International Journal of Scientific Research in Mathematical and Statistical Sciences **Research Paper**
Volume-5, Issue-5, pp.65-71, October (2018) **E-ISSN:** 2348-4519

Empirical Robust Multivariate Regression Parameter Estimation Using Median Approach

O.K Sajana[1*], T.A Sajesh[2]

[1]Department of Statistics, St Thomas' College (Autonomous), Thrissur District, Kerala, India
[2]Department of Statistics, St Thomas' College (Autonomous), Thrissur District, Kerala, India

[*]*Corresponding Author: sajana.kunjunni@gmail.com, Tel.: +919495809687*

Available online at: www.isroset.org

Received:04/Oct/2018, Accepted:14/Oct/2018, Online: 31/Oct/2018

Abstract— Main purpose of multivariate regression analysis is the estimation of model parameters. The use of maximum likelihood method would not be appropriate in estimation problems while data contains outlier or extreme observations. So it is necessary to find a parameter estimation method in which the value of the estimator is not much affected by small changes in the data. This paper introduces robust method for multivariate regression based on robust estimation of location and scatter matrix of predictor and response variables. In this paper Comedian method is taken as a robust estimator of location and scatter. Based on the simulations, the finite-sample efficiency and robustness of the estimator are investigated. Efficiency of proposed robust estimators is compared with maximum likelihood estimator, minimum covariance determinant estimator and orthogonalized Gnanadesikan-Kettenring estimator in terms of mean squared errors. Proposed estimator combines high robustness and high efficiency in estimation. The proposed method is illustrated on a real data set.

Keywords—*Multivariate Regression, Outliers Detection, Comedian Approach, Finite Sample Efficiency.*

I. Introduction

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between dependent variables (response) and independent variables (predictors). More specifically, regression analysis helps to understand how the typical value of the dependent variables changes when any one of the independent variables is varied. As outlined above multivariate regression model allows us to assess the impact of multiple variables on one or more dependent variable in the same model.

Consider the multivariate regression model

$$\mathbf{y_i} = \boldsymbol{\mathcal{B}}^{\mathrm{t}}\mathbf{x_i} + \boldsymbol{\alpha}_{\mathbf{i}} + \boldsymbol{\varepsilon}_{\mathbf{i}}\,, \quad i = 1, \dots, n \tag{1}$$

with $\mathbf{x_i} = (x_{i1}, \dots, x_{ip})^{t} \in \mathbf{R}^{p}$, $\mathbf{y_i} = (y_{i1}, \dots, y_{iq})^{t} \in \mathbf{R}^{q}$, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{iq})^{t} \in \mathbf{R}^{q}$ is the q-dimensional intercept vector and $\boldsymbol{\mathcal{B}} \in \mathbf{R}^{(p\times q)}$ is the $(p\times q)$ slope matrix. The error terms $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1} \ . \ . \ . \ \varepsilon_{iq})^{t}$ are independent and identically distributed random variables with center zero, positive definite and symmetric scatter matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ of size q.

Let us denote the location of the joint ($\boldsymbol{x}$, $\boldsymbol{y}$) variables by $\boldsymbol{\mu}$ and scatter matrix by $\boldsymbol{\Sigma}$. Partitioning $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ yields the notations:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix} \tag{2}$$

Traditionally, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are often estimated by the classical estimation procedures like Maximum Likelihood Method. Let $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ be the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Then the maximum likelihood estimators for $\boldsymbol{\mathcal{B}}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}_{\varepsilon}$ are given by

$$\hat{\boldsymbol{\mathcal{B}}} = \hat{\boldsymbol{\Sigma}}^{-1}_{xx} \tag{3}$$

$$\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\mu}}^{-1}_{y} - \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{t}}\hat{\boldsymbol{\mu}}_x \tag{4}$$

$$\hat{\boldsymbol{\Sigma}}_{\varepsilon} = \hat{\boldsymbol{\Sigma}}_{yy} - \hat{\boldsymbol{\mathcal{B}}}^{\mathrm{t}}\hat{\boldsymbol{\Sigma}}_{xx}\hat{\boldsymbol{\mathcal{B}}} \tag{5}$$

The expressions (3), (4) and (5) are directly depending on the estimates of the location vector and scatter matrix of response and predictor variable respectively. Unfortunately, classical estimators are not robust to the presence of outliers which are the observations in a data that appears to be inconsistent with the remainder of that data set [1]. Consequently, the classical regression techniques are extremely sensitive to the presence of outliers and provide misleading results. As a solution to this problem, one may replace classical estimates of location and scatter by highly robust estimates which are less sensitive to outliers and perform robust analysis. Many robust estimates have been proposed over the years with various properties [2].

An overview of robust multivariate regression techniques is explained in the context of simultaneous equation models by [3]. Application of M-estimator to each coordinate of the responses was investigated and suggested to minimizing the sum of the Euclidean norm of the residuals [4, 5]. But, these two methods are not affine equivariant. Robust estimation procedure of multivariate regression based on the sign covariance matrix was introduced by [6]. The application of Minimum Covariance Determinant (MCD) method which possesses breakdown value was discussed and developed the reweighted versions of Minimum Covariance Determinant estimator in multivariate regression estimation [7, 8]. The MCD methods provide better robust estimates of multivariate regression coefficients compared to many other methods. But for large dimensional data it is computationally expensive and time consuming. This article explores the possibilities of Comedian method, proposed for robust estimation of multivariate regression coefficients explained in [9]. Comedian estimates has high breakdown value and provide better results for large dimensional datasets. The efficiency of the proposed method is evaluated through simulation and the results are compared with that of MLE estimates, Orthogonalized Gnanadesikan-Kettenring (OGK) estimates of and MCD estimates [10].

In section II, robust method adapted for multivariate regression estimation with suitable threshold function is described. Section III consist results of simulated environment in terms of finite sample efficiencies. Section IV includes simulated robustness properties of proposed method. Application of proposed method in real life dataset explained in section V. The conclusion is presented in last section.

II. MATERIALS AND METHODS

Consider the data set **Z** = {$\mathbf{z}_i$; i=1,2,…,n}∈ $\mathbf{R}^{p+q}$ consisting of q response variables and p predictor variable each sample of size n. Then the comedian matrix **COM** (**Z**) is defined as

$$\boldsymbol{COM}(\boldsymbol{Z}) = (COM\ (\boldsymbol{Z_i}, \boldsymbol{Z_j})), i, j\ =\ 1, 2, .., p. \qquad (6)$$

Similarly, multivariate correlation median matrix **δ** is defined as,

$$\boldsymbol{\delta}(\boldsymbol{Z}) = \boldsymbol{DCOM}(\boldsymbol{Z})\boldsymbol{D^T} \qquad (7)$$

where $\boldsymbol{D}$ is a diagonal matrix with diagonal elements $1/MAD(\mathbf{Z}_i)$ (i = 1, …, p).

Consider a pxp matrix **E** whose columns are eigenvectors of **δ** (**Z**). Let **Q** = **D** (**Z**)$^{\mathbf{-1}}$ **E** and $\mathbf{w}_i$ = $\mathbf{Q}^{-1}\mathbf{z}_i$ (i = 1, 2, …, n). Then **W** is an orthogonalized matrix with rows $\boldsymbol{w}_i^{\mathbf{T}}$, ($i$= 1, …, n) and columns $\boldsymbol{W}_j$ (j = 1, …, p). The resulting robust estimates for location $\boldsymbol{\mu}$ and scatter **Σ** are then defined as

$$\boldsymbol{\Sigma_R} = \boldsymbol{Q\Gamma Q^T}\ and\ \boldsymbol{\mu_R} = \boldsymbol{Ql} \qquad (8)$$

where **Γ** = diag(MAD($\boldsymbol{W}_1$)^2, …, MAD($\boldsymbol{W}_p$)^2) and $\boldsymbol{l}$ = (med($\boldsymbol{W}_1$), …, med($\boldsymbol{W}_1$))$^{\boldsymbol{T}}$. The procedure can be iterated, computing $\mathbf{\Sigma_R}$ and $\boldsymbol{\mu}_\mathbf{R}$ for **W** and then expressing them in the original coordinate system. These estimates can be improved on by a reweighting step by using a robust Mahalanobis distance defined as,

$$RD(\boldsymbol{z_i}) = rd_i\ =\ (\boldsymbol{z_i} - \boldsymbol{\mu_R})^{\boldsymbol{T}}\ \boldsymbol{\Sigma_R^{-1}}\ (\boldsymbol{z_i} - \boldsymbol{\mu_R}),$$
$$i\ =\ 1, 2, \dots, n \qquad (9)$$

where $\mathbf{\Sigma_R}$ and $\boldsymbol{\mu}_\mathbf{R}$ are defined in (8). Let M be a weight function, and define $\mathbf{\Sigma_{RW}}$ and $\boldsymbol{\mu}_\mathbf{RW}$ as the weighted mean and covariance matrix, where each $\boldsymbol{z_i}$ has weight m_i = M (d_i), that is,

$$\mathbf{\Sigma_{RW}} = \frac{\sum_i^n m_i \mathbf{z}_i}{\sum_i^n m_i}\ and\ \boldsymbol{\mu}_\mathbf{RW} = \frac{\sum_i^n m_i(\mathbf{z}_{i-\boldsymbol{\mu}_\mathbf{RW}})(\mathbf{z}_{i-\boldsymbol{\mu}_\mathbf{RW}})^T}{\sum_i^n m_i}$$

The simplest weight function M is "hard rejection", with $M(d)$ = $I(d \leq cv)$, where $I(.)$ is the indicator function. We consider

$$cv = 1.4826\ \frac{\chi_p^2(0.95)\text{median}\ (rd_1, \dots, rd_n)}{\chi_p^2(0.5)} \qquad (10)$$

It is showed that reweighted *comedian* estimates are positive definite, possess high-breakdown value and are approximately affine equivariant [9]. Then the robust Comedian estimators for $\boldsymbol{\mathcal{B}}$, **α** and $\mathbf{\Sigma}_\varepsilon$ are obtained by

$$\widehat{\boldsymbol{\mathcal{B}}}_{\boldsymbol{R}} = \ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Rx}}^{-1} \qquad (11)$$
$$\widehat{\boldsymbol{\alpha}}_{\boldsymbol{R}} = \widehat{\boldsymbol{\mu}}_{\boldsymbol{Ry}}^{-1}\ - \widehat{\boldsymbol{\mathcal{B}}}_{\boldsymbol{R}}^{t}\widehat{\boldsymbol{\mu}}_{\boldsymbol{Rx}} \qquad (12)$$
$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{R}\varepsilon} = \ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Ry}} - \widehat{\boldsymbol{\mathcal{B}}}_{\boldsymbol{R}}^{t}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Rx}}\widehat{\boldsymbol{\mathcal{B}}}_{\boldsymbol{R}} \qquad (13)$$

Efficiency of the proposed method is analyzed and evaluated through simulation.

III. EMPIRICAL RESULTS

To investigate the importance and finite sample efficiency of Comedian multivariate regression, the following simulation study is performed. For various sample sizes n and for different choices of p and q, simulated m datasets of size n from the multivariate standard Gaussian distribution N(**0**, $\mathbf{I}_{p+q}$), which corresponds to putting $\boldsymbol{\mathcal{B}}$ = **0** and $\boldsymbol{\alpha}$ = **0**. For each dataset $\mathbf{Z}^{(k)}$, k = 1, . . . ,m, Comedian regression has been

Table 1. Finite sample comparison of Comedian, MLE, MCD and OGK estimations based on Mean Square Error (MSE) for p=q=6

	n			
	50	200	500	1000
Comedian Regression:				
Slope	1.202	1.051	1.014	1.002
Intercept	1.176	1.032	1.003	1.004
$\mathbf{\Sigma}_{\text{diagnoal}}$	2.641	2.126	2.065	2.050
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	0.897	0.983	0.996	0.997
MLE Regression:				
Slope	1.201	1.051	1.014	1.002
Intercept	1.176	1.032	1.003	1.004
$\mathbf{\Sigma}_{\text{diagnoal}}$	2.640	2.126	2.065	2.050
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	0.897	0.983	0.996	0.997
MCD Regression:				
Slope	3.571	1.541	1.245	1.171
Intercept	2.290	1.297	1.137	1.086
$\mathbf{\Sigma}_{\text{diagnoal}}$	6.323	2.883	2.412	2.325
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	3.187	1.540	1.235	1.179
OGK Regression:				
Slope	1.732	1.428	1.353	1.329
Intercept	1.522	1.260	1.225	1.205
$\mathbf{\Sigma}_{\text{diagnoal}}$	5.322	5.774	8.046	12.440
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	0.908	1.056	1.074	1.093

Table 2. Finite sample comparison of Comedian, MLE, MCD and OGK estimations based on Mean Square Error (MSE) for p=q=6, when the data contains 10% outlier.

	n			
	50	200	500	1000
Comedian Regression:				
Slope	1.335	1.144	1.141	1.109
Intercept	1.299	1.121	1.150	1.147
$\mathbf{\Sigma}_{\text{diagnoal}}$	2.935	2.421	2.283	2.226
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	0.996	1.060	1.129	1.092
MLE Regression:				
Slope	12.545	10.731	10.792	10.537
Intercept	54.249	211.368	523.930	1051.373
$\mathbf{\Sigma}_{\text{diagnoal}}$	3424.488	16510.274	42681.683	86414.403
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	3585.411	16937.652	43621.652	88193.685
MCD Regression:				
Slope	3.226	1.516	1.344	1.264
Intercept	2.229	1.344	1.236	1.205
$\mathbf{\Sigma}_{\text{diagnoal}}$	6.288	4.373	5.875	9.717
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	3.336	1.749	1.556	1.459
OGK Regression:				
Slope	1.876	1.515	1.515	1.508
Intercept	1.651	1.365	1.349	1.357
$\mathbf{\Sigma}_{\text{diagnoal}}$	4.988	5.050	6.501	9.192
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	0.964	1.132	1.226	1.248

carried out for yielding the (p × q) slope matrix estimate $\widehat{\square}^{(k)}$, the intercept vector $\widehat{\boldsymbol{\alpha}}^{(k)}$, and the ($q$ × q) covariance matrix estimate $\widehat{\mathbf{\Sigma}}_{\varepsilon}{}^{(k)}$ of the errors.

To measure sample efficiency, mean squared error (MSE) of the proposed estimators are used. As commonly defined, MSE of a univariate component T are given by

$$MSE(T) = n\, ave_l(T^{(l)} - \theta)^2$$

Where, θ is the true value of parameter. The MSE of slope is defined as

$$MSE\left(\widehat{\boldsymbol{\mathcal{B}}}\right) = n\, ave_{j,k}(MSE\left(\widehat{\boldsymbol{\mathcal{B}}}_{j,k}\right))$$

Similarly for the intercept $\widehat{\boldsymbol{\alpha}}$ and for the diagonal and off-diagonal of $\widehat{\mathbf{\Sigma}}_{\varepsilon}$.

Table 1 gives the efficiency comparison results of Comedian regression method from simulated data. The proposed Comedian regression is compared with MLE, MCD and OGK based on Mean square Error (MSE). The MSE of slope matrix, intercept vector, and error covariance matrix were obtained based on different methods are tabulated. All simulations were done with m=1000 replications. The table contains sample sizes between 50 and 1000. In the Table 1, MSE of Comedian regression estimates equals MSE of MLE regression estimates. But he MSE obtained from Comedian regression are much lower than those obtained from MCD regression and OGK regression. Simulations for other sample sizes n and different dimensions p and q gave similar results.

To study the robustness, multivariate data sets contaminated by different type of outliers are simulated. A point ($\mathbf{x}_i$, $\mathbf{y}_i$) that does not follow the linear pattern of the majority of the data but whose $\mathbf{x}_i$ is not outlying is called a vertical outlier. A point ($\mathbf{x}_i$, $\mathbf{y}_i$) whose $\mathbf{x}_i$ is outlying is called a leverage point. Such a point ($\mathbf{x}_i$, $\mathbf{y}_i$) do not follow pattern of the remaining data is term as bad leverage point; otherwise, it is a good leverage point. The data sets are generated with both type of outliers because regression estimators often inefficient in the presence of vertical outliers or bad leverage points. From the data discussed in the beginning of the section 4, 10% of data is replaced as follows. To include vertical outliers, the $\mathbf{x}_i$'s are kept same and q response variables are distributed as N(2$\surd\chi^2_{p+q,\ 0.99}$, 0.1). Here only the response variables are outlying. Further 10% of the data is replaced with bad leverage points for which p independent variables are generated according to N(2$\surd\chi^2_{p,\ 0.99}$, 0.1) and q dependent variables are generated according to N(2$\surd\chi^2_{q,\ 0.99}$, 0.1).

Efficiency comparison results of Comedian regression method for a 10% contaminated data are shown in Table 2. Here also the proposed Comedian regression is compared with MLE, MCD and OGK based on Mean square Error (MSE). From the table, one can see that MSE obtained from Comedian regression are much lower than those obtained from MLE regression, MCD regression and OGK regression. The contamination level is increased to 20% and 40%, the efficiency values are shown in Table 3 and Table 4 respectively. In 20% contaminated data, the efficiency of proposed method is similar to 10% contamination. Comedian

Regression MSEs are greater for small sample sizes and gradually it deceases for 40% contaminated data. Simulations for other sample sizes *n* and different dimensions *p* and *q* gave similar results. The efficiency of Comedian regression is compared in correlated sample by generating correlated multivariate Gaussian responses with correlation r_{jk} = 0.5. The result of efficiency comparison using correlated data is shows the performance of Comedian regression under collinear situations. Again, comparison process is repeated for data with 10% of vertical outliers and 10% of bad leverage points.

IV. Robustness Properties

The robustness properties of the estimator are studied in terms of breakdown and affine equivariance. These robustness properties confirm finite sample results in the previous section. The breakdown point of an estimator is the proportion of outliers an estimator can handle before giving an incorrect result. An empirical method to find the breakdown value was discussed in [9]. To find the breakdown value Comedian regression method, observations of size n generated from the multivariate standard Gaussian distribution N ($\mathbf{0}$, $\mathbf{I}_{p+q}$). The efficiency of the estimates from data sets with and without outliers is compared to find the maximum proportion of contamination tolerable by the Comedian regression method. The study consists of two kind of contamination: vertical outliers and bad leverage points described in the previous section. Various values of *n* (*n* = 100, 1000), *γ* percentage of contamination (*γ* = 10, 20, 30, 40, 45, 48) and different combinations p and q of were selected to identify the empirical breakdown values of the Comedian regression method. Table 5 contains efficiency comparison of the Comedian regression estimates through MSE of the additional 48% contaminated observations. The efficiency values are tabulated for different combinations of variables dimension of sample size *n* =1000. It is possible to observe that the MSE values similar for both type of simulated data, *γ*=48 indicate high level of contamination that the Comedian regression method can robustly detect.

Table 3. Finite sample comparison of Comedian, MLE, MCD and OGK estimations based on Mean Square Error (MSE) for p=q=6, when the data contains 20% outlier

	n			
	50	200	500	1000
Comedian Regression:				
Slope	1.571	1.328	1.269	1.252
Intercept	1.492	1.291	1.258	1.248
$\mathbf{\Sigma}_{\text{diagnoal}}$	3.432	2.667	2.538	2.582
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	1.097	1.228	1.251	1.258
MLE Regression:				
Slope	21.310	18.268	17.437	17.620
Intercept	210.239	839.258	2097.342	4194.708
$\mathbf{\Sigma}_{\text{diagnoal}}$	10923.406	52256.444	134867.700	272745.700
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	11328.196	53570.524	137997.700	278842.700
MCD Regression:				
Slope	3.791	1.617	1.416	1.365
Intercept	10.928	1.448	1.342	1.301
$\mathbf{\Sigma}_{\text{diagnoal}}$	437.533	8.296	16.307	31.234
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	445.059	2.099	1.916	1.894
OGK Regression:				
Slope	2.105	1.680	1.624	1.668
Intercept	1.836	1.495	1.522	1.603
$\mathbf{\Sigma}_{\text{diagnoal}}$	5.103	4.472	5.269	7.086
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	1.051	1.302	1.416	1.531

Generalized versions of regression, scale, affine equivariance and robustness of multiple regression estimators developed in [11]. Consider $\mathbf{T(X,Y)}$=($\widehat{\boldsymbol{\mathcal{B}}}^t$, $\widehat{\boldsymbol{\alpha}}$)^t, $\mathbf{X}$ is (*n*×*p*) matrix and $\mathbf{Y}$ is (*n*×*q*) matrix. The regression equivariance is that if we transformation of the response variables by adding a linear transformation of predictor variables is equivalent to adding the coefficients in the linear transformation to the estimator. The estimator $\mathbf{T}$ is said to be regression equivariant if

$$\boldsymbol{T}(\boldsymbol{X},\boldsymbol{Y} + \boldsymbol{XC} + \boldsymbol{I_n V^t}) = \boldsymbol{T}(\boldsymbol{X},\boldsymbol{Y}) + (\boldsymbol{C^t}\,,\boldsymbol{V})^{\boldsymbol{t}} \qquad (14)$$

Here $\mathbf{C}$ is any (*p*×*q*) matrix, $\mathbf{V}$ is any (*q*×1) vector, and $\mathbf{I}_n$ = (1, 1,…,1)^t ∈ $\mathbf{R}^n$

The estimator $\mathbf{T}$ is said to be y-affine equivariant if

$$\boldsymbol{T}(\boldsymbol{X},\boldsymbol{Y}\,\boldsymbol{M} + \boldsymbol{I_n P^t}) = \boldsymbol{T}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{M} + \left(\boldsymbol{O}_{pq}^{\boldsymbol{t}},\boldsymbol{P}\right)^{\boldsymbol{t}} \qquad (15)$$

The $\mathbf{y}$-affine equivariance of estimator $\mathbf{T}$ means linear transformation of the response variables implies that the estimator $\mathbf{T}$ is transformed in the same manner. Here $\mathbf{M}$ is any nonsingular (*q*×*q*) matrix, $\mathbf{P}$ is any (*q*×1) vector, and $\mathbf{O}_{pq}$ is (*p*×*q*) zero matrix.

The estimator $\mathbf{T}$ is said to be x-affine equivariant if

$$\boldsymbol{T}(\boldsymbol{Y}\,\boldsymbol{N}^t + \boldsymbol{I}_n\,\boldsymbol{D}^t,\boldsymbol{Y}) = \boldsymbol{T}(\boldsymbol{X},\boldsymbol{Y})\,\boldsymbol{M} + \left(\widehat{\boldsymbol{\mathcal{B}}}^t\,\boldsymbol{N}^{-1},\,\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\mathcal{B}}}^t\,\boldsymbol{N}^{-1}\boldsymbol{D}\right)^{\boldsymbol{t}} \qquad (16)$$

Here $\mathbf{N}$ is any nonsingular (*p*×*p*) matrix and $\mathbf{D}$ is any (*p*×1) vector. If the predictor variables are transformed linearly, then $\mathbf{x}$-affine equivariance says that the estimator $\mathbf{T}$ transforms accordingly.

The three equivariance properties are empirically proved with the help of simulated samples in all possible situations by varying parameter and different contamination levels. Table 6 gives empirical evidence to the affine equivariance expressions (14), (15) and (16). The table contains the MSE efficiency of the estimates from transformed data and efficiency of transformed estimates from untransformed data. It is clear that the MSE values are equal when the transformations are given to data and estimate. This indicates the Comedian regression method is affine equivariant. The result is similar when affine equivariance is tested for different possible contamination levels.

One of the important advantages of Comedian regression is that the time consumption for estimation is relatively less compared with MLE, OGK and MCD method. A simulation study is performed to compare the time efficiency of the proposed method. The simulation consist of different sample sizes n (50, 200, 500, 1000) with different combinations of p and q, all simulation done for m=1000 replications. The average time consumptions of different robust methods are tabulated in Table 7. It is possible to see that comedian regression method requires relatively less time for the estimation than other methods.

V. ILLUSTRATION EXAMPLES

Consider the dataset consisting of measurements of properties of Pulp-Fiber and the paper made from them [8]. The dataset comprises of n = 62 observations with p = 4 predictor variables and q = 4 response variables. The predictor variables describe the properties four pulp fiber characteristics: arithmetic fiber length, long fiber fraction, fine fiber fraction and zero span tensile and the response variables measure four properties of paper: breaking length, elastic modulus, stress at failure and burst strength measure property paper made from them. The objective is to establish a relationship between pulp fiber properties and the resulting paper properties.

The Figure1 shows the diagnostic plot of Pulp-Fiber data (robust residual distance versus the robust distance of residuals). The vertical and horizontal cutoff lines shown in the Figure1 is at $\sqrt{\chi^2_{4,\,0.975}} = 3.34$. Observations 56, 58, 59, 60, 61 and 62 lie far from both the cutoff lines, these six observations thus be classified as outliers (bad leverage points). Some observations (28, 51, and 52) lie above the horizontal cutoff lines, these are vertical outliers because they have small residual distance. Considering the fact that the efficiency of Comedian Multivariate regression is relatively high, the suspicious outliers are observations 56, 58, 59, 60, 61 and 62. This computation took only 0.42 seconds in R-Programming.

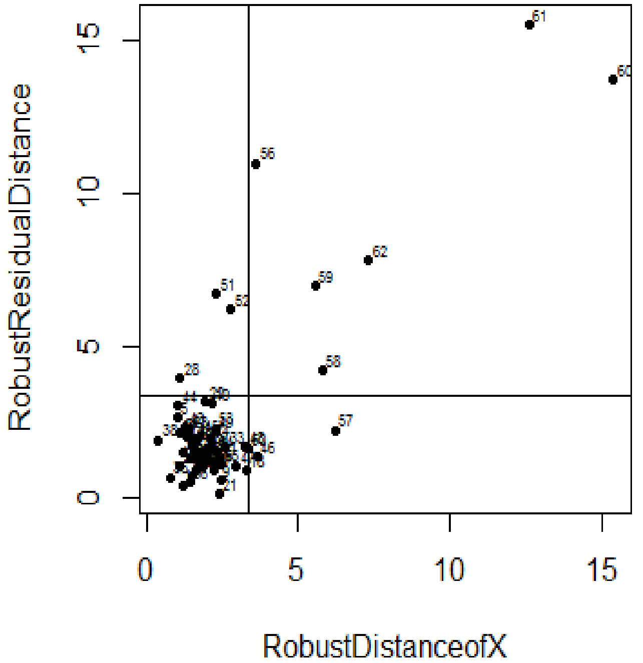

Figure 1: Plot of Robust Residuals versus Robust Distances for the Pulp-fiber data

Table 4. Finite sample comparison of Comedian, MLE, MCD and OGK estimations based on Mean Square Error (MSE) for p=q=6, when the data contains 40% outlier

	n			
	50	200	500	1000
Comedian Regression:				
Slope	12.487	6.554	8.777	13.15665
Intercept	170.516	373.544	1602.672	5444.999
$\mathbf{\Sigma}_{\text{diagnoal}}$	5304.519	16483.99	72856.47	250612.3
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	5516.615	16957.59	74829.62	257271.1
MLE Regression:				
Slope	31.204	26.305	26.876	25.952
Intercept	846.075	3358.234	8386.028	16783.810
$\mathbf{\Sigma}_{\text{diagnoal}}$	24526.030	116970.600	301261.400	608892.000
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	25482.450	120633.600	310316.500	626954.500
MCD Regression:				
Slope	77.393	50.714	50.593	47.986
Intercept	1818.105	9090.076	22660.577	45244.190
$\mathbf{\Sigma}_{\text{diagnoal}}$	50644.300	185423.200	474748.800	962059.400
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	52398.640	193795.500	496380.500	1006114.300
OGK Regression:				
Slope	51.700	30.970	30.826	29.763
Intercept	1010.802	4416.548	11286.298	22811.280
$\mathbf{\Sigma}_{\text{diagnoal}}$	22468.810	122102.500	317716.600	644000.400
$\mathbf{\Sigma}_{\text{offdiagnoal}}$	23621.880	127502.900	331643.200	672213.500

VI. CONCLUSION

The context of robustness classical regression estimation procedures are highly sensitive to the outliers presented in the data set. Several alternative robust methods are available to estimate multivariate regression parameters like MCD and OGK. In this paper an improved alternative robust estimation method is suggested based on Comedian estimation introduced by [9]. Substituting proposed Multivariate robust estimate of location and scatter in the classical expressions of Multivariate regression parameter for slope, intercept and error covariance gives Comedian Multivariate regression method. The performance of Comedian regression is investigated by a comparative study with MLE, MCD and OGK methods. The finite sample efficiency and robustness properties are explained with the help of simulated samples and the MSE result are presented in the tables. The proposed approach gave best finite sample performance in the simulations and also gave highest efficiency compared to the other methods. Moreover, the robustness properties of the proposed approach also exist in simulations with contaminated data sets. The proposed method satisfied robustness properties like high breakdown value and affine equivariance through simulation technique. The time efficiency of the Comedian regression method is remarkably better than the other two methods and it is explained through the average time spent for estimation from 1000 replications of simulation. The Comedian regression method requires almost half time required by the OGK and MCD method. Comedian Multivariate regression on a real data application with the help of diagnostic plots has been illustrated. The plots have been constructed based on robust residual distances. With the help of proposed estimators, it is easy to identify possible outliers contained in the data. The proposed robust regression estimator is suitable for multivariate regression estimation and further regression analysis in real data sets with multiple outliers.

REFERENCE

[1]. V. Barnett, T. Lewis, “*Outliers in Statistical Data*”, 3rd ed. Chichester, John Wiley & Sons, pp. 27-52, 1994.

[2]. T.A. Sajesh, M. R. Srinivasan. “*An Overview of Multiple Outliers in Multidimensional Data*”, Sri Lankan Journal of Applied statistics, Vol. 14, Issue. 2, pp. 87-120, 2013.

[3]. R.A. Maronna, V.J. Yohai, “*Robust estimation in simultaneous equation models*”, Journal of Statistical Planning and Inference, Vol. 57, Issue. 2, pp. 233-244, 1997.

[4]. R. Koenker, S. Portnoy, “*M Estimation of Multivariate Regressions*”, Journal of the American Statistical Association, Vol. 185, Issue. 412, pp. 1060–1068, 1990.

[5]. Z.D. Bai, N.R. Chen, B.Q. Miao, C.R. Rao, “*Asymptotic Theory of Least Distance Estimate in Multivariate Linear Models*”, Statistics, Vol. 21, Issue. 4, pp. 503-51, 1990.

[6]. E. Ollila, H. Oja, T.P. Hettmansperger, “*Estimates of regression coefficients based on sign covariance matrix*”, Journal of Royal Statistical Society: Series B, Vol. 64, Issue. 3, pp. 447-466, 2002.

[7]. P.J. Rousseeuw, “*Least Median of Squares Regression*”, Journal of American Statistical Association, Vol. 79, Issue. 388, pp. 871–880, 1984.

[8]. P.J. Rousseeuw, S. V. Aelst, K. V. Driessen, J. Agull´o, “*Robust multivariate regression*”, Technometrics, Vol. 46, Issue. 3, pp. 293-305, 2004.

[9]. T.A. Sajesh, M.R. Srinivasan. “*Outlier Detection for High Dimensional Data using Comedian Approach*”, Journal of Statistical Computation and Simulation, Vol. 82, Issue. 5, pp. 745-757, 2012.

[10]. R.A. Maronna, R.H. Zamar, “*Robust estimates of location and dispersion for high-dimensional data sets*”, Technometrics, Vol. 44, Issue. 4, pp. 307–317, 2002.

[11]. P.J. Rousseeuw, A.M. Leroy, “*Robust Regression and Outlier Detection*”, New York: John Wiley & Sons, pp. 112-143, 1987.

AUTHORS PROFILE

O.K Sajana is a Research Scholar in the department of Statistics at St Thomas’ college (Autonomous), Thrissur District, Kerala affiliated to University of Calicut. She received post-graduation and Bachelor’s degree in Statistics. She has 3 years of teaching experience. She published 2 research articles in international journals in the field of Statistical quality control. Her current research area is multivariate outlier detection and Statistical inference.

T.A Sajesh received his MSc, M.Phil and PhD degree from Madras University. He is reviewer of Journal of applied Statistics. He is working as Assistant professor at faculty of Statistics, St Thomas’ college (Autonomous), Thrissur District, Kerala. He has 6 years of teaching experience and 4 years of research experience. He has several publications in reputed international journals. His research interest focuses on robust statistical inference, multidimensional outlier detection.

Table 5. Efficiency comparison of breakdown value

	Normal data				48%Contaminated data			
p,q	Slope	Intercept	$\Sigma_{diagnoal}$	$\Sigma_{offdiagnoal}$	Slope	Intercept	$\Sigma_{diagnoal}$	$\Sigma_{offdiagnoal}$
6,6	1.000	1.008	2.037	1.008	1.000	1.008	2.037	1.008
6,10	1.012	1.016	2.020	0.991	1.012	1.016	2.020	0.991
10,6	1.002	0.983	2.11	0.987	1.002	0.983	2.11	0.987
10,10	1.017	1.022	2.107	0.982	1.017	1.022	2.107	0.982
15,15	1.013	1.005	2.244	0.986	1.013	1.005	2.244	0.986

Table 6: MSE comparison for different affine equivariance,
The sample size is n=1000

	Transformed data					
p,q	Regression Equivariance		Y Equivariance		X Equivariance	
	$\mathcal{B}$	$\boldsymbol{\alpha}$	$\mathcal{B}$	$\boldsymbol{\alpha}$	$\mathcal{B}$	$\boldsymbol{\alpha}$
6,6	336.409	332.911	0.389	330.204	0.374	1.839
6,10	333.429	326.539	0.385	336.376	.378	1.931
10,6	335.287	342.927	0.370	338.949	0.375	2.441
10,10	333.188	340.942	0.379	335.341	0.379	2.465
15,15	334.682	330.545	0.375	332.042	0.379	3.064
	Transformed Estimate					
p,q	Regression Equivariance		Y Equivariance		X Equivariance	
	$\mathcal{B}$	$\boldsymbol{\alpha}$	$\mathcal{B}$	$\boldsymbol{\alpha}$	$\mathcal{B}$	$\boldsymbol{\alpha}$
6,6	336.409	332.911	0.389	330.204	0.377	1.841
6,10	333.429	326.539	0.385	336.376	0.379	1.901
10,6	335.163	342.362	0.370	338.949	0.373	2.380
10,10	333.188	340.9425	0.379	335.341	0.377	2.444
15,15	334.682	330.5456	0.375	332.042	0.379	3.019

Table 7: Average time consumption of different method in R Programming.

		Time(s)		
n	*p,q*	Comedian	MCD	OGK
50	4,4	0.017	0.030	0.023
100	4,4	0.021	0.090	0.027
500	4,4	0.030	0.219	0.038
1000	4,4	0.064	0.249	0.086
50	6,6	0.040	0.069	0.059
100	6,6	0.034	0.162	0.052
500	6,6	0.057	0.421	0.082
1000	6,6	0.083	0.340	0.117
50	10,10	0.089	0.159	0.139
100	10,10	0.115	0.467	0.176
500	10,10	0.139	0.982	0.212
1000	10,10	0.209	1.519	0.317