# LEAST ABSOLUTE DEVIATION ESTIMATION OF MULTI-EQUATION LINEAR ECONOMETRIC MODELS

## *A STUDY BASED ON MONTE CARLO EXPERIMENTS*

**MADHUCHHANDA DASGUPTA**

*DEPARTMENT OF ECONOMICS*

SUBMITTED
IN PARTIAL FULFILMENT OF
THE REQUIREMENT OF THE DEGREE OF
**DOCTOR OF PHILOSOPHY IN ECONOMICS**

**NORTH EASTERN HILL UNIVERSITY**
SHILLONG

# NORTH-EASTERN HILL UNIVERSITY
# DEPARTMENT OF ECONOMICS

**May 2004**

I, Ms Madhuchhanda DasGupta, hereby declare that the subject matter of the thesis is the record of the work done by me, that the contents of this thesis did not form basis of the award of any previous degree to me or to the best of my knowledge to anybody else, and that the thesis has not been submitted by me for any research degree in any other University/Institute.

This is being submitted to the North-Eastern Hill University for the degree of Doctor of Philosophy in Economics.

M. DasGupta

**(Madhuchhanda DasGupta)**
Candidate

**(N. Srivastav)**
Professor & Head

*Head*
*Dept. of Economics*
*North Eastern Hill University*
*Shillong*

21.5.04.

**(SK. Mishra)**
Professor of Economics
Supervisor

*Professor*
*Dept. of Economics*
*North Eastern Hill University*
*Shillong*

## *Acknowledgements*

It is my pleasure to acknowledge my heartfelt gratitude and indebtedness to my esteemed teacher and supervisor Prof. S. K. Mishra, Department of Economics, NEHU for his invaluable guidance, wise counsel and ungrudging help in every stage of my research work. I have no hesitation to record here that Prof. Mishra has not only aroused my interest in Econometrics but also motivated me to work on the present topic. My research work, indeed, would not have seen the light of the day, had I not received from Prof. Mishra constant advice, encouragement and also logistical support.

I take this opportunity to express my gratefulness to my revered teachers Prof. N. Srivastav, Head of the Dept. of Economics, NEHU, Prof. E.D. Thomas, Dean, School of Economics, Management & Information Sciences, NEHU, Dr. A.C. Dubey and Dr. B. Mishra for their encouragement and advice.

Thanks are also due to my friends Ms. Chandana Bhattacharjee, Ms. Sharmila Das Talukdar and my colleague Mrs. Mayashree B. Das for their help and moral support.

I owe a great debt of gratitude to my father who has been my pillar of strength and constant source of inspiration and my mother for her silent sacrifice to give me comfort during the course of my work.

Finally, my longstanding thanks are also due to my brother, Joy, for sustained encouragement in moments of despair. I am also grateful to my cousin Duke who has helped me immensely in various ways for the successful completion of the present work.

M. DasGupta

Shillong, (Madhuchhanda DasGupta)

The 21st May, 2004.

# CONTENTS

# LIST OF TABLES

| Table | Tables in Chapter 6 - Performance of LAD in Estimation of Multi- Equation Linear Models | Page |
|---|---|---|

# CHAPTER 1

# INTRODUCTION

**I. The Background**: Every *real empirical or experiential science* (against the ideal science such as mathematics or logic which derives implications from a closed system of axioms) is an organized description of empirical inter-relations among the variables of the system of its concern. For example, physics investigates into the relationship between electrical conductivity of a metallic wire and its length and thickness, and other such relations; chemistry purports to establish the relationship between the rate of various types of chemical reactions and the temperature and pressure that the reactants are subject to, and so on; botany studies the relationship between the growth of the biomass of a species of plants and the nutritional contents of the soil and the physico-chemical properties of the environment that support plant life, and such issues. Similarly, economics is concerned with the relationship between consumption expenditure of households on a specific commodity, income of the household and its size and the price of the commodity, and the likes. On constructive thinking and empirical evidence each branch of *real science* describes such relationships of its concern and puts them in an organic order to make an ever-growing edifice of its literature.

*Real sciences* originate from and feed on the empirical experiences accumulated either by observation or by experimentation. Growth of a *real science* is squarely based on constructive thinking that suggests the scientists to gather information on relevant variables and the suitable methods to process the information so accumulated to establish the relationship among them. Therefore, a *real science* progresses with a hypothesis in

which the values of a dependent variable is associated with the values of a single or several explanatory variables and it is held that variations in (the value of) the dependent variable is associated with the variations in (the values of) the explanatory variables in a particular manner. Here 'variation' is very crucial. It is the cornerstone of all empirical investigations. All empirical knowledge emanate from the analysis of variation. Therefore, the scientist collects information on the dependent variable (say, y) and the explanatory variables (say, $x_1$, $x_2$, $x_3$) taking on several values. Our set of observations may thus be described as:

$$\begin{bmatrix} y_1, x_{11}, x_{12}, x_{13} \\ y_2, x_{21}, x_{22}, x_{23} \\ y_3, x_{31}, x_{32}, x_{33} \\ \ldots, \ldots, \ldots, \ldots \\ y_n, x_{n1}, x_{n2}, x_{n3} \end{bmatrix} \quad \ldots (1.1).$$

The scientist holds or conjectures that $y = a_0 + a_1x_1 + a_2x_2 + a_3x_3$ (where $a_j$ for j=0, 1, 2 and 3 are non-zero numerical constants) is the most possible relationship between y and x's. However, the correct relationship could be $\upsilon = f(\chi_1, \chi_2, \chi_4, \chi_5, \chi_6, \omega_1, \omega_2, \zeta_1, \zeta_2)$, where f(.) is a nonlinear function. In the scientist's scheme, y, $x_1$ and $x_2$ are the most appropriate representatives that measure $\upsilon$, $\chi_1$ and $\chi_2$ respectively. One must note that all hypotheses in a *real science* relate concepts and only few concepts can be perfectly measured by a number (denial of strict *arithmomorphism* as argued by **N. Georgescu-Roegen**, 1971). Therefore, empirically obtained numerical measurements (data) on any conceptual entity is only a representative of that entity. Now, the scientist has mis-specified the correct relationship by deviating from the fact in many possible ways; first, he has not been able to measure the concepts perfectly accurately – he is working only

with the best possible numerical measures of the concepts to be correlated; secondly he has not been able to incorporate all relevant explanatory variables that explain y – he has failed to include representative variables that measure $\chi_4$, $\chi_5$, $\chi_6$, $\omega_1$, $\omega_2$, $\zeta_1$ and $\zeta_2$. Thirdly, he has included $x_3$ among the explanatory variables while it is none of them in fact; and fourthly (may not be lastly), he has thought of y = f(.) a linear function of x while in fact it is not. It is also not unlikely that he has exchanged a dependent variable for an explanatory variable or that his dependent variable and one of the explanatory variables are co-effects of the rest of the explanatory variables. It is also plausible that explaining y = f(x) is incomplete: y= f(x) and z = $\psi$(x,q) jointly would be more appropriate. And all these do not exhaust the list of mis-specifications.

The scientist should therefore compensate for all such possible errors by adding an error (disturbance or residual) term to his hypothesis and express his model as:

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + e \qquad \ldots \quad (1.2)$$

where the last term captures the effects of all mis-specifications committed by him.

Since the error is due to the deviation of the explicitly expressed model and the true relationship and there is no one-to-one correspondence between the scientist's committed specification and that omitted by him, error vector varies from sample to sample even if the values of the variables specified in the model are fixed. This also amounts to say that y varies from sample to sample for fixed values of x's. More explicitly, in the subscripted matrices $[\,]_1$, $[\,]_2$, $[\,]_3$ and $[\,]_4$ detailed out below,

$$\begin{bmatrix} y_{11}, x_{11}, x_{12}, x_{13} \\ y_{12}, x_{21}, x_{22}, x_{23} \\ y_{13}, x_{31}, x_{32}, x_{33} \\ \ldots, \ldots, \ldots, \ldots \\ y_{1n}, x_{n1}, x_{n2}, x_{n3} \end{bmatrix}_1 \quad \begin{bmatrix} y_{21}, x_{11}, x_{12}, x_{13} \\ y_{22}, x_{21}, x_{22}, x_{23} \\ y_{23}, x_{31}, x_{32}, x_{33} \\ \ldots, \ldots, \ldots, \ldots \\ y_{2n}, x_{n1}, x_{n2}, x_{n3} \end{bmatrix}_2 \quad \begin{bmatrix} y_{31}, x_{11}, x_{12}, x_{13} \\ y_{32}, x_{21}, x_{22}, x_{23} \\ y_{33}, x_{31}, x_{32}, x_{33} \\ \ldots, \ldots, \ldots, \ldots \\ y_{3n}, x_{n1}, x_{n2}, x_{n3} \end{bmatrix}_3 \quad \begin{bmatrix} y_{41}, x_{11}, x_{12}, x_{13} \\ y_{42}, x_{21}, x_{22}, x_{23} \\ y_{43}, x_{31}, x_{32}, x_{33} \\ \ldots, \ldots, \ldots, \ldots \\ y_{4n}, x_{n1}, x_{n2}, x_{n3} \end{bmatrix}_4 \quad \ldots \quad (1.3)$$

x's are fixed and yet the y's are different across the matrices. That is to say that the matrices $[\ ]_1$, $[\ ]_2$, $[\ ]_3$ and $[\ ]_4$ differ from one another in that their y columns are not identical (though their x columns are identical). This is so because for the four samples even though x's are identical, hidden ($\chi_4$, $\chi_5$, $\chi_6$, $\omega_1$, $\omega_2$, $\zeta_1$, $\zeta_2$) are different giving rise to different error vectors (e's). This gives rise to stochasticity in the error, e. The error is a random variable following a particular probability distribution. The scientist may have some speculation, conjecture or hypothesis regarding the probability distribution of the error term, e.

**The method of Least Squares**: The scientist aims at obtaining the best values of **a** = $[a_0, a_1, a_2, a_3]'$. Since no value of **a** can satisfy all equations, he has to define the meaning of the qualification '*best*'. It is natural to define '*best*' **a** – call it $\hat{a}$ – such that it yields an error vector, e with the minimum norm. If the scientist's choice is to minimize the Euclidean norm of the error vector, it is tantamount to minimize the squared Euclidean norm of the error vector. Therefore, he chooses to minimize

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (a_0 + a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3})\}^2 .$$

To minimize S he differentiates S partially with respect to $a_i$ (i=0, 1, 2, 3) and sets those partial derivatives equal to zero. This gives the so-called 'normal equations'.

This system of normal equations in this example will have four unknowns in four equations. Assuming linear independence of these equations, the system may be solved for **a**. This method of obtaining '*best*' **a** ($= \hat{a}$) is called the *Principle of Least Squares.*

Under the Gauss-Markov assumptions $\hat{a} = (X'X)^{-1} X'y$ is the best linear unbiased estimator of **a** in y = X**a** + e. These assumptions are: (i) $E(e_i) = 0 \Rightarrow E(y_i) = \sum_{j=0}^{m} a_j x_{ij}; \forall i, i = 1,2,...,n.$ In general, E(e) = 0 $\Rightarrow$ E($y_i$) = X**a**. Here E(.) means the expectation of the random variable (.). (ii) $E(ee') = \sigma^2 I$ or geometrically, the spherical disturbances. (iii) The matrix of sample values of the explanatory variables, X, is a fixed non-stochastic matrix of full rank.

However, reality often breaks away with the Gauss-Markov assumptions. Very often disturbances are non-spherical. Occasionally, in single equation models, the assumption regarding the non-stochasticity of the sample values of the explanatory variables may not be satisfied. That makes the (ordinary) Least Squares estimator, $\hat{a}$, an inconsistent estimator. The violation of the first Gauss-Markov assumption introduces bias into the Least Squares estimator. The estimator continues to be biased even if the sample size is increased indefinitely.

The real world data often consist of disturbances that are non-normally distributed, some of which permit the variate to take on non-negative values only. It has been known since Pareto, that income distribution bears testimony to distribution of the error term with infinite variance. Works of **Meyer & Glauber** (1964), **Fama** (1965) and **Mandlebroth** (1967), among others, confirm that economic data series like prices in commodity and financial markets present a class of distribution with infinite variance. An infinite variance

means "thick tail" which implies that large values or *'outliers'* are present. Sporadic errors in X also are frequently present.

We recall that in defining the 'best' solution the scientist's choice was to minimize the Euclidean norm of the error vector, tantamount to minimization of the squared Euclidean norm of the error vector. That choice yielded the Least Squares estimator. However, this method places a relatively heavy weight on outliers and, therefore, in the presence of outliers, the method can lead to extremely sensitive estimates. This implies that in repeated sampling from a distribution where outliers are prevalent, LS estimates will vary considerably and will fail to establish reliability, becoming extremely sample dependent. Moreover, it is also not possible to obtain a meaningful variance estimate in the cases where the variance does not exist and the LS method in such cases will cease to possess its minimum variance properties. Thus, it becomes necessary to look for an alternative, a more 'robust' estimator, which gives relatively less weight to outliers and hence is not affected by their presence or the distribution that the error term characterizes.

Therefore, one parts with the convention of minimizing the Euclidean norm of the disturbance vector and, instead, defines the 'best' solution, $\tilde{a}$, of **a** in y = X**a** + e that minimizes the *absolute norm* of e (or yields minimal $\sum_{i=1}^{n}|e_i| = \sum_{i=1}^{n}\left|y_i - \sum_{j=0}^{m}\tilde{a}_j x_{ij}\right|$). It may be noted that the objective of the scientist is to obtain the best solutions – the solution that best represents the factual relationship among the variables - the solution that is least affected by the sample aberrations – and not to stick to any convention or prefer one norm to the others. Evidently, since the 'least absolute norm' estimator minimizes the sum of absolute residuals, it is less influenced by outliers and, in the presence of outliers, its sampling variability is less than that of the LS estimator. This estimator is called by

various names such as LAD (Least Absolute Deviation), LAR (Least Absolute Residual) and MAD (minimum Absolute Deviation) estimator.

The idea is not new. Among the multitude of measures of central tendency, the arithmetic mean is obtained by minimizing the Euclidean norm of errors and therefore, it is an estimator of the population mean that minimizes the error variance (or the squared expected Euclidean norm of deviations about the sample mean). However, the median is obtained by minimizing the absolute norm of deviations from the measure of the central tendency. Thus we define arithmetic mean, $\hat{a}$, and median, $\tilde{a}$ in

$$y_i = x_i \mathbf{a} \qquad \ldots (1.4)$$

(where x is a unitary vector or $x_i = 1$ for i =1,2,...,n) such that $\hat{a}$ minimizes $S_2 = \sum_{i=1}^{n} |y_i - x_i\hat{a}|^2 \equiv \sum_{i=1}^{n} (y_i - x_i\hat{a})^2$ for the choice of $\hat{a}$ as an estimator of **a** and $\tilde{a}$ minimizes $S_1 = \sum_{i=1}^{n} |y_i - x_i\hat{a}|$ as an estimator of **a** in (1.4) above. It is well known that $\tilde{a}$ (the median of y) is unaffected by extremal values of the variate y and therefore, the outliers in y, while $\hat{a}$ (the arithmetic mean of y) is oversensitive to the extremal values and outliers. Considering in a more general framework, $\tilde{a}$ and $\hat{a}$ are obtained by minimization of the Minkowski norm $L_p$ for p=1 and p=2 respectively.

The recommendations for using LAD estimator may be traced back to **Gauss** and **Laplace** (1818) as mentioned by **Taylor** (1974). **Edgeworth** (1887, 1888, 1923), **Rhodes** (1930) and **Singleton** (1940) investigated into this method of estimation. But at those times, computational difficulties involved with the method went in its disfavour. However, development of linear programming (LP) and fast computing machines intensified the

interest of researchers in estimation of regression models by minimization of $L_1$ norm. **Charnes, Cooper & Ferguson** (1955) were the first to transform the problem of estimation of parameters of a linear (regression) model into an LP problem. Later, iterative and search methods to obtain the solution were also discovered. **Fisher** (1961), **Ashar & Wallace** (1963), **Meyer & Glauber** (1964), **Rice & White** (1964), **Usow** (1967), **Oveson** (1968), **Robers & Ben-Israel** (1969), **Abdelmalek** (1971, 1974), **Blattberg & Sargent** (1971), **Smith & Hall** (1972), **Barrodale & Roberts** (1973), **Schlossmacher** (1973), **Fair** (1974), **Taylor** (1974), **Nyquist & Westlund** (1977), **Bassett & Koenker** (1978), **Powell** (1984), etc. intensively worked on the $L_1$ estimator of single equation (regression) models. **Powel** (1984) deserves a special mention since he extended the application of LAD estimation to censored regression.

Investigations into the performance of $L_1$ estimator of single equation models clearly establish its superiority to the conventional $L_2$ (that is, Least Squares) estimator when errors contain outliers and hence are thick-tailed. In this regard, **Bassett** and **Koenker** (1978) deserve a special mention. They analytically established that $L_1$ estimator has strictly smaller asymptotic ellipsoids than $L_2$ estimator for linear models from any F for which the sample median is a more efficient estimator of location than the sample mean. Side by side Monte Carlo experiments also were conducted to compare $L_1$ and $L_2$ estimators. These studies also strongly indicated the superiority of $L_1$ to $L_2$ estimator for single equation linear models with disturbances infested with outliers. **Pollard** (1991), **Phillips** (1991), **Chen** (1996), **Hitomi & Kagihara** (2001), etc. are some recent works on LAD estimation and its extension.

Multi-equation linear econometric models, described as YA+XB+E = 0, were first estimated by minimization of $L_2$ norm. These estimators are collectively called the k-class estimators (Indirect Least Squares – ILS, Two-Stage Least Squares - 2-SLS and Limited Information Max Likelihood - LIML). The ILS has very limited application due to its applicability in case of exactly identified equations only. In 2-SLS (suitable to estimating exactly as well as over-identified equations), OLS is used to estimate the matrix of Reduced Form Coefficients (P) at the first stage, which also gives estimated Y ($\hat{Y}$ = XP). In the second stage, Y is replaced by $\hat{Y}$ if and only if it appears as an explanatory variable in any structural equation. After this replacement, OLS is applied to estimate the structural equations, one at a time. Thus, OLS is applied twice, once at each stage. The 2-SLS is also an Instrumental Variable method of estimation. In spite of OLS – the basic building block of 2-SLS – being an ideal estimator if the required conditions for its application are met, 2-SLS is ordinarily a biased but consistent estimator. It was found that $L_2$-based k-class estimator performs extremely poorly when non-normally distributed, hyper-kurtic or outlier-infested errors are met with.

**Glahe & Hunt** (1970) were the first to apply LAD estimation method to multi-equation linear models. **Amemiya** (1980, 1982) extended LAD ($L_1$ estimator) to multi-equation models. **Newey** (1985), **Pagan** (1986), **Fair** (1994) and **Kim & Muller** (2000) are some important studies on 2-SLAD estimation.

**Amemiya** (1982) is certainly a landmark in the history of investigation into the possibility of applying $L_1$ estimation method to multi-equation linear models. He conjectured that if the disturbances in the structural as well as reduced form equations of a multi-equation model follow 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. Based on this conjecture he developed D2SLAD (Double Two-Stage Least Absolute) estimator. Thus he generalized LAD to include 2-SLS as its special case. Amemiya's work is theoretical and his conclusions relate to asymptotic properties (consistency) of LAD in estimating the multi-equation model. He suggested estimation by D2SLAD (Double 2-Stage LAD) in case of full non-normal and outlier infested errors. Amemiya suggested Monte Carlo experiments to be carried out in order to study the properties of 2SLAD estimator vis-à-vis those of D2SLAD and 2-SLS.

What has been described above is in essence an investigation into an estimator parallel to the Basmann-Theil estimator (2-SLS) of a multi-equation system, where instead of the Euclidean norm the absolute norm is minimized. The idea of estimation in two stages is maintained. However, there is another parallel to 2-SLS suggested by **Khazzoom** (1976) that is more akin to the ILS.

Khazzoom investigated into generalization of ILS for an over-identified equation. He estimated 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 by the 2-SLS, he applied generalized inverse of the relevant sub-matrix of reduced form coefficients to obtain the structural coefficients. This may be called the GILS (Generalized Indirect Least Squares) estimator.

It is natural to conjecture that since Khazzoom minimizes the Euclidean norm of errors, his method of estimation may share the shortcomings of ILS and 2-SLS, especially when the error vectors contain outliers. It is plausible that in presence of outliers an

estimator parallel to that of Khazzoom (but minimizes the absolute norm of errors) is more effective.

**II. Objectives of the Study**: The description in the preceding section suggests investigation in several lines. In the two-stage estimation procedure one may replace OLS by LAD

(i) at the first stage only (to obtain reduced form coefficients) and apply OLS at the second stage, or

(ii) at the second stage only (to obtain the structural coefficients while the reduced form coefficients at the first stage are obtained by OLS), or

(iii) at the first stage to obtain reduced form coefficients and consequently at the second stage also to obtain the structural coefficients. Suggestions given by Amemiya need implementation and empirical/experimental corroboration.

Similarly, in case of the Khazzoom estimator that applies OLS to obtain the reduced form coefficients, one may apply LAD instead for the same purpose. The present study basically aims at investigating into these possibilities. In particular, the generic objectives of the present study are given as hereunder:

(i) To study the relative performance of the LAD estimator (vis-à-vis OLS) in case of single equation models containing error vectors of different statistical distributions infested with outliers. It is required since single equation estimation is the basic building block in all methods that concern us.

(ii) To study the relative performance of the LAD estimator applied at different stages (vis-à-vis the standard 2-SLS estimator) in estimating the multi-equation models containing error vectors of different statistical distributions infested with outliers.

(iii) To study the relative performance of the LAD-based Khazzoom type estimator vis-à-vis the original Khazzoom estimator as well as various types of two-stage estimators applying OLS or LAD estimator in estimating the multi-equation models containing error vectors of different statistical distributions infested with outliers.

**III. The Methodology**: Owing to very poor mathematical tractability of LAD estimator we have relied on the *Monte Carlo method* to study the properties of various methods of estimation that we purport to investigate into. The Monte Carlo method of investigating into the properties of a mathematical system is basically experimental and numerical in nature. This method may be loosely described as a statistical simulation method, where statistical simulation is defined in quite general terms to be any method that utilizes sequences of random numbers to perform the simulation. The main idea behind the method is either to construct a stochastic model that is in agreement with the whole problem analytically or to simulate the whole problem directly. In both cases, an element of randomness has to be introduced according to some well-defined rules. Then a large number of trials are performed, the results are observed, and finally a statistical analysis is undertaken in the usual way. The advantages of the method are that even very difficult or analytically intractable problems can often be treated very easily, and desired modifications can be applied without much trouble. Nevertheless, the conclusions emanating from Monte Carlo studies are relatively less (vis-à-vis analytical methods)

accurate and the large number of trials that are necessary. Now, when very fast computers are readily available, the cumber of calculation in large number of trials is no longer a serious matter. This facility may be utilized to increase the number of trials sufficiently large so as to attain an acceptable degree of precision. Thus the trade-off between cumber and precision goes in favour of the Monte Carlo method.

In this investigation we have limited ourselves to deal with linear econometric models. Our single equation models are described as $y = X\mathbf{a} + e$. The multi-equation models are described as $YA + XB + E = 0$. The disturbance term (e or E as the case may be) is stochastic and follows either of the five distributions : (i) Normal, (ii) Cauchy, (iii) $Beta_1$, (iv) $Beta_2$ and (v) Gamma. We have experimented with the cases when the disturbance term is infested with outliers, different in numbers and sizes. In the simulation we have experimented with the models of various sizes (different number of variables or equations) as well as with various sample sizes implying Y and X of varying dimensions.

In case of single equation models described as $y = X\mathbf{a} + e$, (i) first X of the desired dimension is generated. The coefficients, **a**, are assumed. These X and **a** remain fixed in a single bout of experiment. (ii) Then e of desired distribution is generated and $y = X\mathbf{a} + e$ is obtained. The desired number of outliers (of sizes within some desired range) are generated and added to y. Next, the desired method(s) of estimation is (are) applied to obtain $\hat{a}$ which is the estimator of **a**. The step # ii is repeated a large, say m, number of times. This gives m number of $\hat{a}$ 's which are subjected to statistical analysis to assess the properties of the method(s) of estimation.

In case of multi-equation models described as $YA + XB + E = 0$ : (i) first X of the desired dimension is generated and the coefficients A and B are assumed. From A and B

we obtain $\Pi = -BA^{-1}$. (ii) Then E of desired distribution is generated and $Y = X\Pi - EA^{-1}$ is obtained. The desired number of outliers (of sizes within some desired range) are generated and added to Y. Next, the desired method(s) of estimation is (are) applied to obtain $\hat{A}$ and $\hat{B}$ which are the estimators of A and B. The step # ii is repeated a large, say m, number of times. This gives m number of $\hat{A}$'s and $\hat{B}$'s which are subjected to statistical analysis to assess the properties of the method(s) of estimation.

**IV. Organization of the Work**: The subsequent part of this work is organized as follows: First, in chapter-2, the framework of linear econometric models is presented. Presently, estimation of econometric models by minimization of the Euclidean norm of errors is the most popular practice and hence, most of the estimation techniques are based on the Least Squares principle directly or indirectly. Details on these techniques are given therein.

In chapter-3 we present a literature survey on LAD estimation and the relevant numerical algorithms for that purpose. Estimation by minimization of absolute errors is mostly intractable by analytical methods and its performance (vis-à-vis LS-based methods) can in general be assessed only by simulation-based empirical methods such as the Monte Carlo method. In chapter-4, we present the salient features of the Monte Carlo method and give an outline that paves a way to using this simulation method to assess the relative performance of LAD-based estimators in single as well as multi equation econometric models. Since the Monte Carlo method is based on generating random numbers, we also present a description of some relevant statistical distributions that random numbers can follow and the computer-based techniques that may be used to generate random variates following such statistical distributions.

Next, in chapter-5, we conduct Monte Carlo experiments to compare the performance of LAD-based and LS-based methods of estimation of single equation econometric models. The residual terms in these models follow five different statistical distributions viz. normal, Cauchy, Gamma, $Beta_1$ and $Beta_2$. The residual terms may contain outliers different in number and magnitude.

Chapter-6 is an extension of the earlier chapter to the case of multi-equation econometric models. However, unlike in case of single-equation econometric models, we have several LS-based methods of estimation in case of the multi-equation econometric models. That is so because these models are estimated in two stages. Consequently, it is possible to replace LS by LAD at any one or both of these stages, giving rise to many possible combinations and possibilities of a comparative study in their performance. This has been done in Chapter-6.

We present a summary of our results in chapter-7. We have also visualized the lines of future research in application of LAD estimators of econometric models.

# CHAPTER 2

## THE FRAMEWORK OF LINEAR ECONOMETRIC MODELS AND THEIR ESTIMATION

**I. Introduction:** Construction of an econometric model is a pre-requisite for any econometric research. A model abstractly represents reality by bringing out what is relevant to a particular question neglecting all other unnecessary details. These models may be of different sizes and complexities. The most simple of them may involve single relationship among economic variables and, therefore, can be summed up in a single equation such as:

$$y = a_0 x_0 + a_1 x_1 + \ldots + a_k x_k + e$$

where $x_0$ is a unit (column) vector. In the matrix form this model is simply expressed as $y = Xa + e$. Such a model is popularly known as the multi-variate regression model. However, this simple relationship is the most fundamental building block of more complex models known as the multi-equation models that may consist of an impressive array of relations.

Multi-equation models almost invariably exhibit simultaneity relations in different variables of the model. Simultaneity occurs in an econometric model within which more than one causal relationship among the variables is specified. When interrelations exist among the variables, the single equation specification (as illustrated above) with its one way implied causality from predetermined to one endogenous variable is neither an accurate nor sufficient representation. In such cases, the economic variables are determined by a complete system of equations. A complete system is the representation

(model) in which there are as many equations as endogenous variables (whose formation is to be 'explained' by the equations). The equations are usually of four types: equations of economic behaviour, institutional rules, technological laws of transformation and identities. The term 'structural equation' is used to comprise all four types of equations.

Systems of structural equation may be composed entirely on the basis of economic theory or on the dual basis of economic theory combined with systematically collected statistical data for the relevant variable for a given period. These models serve as an indispensable aid to forecasting and an invaluable guide to policy-making whether for a firm or a governmental agency.

Since a characteristic feature of most economic models is interdependence and joint determination of several variables, the use of simultaneous equation model is more meaningful and logical. The seminal papers from which the simultaneous equation model developed establish the importance of joint endogeneity for statistical analysis of economic relationships. **Haavelmo** (1943, 1944) realised that in the presence of jointly endogenous variables, a joint probability distribution is necessary to analyse the data.

**II. The General Simultaneous Equation Model**: Before proceeding further let us develop the general simultaneous equation model. For this let us take a linear simultaneous equation model containing m structural relations. The model may be best described in a matrix form as $YA + XB + E = 0$, where, Y represents m number of endogenous variables, the X represents k number of pre-determined variables, E symbolizes the stochastic disturbance terms m in number and A and B are the matrices of structural coefficients of compatible order. For the purpose of estimation, n is the sample size of data on Y and X. The underlying theory in general will satisfy that only a small

number of all the variables of the model will occur in any one equation, and the matrices A and B will have a considerable number of zeros. Further in each structural equation, one endogenous variable can by convention be regarded as the dependent variable, the coefficient of which is equal to unity. This is known as the normalization rule. By a suitable arrangement of the equations and the endogenous variables in the model, this rule helps to make the main diagonal elements such that $a_{ii} = -1$ for all $i$ in the matrix A. To take care of the constant terms in the equations, the first column of X is a unit vector such that $x_{i1} = 1$ for all $i = 1, 2, \ldots, n$.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix}; \quad B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ b_{k1} & b_{k2} & \cdots & b_{km} \end{bmatrix}; \quad 0 = \begin{bmatrix} 0_{11} & 0_{12} & \cdots & 0_{1m} \\ 0_{21} & 0_{22} & \cdots & 0_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 0_{n1} & 0_{n2} & \cdots & 0_{nm} \end{bmatrix}$$

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{bmatrix}; \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}; \quad E = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1m} \\ e_{21} & e_{22} & \cdots & e_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nm} \end{bmatrix}$$

In the model described above, A is a square matrix and is assumed to be non-singular, since if it were not, one or more of the structural relations would merely be a linear combination of other structural relations, thus being redundant, or if the rows of B did not obey the same linear restrictions as the rows of A, the m structural equations would be inconsistent. The B matrix is generally not a square matrix. The m equations in the structural from jointly determine for each observation, the value of m endogenous variables, given k pre-determined variables, m stochastic disturbance term and m(m+k-1) coefficients of the system. The stochastic disturbance terms are assumed to be identically

and independently distributed over the samples with zero mean and constant covariance matrix.

**III. Estimation of Single Equation Econometric Models**: Single equation econometric models are often estimated by the method of Least Squares that minimizes the squared Euclidean norm of the error vector, S. It is tantamount to minimization of the Euclidean norm of the error vector.

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_k x_{ik})\}^2 .$$

To minimize S we differentiate S partially with respect to $a_i$ (i=0, 1, 2, …,k) and set those partial derivatives equal to zero.

$$\begin{aligned}
&\frac{\partial S}{\partial a_0} = -2\sum_{i=1}^{n} \{y_i - (a_0 + a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3})\}(1) = 0 \\
&\frac{\partial S}{\partial a_1} = -2\sum_{i=1}^{n} \{y_i - (a_0 + a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3})\}(x_{i1}) = 0 \\
&\ldots \quad \ldots \quad \ldots \qquad\qquad \ldots \qquad\qquad \ldots \quad \ldots \\
&\frac{\partial S}{\partial a_k} = -2\sum_{i=1}^{n} \{y_i - (a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_k x_{ik})\}(x_{ik}) = 0
\end{aligned} \qquad \ldots (2.1)$$

Simplifying the above equations we obtain the so-called normal equations as follows:

$$\begin{aligned}
&\sum_{i=1}^{n} y_i = n a_0 + a_1 \sum_{i=1}^{n} x_{i1} + a_2 \sum_{i=1}^{n} x_{i2} + ... + a_k \sum_{i=1}^{n} x_{ik} \\
&\sum_{i=1}^{n} y_i x_{i1} = a_0 \sum_{i=1}^{n} x_{i1} + a_1 \sum_{i=1}^{n} x_{i1}^2 + a_2 \sum_{i=1}^{n} x_{i1} x_{i2} + ... + a_k \sum_{i=1}^{n} x_{i1} x_{ik} \\
&\ldots \quad \ldots \qquad\qquad \ldots \qquad\qquad \ldots \qquad\qquad \ldots \\
&\sum_{i=1}^{n} y_i x_{ik} = a_0 \sum_{i=1}^{n} x_{ik} + a_1 \sum_{i=1}^{n} x_{i1} x_{ik} + a_2 \sum_{i=1}^{n} x_{i2} x_{ik} + ... + a_k \sum_{i=1}^{n} x_{ik}^2
\end{aligned} \qquad \ldots (2.2)$$

This system of normal equations has k+1 unknowns in k+1 equations. Assuming linear independence of these equations, the system may be solved for **a**. This method of obtaining '*best*' **a** ( $= \hat{a}$ ) is called the *Principle of Least Squares.*

**A purely Algebraic approach to obtain** $\hat{a}$ : The linear model described above may also be expressed in terms of (sample) observations. It is an inconsistent system of n linear equations in k+1 variables.

$$
\begin{aligned}
y_1 &= a_0x_{10} + a_1x_{11} + a_2x_{12} + \ldots + a_kx_{1k} \\
y_2 &= a_0x_{20} + a_1x_{21} + a_2x_{22} + \ldots + a_kx_{2k} \\
y_3 &= a_0x_{30} + a_1x_{31} + a_2x_{32} + \ldots + a_kx_{3k} \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
y_n &= a_0x_{n0} + a_1x_{n1} + a_2x_{n2} + \ldots + a_kx_{nk}
\end{aligned}
\qquad \ldots (2.3)
$$

The new variable $x_0 = [x_{10}, x_{20}, x_{30}, \ldots, x_{n0}]'$ appearing in the first term with the coefficient $a_0$ is a unitary (column) vector taking on the value of unity (that is, 1) in all equations. The system of linear equations in (2.3) above is inconsistent since no value of **a** $= [a_0, a_1, a_2, \ldots, a_k]'$ can satisfy the equality relations. Alternatively expressed in matrix notations, the system of equations y = X**a** is an inconsistent system of (linear) equations. Pre-multiplying the equation y = X**a** by G, we have Gy = GX**a**. Now suppose GX = I, an identity matrix, and the solution vector $\hat{a}$ = Gy. If G = $(X'X)^{-1}X'$, then obviously GX = $(X'X)^{-1}X'X$ = I. Here G is the (generalized) inverse of X (**Krishnamurthy & Sen**, 1976, p. 155, pp. 196-198; **Rao & Mitra**, 1971) . Therefore, $\hat{a}$= Gy = $(X'X)^{-1}X'y$. The property of $\hat{a}$ thus obtained is that it minimizes the squared Euclidean norm S = e'e

$=(Y-X\hat{a})'(Y-X\hat{a})$ of the disturbance vector e. It also implies that the Euclidean norm of e is minimized by $\hat{a}$. If we considered y* = (y-e) instead of y, the system y* = **Xa** is over-determined, though consistent. The over-determined system of equations will admit generalized solution

$$\hat{a}^* = (X'X)^{-1}X'y^* = (X'X)^{-1}X'y - (X'X)^{-1}X'e.$$

Now if $(X'X)^{-1}X'e = 0$, we have $\hat{a} = \hat{a}^*$.

Strictly speaking, the method of Least Squares to obtain $\hat{a}$ is a mathematical rather than statistical method because it does not incorporate randomness or probability considerations in obtaining $\hat{a}$ nor does it give us any hint on the statistical properties of $\hat{a}$, although $\hat{a}$, the best estimator of **a**, is obtained by minimizing the sample S = e'e which is stochastic in nature. Since $\hat{a} = (X'X)^{-1}X'y$ where y is stochastic, $\hat{a}$ assumes stochasticity because of the simple fact that the function of a random variable is in turn a random variable.

It may be shown (**Johnston**, pp. 171-174) that under certain assumptions $\hat{a} = (X'X)^{-1}X'y$ is the best linear unbiased estimator (**blue**) of **a** in y = X**a** + e. These assumptions are known as the Gauss-Markov assumptions of (linear) regression model $y = Xa + e$ (**Intriligator**, pp. 106-109). These assumptions are: (i) $E(e_i) = 0 \Rightarrow E(y_i) = \sum_{j=0}^{m} a_j x_{ij}; \forall i, i = 1,2,...,n.$ In general, E(e) = 0 $\Rightarrow$ E($y_i$) = X**a**. Here E(.) means the expectation of the random variable (.). (ii) $E(ee') = \sigma^2 I$. This assumption asserts that each disturbance $e_i$ is distributed with the same positive and finite variance $\sigma^2$ and all

disturbances are pair-wise uncorrelated. This is also called the assumption of homoskedasticity and non-autocorrelation (or geometrically, the spherical disturbances). (iii) The matrix of sample values of the explanatory variables, X, is a fixed non-stochastic matrix of full rank. Although the assumption of normality of e is not required to prove **blue** properties of $\hat{a}$, such an assumption makes $\hat{a}$ the most efficient estimator of **a** in y = Xa + e since in that case $\hat{a}$ is also a maximum likelihood estimator of **a**, attaining the Cramer-Rao bound (**Intriligator**, p. 108).

However, reality often breaks away with the Gauss-Markov assumptions that generously bestow upon the Least Squares estimator, $\hat{a}$, the properties of an ideal estimator. Very often disturbances are non-spherical. It has been shown that if e has a non-spherical distribution, the Aitken estimator (or Generalized Least Squares - **GLS** - estimator) $\tilde{a}$ of **a** is **blue**. The GLS estimator $\tilde{a}$ is given by

$$\tilde{a} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}y \qquad \ldots (2.4)$$

The estimator is 'general' in the sense that for spherical disturbances $\Omega = I$, $\tilde{a}$ reduces to the (ordinary) Least Squares estimator $\hat{a}$, that is, $\tilde{a} = \hat{a}$. In practice, however, $\Omega$ is not known and therefore it has to be estimated from the sample observations (**Intriligator**, pp. 168-173) and consequently the GLS estimator is only consistent (**Theil**, pp. 398-400).

Occasionally, in single equation models, the assumption regarding the non-stochasticity of the sample values of the explanatory variables may not be satisfied. That makes the (ordinary) Least Squares estimator, $\hat{a}$, an inconsistent estimator (**Theil**, p. 608). There are two suitable estimators that deserve a mention in this regard. First, the

Wald-Hooper-Theil estimator (**Hooper, JW & H Theil** (1958)) and the second, the Instrumental Variable method (**Reiersol, O.** (1950)). In multi-equation (or simultaneous equation) models stochasticity of explanatory variables is invariably met with. Therefore, (ordinary) Least Squares estimator is an inconsistent estimator of the structural parameters of multi-equation models. Other methods such as Two-Stage Least Squares, Limited Information Maximum Likelihood method, etc. are applied to estimation of parameters of these models.

The violation of the first Gauss-Markov assumption is no less serious. It introduces bias into the Least Squares estimator. If the estimator continues to be biased even if the sample size is increased indefinitely (although in practice it is often impossible to increase the sample size indefinitely), it is futile to investigate into other desirable properties of the estimator. The presence of outliers in the errors often leads to the violation of the first Gauss-Markov assumption and the method of Least Squares yields unreliable estimates of the parameters. This fact prompts us to investigate into the alternative methods of estimation such as the Least Absolute Deviation Estimation.

**IV. Estimation of Multi-equation or Simultaneous Equation Econometric Models**: It is well-known now that in a system of simultaneous equations the method of (ordinary) least squares yields biased and inconsistent estimators (Intriligator, p. 376) of the structural coefficients. This was first established by Haavelmo (1943, 1947). He showed that simultaneity induces correlation between the regressors and the residuals giving inconsistent and biased estimates. In his path-breaking monograph, "*The probability Approach in Econometrics*" (1944) he developed a method for solving the problem of single equation bias.

**Marschak** (1953) has explained why we must try to establish the structural coefficients. The basic argument is that by its very definition each structural equation describes a specific, distinct and thereby autonomous link in the economic process. This autonomous character of structural relations also means that the structural coefficients are more stable, more like physical constants, than are the reduced form composites. Structural coefficients unlike the reduced form coefficients are more easily judged by intuition and their changes are better capable of reasonable discussion and interpretation. Structural coefficients are more easily comparable with the accumulated empirical evidence, which refers to the economic structure and not to the reduced form.

In the way to establishing the structural coefficients of a simultaneous equation model one has to estimate the reduced form equations. Haavelmo proposed to first estimate the reduced form of a system of simultaneous equations by ordinary least squares and then to derive estimates for the parameters of the structural form by ***indirect least squares***. In the case of exactly identified systems, **Haavelmo** (1943, 1944, 1947) showed that this method is equivalent to the method of maximum likelihood. Considerable theoretical investigations were also carried out by the researchers in **Cowles Foundation,** e.g., **Koopmans** (1950) and **Hood & Koopmans** (1953), etc. on methods of estimating structural parameters of a system of simultaneous linear stochastic equations. The estimates have been derived from maximum likelihood considerations and proved to be asymptotically unbiased and efficient.

Haavelmo's investigations made the 'identification problem' central to the estimation of simultaneous equation models. The identification problem may be cogently described as follows.

From the structural equations YA + XB + E = 0, we may obtain the reduced form equations through pre-multiplying the system of structural equations by $A^{-1}$. That is explicitly, $YAA^{-1} + XBA^{-1} + EA^{-1} = 0$ or $Y = X\Pi + U$, where $\Pi = -BA^{-1}$ and $U = -EA^{-1}$. Now since the reduced form equations are amenable to estimation by Ordinary Least Squares, we obtain the least squares estimates of $\Pi$ as $P = (X'X)^{-1}X'Y$. But $\Pi A = -B$ and hence $P\hat{A} = -\hat{B}$. This gives us a system of k equations in m(m+k-1) unknowns even after applying the normalization rule on A such that $a_{ii} = -1$ for all $i = 1,2,...,m$. Obviously, such a system of equations admits no determinate solution. For any particular structural equation, say $j^{th}$ equation, we have $P\hat{a}_j = -\hat{b}_j$ which is a system of k equations in m+k-1 unknowns. Unless we assume some (at least as many as m-1) elements of $\hat{a}_j$ or $\hat{b}_j$ equal to zero, we cannot estimate other remaining elements of $\hat{a}_j$ and $\hat{b}_j$. Thus, the identifiability or the zero restriction on <u>*at least*</u> m-1 elements of $\hat{a}_j$ and $\hat{b}_j$ is the basic requirement for obtaining the estimates of the coefficients of the $j^{th}$ structural equation. Moreover, the system of equations $P\hat{a}_j = -\hat{b}_j$ must satisfy the rank condition which is the basic algebraic requirement for solving any system of linear equations.

In what follows we describe various methods to estimate the structural equations of a multi-equation (simultaneous equation) model. These methods minimize the Euclidean norm of errors in the dependent variable directly or indirectly. Identifiability of each equation in the model is presumed.

**Indirect Least Squares and Generalized Indirect Least Squares**: The Indirect Least Squares (ILS) method was first proposed by **Girshick** (ref. by Hood & Koopmans, 1953, p. 140). It was generalized by **Khazzoom** (1976), though he did not name it the

Generalized Indirect Least Squares (GILS). For a comprehensive treatment we will describe here the method of GILS only.

From the structural equations YA + XB + E = 0 we derive the reduced form equations Y = XΠ +U and estimate its coefficients by P = $(X'X)^{-1}X'Y$ by OLS. Since $(X'X)^{-1}X' = X^{-g}$ (the generalized inverse of X), we may also write $P = X^{-g}Y$.

The question that arises now is whether we can obtain $\hat{A}$ and $\hat{B}$ from the knowledge of P. This is the problem of identification. If all the columns of A and B can be obtained then A and B can also be completely obtained. If a particular column of A and B (say, $a_j$ and $b_j$) can be obtained from the relation

$$Pa_j = -b_j \qquad \ldots (2.5)$$

then the $j^{th}$ equation is identifiable. Since this is true for any equation and (2.5) holds for any particular column of A and B, dropping the subscript j the equation (2.5) can be rewritten as

$$\underset{k\times m}{P}\ \underset{m\times 1}{a} = \underset{k\times 1}{-b} \qquad \ldots (2.6)$$

The $j^{th}$ element of $a$ is –1 if (2.6) relates to the equation j. The system (2.6) is in k equations and k+m-1 unknowns. So equation (2.6) cannot be solved for $a$ and $b$.

Now, suppose, from a-priori information the values of some (say $k_2$) elements of $b$ and some other elements of $a$ can be obtained such that $k_2$ elements of $b$ and $m_2$ elements of $a$ are known . In that case, out of k elements of $b$ only $k-k_2 = k_1$ elements are unknown. Similarly out of m elements of $a$, $m - m_2 = m_1$ elements are unknown. Thus, we proceed to identification by restriction on the structural coefficients matrix $a$ and $b$.

Pre-multiplying (2.6) by a suitable permutation matrix R we may obtain

$$RP\,a = -\,R\,b \qquad \ldots (2.7)$$

such that $b$ can be partitioned into two sub matrices, first of which, $b_1$ has $k_1$ unknown elements and the second, $b_2$ has $k_2$ known elements.

Correspondingly, RP can also be partitioned into $P_1$ and $P_2$. It may noted that such a permutation effects only a re-shuffling of equations in (2.6) and does not have any bearing on its solution. Consequently, equation (2.7) can thus be re-written in a partitioned form as

$$\left[\frac{P_1}{P_2}\right]a = -\left[\frac{b_1}{b_2}\right]; \quad R^{-1}\left[\frac{b_1}{b_2}\right] = b \qquad \ldots (2.7a)$$

Here, $P_1$, $P_2$, $b_1$ and $b_2$ are $k_1 \times m$, $k_2 \times m$, $k_1 \times 1$ and $k_2 \times 1$ matrices respectively.

Again, pre-multiplying $a$ by a suitable permutation matrix, S, and post-multiplying RP by $S^{-1} = S'$ in equation (2.7a) we obtain

$$\left[\frac{P_1}{P_2}\right]S^{-1}Sa = \left[\frac{P_1}{P_2}\right]S^{-1}\left[\frac{a_1}{a_2}\right]; \qquad S^{-1}\left[\frac{a_1}{a_2}\right] = a \;;\; S^{-1}S = I$$

where $a_1$ contains $m_1$ unknown elements and $a_2$ contains $m_2$ known elements. Equation (2.7a) may now be rewritten as

$$RPS^{-1}Sa = -Rb \qquad \ldots (2.7b)$$

This operation helps to shuffle and partition RP into two column submatrices, without any effect to the solution of the system of equations in (2.6). In the partitioned form (2.7b) can be written as

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = -\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad \ldots (2.8)$$

From (2.8) we obtain two equations,

$$P_{11}a_1 + P_{12}a_2 = -b_1 \quad \ldots (2.9)$$

$$P_{21}a_1 + P_{22}a_2 = -b_2 \quad \ldots (2.10)$$

Now, since $a_2$ and $b_2$ are known, we get

$$P_{21}a_1 = -[b_2 + P_{22}a_2] \quad \ldots (2.11)$$

Here $P_{21}$ is a $k_2 \times m_1$ matrix giving rise to $k_2$ equations in $m_1$ unknowns as in (2.10).

If $P_{21}$ is a square (rectangular) matrix in $k_2$ rows and $m_1$ column and rank $(P_{21}) = r \leq m_1$, then the Moore-Penrose inverse of $P_{21}$ that is $P_{21}^+$ exists and is unique. Using this generalized inverse of $P_{21}$ we obtain from (2.9) and (2.11)

$$a_1 = -P_{21}^+[b_2 + P_{22}a_2] \quad \ldots (2.12)$$

$$b_1 = -[P_{11}a_1 + P_{12}a_2] \quad \ldots (2.13)$$

This formulation and consequent technique in (2.12) and (2.13) is GILS. **The ILS is only a special case of GILS when $k_2 = m_1 = r$ and therefore, $P_{21}^+ = P_{21}^{-1}$.** However, in case $k_2 \geq m_1$ and $r = m_1$, then $P_{21}^+$ yields least squares g-inverse, $P_{21}^{-1}$ or $P_{21}^{-g}$.

**GILS in Another Form:** Rewriting equation (2.8) as

$$\begin{bmatrix} P_{11} & I \\ P_{21} & 0 \end{bmatrix} \begin{bmatrix} a_1 \\ b_1 \end{bmatrix} = -\begin{bmatrix} P_{12}a_2 \\ b_2 + P_{22}a_2 \end{bmatrix} \quad \ldots (2.14)$$

which may be again rewritten as

$$\begin{bmatrix} I & P_{11} \\ 0 & P_{21} \end{bmatrix} \begin{bmatrix} b_1 \\ a_1 \end{bmatrix} = -\begin{bmatrix} P_{12}a_2 \\ b_2 + P_{22}a_2 \end{bmatrix} \quad \ldots (2.15)$$

From (2.15) we obtain

$$\begin{bmatrix} I & P_{11} \\ 0 & P_{21} \end{bmatrix}^{+} \begin{bmatrix} P_{12}a_2 \\ b_2 + P_{22}a_2 \end{bmatrix} = -\begin{bmatrix} b_1 \\ a_1 \end{bmatrix} \quad \ldots (2.16)$$

Now,

$$\begin{bmatrix} I & P_{11} \\ 0 & P_{21} \end{bmatrix} \begin{bmatrix} I & -P_{11}P_{21}^{+} \\ 0 & P_{21}^{+} \end{bmatrix} \begin{bmatrix} I & P_{11} \\ 0 & P_{21} \end{bmatrix} = \begin{bmatrix} I & P_{11} \\ 0 & P_{21} \end{bmatrix} \quad \ldots (2.17)$$

or,

$$\begin{bmatrix} I & 0 \\ 0 & P_{21}P_{21}^{+} \end{bmatrix} \begin{bmatrix} I & P_{11} \\ 0 & P_{21} \end{bmatrix} = \begin{bmatrix} I & P_{11} \\ 0 & P_{21} \end{bmatrix} \quad \ldots (2.18)$$

or,

$$\begin{bmatrix} I & P_{11} \\ 0 & P_{21}P_{21}^{+}P_{21} \end{bmatrix} = \begin{bmatrix} I & P_{11} \\ 0 & P_{21} \end{bmatrix} \quad \ldots (2.19)$$

Since $P_{21}\,P_{21}^{+}P_{21} = P_{21}$ (Theil, pp. 269-70), (2.19) is equal to (2.17).

Thus,

$$\begin{bmatrix} I & P_{11} \\ 0 & P_{21} \end{bmatrix}^{+} = \begin{bmatrix} I & -P_{11}P_{21}^{+} \\ 0 & P_{21}^{+} \end{bmatrix} \quad \ldots (2.20)$$

It goes without saying that since the least squares g-inverse is only a special case of Moore-Penrose inverse, the above proof is true for g-inverse $[\bullet]^{-g}$ as well.

Equation (2.16) may now be written as

$$\begin{bmatrix} I & -P_{11}P_{21}^{+} \\ 0 & P_{21}^{+} \end{bmatrix} \begin{bmatrix} P_{12}a_2 \\ \hline b_2 + P_{22}a_2 \end{bmatrix} = -\begin{bmatrix} b_1 \\ \hline a_1 \end{bmatrix} \quad \ldots (2.21)$$

Hence, $a_1 = -P_{21}^{+}[b_2 + P_{22}a_2]$ and $b_1 = -[P_{11}a_1 + P_{12}a_2]$ which is the same as in (2.12) and (2.13) before.

**Two Stage Least Squares**: The most widely used method for estimating the parameters of an equation of a simultaneous equation model is the 2-SLS developed by H. Theil and independently by Basmann who called it Generalized Classical Linear Estimation method. It is applicable to over-identified/exactly identified equations. In the case of exactly identified equation, the 2-SLS estimates are identical with the ILS estimates given by equations (2.12) and (2.13). The basic idea behind the 2-SLS is to substitute each endogenous explanatory variable (that is correlated with the residuals) by a corresponding composite variable (uncorrelated with the residuals) which is a linear function of all exogenous variables in the model. Since the surrogate (instrumental) variables are uncorrelated with the residuals, the estimates of the parameters will be consistent.

Let us suppose that we are interested in estimating the first equation of the general interdependent system of equations.

$$y_1 = Y_1 a_1 + X_1 b_1 + e_1 \quad \ldots (2.22)$$

In the notation above, $y_1$ is the dependent endogenous variable in the first structural equation and on account of normalization rule the coefficient associated with it is unity. $Y_1$ and $X_1$ are explanatory endogenous and exogenous variable (respectively) that appear in the first structural equation while $a_1$ and $b_1$ are the coefficients associated with them. In the first stage, reduced form regression of $Y_1$ on all the pre-determined variables in the

system (X) is performed. The estimates $\hat{Y}_1$ thus obtained are then used to replace the actual observations on the $Y_1$ variables. Hence,

$$\hat{Y}_1 = X\left(X'X\right)^{-1}X'Y_1 \qquad \ldots (2.23)$$

Application of OLS in the second stage to the equation thus transformed yields the estimating equations.

$$\begin{pmatrix} \hat{Y}_1'\hat{Y}_1 & \hat{Y}_1'X_1 \\ X_1'\hat{Y}_1 & X_1'X_1 \end{pmatrix}\begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} \hat{Y}_1'y_1 \\ X_1'y_1 \end{pmatrix} \qquad \ldots (2.24)$$

The vector $\begin{bmatrix} c \\ d \end{bmatrix}$ now denotes the 2SLS estimator of $\begin{bmatrix} a_1 \\ b_1 \end{bmatrix}$. Since the above system contains the same number of equations and unknowns, i.e. m+k-1 equations in m+k-1 unknowns, the system in general will have a unique solution.

The form in which the 2SLS equations are usually presented can be derived in the following way:

The matrix $Y_1$ can be written as

$$Y_1 = \hat{Y}_1 + E_1 \qquad \ldots (2.25)$$

Now, since the OLS residuals are orthogonal to the estimated value of the dependent variable and to each of the explanatory variables, therefore,

$$\hat{Y}_1'E_1 = 0 \qquad \textit{and} \qquad X_1'E_1 = 0$$

Using equation (2.25) we get,

$$\begin{aligned}\hat{Y}_1'\hat{Y}_1 &= (Y_1 - E_1)'(Y_1 - E_1)\\ &= Y_1' Y_1 - Y_1' \mathrm{E}_1 - E_1' Y_1 + E_1' \mathrm{E}_1\\ &= Y_1' Y_1 - (\hat{Y}_1 + E_1)' E_1 - E_1'(\hat{Y}_1 + E_1) + E_1' \mathrm{E}_1\\ &= Y_1' Y_1 - \hat{Y}_1' \mathrm{E}_1 - E_1' \mathrm{E}_1 - E_1' \hat{Y}_1 - E_1' \mathrm{E}_1 + E_1' \mathrm{E}_1\\ &= Y_1' Y_1 - E_1' \mathrm{E}_1\end{aligned}$$

$$\begin{aligned}\hat{Y}_1' X_1 &= (Y_1 - E_1)' X_1\\ &= Y_1' X_1\\ X_1' \hat{Y}_1 &= X_1'(Y_1 - E_1)\\ &= X_1' Y_1\end{aligned}$$

*and* $\hat{Y}_1' y_1 = (Y_1 - E_1)' y_1$

Thus for $\kappa = 1$ the 2SLS estimator can now be written as

$$\begin{pmatrix} Y_1' Y_1 - \kappa(E_1' \mathrm{E}_1) & Y_1' X_1 \\ X_1' Y_1 & X_1' X_1 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} (Y_1 - \kappa E_1)' y_1 \\ X_1' y_1 \end{pmatrix} \qquad \ldots (2.26)$$

The form shows clearly how the 2SLS estimator differs from the inconsistent OLS estimator, which is given by

$$\begin{pmatrix} Y_1' Y_1 & Y_1' X_1 \\ X_1' Y_1 & X_1' X_1 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} Y_1' y_1 \\ X_1' y_1 \end{pmatrix} \qquad \ldots (2.27)$$

or

$$\begin{pmatrix} Y_1' Y_1 - \kappa(E_1' \mathrm{E}_1) & Y_1' X_1 \\ X_1' Y_1 & X_1' X_1 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} (Y_1 - \kappa E_1)' y_1 \\ X_1' y_1 \end{pmatrix} \text{ for } \kappa = 0 \qquad \ldots (2.28)$$

**Limited-Information Maximum Likelihood Estimator**: The limited information maximum likelihood (LIML) method is another alternative approach that gives consistent

estimates and was developed by **Anderson & Rubin**. They estimated the parameters of an equation by maximizing the likelihood function subject to appropriate constraints using Lagrange multipliers. Hood & Koopmans have used a different method to develop the LIML estimator. They arrived at the estimates of the parameters of an equation by maximizing the likelihood function for the observations on the endogenous variables included in that equation disregarding the over identifying restrictions on other structural equations. This method assumes that the structural disturbances are normally distributed.

The structural equation to be estimated is again

$$y_1 = Y_1 a_1 + X_1 b_1 + e_1 \qquad \ldots (2.29)$$

Equation (2.29) can be written as

$$y_1^* = X_1 b_1 + e_1 \qquad \ldots (2.30)$$

where, $y_1^* = y_1 - Y_1 a_1$. Now $y_1^*$ is a linear combination of the endogenous variables included in the first structural equation. The coefficient $b_1$ in equation (2.30) can be estimated by ordinary least squares method. Thus, the estimator is

$$\hat{b}_1 = \left(X_1' X_1\right)^{-1} X_1' y_1^* \qquad \ldots (2.31)$$

and the sum of squared errors is

$$S_1 = \left(y_1^* - X_1 b_1\right)' \left(y_1^* - X_1 b_1\right)$$

$$= y_1^{*\prime} y_1^* - y_1^{*\prime} X_1 \left(X_1' X_1\right)^{-1} X_1' y_1^* \qquad \ldots (2.32)$$

Now, let the relation which includes all the predetermined variables be given by

$$y_1^* = \mathrm{X}\, \mathrm{b} + \varepsilon_1 \qquad \ldots (2.33)$$

where, X is an nxk matrix of all pre-determined variables and

$$b = \begin{bmatrix} b_1 \\ O^{**} \end{bmatrix}$$

Application of OLS to equation (2.33) gives

$$\hat{b} = (X'X)^{-1} X' y_1^* \qquad \ldots (2.34)$$

and the sum of squared errors is

$$S_2 = (y_1^* - Xb)'(y_1^* - Xb)$$

$$= y_1^{*'} y_1^* - y_1^{*'} X (X'X)^{-1} X' y_1^* \qquad \ldots (2.35)$$

Using equations (2.32) and (2.35) we can obtain the ratio between the two sums as

$$l = \frac{S_1}{S_2} = \frac{y_1^{*'} y_1^* - y_1^{*'} X_1 (X_1'X_1)^{-1} X_1' y_1^*}{y_1^{*'} y_1^* - y_1^{*'} X (X'X)^{-1} X' y_1^*} \qquad \ldots (2.36)$$

The ratio $l$ cannot be less than one simply because that the explanatory power of X cannot be less than $X_1$ and therefore, the variance $S_1$ cannot be less than the variance $S_2$.

To estimate the elements of $a_1$, let us suppose that

$$y_1^* = Y_{1\Delta} a_{1\Delta} \qquad \ldots (2.37)$$

where $Y_{1\Delta} = [Y_1 \quad Y_2 \quad \cdots \quad Y_m]$ *and* $a_{1\Delta} = \begin{bmatrix} 1 \\ a_{12} \\ \vdots \\ a_{1m} \end{bmatrix}$

The ratio $l$ can be written as

$$l = \frac{a_{1\Delta}' W_1^* a_{1\Delta}}{a_{1\Delta}' W_1 a_{1\Delta}} \qquad \ldots (2.38)$$

where,

$$W_1^* = Y_{1\Delta}' Y_{1\Delta} - (Y_{1\Delta}' X_1)(X_1' X_1)^{-1} X_1' Y_{1\Delta}$$

$$W_1 = Y_{1\Delta}' Y_{1\Delta} - (Y_{1\Delta}' X)(X' X)^{-1} X' Y_{1\Delta}$$

The limited information maximum likelihood method is to choose such values of the coefficient that would minimize '$l$'. Thus, on differentiating '$l$' with respect to $a_{1\Delta}$ we have

$$\frac{\partial l}{\partial a_{1\Delta}} = \frac{2(W_1^* a_{1\Delta})(a_{1\Delta}' W_1 a_{1\Delta}) - 2(W_1 a_{1\Delta})(a_{1\Delta}' W_1^* a_{1\Delta})}{(a_{1\Delta}' W_1 a_{1\Delta})^2} \qquad \ldots (2.39)$$

The necessary condition for minimization requires that the first order derivative is equated to zero. Therefore,

$$\frac{2}{(a_{1\Delta}' W_1 a_{1\Delta})}\left[W_1^* a_{1\Delta} - \left(\frac{a_{1\Delta}' W_1^* a_{1\Delta}}{a_{1\Delta}' W_1 a_{1\Delta}}\right) W_1 a_{1\Delta}\right] = 0$$

Using equation (2.38) we get,

$$W_1^* a_{1\Delta} - l W_1 a_{1\Delta} = 0 \qquad \ldots (2.40)$$

Now, for $a_{1\Delta} \neq 0$, the determinant of the matrix formed must be equal to zero, i.e.,

$$\left|W_1^* - l W_1\right| = 0 \qquad \ldots (2.41)$$

The elements of $W_1^*$ and $W_1$ can be obtained from the sample observations (as defined for (2.38). The equation (2.41), therefore, becomes a polynomial of m degree in $l$

which must be solved for the smallest root $\hat{l}$. This may also be considered as an eigen value problem (Krishnamurthy & Sen, pp. 248-249). Substituting the value of $\hat{l}$ thus obtained in equation (2.40) and solving the simplified equation

$$\left(W_1^* - lW_1\right)\hat{a}_{1\Delta} = 0 \qquad \ldots (2.42)$$

the estimator $\hat{a}_{1\Delta}$ can be obtained. Next, the elements of $b_1$ can be obtained by

$$\hat{b}_1 = (X_1'X_1)^{-1}\left(X_1'y_1^*\right) = (X_1'X_1)^{-1}X_1'Y_{1\Delta}\hat{a}_{1\Delta} \qquad \ldots (2.43)$$

Equations (2.42) and (2.43) define the LIML estimates of the structural equation.

The equation-by-equation estimation methods, namely GILS, ILS, 2SLS and LIML discussed so far, are all essentially limited information estimators. In the estimation of any structural equation, these estimators do not take into account complete information on all structural equations in the model. Although the estimators give consistent estimates, they are not asymptotically efficient, as they do not consider the correlation of disturbances across the equation. This deficiency can be overcome by estimating all the equations of the system simultaneously. For this purpose the full information methods such as 3SLS or FIML can be used (Intriligator, pp. 402-416; Johnston, pp. 486-492; Theil, pp. 508-528). Since we have not used these estimators in our investigation, we do not consider it necessary to describe them here.

# CHAPTER 3

# LAD ESTIMATION: A LITERATURE SURVEY

**I. Introduction**: As it has been mentioned before, the Least Squares method of estimation of parameters of linear (regression) models performs well provided that the residuals (disturbances or error) 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 method performs well.

This chapter has two-fold objectives, the first relates to the investigation in the properties of the Least Absolute Deviation (LAD) estimator in estimating regression models and the second relates to the details of the numerical algorithms to compute the regression coefficients by minimizing the sum of absolute disturbances. First, we address to the survey of literature on the study of the properties of the LAD estimator (in single as well as multi-equation models) together with the numerical algorithms to estimate the regression coefficients by this method. Subsequently we deal with the details of different numerical algorithms (linear programming based methods, iterative methods and non-linear search methods) developed for computing the coefficients of a (linear) regression model.

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 system of equations is inherently an inconsistent system

for it cannot have any solution that exactly satisfies all the equations. Therefore, the '*solution*' leaves most of the equations (if not all) unsatisfied by a quantity (of either sign) called the residual or the error terms. 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. Survey of Literature on Properties and Algorithms of LAD Estimator**: 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 (which amounts to minimization of Euclidean Vector 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

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

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 is a lot of mathematical difficulty in working with absolute value functions on account of their lack of amenability to calculus methods.

Econometricians generally take for granted that the error terms implicit in the models are generated by distributions having a finite variance. However, since the time of

Pareto it has been known that the distribution of income bears evidence that the error distribution could have infinite variance. Works of many econometricians, namely, **Meyer & Glauber** (1964), **Fama** (1965) and **Mandlebroth** (1967) on economic data series like prices in financial and commodity markets indicate 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.

An infinite variance means fat tails and fat tails mean a lot of outliers. 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.

The recommendations for using $L_1$ norm may be traced back to **Edgeworth** (1887, 1888, 1923), **Rhodes** (1930) and **Singleton** (1940). Edgeworth and Rhodes pointed out that random sampling and normal distribution, which are needed to justify the method of Least Squares, as an optimal method, often does not exist. 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.

**The Linear Programming based algorithm to LAD Estimation**: With the advent of linear programming as a mathematical method to solve 'corner-optimum' problem that defies

solution by calculus method, 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 pointed out that to fit a linear function

$$\hat{X}_1 = a_0 + a_{12}X_2 + .... + a_{1K}X_K \qquad \ldots (3.1)$$

to the observations

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

($X_{ij}$ 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 (3.1) can be expressed as

$$\begin{aligned} a_0 &= y_1 - Z_1 \\ a_{12} &= y_2 - Z_2 \\ &\ldots\ldots\ldots\ldots \\ a_{1K} &= y_K - Z_K \end{aligned} \qquad \ldots (3.2)$$

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

$$X_{i1} - \hat{X}_{i1} = E_i - V_i ; (i = 1, 2, ..., n) \quad \ldots (3.3)$$

where $E$'s and $V$'s are $2n$ non-negative variables to be determined. Using (3.1), (3.2) and (3.3), the system of n linear equations can be obtained

$$\left.\begin{array}{l} E_1 - V_1 + y_1 - Z_1 + X_{12}y_2 - X_{12}Z_2 + .... + X_{1K}y_K - X_{1K}Z_K = X_{11} \\ E_2 - V_2 + y_1 - Z_1 + X_{22}y_2 - X_{22}Z_2 + .... + X_{2K}Y_K - X_{2K}Z_K = X_{21} \\ ..... \\ E_n - V_n + y_1 - Z_1 + X_{n2}Y_2 - X_{n2}Z_2 + .... + X_{nK}y_K - X_{nK}Z_K = X_{n1} \end{array}\right| \quad \ldots (3.4)$$

Since the problem of finding positive values of E's, $V$'s, $y$'s and $Z$'s satisfying system (3.4) and minimizing the linear form

$$R = \sum_{i=1}^{n} (E_i + V_i) \quad \ldots (3.5)$$

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 (3.1) could be constrained further by adding additional conditions as additional equations to the system of constraints (3.4). If it is desired to specify certain parameters of (3.1) 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 (3.5).

**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 $L_p$ $(1 \leq P \leqq \infty)$ 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** and **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, Robers and Robers and Dantzig. The algorithm is an improved version of the primal algorithm described by Barrodale and Young in 1966. In the improved version, Barrodale and 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.

In the paper, the general $l_1$-linear approximation problem was stated as follows: Let $f(x)$ be given real-valued function defined on a discrete subset $X=\{x_1,x_2,...,x_m\}$ of Euclidean space $E^n$. Given $n(\leq m)$ real-valued function $\phi_j(x)$ defined on $X$, a linear approximating function $L(A,x)=\sum_{j=1}^{n} a_j\phi_j(x)$ was formed for any set $A=\{a_1,a_2,...,a_n\}$ of real numbers. The $l_1$-problem is to determine a best $L(A^*,x)$ which minimizes

$$\sum_{i=1}^{m}|f(x_i)-L(A,x_i)| \qquad \ldots (3.6)$$

The $l_1$-problem (3.6) was restated as a linear programming problem. For the $l_1$-problem it was assumed $\phi_{ji} \equiv \phi_j(x_i), f_i \equiv f(x_i)$, and positive variables $e_i$, $v_i$, $b_j$, $c_j$ were defined by putting $f_i-\sum_{j-1}^{n} a_j\phi_{ji}=e_i-v_i, i=1,2,....,m$ and $a_j=b_j-c_j$ for $j=1,2, ...,n$. Then a best $l_1$-approximation corresponds to an optimal solution to the primal of the linear programming problem:

$$\text{Minimize } \sum_{i=1}^{m}(e_i + v_i)$$

$$\text{subject to } f_i = \sum_{j=1}^{n}(b_j - c_j)\phi_{j,i} + e_i - v_i, i = 1,2,....,m \text{ and } b_j, c_j, e_i, v_i \geq O \qquad \ldots (3.7)$$

It was proved that if the column rank of the *mxn* matrix $\Phi = \{\phi_{ji}\}^T$ is $K(\leq n)$ then their exits a best $l_1$- approximation which interpolates *f(x)* in at least *K* points on *X*.

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 corresponding to (3.7), 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.

An inspection of the linear programming problem reveals that (i) an initial basic feasible solution is immediately available and (ii) only *n* columns are needed to store the information contained in the right-hand side of the equality constraints. Thus denoting the columns of the simplex tableau corresponding to (3.7) by *R*, $b_j$, $c_j, e_i$ and $v_i$, an initial basis is provided by $e_1, e_2, ...., e_m$ whenever each $f_i$ is positive. If an $f_i$ is negative, the sign of the corresponding row is changed and $e_i$ is replaced by $v_i$ in the basis. It is also assumed

that $b_j = -c_j$ and $e_i = -v_i$ , and that the sum of the marginal (or reduced) costs of $b_j$ and $c_j$ is zero and of $e_i$ and $v_i$ is $-2$. Thus the condensed form of the simplex method (in which the basis is suppressed) can be applied to just $n$ columns, which initially contain $b_1, b_2, \ldots, b_n$. Within an array of dimensions $(m+2)x(n+2)$ which initially contains the data in Table-1 (including labels for the basic and nonbasic vectors), the simplex iterations can be performed . The program requires additional arrays totaling $3m+n$ words: these are used as workspace and for storing output information.

**Table-1. Initial condensed simplex tableau for the algorithm (assuming each $f_i$ is nonnegative)**

| Basis | R | $b_1$ | $b_2$ | ... | $b_n$ |
|---|---|---|---|---|---|
| $e_1$ | $f_1$ | $\phi_{1,1}$ | $\phi_{2,1}$ | ... | $\phi_{n,1}$ |
| $e_2$ | $f_2$ | $\phi_{1,2}$ | $\phi_{2,2}$ | ... | $\phi_{n,2}$ |
| . | . | . | . | ... | . |
| $e_m$ | $f_m$ | $\phi_{1,m}$ | $\phi_{2,m}$ | ... | $\phi_{n,m}$ |
| Marginal Cost | $\sum_{i=1}^{m} f_i$ | $\sum_{i=1}^{m} \phi_{1,i}$ | $\sum_{i=1}^{m} \phi_{2,i}$ | ... | $\sum_{i=1}^{m} \phi_{n,i}$ |

The algorithm is implemented in two stages. In Stage-1 the choice of pivotal column is restricted to the vectors $b_j$ and $c_j$ during the first $n$ iterations. The vector with the largest marginal cost is chosen to enter the basis. From among the basic vectors $e_i$ and $v_i$ the vector which causes the maximum reduction in the objective function is chosen to leave the basis. At the end of Stage-1, the rank $K(\leq n)$ of the matrix $\Phi$ is determined by the total number of vectors $b_j$, $c_j$ in the basis. Since $K$ of the vectors $e_i$ (or $v_i$) have been removed from the basis, the current simplex tableau represents an approximation, which interpolates at least $K$ data points. If the approximation interpolates more than $K$ data

points, then the simplex tableau is degenerate: this does not cause any problems in practice.

In Stage-2 non-basic $e_i$ or $v_i$ are interchanged with basic $e_i$ or $v_i$: the basic $b_j$ and $c_j$ vectors are not allowed to leave the basis during this stage. The vectors entering and leaving the basis are chosen in a similar manner as in Stage-1. The algorithm terminates when all the marginal costs are negative. In Stage-2 each simplex tableau corresponds to an approximation, which interpolates $K$ data points (assuming nondegeneracy). $K$-$1$ of these points remains fixed in each iteration. The vector entering the basis determines which point is to be dropped from the interpolating set while the vector leaving the basis determines the new point of interpolation.

Since some of the basic vectors $b_j$ and $c_j$ can have negative values associated with them the final tableau at the end of Stage-2 may be infeasible. The solution can be made feasible and optimal by interchanging such basic vectors $b_j$ (or $c_j$) with the corresponding nonbasic vectors $c_j$ (or $b_j$).

The main modification of the simplex method is in choosing the vector $e_i$ or $v_i$ to leave the basis. The vector that causes the maximum reduction in the objective function is chosen in both the stages.

The study conducted by **Oveson** (1968) on the LAD estimator gave a new thrust to the investigation into the properties and applicability of the estimator. It had also been almost well 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 call 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.

**Indirect or Iterative Algorithms**: **Spyropoulos, Kiountouzis & Young** (1973) and **Abdelmalek** (1974) made a progress in this direction. **Schlossmacher** (1973) and **Fair** (1974) also proposed an improved algorithm for $L_1$ estimation that is very similar to iterative weighted least squares. 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.

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

$$f(e_i) = (2\lambda)^{-1} \exp\left\{-\frac{|e_i|}{\lambda}\right\} \qquad \ldots (3.8)$$

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 (3.8) was confirmed in a Monte Carlo study by **Smith** and **Hall** (1972).

**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) $\underset{n \to \infty}{Lim}\, n^{-1}X'X = Q$ 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 LAE (alias LAD) 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 LAE 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 $\underset{n \to \infty}{Lim}\, n^{-1}X'X = Q$ 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 LAE estimator was stated and proved. It was proved that the LAE 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 LAE 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 LAE's median-type robustness and insensitivity to outlying observations.

**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.

**Pollard** (1991) presented an alternative approach for studying the asymptotic theory of Least Absolute Deviation (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.

**Phillips** (1991) presented the asymptotic theory for the LAD estimator in a regression model setting 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.

**Weiss** (1991) established that it was possible to use the Least Absolute Deviation (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

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

$z_t$ = 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 $\{x_{t-i}\}(i \geq 0)$ and $\{e_{t-i}\}(i \geq 1)$.

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.

Combining recent advances in interior point methods for solving linear programs with a new statistical preprocessing approach for $l_1$-type problems, **Portnoy and 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.

**Breidt, Davis and 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.

**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 and 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.

**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.

**Estimation of Multi-equations Models**: 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 **H Theil** (1953) used repeated least squares to estimation of parameters of a regression equation in the multi-equation model. **RL Basmann** (1957) used least squares repeatedly for estimating parameters of a multi-equation linear econometric model. **H Theil** (1961) developed the method of Two-Stage Least Squares (2-SLS). Very soon **A Zellner** and **H Theil** (1962) developed Three-Stage Least Squares (3-SLS) also.

It would be befitting to describe the nature and the issues related with the estimation of multi-equation (linear) models once again (we have already described them in chapter-2). We are of the opinion that recapitulation 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 ( $Y = \Upsilon + \varepsilon$ , where $\Upsilon$ 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 ($X_j = \chi_j + \nu_j$). 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 $Y = \Upsilon + \varepsilon + E$. It is assumed that $\nu_j$ (in the pre-determined variables comprising X) and $E_j$ are orthogonal (linearly independent).

Our objective is to estimate A and B. First, we simplify our 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, our 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 $\hat{b}_{ir} \in \hat{b}_r$ (or both). The most straightforward way to do that is to set some $\mu_r$ unknowns ($\hat{a}_{sr} \in \hat{a}_r$ or $\hat{b}_{ir} \in \hat{b}_r$ 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 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,

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

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

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 (**Krishnamurthy & Sen**, pp. 153-185). 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 (**Krishnamurthy & Sen**, pp. 191-193). 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 $k_2 \times m_1$ 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* (2-SLS). From the procedure and the conditions governing its application it is clear that 2-SLS 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 2-SLS 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 2-SLS. **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 2-SLS, 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 $\hat{b}_1 = -(P_{11}\hat{a}_1 + P_{12}a_2)$. This is $GILN_2$ (alias GILS or Generalized Indirect Least Euclidean-Norm) estimator because ILS (Indirect Least Squares) estimator, which is applicable only

in case of an exactly identified structural equation, is defined as c and $\hat{b}_1 = -(I'_{11}\hat{a}_1 + I'_{12}a_2)$.

If $P_{21}$ is a square matrix of full rank, $P_{21}^{-g} = P_{21}^{-1}$. Therefore, $GILN_2$ is applicable in case of any structural equation exactly or over identified. In Khazzoom estimator $P = (X'X)^{-1}X'Y$ is the matrix of Reduced Form coefficients estimated by OLS.

**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. Our survey suggests that 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 were perhaps the first to extend $L_1$ estimation to multi-equation models, compare 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

$$Y_1 + A_{12}Y_2 + B_{11}Z_1 + B_{12}Z_2 + B_{10} = E_1 \qquad \ldots (3.9)$$

$$Y_1 + A_{22}Y_2 + B_{23}Z_3 + B_{24}Z_4 + B_{20} = E_2 \qquad \ldots (3.10)$$

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 $\sigma_u^2 = (\sigma_u + i)^2$ where $\sigma_u = 5$ 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 $T_{\hat{\theta}^K} = \left(\hat{\theta}^K - \theta^*\right) / \hat{\sigma}_{\hat{\theta}^K}$, 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 $(S_{q1})$ 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 $\underset{n \to \infty}{Lim}\, n^{-1}X'X$ 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 $\tilde{a}$ that minimizes $S = \sum_{i=1}^{n} |Y_i - X_i' \tilde{a}| - \sum_{i=1}^{n} |E_i|$, where $X_i'$ 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 $\tilde{a}$. 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 $\tilde{a}$ 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 $\sum_{i=1}^{n} \left| Y_i - X_i' \hat{\pi} A - X_{1i}' \tilde{a} \right|$, 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 $q \to 0$. It appears that no study has yet been carried out to implement the suggestions made by Amemiya.

We have already described the work of **Khazzoom** (1976) who generalized Indirect Least Squares estimator for (exactly or over-) identified equations. From this, it may follow 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) or absolute. It appears that this possibility is hitherto unexplored.

**Fair** (1994) estimated the US model by 2SLS, 2SLAD, 3SLS and FIML. 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.

**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.

**III. Details of different Computational Algorithms for LAD Estimation:** Now we provide the details of the numerical algorithms mentioned in the preceding section of this chapter. The Numerical algorithms for the purpose may broadly be classified into three heads: (i) those based on LP formulation, (ii) those based on iterative schemes - using the fixed point theorem which says that a continuous and bounded function onto itself has at least one point on which x = f(x) and (iii) those based on the multi-variate search methods of non-linear programming.

To begin, let us have a closer look into the problem of minimum absolute deviation estimation of regression models: Let the linear regression equation to be estimated be

$$y = Xa + e \qquad \ldots (3.11)$$

The Least Absolute Deviation (LAD) estimator is given by that value of $\tilde{a}$, which minimizes

$$S = |e_i| = \sum_{i=1}^{n} \left| y_i - \sum_{j=0}^{m} \tilde{a}_j x_{ij} \right| \qquad \ldots (3.12)$$

This function is defined in a similar way to the least squares estimator. But, unlike the least squares estimator that minimizes the sum of the squared residuals, the LAD estimator minimizes the sum of their absolute values of the residuals. Consequently, relative to the least squares estimator, it is less influenced by outliers and in repeated sampling from a distribution where outliers are prevalent, its sampling variability is likely to be less than that of the LS estimator.

Since equation (3.12) represents an absolute value function, it is not differentiable at every point. Hence, differentiation, which is conventionally used for minimization in the least squares method, cannot be applied here. In fact, the minimum point of S is a vertex and hence non-differentiable. The problem, therefore, has to be solved by devices other than calculus.

To probe into this aspect further, let us take the simple linear regression model with one independent variable and an intercept, namely,

$$Y_i = a_0 + a_1 X_i + e_i, \qquad \text{i=1, 2,...,n} \qquad \ldots (3.13)$$

Now, we need to find $\hat{a}_0$ and $\hat{a}_1$ so that

$$S = \sum \left| Y_i - \hat{Y}_i \right| \qquad \ldots (3.14)$$

is minimized. Here, $\hat{Y}_i = \hat{a}_0 + \hat{a}_1 X_i$.

Suppose we choose $\hat{a}_0$ such that the regression line includes the point of means. This implies that the origin is translated to the point $(\bar{Y}, \bar{X})$. Then by changing the co-ordinates

$$y_i = Y_i - \bar{Y}$$

and $x_i = X_i - \bar{X}$

equations (3.13) and (3.14) can be written as

$$y_i = a_1 x_i + e_i \qquad \ldots (3.15)$$

(since $\hat{a}$ = 0 in the new co-ordinate system)

and $$S = \sum |y_i - \hat{y}_i| = \sum |y_i - \hat{a}_1 x_i| \qquad \ldots (3.16)$$

An element of S is represented as

$$S_i = |y_i - \hat{a}_1 x_i| \qquad \ldots (3.17)$$

This is a broken line ( Y = $|X|$ ) in the (s, $a_1$) plane and is composed of two half- lines with a minimum equal to zero at

$$\hat{a}_i = \frac{y_i}{x_i} \quad \text{(obtained by equating } S_i \text{ to zero).} \qquad \ldots (3.18)$$

It has been found that $S_i$ is always convex upwards since the slope of the half-line to the left of $\hat{a}_i$ is equal to - $|x_i|$ and to the right of $\hat{a}_i$ is equal to $|x_i|$. As a result, $S = \sum S_i$ considered in the (s, $a_1$) plane, being the sum of connected half- lines will consist of connected line segments. The slope of S at any point $a_1$ will be the aggregate of the slopes of $S_i$ at that value of $a_1$. Now, since the slopes change only at the minimum

points of $S_i$, the slope of S also can change only at these points. S is also convex since each $S_i$ is convex.

It can be inferred from the above that only points corresponding to the minimum points of the individual $S_i$ are required to be considered in minimizing S which are finite in number and have value equal to $y_i/x_i$. Hence, the regression line will pass through the point of means and also the observation corresponding to the minimizing i. Therefore, determination of the regression line is by the point of means and the observation associated with the minimizing $\hat{a}_i$.

Let us now consider the case where the regression line does not pass through the point of means. Here the estimation problem is to find $\hat{a}_0$ and $\hat{a}_1$ such that

$$S = \sum |Y_i - \hat{a}_0 - \hat{a}_1 X_i| \qquad \ldots (3.19)$$

is minimized. A typical element of S may be represented as

$$S_i = \sum |Y_i - \hat{a}_0 - \hat{a}_1 X_i| \qquad \ldots (3.20)$$

It is composed of two half –planes in (S, $a_0, a_1$) space that intersect in the ($a_0, a_1$) plane (in the three dimension case).

Thus, a family of ($a_0, a_1$) corresponding to the minimum of $S_i$ is found to exist and is given by the equation

$$Y_i - a_0 - a_1 X_i = 0 \qquad \ldots (3.21)$$

S being a sum of half-planes, its surface is composed of connected plane segments. It is convex upwards because the $S_i$ are convex. S, therefore, will have a minimum unique region, which will be a point if the $S_i$ which intersects at the ($a_0, a_1$) plane are not parallel to one another, or a line if the family ($a_0, a_1$) obtained on mininimizing the $S_i$ are the same

for each $S_i$ or a closed polygon if different families of $(a_0, a_1)$ are obtained for different $S_i$. A point represents a unique solution to the estimation problem, whereas a line or closed polygon represents multiple solutions.

Since the direction cosines of the constituent half-planes of the $S_i$ change only at their lines of intersection, the boundaries of the plane segments that compose S will be directly above the set of lines

$$Y_i - a_0 - a_1 X_i = 0, \qquad i=1,2,\ldots,n \qquad \ldots (3.22)$$

in the $(a_0, a_1)$ plane. A unique solution to the problem must be above one of the points of intersection of the lines in equation (3.22). On the other hand, a non-unique solution will lie either on a closed interval of a line in equation (3.22) constituted by the intersection of that line with two other parallel lines, or on a closed polygon formed by three or more of the lines.

The three dimensional case discussed so far can be extended to n independent variables. Here each $S_i$ is composed of two n-dimensional half-hyperplanes in the $(S, \tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n)$ space intersecting in the $(\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n)$ hyperplane and is convex upward. Since $S = \sum S_i$, therefore, S will form a polygonal surface whose edges will lie above the (n-1) dimensional hyperplanes which are the result of the intersections of the half-hyperplanes of S with the $(\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n)$ hyperplane. S is convex since the $S_i$ are convex. When the solution to the estimation problem is unique, the minimum of $S_i$ will lie above the intersection of n of these (n-1) dimensional hyperplanes. These n observations determine the regression hyperplane.

In the cases with larger dimensions it may be impossible to find a solution to the estimating problem due to the manifold increase in the complexity of the calculations.

Therefore, in order to find an optimal solution, the problem can be converted to a linear programming problem and a solution sought.

**LAD Estimation as a Linear Programming Problem**: Linear Programming (developed in the late 1940's) immediately attracted the attention of econometricians seeking optimization of a linear function with corner solution at a vertex. Therefore, LAD estimation was originally converted to an equivalent linear programming problem.

In matrix form, the linear programming problem is to minimize

$$A = P'Z \qquad \ldots (3.23)$$

subject to the constraints

$$BZ \geq a_1 \qquad \ldots (3.24)$$

$$Z \geq 0 \qquad \ldots (3.25)$$

The function that is being minimized in equation (3.12) is first converted into an equivalent linear form by writing $e_i$ as the difference between two non- negative variables.

$$e_i = v_i - w_i, \qquad v_i, w_i \geq 0 \qquad \ldots (3.26)$$

and then minimizing the quantity

$$S^* = \sum v_i + \sum w_i \qquad \ldots (3.27)$$

Although S* of equation (3.27) and S of equation (3.12) are different functionals, the values of v and w that minimize S* give the value of e that minimizes S. This implies that equation (3.27) is equivalent to equation (3.12). In the optimal solution also S* and S are equal.

The constraints in equation (3.24), which are given by n observations, can be re-written in the form of equation (3.28)

$$\begin{bmatrix} x_{11} & -x_{11} & x_{12} & -x_{12} & \cdots & x_{1K} & -x_{1K} & 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ x_{21} & -x_{21} & x_{22} & -x_{22} & \cdots & x_{2K} & -x_{2K} & 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & -x_{n1} & x_{n2} & -x_{n2} & \cdots & x_{nK} & -x_{nK} & 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \times \begin{bmatrix} \tilde{a}_1 \\ \tilde{a}_1^* \\ \tilde{a}_2 \\ \tilde{a}_2^* \\ \vdots \\ \tilde{a}_K \\ \tilde{a}_K^* \\ v_1 \\ w_1 \\ v_2 \\ w_2 \\ \vdots \\ v_n \\ w_n \end{bmatrix}$$

$$= \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \ldots(3.28)$$

Here e has been replaced by v – w and the variable x appears twice, once positively and once negatively. $\tilde{a}_1^*, \tilde{a}_2^*, \ldots, \tilde{a}_K^*$ are the coefficients of the K new x variables. The independent variables x and -x correspond to the matrix B in equation (3.24), the regression coefficients $\tilde{a}_1, \tilde{a}_1^*, \tilde{a}_2, \tilde{a}_2^*, \ldots, \tilde{a}_K, \tilde{a}_K^*$ (which are non-negative) to the vector Z and y to the vector of constraints $a_1$. The residual term $v_1, w_1, v_2, w_2, \ldots, v_n, w_n$ represent the introduction of 2n "slack" activities. The restriction in equation (3.25) corresponds to the vector of non-slack activities level.

In the problem there are altogether 2(K+n) activities. Succeeding pairs of activities, namely, $(\tilde{a}_j, \tilde{a}_j^*)$ and $(v_i, w_i)$ of the matrix Z in equation (3.28) are dependent

and therefore in any solution, only one member of each pair for each i and j can be non-zero. This is necessary that the activities in any basis must be independent. Because only one of the regression coefficient can be non- zero, the negative of each independent variable has been included to allow the coefficients of the independent variable to be of either sign. The values taken by S* in equation (3.27) and S in equation (3.12) will be equal since both $v_i$ and $w_i$ cannot be non-zero. Since the activities corresponding to the independent variables have zero "cost" in the minimand, the independent variables will always be present in the optimal solution. This implies that the optimal solution will contain n-K residuals (i.e., slack activities) and K residuals will therefore be zero. This corresponds to what was earlier observed that the regression hyperplane is determined by a subset of n observations.

The above formulation of the problem will always yield a solution if degeneracy is not considered. However, if the number of observations is large, the computational procedure will be cumbersome and lengthy. Wagner has shown that by considering the dual of the above formulation and solving it as a problem in bounded variables using special simplex procedures developed by Charnes and Lemke (1954), Dantzig (1955) and Wagner (1958), the number of activities and constraints can be substantially reduced.

Although linear programming is a computational method that can be used in $L_1$ estimation, the method is known to be computationally unwieldy. There are iterative methods as well as multivariate search methods that can be used to minimize S and hence determine that value of $\tilde{a}$ when S is minimum.

**The Iterative Algorithm**: **Schlossmacher** (1973) and **Fair** (1974) suggested an iterative method in the case of a linear model, $Y = Xa + e$. In this method, the steps for computing are as follows:

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

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.

While using this algorithm, two points should be taken care of. First, that X includes the unitary vector often designed for taking care of the constant term in the model and second, that in case $|e_i|$ (for any i ) is very small (smaller than say, 0.00001) than that $e_i$ is set equal to a small value (say, 0.00001 as done by Fair or completely ignored i.e. set equal to zero as done by Schlossmacher). 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. Though 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.

**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.

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.

$$\sum_{i=1}^{n} w_i\left(y_i - x_i'a\right)^2 \qquad \ldots (3.29)$$

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 $\tilde{a}$

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

Step III: Using weighted least squares $\hat{a} = (XWX)^{-1}XWy$

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

Step IV: Repeat the process until $\hat{a}$ converges.

**LAD Estimation as Multivariate Search Problem**: Multivariable search problems start from some base point and move towards the optimum based on sequential improvement in the value of the objective function. The step size may be fixed or accelerated and decelerated subject to a set of rules. There are several procedures based on this technique.

**(i) Fletcher-Powell Algorithm:** The method requires several alternatives starting points as multi-modal functions are suspected. To minimize: $F(a_1, a_2, \ldots, a_m)$ the steps are as follows:

Step I: Select a starting point.

Step II: Determine the search direction. In normalized form this can be written as

$$M_i^{(k)} = \left\{ \frac{-\sum_{j=1}^{N} H_{ij}\left(\frac{\delta F}{\delta a_j}\right)}{\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{N} H_{ij}\left(\frac{\delta F}{\delta a_j}\right)\right)^2\right]^{\frac{1}{2}}} \right\}^k, \text{ where, } i = 1, 2, \ldots, m$$

here , k= iteration index (starting point k=0),

$M_i$ = direction vector components,.

$\delta F/\delta a_j$ = slope vector components,

$H_{ij}$ = elements of a symmetric positive definite matrix (mxm), (initially an identity matrix).

The initial direction therefore, is the path of steepest descent.

Step III: Conduct a one-dimensional search in the direction chosen in step II in the following manner until a minimum is located utilizing the relation $a_{i(new)} = a_{i(old)} + SM_i; i = 1,2,...,m$ where S is the step size in the direction of search.

(1) evaluate the objective function first.

(2) then, incrementing a distance $\Delta a$ to the independent variable, evaluate the objective function again. If a function improvement is obtained then double the step size for the next function evaluation. But, if a function improvement is not obtained in the first step, reverse the direction by locating the next point at a distance - $\Delta a$ from the starting point.

(3) after the first step, double the step size if a function improvement is obtained and halve it if a worse function evaluation is obtained.

(4) when a local optimum is encountered, the procedure will yield three points $(a_k, a_{k-1}, a_{k-2})$ straddling the optimum. Locate an additional point $a_{k+1}$ such that $a_{k+1} = a_{k-1} + \Delta a/2$ which is the minimum in the direction of search, where $\Delta a$ is the current step size. Then retain the best three points (call them $a_1$, $a_2$, $a_3$).

(5) then a quadratic equation, $\phi$, is curve-fitted to the three retained points. The optimum location, $a^*$, located by setting $\delta\phi/\delta a = 0$, is

$$a^* = \frac{1}{2}\left[\frac{(a_2^2 - a_3^2)F(a_1) + (a_3^2 - a_1^2)F(a_2) + (a_1^2 - a_2^2)F(a_3)}{(a_2 - a_3)F(a_1) + (a_3 - a_1)F(a_2) + (a_1 - a_2)F(a_3)}\right]$$

(6) then compare the objective function at $a^*$ with the best previous point subject to a convergence limit,

$$\left|a^* - a_i\,(best)\right| \le \lim$$

If the above criterion is satisfied, the procedure stops. If not, worst point is replaced by $a^*$ and a new quadratic surface is fitted and local optimum is obtained.

Step IV: Make a convergence check. If convergence is achieved, terminate the procedure. If convergence is not achieved, choose a new search direction as per step II. Calculate $H^{(k+1)}$ as follows:

$$H^{(k+1)} = H^{(k)} + A^{(k)} - B^{(k)}$$

where,

$$A^{(k)} = \frac{\Delta a^{(k)}\left(\Delta a^{(k)}\right)^t}{\left(\Delta a^{(k)}\right)^t\left(\Delta G^{(k)}\right)}$$

$$B^{(k)} = \frac{H^{(k)}\Delta G^{(k)}\left(\Delta G^{(k)}\right)^t H^{(k)}}{\left(\Delta G^{(k)}\right)^t H^{(k)}\Delta G^{(k)}}$$

$\Delta a^{(k)} = a^{(k+1)} - a^{(k)}$ (difference in location between iterations)

$\Delta G^{(k)} = \frac{\delta F^{(k+1)}}{\delta a} - \frac{\delta F^{(k)}}{\delta a}$ (difference in slopes between iterations)

Perform a new one dimensional search in the new direction. Repeat the process until convergence is obtained.

Although the above algorithm can effectively locate the minimum, the fact that this method uses derivates, would possibly be its drawback. Since LAD is essentially an absolute value function, it is not differentiable exactly at its minimum point. The minimum of an absolute value function is located in a vertex of the simplex.

Therefore, derivative-free search methods of finding the minimal value of S and subsequently the vector $\tilde{a}$ are recommended. Such derivative-free search methods have been developed by R. Hooke, TA Jeeves, HH Rosenbrock and others. Two such algorithms which can effectively be used to minimize the absolute value function and are also relatively easy to program on a computer are discussed below.

**(ii). Nelder-Meade Algorithm**: JA Nelder and R Meade developed an algorithm using a regular geometric figure called a simplex, consisting of m+1 vertices, to find the minimum of a multivariable unconstrained non-linear function (**Kuester & Mize**, 1973). This simplex method utilizes reflected, expanded and contracted points to locate the minimum.

To minimize the function: $F(a_1, a_2, ..., a_m)$, the steps are as follows:

Step I: First, pick a starting point, say $\hat{a}_1$

Step II: Next, construct a simplex that consists of the starting point and the following additional points as:

$$\hat{a}_j = \hat{a}_j + \xi_j \ , \qquad j = 2,3,4,...,m+1$$

where, $\xi_j$ is determined from the following table:

| $j$ | $\xi_{1,j}$ | $\xi_{2,j}$ | $\cdots$ | | $\xi_{m-1,j}$ | $\xi_{m,j}$ |
|---|---|---|---|---|---|---|
| 2 | $p$ | $q$ | $\cdots$ | $\cdots$ | $q$ | $q$ |
| 3 | $q$ | $p$ | $\cdots$ | $\cdots$ | $q$ | $q$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | | $\vdots$ | $\vdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | | $\vdots$ | $\vdots$ |
| $m$ | $q$ | $q$ | $\cdots$ | $\cdots$ | $p$ | $q$ |
| $m+1$ | $q$ | $q$ | $\cdots$ | $\cdots$ | $q$ | $p$ |

where, $p = \frac{\sigma}{m\sqrt{2}}\left((m+1)^{\frac{1}{2}} + m - 1\right)$ ; $q = \frac{\sigma}{N\sqrt{2}}\left((m+1)^{\frac{1}{2}} - 1\right)$; m=total number of variables and $\sigma$ = side length of simplex.

Step III: Calculation of the Centroid.

Once the simplex is formed, the objective function is evaluated at each point. The worst point (highest value of objective function) is replaced by a new point. Three operations are used – reflection, contraction and expansion.

A reflection point is first located thus:

$a_{i,j}$ (reflected) = $\bar{a}_{i,c} + \alpha\left(\bar{a}_{i,c} - a_{i,j}(worst)\right)$ ; i = 1,2,...,m

where, $\alpha$ is a positive constant, $\bar{a}_{i,c}$ are the centroid co-ordinates of all points excluding the worst point. It is calculated as

$$\bar{a}_{i,c} = \frac{1}{K-1}\left[\sum_{j=1}^{K} a_{i,j} - a_{i,j}(worst)\right], \quad \text{i= 1,2,...,m}$$

Set K = m+1.

Step IV: Check Reflected Point:

If the reflected point has the worst objective function value of the current points, a contracted point is located as follows:

$a_{i,j}$ (contracted)= $\bar{a}_{i,c} - \beta\ \left(\bar{a}_{i,c} - a_{i,j}(worst)\right)$, i = 1,2,...,m and $\beta$ lies between 0 and 1.

If the reflected point, though better than the worst point, is still not the best point, a contracted point is calculated from the reflected point as follows.

$a_{i,j}(contracted) = \bar{a}_{i,c} - \beta\left(\bar{a}_{i,c} - a_{i,j}(reflected)\right)$, i = 1,2, ...,m.

The objective function is now evaluated at the contracted point. If an improvement over the current points is achieved, the process is restarted. If an improvement is not achieved, the points are moved one-half the distance towards the best points:

$$a_{i,j}\ (\text{new}) = \left(a_{i,j}(best) + a_{i,j}(old)\right)/2, \qquad i = 1,2, \ldots,m.$$

The process is then restarted.

Step V: Calculate Expansion Point: If the reflected point calculated in step III is the best point, an expansion point is calculated as follows:

$$a_{i,j}\ (\text{expansion}) = \bar{a}_{i,c} + \gamma\left(a_{i,j}(reflected) - \bar{a}_{i,c}\right), \qquad i = 1,2,\ldots,m$$

where, $\gamma$ is a positive constant. If the expansion point is an improvement over the reflected point, the expansion point replaces the reflected point and the process restarted. If the expansion point is not an improvement over the reflected point, the reflected point is retained and the process restarted.

Step VI: The procedure is terminated when the convergence criterion is satisfied or a specified number of iterations has been exceeded.

**(iii) Hooke-Jeeves Algorithm**: The algorithm developed by R Hooke and TA Jeeves is based on the direct search method (**Kuester & Mize**, 1973). The method assumes a unimodal function and is derivative free. If more than one minimum exists or the shape of the surface of the simplex is unknown, several sets of starting points are required. The steps involved are:

Step I: Select a base point and evaluate the objective function.

Step II: Make local searches in each direction by stepping $a_i$ a distance $S_i$ to each side and evaluate the objective function to ascertain whether a lower function value can be obtained.

Step III: If there is no improvement i.e., the function does not decrease, reduce the step size and make searches from the previous best point.

Step IV: If the value of the objective function has decreased, locate a " temporary head" $a_{i,0}^{(k+1)}$ using the two previous base points $a_i^{(k+1)}$ and $a_i^{(k)}$ such that

$$a_{i,0}^{(k+1)} = a_i^{(k+1)} + \beta\left(a_i^{(k+1)} - a_i^{(k)}\right)$$

where, the variable index i varies from 1 to m; 0 denotes the temporary head; k is stage index ( a stage is the end of m searches) and $\beta \geq 1$.

Step V: If the temporary head gives a lower function value, make a new local search for the temporary head and locate a new head. Check the value of the function F. Continue this method of expansion as long as the function value keeps decreasing.

Step VI: If the temporary head does not result in a lower function value, make a search from the previous best point.

Step VII: Continue the procedure until the convergence criterion is satisfied. Once convergence is achieved, the procedure is terminated.

**Random Walk Method**: This method is based on generating a sequence of improved approximations to the minimum, each derived from the preceding approximation (**Rao**, 1978, pp. 252-254). 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.

**Random Walk Method with Direction Exploitation**: 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 (**Rao**, 1978, pp. 255-256). 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

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

Random walk methods can work even if the objective function F is discontinuous or non-differentiable at some of the points. This property makes this method quite suitable for minimizing the absolute norm of residuals.

# CHAPTER 4

# THE METHODOLOGY OF MONTE CARLO EXPERIMENTS

**I. Introduction**: There are two approaches to studying the properties of a mathematical system (representing a real-world system): the first, *analytical* and the second, *experimental*. In the analytical approach, the mathematical system is described in great details (so as to represent the real-world system as accurately as possible). All the axioms and assumptions underlying the system are exhaustively and explicitly described with utmost accuracy. All definitions relevant to the system are accurately and explicitly stated. All the operations that the mathematical system would be subjected to are accurately and exhaustively described in their details. Then, appropriate mathematical methods are applied to draw conclusions from the axioms, assumptions and statements that describe the said mathematical system. All the conclusions drawn in this process are the implications of the mathematical system so described. Nevertheless, mathematical methods applied to draw conclusions have some impact on the conclusions. A careful attempt is made, therefore, to apply mathematical methods such that they modify the conclusions as little as possible. These conclusions are deductive in nature and have general applicability.

However, complicated systems cannot be described in very accurate and exhaustive details. Even if the system is described in details, appropriate methods to draw conclusions may be either unavailable or intractable due mainly to the complexity of the system. To investigate into the properties of such a mathematical system by analytical methods is very difficult, if not impossible. Experimental methods are applied to investigate into the properties of such systems. The conclusions drawn from experimental

methods provide a great insight into the nature of the system. However, these conclusions have lesser degree of certainty as well as applicability. These conclusions are inductive inferences – generalizations based on the study of samples – and, therefore, probabilistic in nature.

**II. The Monte Carlo Method**: The Monte Carlo method of investigating into the properties of a mathematical system is basically experimental and numerical in nature. Numerical methods that are known as 'Monte Carlo methods' can be loosely described as statistical simulation methods, where statistical simulation is defined in quite general terms to be any method that utilizes sequences of random numbers to perform the simulation. Monte Carlo methods have been used for centuries, but only in the past several decades has the technique gained the status of a full-fledged numerical method capable of addressing the most complex applications. The credit for inventing the Monte Carlo method often goes to Stanislaw Ulam, a Polish born mathematician who worked for John von Neumann on the United States' Manhattan Project during World War II. Ulam is primarily known for designing the hydrogen bomb with Edward Teller in 1951. Quoted in **Eckhardt** (1987), Ulam describes the incident as:

> The first thoughts and attempts I made to practice [the Monte Carlo Method] were suggested by a question which occurred to me in 1946 as I was convalescing from an illness and playing solitaires. The question was what are the chances that a Canfield solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays. This was already possible to envisage with the beginning of the new era of fast computers, and I immediately thought of problems of neutron diffusion and other questions of mathematical physics, and more generally how to change processes described by certain differential equations into an equivalent form interpretable as a

succession of random operations. Later ... [in 1946, I] described the idea to John von Neumann, and we began to plan actual calculations.

Working with John von Neumann and Nicholas Metropolis, he developed algorithms for computer implementations, as well as exploring means of transforming non-random problems into random forms that would facilitate their solution via statistical sampling. This work transformed statistical sampling from a mathematical curiosity to a formal methodology applicable to a wide variety of problems. It was Metropolis who named the new methodology after the casinos of Monte Carlo. Ulam and Metropolis published their paper on Monte Carlo method in 1949.

Monte Carlo method is now used routinely in many diverse fields, from the simulation of complex physical phenomena such as radiation transport in the earth's atmosphere and the simulation of the esoteric sub-nuclear processes in high energy physics experiments, to the mundane, such as the simulation of simple games. The analogy of Monte Carlo methods to games of chance is a good one, but the *game* is a physical system, and the outcome of the game is not a pot of money or stack of chips (unless simulated) but rather a solution to some problem. The *winner* is the scientist, who judges the value of his results on their intrinsic worth, rather than the extrinsic worth of his holdings.

Statistical simulation methods may be contrasted to conventional numerical discretization methods, which typically are applied to ordinary or partial differential equations that describe some underlying physical or mathematical system. In many applications of Monte Carlo, the physical process is simulated directly, and there is no need to even write down the differential equations that describe the behavior of the system. The only requirement is that the physical (or mathematical) system be described

by probability density functions (pdf's). Assuming that the evolution of the physical system can be described by probability density functions, then the Monte Carlo simulation can proceed by sampling from these pdf's, which necessitates a fast and effective way to generate random numbers uniformly distributed on the interval [0,1]. The outcomes of these random samplings, or trials, must be accumulated or tallied in an appropriate manner to produce the desired result, but the essential characteristic of Monte Carlo is the use of random sampling techniques (and perhaps other algebra to manipulate the outcomes) to arrive at a solution of the physical problem. In contrast, a conventional numerical solution approach would start with the mathematical model of the physical system, discretizing the differential equations and then solving a set of algebraic equations for the unknown state of the system.

Nevertheless, this general description of Monte Carlo methods may not directly apply to some applications. It is natural to think that Monte Carlo methods are used to simulate random, or stochastic, processes, since these can be described by pdf's. However, this coupling is actually too restrictive because many Monte Carlo applications have no apparent stochastic content, such as the evaluation of a definite integral or the solution of a system of linear equations. However, in these cases and others, one can pose the desired solution in terms of pdf's, and while this transformation may seem artificial, this step allows the system to be *treated* as a stochastic process for the purpose of simulation and hence Monte Carlo methods can be applied to simulate the system. Therefore, one takes a broad view of the definition of Monte Carlo methods and includes in the Monte Carlo rubric all methods that involve statistical simulation of some underlying system, whether or not the system represents a real physical process. The range of applications that have been addressed with statistical simulation techniques is enormous, from the simulation of

galactic formation to quantum chromodynamics to the solution of systems of linear equations.

**Major Components of a Monte Carlo Algorithm**: Given the working definition of Monte Carlo method, let us now describe briefly the major components of a Monte Carlo method. These components comprise the foundation of most Monte Carlo applications. The primary components of a Monte Carlo simulation method include the following:

*Probability distribution functions (pdf's)* - the physical (or mathematical) system must be described by a set of pdf's.

*Random number generator* - a source of random numbers uniformly distributed on the unit interval must be available.

*Sampling rule* - a prescription for sampling from the specified pdf's, assuming the availability of random numbers on the unit interval, must be given.

*Scoring (or tallying)* - the outcomes must be accumulated into overall tallies or scores for the quantities of interest.

Additionally, an estimate of the statistical error (variance) as a function of the number of trials and other quantities have to be determined. To enhance the speed at which the experiments are carried out, methods for reducing the variance in the estimated solution (resulting into reduction of the computational time for Monte Carlo simulation) are applied. For complicated and large systems, the *Parallelization and vectorization*- based algorithms are used such as to allow the Monte Carlo method to be implemented efficiently on advanced computer architectures.

Thus, the Monte Carlo method is an artificial sampling method, which may be used for solving complicated problems in analytic formulation and for simulating purely statistical problems. The method is being used more and more in recent years, especially

in those cases in recent years, where the number of factors included in the problem is so large that an analytical solution is either extremely involved or rather impossible. The main idea behind the Monte Carlo method is either to construct a stochastic model that is in agreement with the whole problem analytically or to simulate the whole problem directly. In both cases, an element of randomness has to be introduced according to some well-defined rules. Then a large number of trials are performed, the results are observed, and finally a statistical analysis is undertaken in the usual way. The advantages of the method are, above everything that even very difficult problems can often be treated very easily, and desired modifications can be applied without much trouble. However, the disadvantages of Monte Carlo method are the relatively poor (vis-à-vis analytical methods) precision and the large number of trials that are necessary. Now, when very fast computers are readily available, the cumber of calculation in large number of trials is no longer a serious matter. This facility may be utilized to increase the number of trials sufficiently large so as to attain an acceptable degree of precision.

**Monte Carlo Method and Study of Properties of Estimators of a Linear Model**: A number of attempts have been made to investigate into the small sample properties of various estimators of linear model. We find a detailed discussion on the application of Monte Carlo method to study the properties of the estimation methods of multi-equation models in **Intriligator**, 1978, pp. 416-420.

In our endeavour to gauge the performance of LAD estimator vis-a vis LS estimator, we propose to use of Monte Carlo experiments. The essence of the Monte Carlo study is that instead of estimating unknown parameters using a specific technique, known parameters, which are chosen before hand, are estimated using different techniques. Comparisons between estimated and true parameters are subsequently used to

make inference about the different techniques. It simulates the process of estimating parameters using a controlled setting in which the true parameter values are known, like a 'laboratory' in which controlled experiments on econometric estimators can be studied.

**III. Random Numbers, their Distributions and Generation**: As mentioned before, random numbers play a very important role in the Monte Carlo method. Random numbers are almost always generated on computers by using some mathematical method. Such random numbers (generated by using some mathematical formula) are in fact **pseudo-random numbers**, so called because there exists a rule (the mathematical formula) for generating them and they have a well-defined periodicity. However, if the period is large enough and the numbers so generated satisfy certain statistical tests, then these numbers are random for all practical purposes. There are three leading methods for generating rectangular or uniformly distributed (pseudo) random numbers, namely the Mid- square method, the Fibonacci method and the Power method. These methods can be combined or hybridized also. Random numbers with various statistical distributions are obtained by suitable transformation of rectangular distributed random numbers. In most cases, first the rectangular distributed random numbers are transformed into normally distributed (standard) random numbers by an application of the Central Limit Theorem (i.e., the average of sufficiently large number of rectangular distributed random numbers has normal distribution) and suitable scaling.

Alternatively, if $r_1$ and $r_2$ are the two uniformly distributed random numbers lying between zero and unity, then $u = \sqrt{-2\ln(r_1)}\cos(2\pi r_2)$ follows standard normal distribution. Here $\pi = 4\tan^{-1}(1) \cong 3.1415926535897932$ (Texas Instruments, p. 54). Then these normally distributed random numbers are transformed into other random numbers with desired distribution.

**Pseudo-random Number Generators**: Some of the pseudo-random number generators are discussed below.

(i) ***Generator 1. (Mid- square method):*** A Sequence of the pseudo-random numbers $\{r_i\}$, i = 1,2, ..., where each $r_i$ is a 38 bit number is generated by the recursive algorithm defining

$r_{i+1}$ = mid (part of) 38 bits of $r_i^2$

$r_1$= 01 0000 1011 1011 1011 1111 1010 0100 1101 1110

i = 1,2,... to any large integer.

This sequence terminates in zero for i = 750000 and has been very thoroughly tested for randomness and found useful in many practical problems.

The sequence of random numbers generated by this method may exhaust in the case of large problems. In such circumstances, the same sequence can be used repeatedly without repetition of results, provided the initial $r_1$ is not used at the identical place it was first called.

Alternatively, the random numbers generated can be increased two or four times by shifting the available random numbers. This procedure involves using in sequence not only $r_i$ but also the fractional part of $2^{\alpha_1} r_i$, $2^{\alpha_2}\left(2^{\alpha_1} r_i\right)$, etc as random numbers before squaring $r_i$ to obtain $r_{i+1}$. (It should however, be noted that through such shifts the left most digit is lost).

(ii) ***Generator 2 (Mid-square-bit method):*** Algorithm for generating a sequence of pseudo- random numbers $\{r_i\}$, i=1, 2, ... is given as follows (**Krishnamurthy & Sen**, p. 303).

Let $r_1$, be an 8 bit random number.

(b). Pick up its $8^{th}$ (or $9^{th}$) bit, place it in the $8^{th}$- bit position of $r_2$

(c). Take $r_{11}$ = mid (part of) 8 bits of $r_1^2$. Square $r_{11}$. Pick up the $8^{th}$ (or $9^{th}$) bit of $r_{11}^2$, place it in the $7^{th}$- bit position of $r_2$.

(d). Take $r_{12}$= mid (part of) 8 bits of $r_{11}^2$. Square $r_{12}$. Pick up the $8^{th}$ (or $9^{th}$) bit of $r_{12}^2$, place it in the $6^{th}$ bit position of $r_2$ .

(e) Continue the process five more times to obtain an 8-bit $r_2$.

Using this algorithm, $r_3$ can now be generated form $r_2$. The process can be continued to generate $r_i$. Although, such generation is computationally more time consuming, the method has been tested and found to be very useful.

***Generator 3 (Power residue method):*** The power residue method is the most commonly used method to generate pseudo random sequences. A sequence of integers $r_1$, $r_2$, ... is defined recursively by

$$r_{i+1} \equiv \alpha r_i (\text{mod } \mu)$$

where $r_1$ = a starting value, $\alpha$ and $\mu$ are certain integers and the congruential notation means that $r_{i+1}$ is the ( positive ) remainder when $\alpha r_i$ is divided by $\mu$. Now, the sequence $r_1$, $r_2$ ,... will be a periodic sequence whose period cannot exceed $\mu$, since division by $\mu$ can produce at most $\mu$ different remainders ( viz, 0, 1, 2, ...,$\mu$-1 ). The integers $\alpha$ and $\mu$ should be selected in such a manner that they are capable of producing a very long period relative to the number of random numbers required in computation.

The Power residue algorithm for the generation of pseudo- random numbers is:

(1) Let the given computer have a word length of K bits. Also let the arithmetic be carried out with the binary point to the extreme right or, equivalently, let the arithmetic be the integer.

(2) Choose $\alpha$ = an integer of the form $8x \pm 3$ and close to $2^{K/2}$

(3) Chose $\mu = 2^K$

(4) Choose $r_1$ = any odd integer. Now $\alpha r_1$, =2K- bit number.

(5) Discard the K high-order bit. The residue $\mu_2$ is then the K low-order bit number.

The process is iterated. To obtain a pseudo random number $r_i$ in [ 0,1] , the binary point is considered to be at the extreme left. The period of the sequence $r_1$ , $r_2$ .... is $2^{K-2}$ for a 32-bit computer, this sequence has a period $2^{30} \doteq 10^9$.

On most of the Personal Computers, we may generate uniformly distributed random numbers lying between 0 and 1 by the following algorithm:

(1). Declare IU and IV as the two 2-byte integers.

(2). Feed a seed, IU (preferably odd and of five digits, but less than 32766).

(3). Define IV=259 x IU

(4). If IV $\geq$ 0 then R=IV; Re-initiate IU = IV; Standardize R = R x 0.3051851E-04.

Store/print the random number R

Go to step (3) to generate the next random number with the re-initiated IU.

**Else**

(5). Replace IV by IV + 32767+1; R = IV; Re-initiate IU = IV;

Standardize R = R x 0.3051851E-04.

Store/print the random number R

Go to step (3) to generate the next random number with the re-initiated IU.

**Distributions of Random Numbers**: There are different types of random numbers, some discrete and others continuous. They follow different types of distribution. Data collected from nature following random processes throw up different types of distribution. Karl Pearson identified eight types of empirical distributions (Kapur & Saxena, pp. 296-300).

Additionally, there are umpteen number of theoretical distributions. Here we give the salient features of some distributions relevant for our purpose.

***The Normal Distribution***: A continuous random variable, e, has the normal distribution if it has the probability density function (pdf)

$$f(e) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}\left(\frac{e-\mu}{\sigma}\right)^2\} \text{;where} \quad -\infty < e < \infty$$

In the standard normal distribution the mean is taken as zero and the standard deviation as unity. The pdf now reduces to

$$f(e) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}e^2\right)$$

The important features of this function are :

1. The normal probability curve is bell shaped and symmetrical about the ordinate at e= $\mu$.
2. The ordinate decreases rapidly as e increases numerically. Curve extends to infinity on either side of the mean. The maximum ordinate is at e = $\mu$ and is given by Ymax = $\frac{1}{\sigma\sqrt{2\pi}}$.
3. The ordinate at e = $\mu$ divides the area under the normal curve into two equal parts. Thus the median of the distribution coincides with the mean and the mode.
4. The two points of inflexion of the normal curve are equidistant from the mean.
5. For large values of $\sigma$, the normal curve tends to be flatter, while for small values of $\sigma$, it has a small peak.

6. No portion of the curve lies below the X-axes since normal probability function can not be negative.

***Non-normal Distributions***: The random nature of the error term has prompted econometricians to 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 characterised by infinite variance. Thus, the possibility of the error term following a non-normal distribution exists. In such cases, the least squares estimator will give less efficient estimates. Non-normal distributions are those, which acquire skewed, platykurtic or leptokurtic shapes when plotted i.e., they are asymmetrical about the mean or deviate significantly from the bell-shaped normal distribution. Three important and well-known non-normal distributions that have been considered in the study are discussed below.

***Cauchy Distribution:*** This distribution is named after its discoverer KL Cauchy (1789-1857). The distribution is useful in statistical theory as it is an example of a distribution for which a number of theorems fail and hence emphasizes the need for scrutiny of the conditions under which statistical theorem holds.

The pdf of the Cauchy-distributed variate, e, is given by:

$$f(e)=\frac{1}{\pi}\cdot\frac{1}{1+(e-\mu)^2}, \qquad -\infty < e < \infty$$

The pdf curve is symmetrical about the point $e = \mu$ and thus $e = \mu$ gives us the median of the distribution. Also the density function is obviously maximum at $e = \mu$ so that this point is also the mode. The mean is given by:

$$\mu_1' = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{ede}{1+(e-\mu)^2} = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{(e-\mu)de}{1+(e-\mu)^2} + \frac{\mu}{\pi}\int_{-\infty}^{\infty}\frac{de}{1+(e-\mu)^2}$$

Now since $\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{de}{1+(e-\mu)^2} = 1$, we have

$$\mu' = \mu + \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{ydy}{1+y^2} = \mu + \frac{1}{\pi}\left[\lim_{e,e'\to\infty}\int_{-e}^{e'}\frac{ydy}{1+y^2}\right]$$

If $e, e'$ approach infinity independently, the second integral does not converge to a limit and thus, in the most general sense, the mean of the Cauchy distribution does not exist. However, if $e' = e$, the integral vanishes and thus if we take the principal value of the integral, the mean of the Cauchy distribution is at $\mu$. Thus, in this sense the median and mode of the distribution coincide and the distribution is symmetrical.

The variance of the distribution is $\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{(e-\mu)^2}{1+(e-\mu)^2}de$ and the integral does not converge. Thus for Cauchy distribution, the second and higher moments about the mean do not exist. The cumulative distribution function (cdf) of the distribution of $e$ is :

$$F(e) = \frac{1}{\pi}\int_{-\infty}^{e}\frac{de}{1+(e-\mu)^2} = \frac{1}{\pi}\left[\tan^{-1}(e-\mu) + \frac{\pi}{2}\right]$$

The characteristic function of the distribution about the mean is

$$\phi(t) = E\left[\exp.\{it(e-\mu)\}\right] = \frac{1}{\pi}\int_{-\infty}^{\infty}\exp.\{it(e-\mu)\}\left\{\frac{1}{1+(e-\mu)^2}\right\}de$$

$$= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\exp(ity)}{1+y^2} dy = \exp\{-|t|\}.$$

The Cauchy-distributed variate, e, relates to normally distributed variate in an interesting manner which is relevant to generate e from the latter. If $u_1$ and $u_2$ are independent normal variates with means $m_1$ and $m_2$ and variances $\sigma_1^2$ and $\sigma_2^2$ then the variate $e = \frac{u_1 - m_1}{u_2 - m_2}$ is Cauchy distributed (**Kapur & Saxena**, 1982; p. 427). In particular, the quotient of two independent standard normal variates is Cauchy distributed.

**Gamma Distribution**: The continuous random variate, e, which is distributed with pdf

$$f(e) = \frac{m^l}{\Gamma(l)} e^{l-1} \exp(-me) \quad , e \geq 0, l \geq 0, m > 0$$

is called a Gamma Variate with parameters $l$ and $m$, and the distribution is called the Gamma distribution. The function $f(e)$ represents a probability density, since its integral over the range $(0, \infty)$ is unity.

The moment generating function (mgf) with respect to origin for gamma distribution with $m=l$ is

$$M_{e=0}(t) = \int_0^{\infty} \exp(te) \frac{1}{\Gamma(l)} e^{l-1} \exp(-e) \, de = \frac{1}{\Gamma(l)} \int_0^{\infty} e^{l-1} \exp\{-e(t-1)\} de = (1-t)^{-l},$$

provided that $|t| < 1$

Differentiating $M_{e=0}(t)$, r times with respect to $t$ and putting $t=0$, we find

$$\mu_r' = l(l+1).....(l+r-1)$$

In particular, $\mu_1' = l$, $\mu_2' = l(l+1)$ and $\mu_2 = l, \mu_3 = 2l$ and $\mu_4 = 6l + 3l^2$. Hence,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{4}{l} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{6}{l}$$

Hence as $l \to \infty, \beta_1 \to 0$ and $\beta_2 \to 3$ so that Gamma distribution tends to normal distribution as the parameter tends to infinity.

**Two Important Properties of Gamma Variates**: Gamma variates have two important properties relevant to our investigation; (i) the sum of two independent Gamma Variates with parameters $l$ and $m$ is a Gamma Variate with parameter $l+m$ and (ii) if u is a standard normal variate, then $e = \frac{1}{2}v^2$ is a Gamma variate with parameter $\frac{1}{2}$ (**Kapur & Saxena**, pp. 288,289). Using these properties, it is possible to generate Gamma distributed random numbers from standard normal random numbers.

**Beta Distributions**: There are two types of Beta-distributed random variables, the first is called $\beta_1$-distributed, and the second is called $\beta_2$-distributed. The continuous random variate, e, which is distributed with pdf

$$f(e) = \frac{1}{B(l,m)} e^{l-1}(1-e)^{m-1}, 0 \le e \le 1;\ l,m > O$$

$$= 0, \text{ elsewhere}$$

is a Beta variate of the first kind with parameters $l$ and $m$ and is referred to as a $\beta_1(l,m)$ variate. The mgf with respect to the origin for this distribution does not have a simple form but the moments are easily found directly.

$$\mu_r' \text{ about origin } = \frac{1}{B(l,m)} \int_0^1 e^{r+l-1}(1-e)^{m-1} de$$

$$= \frac{B(r+l,m)}{B(l,m)}$$

$$= \frac{l(l+1)...(l+r-1)}{(l+m)(l+m+1)...(l+m+r-1)}, r = 1,2,.....$$

In particular, $\mu_1' = \frac{l}{l+m}$, $\mu_2' = \frac{l(l+1)}{(l+m)(l+m+1)}$, $\mu_2 = \frac{lm}{(l+m)^2(l+m+1)}$

Putting $r = 3$ and $r = 4$ we successively obtain $\mu_3$ and $\mu_4$ and hence we can show that

$$\mu_3 = \frac{2lm(m-l)}{(l+m)^3(l+m+1)(l+m+2)}$$

$$\mu_4 = \frac{3lm\{lm(l+m-6)+2(l+m)^2\}}{(l+m)^4(l+m+1)(l+m+2)(l+m+3)}$$

In case $l=l=m$, $f(e) = \frac{1}{B(1,1)}.1 = 1, 0 < e < 1$

which is the probability function of uniform distribution in the range *(0,1)*

Beta Distribution of the Second Kind: The continuous random variate, e, which is distributed with pdf

$$f(e) = \frac{1}{B(l,m)}.\frac{e^{l-1}}{(1+e)^{l+m}}, e \geq 0;\ l,m > 0$$

$$= 0 \text{ for } e < 0.$$

is a Beta variate of second kind with parameters $l$ and $m$ and may be referred to as $\beta_2(l,m)$ variate. The function given above represents a probability density since

$$\int_0^\infty f(e)de = \int_0^\infty \frac{1}{B(l,m)}.\frac{e^{l-1}}{(1+e)^{l+m}}de$$

$$= \frac{1}{B(l,m)}\int_0^1 y^{m-1}(1-y)^{l-1}dy$$

on putting $\quad 1+e = y^{-1},\ de = \frac{dy}{y^2}$

$$\therefore \int_0^\infty f(e)de = 1 \qquad \text{since } B(l,m) = B(m,l)$$

$$\text{Also } \mu_r' \text{ about origin} = \int_0^{\infty} \frac{1}{B(l,m)} \cdot \frac{e^{l+r-1}}{(1+e)^{l+m}} de$$

$$= \frac{1}{B(l,m)} \int_0^1 y^{m-r-1}(1-y)^{l+r-1} dy, \quad 1+e = y^{-1}$$

$$= \frac{1}{B(l,m)} \int_0^1 Z^{l+r-1}(1-Z)^{m-r-1} dz \quad \text{where } Z=1-y$$

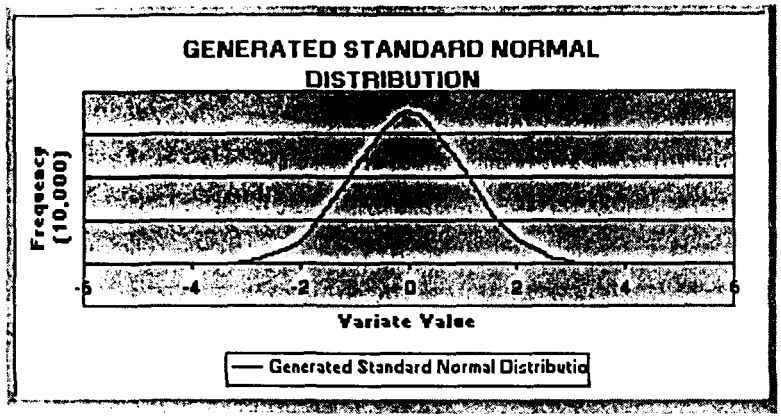
$$\therefore \mu_r' = \frac{B(l+r,m-r)}{B(l,m)} = \frac{l(l+1)...(l+r-1)}{(m-1)(m-2)...(m-r)}; \; r < m.$$

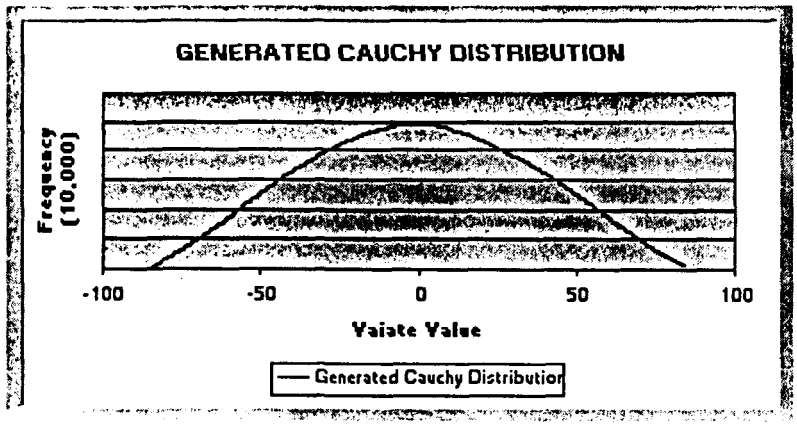
Beta-distributed random numbers are related to Gamma-distributed random numbers in a very interesting manner that may be exploited to generate Beta-distributed variates from the Gamma-distributed variates. From two independent Gamma variates, $g_1$ and $g_2$ with parameters $l$ and $m$ respectively, we may obtain $e_1 = g_1/(g_1+g_2)$, which is a $\beta_1(l,m)$ distributed variate, and $e_2 = g_1/g_2$, which is a $\beta_2(l,m)$ distributed variate (**Kapur & Saxena**, 1982; p. 292).

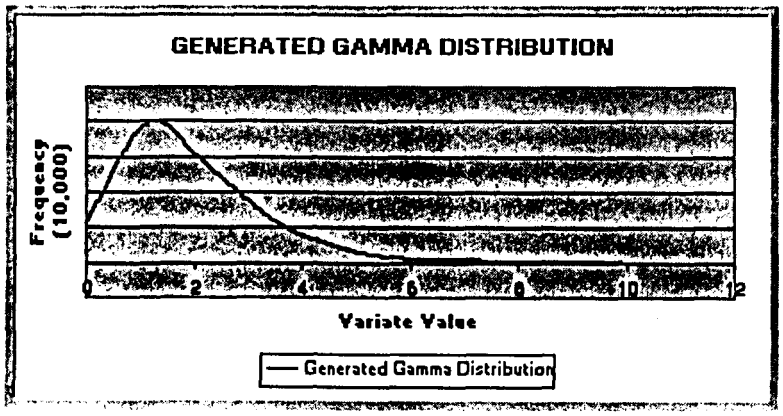
**Generation of Random numbers following different Distributions**: Simulation requires generation of a large number of random numbers that follow the desired distribution. The enterprise starts with the generation of uniformly distributed random numbers lying between 0 and 1. Uniformly distributed random numbers may be transformed into normally distributed random numbers, $x \sim N(0,1)$, by the transformation $x = \sqrt{\{-2\ln(u_1)\}}\{cos(2\pi u_2)\}$ where $u_1$ and $u_2$ are uniformly distributed independent random numbers lying between (0,1) and x is the standard normal variate (**Knuth** (1969), **Texas Instruments Inc** (1979), p 54). Alternatively, one may generate $x \sim N(0,1)$ from uniformly distributed u(0,1) numbers, by using the Central Limit Theorem (**Gillett** 1979, p. 519; **Kapur & Saxena,** 1982, p. 386). However, this method is less accurate and time consuming than Knuth's method. Normally distributed variate, x, may be used to generate Gamma distributed variate, g, since, if x is a standard normal variate, then g

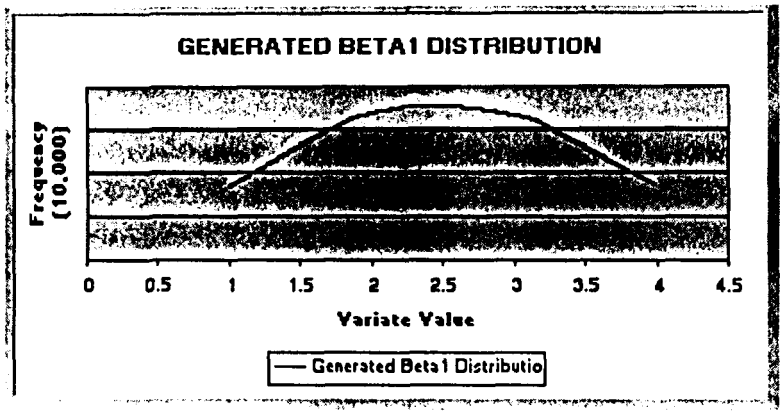
$=\frac{x^2}{2}$ is a Gamma variate with parameter $\frac{1}{2}$ (**Kapur & Saxena**, 1982; p 288). Due to the additive property of Gamma variates, if $x_i$ ($i$=1,2,...,n) are n independent normal variates with means $m_i$ and standard deviations $\sigma_i$ then $g = \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - m_i)^2}{\sigma_i^2}$ is a Gamma variate with parameter $\frac{1}{2}n$ (**Kapur & Saxena**, 1982; p. 289). From two independent normally distributed variates $x_1$ and $x_2$ we may obtain a Cauchy distributed variate, since, if $x_1$ and $x_2$ are independent normal variates with means $m_1$ and $m_2$ and variances $\sigma_1^2$ and $\sigma_2^2$ then the variate $c = \frac{x_1 - m_1}{x_2 - m_2}$ is Cauchy distributed (**Kapur & Saxena**, 1982; p. 427). In particular, the quotient of two independent standard normal variates is Cauchy distributed. From two independent Gamma variates, $g_1$ and $g_2$ with parameters $l$ and $m$ respectively, we may obtain $v_1 = \frac{g_1}{g_1 + g_2}$, which is a $\beta_1(l,m)$ distributed variate, and $v_2 = \frac{g_1}{g_2}$, which is a $\beta_2(l,m)$ distributed variate (**Kapur & Saxena**, 1982; p. 292). In general, starting from uniformly distributed variates, we may obtain a variate with almost any kind of distribution by a sequence of suitable transformations.

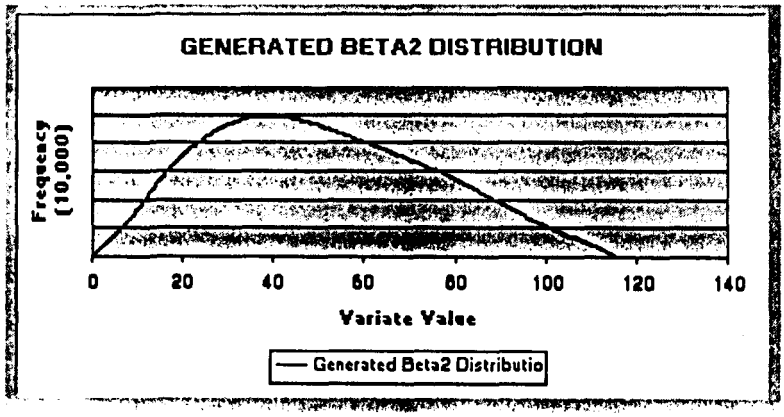
Graphic presentations of different distributions generated on computer (using the procedure mentioned above) give a fairly representative view of their nature. Among these, Gamma is a very skewed distribution for small shape and scale parameters. Gamma variates are non-negative. Beta variates are non-negative and largely platy-kurtic.


GENERATED STANDARD NORMAL DISTRIBUTION
Frequency (10,000)
-6
-4
-2
0
2
4
6
Variate Value
Generated Standard Normal Distributio


GENERATED CAUCHY DISTRIBUTION
Frequency (10,000)
-100
-50
0
50
100
Vaiate Value
Generated Cauchy Distributior


GENERATED GAMMA DISTRIBUTION
Frequency (10,000)
0
2
4
6
8
10
12
Variate Value
Generated Gamma Distributior


GENERATED BETA1 DISTRIBUTION
Frequency (10,000)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Variate Value
Generated Beta1 Distributio


GENERATED BETA2 DISTRIBUTION
Frequency (10,000)
0
20
40
60
80
100
120
140
Variate Value
Generated Beta2 Distributio

# CHAPTER 5

## PERFORMANCE OF LAD IN ESTIMATION OF SINGLE-EQUATION LINEAR MODELS

**I. Introduction**: In this chapter we study Least Absolute Deviation (LAD) estimator as an alternative to OLS estimator of parameters of a single linear regression model. To estimate the regression parameters by LAD we use Fair-Schlossmacher algorithm. It is an iterative method. It works like a weighted Least Squares method of estimation, where weights are inverse of the absolute value of the residuals obtained by OLS. These weights change from iteration to iteration until convergence (up to a pre-determined accuracy level) is reached. This algorithm works better than most of the search methods of non-linear optimization.

After we obtain the estimates of regression coefficients by the Fair-Schlossmacher algorithm, we try to investigate if these estimates may further be improved (yield smaller value of the loss function). We have tried with several methods (Hooke-Jeeves, Nelder-Mead, Rosenbrock and Powel) but with no success. However, the Random Walk method has succeeded in improving the parameter estimates most of the time. It may be mentioned that the Random Walk method by itself does not minimize the loss function effectively. However, it improves the Fair-Schlossmacher estimates quite effectively. This procedure – first obtaining LAD estimates by Fair-Schlossmacher method and thereafter improving it by the Random Walk method – has been named as the LADRW estimation. Therefore, this method also is a candidate method of estimation.

**II. Specification of Single Equation Model:** A single equation model is specific in the sense of (i) the number of explanatory variables (except the constant) in the model, (ii) distribution (or probability density function) of disturbance vector in the model including the magnitude of variance of the disturbance vector, (iii) presence (number and size) of outliers in the disturbance vector and lastly (iv) the sample size or the number of cases (or no. of observations). Specification about linearity or otherwise also is important. However, in our experiments we have considered linear models only. Other particulars of the specification are given as follows.

**a). Size of the Single Equation Model**: In our endeavor to ascertain the efficiency of the LAD estimator, Monte Carlo experiments have been conducted using three models of different sizes.

(i) **Model 1**: Model#1 comprises of 1 explanatory variable and additionally a constant. The Model #1 is:

$$y_i = \sum_{j=1}^{1+1} a_j x_{ij} + e_i; i = 1, n;\ x_{i,m+1} = 1\ \forall i(1,n)$$

(ii) **Model 2**: Model#2 has 3 explanatory variables and an additional constant. The model #2 is:

$$y_i = \sum_{j=1}^{3+1} a_j x_{ij} + e_i; i = 1, n;\ x_{i,m+1} = 1\ \forall i(1,n)$$

(iii) **Model 3**: Model#3 has 5 explanatory variables and an additional constant. The model#3 is:

$$y_i = \sum_{j=1}^{5+1} a_j x_{ij} + e_i; i = 1, n;\ x_{i,m+1} = 1\ \forall i(1, n)$$

In the above models, the y's represent the dependent variables and the x's represent the explanatory variables. The error (residual or disturbance) term is e. The sample size is n.

**b). Sample Size**: Samples of two different sizes have been taken in the study

(i) Sample of size (no. of cases or observations) = 20.

(ii) Sample of size (no. of cases or observations) = 50.

**c). Distribution of Disturbance term**: To study the effect of the probability distribution of the error term on the efficiency of the candidate estimator vis-à-vis OLS, we have considered five probability distributions. They are (i) Normal, (ii) Cauchy, (iii) Gamma, (iv) $Beta_1$, and (v) $Beta_2$. Gamma, $Beta_1$ and $Beta_2$ distributed errors have been generated with parameter of size =2. To help the visualization of the nature of these distributions, the figures in the preceding chapter give their pictorial view. As already mentioned earlier, these graphs were obtained by frequency classification and plotting of 10,000 random numbers (of specified distribution) generated by a computer program.

**d). Standard Deviation of the Disturbance term**: The standard deviation of error used in the study is 0.1.

**e). Number and size of Outliers**: Outliers of different number and magnitude have also been introduced to the error vector to examine whether their presence, number and size affect the efficiency and bias of the estimator under study.

(i) The numbers of outliers used in the study are 0 (i.e. absence of outliers), 1, 3 and 5.

(ii) The magnitudes of outliers (when present) are 0.5, 1 and 2 times the mean value of the dependent variable ( $\bar{Y}_t$ ) in a particular model.

**III. Numerical Details**: Monte Carlo experiments are basically numerical methods Additionally, they are computer-based since it is practically impossible to conduct such experiments manually. Once the models are specified, random numbers are generated in accordance with the specifications and the parameters are estimated by the candidate methods of estimation. Necessary information regarding the experiments conducted in this study is given below.

1. We have generated uniformly distributed random numbers lying between zero and unity (0,1). With the random numbers, the matrix of explanatory variables, x(n,m-1), has been formed. The number of rows (cases or observations, n) as well as the number of columns (explanatory variables less one = m-1) are as specified in a particular model. The last (that is $m^{th}$) column of x is a unit vector (=1 for all i; i=1,n). To scale up the magnitude of the numerical values constituting the matrix X (except for the last column), we have multiplied the elements, $x_{ij}$ (i=1,n; j=1,m-1) by the scalars 10. Appropriate scaling is needed to keep up accuracy in the numerical analysis.

2. Then we have chosen (subjectively) appropriate number of regression coefficients, including the constant and used them as parameters to obtain $y_i = \sum_{j=1}^{m} a_j x_{ij}; i = 1, n$. This gives us the non-stochastic y vector.

3. Next, the disturbance vector, $e_i$ ; i=1,n of appropriate distribution has been generated and added to y to obtain the stochastic dependent variable *y*.

4. The accuracy level of the LAD estimates has been fixed at 0.0001.

5. The tolerance level for accuracy which is required for convergence of parameter estimates has been fixed at 0.0001.

6. We have estimated the regression parameters, **a**, by OLS.

7. We have estimated the regression parameters, **a**, by LAD estimator. To obtain the LAD estimates, 100 iterations have been performed. It is assumed that such a number of iteration is sufficient for an accurate estimation.

8. The experiment conducted has been replicated 100 times – each time drawing fresh samples of random disturbance vectors (while keeping the structure fixed), adding them to the dependent variable (non-stochastic y as in 3 above) and estimating the parameters by the candidate methods of estimation OLS and LAD).

The step-wise procedure may be described as follows:

(i) Generate uniformly distributed random numbers to obtain the x(n,m) scale it up by a predetermined positive scalar (say 10) and augment x(n,m) by a unit vector to obtain x(n,m+1). Here n and m are pre-specified.

(ii) Generate y = x**a**.

(iii) Using another seed, generate a set of random variables of desired distribution (agreeing with the order of y vector) and add them to y. The dependent variable thus obtained is now stochastic.

(iv) Decide on the number and the sizes of outliers and add them to $y_i$ (i=1,n) at randomly chosen locations.

(v) Estimate the parameter vector **a** (call it $\hat{a}$) by different estimators (OLS and LAD).

(vi) Repeat steps (iii) to (v) for the desired number of times (100 here) for replication.

(vii) Use the results in (vi) to obtain RMS of deviations of the estimated $\hat{a}$ from the true **a** while $\hat{a}$ is estimated by the different estimators (OLS and LAD). Obtain relative RMS (relative to OLS =1) for different estimators.

Step (5) is repeated for different distributions (of error vectors) with 0.1 magnitude of standard deviations, different numbers and magnitudes of outliers and different model sizes of different sample sizes (no. of cases or observations).

**IV. Details of Specification-wise no. of experiments**: It is facilitating to describe the details of specification-wise number of experiments conducted in our study. In all, we have conducted 300 experiments (each one with 100 replicates yielding one RMS for each candidate estimator, OLS and LAD). These details are as follows:

**Size of Model**: Model#1 (100), Model#2 (100), Model#3 (100)

**Sample Size**: Small i.e. 20 observation (150), Larger, i.e. 50 observations (150)

**Distributions**: Normal (60), Cauchy (60), Gamma (60), $B_1$ (60), $B_2$ (60)

**Number of Outliers**: None (120), One (60), Three (60), Five (60)

**Size of Outliers**: None/zero (120), Small (90), Large (90). In the experiments we have used four different factors (call k) for generating outliers of different sizes. For small samples we use two values of k (0.5 and 1.0). For large samples we use k=1.0 and k= 2.0. The size of outliers is k*mean(y) or k times the mean of the dependent variable. The specified number of outliers are added to different observations of the dependent variable (y) at randomly selected locations (cases or observation). The value of the decision

variable OSIZE takes on these values of k. Thus OSIZE takes on three different values – 0.5, 1.0 and 2.0 depending on the specified conditions.

**V. Findings of Monte Carlo Experiments**: In accordance with the design of Monte Carlo Experiments we have run the simulation program for 300 times with the specifications as given in **Details of Specification-wise no. of experiments** in the earlier section of this chapter. The computer program (source codes of computing programs in FORTRAN 77) is appended. The program estimates 100 replicates (with different disturbance vectors of a specified distribution and different outliers specific in number and size) of regression parameters of a specified model by OLS, LAD (obtained by the Fair-Schlossmacher algorithm) and LADRW (Random Walk method of optimization applied on the iterative estimates obtained by the Fair-Schlossmacher algorithm of LAD estimation) and the RMS of the regression coefficients (with the true parameters as a reference) of the latter two estimators with respect to those of the OLS estimator. Thus for each run we obtain two values of relative norms, one for LAD and the other for LADRW estimator. We have also converted the relative norms of each candidate estimator (LAD and LADRW) into two binary variables (one each for LAD and LADRW such that if the relative norm of a candidate estimator is smaller than the OLS norm then the binary variable takes on a value of unity or zero otherwise. In this scheme, the *unity* value of the related binary value signifies *success* (that is, the candidate estimator has succeeded in estimating the parameters of the regression equation more efficiently than does the OLS estimator) while the value of *zero* signifies that OLS estimator dominates (or is as good as) the candidate estimator and it is considered as a *failure*.