Int. J. Agricult. Stat. Sci., Vol. 5, No. 2, pp. 355-372, 2009 *ISSN : 0973-1903*

LEAST SQUARES VERSUS LEAST ABSOLUTE DEVIATIONS ESTIMATION IN REGRESSION MODELS

S. Eakambaram* and R. Elangovan

Department of Statistics, Annamalai University, Annamalai Nagar – 608 002, India.
*E-mail: eakambaram.s@gmail.com

Abstract

The popularity of regression analysis has risen dramatically in recent years. The once computationally difficult techniques and now conveniently available to researchers across a wide range of disciplines. The Least Absolute Deviations (LAD) regressions is a field of great interest in regression analysis. The applications of this method are available in the recent literature covering various aspects of modelling, computational efficiency, error analysis, gross identification etc. LAD estimation of conditional medians often is preferable to Least Squares (LS) estimation of conditional means for regression analysis. LAD regression has greater power than LS regression for asymmetric error distributions and heavy tailed, symmetric error distributions and has greater values of the dependent variable. It is the simplest available estimate that is not unduly influenced by large changes in the original distribution and minimizes the maximum asymptotic bias due to skewed outliers. Previous research has indicated that the LAD estimators tend to be more efficient than that of LS estimators in the present of large disturbances. The object of the paper is to provide properties of LS and LAD in regression models, and also is concerned with the application of LAD method, verification of results empirically using simulation results. Numerical examples are provided by comparing LS versus LAD estimation in regression models. Simulation results show that LAD estimation gives a significant advantages in efficiency over OLS estimation.

Key words : Regression Analysis, Ordinary Least Squares, Least Absolute Deviations Regressions, Skewed Outliers.

1. Introduction

Statistical methods are essential for all scientific and industrial researches, especially in the filed of Economics, Sociology, Medicine, Astronomy, Psychology, Business Management, Education, Engineering to name a few. One of the most powerful statistical method is the application of Least Squares. Searle (1971) says that "it is designed for situations where a variable is thought to be related to one or more other measurements made usually on the same object". In linear model analysis, a method to obtain valid inferences is the Least Square (LS) approach. This has been developed in all its details and is universally used. The assumption of normality of the random variables is made in this analysis. A few non-normal variables can be transformed into normal by using certain transformation techniques. If the observation follow the Poisson distribution, then the square root transformation

*This work already has been presented in the 96th Indian Science Congress held at North Easter Hill University, Shillong, from January 3-7,2009.

$Y_{ij}^{*} = \sqrt{Y_{ij}}$ or $Y_{ij}^{*} = \sqrt{1+Y_{ij}}$ would be used. For binomial data expressed as fractions, the arcsine transformation $Y_{ij} = \arcsin\sqrt{Y_{ij}}$ is useful [Montgomery (1997)]. This transformation has a wide use in Engineering. For instance in the paper "Modelling of Process Parameters in Dynamic Form-Cladding" by Raghukandan *et al.* (1998) arcsine transformation is used. For more discussion on transformation, kindly refer to Draper and Hunter (1969). In many situations, the distribution is highly skewed such as Lognormal, Pareto etc. and the use of LS theory leads to highly unreliable results. Relating to power system in Electrical Engineering Faico and Assis (1988) said "If one or more measurements are contaminated by Gross error (non - normal errors) then the hypothesis of normalized residuals are no longer correct". They have stated that in such a situation the identification of this type of errors could be done using LAD method, more efficiently. The testing of hypothesis done using F-test as done in Tauchi's experiments assume normality. In non-normal situation a modified approach is given by Rao *et al.* (1990).

The Least Absolute Deviations (LAD) regression is a field of great interest in regression analysis. The applications of this method are available in the recent literature covering various aspects of modelling, computational efficiency, error analysis, gross error identification etc., Linear models, their variants and extensions are among the most useful and widely used statistical tools for social research. This paper aims to provide an accessible, in-depth, modern treatment of regression analysis and closely related methods. The linear regression model is given by the equation

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \tag{1.1}$$

for i = 1, 2, ..., n sampled observations. In equ. (1.1), y_i is the dependent variable, the x_{ji} are regressors and ε_i is an unobservable error. The β_j are unknown parameters to be estimated from the data. It is assumed that the errors are normally and independently distributed with zero expectation (mean) and constant variance σ^2 i.e. $\varepsilon_i \sim NID(0, \sigma^2)$. If the assumptions that error are normally distributed appears to be violated and if any prior knowledge is available about error distribution, the maximum likelihood argument could be used to obtain estimates based on the criterion other than least squares namely minimization of $\sum_{i=1}^{n} |e_i|$, sum of absolute errors, the linear regression model stated in equ. (1.1) is a particular method of minimizing L_p norm. Taking the linear model of the form stated in equation (1.1),

$Y_i = X_i' \beta + e_i$, the Lp norm is

$\left[\sum |e_i|^p\right]^{1/p}$ with $p \geq 1$. Minimising this, one get estimate of β and also use this for other inferential purposes. Put p=2, we get the classical approach to regression problems, namely least square estimates, though they possess properties such as simplicity, they are not well suited if there are violation in the basic assumptions as stated above. When $p = \infty$ one get the well known minimax or Chebycheff's method. When p = 1, the L_1 norm method is obtained and is known as the Least Absolute Deviation method in literature. The methods of minimizing the sum of absolute and squared deviations from hypothesized linear models have vied for statistical favour for more than 250 years. This venerable method which Laplace called the 'method of situations' has had a bewildering variety of names. Recently, it has accumulated a large array of acronyms: LAE, LAD, MAD, MSAE and others. It is also frequently referred to as L_1-regression and less frequently as median regression. A more detailed discussion on LS procedures and its estimates can be seen in Searle (1971).

2. Methods

As was pointed out in Section 1, the functional relationship of Y and X is of the following form

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2.1}$$

Which is known as simple linear regression of Y on X, β_0 and β_1 are called parameters and should be estimated. Equation (2.1) means that for a given X_i, a corresponding Y_i consists of $\beta_0 + \beta_1 X_i$ and an ε_i by which can observation may fall off the true regression line. On the basis of the information available from the observations, we would like to estimate β_0 and β_1. The term ε is a random variable and is called "error term". By equ. (2.1), we can write

$$Y_i - \beta_0 - \beta_1 X_i = \varepsilon_i \tag{2.2}$$

Finding β_0 and β_1 from (X_i, Y_i), i = 1, 2, ..., n is called estimation of the parameters. There are different methods of obtaining such estimates. In this sections that follow, we will consider methods that estimate the parameters by minimizing sum of squared deviations, mean absolute deviations and maximum of absolute deviations. A more recent history of regression analysis and its applications can also be seen in Draper and Smith (1981).

Minimizing Sum of Squared Deviations (Least Squares Regression)

This method is based on choosing β_0 and β_1 so as to minimize the sum of squares of the vertical deviations of the data points from the fitted line. The sum of squares of deviations (SSD) from the line is

$$SSD = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2 \tag{2.3}$$

Then, we would choose the estimates β_0 and β_1 so that if we substitute them in equ. (2.3) the sum of squared deviations is minimum. Differentiating equ. (2.3) with respect to β_0 and β_1 and setting the resultant partial derivatives to zero, we have

$$\left.\begin{aligned} \frac{\partial SSD}{\partial \beta_0} &= -2\sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 X_i\right) \\ \frac{\partial SSD}{\partial \beta_1} &= -2\sum_{i=1}^{n} X_i\left(Y_i - \beta_0 - \beta_1 X_i\right) \end{aligned}\right\} \tag{2.4}$$

and hence,

$$\left.\begin{aligned}\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1 X_i\right)=0\\ \sum_{i=1}^{n}X_i\left(Y_i-\beta_0-\beta_1 X_i\right)=0\end{aligned}\right\} \tag{2.5}$$

From (2.5), we have

$$\left.\begin{aligned}\beta_0 n+\beta_1\sum_{i=1}^{n}X_i=\sum_{i=1}^{n}Y_i\\ \beta_0\sum_{i=1}^{n}X_i+\beta_1\sum_{i=1}^{n}X_i^2=\sum_{i=1}^{n}X_iY_i\end{aligned}\right\} \tag{2.6}$$

Equations (2.6) are called normal equations. From equ. (2.6), we have

$$\hat{\beta}_1=\frac{\Sigma X_iY_i-\left(\Sigma X_i\right)\left(\Sigma Y_i\right)/n}{\Sigma X_i^2-\left(\Sigma X_i\right)^2/n}$$

and $\hat{\beta}_0=\overline{Y}-\hat{\beta}_1\overline{X}$, where $\overline{Y}$ and $\overline{X}$ are $\sum_{i=1}^{n}Y_i/n$ and $\sum_{i=1}^{n}X_i/n$, respectively. $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained in this direction are called least-squares estimates of β_0 and β_1, respectively. Thus, we can write the estimated regression equation as

$$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X,$$

which is called the prediction equation.

Minimizing Mean Absolute Deviations (MINMAD Regression)

For the simple linear regression model, namely.

$Y = \beta_0 + \beta_1 X + \varepsilon,$

We have observed data on X and Y given by (X_i, Y_i), i = 1, 2, ..., n. We are interested in finding the coefficients β_0 and β_1 such that

$$\frac{1}{n}\sum_{i=1}^{n}\left|Y_i-\beta_0-\beta_1X_i\right| \tag{2.7}$$

is minimized. The expression in equ. (2.7) is known as the mean absolute deviation from the observed and the predicted values of the dependent variable. It can easily be seen that minimization of equ. (2.7) is the same as minimizing

$$\sum_{i=1}^{n}\left|Y_i-\beta_0-\beta_1X_i\right|, \tag{2.8}$$

which is the sum of absolute deviations. First we consider the problem with an additional restriction that β_0 and β_1 as well as minimizing that sum of the absolute deviations should be such that

$Y_0 = \beta_0 + \beta_1 X_0$ for a given pair (X_0, Y_0)

Minimizing Maximum of Absolute Deviations (MINMAXAD Regression)

Consider estimating the parameters β_0, β_1 by minimizing the maximum of absolute deviations. Under this criterion, we have the objective of finding β_0, β_1 such that β_0, β_1 is a solution to

$$\underset{(\beta_0,\beta_1)}{\text{Minimize}}\left[\max_{1\le i\le n}\left|Y_i-(\beta_0+X_i\beta_1)\right|\right]$$

First, we discuss the problem without the β_0 term, i.e., $Y = \beta X + \varepsilon$, $f_1(\beta) = |Y_i - X_i\,\beta|$ for any i, consists of two straight lines with minimum at $(Y_i / X_1, 0)$ and slopes $-|X_i|$, $|X_i|$. The least-squares regression criterion used to minimize $\sum_{i=1}^{n} d_i^2$, and d_i is the deviation from the observed and predicted values of the dependent variable corresponding to the ith observation. Now notice that $\sum_{i=1}^{n} d_i^2$ can be thought of as the variance of the deviations, namely, $1/n\ \Sigma(d_i - d)^2$,

Where, $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)$

But $\bar{d} = 0,$ for the least-squares regression line as it passes through the point $\left(\bar{X},\bar{Y}\right)$. $\Sigma(d_i - d_j)^2$ can be equivalently written as $(1/n)\ \sum_{i<j}\left(d_i-\bar{d}\right)^2$. Thus minimizing $\sum_{i<j}\left(d_i - d_j\right)^2$ is a possible criterion for finding a regression line. In case $\bar{d} = 0,$ this is equivalent to the least-squares regression line. Replacing $(d_i - d_j)^2$ by $|d_i - d_j|$, we obtain another criterion, namely, the sum of the absolute difference between deviations.

Minimizing Sum of Absolute Differences Between Deviations (MINSADBED Regression)

Consider the estimation of the parameters β_0 and β_1, minimizing the sum of absolute differences between deviations; that is,

$$\text{Minimize}\sum_{i<j}\left|d_i - d_j\right|$$

From the expression for d_i, d_j we have

$$\sum_{i<j}\left|d_i - d_j\right| = \sum_{i<j}\left|\left(Y_i-\beta_0-\beta_1X_i\right)-\left(Y_j-\beta_0-\beta_1X_j\right)\right|$$

$$= \sum_{i<j}\left|\left(Y_i - Y_j\right) - \beta_1\left(X_i - X_j\right)\right|$$

Let $Y_{ij} = Y_i - Y_j$ and $X_{ij} = X_i - X_j$, $\quad i < j$.

Then we have

$$\sum_{i<j}\left|d_i - d_j\right| = \sum_{i<j}\left|\left(Y_{ij} - \beta_1 X_{ij}\right)\right|$$

Then the objective function in this case is similar to that of the minimizing mean absolute deviations regression, with β_1 alone to be estimated, as the β_0 term cannot be estimated using this method. So, we can apply the procedure developed for MINMAD with $n(n-1)/2$ data points Y_{ij}, X_{ij} to solve this problem. One way of obtaining an estimate of β_0 is to force the line to pass through $\left(\overline{X}, \overline{Y}\right)$ and take the corresponding constant term as an estimate of β_0; that is, $\hat{\beta}_0 = \overline{Y} - \overline{X}\hat{\beta}_1$.

There are other ways to estimating β_0. One such estimate is

$$\hat{\beta}_0 = \underset{i<j}{\text{Median}}\,\frac{1}{2}\left(Y_i + Y_j\right)$$

Minimizing Sum of Absolute Differences between Absolute Deviations (MINSADBAD Regression)

We consider the estimation of parameters β_0 and β_1 in the simple regression by minimizing the sum of absolute differences between absolute deviations.

$$\underset{\beta_0,\beta_1}{\text{Minimize}} \sum_{i<j}\left|\left|d_i\right| - \left|d_j\right|\right|$$

Where, $d_i = Y_i - (\beta_0 + \beta_1 X_i)$

Least Squares Estimation with Consistent Linear Restrictions

Suppose we wish to estimate β using the least-squares method where β is subjected to consistent linear-equality restrictions. Then, we have the problem

$$\text{Minimize } \varepsilon'\varepsilon \tag{2.9}$$

Subject to $A\beta = C$

Where A is a known $q \times p$ matrix of rank r, C is a known column vector, and $\varepsilon'\varepsilon$ is defined as before. This problem is approached using Lagrangian multipliers.

The Lagrangian function is

$$L = \varepsilon'\varepsilon + (\beta'A' - C')\lambda \tag{2.10}$$

Now, equating $\frac{\partial L}{\partial \beta} = 0$ and $\frac{\partial L}{\partial \lambda} = 0$, we obtain,

$$\left.\begin{aligned} -2X'Y + 2X'X\beta + A'\lambda = 0 \\ A\beta = C \end{aligned}\right\} \tag{2.11}$$

Let the solution of equ. (2.11) be $\hat{\beta}_R$ and $\hat{\lambda}_R$, the subscript R denoting the restricted problem.

$$\hat{\beta}_R = (X'X)^{-1}X'Y - \frac{1}{2}(X'X)^{-1}A'\lambda_R \tag{2.12}$$

$$= \hat{\beta} - \frac{1}{2}(X'X)^{-1}A'\hat{\lambda}_R$$

As $\hat{\beta}_R$ satisfies $A\beta = C$

$$C = A\hat{\beta}_R$$

$$= A\hat{\beta} - \frac{1}{2}(X'X)^{-1}A'\hat{\lambda}_R$$

We have assumed that A is of rank r, $(X'X)^{-1}$ is positive definite as $(X'X)$ is positive definite. And so $A(X'X)^{-1}$ is also positive definite. Hence,

$$-\frac{1}{2}\hat{\lambda}_R = \left[A(X'X)^{-1}A'\right]^{-1}\left(C - A\hat{\beta}\right) \tag{2.13}$$

Thus, $\hat{\beta}_R$ can be obtained from equ. (2.11) as

$$\hat{\beta}_R = \hat{\beta} + (X'X)^{-1}A'\left[A(X'X)^{-1}A'\right]^{-1}\left(C - A\hat{\beta}\right) \tag{2.14}$$

It can be shown that $\hat{\beta}_R$ actually is optimal. It is known that $\hat{\beta}$ and $\hat{\beta}_R$ are unbiased estimates of β, if E(ε) = 0. Further, if the variance-covariance matrix of ε is given by $\sigma^2 I_n$, then for any linear combination, $a'\hat{\beta}$ is the minimum variance unbiased linear estimate of $a'\beta$ for every a. Moreover, if the $\varepsilon \sim N(0, \sigma^2 I_n)$, maximizing the likelihood function is equivalent to minimizing the quantity $\varepsilon'\varepsilon$. So, in this case least squares estimate $\hat{\beta}$ is also the maximum likelihood estimate.

Minimizing Mean Absolute Deviations Regression

Consider the problem of minimizing $\Sigma|d_i|$ with respect to β where d_i is the deviation from the observed and predicted values of Y_i, the ith observation. The problem is the same as minimizing mean absolute deviation; it is alternatively known as the L_1 – norm minimization problem. The problem can be stated as follows:

Minimize $\Sigma|d_i|$ (2.15)

Subject to $X\beta + d = Y$

d, β unrestricted in sign.

Noting the fact that $|d_i| = d_{1i} + d_{2i}$. When d_{1i} and d_{2i} are non-negative and $d_i = d_{1i} - d_{2i}$, we can reformulate the problem as:

Minimize $\Sigma d_{1i} + \Sigma d_{2i}$

Subject to $X\beta + d_1 - d_2 = Y$

β unrestricted in sign

$$d_1, d_2 \geq 0 \tag{2.16}$$

For detailed study on these methods, refer to Arthanari and Dodge (1981).

3. LS Approach

Consider the model

$$Y = X\beta + e, \quad e \sim N(0, \sigma^2 I) \tag{3.1}$$

Where Y is a $n \times 1$ vector of observed values of dependent variables.

X is a $n \times k$ matrix of values of independent variables and,

e is the vector of random variables assumed to have a $N(0, \sigma^2 I)$ distribution.

Under LS $\hat{\beta}$ is the estimate which minimize

$$(Y - X\beta)'(Y - X\beta) \tag{3.2}$$

and $\hat{\beta}_H$ is the estimate which minimizes

$$(Y - X\beta)'(Y - X\beta) \text{ under H} \tag{3.3}$$

$$\text{Let } B' = (Y - X\hat{\beta})'(Y - X\hat{\beta}) \tag{3.4}$$

and $A' = (Y - X\hat{\beta}_H)'\ (Y - X\hat{\beta}_H)$

then

$$C'\left[\frac{A' - B'}{B'}\right] \tag{3.5}$$

has a Snedecor's F under the hypothesis $\beta = 0$. Where C' is a proper constant.

4. LAD Approach

If the assumption that errors are normally distributed appears to be violated, and if any prior knowledge is available about error distribution, the maximum likelihood argument could be used to obtain estimates based on the criterion other than LS, namely, minimization of the sum of absolute errors $\sum_{i=1}^{n} |e_i|$, where e_i's are components of $e = Y - X\beta$. Let $\tilde{\beta}$ denote the estimates obtained through LAD method by minimizing $\Sigma|e_i|$. Let $B = \Sigma|\tilde{e}_i|$, where e_i's are the components of $(Y - X\tilde{\beta})$. Let $\tilde{\beta}_H$ be the LAD estimates obtained through LAD method by minimizing $\Sigma|e_i|$, subject to H. Let $A = \Sigma|\tilde{e}_{iH}|$, where $\tilde{e}_{iH}$'s are the components of $(Y - X\tilde{\beta}_H)$.

Rao *et al.* (1990) have shown that

$$C\left[\frac{A-B}{A}\right] \tag{4.1}$$

has an asymptotic χ^2 distribution. Where C is a proper constant. For a detailed computational algorithm, refer to Arthanari and Dodge (1981), the results are given in Table 1.

Example 4.1 : The following examples provide an empirical behaviour of the classical least square and the LAD for the simple regression model and assuming the error distribution as log-normal, normal, exponential and chi-square. A simplex Linear Programming (LP) algorithm given by Arthanari and Dodge (1981) was used to make LAD regression estimates. For generating random numbers, the algorithm suggested by Rubinstein (1981) was used.

Log Normal Distribution

Let X be from N(μ, σ^2). Then Y = e^x has the lognormal distribution with p.d.f.

$$f_y(y)=\begin{cases}\dfrac{1}{\sqrt{2}\,\pi\sigma_y}\,e^{\left[\frac{-(\ln y-\mu)^2}{2\sigma^2}\right]}, & 0<y<\infty\\ 0 \quad \text{otherwise}\end{cases}$$

Algorithm

1. Generate Z from N (0,1)
2. $X \leftarrow \mu + \sigma Z$
3. $Y \leftarrow e^x$
4. Deliver Y

For simulating data and testing, where in to get MINMAD results, Arthanari's algorithm was used and the results were shown in Table 1.

Normal Distribution

Let X be a random variable from N (μ, σ^2) with p.d.f.

$$f_x(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty<x<\infty$$

Here μ is the mean and σ^2 is the variance.

Since X = $\mu + \sigma Z$, where Z is the standard normal variable denoted by N (0,1).

Algorithm

1. Generate U from U(0,1)
2. $X \leftarrow F_x^{-1}(U)$
3. Deliver X.

The inverse transform method cannot be applied to the normal distribution and some alternative procedures have to be employed. More about generation from normal distribution can be found in Fishman (1978).

For simulating data and testing, where in to get MINMAD results, Rubinstein (1981), Arthanari and Dodge (1981), algorithm is used and the results are shown in Tables 2 and 3.

Exponential Distribution

Let X be an exponential variable with p.d.f.

$$f_x(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & 0 \le x \le \infty, \ \beta > 0 \\ 0, & \text{Otherwise} \end{cases}$$

denoted by exp (β)

By inverse transform method

$$U = F_x(x) = 1 - e^{-x/\beta}$$

So that

$$X = -\beta \ 1n \ (1-U)$$

Since 1 – U is distributed in the same way as U, we have

$$X = -\beta \ 1n \ U.$$

Algorithm

1. Generate U from U(0,1)
2. $X \leftarrow -\beta \ 1n \ U$
3. Deliver X

For stimulating data and testing, where in to get MINMAD results, Rubinstein (1981), Arthanari and Dodge (1981) algorithm's is used and results are shown in Tables 2 and 3.

Chi-square distribution

Let $Z_1, ..., Z_k$ be from N (0,1). Then

$$Y = \sum_{i=1}^{k} Z_i^2 \tag{4.1}$$

has the chi-square distribution with k- degrees of freedom and is denoted by χ^2 (k).

For generating a chi-square variate from χ^2 (k) is to generate k standard normal random variables and then apply equ. (4.1).

Consider two cases

Case I: If k is even, then Y can be computed as

$$Y = -2 \ 1n\left(\underset{i=1}{\overset{k/2}{\pi}} U_i\right) \tag{4.2}$$

Formula in equ. (4.2) requires k/2 uniform variates compared to k in equ. (4.1). It also requires one logarithmic transformation, compared to k logarithmic and k cosine or sine transformations for generating Z_i from N (0,1), i = 1, 2, ..., k.

Case II: If k is odd, then

$$Y = -2\ 1n\left(\underset{i=1}{\overset{\frac{k}{2}-\frac{1}{2}}{\pi}} U_i\right) + Z^2 \tag{4.3}$$

Where Z is from N (0, 1) and U_i from $\vartheta(0, 1)$. For k > 30 the normal approximation for chi-square variates can be used based on the following formula suggested by Naylor (1966).

$$Z = \sqrt{2Y} - \sqrt{2k-1}$$

Solving for Y, the chi-square variate, we obtain

$$Y = \frac{\left(Z + \sqrt{2k-1}\right)^2}{2} \tag{4.4}$$

For stimulating data and testing, where in to get MINMAD results, Rubinstein (1981), Arthanari and Dodge (1981) algorithm's is used and the results are shown for simple LAD and LS regression, multiple LAD and LS regression given in Tables 2 and 3.

Table 1 : For testing $H_0 : \beta = 0$ and for n = 200 in simple LAD and LS regression ($Y = X\beta + e$); lognormal with $\mu = e^{-0.5}$.

Error distribution		α for LS		α for LAD	
	β	Prob.	F	Prob.	Chi-square
Lognormal ($\sigma^2 = 0.1$)	0	0.025	1.9396	.194	11.1120
	.5	0.043	2.4096	.139	12.3546
	1	0.256	2.128	.168	5.3185
Lognormal ($\sigma^2 = 2.7$)	0	.039	1.2798	.3199	10.0980
	.5	.223	1.132	.358	5.6804
	1	.545	.397	.681	3.0760
Lognormal ($\sigma^2 = 3.5$)	0	.019	1.5898	.2495	11.7586
	.5	.1635	.4452	.6439	6.5051
	1	.304	.3961	.6806	4.8290

For the data generated, both F test and the chi-square statistic, given in equ. (3.5) and equ. (4.1) were computed and their probabilities under the null hypothesis were calculated. For the latter statistic, the asymptotic distribution was assumed to be true, even in small samples. In other words, the χ^2 test adopted is not an exact test, but an asymptotic one. For computing the χ^2 statistic, LAD have to be computed, after computing the estimates of the parameters. For simulating data and testing, where in to get LAD results, Arthanari's algorithm was used. From Table 1, it is observed that Rao's statistic is quite sensitive, when the values of the difference in population means and standard deviations are very close to each other and also when the value of the standard deviation is greater than the maximum difference in means. From Tables 2 and 3 it is clear that LAD approach shows better results compared to LS approach for the given level α.

Table 2 : For testing $H_0 : \beta = 0$; for $\beta_1 = 0, 0.5$ and 1; for n = 100 n =200 in simple LAD and LS Regression ($Y=\beta_0 + \beta_1 X_1 + \varepsilon$); where X = 1, 2, ..., 7.

Error distribution	β_1	n = 100				n = 200			
		α for LAD		α for LS		α for LAD		α for LS	
		0.01	**0.10**	**0.01**	**0.10**	**0.01**	**0.10**	**0.01**	**0.10**
Normal ($\sigma^2 = 4$)	0	0.003	0.103	0.001	0.101	0.005	0.099	0.015	0.137
	0.5	0.483	0.791	0.603	0.895	0.757	0.955	0.897	0.989
	1	0.987	0.999	0.997	0.999	0.999	0.999	0.999	0.999
Exponential ($\sigma^2 = 4$)	0	0.011	0.095	0.007	0.109	0.007	0.095	0.009	0.103
	0.5	0.765	0.971	0.617	0.897	0.995	0.991	0.913	0.981
	1	0.999	0.999	0.989	0.999	0.999	0.999	0.999	0.999
Chi-square ($\sigma^2 = 4$)	0	0.013	0.111	0.013	0.115	0.017	0.103	0.103	0.109
	0.5	0.755	0.937	0.629	0.885	0.961	0.997	0.901	0.991
	1	0.997	0.999	0.993	0.999	0.999	0.999	0.999	0.999

Table 3 : For testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ and $H_0 : \beta_3 = 0/\beta_1, \beta_i$; for $\beta_3 = 0, 0.5$ and 1; $\beta_i = \beta_2 = 0$; in multiple LAD and LS regressions ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$), for n = 100 and x uniformly distributed (1,7).

Error distribution	β_3	$H_0 : \beta_i = \beta_2 = \beta_3 = 0$				$H_0 : \beta_3 = 0/ \beta_i, \beta_2$			
		α for LAD		α for LS		α for LAD		α for LS	
		0.01	**0.10**	**0.01**	**0.10**	**0.01**	**0.10**	**0.01**	**0.10**
Normal ($\sigma^2 = 4$)	0	0.007	0.073	0.005	0.101	0.015	0.115	0.009	0.117
	0.5	0.419	0.739	0.543	0.851	0.581	0.895	0.735	0.943
	1	0.989	0.999	0.997	0.999	0.995	0.999	0.999	0.999
Exponential ($\sigma^2 = 4$)	0	0.009	0.093	0.007	0.099	0.013	0.101	0.009	0.089
	0.5	0.793	0.949	0.625	0.871	0.880	0.971	0.783	0.937
	1	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Chi-square ($\sigma^2 = 4$)	0	0.011	0.117	0.011	0.099	0.007	0.135	0.003	0.099
	0.5	0.763	0.935	0.579	0.875	0.875	0.997	0.759	0.949
	1	0.999	0.999	0.997	0.999	0.999	0.999	0.999	0.999

5. Comparing LS and LAD

For a number of reasons, LS regression analysis is perhaps the single most widely used statistical technique by practitioners in industry, government and academia. Despite the high regard in which LS estimation is held in the statistical theory, there are situations where in other criteria are more appropriate for parameter estimation in simple linear models. An important problem involves the identification and handling of observation that are "outliers." Whatever the sources of the anomalies, it is usually desirable that outliers be identified and that they not have an unduly large influence on model parameter estimates. Least – squares estimation falls short on both counts. A number of estimations procedures which are more robust to departures from the usual least squares assumptions have been discussed in the statistical literature Adichi (1967), Andrews (1974), David (1976), Huber (1973), Wagner (1959). One method, LAD estimation, the most promising for applied work due to a combination of robustness properties and computational case, some interesting results are also found in Weiss (1988).

LAD Estimation

The model under consideration will be of the form

$Y = X\beta + \varepsilon$

Where,

Y is an $n \times 1$ vector of observations on the regress and

X is an $n \times p$ matrix of values of the p regressors

β is a $p \times 1$ vector of parameters and

ε is an $n \times 1$ vector of random disturbances.

Residuals are defined as

$$e = (e_1, e_2, \ldots, e_n)' = Y - X\hat{\beta},$$

Where $\hat{\beta}$ is the estimator of β.

The L_h estimators of β is the $\hat{\beta}$ that minimizes $\sum_{i=1}^{n} |e_i|^h$.

It is to be noted that LAD (h=1) and OLS (h=2) are special cases of L_h estimation.

6. Results

The following simulation results constitute a more systematic investigation of LS versus LAD in regression models, with disturbance taken from a normal distribution having zero mean and a given standard deviation σ^*. Disturbances associated with ordinary points follow a normal distribution with standard deviation σ, where $\sigma<\sigma^*$. For simulation Rubinstein (1981) algorithm is used. Consider a model of the form

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \tag{6.1}$$

in which

$\beta_1 = 6.0$; $\beta_2 = 3.0$; $\beta_3 = -2.0$

This will be referred to as the three variable model, each of the two non-constant regressions, X_{i2} and X_{i3} took an integer values from -3 to 3 in all combinations giving, a complete balanced design with n=25 for each replication of the experiment, 25 disturbances were generated normal pseudo-random number generator with zero mean. For simulating normal pseudo-random number with zero mean, the algorithm suggested by Rubinstein (1981) is used. The standard deviation, σ, for ordinary disturbances is equal to 1. From the disturbances, values of the dependent variable are calculated according to equ. (6.1) and the parameters, β_1, β_2 and β_3 were estimated by each of the three methods LAD, OLS and Generalized Least Squares (GLS). If the outliers are associated with extreme values of the regressors the effect on the estimates will be greater that if they are associated with middle values. To remove this source of variation, the variances of disturbances were permitted randomly after each 10 replications. Each sample result was based on 100 replications. Denote the sample standard deviations of $\left(\hat{\beta}_j\right)$ when estimated by OLS and LAD as $\hat{\sigma}_{OLS}\left(\hat{\beta}_j\right)$ and $\hat{\sigma}_{LAD}\left(\hat{\beta}_j\right)$ respectively. These sample S.D are calculated according to the usual formula

$$\hat{\sigma}\left(\hat{\beta}_j\right) = \left[\sum_{k=1}^{N}\left(\hat{\beta}_{jk} - \hat{\beta}_j\right)^2 / (N-1)\right]^{1/2}$$

Where $\hat{\beta}_{ij}$ is the kth sample estimate of β_j, N is the number of sample estimates and

$$\hat{\beta}_j = \frac{1}{N}\sum_{k=1}^{N}\hat{\beta}_{jk}.$$

An obvious alternative would be to base the estimate on deviations from the true population means, β_j, rather than from the sample means. The difference $\beta_j - \hat{\beta}_j$ were so that applying the alternative method of measuring deviations always produced standard deviation estimates within 0.1 percent of $\hat{\sigma}\left(\hat{\beta}_j\right)$. For each of the p parameters and each estimation method, the sample standard deviation is calculated from 1000 parameter estimates $\hat{\beta}_j$, the sample standard deviation for $\hat{Y}$ is calculated as

$$\hat{\sigma}\left(\hat{Y}\right) = \left[\frac{1}{nN-1}\sum_{i=1}^{n}\sum_{k=1}^{N}\left\{\hat{Y}_{ik} - \mu_i - \left(\overline{\overline{Y}} - \overline{\mu}\right)\right\}^2\right]^{1/2}$$

Where,

$$\hat{Y}_{jk} = \text{the kth sample value of } \hat{Y}_i\,; \qquad \mu_i = \sum_{j=1}^{P}\beta_j X_{ij}$$

$$\overline{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mu_i\;; \quad \text{and} \quad \overline{\overline{Y}} = \frac{1}{nN}\sum_{i=1}^{n}\sum_{k=1}^{N}Y_{ik}.$$

Thus, the standard deviation of $\hat{Y}$ for each estimation method is based on 100n (2500 in this case), is a estimates of Y_i. The relative efficiency of LAD estimator of β_j relative to the LS is then given by

$$\frac{\hat{\sigma}_{LS}\left(\hat{\beta}_j\right)}{\hat{\sigma}_{LAD}\left(\hat{\beta}_j\right)}$$

And similarly for $\hat{Y}$, since both estimation methods give unbiased estimators, relative efficiency offers a good basis for comparing the methods values greater than 1.0 indicate

LAD is outperforming OLS, atleast in this relative variance sense. The results were shown in Tables 4 to 11 and the comparison of the three estimation methods based on the standard deviation $\hat{Y}$ is shown Fig. 1.

Table 4 : Relative efficiencies of LAD estimator for a three variable model in equ. (6.1) with n=25 and $\sigma^* = 10$.

No. of outliers	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{Y}$
0	0.92	0.93	0.92	0.92
1	1.80	1.84	1.86	1.83
2	2.37	2.32	2.47	2.39
3	2.77	2.61	2.75	2.71
4	2.75	2.68	2.97	2.80
5	2.99	2.88	3.00	2.96
6	3.00	2.98	3.00	2.99
7	2.99	3.00	2.99	3.00
8	2.73	2.78	2.80	2.77
9	2.71	2.68	2.73	2.71
10	2.61	2.71	2.61	2.64
11	2.58	2.57	2.63	2.59
12	2.47	2.52	2.57	2.52
13	2.20	2.23	1.98	2.14
14	2.10	2.03	1.78	1.97
15	1.80	1.86	1.73	1.80
16	1.56	1.54	1.57	1.55
17	1.32	1.38	1.43	1.38
18	1.29	1.23	1.19	1.24
19	1.21	1.19	1.08	1.16
20	1.15	1.04	1.03	1.07
21	0.93	0.96	0.96	0.95
22	0.89	0.87	0.89	0.88
23	0.84	0.86	0.83	0.84
24	0.79	0.84	0.81	0.81
25	0.73	0.80	0.79	0.77

Table 5 : Sample standard deviation $\hat{Y}$ for three Estimation methods for model in equ. (6.1) with n=25 and $\sigma^* = 10$.

No. of outliers	OLS	GLS	LAD
0	0.41	0.41	0.49
1	0.88	0.47	0.55
2	1.06	0.49	0.57
3	1.21	0.50	0.61
4	1.32	0.52	0.64
5	1.48	0.54	0.68
6	1.51	0.56	0.71
7	1.84	0.57	0.73
8	1.89	0.59	0.76
9	1.92	0.61	0.78
10	2.22	0.63	0.84
11	2.31	0.70	0.86
12	2.35	0.72	0.92
13	2.48	0.74	1.11
14	2.50	0.76	1.23
15	2.61	0.79	1.43
16	2.73	0.80	1.71
17	2.84	0.86	1.74
18	2.89	0.91	1.79
19	3.00	0.93	2.04
20	3.06	1.03	2.42
21	3.11	1.50	2.89
22	3.20	1.69	3.23
23	3.27	2.23	3.40
24	3.34	2.68	3.80
25	3.60	3.48	4.32

From Table 4 it is obvious that the LAD estimation is superior to OLS as the number of outliers ranges between 1 to 19. When the homoscedasticity assumption is satisfied, the efficiency of LAD estimates is about 92 percent. Table 5 gives actual values of sample standard deviations of $\hat{Y}$ for the three estimation methods and the data listed in Table 5 are also graphically depicted and is shown in Fig. l. It is observed that when the variances of the disturbances are known, GLS gives best estimate. When the variances of the disturbances are not known, LAD estimation looks to be a good choice. Table 6 shows that with a moderate number of outliers as compared to Table 1, the advantages of LAD over LS estimators increases as σ^* increases. Even with $\sigma^* = 5$ the advantage of LAD estimator is

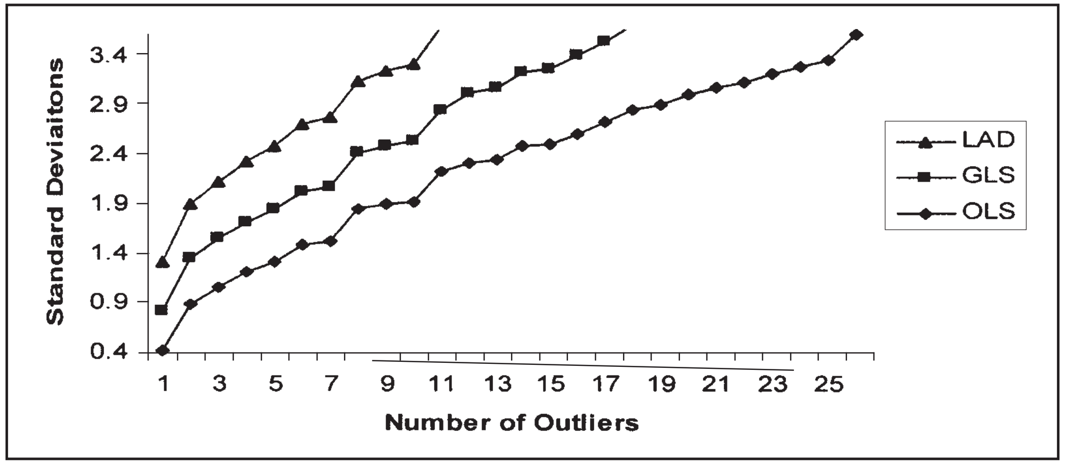

Fig. 1 : Sample standard deviation of $\hat{Y}$ for three estimation methods.

Table 6 : Relatives efficiencies of LAD estimators, $\hat{Y}$ for the model in equ. (6.1) with n=25 and σ^* varied.

No. of outliers	$\sigma^* = 5$	$\sigma^* = 10$	$\sigma^* = 25$
0	0.89	0.92	0.91
1	1.16	1.83	3.93
2	1.37	2.39	5.35
3	2.47	2.71	6.01
4	2.60	2.80	6.72
5	2.66	2.96	6.96
6	2.71	2.99	2.99
7	2.78	3.00	7.02
8	2.73	2.77	6.54
9	2.01	2.71	5.78
10	1.93	2.64	5.54
11	1.78	2.59	5.18
12	1.60	2.52	4.99
13	1.53	2.14	4.58
14	1.38	1.97	3.53
15	1.32	1.80	2.99
16	1.29	1.55	2.18
17	1.22	1.38	1.99
18	1.18	2.24	1.67
19	1.14	1.16	1.32
20	1.03	1.07	1.05
21	0.94	0.95	1.01
22	0.89	0.88	0.96
23	0.84	0.84	0.93
24	0.80	0.81	0.88
25	0.78	0.77	0.79

Table 7 : Relatives efficiencies of LAD estimators, $\hat{Y}$ for the models in equ. (6.1) with σ^* =10 and a varied number of observations.

No. of outliers	n=16	n=25	n=49
0	0.88	0.92	0.89
1	2.05	1.83	1.53
2	2.62	2.39	1.89
3	2.85	2.71	2.09
4	2.71	2.80	2.21
5	2.76	2.96	2.56
6	2.83	2.99	2.61
7	2.03	3.00	2.93

Table 8 : Relatives efficiencies of LAD estimators, $\hat{Y}$ for the models in equ. (6.1) with n=25 and σ^*= 10 comparing balanced designs.

No. of outliers	Random design	Balanced design
0	0.92	0.92
1	0.78	1.83
2	2.18	2.39
3	2.65	2.71
4	2.84	2.80
5	2.94	2.96
6	2.96	2.99
7	2.85	3.00

Table 9 : Relative efficiencies of LAD estimators, $\hat{Y}$ for n=25 and $\sigma^{*} = 10$ and a varied number of regressors.

No. of outliers	p=2	p=3	p=4	p=5
0	0.88	0.92	0.92	0.88
1	0.76	1.83	1.85	1.85
2	2.45	2.39	2.43	2.38
3	2.58	2.71	2.80	2.80
4	2.86	2.80	2.83	2.81
5	2.92	2.96	2.90	2.83
6	3.06	2.93	2.93	2.87
7	2.96	3.00	2.98	2.86

Table 10 : Relatives efficiencies of LAD estimators, $\hat{Y}$ for the models in equ. (6.1) with n=25 and standard deviations uniformly over various intervals.

Range of S.D's	Relative efficiency
1.0 - 5.0	0.96
1.0 - 10.0	1.11
1.0 - 25.0	1.23
1.0 - 50.0	1.36
1.0 - 75.0	1.45
1.0 - 100	1.69
0.5 - 5.0	1.49
0.1 - 1.0	1.40
0.1 - 2.0	1.38
0.1 - 3.0	1.21
0.1 - 4.0	1.13
0.1 - 5.0	1.47

Table 11 : Relative efficiencies of LAD estimators when OLS assumptions are satisfied.

p	n	σ	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{Y}$
2	16	1.0	0.991	0.972				0.992
2	16	10.0	0.983	0.931				0.957
2	25	1.0	0.953	0.936				0.944
3	16	0.5	0.912	0.923	0.914			0.916
3	16	1.0	0.831	0.801	0.812			0.815
3	16	10.0	0.840	0.811	0.822			0.824
3	25	0.5	0.891	0.882	0.792			0.855
3	25	1.0	0.831	0.822	0.834			0.829
3	25	10.0	0.792	0.813	0.822			0.809
3	49	1.0	0.853	0.881	0.891			0.875
3	49	10.0	0.801	0.805	0.808			0.805
4	25	1.0	0.813	0.876	0.875	0.812		0.844
5	25	1.0	0.812	0.803	0.813	0.822	0.813	0.813

clear. Table 7 gives the relative efficiencies for upto seven outliers. Table 7 shows that in the presence of a smaller number of outliers, the relative advantage of LAD estimation become greater as the number of data points decreases. Table 8 compares the efficiencies of the model with balanced design and randomized design. Randomization of the design does not appear to have any importance influence on the efficiencies. Table 9 gives the results for two, three, four and five independent variables. Table 9 reveals that little variations in the model, especially in the presence of outliers. This result shows that LAD estimation can be useful compared to LS. Table 10, shows relative efficiencies of LAD estimators for various ranges of standard deviations. Although LAD estimation is still more efficient for large ranges of σ, it does not achieve the great advantages in the cases involving clear-cut outliers. Table 11 is based on least 100 replications. A few values were estimated from as many as 2500 replications in order to establish with high confidence that all the efficiencies are not equal.

7. Conclusion

It is observed that from the example, Rao's statistic is quite sensitive, when the values of the difference in population means and standard deviations are close to each other and also when the value of standard deviations is greater that the maximum difference in means. LAD estimators are not inefficient to LS estimator with LS assumptions are satisfied, but are dramatically more efficient in many situations where large disturbances are present. It was found that regardless of the number of regressors, the number of observations, the standard deviation of the disturbances, when the LS assumptions including normality assumptions were satisfied, the efficiency of LAD estimators relative to OLS estimators varies only 92 percent. The lowest efficiency for any LAD estimator was about 78 percent. Hence it is worthwhile to compare the exact probabilities of the test statistic and to a final judgment. It was also absorbed that LAD procedure, on the whole, provide the most attractive preliminary estimator for robust regression when compared to LS estimator. LAD estimator, making it particularly attractive relative to LS when the regression process is though to be particularly long-tailed.

Acknowledgement

The authors thank to Prof. V. Ramachandran (Retd.) and Prof. R. Sathiyamoorthi (Retd.), Department of Statistics for their valuable suggestion in completing this paper.

References

Andrews, D. F. (1974). A Robust Method of Multiple Linear Regression. *Technometrics*, **16**, 523-531.

Athanari, T. S. and Y. Dodge (1981). *Mathematical Programming in Statistics.* John Wiley and Sons, New York.

Bai, Z. D. and C. Radhakrishna Rao (1990). Least Absolute Analysis of Variance. *Sankhya*, **52**, Series B, Part 2, pp. 166-177.

Brecht, H. David (1976). Regression Methodology with Gross Observations Errors in the Explanatory Variables. *Decision Sciences*, **7**, 57-65.

Draper, N. R and H. Smith (1981). *Applied Regression Analysis*. John Wiley and

Draper, N. R, and W. G. Hunter (1969). Transformations: Some Examples Revisited. *Technometrics*, **11**, 23-40.

Faica, D. M. and S. M. Assis (1988). Linear Programming State Estimation Error Analysis and Gross Error Identification. *IEEE Transaction on Power Systems*, **3(3)**, 809-815.

Fishman, G. S. (1978). *Principles of Discrete Event Simulation*, Wiley, New York.

Huber, Peter J. (1973). *Robust regression: Asymptotics*, Conjectures and Monte Carlo. *Annals of Statistics*, **1**, 799-821.

Montgomery, D. C. (1977). *Design and Analysis of Experiments*. John Wiley and Sons, New York.

Naylor, T. H. *et al.* (1966). *Computer Simulation Techniques*, Wiley, New York.

Raghukandan, K., M. Rathinasabapathi and P. V. Vaidyanathan (1998). Modelling of Process in Dynamic Form-Cladding. *Metals and Materials*, **4(5)**, 1057-1061.

Robinstein, R. Y. (1981). *Simulation and Monte Carlo Method.* John Wiley and Sons, New York.

Searle. S. R. (1971). *Linear Models*. 2nd Ed., John Wiley and Sons., New York. Sons, New York.

Wagner, M. Harvey (1959). Linear Programming Techniques for Regression Analysis. *Journal of the American Statistical Association*, **54**, 206-212.

Weiss, A. (1988). A Comparison of Ordinary Least Squares and Least Absolute Error Estimation. *Econometric Theory*, **14(3)**, 517-527.