshold are the same as those identified as the bad outliers by various research studies.

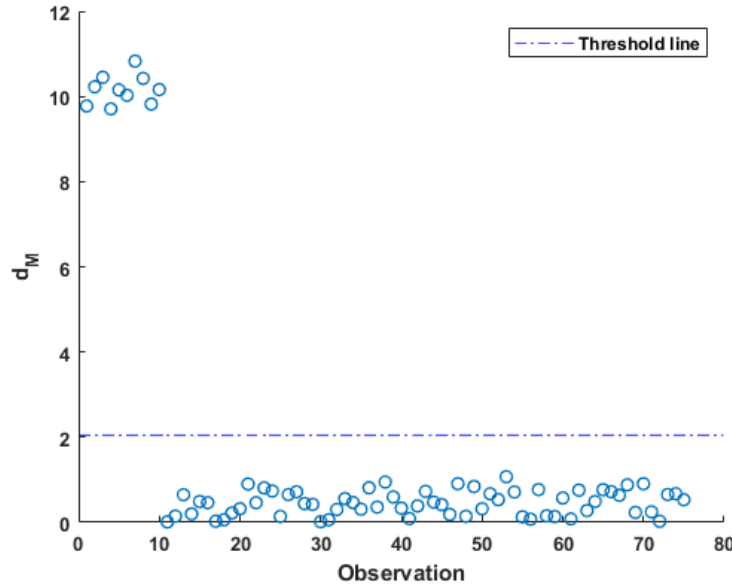


Figure 3. Visualization of outliers - HBK data.

Stack loss data consists of 21 observations which provides details of amount of ammonia lost (y) by the vending machines related to the three factors, air flow to the plant (x_1), cooling water inlet temperature (x_2) and acid concentration (x_3). The observations 1, 3, 4 and 21 are found to be the outliers. The parameter estimates obtained from OLS method for data without outliers and those obtained from Huber, Hampel, Andrew, Tukey and the proposed estimator for observations with outliers are listed in Table 8.

Table 8. Estimated coefficients - Stack loss data.

Coefficients	OLS	Huber	Hampel	Andrew	Tukey	Proposed Estimator
β_0	-37.6525	-40.6142	-39.9214	-37.2154	-37.4100	-36.8297
β_1	0.7977	0.7419	0.7162	0.8262	0.7200	0.8320
β_2	0.5773	1.2035	1.2953	0.5279	0.9636	0.4819
β_3	-0.0671	-0.1408	-0.1526	-0.0715	-0.1102	-0.0758

Figure 4 gives the observation against d_M plot where the horizontal line represents the threshold, t . All the outliers have been detected which lies above the threshold line.

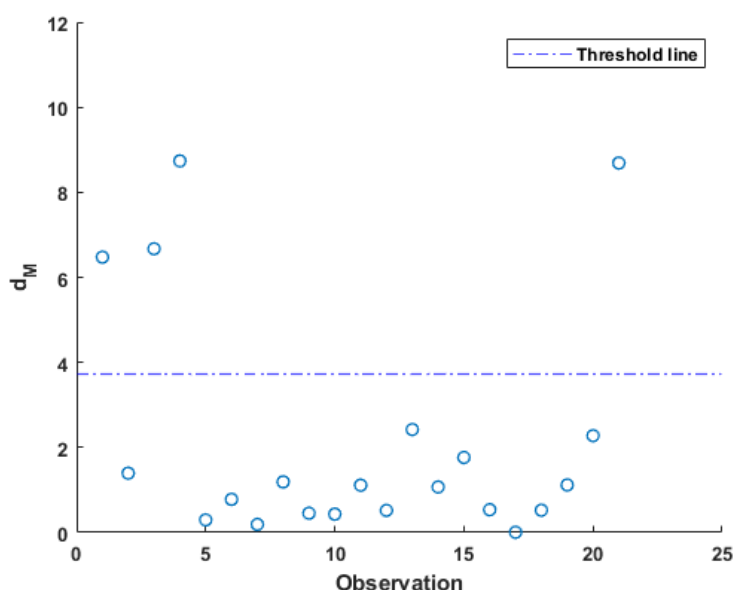


Figure 4. Visualization of outliers - Stack loss data.

Delivery time data is related to the two factors, the number of cases of products stocked (x_1) and the distance walked by the route driver (x_2) affecting the delivery time (y) of vending machine. Among the 25 observations, 1, 4, 9, 20, 22, 23 and 24 are found to be the outliers. The parameter estimates obtained from OLS method for data without outliers and those obtained from Huber, Hampel, Andrew, Tukey and the proposed estimator for observations with outliers are listed in Table 9

Table 9. Estimated coefficients - Delivery time data.

Coefficients	OLS	Huber	Hampel	Andrew	Tukey	Proposed Estimator
β_0	3.719	3.327	3.864	4.467	3.3521	3.6755
β_1	1.406	1.529	1.427	1.463	1.4371	1.3471
β_2	0.016	0.014	0.014	0.011	0.0144	0.016

Figure 5 gives the observation vs d_M plot where the horizontal line represents the threshold. The seven observations corresponding to the residuals falling above the threshold line is considered to be the outliers.

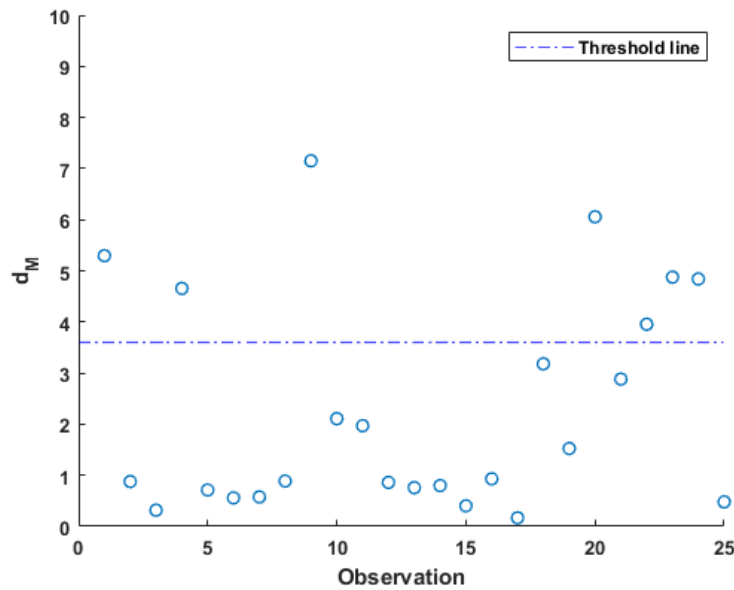


Figure 5. Visualization of outliers - Delivery time data.

4.3. Performance analysis

In this subsection, we present the observations from the investigation about the breakdown point and efficiency to measure the robustness of the proposed estimator.

4.3.1. Breakdown point. Breakdown point is a significant property that has practical concern while devising a robust estimator. The definition of it dates back to the initial studies of robustness and differ slightly over different researchers. We consider in our study the general definition [13, 18] for finite sample breakdown point for an estimator T over a sample X with n observations given by,

$$\epsilon_n^*(T, X) = \min \left\{ \frac{m}{n} : \sup_m D(T(X), T(X')) = \infty \right\} \quad (4.1)$$

where $m = 1, 2, \dots, n$ represents the amount of contamination in the dataset and D is any appropriate distance measure between the clean dataset X and the contaminated dataset X' . For our study we choose D as the root mean square of the residuals of the estimation $T(X)$ and $T(X')$ respectively.

The early attempts on robustness by the M estimators, L and R estimators could not improve the breakdown point over least square estimators as expected. However, the later studies focused on improving it, could accomplish upto 50% but only few studies achieved simultaneously with good efficiency [10].

We investigate the breakdown point for the proposed method. For this purpose, we carried out simulation iteratively considering various number of observations. In each case, the breakdown point is computed using (4.1) taking number of outliers m , ranging from 1 to n . Despite the number of iterations, we depict one with $n = 100$ for the visualization of the breakdown point in the Figure 6 and is evident that it is more than 50%.

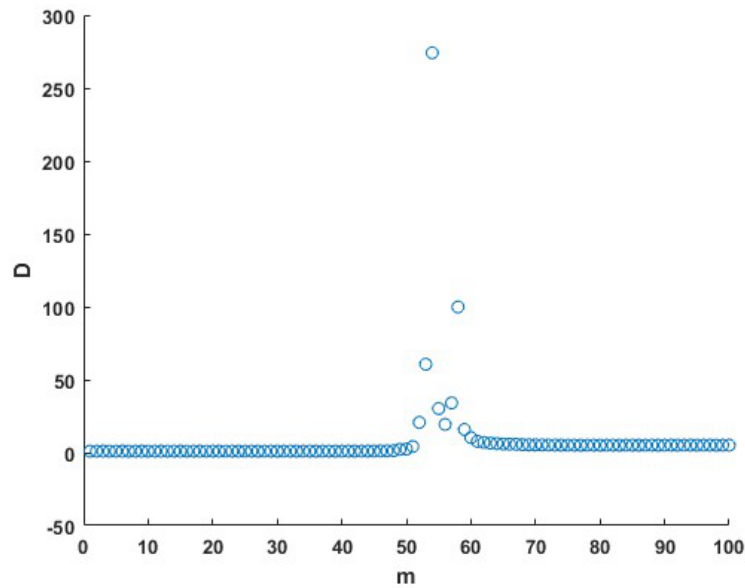


Figure 6. Breakdown point.

4.3.2. Statistical efficiency. We recall the definition of the efficiency E given in Section 4 which is understood to be the best when it is close to unity. We compute it for the proposed method using,

$$E = \frac{RMS_{OLS}}{RMS_{Robust}} = \frac{\sum_{i=1}^n r(ols)_i^2/n}{\sum_{i=1}^n r(robust)_i^2/n} \tag{4.2}$$

Thus the efficiency measure, E is expected to be close to unity. The efficiency of the new estimator for three different data sets are shown in Table 10.

Table 10. RMS and efficiency.

Dataset	(OLS)	RMS(Proposed Estimator)	Efficiency
Delivery Time data	11.082	11.036	1.004
Stack Loss data	11.499	11.772	0.981
HBK data	13.905	13.898	1.001

Nevertheless, the simplicity of the computation, the efficiency of the proposed robust estimation technique performs on par with many of the well known existing M-estimation approaches. A comparative study is been presented in Table 11 with the efficiency of the proposed estimator along with that of the estimators under consideration.

Table 11. Comparison of efficiency.

Data Set	Efficiency						
	Huber	Hampel	Andrew	Tukey	LWS	L1-Norm	Proposed Estimator
Delivery time data	1.185	0.885	0.983	1.004	2.664	1.339	1.004
Stack loss data	1.197	0.979	0.985	0.848	1.674	139.45	0.981
HBK data	2.103	1.170	0.999	0.983	1.061	0.356	1.001

4.4. Sensitivity analysis

It is understood that there are three hyper-parameters affecting the results of our study. One among them is the multiple of mean absolute deviation of residuals involved in the calculation of threshold t which will alter the detection and swamping probabilities. The second is the number of outliers in a given dataset which will affect the mean square error of residuals. And if we could define a distance metric in p -dimensional space, the number of outliers will influence the distance given by L2-norm, denoted by N , between the parameter vectors of the dataset involving different number of outliers and the respective dataset without outliers.

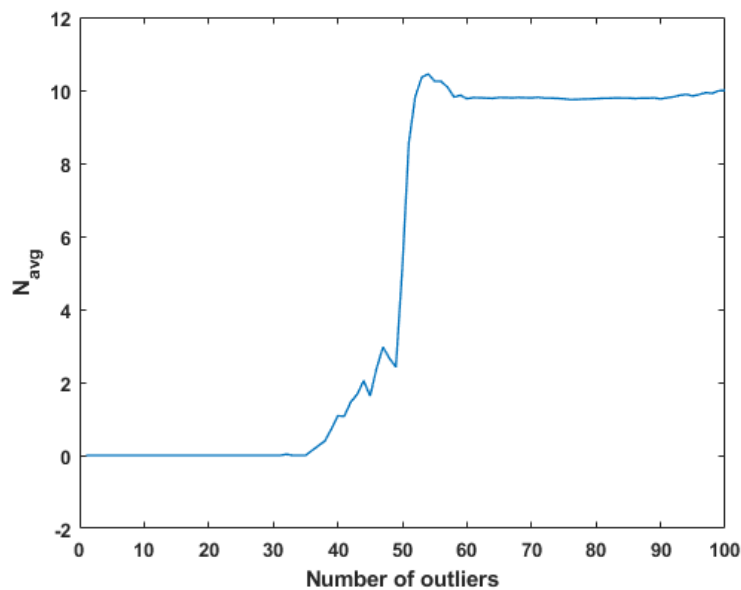


Figure 7. Sensitivity analysis on coefficient estimates.

The constant used as the multiple of mean absolute deviation of residuals in the threshold t , the first hyper parameter is highly sensitive in the sense that the lesser or more the constant than the chosen one will worsen the detection or swamping probabilities. Referring Table 1, 1.57 is observed to be the optimal value considered for the computation of the threshold. We have performed two different sensitivity analyses based on the second hyper parameter, the number of outliers. The sensitivity pertaining to root mean square error of the residuals of the second hyper parameter is depicted in Figure 6 and shows how well the model captures the information from the data when there is an increase in the number of outliers. The figure shows that till 50% of contamination, there is no significant increase in root mean square error of residuals. Further, the sensitivity due to the number of outliers is studied using the distance between the parameter vectors which is presented in Figure 7. The figure displays the average of distances N say, N_{avg} obtained by iterations. It is obvious that N is expected to be small but the figure reveals the distance is acceptably small till approximately 40% of the data gets contaminated.

5. Conclusion

A hybrid approach of M and R estimators using an iterative procedure is proposed to detect outliers and to estimate the regression parameters for linear models. The simplicity of the computation of the proposed method compared to M-estimators is that the weights are based on the order of the residuals instead of deriving it from an arbitrary function.

This algorithm down weights those observations corresponding to residuals far from its median. As the very purpose of any robust estimation is to lessen the influence of the observations corresponding to the residuals of extreme magnitudes, the normal density curve of the residual argument is a natural choice for a weight function. A Monte Carlo simulation and a study on different empirical data sets were conducted to analyze the performance of the proposed estimator. The detection and swamping of outliers is also well comparable with the considered estimators. It is also observed from the simulation study that the efficiency, a statistical measure of the proposed method is close to the better performing M-estimators given by Andrew and Tukey. It comes to light that the breakdown point, a critical statistic for the proposed method is found to be 50% and is as high as the maximum of the existing estimators. A sensitivity analysis was carried out based on two hyper parameters namely the number of outliers and the multiple used in the threshold value. It is evident from the analysis that the estimation is robust upto approximately 40% of contamination. The results are depicted clearly using tables and scatter plots. Nevertheless, the simplicity of the computation, the proposed hybrid approach of robust estimation technique performs on par with many of the well known existing approaches.

In our study, we consider errors following normal distribution with artificially implanted outliers. However, the proposed method can be extended to a model involving errors of mixed distributions [7, 24] and a suitable data driven tuning constant.

Acknowledgements

Authors expresses sincere gratitude to the editor and reviewers for their valuable suggestions aimed at enhancing the quality and presentation of the paper.

Author contributions. All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

Conflict of interest statement. The authors declare that they have no competing interests.

Funding. No funding was obtained for the current study.

Data availability. Not applicable.

References

- [1] D.F. Andrews, *A robust method for multiple linear regression*, Technometrics. **16** (4), 523-531, 1974.
- [2] R. Baby, C.S. Kumar, K.K. George and A. Panda, *Noise compensation in i-vector space using linear regression for robust speaker verification*, 2017 International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT). 161-165, 2017.
- [3] A.E. Beaton and J. WTukey, *The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data.*, Technometrics. **16** (2) 147-185, 1974.
- [4] G.B. Begashaw and Y.B. Yohannes , *Review of outlier detection and identifying using robust regression model*, International Journal of Systems Science and Applied Mathematics. **5** (1), 4-11, 2020.
- [5] D.Q.F. De Menezes, D.M. Prata, A.R. Secchi and J.C. Pinto, *TA review on robust M-estimators for regression analysis*, Comput. Chem. Eng. **147**, 107254. 2021.

- [6] F.Y. Edgeworth, *On observations relating to several quantities*, *Hermathena*. **6** (13), 279-285, 1887.
- [7] L. Fu, Y.G. Wang and F. Cai, *A working likelihood approach for robust regression*, *Stat. Methods Med. Res.* **29** (12), 3641-3652, 2020.
- [8] F.R. Hampel, *The influence curve and its role in robust estimation*, *J. Am. Stat. Assoc.* **69** (346), 383-393, 1974.
- [9] D.M. Hawkins and D. Bradu, *Location of several outliers in multiple-regression data using elemental sets*, *Technometrics*. **26** (3), 197-208, 1984.
- [10] S. Hekimolu and R.C. Erenoglu, *A new GM-estimate with high breakdown point*, *Acta Geod. Geophys.* **48**, 419-437, 2013.
- [11] P.W. Holland and R.E. Welsch, *Robust regression using iteratively reweighted least-squares*, *Commun. Stat. - Theory Methods* **6** (9), 813-827, 1977.
- [12] P.J. Huber and R.E. Welsch, *Robust regression: Asymptotics, conjectures and monte carlo*, *Ann. Stat.* **1** (5), 799-821, 1973.
- [13] M. Hubert and M. Debruyne, *Breakdown value*, *WIREs Comp Stat.* **1**, 296-302, 2009.
- [14] L.A. Jaeckel, *Estimating regression coefficients by minimizing the dispersion of the residuals*, *Ann. Math. Statist.* **43** (5), 1449-1458, 1972.
- [15] J. Jureckova, *Nonparametric estimate of regression coefficients*, *Ann. Math. Statist.* **42** (4), 1328-1338, 1971.
- [16] J. Kalina, *Regularized least weighted squares estimator in linear regression*, *Commun. Stat. - Simul. Comput.* 2024.
- [17] D.M. Khan, M. Ali, Z. Ahmad, S. Manzoor and S. Hussain, *A new efficient re-descending M-estimator for robust fitting of linear regression models in the presence of outliers*, *Math. Probl. Eng.* 3090537, 2021.
- [18] D.C. Montgomery, E.A. Peck and G.G. Vining, *Introduction to Linear Regression Analysis*, Fifth Edition, 2013.
- [19] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, 2005.
- [20] P.J. Rousseeuw and M. Hubert, *Robust statistics for outlier detection*, *WIREs Data Mining Knowl Discov.* **1**, 73-79, 2011.
- [21] P.J. Rousseeuw and M. Hubert, *Anomaly detection by robust statistics*, *WIREs Data Mining Knowl Discov.* **8**, e1236, 2018.
- [22] A. Shyna, C. UshadeviAmmu, A. John, C. Kesavadas and B. Thomas, *Deep-ASL enhancement technique in arterial spin labeling MRI - A novel approach for the error reduction of partial volume correction technique with linear regression algorithm*, *J. Comput. Sci.* **58**, 101546, 2022.
- [23] R. Tibshirani, *Regression shrinkage and selection via the lasso*, *J. R. Stat. Soc., B: Stat. Methodol.* **58** (1), 267-288, 1996.
- [24] Y.G. Wang, X. Lin, M. Zhu and Z. Bai, *Robust estimation using the huber function with a data-dependent tuning constant*, *J. Comput. Graph. Stat.* **16** (2), 468-481, 2007.
- [25] J.W. Wisnowski, D.C. Montgomery and J.R. Simpson, *A comparative analysis of multiple outlier detection procedures in the linear regression model*, *Comput. Stat. Data Anal.* **36** (3), 351-382, 2001.
- [26] C. Yu and W. Yao, *Robust linear regression: A review and comparison*, *Commun. Stat. - Simul. Comput.* **46** (8), 6261-6282, 2017.