# Least absolute deviation estimation of linear econometric models: A literature review

Dasgupta, Madhuchhanda and Mishra, SK

North-Eastern Hill University, Shillong

1 June 2004

# Least Absolute Deviation Estimation of Linear Econometric Models : A Literature Review

**M DasGupta**
**SK Mishra**
**Dept. of Economics**
**NEHU, Shillong, India**

**I. Introduction**: The Least Squares method of estimation of parameters of linear (regression) models performs well provided that the residuals (disturbances or errors) are well behaved (preferably normally or near-normally distributed and not infested with large size outliers) and follow Gauss-Markov assumptions. However, models with the disturbances that are prominently non-normally distributed and contain sizeable outliers fail estimation by the Least Squares method. An intensive research has established that in such cases estimation by the Least Absolute Deviation (LAD) method performs well. This paper is an attempt to survey the literature on LAD estimation of single as well as multi-equation linear econometric models.

Estimation of the parameters of a (linear) regression equation is fundamentally a problem of finding solution to an over-determined and inconsistent system of (linear) equations. The over-determined and inconsistent system of equations cannot have any solution that exactly satisfies all the equations. Therefore, the '*solution*' leaves the equations (not necessarily all) unsatisfied by some quantity (of either sign) called the residual, disturbance or error. It is held that these residuals should be as small as possible and this fact determines the quality of the '*solution*'. It is accomplished by minimization of a particular norm of the residual vector, $\|e\|$, in the sample.

**II. The Origins**: The method to solve an over-determined system of (linear) algebraic equations dates back to **KF Gauss** and **PS Laplace** as mentioned by **Taylor** (1974). These mathematicians suggested (and used) the method of Least Squares, which minimizes the sum of square of residuals in the equation (tantamount to minimization of Euclidean norm of the residuals). They also suggested (and used) the method of Least Absolutes, which minimizes the sum of absolute residuals in the equations (which amounts to minimization of absolute norm of the residuals). In the sense of ***Minkowski norm***, the methods of Least Squares ($L_2$) and Least absolute ($L_1$) are expressed as

$$\text{Min (S)} = \min_{\hat{a}} \left[ \left( \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{k} \hat{a}_j X_{ij} \right|^p \right)^{\frac{1}{p}} \right] \text{ for } p = 2 \text{ and } p = 1, \text{ respectively.}$$

The Least Squares method is computationally convenient because minimization of *Euclidean norm* is amenable to calculus methods. This convenience led to its popularity. On the other hand, there are mathematical difficulties in working with absolute value functions on account of their lack of amenability to calculus methods.

**III. Justification to LAD Estimation**: Econometricians generally take for granted that the error terms in the econometric models are generated by distributions having a finite variance. However, since the time of Pareto the existence of error distributions with infinite variance is known. Works of many econometricians, namely, **Meyer** & **Glauber** (1964), **Fama** (1965) and **Mandlebroth** (1967), on economic data series like prices in financial and commodity markets confirm that infinite variance distributions exist abundantly. The distribution of firms by size, behaviour of speculative prices and various other recent economic phenomena also display similar trends. Further, econometricians generally assume that the disturbance term, which is an influence of innumerably many factors not accounted for in the model, approaches normality according to the Central Limit Theorem. But **Bartels** (1977) is of the opinion that there are limit theorems, which are just likely to be relevant when considering the sum of number of components in a regression disturbance that leads to non-normal stable distribution characterized by infinite variance. Thus, the possibility of the error term following a non-normal distribution exists.

An infinite variance means fat tails. Fat tails also may mean a lot of outliers in the disturbances. Since the method of least squares places heavy weights on the error terms, we look to an alternative, more robust estimator, which minimizes the absolute values and not the squared values of the error term. The Least Absolute Deviation (LAD) estimator, suggested by **Gauss** and **Laplace**, is such an estimator that minimizes the absolute value of the disturbance term. This estimator measures the error term as the absolute distance of the estimated values from the true values and belongs to the median family of estimators.

**Edgeworth** (1887, 1888, 1923), **Rhodes** (1930) and **Singleton** (1940) emphasized $L_1$ approximation (estimation). They pointed out that random sampling and normal distribution, which are needed to justify the method of Least Squares as an optimal method, often do not materialize. In other circumstances least squares may give undue weights to extreme observations. The method proposed by Edgeworth applies to only two variables. The methods of Rhodes and Singleton, while extending the proposals of Edgeworth to more than two dimensions, become extremely unwieldy as the dimension of the model increases.

**IV. The Advent of Operations Research and its Impact on LAD Estimation**: The post World War II era opened with the development Operations Research and Linear Programming. With the development of mathematical methods to solve 'corner-optimum' problem (that defied optimization by the classical calculus methods), feasibility of parameter estimation by minimization of the sum of absolute residuals received a major breakthrough. **Charnes, Cooper** & **Ferguson** (1955) showed that a certain management problem involving a minimization of absolute values could be transformed to standard linear programming form by employing the device of representing a deviation as the difference between two non-negative variables. The paper by **Charnes *et al***. (1955) is considered to be a seminal paper for giving a new lease of life to $L_1$ regression. **Fisher** (1961) showed how a curve can be fitted with minimum absolute deviation (rather than squared deviations) using linear programming. His article reviewed the formulation of this application of linear programming. Fisher showed how to fit a linear function

$$\hat{X}_1 = a_0 + a_{12}X_2 + .... + a_{1K}X_K \qquad \ldots (1)$$

to the observations

$$\begin{bmatrix} X_{11} & X_{12} & ... & X_{1K} \\ X_{21} & X_{22} & ... & X_{2K} \\ ..... & ..... & ... & ... \\ X_{n1} & X_{n2} & ... & X_{nK} \end{bmatrix}$$

( $X_{ij}$ representing observation $i$ on variable $j$, and $n > K$ ) by minimizing the sum of absolute deviations

$$S = \sum_{i=1}^{n} \left| X_{i1} - \hat{X}_{i1} \right|$$

The parameters in equation (1) can be expressed as

$$\begin{aligned} a_0 &= y_1 - Z_1 \\ a_{12} &= y_2 - Z_2 \\ &..... \\ a_{1K} &= y_K - Z_K \end{aligned} \qquad \ldots (2)$$

where *y's* and *Z's* are *2K* non-negative variables that are to be determined. And the residuals for observation *i* be expressed as

$$X_{i1} - \hat{X}_{i1} = E_i - V_i ; (i = 1, 2, ..., n) \qquad \ldots (3)$$

where *E's* and *V's* are *2n* non-negative variables to be determined. Using (1), (2) and (3), the system of n linear equations can be obtained

$$\left.\begin{aligned} &E_1 - V_1 + y_1 - Z_1 + X_{12}y_2 - X_{12}Z_2 + .... + X_{1K}y_K - X_{1K}Z_K = X_{11} \\ &E_2 - V_2 + y_1 - Z_1 + X_{22}y_2 - X_{22}Z_2 + .... + X_{2K}Y_K - X_{2K}Z_K = X_{21} \\ &..... \\ &E_n - V_n + y_1 - Z_1 + X_{n2}Y_2 - X_{n2}Z_2 + .... + X_{nK}y_K - X_{nK}Z_K = X_{n1} \end{aligned}\right| \qquad \ldots (4)$$

Since the problem of finding positive values of E*'s*, *V's*, *y's* and *Z's* satisfying system (4) and minimizing the linear form

$$R = \sum_{i=1}^{n} (E_i + V_i) \qquad \ldots (5)$$

is a linear programming problem in *2(n+K)* variables in *n* constraints, the simplex method of **Dantzig** may be applied directly and the solution obtained. For automatic computation a suitable unit basis for the first stage can be formed from the set of E*'s* and *V's*. At the minimum *R=S*.

Fisher opined that the method of fitting a linear regression function by minimizing the sum of absolute deviation is also flexible. In support of this, he pointed out that the regression function (1) could be constrained further by adding additional conditions as additional equations to the system of constraints (4). If it is desired to specify certain parameters of (1) as positive, the original parameters without conversion to a difference

can be used as a variable in the linear programming problem. A non-linear regression function is additive in the sense that it can be written in the form

$$X_1 = f_2(X_2) + \ldots\ldots + f_K(X_K).$$

It can be transformed into a linear function by making transformations of the independent variable as is done in least squares regression. If unequal weighting of the observed data is desired in the fitting, the desired weights, rather than unit weighting, should be inserted as coefficients in the objective function (5).

**Ashar & Wallace** (1963) studied the statistical properties of regression parameters estimated by minimization of $L_1$ norm. **Huber** (1964) explored the properties of $L_1$ regression in its robustness to wild fluctuations in the magnitude of residual elements.

**Meyer & Glauber** (1964) for the first time directly compared $L_1$ and $L_2$ regression estimators. They estimated their investment model by minimization of $L_1$ as well as $L_2$ norm and tested the regression equations obtained on post-sample data by using those equations to forecast the nine (in some cases eleven) observations subsequent to the period of fit. They found that with very few exceptions, the equations estimated by $L_1$ minimization outperformed the ones estimated by $L_2$ minimization even on criteria (such as sum of squared forecast errors) with respect to which, $L_2$ regression is ordinarily thought to be remarkably suitable or optimal.

**Rice & White** (1964) compared $L_1$,, $L_2$ and $L_\infty$ (minimization of maximum deviation) norms for a single equation model. In their paper they observed that for the important problem of smoothing and estimation in the presence of wild points (outliers), the $L_1$ norm appears to be markedly superior among the $L_p$ $(1 \leq P \leq \infty)$ norms.

**Usow** (1967) studied the $L_1$ approximation for discrete functions and the discretization effects while the functions in original are continuous. This paper is more concerned with approximation of functions using $L_1$ norm rather than estimation of regression parameters and their statistical properties, though its mathematical approach has a significant bearing on the development of statistical theory relating to properties of $L_1$ regression.

In 1973 **Barrodale** & **Roberts** presented an algorithm for $L_1$-approximation by modifying the simplex method of linear programming, which is computationally superior to the alogorithms given by Usow and **Robers & Robers**. The algorithm is an improved version of the primal algorithm described by **Barrodale & Young** in 1966. In the improved version, Barrodale & Roberts were able to significantly reduce the total number of iterations required by discovering how to pass through several neighbouring simplex vertices in a single iteration.

Authors like **Wagner** and **Robinowitz** (for example) have suggested that the dual of the problem should be solved when m is large. However, it is found that an application of bounded-variable simplex method of Dantzig to the dual problem leads to a less efficient algorithm in general than solving the primal problem by their version of the standard form of the simplex method.

Although several alternatives to the standard form of the simplex method can be used to solve a linear programming problem (two forms of the revised simplex method, the primal-dual algorithm, the dual simplex algorithm, etc.), the denseness of the condensed tableau, the availability of an initial basic feasible solution, and the simplicity with which the idea of passing through several vertices in a single iteration can be implemented, combine together to make the standard form of the simplex method the most economical algorithm for $l_1$-problem.

The study conducted by **Oveson** (1968) in his Doctoral research on the LAD estimator gave a new thrust to the investigation into the properties and applicability of the estimator. It was almost fully established that in the presence of errors generated by thick-tailed distribution, $L_1$ regression performed better than $L_2$ regression.

In 1969 **Robers and Ben-Israel** applied a new method for linear programming to the dual formulation of the $l_1$-problem. Their new method (which they called interval linear programming) is capable of solving any bounded-variable linear programming problem, and so it is natural to apply it to the $l_1$-problem in particular.

In 1971 **Abdelmalek** described an algorithm, which determines best $l_1$-approximations as the limit of best *lp*-approximations as $P \to 1^+$. His technique thus obtains a solution to a linear problem by solving a sequence of non-linear problems.

**Pollard** (1991) presented an alternative approach for studying the asymptotic theory of LAD estimator in a simple regression context. The approach was built on the convexity of the LAD criterion function to construct a quadratic approximation whose minimand is close enough to the LAD estimator for the latter to share the same asymptotic normal distribution.

**Taylor** (1974) gave the condition under which the $L_1$ norm estimator is unbiased and consistent and discussed some of the problems encountered when trying to establish a distribution theory, under the assumptions that (i) $e_i$ are independent, identically distributed random variables with a continuous distribution function $F$ and median zero, and (ii) $\lim_{\to\infty}(n^{-1}X'X) = Q$ is a positive definite matrix. **Nyquist** & **Westlund** (1977) compared the two estimators ($L_1$ and $L_2$ norm estimators) with regard to their statistical properties.

In 1978, **Bassett** & **Koenker** developed the asymptotic theory of Least absolute error regression. Their article resolved a long-standing open question concerning the LAD (alias LAE)) estimator by establishing its asymptotic normality under general conditions, thereby extending a result of **PS Laplace** to the general linear model. *The result confirmed that for the general linear model the LAD estimator is a natural analog of the sample median*. The authors proved that in the general linear model with independent and identically distributed errors and distribution function F, the estimator which minimizes the sum of absolute residuals is demonstrated to be consistent and asymptotically Gaussian with covariance matrix $W^2Q^{-1}$, where $\lim_{\to\infty}(n^{-1}X'X) = Q$ and $W^2$ is the asymptotic variance of the ordinary sample median from samples with distribution F.

This indicated that for any error distribution for which median is more efficient than the mean as an estimator of location, the least absolute error estimator has smaller asymptotic ellipsoids than the least squares estimator and therefore is more efficient than LS estimator. In the paper, a number of equivariance properties of LAD estimator was stated and proved. It was proved that the LAD estimator is affine equivariant, scale and shift equivariant and equivariant to reparameterization of design. Though Least Squares estimator shares the same properties, typically robust alternatives to least squares are not equivariant in one or more of the above senses.

It was also observed that in a scatter of sample observations in $\mathbb{R}^2$ with the LAD solution line slicing through the scatter, as long as the moving observations lie on the same side of the original line, the solution is unaffected. This property is not shared by least squares, and although obvious in the case of median, it seems to capture part of the intuitive flavour of LAD's median-type robustness and insensitivity to outlying observations.

**Phillips** (1991) presented the asymptotic theory for the LAD estimator (of a regression model) by using generalized functions of random variables and generalized Taylor series expansions. The approach was justified by the smoothing that was delivered in the limit by the asymptotics, whereby the generalized functions were forced to appear as linear functionals wherein they became real valued. He studied models with fixed random regressors, and autoregressions with infinite variance errors and a unit root. His approach enabled the development of higher order asymptotic expansion of the distribution of the LAD estimator. The results obtained also showed that the LAD estimator converges at a faster rate in the unit root model for $0 < \alpha < 2$ than the OLS estimator.

**Sakata** (2001) proposed a general estimation principle based on the assumption that instrumental variables (IV) do not explain the error term in a structural equation. He opined that unlike the IV estimators such as two-stage least squares estimator, the estimators based on the proposed principle are independent of the normalization constraint. Based on this new principle, he proposed the $L_1$IV estimator, which is an IV estimation counterpart of the LAD estimator. The author investigated the asymptotic properties of the $L_1$IV estimator. A consistent estimator of its asymptotic covariance matrix and a consistent specification test based on the $L_1$IV estimator were proposed. The problem of identification in $L_1$IV estimation was also discussed.

**V. Iterative Algorithms for LAD Estimation**: Once LAD estimation is justified and its edge over the OLS estimation (in an appropriate condition) is established, an efficient algorithm to obtain LAD estimates has a practical significance. A progress in this direction was made by **Spyropoulos, Kiountouzis** & **Young** (1973) and **Abdelmalek** (1974). **Schlossmacher** (1973) and **Fair** (1974) also proposed an improved algorithm for $L_1$ estimation that is very similar to iterative weighted least squares. Given a linear econometric model, $Y = Xa + e$

Step I: Obtain the LS estimates of $\hat{a}$, $\hat{Y}$ and $\hat{e}$ using the formulae

$$\hat{a} = (X'X)^{-1}X'Y; \qquad \hat{Y} = X\hat{a}; \qquad \hat{e} = Y - \hat{Y}$$

Step II: Compute $\hat{W}_{ij} = \dfrac{1}{|\hat{e}_i|}$ for i=j , else $\hat{W}_{ij} = 0$ i, j = 1,2, …,n.

Step III: Compute $\hat{a} = (X'\hat{W}X)^{-1} X'\hat{W}Y$; $\hat{Y} = X\hat{a}$; $\hat{e} = Y - \hat{Y}$

Step IV: If the values of $\hat{a}$ are stable (convergence has been reached to the pre-assigned accuracy) then stop, otherwise go to step II.

The asymptotic variance-covariance matrix of $\hat{a}$ is given (**Taylor**, 1974) by

$$\hat{v} = 0.25\left\{\hat{f}(0)\right\}^{-2}(X'X)^{-1}, \qquad \text{where,} \qquad \hat{f}(o) = (p-q-1)/\{n(\hat{e}_p - \hat{e}_q)\}$$

here p and q are integers such that $p \geq \dfrac{n}{2}+1 = q$ for even n or $p = \dfrac{n}{2}+0.5 = q$ for odd n. Further that $e_1 \leq e_2 \leq e_s \leq .... \leq e_n$ are ordered $L_1$ residuals. Although the best values of p and q cannot be ascertained, it has been suggested that 3n/4 and n/4 are the most appropriate values of p and q respectively as these values are not much affected by extreme values.

A common problem with the iterative least squares procedure is that, in any given iteration, some of the residuals may be zero or very close to zero, thereby, making construction of weights difficult. Fair and Schlossmacher dealt with this problem in different ways. When a residual was less than 0.00001, Fair set it equal to 0.00001 while Schlossmacher ignored the observation, at least for the given iteration, by setting the weight equal to zero. Although the two solutions are to some extent contradictory, both authors reported satisfactory results in their empirical work.

There is yet another method suggested by **Anscombe** (1967). The method minimizes squared deviations for small errors, absolute deviations for moderate errors and rejects observations with large errors. The estimator is obtained by minimizing the weighted least squares function $\sum_{i=1}^{n} w_i (y_i - x_i'a)^2$, where, $w_i = 1$ if $|\hat{e}_i| \leq m_1$; or $w_i = m_1/|e_i|$ if $m_1 < |\hat{e}_i| \leq m_2$; or $w_i = 0$ if $|\hat{e}_i| > m_2$; $m_1$ and $m_2$ are either pre-assigned multiples of the standard deviation of $e_i$ or pre-assigned constants. The steps to be followed for minimizing Equation (3.29) iteratively are:

Step I: Calculate either the LS or LAD estimates for $a$ say $\hat{\alpha}$

Step II: First calculate the errors $\hat{e}_i = y_i - x_i'\hat{\alpha}$ and then the corresponding weights.

Step III: Using weighted least squares $\hat{a} = (X'WX)^{-1} X'Wy$

where $W$ = diagonal $(W_1, W_2 ..... W_n)$, estimate $a$.

Step IV: Replace $\hat{\alpha}$ by $\hat{a}$

Step V: Repeat the process II through V until abs($\hat{\alpha}$ - $\hat{a}$ ) < $\varepsilon$, a very small pre-assigned positive number (for accuracy of estimation).

Along these efforts, a number of works using Monte Carlo method of simulation to compare the sampling properties of $L_1$ regression with new alternatives to $L_2$ norm estimators were done. **Blattberg** & **Sargent** (1971) pointed out that if the disturbances follow a two-tailed exponential distribution with density function

$$f(e_i) = (2\lambda)^{-1} \exp\left\{-\frac{|e_i|}{\lambda}\right\} \qquad \ldots (6)$$

then maximization of the likelihood function is equivalent to minimization of $\sum_{i=1}^{n} |e_i|$ and so the least absolute deviation estimator becomes the maximum likelihood estimator. The superiority of $L_1$ norm estimator over $L_2$ norm estimator in finite samples, when errors follow the density in (6) was confirmed in a Monte Carlo study by **Smith** and **Hall** (1972).

**Huber** (1973) put forward an estimator that minimizes appropriately weighted squared deviations for small residuals, and absolute deviations for large residuals. The estimator minimizes

$$\sum_{i=1}^{n} f\left(y_i - x_i' a\right)$$

where, $f(e_i) = \frac{1}{2} e_i^2$ for $|e_i| < \delta$

$= \delta |e_i| - \frac{1}{2}\delta^2$ for $|e_i| \geq \delta$, and $\delta$ is a pre- assigned constant.

A set of normal equations that can be solved iteratively starting with either LS or LAD estimator for $a$ has been suggested by Huber. However, for any given observation the appropriate functions can change from iteration to iteration.

Combining recent advances in interior point methods for solving linear programs with a new statistical preprocessing approach for $l_1$-type problems, **Portnoy & Koenker** (1997) obtained a 10 to 100 fold improvement in computational speeds over current (simplex-based) $l_1$ algorithms in large problems, demonstrating that $l_1$ methods can be made competitive with $l_2$ methods in terms of computational speed throughout the entire range of problem sizes.

The iterative computational methods of estimation mentioned above often yield results that are close to optimum. It may be useful to apply random walk methods of optimization to refine the results further. The random walk method is based on generating a sequence of improved approximations to the minimum, each derived from the preceding approximation. Thus if $a_i$ is the approximation to the minimum obtained in the $(i-1)^{th}$ stage (or step or iteration), the new or improved approximation in the $i^{th}$ stage is found from the relation

$$a_{i+1} = a_i + \lambda u_i$$

where $\lambda$ is a prescribed scalar step length, and $u_i$ is a unit random vector generated in the $i^{th}$ stage. The detailed procedure of this method is given by the following steps.

1. Start with an initial point $a_i$ and a scalar step length that is sufficiently large in relation to the final accuracy desired. Find the functional value $F_1 = F(a_1)$.
2. Set the iteration number i =1.
3. Generate a set of n random numbers and formulate the unit random vector $u_i$.
4. Find the new value of the objective function as $F = F(a_1 + \lambda u)$.
5. Compare the values of F and $F_1$. If $F < F_i$, set $a_i = a_1 + \lambda u$, and $F_1$=F , and repeat step 3 through 5. If $F \geq F_1$, just repeat step 3 through 5.
6. If a sufficiently large number of iterations (N) cannot produce a better point, $a_{i+1}$, reduce the scalar length $\lambda$ and go to step 3.
7. If an improved point could not be generated even after reducing the value of $\lambda$ below a small number $\varepsilon$, take the current point $a_i$ as the desired optimum point, and stop the procedure.

In the random walk method described above, we proceed to generate a new unit random vector $u_{i+1}$ as soon as we find that $u_i$ is successful in reducing the function value for a fixed step length $\lambda$. However, we may expect to achieve a further decrease in the function value by taking a longer step length along the direction $u_i$. Thus the random walk method can be improved if each successful direction is exploited until it fails to be useful. This can be achieved by using any of the one-dimensional minimization method. According to this procedure, the new point $a_{i+1}$ is found as

$$a_{i+1} = a_i + \lambda_i^* u_i$$

where $\lambda_i^*$ is the optimal step length found along the direction $u_i$ so that

$$F_{i+1} = F(a_i + \lambda_i^* u_i = \min_{\lambda_i} F(a_i + \lambda_i u_i)\,.$$

It has been found (**Dasgupta**, 2004) that the random walk method improves the estimates obtained by Fair-Schlossmacher algorithm.

**VI. Extensions of the LAD Estimator: Powell** (1984) proposed an alternative to maximum likelihood estimation of the parameters of the Censored Regression Model. He generalized the Least Absolute Deviations estimation for the standard linear regression model. The estimator was found by minimizing $\sum \left| y_i - \max\left(0, x_i s \beta\right) \right|$.

In the paper, he showed that the Censored Least Absolute Deviation (CLAD) estimator is robust to heteroskedasticity and is consistent and asymptotically normal for a wide class of error distribution. Consistency of the asymptotic covariance matrix was also proved. As a consequence, tests of hypothesis concerning the unknown regression coefficient can be constructed which are valid in large samples. He also opined that the Censored LAD estimator can be computed using “direct search” methods developed for nonlinear programming.

**Weiss** (1991) established that it was possible to use the LAD estimator to estimate the parameters of a nonlinear dynamic model. He considered a model given by

$$y_t = g(x_t, \beta_0) + e_t$$

where g = a known function

$x_t = (y_{t-1}, ....., y_{t-p}, z_t)$

$z_t =$ vector of exogenous variables

$\beta_0 = (k \times 1)$ vector of unknown parameter

$e_t =$ unobserved error term which satisfies median $(e_t / I_t) = 0$

$I_t = \sigma -$ algebra (information set at period t) generated by $\{x_{l-i}\}(i \geq 0)$ and $\{e_{l-i}\}(i \geq 1)$.

The Nonlinear Least Absolute Deviations (NLAD) estimator was defined as the solution of the problem:

$$\min_{\beta}\{Q_T(\beta)\} \equiv \min_{\beta}\{\frac{1}{T}\sum_{t=1}^{T}|y_t - g(x_t, \beta)|\}$$

The author investigated the model and proved theoretically that the NLAD estimator $\hat{\beta}$ was consistent and asymptotically normal under certain assumptions.

**Chen** (1996) investigated the linear regression model

$$Y_i = x_i'\beta_0 + e_i \; ; \qquad 1 \leq i \leq n, \quad n \geq 1$$

under the assumptions that the random error $e_i$ belongs to a certain class $F$ of distributions in $\mathbb{R}^{\infty}$, that each $e_i$ has a unique median zero and for each $e_i$ there must be at least linear accumulation of probability in the vicinity of zero. He showed that the sufficient condition $d_n \equiv \max_{1 \leq i \leq n} x_i'\left(\sum_{j=1}^{n} x_j x_j'\right)^{-1} x_i = O\left(1/\log n\right)$ for strong consistency of the LAD estimate $\hat{\beta}_n$ of $\beta_0$ given by **Chen et al.** (1992) fails. The author proved that for any constant sequence $D_n \uparrow \infty$, the condition $d_n = O(D_n / \log n)$ is no longer sufficient.

**Breidt, Davis & Trindade** (2000) studied the Least Absolute Deviation estimation for All-Pass time series models. An All-Pass time series model is an autoregressive moving average model in which all the roots of the autoregressive polynomial are reciprocals of roots of the moving average polynomial and vice versa. The uncorrelated (white noise) time series generated by the All-Pass models are not independent in the non-Gaussian case. The authors opined that an approximation to the likelihood of the model in the case of Laplace (two-sided exponential) noise yields a modified absolute deviation criterion, which can be used even if the underlying noise is not Laplacian. They established the asymptotic normality for LAD estimators of the model parameters under general conditions. Behaviour of the estimators in finite samples was also studied via simulation.

**Kim and Muller** (2000) presented the asymptotic properties of two-stage quantile regression estimators. In their paper, they derived the asymptotic representation of the estimators and proved the asymptotic normality with quantile regression predictions. The

asymptotic variance matrix and asymptotic bias were discussed. They also analysed the asymptotic normality and the asymptotic covariance matrix with LS predictions. The results obtained permitted valid inferences in structural models estimated by using quantile regressions, in which the possible endogeneity of some explanatory variables was treated via ancillary predictive equations. Simulation results illustrated the usefulness of this approach.

**Furno** (2000) compared the performance of LAD and OLS in the linear regression model with random coefficient autocorrelated (RCA) errors. The presence of thick tailed error distribution led to the estimation of the RCA model by Least Absolute Deviation (LAD) estimator. It is known that when error follows a double exponential distribution, LAD coincides with maximum likelihood. In all other cases, the estimator is less affected by observations coming from tails, since it minimizes the absolute value and not the squared value of the residuals. In case of leptokurtic error distribution, the LAD estimator is particularly useful. Furno proved that the LAD estimator for randomly autocorrelated errors is asymptotically normal. The more general random coefficient ARMA models for the error term was also considered in the study and the resulting heteroskedasticity was analysed. Monte Carlo experiments revealed that LAD improved upon OLS in case of RCA errors, both in terms of bias reduction and efficiency gains. However, in the case of constant autocorrelation model, the results confirmed that LAD is not advantageous, especially in small samples, since its sampling distribution differs from the asymptotic one.

**Hitomi & Kagihara** (2001) proposed a NSLAD (nonlinear Smoothed LAD) estimator that is practically computable and has the same asymptotic properties as the NLAD estimator in Weiss' (1991) nonlinear dynamic model. Monte Carlo experiments were conducted to compare the performance of the NSLAD and the nonlinear least-squares (NLS) estimators. In the study two types of error distributions were considered – standard normal distribution where the NLS estimator becomes MLE and the Laplace distribution where the NLAD estimator is MLE. The results reported indicate that as the sample size increases the bias becomes negligible and the difference between NSLAD and NLS estimators ceases. While the NLS estimator was found to have a smaller standard deviation when the error term's distribution was standard normal, the NSLAD estimator had a smaller standard deviation when the error term followed Laplace distribution. No difference was found in the performance of the two estimators with respect to median and quartiles. Although NLS had a marginal edge over NSLAD as far as computation time was concerned, NSLAD was found to take relatively lesser time when the error term followed Laplace distribution.

**VII. Estimation of Multi-equation Models by Minimization of (squared) Euclidean Norm**: By the middle of 1960's, multi equation econometric models and techniques used for estimating their parameters had already gained a solid ground. The method of limited information maximum likelihood (LIML) was developed in the late 1940's (**Haavelmo**, 1947). But the use of least squares method for estimation of parameters of a multi-equation econometric model had to wait until **Theil** (1953) used repeated least squares to estimation of parameters of a regression equation in the multi-equation model. **Basmann**

(1957) used least squares repeatedly for estimating parameters of a multi-equation linear econometric model. **Theil** (1961) developed the method of Two-Stage Least Squares (2SLS).

It would be befitting to describe the nature and the issues related with the estimation of multi-equation (linear) models which will help us in pin-pointing the nodes at which the least squares technique may be replaced by LAD.

A multi-equation (linear) system may be described as YA + XB + E = 0, where, Y(n,m) is a matrix representing m number of *endogenous variables* each in n number of observations, X(n,k) is a matrix representing k number of *predetermined* variables each in n observations, E(n,m) is a matrix representing m number of *stochastic vectors* (error terms in the model) each in n elements and 0(n,m) is a null matrix in n rows and m columns. Associated with Y and X there are the coefficient matrices, A(m,m) and B(k,m) respectively, called the *structural coefficients matrices.* It is assumed that the model YA + XB + E = 0 is complete, which implies that the model has as many (linearly independent) equations as the number of endogenous variables and the matrix A is a regular (not singular) matrix. While Y is a matrix of stochastic vectors ( $Y = \mathbb{Y} + \varepsilon$ , where $\mathbb{Y}$ is the matrix of true endogenous variables and $\varepsilon$, different from E, is the matrix of disturbances), some, but not all, of the vectors in X may be stochastic ($X_j = \mathbb{X}_j + \nu_j$). In case X is a non-stochastic vector, it is called an exogenous variable. It is also pertinent to note that the structural coefficient matrix A has a special structure such that the elements in its principal diagonal are all minus unity (-1) or $a_{ij} = -1 \ \forall \ i = j$. Further, depending on the nature of the model, A may be diagonal (that is $A = -I$, a negatively signed identity matrix), lower triangular (where $a_{ij} = 0 \ \forall \ i < j$) or upper triangular (where $a_{ij} = 0 \ \forall \ i > j$) characterizing a recursive model, block-diagonal, or finally a regular one (which characterizes a true simultaneous model).

Empirically, we collect data on Y and the exogenous variables (that may make a full or partial X). Thus, empirical Y has two strains of error or $Y = \mathbb{Y} + \varepsilon + E$. It is assumed that $\nu_j$ (in the pre-determined variables comprising X) and $E_j$ are orthogonal (linearly independent).

The objective is to estimate A and B. First, we simplify the model by a transformation of its equations (called structural equations) into another type of equations (called reduced form equations) in which $Y - XP - \eta = 0$. This transformation is effected by post-multiplying our structural model YA + XB + E = 0 by the inverse of the coefficients matrix A. That is:

$YAA^{-1} + XBA^{-1} + EA^{-1} = 0A^{-1}$ or $Y - X\Pi - \eta = 0$, where, $\Pi = -BA^{-1}$ and $\eta = -EA^{-1}$.

The reduced form model $Y - X\Pi - \eta = 0$ may be rewritten as $Y = X\Pi + \eta$. Having assumed that the stochastic terms in X, if any, and E are orthogonal, it is obvious that vectors in $\eta$ and $\nu$ are orthogonal across $\eta$ and $\nu$. This result prompts us to estimate $\Pi$ by a suitable method such as the method of ordinary least squares(OLS). The OLS estimator of $\Pi$ is given by:

$$P = (X'X)^{-1}X'Y \quad \text{or} \quad P = \{(X'X)^{-1}X'\}Y.$$

In $P = \{(X'X)^{-1} X'\}Y$, factor $\{(X'X)^{-1} X'\}$ has a special interpretation. It is the generalized inverse (more exactly, the least squares g-inverse) of X. That is, $X^{-g} = [X'X]^{-1}X'$, such that $X^{-g}X = \{(X'X)^{-1} X'\}X = (X'X)^{-1} X'X = I$ and $XX^{-g} = X\{(X'X)^{-1} X'\} = I_d$, an idempotent matrix. Having obtained $P$, one may obtain the expected Y by the relationship $\hat{Y} = XP$.

However, the original objective was to estimate A and B, and instead, we have estimated $\Pi = -BA^{-1}$. The question is: can we obtain, through some algebraic manipulation, the estimated A and B (that is, $\hat{A}$ and $\hat{B}$), and if the answer is in an affirmative, then under what conditions can we obtain $\hat{A}$ and $\hat{B}$ ? This is the problem of identification.

It is obvious that if each column of A as well as B could be known, A and B in full can be known. Hence, we will try to answer the question posed above for a particular equation (say, $r^{th}$ one) in the model YA + XB + E = 0. Since $\Pi = -BA^{-1}$ , it implies $P = -\hat{B}\hat{A}^{-1}$ or $\hat{P}\hat{A} = -\hat{B}$. For the $r^{th}$ equation, only the respective $r^{th}$ columns of $\hat{A}$ and $\hat{B}$ would be used. Thus, for the $r^{th}$ equation we solve the system of equations given by $P\hat{a}_r = -\hat{b}_r$, where $\hat{a}_r$ and $\hat{b}_r$ are referring to the $r^{th}$ columns of the expected A and B matrices respectively.

Since A is an m x m matrix and B is a k x m matrix, $\Pi = -BA^{-1}$ is a k x m matrix. Therefore, the expression $P\hat{a}_r = -\hat{b}_r$ is a system of k (linear) equations involving m+k unknowns. It is obvious that we cannot (uniquely) determine m+k unknowns and thus the system of equations $P\hat{a}_r = -\hat{b}_r$ is indeterminate.

We may proceed further by augmenting the system with m number of additional (independent) equations in $\hat{a}_{sr} \in \hat{a}_r$ or $\hat{b}_{tr} \in \hat{b}_r$ (or both). The most straightforward way to do that is to set some $\mu_r$ unknowns ($\hat{a}_{sr} \in \hat{a}_r$ or $\hat{b}_{tr} \in \hat{b}_r$ or both) equal to zero. It amounts to zero restriction on some $\mu_r$ structural coefficients in the $r^{th}$ structural equation. It is obvious that $\mu_r \geq$ m, else the problem is indeterminate. In case $\mu_r$ = m, we have as many equations as the unknowns, and further assuming that no equation is linearly dependent on the others, the unknowns (remaining after the zero restriction) can uniquely be determined. In this case we say that the equation $P\hat{a}_r = -\hat{b}_r$ is *exactly identified*. However, if $\mu_r >$ m, we have the equations larger in number than the unknowns, and the system of equations is over-determined. Generally, such an over-determined system is also inconsistent. That is to say that the solutions (values of the unknowns obtained from such an over-determined system of equations) do not satisfy all the equations. In this case we say that the equation $P\hat{a}_r = -\hat{b}_r$ is *over-identified*.

To formalize what we have mentioned above, let us categorize the elements of $\hat{a}_r$ and $\hat{b}_r$ into two (disjoint) categories, namely, the unknown ones and the known ones. We will use the subscripts 1 and 2 (respectively) to identify them. Thus, $[\hat{a}_r]_1$ and $[\hat{b}_r]_1$ are the partitioned vectors (columns) of $[\hat{a}_r]$ and $[\hat{b}_r]$, whose elements are some (say, $m_1$ and $k_1$ respectively) unknown quantities. Similarly, $[\hat{a}_r]_2$ and $[\hat{b}_r]_2$ are the partitioned vectors (columns) of $[\hat{a}_r]$ and $[\hat{b}_r]$, whose elements are ($m_2 = m-m_1$ and $k_2 = k-k_1$ respectively) known quantities. Note that in order to avoid the under-identifiability of the $r^{th}$ structural equation it is necessary that $k_1 + m_1 = \mu_r \geq m$.

We have mentioned earlier that the elements in the principal diagonal of matrix A are all minus unity ($a_{ii} = -1 \ \forall \ i$). Presently, we are concerned with the $r^{th}$ column of the matrix A. Thus, the $r^{th}$ element of $[\hat{a}_r] = a_{rr} = -1$. In our scheme of categorized partition, therefore, the element $a_{rr}$ would belong to $[\hat{a}_r]_2$. Further, due to zero restriction on the coefficients all the rest elements of $[\hat{a}_r]_2$ are zero and all the elements of $[\hat{b}_r]_2$ are zero.

In the said scheme of categorized partition it would be helpful (**Mishra**, 1997) to use the permutation matrix operation on $[\hat{a}_r]$ and $[\hat{b}_r]$. Let G(m,m) be the permutation matrices obtained by permutating the columns of the identity matrix I(m,m) and let H(k,k) be the permutation matrix obtained by permutating the rows of the identity matrix I(k,k) such that:

$$G[\hat{a}_r] = \frac{[\hat{a}_r]_1}{[\hat{a}_r]_2} \quad \text{and} \quad H[\hat{b}_r] = \frac{[\hat{b}_r]_1}{[\hat{b}_r]_2}$$

Therefore, the system of equations $P\hat{a}_r = -\hat{b}_r$ is transformed (rearranged) as follows:

$$H\,P\,G^{-1}G[\hat{a}_r] = H[\hat{b}_r].$$

Due to pre-multiplication of $P$ by H, the rows of $P$ are permutated in correspondence with $H[\hat{b}_r]$ and due to post-multiplication of $P$ by $G^{-1}$, the columns of $P$ are permutated in accordance with $G[\hat{a}_r]$. Let us rename $H\,P\,G^{-1}$ as Q. It is to be noted that Q is numerically known since $P$, G and H are all known. Then,

$$Q\frac{[\hat{a}_r]_1}{[\hat{a}_r]_2} = \frac{[\hat{b}_r]_1}{[\hat{b}_r]_2}, \text{ or}$$

$$\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} \dfrac{[\hat{a}_r]_1}{[\hat{a}_r]_2} \end{bmatrix} = \begin{bmatrix} \dfrac{[\hat{b}_r]_1}{[\hat{b}_r]_2} \end{bmatrix}.$$

In this scheme, $Q_{11}$ is a $k_1$ x $m_1$ matrix, $Q_{12}$ is a $k_1$ x $m_2$ matrix, $Q_{21}$ is a $k_2$ x $m_1$ matrix and $Q_{22}$ is a $k_2$ x $m_2$ matrix. This gives us two equations:

$$[Q_{11}]\,[\hat{a}_r]_1 + [Q_{12}]\,[\hat{a}_r]_2 = [\hat{b}_r]_1$$

$$[Q_{21}]\,[\hat{a}_r]_1 + [Q_{22}]\,[\hat{a}_r]_2 = [\hat{b}_r]_2$$

Since $[\hat{a}_r]_2$ and $[\hat{b}_r]_2$ are known (more specifically, only one element of $[\hat{a}_r]_2$ is -1 and other elements are zero, and all the elements of $[\hat{b}_r]_2$ are zero), $[Q_{21}]^{-g}[[\hat{b}_r]_2 - [Q_{22}][\hat{a}_r]_2] = [\hat{a}_r]_1$. In particular, since $[\hat{b}_r]_2 = [0]$, we have $[\hat{a}_r]_1 = -[Q_{21}]^{-g}[Q_{22}][\hat{a}_r]_2$. Once $[\hat{a}_r]_1$ is obtained, one may subsequently obtain $[\hat{b}_r]_1$ in $[Q_{11}][\hat{a}_r]_1 + [Q_{12}][\hat{a}_r]_2 = [\hat{b}_r]_1$ by substitution.

In case $[Q_{21}]$ is a square matrix of full rank (the $r^{th}$ equation is exactly identifiable), $[Q_{21}]^{-g} = [Q_{21}]^{-1}$. Obtaining $[\hat{a}_r]_1 = -[Q_{21}]^{-1}[Q_{22}][\hat{a}_r]_2$ and subsequently $[\hat{b}_r]_1$ by substitution (applicable only if the $r^{th}$ structural equation is exactly identifiable) is called the method of *Indirect Least Squares*.

However, if $[Q_{21}]$ is not a square matrix or it is deficient in rank $[Q_{21}]^{-1}$ would not exist. The restriction of $\mu_r \geq m$ together with the assumption that $[Q_{21}]$ is of a rank $m_1$ guarantees that $[Q_{21}]^{-g}$ (or the least squares generalized inverse of $[Q_{21}]$) exists (**Theil,** 1971, pp. 268-273). That is to say that the least squares solution of $[\hat{a}_r]_1$ exists. In the worst case, when the rank of $[Q_{21}]$ is $< m_1$, only the proper Moore-Penrose inverse of [Q21] or $[Q_{21}]^{+}$ exists. In that case $[\hat{a}_r]_1$ cannot be known or estimated uniquely.

We have seen that in case $[Q_{21}]$ is a square matrix, $[Q_{21}]^{-1}$ exists (provided that $[Q_{21}]$ has a full rank of $m_1$). Since $[Q_{21}]$ is a $k_2$ x $m_1$ matrix, its being a square matrix implies that $k_2 = m_1$. Now $k_2$ means the number of elements in $[\hat{b}_r]_2$ all set to zero, which in turn implies the number of pre-determined variables appearing in the model YA + XB + E = 0, but absent from the $r^{th}$ equation. Similarly, $m_1$ means the number of endogenous variables with unknown structural coefficients that appear in the $r^{th}$ equation. We have seen that exactly one more endogenous variable ($y_r$) appears in the $r^{th}$ equation, but its coefficient is minus unity (-1) due to which fact $a_{rr} = -1$. Therefore, it is said that the *necessary condition for exact identification* of the $r^{th}$ equation is that the number of endogenous variables appearing in it is equal to the number of pre-determined variables absent from it plus one. The sufficient condition for exact identification is, of course, that $[Q_{21}]$ has a full rank of $m_1$.

In case of an over-identification where $k_2 \geq m_1$ the number of endogenous variables appearing in the particular equation (the $r^{th}$ one) must be less than the number of pre-determined variables absent from the model plus one. This is the necessary condition for over-identification. The sufficient condition is that $[Q_{21}]$ has a full rank of $m_1$.

Therefore, the $r^{th}$ equation would be under-identified if and only if either (or both) of the two conditions is (are) satisfied: (i) $k_2 < m_1$ (ii) rank of $[Q_{21}]$ is deficient or rank($[Q_{21}]$) $< m_1$. It is obvious that in the case where $k_2 < m_1$, rank($[Q_{21}]$) $\leq k_2 < m_1$. Therefore, $k_2 < m_1$ guarantees under-identification. However, the rank of $[Q_{21}]$ might be deficient (rank($[Q_{21}]$) $< m_1$) even if $k_2 \geq m_1$ provided that there are enough number of linear dependencies in the equation system $[Q_{21}][\hat{a}_r]_1 + Q_{22}][\hat{a}_r]_2 = [\hat{b}_r]_2$.

We have seen that in case of over-identification $k_2 > m_1$, due to which the system of equations described by $[Q_{21}]\,[\hat{a}_r]_1 + Q_{22}][\hat{a}_r]_2 = [\hat{b}_r]_2$ has the number of equations larger than the number of unknowns to be determined. Consequently, $[Q_{21}]^{-1}$ is not defined. It is natural to think of obtaining $[Q_{21}]^{-g}$ and $[\hat{a}_r]_1 = -[Q_{21}]^{-g}[Q_{22}]\,[\hat{a}_r]_2$. However, Henri Theil and R L Basmann appear not to have been attracted by this route to estimation of $[\hat{a}_r]_1$ and subsequently obtaining $[\hat{b}_r]_1$ in $[Q_{11}]\,[\hat{a}_r]_1 + [Q_{12}]\,[\hat{a}_r]_2 = [\hat{b}_r]_1$. Instead, they obtain expected Y (say, $\hat{Y}$) by the reduced form equations (that is $\hat{Y} = X\,P$). Then, in any particular (over-identified) equation, say the $r^{th}$ equation, each $y_s$ (with un-determined coefficients, $s \neq r$) is replaced by the corresponding $\hat{y}_s$ such that $\hat{Y}a_r + Xb_r = 0$. Since $y_r$ in the $r^{th}$ equation appears with a known coefficient ($a_{rr} = -1$), $y_r$ is not replaced by $\hat{y}_r$. Then estimation of $a_r$ and $b_r$ by OLS is permissible as the error in $y_r$ (dependent endogenous variable) is no longer correlated with the errors in the explanatory variables ($Y_s$ or $X_r$) and the Gauss-Markov conditions are satisfied. By OLS, therefore, the unknown coefficients in $a_r$ and $b_r$ appearing in $\hat{Y}a_r + Xb_r = 0$ are estimated. Thus, first OLS is used to obtain P and subsequently, OLS is used once again on $\hat{Y}a_r + Xb_r = 0$ to obtain the unknown coefficients in $a_r$ and $b_r$. On account of applying OLS at two stages, the method is called the *Two-Stage Least Squares* (2SLS). From the procedure and the conditions governing its application it is clear that 2SLS is an Instrumental variable approach to estimation of $Y_r a_r + X_r b_r = 0$, where each $y_s$ (with un-determined coefficients, $s \neq r$) is replaced by the instrumental variable $\hat{y}_s$. It follows from this the 2SLS estimator is (usually) biased but consistent.

It is natural to explore the possibility of obtaining $[\hat{a}_r]_1$ and $[\hat{b}_r]_1$ in an over-identified structural equation by using the least squares inverse of $[Q_{21}]$, that is $[Q_{21}]^{-g}$ as mentioned earlier. However, it took a long time to attract one's attention since **Basmann** (1957) and **Theil** (1961) developed 2SLS. **Khazzoom** (1976) investigated into generalization of ILS (evidently ignored by Theil and Basmann) for an over-identified equation. Khazzoom estimates reduced form equations of a multi-equation linear econometric model by OLS but (in the second stage) instead of estimating the (modified) structural equations by OLS (or the Instrumental variable method) as done in the 2SLS, he applies generalized inverse of the relevant submatrix of reduced form coefficients to obtain the structural coefficients. More explicitly, for the model YA+XB+E=0 (the reduced form equations being $Y = X\Pi + U$, $\Pi = -BA^{-1}$ and $P = \hat{\Pi}$), in the relationship $Pa_j = -b_j$ for any ($j^{th}$) structural equation, we have

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = -\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$ where $a_1$ and $b_1$ are unknown structural coefficients, $a_2 = (0\ 0\ \ldots\ 0\ -1)'$ and $b_2 = (0\ 0\ \ldots\ 0\ 0)'$. From this we obtain $\hat{a}_1 = -P_{21}^{-g}(b_2 + P_{22}a_2)$ and $\hat{b}_1 = -(P_{11}\hat{a}_1 + P_{12}a_2)$.

**VIII. Estimation of Multi-equation Models by LAD**: So far we have seen how the method of Least squares is applied to estimation of the structural coefficients in a multi-

equation (linear) model. Now we turn to the application of LAD to estimation of the same. The $L_1$ norm (LAD) estimation entered the domain of multi-equation model with the paper published by **Glahe** & **Hunt** (1970). Since then works on searching a suitable, fast and convenient numerical method (algorithm)) for $L_1$ norm estimator continued. Glahe & Hunt compared the estimated parameters with those estimated through $L_2$ estimation and use of Monte Carlo Methods for their performance appraisal.

In their paper, a distribution sampling study comprising of four major experiments has been described. All the experiments have been based upon the exactly specified, over-identified simultaneous equation model

$$Y_1 + A_{12}Y_2 + B_{11}Z_1 + B_{12}Z_2 + B_{10} = E_1$$
$$Y_1 + A_{22}Y_2 + B_{23}Z_3 + B_{24}Z_4 + B_{20} = E_2$$

where $Y_1$ and $Y_2$ are jointly determined endogenous variables; $Z_1$, $Z_2$, $Z_3$ and $Z_4$ are exogenous variables; $E_1$ and $E_2$ are the random error terms which are assumed to be normally and independently distributed with a zero mean and standard deviation of ten (except in the experiment involving heteroskedasticity).

A single structure for the basic model presented above was used throughout. For the exogenous variables, economic time series data covering the period 1960-1964 were chosen. The values chosen were quarterly values for farm income ($Z_1$), farm equipment price index ($Z_2$), personal income ($Z_3$) and adjusted money supply ($Z_4$). Except for the experiment involving multi-collinearity, the data were randomly shuffled to purge the inherent multi-collinearity present in most economic time series data.

The structural equations were transformed to the reduced form equation to generate data. A random normal deviate generator was used to generate errors for sample sizes ten and twenty. With these data the values for the endogenous variables were calculated. Keeping the vectors of exogenous variables constant for each set of data, fifty sets of data were generated for each sample size.

In each experiment six estimators were tested. These estimators were direct least squares (DLS), direct least absolute (DLA), two-stage least squares (TSLS), two-stage least absolute (TSLA), least squares no restrictions (LSNR), and least absolute no restrictions (LANR). (Direct application of least squares or the method of least absolute to the reduced form yielded LSNR and LANR estimators). The first two pairs were used to compute moments of the distribution of each parameter estimate, for sample sizes ten and twenty, based upon the fifty replications. All three pairs were used to compute conditional predictions of each of the jointly determined variables.

Each of the four major experiments conducted was divided into sub-categories where small sample sizes of ten and twenty were tested. The first experiment was conducted using the classical simultaneous equation model. Normally and independently distributed error terms with mean = zero and standard deviation = 10, uncorrelated exogenous variables and correct specification of the model were used. The second experiment considered a level of multi-collinearity among the explanatory variables. Heteroskedasticity was considered in the third experiment. The variance considered was a

monotonic function increasing over time given by $\sigma_u^2 = (\sigma_o + i)^2$ where $\sigma_o = 5$ and $i = 0,1,....,N-1$. In the fourth experiment misspecified model was investigated. The model was misspecified by including an additional exogenous variable and a parameter with a true value of zero in the estimation sequence. The endogenous variables were generated in the same manner. The computational method used in $L_1$ estimation was based on Usow $L_1$ Fit Algorithm, developed by **Usow**.

The study was concerned with two major objectives - the estimation of structural parameters and conditional prediction. Examining the means and standard deviations of the estimates of structural parameters some summary statistics were prepared. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were used for an evaluation of the performance of the estimators on the basis of smallest bias and smallest standard deviation. Rankings of the actual results by smallest RMSE and MAE were prepared and from those rankings summaries and summary statistics were calculated.

To test the consistency of the total rankings of the estimators, Kendall's coefficients of concordance, *W*, was used. The hypothesis that there was no difference between estimators (when paired) in the number of times one estimator produced smaller MAE's than another one in each experiment was tested using the Cochran Q test. The hypothesis was accepted at 0.05 level. The Wilcoxin matched-pairs signed-ranks test was used to compare the $L_1$ and $L_2$ estimators to determine whether or not one was significantly different from another one. To check for the normality of sample distributions of the *studentized* ratios of structural-coefficient estimates, Kolmogoroff-Smirnov test as explained by Birnbaum was used. The ratio used has been given by $T_{\hat{\theta}^K} = \left(\hat{\theta}^K - \theta^*\right) \Big/ \hat{\sigma}_{\hat{\theta}^K}$ , where $\theta^*$ is the hypothesized value of $\theta$.

The result of the four experiments showed that the two direct methods were best overall estimators for making conditional predictions, whether MAE or RMSE criterion were used and was true for both sample sizes. When errors were normally distributed and no substantial multicollinearity was present, none of the reduced form estimators was "poor". But in the presence of multicollinearity DLS fell off sharply in predictive ability. When other problems existed, LSNR or LANR proved to be more reliable since they were the methods with least variability of the six studied.

It was also observed that LANR performed as well as LSNR and both outperformed the solved reduced-form methods. The structural estimators, DLA and TSLA, did not outperform DLS and TSLS. They did succeed in doing as well as the least squares estimators in many respects. The authors, therefore, concluded that $L_1$ norm estimator should prove equal or superior to $L_2$ norm estimators for model using a structure similar to the one used in the study. They, however, held that with an increase in sample size the superiority of the $L_1$ norm estimator loses its edge over $L_2$ norm estimator.

**Amemiya** (1980) developed the two-stage least absolute deviation estimator, which is rather analogous to two-stage least squares by **Theil** (1961). **Amemiya** (1982) further extended the method to provide it a mathematical and statistical basis in the

direction of consistency and related statistical properties. In this paper (of 1982) he defined a class of estimators called the two-stage least absolute deviation estimators (2SLAD) and derived their asymptotic properties. The problem of finding the optimal member of the class was also considered.

**Amemiya** (1982) also pointed out that in structural equations and reduced form equations as given below:

$$YA+XB+E= ZC+E \text{ and}$$

$$Y=X\pi +V; \text{ where } Z = (Y,X) \text{ and } C = \begin{pmatrix} A \\ B \end{pmatrix}$$

how one defines the LAD (least absolute deviation) estimator analogue of 2SLS (two-stage least squares estimator) in the estimation of C ? Amemiya points out that the authors of all previous studies (before Amemiya wrote that article) on the subject defined LAD as the value of C that minimized

$$S_a = \sum \left| Y_1 - P_1' \, ZC \right| \text{ , where } P = X \left( X'X \right)^{-1} X'$$

It was rather natural to define LAD that way since then. Theil interpreted 2SLS so as to minimize $S_L = \sum \left( Y_1 - P_1' \, ZC \right)^2$ . However, if one wanted to use an interpretation of 2SLS as the instrumental variable estimator minimizing $S_{1L} = \sum \left( P' \, Y - P' \, ZC \right)^2$ , one would define 2SLAD analogously to minimize $S_{1A} = \sum \left| P'Y - P' \, ZC \right|$ . Combining the above two ideas, 2SLAD can be defined as a class of estimators obtained by minimizing

$$S_{qA} = \sum \left| qY + (1-q) P' \, Y - P' \, ZC \right|$$

where q is the parameter to be determined by the researcher. The minimization of

$$S_{qL} = \sum \left\{ qY + (1-q) P' \, Y - P' \, ZC \right\}^2$$

yields 2SLS for any value of q whereas minimization of its absolute analogue $(S_{qA})$ depends crucially on the value of *q*. If *q=0*, it yields the estimator which is asymptotically equivalent to 2SLS. Thus, in the asymptotic sense the class of 2SLAD estimator contains 2SLS as a special case. This finding by Amemiya has a very powerful generalizing effect on the estimators.

In the article, Amemiya proved the strong consistency and the asymptotic normality of the LAD estimator in the standard regression model. Though the asymptotic normality was proved by Bassett and Koenker prior to Amemiya, the method used by Amemiya is simple to understand and more easily generalizable to other models such as simultaneous equation models or non-linear regression models.

Given the standard regression model $Y = Xa + E$, where $X$ is a *nxk* matrix of bounded constants such that $\lim_{\to\infty}(n^{-1}X'X)$ is a finite positive-definite matrix and $E$ is a n-vector of i.i.d random variables, the LAD estimator has been defined to be a value of $\hat{a}$ that minimizes $S = \sum_{i=1}^{n} \left| Y_i - X_i' \, \hat{a} \right| - \sum_{i=1}^{n} \left| E_i \right|$, where $X_t'$ is the i[th] row of X. The second term of the right-hand side of the equation does not affect the minimization since it is

independent of $\hat{a}$. It was added to facilitate proof of consistency without assuming the existence of a finite first moment. The strong consistency of LAD was proved by showing that $n^{-1}S$ converges almost surely uniformly in $\hat{a}$ to a function which attains the minimum at $a$, the true value. Strong consistency of 2SLAD for any value of $q>0$ followed from the strong consistency of LAD. Asymptotic normality of 2SLAD was proved only for the case where *E* and *V* are normally distributed.

In the 2SLAD estimation studied, it was assumed that the minimization of the sum of absolute deviation is applied only to a specific equation to be estimated and not to all the reduced form equations. In other words, LAD was applied only in the second stage of regression and not in the first. The author, however, opined that if *V* as well as *E* follows a non-normal distribution, it would be better to apply LAD to the reduced form equation as well as to the structural equation to be estimated.

Applying LAD to each of the reduced form equations, $Y = X\pi + V$, $\hat{\pi}$ was obtained and then minimizing $\sum_{i=1}^{n} \left| Y_i - X_i' \hat{\pi} A - X_{1i}' \hat{a} \right|$, the double two-stage least absolute deviation estimator (D2SLAD) was developed. It was shown that even under the fully non-normal case D2SLAD is far inferior to 2SLAD for *q* between 0.2 and 0.5 for realistic values of the parameters. This result showed that in applying the LAD estimation to 2SLS, it is much more important to use LAD in the second stage than in the first stage.

Since $\hat{\pi}$ is a strongly consistent estimator of $\pi$, the strong consistency of D2SLAD followed easily from the strong consistency of LAD. Asymptotic normality of D2SLAD, however, has not been proved.

Asymptotic variance of 2SLAD and D2SLAD in a partially non-normal case (where E follows a mixture of normal distributions and *V* is normal) and fully non-normal case (where both *E* and *V* follow a mixture of non-normal distributions) were obtained. It was observed that when all the error terms follow a mixture of normal distributions, 2SLAD with a small value of *q* somewhere between 0 and 0.5 is recommended and it does not pay to use the more complicated D2SLAD.

Amemiya suggested Monte Carlo experiments to be carried out in order to study the properties of 2SLAD estimator (by minimization of $S_{qA}$) which may be compared with the properties of the estimator obtained by minimization of $S_{qL}$ and $S_{qA}$ for $q=0$.

**IX. Comparative Studies**: Earlier, a reference has been made to the work of Glahe & Hunt who compared the results of LS-based with LAD-based estimators of structural equations. **Fair** (1994) estimated the US model by 2SLS, 2SLAD, 3SLS (Three Stage Least Squares) and FIML (Full Information Max Likelihood) methods (see, **Theil**, 1971). Median unbiased (MU) estimates were also obtained for eighteen lagged dependent variable coefficients. The 2SLS asymptotic distribution was compared to the exact distribution and was found to be close. A comparative study of four sets of estimates, that is, 2SLS, 2SLAD, 3SLS and FIML was made. The results obtained showed that the

estimates are fairly close to each other with the FIML being the farthest apart. The 3SLS estimator was found to be more efficient than the 2SLS estimator. The 2SLS standard errors were on an average 28 percent larger than the 3SLS standard errors. And the 3SLS standard errors were on average smaller (19 percent) than the FIML standard errors. To compare the different sets of coefficient estimates, the sensitivity of the predictive accuracy of the model to the different sets was also examined. The RMSEs were found to be very similar across all the five sets of estimates. No one set of estimates dominated the other and in general the differences were found to be quite small. The author also compared the US model to the VAR5/2, VAR4 and AC models. The US model was found to do well in the tests relative to the VAR and AC models.

In view of the possibilities of replacing OLS with LAD estimator at either or both stages (parallel to 2SLS) of estimation of the structural equations of a multi-equation linear model, **Mishra & Dasgupta** (2003) conducted Monte Carlo experiments to compare 2SLS (alias LS-LS) with LS-LAD, LAD-LS and LAD-LAD estimates of structural coefficients while the disturbances in the structural equations were normal, $Beta_1$, $Beta_2$, Gamma and Cauchy distributed with and without the presence of outliers.

We have already described the work of **Khazzoom** (1976) who generalized Indirect Least Squares estimator for (exactly or over-) identified equations. It appears that Khazzoom's work is relatively less acknowledged. However, it deserves comparison with other methods of estimation. One may also conjecture that if LAD performs better than OLS in estimating the matrix of reduced form coefficients, application of generalized inverse on such matrix (of reduced form coefficients) would be better than the GILS suggested by Khazzoom. A more generalized name – ***Generalized Indirect Least Norm*** (**GILN**) - may be given to the family of such methods for the minimand norm may be Euclidean (as suggested by Khazzoom, alias $GILN_2$) or absolute, giving $GILN_1$. Mishra & Dasgupta also compared $GILN_1$ with $GILN_2$, 2SLS, LS-LAD, LAD-LS and LAD-LAD estimates of structural coefficients. The results showed that LAD-LAD estimator performs better then 2SLS if errors are non-normal or outliers are present.

## BIBLIOGRAPHY

**Abdelmalek, NN** (1971). "Linear $L_1$ Approximation for a Discrete Point Set and $L_1$ Solutions of Overdetermined Linear Equations", *Journal of Assoc. Comput. Mach.*,18 (41-47).

**Abdelmalek, NN** (1974). "On the Discrete Linear $L_1$ Approximation and $L_1$ Solution of Over- determined Linear Equations", *Journal of Approximation Theory,*11(35-53).

**Amemiya, T** (1980). "The Two Stage Least Absolute Deviations Estimators", Technical Report No. 297, Institute for Mathematical Studies in the Social Sciences, Stanford University, December.

**Amemiya, T** (1982). "Two-Stage Least Absolute Deviations Estimators", *Econometrica*, 50: 1-3 (689-711)

**Anscombe, FJ** (1967). "Topics in the Investigation of Linear Relations Fitted by the Method of Least Squares", *Journal of the Royal Statistical Society,* Series B, 29 (1-52).

**Ashar, VC & TD Wallace** (1963). "A Sampling Study of Minimum Absolute Deviation Estimators", *Operations Research*, 11 (747-758).

**Barrodale, I**. (1968). "$L_1$ Approximation and the Analysis of Data", *Appl. Statist*. 17 (51-57).

**Barrodale, I & FDK Roberts** (1973). "An Improved Algorithm for Discrete $L_1$ Approximation ", SIAM. *Journal of Numerical Analysis*, 10 (839-848).

**Barrodale, I & A Young** (1966). "Algorithms for Best $L_1$ and $L_\infty$ Linear Approximations on a Discrete Set", Numer. Math., 8 (295-306).

**Bartels, R** (1977). "On the Use of Limit Theorem Arguments in Economic Statistics", *American Statistician*, 31 (85-87).

**Basmann, RL** (1957). "A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation", *Econometrica*, 25 (77-84).

**Bassett, G Jr & R. Koenker** (1978). "Asymptotic Theory of Least Absolute Error Regression", *Journal of American Statistical Association*, 73 (618-622).

**Blattberg, R & T Sargent** (1971). "Regression with Non-Gaussian Stable Disturbances: Some Sampling Results", *Econometrica*, 39 (501-510).

**Breidt, FJ, RA Davis** & **A Trindade** (2000). "Least Absolute Deviation Estimation for All-Pass Time Series Models", *Annals of Statistics*., (in press). www.stat.colostate.edu/~j**breidt**/technical_reports.htm

**Charnes, A, WW Cooper & RO Ferguson** (1955). "Optimal Estimation of Executive Compensation by Linear Programming" *Management Science*, 1 (138-151).

**Chen, XR, ZD Bai, LC Zhao & YH Wu** (1992). "Consistency of Minimum $L_1$- norm Estimates in Linear Models" in *Development of Statistics: Recent Contributions from China* (249-260), Pitman Research Notes in Mathematical Series 258, Longman Scientific and Technical, Longdon.

**Chen, X** (1996). "On a Problem of Strong Consistency of Least Absolute Deviation Estimates". *Statistica Sinica*, 6 (481-489).

**Dantzig, GB** (1955). "Upper Bounds, Secondary Constraints, and Block Triangularity in Linear Programming". *Econometrica*, 23 (166-173).

**Dantzig, GB** (1963). *Linear Programming and Extensions*. Princeton University Press, Princeton, N. J.

**Dasgupta, M** (2004). *Least Absolute Deviation Estimation of Multi-Equation Linear Econometric Models: A Study Based on Monte Carlo Experiments*. Doctoral Dissertation, Department of Economics, NEHU, Shillong ( unpublished ).

**Edgeworth, FY** (1887). "On Discordant Observations" , *Phil. Magazine,* 24 ( 364-375 ).

**Edgeworth, FY** (1888). "On a New Method of Reducing Observations Relating to Several Quantities", *Phil. Magazine*, 5, 25 (185-191).

**Edgeworth, FY** (1923). "On the Use of Medians for Reducing Observations Relating to Several Quantities" , *Phil. Magazine* 6th Ser. (1074-1088).

**Fair, RC** ( 1974). "On the Robust Estimation of Econometric Models", *Annals of Economic and Social Measurement*, 3 ( 667-677).

**Fair, RC** (1994). *Estimating and Testing the US Model*, Yale Univ. USA.

**Fama, EF** (1965). "The Behaviour of Stock Market Prices ", *Journal of Business*, 38 (34-105).

**Fisher, WD** (1961). "A Note on Curve Fitting with Minimum Deviation by Linear Programming", *Journal of American Statistical Association*, 56 ( 359-361 ).

**Furno, M** (2000). "LAD Estimation with Random Coefficient Autocorrelated Errors". www.eco.unicas.it/docente/**furno**/Text-rca.pdf

**Girshick, MA & T, Haavelmo** (1947). "Statistical Analysis of the Demand for Food: Examples of Simultaneous Estimation of Structural Equations", *Econometrica*, 15: 2 (79-110)

**Glahe, FR & JG Hunt** (1970). "The Small Sample Properties of Simultaneous Equation Least Absolute Estimators vis-a-vis Least Squares Estimators", *Econometrica,* 5: 38 (742-753).

**Haavelmo, T** (1943). "The Statistical Implications of a System of Simultaneous Equations", *Econometrica*, 11 (1-12).

**Haavelmo, T** (1944). "The Probability Approach in Econometrics", *Econometrica*, 12, Supplement.

**Haavelmo, T** (1947). "Methods of Measuring the Marginal Propensity to Consume". *Journal of American Statistical Association*, 42 (105-22).

**Hill, RW & PW Holland** (1977). "Two Robust Alternatives to Least Squares Regression", *Journal of American Statistical Association*, 72 (828-833).

**Hitomi, K** & **M Kagihara** (2001). "Calculation Method for Non-Linear Dynamic Least Absolute Deviation Estimator". *J. of Japan Statist. Society*, 31-1 (39-51).

**Hood, WC & TC, Koopmans** (Eds) (1953). Studies in Econometric Method, Cowles Foundation Monograph 14, John Wiley, New York.

**Huber, PJ** (1964). "Robust Estimation of a Location Parameter", Annals of Mathematical Statistics, 35 (73-101).

**Huber, PJ** (1972). "Robust Statistics: A Review", *Annals of Mathematical Statistics*, 43 (1041-1067).

**Huber, PJ** (1973). "Robust Regression: Asymptotics, Conjectures, and Monte Carlo", *The Annals of Statistics*, 1 (799-821).

**Khazzoom, JD** (1976). "An Indirect Least Squares Estimator for Over-identified Equations", *Econometrica*, 44, 4 (741-750).

**Kim, T H** & **C Muller** (2000). "Two-Stage Quantile Regression"_, *Discussion Paper in Economics* #00/1, Nottingham University.

**Mandlebroth, BB** (1963). "The Variation of Certain Speculative Prices", *Journal of Business*, 36 (394-419).

**Mandlebroth, BB** (1967). "The Variation of Some Other Speculative Prices", *Journal of Business*, 40 (393-413).

**Meyer, JR & RR Glauber** (1964). *Investment Decisions: Economic Forecasting and Public Policy*, Harvard Business School Press, Cambridge, Massachusetts.

**Mishra, SK** (1997). "Generalisation of Indirect Least Squares to Estimation of Over-identified Equations of a Multi-Equation Linear Econometric Model", *Assam Statistical Review*, 9:1, (57-66).

**Mishra, SK & M Dasgupta** (2003). "Least Absolute Deviation Estimation of Multi-Equation Linear Econometric Models: A Study Based on Monte Carlo Experiments" http://ssrn.com/abstract=454880

**Nyquist, H & A Westlund** (1977). *$L_1$ Versus $L_2$ Norm Estimation in Interdependent Systems when Residual Distributions are Stable*, Mimeo, Dept. of Statistics, Univ. of Umea.

**Oveson, RM** (1968). *Regression Parameter Estimation by Minimizing the Sum of Absolute Errors*, Doctoral Dissertation, Harvard University, Cambridge, Mass (unpub).

**Phillips PCB** (1991). "A Shortcut to LAD Estimator Asymptotics". *Econometric Theory*, Vol. 7-4 (450-463).

**Pollard, D** (1991). **"**Asymptotics for Least Absolute Deviation Regression Estimators". *Econometric Theory*, . 7-2 (186-199).

**Portnoy S** & **R Koenker** (1997). **"**The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-error versus Absolute-error Estimators". *Statistical Science*, 12 (279-300).

**Powell, JL** (1984). "Least Absolute Deviations Estimation for the Censored Regression Model". *J. of Econometrics*, 25 (303-325).

**Rabinowitz, P** (1968). "Applications of Linear Programming to Numerical Analysis", *SIAM Review, 10 (121-159).*

**Rhodes, EC** (1930). "Reducing Observations by the Method of Minimum Deviations", *Phil. Magazine* 7th Ser. ( 974-992 ).

**Rice, JR & JS White** (1964). "Norms for Smoothing and Estimation", *SIAM Review*, 6 (243-256).

**Robers, PD & A Ben-Israel** (1969). "An Interval Programming Algorithm for Discrete Linear $L_1$ Approximation Problems", *Journal of Approximation Theory*, 2 (323-336).

**Robers, PD & SS Robers** (NA), "Discrete Linear $L_1$ Approximation by Interval Linear Programming", *Manuscript*. Ref. by Barrodale, I & FDK Roberts (1973)

**Sakata, S** (2001). "Instrumental Variable Estimation Based on Mean Absolute Deviation". Papers of Department of Economics, University of Michigan, USA citeseer.ist.psu.edu/**sakata**01**instrumental**.html (Submitted to Journal of Econometrics, www.econ.ubc.ca/s**sakata**/public_html/cv/cv.html), www.econ.lsa.umich.edu/wpweb/ l1iv.pdf, 67 pages.

**Schlossmacher, EJ** (1973). "An Iterative Technique for Absolute Deviations Curve Fitting", *Journal of the American Statistical Association*, 68 (857-859).

**Singleton, RR** (1940). "A Method for Minimizing the Sum of Absolute Errors", *Am. Math. Statist*. 11 ( 301-310 ).

**Smith, VK & TW Hall** (1972). "A Comparison of Maximum Likelihood Versus BLUE Estimators", *The Review of Economics and Statistics*, 54 (186-190).

**Spyropoulous, K, E Kiountouzis & A Young** (1973). "Discrete Approximation in the $L_1$ Norm", *Computer Journal*, 16 (180-186).

**Theil, H** (1953). Repeated Least Squares Applied to Complete Equation Systems, Central Planning Bureau (Mimeo), The Hague.

**Theil, H** (1957). "Specification Errors and the Estimation of Economic Relationships", R*eview of International Statistical Institute*, 25 (41-51).

**Theil, H** (1961). *Economic Forecasts and Policy*, 2nd ed., North Holland Publishing Co., Amsterdam.

**Theil, H** (1971). *Principles of Econometrics*, John Wiley and Sons, Inc. New York.

**Taylor, LD** (1974). "Estimation by Minimizing the Sum of Absolute Errors" in P. Zerembka (ed.), *Frontiers of Econometrics*, Academic Press, New York.

**Usow, KH** (1967). "On $L_1$ Approximation for Discrete Functions and Discretization Effects", *SIAM Journal of Numerical Analysis*, 4 (233-244).

**Wagner, HM** (1958). "A Monte Carlo Study of Estimates of Simultaneous Linear Structural Equations", *Econometrica*, 26 (117-133).

**Wagner, HM** (1959). "Linear Programming Techniques for Regression Analysis", *Journal of American Statistical Association*, 56 (206-212).

**Weiss, AA** (1991). **"**Estimating Non-linear Dynamic Models using Least Absolute Error Estimation". *Econometric Theory*, 7 (46-68).