Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 7 (22): 3469-3476, 2007
ISSN 1812-5654

A Simulation Study on Robust Alternatives of Least Squares Regression

M. Mohebbi, K. Nourijelyani and H. Zeraati
Department of Epidemiology and Biostatistics, School of Public Health,
Medical Sciences/University of Tehran, Tehran, Iran

Abstract: We applied four methods of linear regression; the least squares, Huber M, least absolute deviations and nonparametric to several distributional assumptions. The same sets of simulated data were used and MSE, MAD and biases of these methods were compared. The least absolute deviations, Huber M and nonparametric regression shown to be more appropriate alternatives to the least squares in heavy tailed distributions while the nonparametric and LAD regression were better choices for skewed data. However, no best method could be suggested in all situations and using more than one method of exploratory data analysis is recommended in practice.

Key words: Robust regression, simulation study, outlier data, skewed data, least square, LAD regression, Huber M Method

INTRODUCTION

Modelling data by the means of linear least squares method is very important and crucial. Frequently, however, the well-known least squares regression procedure is only optimal under certain distributional assumption of errors. In practice, this assumption may not hold because of possibility of the skewness or presence of outliers in data. In theory, when the assumption of normality does not meet, the standard least squares estimation for the regression coefficients β will be biased and/or non-efficient, see for example (Hampel *et al*., 1986).

When the assumption of normality is not met in a linear regression problem, several alternative methods of the standard Least Squares (LS) regression have been proposed (Draper and Smith, 1998; Kutner *et al*., 2004; Ortiz *et al*., 2006). Among these, three methods are in widespread application in many branches of applied science. Theses methods are robust M-estimation, (Huber, 1964), Least Absolute Deviations (LAD) method, (Dielman and Pfaffenberger, 1982) and (Bloomfield and Steiger, 1983) and nonparametric (rank based) methods, (Adichie, 1967; Jureckova, 1971; Jaeckel, 1972).

There are many statistical tests as well as visual procedures to assess potential deviances from standard assumption of an ordinary least squares regression model. Among these (Cook and Weisberg, 1999; Weisberg, 2005; Chatterjee and Hadi, 2006; Montgomery *et al*., 2007) described standard procedures for such assessments.

In spite of these diagnosis procedures and availability of alternative methods to ordinary least squares, there are few recommendations regarding the conditions in which each method has better efficacy, but all of these are based only on personal experiences and so a formal comparison of these methods is not available. The aim of this research is a formal comparison of these methods through a simulation study.

Here, we report some summary results of a numerical study undertaken to compare the properties of three alternatives to standard least squares method for simple and multiple linear regression analysis. In our simulation study, we generate independent data sets which contain outlier by using heavy tail distribution (compare to standard normal distribution) such as Laplce, logistic and mixtures of normal and Laplace distribution and skewed independent data sets from gamma distribution family with different shape parameter and compare the performance of all four candidate regression methods with generated data from standard normal distribution. Comparison was done using MSE (mean squared errors), MAD (mean absolute deviations) and bias criteria.

BACKGROUND INFORMATION

Consider the following regression linear model

$$Y = X\beta + e \qquad (1)$$

Where, Y is an $(n \times 1)$ vector of observations with the design matrix X of order $n\times p$ such that $X_{i1} = 1$, $i = 1..., n$.

Corresponding Author: K. Nourijelyani, Department of Epidemiology and Biostatistics, School of Public Health, Medical Sciences/University of Tehran, P.O. Box 14185-463, Tehran, Iran
Tel: +98-21-88989124 Fax: +98-21-88989127

We consider β as a (p×1) vector parameter and e as an (n×1) vector of independent, identically distributed errors with some distribution function F, which is generally considered as unknown; we only assume that F belongs to some family of distribution functions. The problem is that of estimating the parameter β.

Standard linear model estimation assumes e to be multivariate normally distributed with mean zero and variance-covariance matrix $\sigma^2 I_n$.

There are various methods for estimation of regression coefficients in the above linear regression model. The most commonly used is the method of least squares; it has the best performance if the errors have a normal probability distribution. We can obtain the least squares estimator by minimizing the sum of squares:

$$\sum_{i=1}^{n}(Y_i - x_i'\beta)^2 \tag{2}$$

in the model 1.1 with respect to β, where x'_i is the ith row of X. If X is of full rank, then the least squares estimator is:

$$\hat{\beta}_{LS} = (X'X)^{-1}X'Y \tag{3}$$

The next popular method of estimation is the method of least absolute deviations (LAD) estimation. Similar to the least squares estimator, a LAD estimator of vector β is obtained by minimizing the expression:

$$\sum_{i=1}^{n}|Y_i - x_i'\beta| \tag{4}$$

with respect to β.

The l_1 fit function in the MASS library from S-Plus software (Venables and Ripley, 2002) can be used to find parameter estimations. This function uses the Barrodale-Roberts algorithm (Barrodale and Roberts, 1973, 1974; Bloomfield and Steiger, 1983) and is a specialized linear programming algorithm.

In the circumstances that the distribution of F is not normal or cannot be approximately assumed normal, we should look for alternative estimation procedures, less sensitive to deviations from normality assumption. We obtain an M-estimator M_n by minimizing:

$$\sum_{i=1}^{n}\rho(Y_i - x_i'\beta) \tag{5}$$

with respect to β, where ρ is an appropriate function, usually convex. If φ(x) = dρ(x)/dx is continuous, then M_n is one of the roots of the system of equations:

$$\sum_{i=1}^{n}x_i\varphi(Y_i - x_i'\beta)=0 \tag{6}$$

More specifically, the well-known Huber M-estimator, which employs the following φ function could be used:

$$\varphi(x)=\begin{cases} x & \text{if } |x| \le c \\ c.\text{sign}\,x & \text{if } |x| \ge c \end{cases}$$

For parameter estimation, the rlm function again from the MASS library in S-Plus can be used. This function uses iteratively weighted least squares method for solving Eq. 6 and is fully described in (Huber, 1981; Hampel *et al*., 1986; Marazzi, 1993).

The nonparametric estimate of regression coefficients is obtained by minimizing:

$$\sum_{i=1}^{n}[\text{rank}(Y_i - x_i'\beta) - \frac{n+1}{2}](Y_i - x_i'\beta) \tag{7}$$

This also can be done iteratively, starting with the vector of the least squares estimates and finding the vectors that give the smaller values in Eq. 7. Then, the nonparametric estimate $\hat{\beta}_0$ is obtained as the median of the differences:

$$y_i - (\hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip})$$

Estimation of regression coefficients by minimizing (7) was proposed by Jaeckel (1972). In fact, Jaeckel's estimations are essentially the same as those of (Birks and Dodge, 1993). The RRegress function in Minitab software can be used for parameters estimation in the case of nonparametric rank based regression.

DESIGN OF SIMULATION EXPERIMENTS

In this section, we will describe the design of our study. Sample size (n), number of independent variables (p), estimation method and distribution of errors are the study variables. We will consider p = 1, 2 and 3 and n equal to 10, 20, 30, 50, 100, 250 and 500. For each n and p, we generate an n×(p+1) matrix with the first column of 1's and the next columns are taken from the *uniform* distribution on [-10, 10]. For simple linear regression (p = 1), we consider β = (5,-3)′, for multivariate regression (p = 2), a value of β = (5,-3,1)′ was chosen and for (p = 3), we set β = (5,-1,3,1)′.

The notation and parameters of distributions family, which used in the simulation process, are introduced in Table 1. The errors were simulated from the following

Table 1: Distribution family for simulating errors

Family of distribution	Notation and parameters	pdf f(x)
Normal	x ~ N(μ, σ²)	$\frac{1}{\sqrt{2\pi}\sigma}e^{-[(x-\mu)/\sigma]^2/2}$
Logistic	x ~ LOG(θ, η) 0<θ	$\frac{1}{\theta}\frac{\exp[(x-\eta)/\theta]}{\{1+\exp[(x-\eta)/\theta]\}^2}$
Laplace	x ~ LA(θ, η) 0<θ	$\frac{1}{2\theta}e^{-\lvert x-\eta\rvert/\theta}$
Gamma	x ~ GAM(θ, κ) 0<θ, 0<κ	$\frac{1}{\theta^{\kappa}\Gamma(\kappa)}x^{\kappa-1}e^{-x/\theta}$ 0 < x

densities: N(0, 1), LOG(0,1), LA(0,1), GAM(1, 0.1), GAM (1, 0.5), GAM(1, 1), GAM(1, 2). In addition, we considered three mixtures of Normal and Laplace distribution as follows: 0.95 N(0, 1) + 0.05 LA(0, 1), 0.90 N(0, 1) + 0.10 LA(0, 1) and 0.85 N(0, 1) + 0.15 LA(0, 1).

In each case, 1000 replications were simulated and regression coefficient of LS, MAD, Huber M-estimate and nonparametric method were calculated. For comparing the properties of the estimation procedures, we focused on the Mean Squared Errors (MSE), Mean Absolute Deviations (MAD) and bias of the estimated coefficients. The following three criteria are used:

$$\text{MSE} = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\beta}_i - \beta)'(\hat{\beta}_i - \beta) \qquad (8)$$

$$\text{MAD} = \frac{1}{1000}\sum_{i=1}^{1000}\sum_{j=1}^{p}\left|\hat{\beta}_{ij} - \beta_j\right| \qquad (9)$$

$$\text{Bias} = \sum_{j=1}^{p}\left|\frac{\sum_{i=1}^{1000}\hat{\beta}_{ij}}{1000} - \beta_j\right| \qquad (10)$$

Where, p is the number of independent variables.

SIMULATION RESULTS

Overall results of the methods under study and corresponding MSE, MAD and bias of 1000 simulation for each estimation method are presented in Fig. 2-21. These figures illustrate the results of MSE, MAD and bias for the case of p = 1 parameter. Graphs for p = 2 and 3 parameters regressions were similar to those of p = 1 parameter regression and so were not shown in this study.

We wrote an S-Plus function for calculating MSE, MAD and bias criteria for LS, Huber M and LAD methods. Parameters in LS method were estimated by lm function in S-Plus software. Parameters in Huber M method were estimated by rlm function and in LAD method by l1fit function, all are available in MASS library from S-Plus software (Venables and Ripley, 2002). The MSE, MAD and bias criteria for nonparametric method were calculated by a function in R wrote by authors.

Depending on different choices of p (number of parameters), n (sample size) and type of distribution, running time of simulation varied between 2 to 130 min.

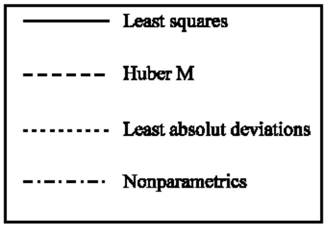

Fig. 1: Legend for the graphs in the result section

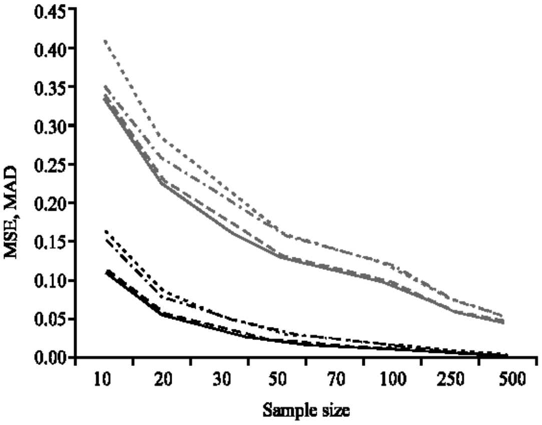

Fig. 2: Results of MSE (black lines) and MAD (grey lines) for 1000 simulations from N(0, 1) distribution

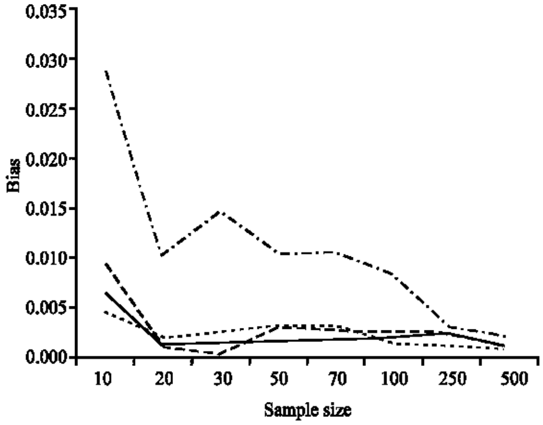

Fig. 3: Results of bias for 1000 simulations from N(0, 1) distribution

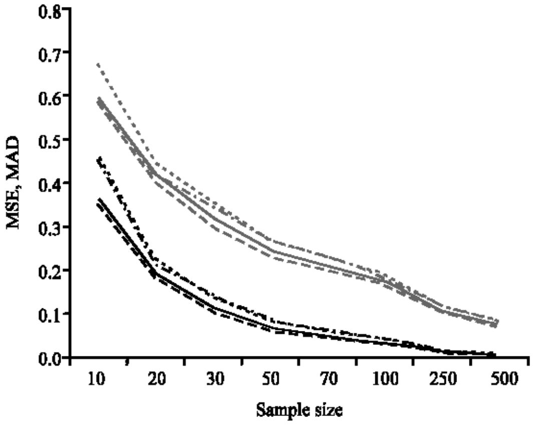

Fig. 4: Results of MSE (black lines) and MAD (grey lines) for 1000 simulations from LOG (0, 1) distribution

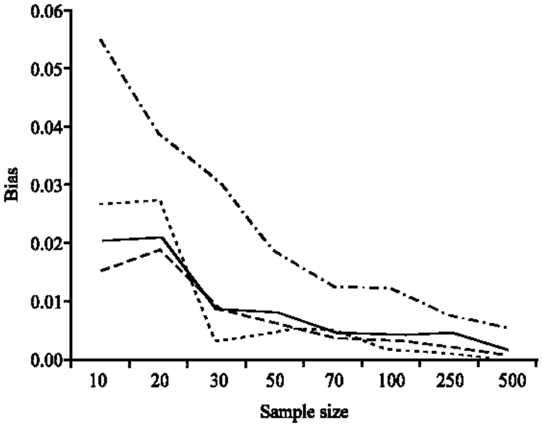

Fig. 5: Results of bias for 1000 simulations from LOG(0, 1) distribution

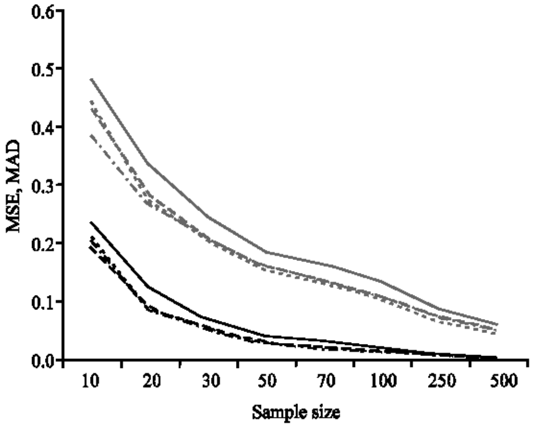

Fig. 6: Results of MSE (black lines) and MAD (grey lines) for 1000 simulations from LA (0, 1) distribution

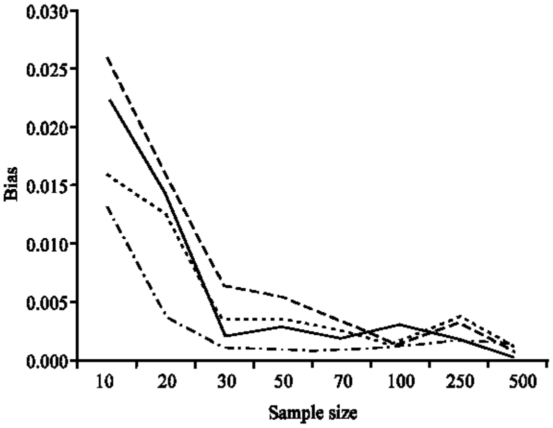

Fig. 7: Results of bias for 1000 simulations from LA(0, 1) distribution

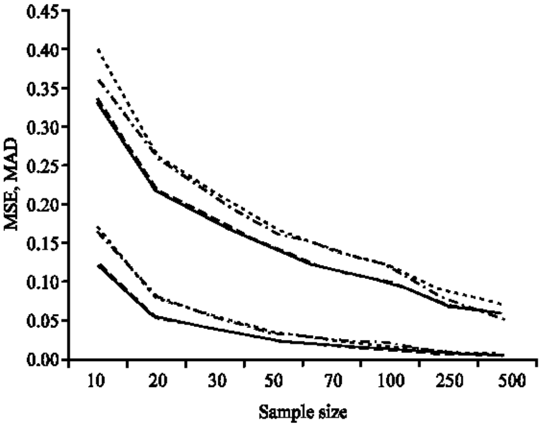

Fig. 8: Results of MSE (black lines) and MAD (grey lines) for 1000 simulations from 0.95 N(0, 1) + 0.05 LA(0, 1) distribution

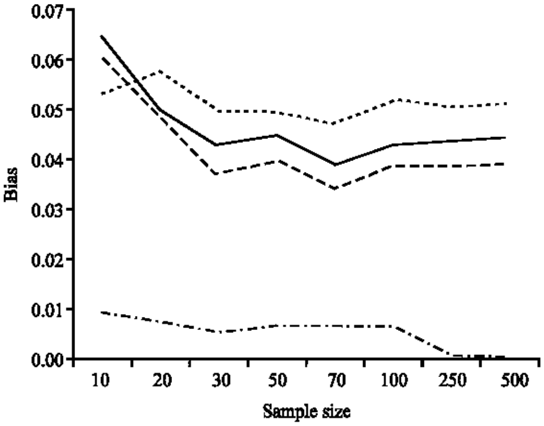

Fig. 9: Results of bias for 1000 simulations from 0.95 N(0, 1) + 0.05 LA(0, 1) distribution

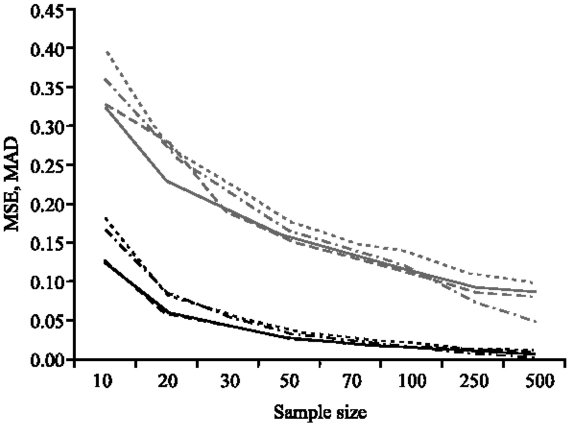

Fig. 10: Results of MSE (black lines) and MAD (grey lines) for 1000 simulations from 0.90 N (0, 1) + 0.10 LA(0, 1) distribution

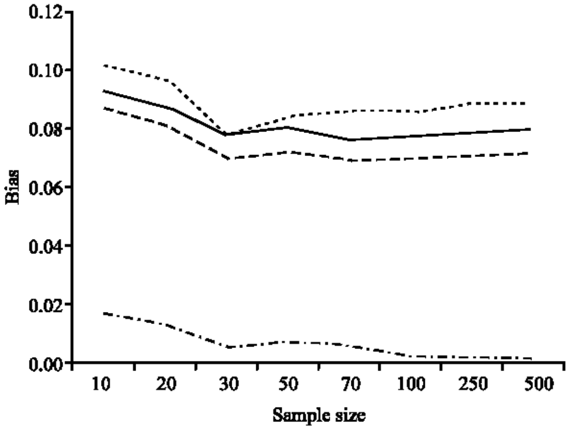

Fig. 11: Results of bias for 1000 simulations from 0.90 N(0, 1) + 0.10 LA(0, 1) distribution

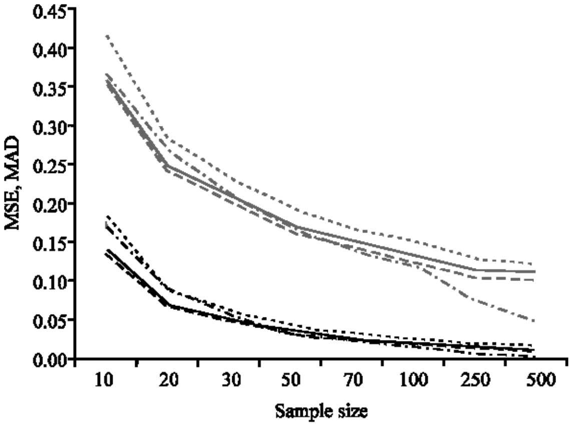

Fig. 12: Results of MSE (black lines) and MAD (grey lines) for 1000 simulations from 0.85 N (0 1) + 0.15 LA(0, 1) distribution

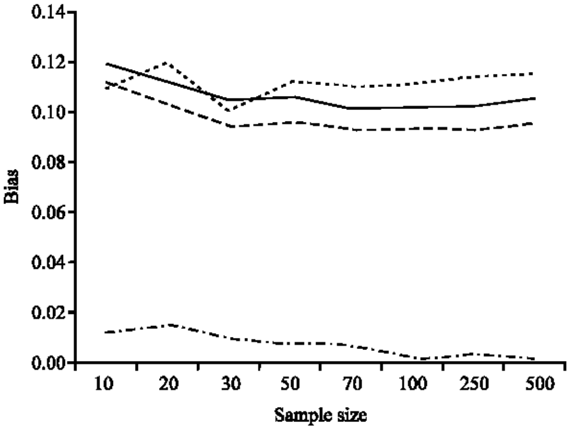

Fig. 13: Results of bias for 1000 simulations from 0.85 N(0, 1) + 0.15 LA (0, 1) distribution

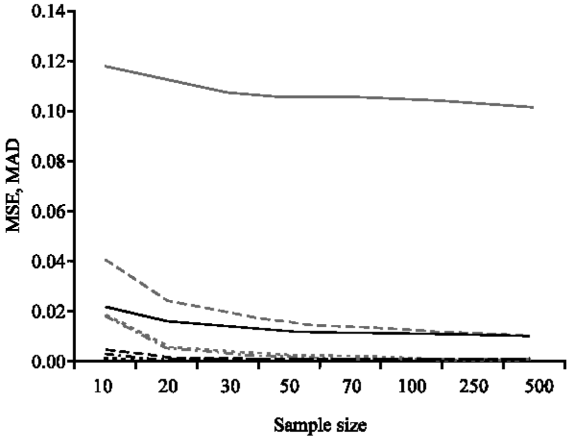

Fig. 14: Results of MSE (black lines) and MAD (grey lines) for 1000 simulations from GAM (1, 0.1) distribution

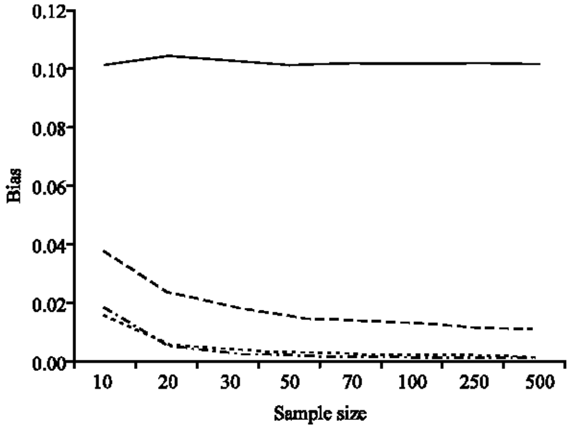

Fig. 15: Results of bias for 1000 simulations from GAM (1, 0.1) distribution

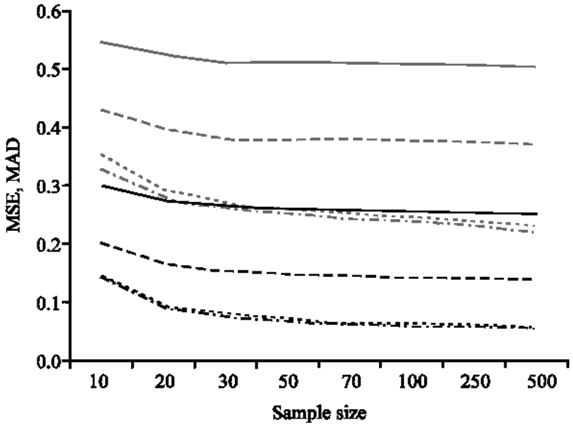

Fig. 16: Results of MSE (black lines) and MAD (grey lines) for 1000 simulations from GAM (1, 0.5) distribution

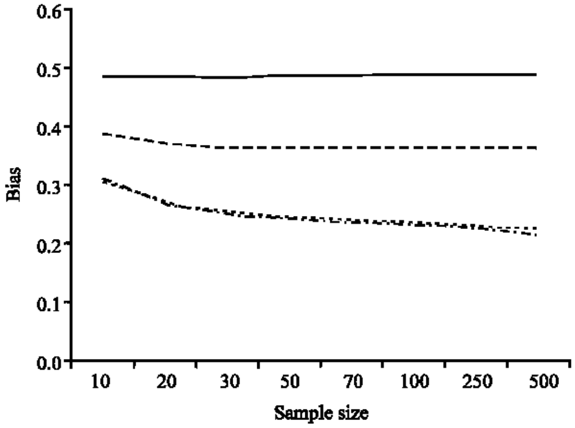

Fig. 17: Results of bias for 1000 simulations from GAM (1, 0.5) distribution

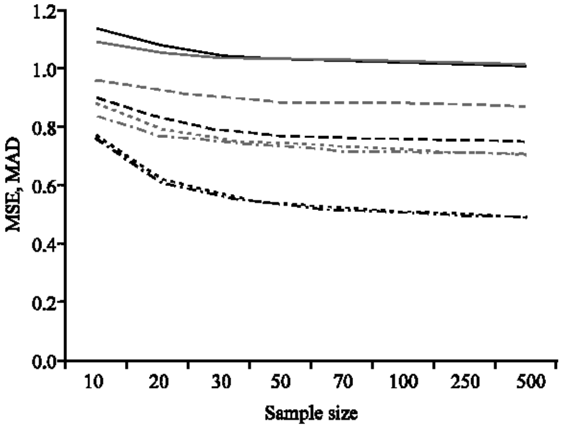

Fig. 18: Results of MSE (black lines) and MAD (grey lines) for 1000 simulations from GAM (1, 1) distribution

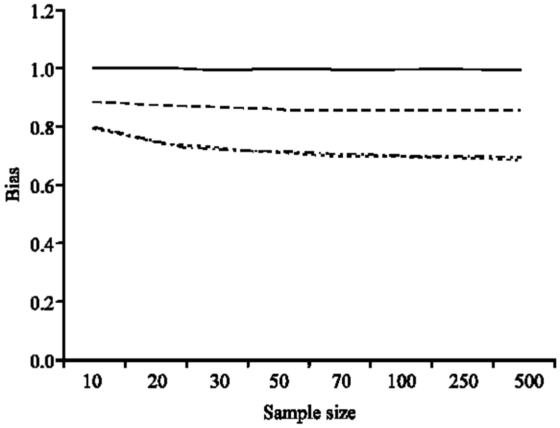

Fig. 19: Results of bias for 1000 simulations from GAM (1, 1) distribution

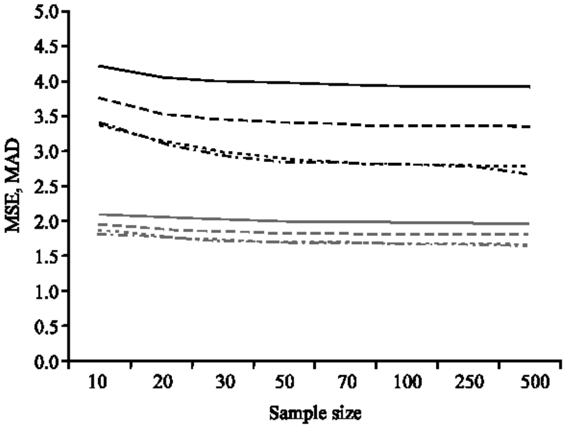

Fig. 20: Results of MSE (black lines) and MAD (grey lines) for 1000 simulations from GAM (1, 2) distribution

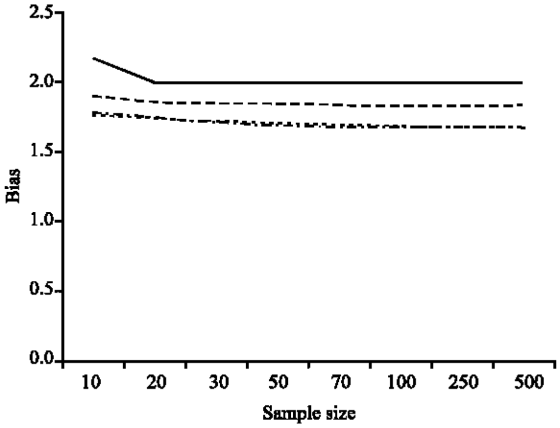

Fig. 21: Results of bias for 1000 simulations from GAM (1, 2) distribution

The general simulation results could be summarized as follows:

As expected, for the studied normal distributions, the MSE and MAD of LS were the smallest, followed by the MSE and MAD of Huber M, Nonparametric and LAD methods, respectively. In addition, in this case, the bias of LS was the smallest followed by the values of biases of Huber M, LAD and nonparametric methods, respectively. However, for the studied logistic distributions, the MSE and MAD of Huber M was the smallest. This followed closely by the MSE and MAD values of LS method.

The MSE and MAD of nonparametric, for this situation were much greater which followed by the MSE and MAD of LAD method. Furthermore, although biases of Huber M, LAD and LS were significantly smaller than the bias of nonparametric, but their patterns as shown in Fig. 6 and 8 were intermingle and so no methods had a preferable bias in this situation. In the cases of Laplace distributions, the MSE and LAD values of LAD, Huber M and nonparametric, were much close to each other but these values for the LS method was significantly larger. As indicated in Fig. 7, the bias of nonparametric in case of LA (0, 1) showed a smaller pattern as compared to the other methods but the general pattern of the bias values for all methods were intermingle so that no preferred method could be chosen based on bias criterion. For the studied mixture family, the LS and Huber M are close to each other and perform better than LAD and nonparametric, with respect to MSE and MAD criteria. In this situation, nonparametric bias was significantly smaller than the bias of other three methods. Although Huber M, LS and LAD performed similarly with regard to the bias criterion, but still as shown in Fig. 9, 11 and 13, the bias of Huber M was slightly lower. In gamma family, for all three criteria, the nonparametric and LAD were close to each other, but inferior to the Huber M. The LS method as also indicated in Fig. 14-21 performed much worst in these situations. We have done a similar set of simulation for $p = 2$ and $p = 3$ parameter regression cases, the results were generally similar for these cases with the following exceptions. For the simulated logistic distributions in the case of $p = 2$ parameter, the least squares method are better than the nonparametric and LAD methods. For the mixture family in $p = 2$ and 3 parameter cases, the biases had no regular pattern. Finally, for the studied Laplace distribution with $p = 2$ and 3 parameter cases, the Huber M method performs better than LAD with respect to all three criteria.

CONCLUSIONS

A parametric estimation method is one based on the assumption that the random errors in the data have a particular type of distribution. Robust M-estimation is an alternative to the parametric estimation when the errors have a distribution that is not necessarily normal but close to normal. One optimal property of the LAD estimates of the regression coefficients is, by their definition, that they are the estimates that give the smallest sum of absolute residuals. In addition, if we assume that the population of errors has a Laplace (or double exponential) distribution, then the LAD estimate is the maximum likelihood estimate (Birks and Dodge, 1993). The strength of LAD estimation is in its robustness with respect to the distribution of response variable. A nonparametric procedure performs reasonably well for almost any possible distribution of the errors. Many such procedures, including the one described here, are based on the idea of using the ranks of the data instead of the actual data values.

In this research performance of four popular regression methods for two important classes of distributions namely symmetric and skewed were investigated. Our choices for symmetric distribution were so that their kurtoses were more than that of standard normal distribution (i.e., heavy tails distributions). This gave us the opportunity to investigate our regression methods with presence of outliers. Present results indicated that when outliers exist, other alternatives of the LS are more appropriate (Fig. 2- 21). Choosing a more efficient alternative to the LS method is closely related to the type of data and so it is advisable to use several alternative methods in data analysis. In cases of skewed distributions, the performance of LS was inferior as compared to other methods. Based on our simulated distributions in this research, the nonparametric and LAD methods were more suitable for the studied Gamma family.

In almost all symmetric distributions investigated here, the MSE and MAD are close for the sample sizes larger than 100 and so none of the estimation methods were superior in such circumstances. However, this has not been true for the cases of the studies skewed distributions where the LS method shown to be far inferior from the other methods of estimation. In general, the bias criterion, as compared with the other criteria, shown to have more fluctuation and this fluctuation persist even for large sample size. This instability of biases created some difficulty and confusion in finding the optimum estimation in some situation.

In this research, four well-known methods of linear regression were studied. However, in order to investigate for possibilities of more suitable methods further studies are needed.

REFERENCES

Adichie, J.N., 1967. Estimation of regression parameters based on rank tests. Ann. Math. Stat., 38: 894-904.

Barrodale, I. and F.D.S. Roberts, 1973. An improved algorithm for discrete L_1 linear approximation. SLAM J. Numerical Anal., 10: 848-893.

Barrodale, I. and F.D.S. Roberts, 1974. Solution of an overdetermined system of equations in the L1 norm. Communication of the ACM, 17: 319-320.

Birks, D. and Y. Dodge, 1993. Alternative Methods of Regression. Wiley, New York.

Bloomfield, P. and W. Steiger, 1983. Least absolute Deviations: Theory, Applications and Algorithms. Boston: Birkhauser.

Chatterjee, S. and A.S. Hadi, 2006. Regression Analysis by Example. 4th Edn. Wiley, New York.

Cook, R. and S. Weisberg, 1999. Applied Regression Including Computing and Graphics, Wiley, New York.

Dielman, T. and R. Pfaffenberger, 1982. LAV (Least Absolute Value) Estimation in Linear Regression: A Review. In: Optimization in Statistics, Zanakis, S.H. and J.S. Rutagi (Eds.). New York: North-Holland.

Draper, N. and H. Smith, 1998. Applied Regression Analysis. 3rd Edn. Wiley, New York.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw and W.A. Sathel, 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

Huber, P., 1964. Robust estimation of a location parameter. Annal. Math. Stat., 35: 73-101.

Huber, P., 1981. Robust Statistics. Wiley, New York.

Jaeckel, L.A., 1972. Estimating regression coefficients by minimizing the dispersion of the residuals. Annal. Math. Stat., 43: 1449-1458.

Jureckova, J., 1971. Nonparametric estimate of regression coefficients. Annal. Math. Stat., 42: 1328-1338.

Kutner, M., C.J. Nachsteim and J. Neter, 2004. Applied Linear Regression Models. 4th Edn. McGraw-Hill, New York.

Marazzi, A., 1993. Algorithms, Routines and Functions for Robust Statistics. Wadsworth and Brooks/Cole.

Montgomery, D., A. Peck and G. Vining, 2007. Introduction to Linear Regression Analysis. 4th Edn. John Wiley, New York.

Ortiz, M., L. Sarabia and A. Herrero, 2006. Robust regression techniques: A useful alternative for the detection of outlier data in chemical analysis. Talanta, 70: 499-512.

Venables, W.N. and B.D. Ripley, 2002. Modern Applied Statistics with S. 4th Edn. Springle, New York.

Weisberg, S., 2005. Applied Linear Regression. 3rd Edn. Wiley, New York.