UNIVERSITY OF CAPE TOWN

DEPARTMENT OF MATHEMATICAL STATISTICS

****

A CONTRIBUTION TO ADAPTIVE ROBUST ESTIMATION

by

G.D.I. BARR

****

A thesis prepared under the supervision of Professor A.H. Money assisted by Dr J.F. Affleck-Graves and Dr M.L. Hart in fulfilment of the requirements for the degree of Doctor of Philosophy in Mathematical Statistics

****

A C K N O W L E D G E M E N T S

I would like to thank Professor A H Money for his constant encouragement and help throughout and for guiding me to this extremely interesting field of research. Many thanks are also due to Doctors J F Affleck-Graves and M L Hart. Together my three superivsors sacrificed innumerable hours in discussing aspects of my thesis with me and I am extremely grateful to them.

I would also like to thank the members of staff of the Department of Mathematical Statistics at the University of Cape Town for the encouragement they have given me. In particular I would like to mention Professor C G Troskie who suggested the Stigler study of Part II and Dr L G Underhill who pointed out the existence of the S_U-S_B distribution programs. Many thanks also to Mr R Gonin of the Medical Research Council who helped me in the early stage with some programming problems as did members of the staff of Computing Services of the University of Cape Town. In addition I am indebted to Dr D Hawkins of the CSIR who contributed some useful points at the conference at the Rand Afrikaans Universiteit in November 1980, and Professor D Nel of the Universiteit van die Oranje Vrystaat who helped me with some theoretical work. I would also like to thank Professor N Laubscher of the University of Port Elizabeth for his help and two unnamed referees for helpful comments on two papers which have been submitted for publication.

Special thanks to Mrs M I Cousins, who typed the major part of this thesis, for her patience and perserverance in producing such well typed work. Mrs S Higgins also helped with typing - many thanks to her. Ms E Maclagan proof read the thesis and I am most grateful for the time she put in.

Finally, I am indebted to the University of Cape Town and the Council for Scientific and Industrial Research for the research bursaries awarded in the first year of this study which enabled me to undertake the project in the first place.

Signed by candidate

G D I Barr
February 1981

TABLE OF CONTENTS

INTRODUCTION

This study initially set out to consider the possibility of constructing an adaptive robust estimation procedure for the standard linear regression model when the disturbance vector deviated from normality, however, after the initial success in that field it seemed only appropriate that the approach be extended to robust location parameter estimation. This is a particular case of the regression model and an area in which a number of different estimators have been proposed and a great deal of comparative research work done. Due to the wider scope of such research the greater part of the thesis is devoted to this field of research, which led to many interesting and useful results and conclusions.

In order to follow the usual progression in the literature, i.e. from location parameter estimation to regression vector parameter estimation, the presentation will not be in the researched chronological order, but will hopefully lead the reader through a more logical sequence of facts.

Part I comprises a survey of literature which is most relevant to this topic - essentially the significant contributions in robust parameter estimation, excluding the non-parametric methods.

In Part II the estimation of the location parameter is considered, a new adaptive scheme proposed and a comparative study undertaken whereby the performance of the new proposed estimator is compared with that of several commonly used robust estimators. The distribution of this estimator is considered and its relationship with a certain statistic examined. This in turn led to the statistic being considered as a possible estimator in its own right.

Part III comprises the research into the development of an adaptive robust estimation procedure for the regression model. In addition an idea proposed by Lloyd (1952) for the estimation of the location parameter is considered as a possibility for extension to the regression model.

PART I

CHAPTER 1

SURVEY OF PARAMETRIC ROBUST ESTIMATION

1. INTRODUCTION

A great deal of research effort has been devoted to the fundamental statistical problem of fitting equations to data. The coefficients of the fitted relationship may be of interest in their own right, or the estimated equation may be used for prediction purposes.

Basically the problem is to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_m$ in the relation:

$$\underline{y} = \beta_0 + \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2 + \ldots + \beta_{m-1} \underline{x}_{m-1} + \underline{e} \tag{1.1}$$

where the $\underline{x}_i$'s may be completely different variables or may be polynomials or trigonometric functions of the same variable.

$\underline{e}$ is generally referred to as the random error or disturbance, and the nature and distribution of this error are of considerable importance as far as the estimation of the above parameters is concerned.

The one case which is of particular interest is that where no $\underline{x}$'s enter into the model and $\underline{\beta}$ (the vector of β coefficients) is simply a population location parameter viz.

$$\underline{y} = \underline{1}\theta + \underline{e}$$

(β_0 is usually denoted by θ in the literature.)

Since there is no real concept of dependence or independence in this model it is usually written as

$$\underline{x} = \underline{1}\theta + \underline{e} \tag{1.2}$$

The most widely used method for estimating the coefficients of a functional relationship is undoubtedly the method of least squares. Justification for the frequent use of this method is provided by the following theorem:

Theorem *Gauss-Markoff Theorem (Johnston (1972))*

If we have $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{m-1} x_{m-1i} + e_i$ $(i = 1,2,\ldots,n)$, *where the* e_i *have zero expectation, constant variance* σ^2 *and zero covariances, and the* x_i *are non-stochastic, then the least squares estimator,*

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$$

(where X *is the* $n\times m$ *matrix of observations on the m independent variables and* $\underline{y}$ *is the* $n\times 1$ *vector of observations on the dependent variable), is the best linear unbiased estimator (BLUE) of* $\underline{\beta}$ *with variance-covariance matrix* $\sigma^2(X'X)^{-1}$.

Note that for the model (1.2) the BLUE of θ is:

$$\hat{\theta} = (\underline{1}'\underline{1})^{-1}\,\underline{1}'\underline{x} = \overline{X},$$

the sample mean, with variance:

$$\sigma^2(\underline{1}'\underline{1})^{-1} = \frac{\sigma^2}{n}$$

The above theorem is of considerable importance as it proves that under the stipulated condition the least squares estimator is the best linear unbiased estimator. However, it is important to note firstly, that the estimator is only best in the class of estimators which are linear functions of the <u>unordered</u> data values - thus estimators such as the median or mid-range are excluded as possibilities for the model (1.2), and, secondly, that the result does not consider non-linear alternatives. The attractiveness (from a statistical point of view), of working with linear estimators has made least squares the most popular method of estimation in practice.

In fact, when $\underline{e}$ has a normal distribution least squares provides the maximum likelihood estimator for $\underline{\beta}$, and for the model (1.2) least squares provides the Minimum Variance Bound (MVB) estimator (Kendall and Stuart, Vol II (1973)) so that no estimator, linear or non-linear has smaller variance.

However, it has been shown in the literature (e.g. Andrews (1974)) that these estimators possess some unattractive properties for certain non-normal (non-Gaussian) error distributions. When outlying data points ("outliers") occur in data sets due to errors in observation or long tailed error distributions, they have an unusually large influence on the least squares estimator because they are weighted according to the square of their distance to the fitted parameter or line.

In addition, because outliers tend to pull the fitted line towards them, one might believe, through examination of the estimated residuals that the actual residuals were in fact normally distributed.

In order to circumvent this problem a number of robust (Box (1953) first coined the term "robustness") estimation procedures have been developed. Such estimators are constructed so as to perform well across a wide range of different parent error distributions.

Seminal work on the robust estimation of the location parameter for a symmetric distribution was done by Hodges and Lehmann (1963) and Huber (1964). Work was also done in the early 1960's on the robust estimation of the scale parameter. However, since the interval $(\mu-k\sigma, \mu+k\sigma)$ with k a constant does not contain a constant proportion of the population for different distributions, no natural measure of scale exists, and thus research in this area has been relatively restricted.

Parametric work on robust estimation, since that of Huber's first study, can be broken up into three distinct areas. Firstly, there is M-estimation which is based on maximum likelihood considerations, secondly, L-estimation which considers (in the location parameter case) estimation using linear functions of the order statistics of the sample, and finally L_p-norm estimation (minimization of the p^{th} powers of the residuals), a method which has received some attention in the regression situation but little in the case of location parameter estimation.

CHAPTER 2

M-ESTIMATION

2.1 A BRIEF SURVEY OF SOME OF THE M-ESTIMATION METHODS PROPOSED

Here we are to consider a symmetric probability density function $f(x)$ with finite unique (through symmetry) location parameter θ. For convenience we may write this density as $f(x-\theta)$.

Letting $X_1, X_2, \ldots, X_n$ represent a random sample from $f(x)$, estimation of θ may be carried out (for known $f(\bar{x})$) using the method of maximum likelihood (Fisher (1921)) denoted below as M-estimation.

Denoting the natural logarithm of the likelihood function by $L(\theta)$ we have:

$$\ell n\ L(\theta) = \sum_{i=1}^{n} \ell n\ f(x_i-\theta)$$

In the notation of the literature this equals:

$$-\sum_{i=1}^{n} \rho(x_i-\theta)$$

We maximise the logarithm of the likelihood by differentiation with respect of θ to yield:

$$\frac{d}{d\theta}\ \ell n\ L(\theta) = \sum_{i=1}^{n} \rho'(x_i-\theta)$$

In the notation of the literature this equals:

$$\sum_{i=1}^{n} \psi(x_i-\theta)$$

$$(\rho'(x) = \psi(x))$$

The θ that maximizes $\ell n\ L(\theta)$ is the solution of

$$\sum_{i=1}^{n} \psi(x_i-\theta) = 0 \quad .$$

In order to clarify the use of M-estimation consider in the first place the M-estimation of θ for the normal (Gaussian) and Laplace (doubly-exponential) distributions which yields the sample mean and median respectively.

1) Normal:

$\rho(y) = \frac{y^2}{2} + c \quad ; \quad c$ a constant

$\psi(y) = y$

$\sum_{i=1}^{n} \psi(x_i-\theta) = 0$ yields

$\hat{\theta} = \bar{x}$

2) Laplace:

$\rho(y) = k|y| + c \quad ; \quad k,c$ constants

$\psi(y) = \text{sign}(y).k$

$\sum_{i=1}^{n} \psi(x_i-\theta) = 0$ yields

$\hat{\theta}$ = sample median

Clearly $\bar{X}$ tends to be much more influenced by outliers than the sample median. In robust estimation one is essentially looking for an estimator that performs adequately over the widest possible range of distributions. (The loss of efficiency under a normal distribution for an estimator which is efficient for another distribution is often referred to in robust theory as the premium of that particular estimator.)

With this in mind, Huber (1964) proposed an M-estimator that was asymptotically efficient for a distribution which followed a normal in the central region and a Laplace in the tails. [It should be noted, however, that while $\overline{X}$ is an efficient estimator of θ for normally distributed data with finite sample size, the median is not an efficient estimator of θ in finite samples from a Laplace distribution.] (See for example Sarhan (1954).)

Huber's ρ function was thus:

$$\begin{aligned} \rho(y) &= \frac{y^2}{2} && |y| \leq a; \quad a \text{ a constant} \\ &= a|y| - \frac{a^2}{2} && |y| > a \quad , \\ \therefore \quad \psi(y) &= -a && y < -a \quad , \\ &= y && -a \leq y \leq a \quad , \\ &= a && y > a \quad . \end{aligned}$$

The equation:

$$\textstyle\sum_{i=1}^{n} \psi(x_i - \theta) = 0$$

is then solved for θ by iterative means to yield the Huber M-estimate of θ.

Since the scheme will not be scale invariant for any predetermined constant a, the raw data values are scaled in practice by some robust estimate of scale (d) and the modified equation:

$$\textstyle\sum_{i=1}^{n} \psi\left(\frac{x_i - \theta}{d}\right) = 0$$

solved by iterative means.

A robust scale statistic which is often used is:

$$d = \underset{i}{\text{median}}|x_i - \text{median}\,(x_i)|/0.6745$$

(The factor 0.6745 makes this an approximately unbiased estimate of scale when the data is distributed normally - see Holland and Welsch (1977).)

Using the scale invariant form with d above a can be selected so as to give varying efficiencies for different distributions. We know that in the limiting situation where $a \to \infty$, the estimator yields $\overline{X}$, whereas if a is made very small the estimator gets close to the sample median. If σ is known, then selection of a equal to 1.5 yields an estimator with asymptotic efficiency greater than 95% (Huber (1964)). This estimator performed well in the Princeton Study (Andrews et al (1972)), the premium being small with good protection provided against heavy tailed distributions.

Two possible iterative schemes for finding $\hat{\theta}$, given some starting value $\hat{\theta}^*$ are given below. For the r^{th} iteration they are:

1) $$\hat{\theta}_r = \hat{\theta}_{r-1} + d\,\frac{\sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\theta}_{r-1}}{d}\right)}{\sum_{i=1}^{n} \psi'\left(\frac{x_i - \hat{\theta}_{r-1}}{d}\right)},$$

2) $\hat{\theta}_r = \hat{\theta}_{r-1} + \frac{d}{n} \sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\theta}_{r-1}}{d}\right)$.

The scheme is continued until $\left|\frac{\hat{\theta}_r - \hat{\theta}_{r-1}}{\hat{\theta}_r}\right| < \varepsilon$

for some tolerance ε.

The first scheme represents Newton's method and the second is attributed to Huber (1975). The Huber method is simpler but tends to require more iterations than Newton's method which tends, however, to be more difficult to implement because it requires the first derivative of ψ. Some robust estimate of θ is taken for θ^*, most usually the sample median. Huber's idea is thus, in essence, to "squash in" points which lie greater than some predetermined distance away from an appropriate θ so that they lie exactly this distance away from θ. The chosen value of a may depend on some prior information or expectation of the underlying distribution of X. If, for example, one suspected that the data was drawn from a distribution with long tails, the use of an a smaller than that for an approximately normal distribution would be appropriate.

An extension of Huber's idea is one due to Hampel (1974) where:

$$\psi(y;\ a;\ b;\ c) = \mathrm{sign}(y)\begin{cases} |y| & ; \quad 0 \le |y| < a, \\ a & ; \quad a \le |y| < b, \\ \frac{c-|y|}{c-b}a & ; \quad b \le |y| < c, \\ 0 & ; \quad c \le |y| \end{cases}$$

for some predetermined a, b, c.

Andrews has suggested (Andrews et al (1972))

$$\psi(y) = \sin(y/d) \quad ; \quad |y| \leq d\pi,$$
$$= 0 \quad ; \quad |y| > d\pi.$$

This estimator with $d = 2.1$ was used in the Princeton study (Andrews et al (1972)). However, more robust forms of this estimate with $d = 1.5$ or 1.8 have recently been proposed (Andrews (1974)).

An estimator which is very similar to, and an adequate substitute for, the Andrews estimation, is the Tukey Biweight (Beaton and Tukey (1974)) with

$$\psi(y) = y\left[1 - \left(\frac{y}{e}\right)^2\right]^2 \quad ; \quad |y| \leq e,$$
$$= 0 \quad ; \quad |y| > e,$$

e is usually taken to be 5.0 or 6.0.

It is noted by Hogg (1979) that since the ρ functions associated with the Andrews and Tukey functions are not convex it is possible that there may be convergence problems in the iterative scheme.

Other M-estimators include those due to Dennis and Welsch (1976) with:

$$\psi(y) = y \exp\left(- \left(\frac{y}{\delta}\right)^2\right) \quad ; \quad \delta \text{ a constant}$$

and Fair (1974) with:

$$\psi(y) = y\left(1 + \frac{|y|}{g}\right)^{-1} \quad ; \quad g \text{ a constant.}$$

The predetermined constants in the ψ functions are known as tuning constants in the literature.

The following table, extracted from Holland and Welsch (1977), gives values for the tuning constants for 95% efficiency at the unit normal distribution for some of the M-estimators described above.

Weight Function	A	T	F	H	W
Tuning Constant	1.339	4.685	1.400	1.345	2.985

where A is the Andrews scheme
T is the Tukey biweight
F is the Fair scheme
H is the Huber scheme
and W is the Dennis and Welsch scheme .

2.2 ASYMPTOTIC DISTRIBUTION OF THE M-ESTIMATORS

Expansion of $\sum_{i=1}^{n} \psi((X_i-\hat{\theta})/d = 0$

as a Taylor series about θ, gives for the case $d = \sigma$ (the population standard deviation):

$$\sqrt{n}(\hat{\theta}-\theta) \simeq \frac{\sigma\left[\Sigma\ \psi\left(\frac{X_i-\theta}{\sigma}\right)\right]/\sqrt{n}}{\left[\Sigma\ \psi'\left(\frac{X_i-\theta}{\sigma}\right)\right]/n}$$

It may be shown (Hogg (1979)) that this expression has a limiting normal distribution with mean zero and variance :

$$\frac{\sigma^2\ E\left[\psi^2\left(\frac{X-\theta}{\sigma}\right)\right]}{E\left[\psi'\left(\frac{X-\theta}{\sigma}\right)\right]^2}$$

This may be approximated by:

$$s_\psi^2 = \frac{d^2\left(\frac{1}{n}\sum_{i=1}^n \psi^2\left[(x_i-\hat{\theta})/d\right]\right)}{\left(\frac{1}{n}\sum_{i=1}^n \psi'\left[(x_i-\hat{\theta})/d\right]\right)^2} .$$

Thus for large n:

$\frac{\sqrt{n}\ (\hat{\theta}-\theta)}{s_\psi}$ has an approximate $N(0,1)$ distribution and statistical inferences may be made about θ. Gross (1976) has examined confidence intervals based on this idea.

2.3 M-ESTIMATION OF THE $\underline{\beta}$ VECTOR IN THE REGRESSION MODEL

The problem of finding M-estimates of the $\underline{\beta}$ vector in the regression model is handled in a parallel manner to that of estimating the location parameter θ. The essential difference is merely that in the iterative scheme it is the residuals of the regression fit which are transformed according to the particular ψ function at each iteration and not the raw data set.

The M-estimate of $\underline{\beta}$, $\hat{\underline{\beta}}$ say, maximizes:

$\sum_{i=1}^n \rho\left(\frac{y_i - x_i\hat{\underline{\beta}}}{d}\right)$; where x_i is the i^{th} observation on the set of m independent variables.

A necessary condition for maximization is that $\hat{\underline{\beta}}$ satisfies

$$\sum_{i=1}^n x_{ij}\,\psi\left(\frac{y_i - x_i\hat{\underline{\beta}}}{d}\right) = 0 \quad , \qquad \forall_j = 1,\ldots,m$$

The solution of this system requires, as with the location parameter case, an initial estimate $\hat{\underline{\beta}}^*$ of $\underline{\beta}$ and a robust estimate of scale d. In a parallel way to the iterative scheme for evaluating θ, the L_1 (minimization of the sum of absolute deviations) estimate of $\underline{\beta}$ is usually taken as an adequate robust initial estimate of $\underline{\beta}$. Using this estimate, d is usually calculated from

$$d = [\text{median}|(y_i - x_i\hat{\underline{\beta}}^*) - \text{median}\ (y_i - x_i\hat{\underline{\beta}}^*)|]/(0.6745)$$

Defining the weight function:

$$w(z) = \psi(z)/z$$

and letting $< >$ denote an $n\times n$ diagonal matrix, the following three iterative schemes are commonly used to calculate M-estimates of $\underline{\beta}$ for some predetermined ψ function and tuning constant. For the r^{th} iteration they are:

1) $$\hat{\underline{\beta}}_r = \hat{\underline{\beta}}_{r-1} + d\left(X' < \psi'\left(\frac{\underline{y}-X\hat{\underline{\beta}}_{r-1}}{d}\right) > X\right)^{-1} X'\ \psi\left(\frac{\underline{y}-X\hat{\underline{\beta}}_{r-1}}{d}\right),$$

2) $$\hat{\underline{\beta}}_r = \hat{\underline{\beta}}_{r-1} + d(X'X)^{-1}\ X'\ \psi\left(\frac{\underline{y}-X\hat{\underline{\beta}}_{r-1}}{d}\right),$$

3) $$\hat{\underline{\beta}}_r = \hat{\underline{\beta}}_{r-1} + \left(X' < w\left(\frac{\underline{y}-X\hat{\underline{\beta}}_{r-1}}{d}\right) > X\right)^{-1} X' < w\left(\frac{\underline{y}-X\hat{\underline{\beta}}_{r-1}}{d}\right) > (\underline{y}-X\hat{\underline{\beta}}_{r-1})\Big),$$

where X is the $n\times m$ matrix of the independent variables and $\underline{y}$ the $n\times 1$ vector of the dependent variable.

The first method is Newton's, the second is due to Huber (1975) and the third due to Beaton and Tukey (1974). In the above schemes d is usually calculated once before starting the iterations. Dutter (1977) has, however, proposed a modification of Huber's method with d being calculated at each iteration using the scheme:

$$d_r^2 = \frac{1}{(n-m)\ E(\psi^2)} \sum_{i=1}^{n} \left[\psi\left(\frac{y_i - x_i \underline{\hat{\beta}}_{r-1}}{d_{r-1}} \right) \right]^2 d_{r-1}^2 \quad .$$

The r^{th} iteration for $\underline{\hat{\beta}}$ is then

$$\underline{\hat{\beta}}_r = \underline{\hat{\beta}}_{r-1} + q\ d_r\ (X'X)^{-1}\ X'\ \psi\left(\frac{\underline{y} - X\underline{\hat{\beta}}_{r-1}}{d_r} \right)$$

and (taking the Huber scheme as illustrative):

$$q = \min\left[\frac{1}{\phi(a) - \phi(-a)}\ ,\ 1.9 \right]$$

where ϕ is the standardised normal distribution function.

This procedure is continued until

$$\left| q\ d_r\ (X'X)^{-1} X'\ \psi\left(\frac{\underline{y} - X\underline{\hat{\beta}}_{r-1}}{d_r} \right) \right| < \varepsilon\ d_r \sqrt{<X'X>j}$$

for all $j = 1, 2, \ldots, m$

($<X'X>j$ is the j^{th} diagonal element of $X'X$)

and

$$\left| \frac{d_r - d_{r-1}}{d_r} \right| < \varepsilon$$

for some tolerance ε.

For monotone ψ-functions (e.g. Huber and Fair) Huber (1973) has shown that under certain conditions (one is a known spread $d = \sigma$) iterated solutions derived from the Huber scheme above have an asymptotically normal distribution with variance-covariance matrix:

$$\sigma^2 \frac{E_F(\psi^2)}{[E_F(\psi')]^2} (X'X)^{-1} \quad ,$$

where F is the cumulative distribution function of the underlying error distribution. An approximation to this variance-covariance matrix is (Hogg (1979)):

$$\frac{(nd^2)\left\{\frac{1}{n} \sum_{i=1}^{n} \psi^2\left(\frac{y_i - x_i\hat{\underline{\beta}}}{d}\right)\right\} (X'X)^{-1}}{(n-m)\left\{\frac{1}{n} \sum_{i=1}^{n} \psi'\left(\frac{y_i - x_i\hat{\underline{\beta}}}{d}\right)\right\}^2} .$$

In conclusion, it is seen that the M-estimation method is, using any of a range of ψ-functions, equally applicable to estimation of θ and $\underline{\beta}$. This flexibility, along with the progress made in the distribution theory, probably makes it the most commonly used robust procedure.

CHAPTER 3

ESTIMATION USING LINEAR COMBINATIONS OF ORDER STATISTICS (L-ESTIMATION)

3.1 REVIEW OF WORK ON L-ESTIMATION OF θ IN SYMMETRIC DISTRIBUTION

Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the order statistics for a sample of size n from a continuous symmetric density with location parameter θ. Linear combinations of these order statistics can be constructed to estimate θ and are often known as L-estimators; such an estimator is denoted by:

$$Z_n(\underline{c}) = \sum_{i=1}^{n} c_i \, X_{(i)}$$

with the general conditions

$$\sum_{i=1}^{n} c_i = 1$$

$$c_{n-i+1} = c_i, \qquad \text{for} \quad i = 1,2,\ldots,[\tfrac{n}{2}]$$

The most common examples of such estimators are:

1) The sample mean, $\bar{X}$ where $c_i = \frac{1}{n}$, $i = 1,2,\ldots,n$.

2) Sample median, where

$$\left.\begin{array}{l} c_{[\frac{n}{2}]} = 1 \\ c_i = 0, \ i \neq [\frac{n}{2}] \end{array}\right\} \text{ if } n \text{ odd}$$

$$\left.\begin{array}{l} c_{\frac{n}{2}} = c_{\frac{n}{2}+1} = \frac{1}{2} \\ c_i = 0, \ \text{for all other } i \end{array}\right\} \text{ if } n \text{ even}$$

3) Midrange, where

$$c_1 = c_n = \tfrac{1}{2},$$

$$c_i = 0, \quad \text{otherwise.}$$

The pioneering work on the L-estimation of the location parameter and the scale parameter using best (minimum variance) linear combinations of order statistics was done by Lloyd (1952). This work was closely followed by a series of papers by Sarhan (1954, 1955a, 1955b) which explored the applicability of such estimators to a variety of distributions. Lloyd applied the method of generalized least squares to the ordered sample of a known distribution (which depended on location and scale parameters only) to find estimators of the location parameter and scale parameter. These estimators would be linear (in the ordered sample) and unbiased and have minimum variance in the class of such estimators.

Lloyd considered the ordered model:

$$\underline{Y} = \underline{1}\,\theta + \underline{u}$$

where $\underline{Y} = \begin{pmatrix} X_{(1)} \\ X_{(2)} \\ \vdots \\ X_{(n)} \end{pmatrix}$

and $\underline{u}$ is the ordered disturbance term

with $E(\underline{u}\underline{u}') = \sigma^2 V$

where V is a positive definite matrix which is known if the underlying disturbance is known.

Generalized least squares (Aitken (1935)) yields:

$$\hat{\theta} = (\underline{1}'\, V^{-1}\, \underline{1})^{-1}\ (\underline{1}'\, V^{-1}\, \underline{Y})$$

the BLUE (best linear unbiased estimator) for the ordered sample or BLSS (best linear systematic statistic) with variance

$$\sigma^2 (\underline{1}'\, V^{-1}\, \underline{1})^{-1}.$$

It can be shown that this estimator will have smaller variance than the least squares estimator $(\overline{X})$ if the reciprocal of the smallest eigenvalue of V^{-1} is less than unity. This will be expanded upon in Part III.

Sarhan published tables for the optimal weights of the order statistics for a number of distributions up to a sample size of 5. The distributions he considered were the Uniform, Triangular, Normal, Laplace, (Doubly Exponential), Exponential, U-shaped (Beta (3,1))*, Parabolic (Beta (2,2)) and his so-called "skewed distribution" (Beta (3,2)).

Apart from the Normal and Exponential distributions which yield $\bar{X}$ as the BLSS (the Cramer-Rao lower bound), explicit forms of the matrix of covariances of the order statistics as a function of n were attainable only for the Uniform distribution. Numerical results for the covariance matrices of the other distributions were published for sample sizes up to 5. Govindarajulu (1966) published order statistic covariances for the Laplace for sample sizes of 2 to 20 (increments of 1) and Barnett (1966) published those for the Cauchy for sample sizes of 5 to 16 (increments of 1) and 18 and 20.

Given the distribution, it is usually algebraically feasible to derive the covariance structure of the order statistics and thus the BLSS if the sample size is not too large.

The general problem in robust estimation, as pointed out above, is to devise an estimator which performs well across a range of distributions. If the class of distributions includes the Normal, Laplace and the Cauchy (or any subset of these three) and the Uniform distribution, it is impossible to formulate one estimator whose asymptotic efficiency to the

* where Beta $(\alpha,\beta) = \dfrac{x^{\alpha-1}(1-x)^{\beta-1}\ \Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$, $0 < x < 1$.

Cramér-Rao lower bound is not zero for one alternative. This follows from the fact that the efficient estimators for the Normal, Laplace and Cauchy have variances of the order of n^{-1}, whereas that of the Uniform (Rectangular) has variance of the order of n^{-2}. (This situation does of course give considerable justification for the use of adaptive schemes in robust estimation.)

Given such constraints, Crow and Siddiqui (1967) set out to determine the L-estimator which maximizes the minimum (over a class of distributions) of the relative efficiency (in general with respect to the BLSS) for a range of sample sizes. The class of distributions examined included the normal, double exponential, Cauchy, parabolic, triangular and uniform. Crow and Siddiqui discuss four distinct L-estimators with the notation: $p = \frac{1}{2} - \frac{r}{n}$, where r is a non-negative integer less than $\frac{n}{2}$:

a) Winsorized mean (Tukey, 1962),

$$W_n(p) = n^{-1}\left[(r+1)(X_{(r+1)} + X_{(n-r)}) + \sum_{i=r+2}^{n-r-1} X_{(i)}\right], \quad r < (n-1)/2$$

$$W_n(\tfrac{1}{2n}) = X_{(\alpha+1)} \quad \text{if} \quad n = 2\alpha+1 \quad (\alpha \text{ a positive integer}).$$

b) Trimmed means (Tukey (1962)),

$$T_n(p) = (n-2r)^{-1}\sum_{i=r+1}^{n-r} X_{(i)}$$

c) Linearly weighted means,

(1) $n = 2\alpha$

$$L_n(p) = (\tfrac{1}{2})(\alpha-r)^{-2}\left[X_{(r+1)}+X_{(n-r)}+3(X_{(r+2)}+X_{(n-r-1)})+\cdots\right.$$

$$+ (2i-2r-1)(X_{(i)}+X_{(n-i+1)}) + \cdots$$

$$+ (2\alpha-2r-1)(X_{(\alpha)}+X_{(\alpha+1)})\Big];$$

(2) $n = 2\alpha + 1$

$$L_n(p) = [(\alpha-r)^2+(\alpha-r+1)^2]^{-1}\Big[X_{(r+1)}+X_{(n-r)}+ 3(X_{(r+2)}+X_{(n-r+1)})$$

$$+\cdots+ (2i-2r-1)(X_{(i)}+X_{(n-i+1)}) + \cdots$$

$$+(2\alpha-2r-1)(X_{(\alpha)}+X_{(\alpha+2)})+(2\alpha-2r+1)X_{(\alpha+1)}\Big] .$$

d) Median and two other symmetric order statistics,

(1) $n = 2\alpha$

$$Y_n(p,a) = a(X_{(r+1)}+X_{(n-r)})+(\tfrac{1}{2}-a)(X_{(\alpha)}+X_{(\alpha+1)});$$

where a is a constant

(2) $n = 2\alpha + 1$

$$Y_n(p,a) = a(X_{(r+1)}+X_{(n-r)})+(1-2a)X_{(\alpha+1)} .$$

Of these a) and b) had received considerable attention in the literature and c) and d) were tentative proposals. The trimmed mean and winsorized mean have both been shown under certain conditions to be asymptotically normal (Bickel (1965)).

Stigler (1973) showed that necessary and sufficient conditions for the trimmed mean to be asymptotically normal could be stated alternatively as follows: "... *that the sample be trimmed at sample percentiles such that the corresponding population percentiles are uniquely defined.*"

Clearly, the only situations in which the above would not hold true, would be when sampling either from a discrete population or from a continuous population with gaps in the sample space. The standard deviation of the trimmed mean can be estimated (Tukey and McLaughlin (1963)) by:

$$S_{T_n(p)} = \sqrt{W_n^2(p)/h(h-1)}\ ; \qquad h = n-2r$$

where $W_n^2(p) = (r+1)\left((X_{(r+1)}-T_n(p))^2+(X_{(n-r)}-T_n(p))^2\right)+\sum_{i=r+2}^{n-r-1}(X_{(i)}-T_n(p))^2$.

Crowe and Siddiqui established the asymptotic normality of c) and d). Although Crowe and Siddiqui hint at the use of a more generalized weighting function than that in c) - "*one might consider weights varying as the k^{th} power of the subscript* ", they make the remark that weights derived from such a function of bounded variation are asymptotically unique as n becomes infinite (this allows them to establish asymptotic normality). As might be expected, the linearly weighted mean of Crowe and Siddiqui, which always weights central order statistics more than outlying ones, performs poorly with platykurtic distributions. In a similar vein to Crowe and Siddiqui's linearly weighted L-estimator, Stigler (1973) has proposed the use of a "smoothly trimmed mean":

$$n^{-1}\sum_{i=1}^{n} J(i/(n+1))X_{(i)}, \quad \text{with}$$

$$\begin{aligned} J(u) &= (u-\alpha)(0.5-\alpha)^{-1}, && \alpha \le u \le 0.5, \\ &= (1-\alpha-u)(0.5-\alpha)^{-1}, && 0.5 \le u \le 1-\alpha, \\ &= 0 , && \text{otherwise}, \end{aligned}$$

(α a constant less than 0.5),

for the case when the conventional trimmed mean fails the test of asymptotic normality.

It is worth noting incidentally that a mean of the trimmings of some trimmed mean (outmean) is not given serious consideration in the literature despite its excellent performance in the Stigler (1977) comparison. Similarly, the possibility of a weighting function which gives greater weight to the extreme order statistics than other more central ones is not given prominence.

The choice of r with L-estimators a) and b) is obviously crucial. For example, in the case of the calculation of a Winsorized mean of a sample with suspected outliers, too little Winsorization (too small an r) will result in overweighting outlier observations, whilst too large an r will result in the neglect of non-contaminating observations with consequent loss of information.

Crowe and Siddiqui show that for the distribution they studied (see above) the maximin (or guaranteed) efficiencies for estimators a), b) and c), (with p between 0.3 and 0.5) were fairly similar, the trimmed mean being the best on average although as p varies $(p = \frac{1}{2} - \frac{r}{n})$ the efficiency declines much less for $L_n(p)$ than for $T_n(p)$ or $W_n(p)$ for the Laplace and Cauchy distributions. Crowe and Siddiqui calculate guaranteed efficiency for their proposed estimator $Y_n(p,a)$ and show that the best $Y_n(p,a)$ (across p and a) is approximately as efficient as the best $W_n(p)$, $T_n(p)$ or $L_n(p)$ for the distributions they consider. $Y_n(p,a)$ does however tend to be somewhat sensitive to changes in the values of p and a. Crowe and Siddiqui recommend, as a rule of thumb, for calculating a robust estimate of the location parameter from a small sample (less than 20) with no prior information on the sampled distribution, that one use either $T_n(p)$ with $r = \frac{n}{4}$, or $L_n(p)$ with $r = \frac{n}{6}$ or $Y_n(p,a)$ with $r = \frac{n}{4}$ and $a = \frac{1}{4}$. Estimated variance (in the finite case) is then given by dividing a conventional sample estimate of the variance by the product of n and the guaranteed efficiency. Approximate confidence intervals are then easily calculated using the asymptotic normality of the estimators.

In a similar vein Gastwirth (1966) has proposed an estimator which is a linear sum of the $33\frac{1}{3}^{rd}$, 50^{th} and $66\frac{2}{3}^{rds}$ percentiles of the sample with weights 0.3, 0.4 and 0.3 respectively. Tukey in Andrews et al (1972) has proposed a weighted average of the first, second and third quartiles with weights $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ respectively.

Chan and Rhodin (1979) have extended the work of Crow and Siddiqui and examined the class of asymptotically best linear estimators based on k symmetrically ranked order statistics (symmetrical k-ABLES) where $k \leq 5$. The system of distributions considered consists of Tukey's lambda family and the normal, Laplace and Cauchy distributions. Tables are provided so that for any subset of this set of distributions the k-ABLE giving the highest guaranteed relative asymptotic efficiency may be obtained. Chan and Rhodin show that the optimal k-ABLE (as described above) compares favourably with optimal trimmed and Winsorized means for subsets of the distributions considered.

3.2 L-ESTIMATION OF THE "LOCATION PARAMETER" FOR UNSYMMETRICAL DISTRIBUTIONS

The problem of estimating central values of skewed distributions has not received much attention in the literature since such a parameter is clearly non-unique.

Hogg (1974) considers the possibility of an unsymmetrical trimmed mean estimator:

$$T_n(p_1,p_2) = (n-r_1-r_2)^{-1} \sum_{i=r_1+1}^{n-r_2} X_{(i)}$$

where
$$p_1 = \tfrac{1}{2} - \frac{r_1}{n}$$
$$p_2 = \tfrac{1}{2} - \frac{r_2}{n}$$

and r_1 and r_2 are non-negative integers less than n.

Statistical inferences can then be made about the expected value of such an estimator, say $\theta(p_1,p_2)$, if some approximate error structure can be derived. Hogg suggests an unsymmetrical Winsorized sum of squares:

$$S_{T_n(p_1,p_2)} = \sqrt{\omega_n^2(p_1,p_2)/h(h-1)}$$

where
$$\omega_n^2(p_1,p_2) = (r_1+1)(X_{(r_1+1)} - T_n(p_1,p_2))^2$$
$$+ (r_2+1)(X_{(n-r_2)} - T_n(p_1,p_2))^2$$
$$+ \sum_{i=r_1+2}^{n-r_2-1} (X_{(i)} - T_n(p_1,p_2))^2$$
$$- \left[r_1(X_{(r_1+1)} - T_n(p_1,p_2)) + r_2(X_{(n-r_2)} - T_n(p_1,p_2))\right]^2 / n$$

and $h = n - r_1 - r_2$.

of the parameters (such as a and p in the case of $Y_n(p,a)$) would be to make them a function of some measure of tail stretch.

Hogg (1967) has done this for the trimmed mean; he proposed:

$$T = \begin{cases} T_n^c(\tfrac{1}{4}), & \hat{k} < 2.0, \\ T_n(\tfrac{1}{2}), & 2.0 \le \hat{k} \le 4.0, \\ T_n(\tfrac{1}{4}), & 4.0 < \hat{k} \le 5.5, \\ T_n(0) & 5.5 < \hat{k}, \end{cases} \tag{3.3.1}$$

where $T_n^c(p)$ is the mean of the trimmings of $T_n(p)$ and $\hat{k}$ is an estimator of sample kurtosis.

Hogg (1972) then proposed a new indicator of tail stretch, namely:

$$Q = [\bar{U}(0.05) - \bar{L}(0.05)]/[\bar{U}(0.5 - \bar{L}(0.5)]$$

where $\bar{U}(\beta)$ is the average of the largest $n\beta$ order statistics (fractional items are used if $n\beta$ is not an integer) and $\bar{L}(\beta)$ has a similar definition using the smallest items.

Hogg (1974) then proposed:

$$T_1 = \begin{cases} T_n^c(\tfrac{1}{4}), & Q < 2.0 \\ T_n(\tfrac{1}{2}), & 2.0 \le Q \le 2.6 \\ T_n(\tfrac{5}{16}), & 2.6 \le Q \le 3.2 \\ T_n(\tfrac{1}{8}) & 3.2 < Q \end{cases}$$

In comments on Stigler (1977) Hogg raises the possibility of having an adaptive symmetrically trimmed mean estimator based on sample measures of skewness and kurtosis, and states that those based on sample skewness alone are probably better than ones based on sample kurtosis alone. He suggested:

$$T_2 = \begin{cases} T_n^c(\frac{1}{4}) , & (b_1)^{\frac{1}{2}} < 1.0 , \\ T_n(.1) , & 1.0 \leq (b_1)^{\frac{1}{2}} < 2.0 , \\ T_n(\frac{1}{4}) , & 2.0 \leq (b_1)^{\frac{1}{2}} , \end{cases}$$

where b_1 is a sample estimate of β_1, the squared skewness.

Harter (1972) has proposed an estimator of location based on classifying the sample as coming from either a uniform, normal or Laplace distribution - the appropriate maximum likelihood estimator is then used. He considers the scheme:

$$\begin{array}{lr} X_{\text{midrange}} , & \hat{k} < 2.2 , \\ \overline{X} , & 2.2 \leq \hat{k} < 3.8 , \\ X_{\text{median}} & \hat{k} \geq 3.8 \end{array}$$

This proposal is examined in more detail in Harter (1979) and several classification schemes based on sample kurtosis, Hogg's Q-statistic and sample likelihood are considered. Harter showed in this study that for estimation of the location parameter, criteria based on Hogg's Q-statistic performed best (marginally better than those based on sample kurtosis). Jaeckel (1971) has considered an adaptive trimmed mean, selecting that $T_n(p)$ which minimizes $\dfrac{\omega_n^2(p)}{(2p)^2}$ a statistic closely related to the variance of $T_n(p)$. This estimator performed well in the Princeton study for large $(n > 20)$ samples. Chan and Rhodin (see above) consider an adaptive form of the k-ABLE. By calculating:

$$p = \frac{F_n^{-1}(0.95) - F_n^{-1}(0.5)}{F_n^{-1}(0.75) - F_n^{-1}(0.5)} ,$$

where $F_n(k) = (\# X_i \leq k)/n$,

[Note the similarity to Hogg's Q-statistic.]

one may choose some subset from the class of distributions they consider, the numbers included in such a subset reflecting "*our uncertainty about the shape of the distribution*" (Chan and Rhodin). When this subset of distributions has been selected their tables yield guaranteed relative asymptotic efficiencies for each k-ABLE $(2 \leq k \leq 5)$. The k-ABLE with the highest guaranteed relative asymptotic efficiency is selected and tables reveal the relevant spacings and weights for the selected k-ABLE.

Hogg (1974) cites the work of Fisher (1972) who uses an adaptive asymmetrical trimmed mean estimation method for determining the distribution (out of a set of k) which has the largest mean. As her test statistic Fisher uses the (Q , Q_2) vector where Q is as defined above and:

$$Q_2 = [\bar{U}(0.05) - T_n(0.25)]/[T_n(0.25) - \bar{L}(0.05)] ,$$

$\bar{U}$ and $\bar{L}$ as defined above.

Essentially Fisher's conclusion for skewed data sets is that as the distribution becomes more skewed (to the right say), as measured by Q_2, the optimal value of p_2 becomes smaller. This conclusion ties in with the results of Sarhan above for his beta distribution.

In conclusion it is worth noting that adaptive L-estimators of location have received relatively little attention, one of their main drawbacks being, as Tukey (Andrews et al (1972)) puts it, that they only "*come into their own for fairly large samples, probably beyond* $n = 50$." It is argued that only then are they testably representative of their parent populations. Hogg (1974) has, however, stated that "*It is my personal opinion that these blatantly adaptive estimators can be constructed to be more effective than the nonadaptive robust estimators at much smaller sample sizes, say* $n = 15$ *or 20.*" One area of the development of these adaptive L-estimators of location to which attention is given in this thesis is that in which the weights are made continuous functions of the test statistic. Hogg (1974) has mentioned that such a scheme could have appeal.

3.4 L-ESTIMATION OF THE $\underline{\beta}$ VECTOR FOR THE REGRESSION MODEL

The extension of L-estimation to the regression case does not occur in an obvious way. Hogg (1979) has however considered the following scheme; since use of $\rho(x) = |x|$ (in the M-estimation case) as described above yields the sample median as an estimate of location and the median plane in the regression case, extension to other percentiles is made by taking:

$$\begin{aligned} \rho(y) &= -(1-p)y \;, & y < 0 \\ &= py \;, & y \geq 0 \end{aligned}$$

This yields the $(100p)^{th}$ percentile in the location case and thus the $(100p)^{th}$ percentile plane in the regression situation.

A possible extension, using this idea, of an L-estimator (proposed for the location case in Part II) to the regression situation is made in Part III.

CHAPTER 4

L_p-NORM ESTIMATION

In general the coefficients of the functional relationship (1.1) may be estimated by minimising the sum of the p^{th} powers of the deviations of the estimated values from the observed values of the dependent variable (L_p— approximation). As discussed above situations do arise in practice for which the use of ordinary least squares (L_2) is unrealistic. Alternatives to least squares, using L_p methods, such as minimising the sum of the absolute deviations (L_1— approximation), can be traced back to Fourier in 1820 and Edgeworth (1887; 1888). These earlier works did not engender much enthusiasm and although mild interest was revived again in the 1920 to 1940 period by Edgeworth (1923), Rhodes (1930) and Singleton (1940), it was only after the appearance of the papers of Karst (1958) and Wagner (1959) and the development of high speed computers that research in this field started to gather momentum.

4.1 L_p— APPROXIMATION

Suppose our model is of the form (1.1). In practice $n > m$ observations are taken and $\underline{\beta}' = (\beta_0, \beta_1, \ldots, \beta_{m-1})$ is estimated by that particular $\underline{b}' = (b_0, b_1, \ldots, b_{m-1})$ say $\underline{\tilde{\beta}} = (\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_{m-1})$ which minimizes the distance function:

$$\sum_{i=1}^{n} |y_i - \sum_{j=o}^{m-1} b_j x_{ji}|^p , \tag{4.1}$$

where $x_{oi} = 1$ for $i = 1,2,...,n$.

The above problem can be formulated as a mathematical programming problem as follows. Let the i^{th} error e_i be written as $e_i = u_i - v_i$ where $u_i \geq 0$ and $v_i \geq 0$; i.e. the variables u_i and v_i represent positive and negative deviations respectively for the i^{th} observation. Then the L_p approximation problem reduces to (Kiountouzis (1972)):

$$\text{minimize } \sum_{i=1}^{n} (u_i^p + v_i^p) \tag{4.2}$$

subject to

$$u_i - v_i + \sum_{j=o}^{m-1} b_j x_{ji} = y_i \quad (i = 1,2,...,n)$$

$$u_i \geq 0, \; v_i \geq 0 \quad (i = 1,2,...,n), \quad \text{and}$$

$$b_j, (j = 0,1,2,...,m-1) \text{ unrestricted in sign.}$$

It should be noted that this formulation is extremely flexible as it allows any observed constraint to be added (e.g. certain coefficients may be pre-specified to be non-negative or weights can be given to various errors). In particular, for prediction purposes Narula and Wellington (1977) have proposed the use of the minimization of the sum of the absolute relative errors. (i.e. $\Sigma_i |e_i/y_i|$.) Although their formulation is only for the case $p = 1$ it can easily be extended to other values of p.

Justification of L_p approximation comes from the following theorem:

Theorem 4.1.1 Kiountouzis (1971)

If our model for n observations takes the general form

$$y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_{m-1} X_{m-1i} + e_i \qquad (i = 1,2,\ldots,n)$$

and the random errors e_i satisfy the following conditions

(i) *The errors e_i are contained only in the results of the measurements y_i.*

(ii) *The errors e_i and e_j relating to any two different observations i and j are mutually independent.*

(iii) *The errors follow a distribution with p.d.f.*

$f(e) = h \exp\{-k|e|^p\}$, where h and k are constants and $1 \leq p \leq \infty$.

(iv) *No other information concerning the coefficients, β_j is available.*

Then the "best" $\underline{b}$ is a vector $\tilde{\underline{\beta}} \in E^{m-1}$ s.t.

$$\sum_{i=1}^{n} |y_i - \sum_{j=0}^{m-1} b_j x_{ji}|^p \quad \text{is a minimum.}$$

Thus for example the "best" value of $\underline{\beta}$ when e has a Laplace, normal and uniform distribution is the L_p-estimator with p equal to 1.0, 2,0 and ∞ respectively. For the model 1.2 ($m = 1$) these correspond to the maximum likelihood L-estimators median, mean and midrange respectively. There is thus an important link between L_p-estimation and L-estimation and this will be discussed at length in Part II.

It is important to stress here that only in the case when $p = 2$ can the L_p-estimator be written as an explicit function of the (unordered) $\underline{y}$ vector. Thus in general the $\underline{\beta}$ vector in the models (1.1) and (1.2) must be found by iterative methods. The problems involved in finding the $\underline{\beta}$ vector for the important cases of $p = 1$; $1 < p < 2$; ∞ are discussed below.

4.2 MINIMISATION OF THE SUM OF THE ABSOLUTE ERRORS (MSAE) (p = 1)

For $p = 1$ (4.2) becomes:

$$\text{minimise} \quad \textstyle\sum_{i=1}^{n}(u_i + v_i) \tag{4.3}$$

subject to

$$u_i - v_i + \textstyle\sum_{j=o}^{m-1} b_j x_{ji} = y_i, \quad (i = 1,2,\ldots,n)$$

$$u_i \geq 0, \quad v_i \geq 0, \quad (i = 1,2,\ldots,n), \quad \text{and}$$

$$b_j, (j = 0,1,2,\ldots,m-1) \quad \text{unrestricted in sign.}$$

It therefore follows that L_1 approximation can be formulated as a linear programming problem in $2n+m$ variables of which m variables (viz $b_0, b_1, \ldots, b_{m-1}$), are unrestricted in sign.

As with any linear programming problem a dual problem exists and can be formulated as follows (Wagner (1959)). Denote the dual variables by f_i. The dual of (4.3) is

$$\text{maximise} \quad \textstyle\sum_{i=1}^{n} f_i y_i \tag{4.4}$$

subject to

$$-1 \leq f_i \leq 1$$

$$\textstyle\sum_{i=1}^{n} f_i = 0$$

$$\textstyle\sum_{i=1}^{n} f_i x_{ji} = 0 \quad , \qquad (j = 1,2,\ldots,m-1).$$

Wagner (1959) also showed that by transforming with $w_i = f_i+1$ the dual problem may be formulated in terms of the non-negative variables w_i. This is computationally desirable and results in the following formulation:

$$\text{maximise } \sum_{i=1}^{n} w_i y_i \quad , \tag{4.5}$$

subject to

$$0 \leq w_i \leq 2 \quad ,$$

$$\sum_{i=1}^{n} w_i = n \quad ,$$

$$\sum_{i=1}^{n} w_i x_{ji} = \sum_{i=1}^{n} x_{ji} \quad , \quad (j = 1,2,\ldots,m-1)$$

Some properties of L_1 – estimation arising out of the linear programming formulation are now presented (Kiountouzis (1971), Appa and Smith (1973), Gentle, Kennedy and Sposito (1977)).

1. At least one L_1 hyperplane giving minimum sum of absolute deviations passes through r of the n points, where r is the rank of the observation matrix X. Usually X is of full rank and thus $r = m$, the number of coefficients to be estimated.

2. The solution to (4.1) is a hyperplane such that:

   $$|n^+ - n^-| \leq m$$

   where n^+ and n^- are the number of observations above and below the hyperplane respectively.*

3. Multiple optimal solutions can occur, i.e. two or more different hyperplanes give the same minimum sum of absolute deviations.

*If there does not exist a hyperplane which passes through more than m data points.

4. Variations in y do not change the optimal values of the coefficients as long as no observation crosses the optimal hyperplane. This property makes the L_1– estimator resistant to wild points.

5. Linear dependence among the independent variables will not cause any failures in the estimation procedure.

6. The L_1 hyperplane can be regarded as an estimate of the median of the conditional distribution of the $\underline{y}$ given the x's. This can be most easily seen by considering the case $m = 1$. Here the model becomes $\underline{y} = \underline{1}\beta_0 + \underline{e}$, and the L_1 estimator is the median of the set $y_1, y_2, \ldots, y_n$. For n odd it is equal to one of the y values, and for n even it lies on the closed interval between the two neighbouring middle values.

4.3 MINIMISATION OF THE MAXIMUM ABSOLUTE ERROR $(p = \infty)$

For $p = \infty$ the L_∞ minimisation criterion is equivalent to choosing the coefficients $\underline{b} = (b_0, b_1, \ldots, b_{m-1})$ as follows:

$$\underset{b_j}{\text{minimise}} \{\underset{i}{\text{maximum}} |y_i - \textstyle\sum_{j=0}^{m-1} b_j x_{ji}|\} \tag{4.6}$$

In 1799 Laplace proposed the above procedure, which was subsequently studied in detail by P.L. Chebychev and as a consequence is usually referred to as Chebychev approximation. A comprehensive account of the theory of Chebychev approximation is given in Rice (1964).

Denoting the maximum absolute deviation by D, the linear programming formulation of (4.1) is:

$$\text{minimise} \quad D \tag{4.7}$$

subject to

$$D \geq y_i - \sum_{j=o}^{m-1} b_j x_{ji} \,, \quad (i = 1,\ldots,n)$$

$$D \geq -y_i + \sum_{j=o}^{m-1} b_j x_{ji} \,, \quad (i = 1,2,\ldots,n)$$

where $D \geq 0$ and b_j, $(j = 0,1,2,\ldots,m-1)$ are unrestricted in sign.

The dual problem associated with the above can be formulated as follows (Wagner (1959)).

Let the dual variables corresponding to constraints $D \geq y_i - \sum_{j=o}^{m-1} b_j x_{ji}$ and $D \geq -y_i + \sum_{j=o}^{m-1} b_j x_{ji}$ be w_i and z_i respectively. Then the dual linear programming problem is:

$$\text{maximise} \quad \sum_{i=1}^{n} y_i(w_i - z_i) \tag{4.8}$$

subject to

$$\sum_{i=1}^{n}(w_i + z_i) = 1$$

$$\sum_{i=1}^{n}(w_i - z_i) = 0$$

$$\sum_{i=1}^{n}(w_i - z_i)x_{ji} = 0 \qquad j = (1,2,\ldots,m-1)$$

$$w_i \geq 0, \quad z_i \geq 0 \quad (i = 1,2,\ldots,n).$$

Some properties of L_∞ approximation arising out of the linear programming formulation of the problem and its associated dual are presented below (Kiountouzis (1971), Appa and Smith (1973)).

1. There exists one optimal hyperplane which is vertically equidistant from at least $m+1$ of the observations, the distance given by the optimal value of D.

2. The $m+1$ observations determining the optimal hyperplane must lie on the convex hull of the n observations.

3. The L_∞ hyperplane is a kind of mid-range estimate, which is most easily seen by considering the special case $m = 1$. Here the model is $\underline{y} = \underline{1}\beta_0 + \underline{e}$ and the L_∞ estimate becomes

$$\tilde{\beta}_0 = \tfrac{1}{2}(\min_i(y_i) + \max_i(y_i)).$$

4.4 OTHER VALUES OF p. (i.e. values other than $p = 1; 2;$ or ∞)

The mathematical programming formulation is given by (4.2). In general any value of $p \geq 1$ can be used but it is only in the special cases of $p = 1; 2$ or ∞ that an unique minimum for the L_p distance function can be found.*

For $p \neq 1; 2$ or ∞, iterative procedures and non-linear programming methods must be used (Fletcher and Powell (1963), Kiountouzis (1971), Forsythe (1972)). The inability to find an exact solution should not be regarded as a major disadvantage but as an encouragement to the development of "good" algorithms.

While theoretically any value of $p > 0$ can be used, in practice $p < 1$ is not of interest (Rice (1964)).

---

*If X is of full rank.

The traditional method is least squares $(p = 2)$ and in cases where this is not appropriate it is usually preferable (so as to avoid giving too much weight to "wild points") to use a value of p such that $1 \leq p < 2$. Values of $p > 2$ other than the special case where $p = \infty$ are not usually considered in the literature.

4.5 MAIN CONCLUSIONS OF LITERATURE ON L_p– ESTIMATION

Some of the main findings which have appeared in the literature on the more general properties of L_p– estimation are briefly summarised below.

(i) For a symmetric error distribution the L_p– estimates of $\underline{\beta} = (\beta_0, \beta_1, \ldots, \beta_{m-1})$ are unbiased for all values of p (Forsythe (1972), Kiountouzis (1973), Harvey (1978)).

(ii) It appears that as the tails of the error distribution become longer ("fatter') than those of the normal distribution, values of $p < 2$ provide better estimates than least squares (Blattberg and Sargent (1971), Forsythe (1972), Kiountouzis (1973), Harter (1977)).

(iii) For distributions with shorter tails than those of the normal distribution, it appears as if $p = \infty$ may be more appropriate (Harter (1977)). However for all other values of p the problem of determining the distribution of the $\tilde{\beta}$'s is extremely complex (being a nonlinear combination of y) and is perhaps mathematically intractable (Ashar and Wallace (1963), Havlicek (1968), Narula and Wellington (1977)).

(v) There are numerous algorithms for determining the estimates for the cases $p = 1$ and $p = \infty$ (e.g. Barrodale and Young (1966), Bloomfield and Steiger (1977)) which are superior to the equivalent Linear Programming solution as far as computer storage requirements and solution time are concerned.

4.6 ADAPTIVE L_p– ESTIMATION IN THE LOCATION PARAMETER AND REGRESSION CASE

L_p– estimation of $\beta_0(\theta)$, the location parameter, has received very little attention in the literature in comparison to L_p– estimation in the more general regression case. Harter has, however, proposed an adaptive scheme using L_p– estimation with p equal to 1.0, 2.0 or ∞, which he applied to the location parameter as well as to the regression situation. [His adaptive scheme for the location parameter case has already been discussed in Chapter 3 but will be repeated here for continuity.] For the location parameter case he used:

$$\begin{array}{lll} L_\infty & ; & \hat{k} < 2.2 \, , \\ L_2 & ; & 2.2 \le \hat{k} \le 3.8 \, , \\ L_1 & ; & \hat{k} > 3.8 \, , \end{array}$$

where $\hat{k}$ is the sample kurtosis of the original sample. Extrapolation to the regression situation was more complicated as the kurtosis of the residual vector has to be estimated. Originally Harter proposed the fitting of an L_2 – regression line in order to arrive at an initial set of residuals. [This method was in fact criticized by Hogg (1974) who suggested a more robust initial fit such as an L_1 – regression although in a rejoinder Harter states that "*one would hope that the final result will not be unduly influenced by the initial estimate* (of the kurtosis of the residuals) *unless it is an extreme one.*"]

Once the set of residuals has been determined the sample kurtosis is calculated and an L_p – regression computed according to the same scheme as that for the location parameter.

For the regression case Hogg (1974) has advocated a similar L_p adaptive approach but suggests the use of Q (or Q_1) [see Chapter 3] as a measure of tail stretch. He does not, however, restrict the values of p to three, but proposes that p "*be taken somewhere between 1.0 and 1.5 for distributions with long tails, around 2 for distributions with moderate tails, and greater than 4 (possibly* ∞) *for short-tailed distributions.*" In this respect, his ideas follow those of Forsythe (1972) who studied the estimation of β_0 and β_1 in a one dimensional regression model by L_p – methods with p = 1.25, 1.50 and 1.75. In Forsythe's study the distribution of the error vector was a standard normal contaminated with a normal with non-zero mean (μ) and standard deviation equal to 4.0 Forsythe found that for the symmetric case (μ = 0.0), lower values of p became more suitable as the contamination increased; this was also true, but to a greater extent, for the skewed case ($\mu = 4.0$). Forsythe concluded that a value of $p = 1.5$ could be a useful compromise for error distributions ranging from the normal to the Laplace.

The L_p– estimation of $\underline{\beta}$ vectors using an adaptive approach and the problem of such estimators distribution is as Hogg (1974) states a "*fruitful area for future research.*" The main thrust of this thesis is in this field and provides, hopefully, a body of research which is both useful in its own right and provides a solid platform for further work in this area.

PART II

I N T R O D U C T I O N

This section covers the work done on robust procedures for the estimation of the location parameter for symmetric and non-symmetric data sets. An adaptive L_p– procedure first developed for the regression case is proposed as a new estimator and the similarity of this method of estimation to L-estimation is considered in a theoretical and practical sense and an adaptive L-estimator is proposed. A study is then undertaken in which the adaptive L_p– and L – procedures are compared with a selected range of alternative robust estimators.

The problem of non-symmetric data sets is then examined and modifications of the procedures mentioned above are considered. Finally the performance of such a procedure on the data sets published by Stigler (1977) is examined and compared to the performance of the set of estimators used in that paper.

CHAPTER 1*

ESTIMATION OF THE LOCATION PARAMETER (θ) FOR SYMMETRIC DISTRIBUTIONS

1.1 INTRODUCTION

As exemplified in Part I an impressive array of robust estimators of location has been proposed in the literature. These estimators are primarily constructed so as to have superior statistical properties to the least squares estimator $\bar{X}$ when the underlying distribution deviates from normality. In particular, these estimators usually exhibit the characteristic of being less sensitive than $\bar{X}$ to outliers or bad-data points. In other words, these estimators are usually constructed to be "better" (according to some criterion) than $\bar{X}$ when the underlying distribution is leptokurtic with respect to the normal. Less attention has been given to the performance of estimators which outperform $\bar{X}$ when the underlying distribution is either platykurtic <u>or</u> leptokurtic with respect to the normal.

1.2 MEASURES OF LOCATION

---

*A paper based on parts of sections 1.1 and 1.5 of this chapter has been accepted for publication by the South African Journal of Statistics.

In the statistical literature numerous population parameters have been proposed as a measure of the location of the distribution. In particular, the mean $(E(X))$, median and midrange have received special attention. Hence, in estimating the location parameter, the first problem is to establish what one considers to be the most appropriate measure of location for the particular data set under consideration. Once the appropriate measure has been chosen the problem of estimation can then be tackled.

Initially, to avoid the subjective choice of the most suitable location parameter consideration was limited to symmetric populations so that:

$$E(X) = X_{median} = X_{midrange} = \theta .$$

1.3 THE L_p– METHOD

Harter and Hogg (see Part I) have proposed stepwise adaptive (according to tail stretch) procedures for the estimation of θ.

There does, however, seem no theoretical justification for using these stepwise adaptive procedures in place of a scheme in which there is a continuous trade off between tail stretch and estimator and such procedures have received little attention in the literature - see Prescott (1978), however, for a proposed adaptive trimmed mean estimator based on a continuous trade off between tail length and the estimator*.

*see also de Wet, T. and Van Wyk, J.W.J. (1979b)

Reconsideration of the Huber type approach (see Part I) shows that minimisation of:

(a) $\sum_{i=1}^{n}|x_i-\theta|$ yields $\hat{\theta} = \hat{X}_{median}$ (the maximum likelihood estimate of θ for the Laplace distribution).

(b) $\sum_{i=1}^{n}(x_i-\theta)^2$ yields $\hat{\theta} = \bar{X}$ (the maximum likelihood estimate of θ for the normal distribution).

(c) $\sum_{i=1}^{n}|x_i-\theta|^{\infty}$ yields $\hat{\theta} = \hat{X}_{midrange}$ (the maximum likelihood estimate of θ for the uniform distribution).

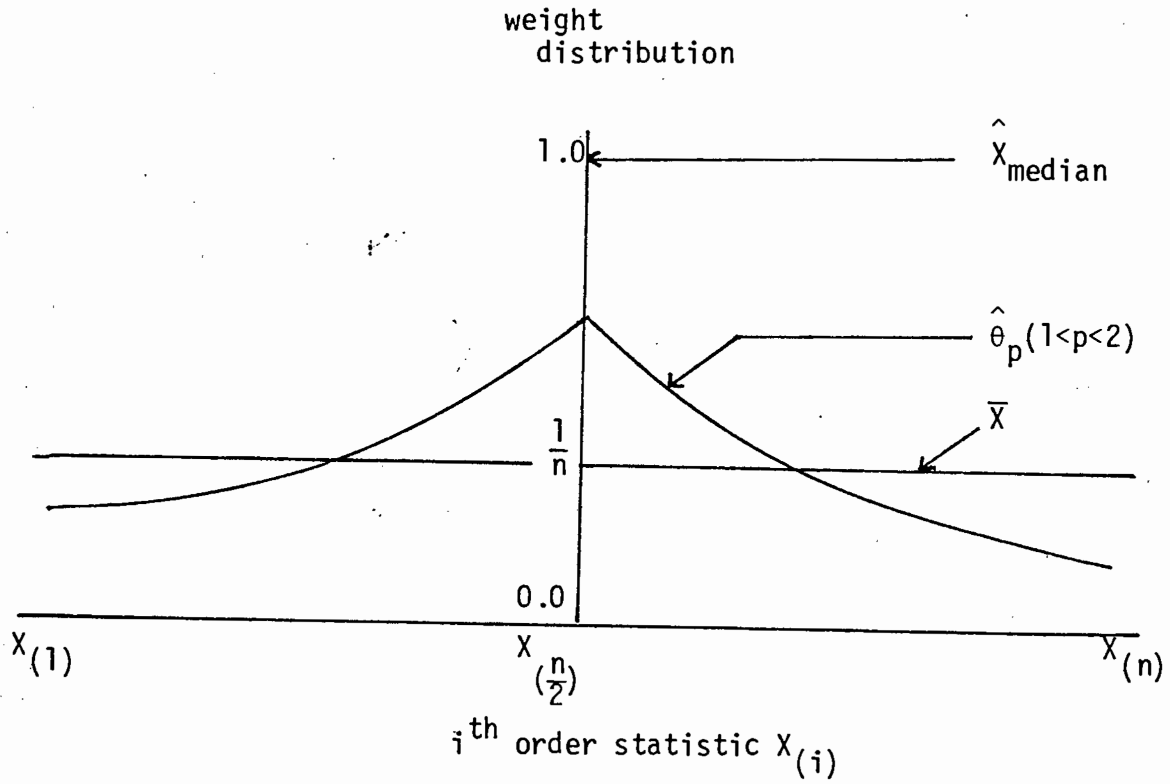
If we consider the minimisation of:

$$\sum_{i=1}^{n}|x_i-\theta|^p \qquad (1.3.1)$$

where p is a continuous function of tail stretch, it would appear that one may be able, at least in terms of maximising minimum efficiency, to improve on the blanket use of either $\bar{X}$ or $\hat{X}_{median}$ and even the use of M-estimators and adaptive L-estimators.

It is clear that using a value of $p = 2$ assigns equal weight to all observations in the calculation of $\hat{\theta}$; use of $p = 1$ assigns all the weight to the middle observation (in the case of n odd) and none to the others; use of $p = \infty$ assigns half weight to each of the end observations and none to the other. Use of $1 < p < \infty$ therefore assigns weight to all the observations; the closer p is to 1 the more the weights shift towards the central values (see Figure 1).

FIGURE 1



Such a procedure may have more intuitive appeal than the truncated mean or Huber approach because in those procedures information regarding the magnitude of data points falling in the tails is lost in the sense that the influence curves (Hampel (1974)) of such estimators are constant (non-zero) past the relevant cut off points. Thus although the existence of outliers influence such estimators and in fact as Hampel notes for the trimmed mean, exert "*the maximum possible influence on each side*" the actual values of such outliers in the tails is irrelevant.

The L_p– approach (except for $p = 1$ and $p = \infty$) assumes no cut off values and allocates weight to all the data points.

1.3.1 Selection of a suitable p based on tail stretch

In this section and throughout this part kurtosis $\left(\frac{\mu_4}{\mu_2^2}\right.$; where μ_i is the i^{th} moment about the mean$\left.\right)$ is used to measure tail stretch. Various alternatives have been proposed e.g. Hogg's Q statistic (Hogg (1974)) but it appears from the literature that none have any clear cut advantage over kurtosis which would render it more useful in the following studies. The use of alternative measures of tail stretch is a fruitful area for future research.

A formula relating the value of p and kurtosis which was based on a simulation study of the robust estimation of the parameters of a regression model with a symmetrical error distribution is established in Part III, Chapter 1. When the kurtosis is known, this has been shown to be superior to either L_2– or L_1–estimation in terms of maximin efficiency based on empirical generalized variance (over a range of distributions). This formula

$$p = 1 + \frac{9}{k^2} \qquad (1.3.2)$$

proposes the use of least squares under a normal distribution, and a continuous decrease in p for increasing k with a limiting value of $p = 1$.

In the general case, when k is unknown, use of the sample estimate of k, $\hat{k}$ instead of k has produced parallel comparative advantage over L_2 and L_1 regression (Part III, Chapter 2).

Estimation of the location parameter is merely a special case of the regression problem and hence the above procedure could be used. This would lead to the following procedure.

<u>The L_p Method of Estimating the Location Parameter of a Symmetric Distribution</u>

Minimise $\sum_{i=1}^{n} |x_i - \theta|^p$

where $p = 1 + \frac{9}{\hat{k}^2}$ (1.3.3)

and $\hat{k} = b_2 = 3 + \frac{k_4}{k_2^2}$ *

k_i being the unbiased estimate of the i^{th} cumulant. The comparative performance of this estimator will be examined in detail in the study of section 1.5.

1.4 ADAPTIVE L-ESTIMATION OF LOCATION WHEN THE UNDERLYING DISTRIBUTION IS SYMMETRIC

1.4.1 Introduction

It was noted above that L_p-norm estimates of the location parameter are closely related to symmetrically weighted

*See Appendix A.

combinations of the order statistics of the underlying data. Thus, for example, when $p = 1$ the estimator is simply the median (central order statistic) with all the other order statistics weighted at zero. When $p = 2$, the estimator is equal to the sum of the order statistics weighted by $\frac{1}{n}$; when $p = \infty$ the estimator is equal to the sum of the first and last order statistic each weighted by $\frac{1}{2}$.

It was thought that consideration of such L-statistics could serve as useful approximations to the L_p-norm estimator and give insight into the properties of the L_p-estimator. This relationship will be developed in Chapter 2.

The possibility of discovering properties of the L_p-estimator via their relationship with L-estimators led to a second study, namely the use of adaptive L-estimation as an estimation procedure in its own right.

Before considering a definite functional form for the weighting function of an adaptive L-estimator of θ, it is worth reiterating the required features of such an adaptive L-estimator. Essentially, that the optimal L-estimator of θ for long tailed distributions will involve weighting the central ordered sample elements more than the extreme elements, and vice-versa for short tailed distributions. When the underlying data is normal, the best linear estimator weights each element the same.

1.4.2 Establishment of the adaptive weighting distribution function

In this section the functional relationship between tail stretch and the optimal L-estimator of location is examined. As before, in order that the measure of the location parameter is unambiguous, we first consider only data sets from underlying distributions which are symmetric. A symmetric distribution implies symmetry of the matrix of covariances of the ordered sample and this will ensure that the minimum variance L-estimator of the location parameter is symmetrically weighted.

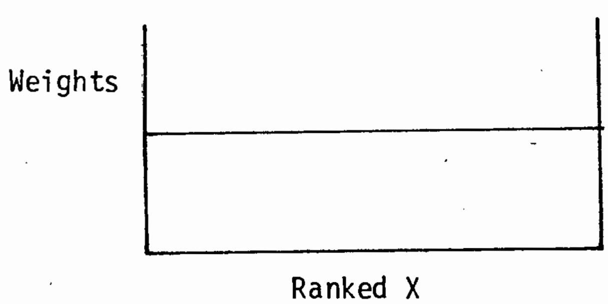
1.4.2.1 The beta weighting function

A weighting function which lent itself well to this study was the beta function:

$$\beta(p,q) = x^{p-1}(1-x)^{q-1} \quad , \quad 0 \leq x \leq 1.$$

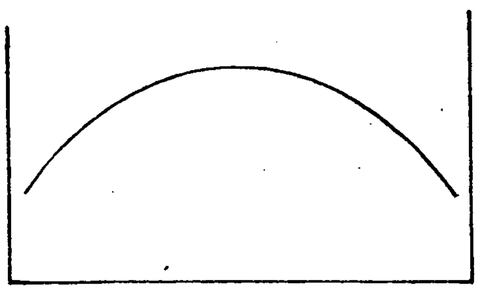
With p put equal to q three distinct forms of the function are easily recognised each with different implications for the weighting of the L-estimator.

(i) $p = q = 1$



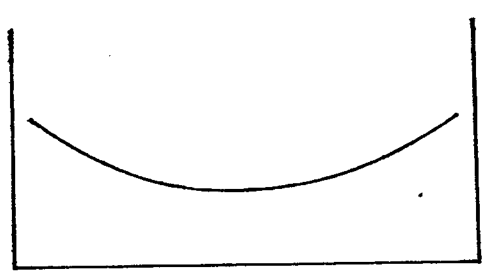
which implies equal weights to all elements of the sample.

(ii) $p = q > 1$



which implies greater weight given to central elements than extreme elements.

(iii) $p = q < 1$



which implies greater weight given to extreme elements than to central elements.

(This function has the advantage that, if required, asymmetrical weighting can be assigned with $p \neq q$ (see Chapter 3).)

1.4.2.2 The relationship between the beta function and sample kurtosis

The problem of the relationship between tail stretch and the optimal weighting function is very similar to the problem of the relationship between the optimal choice of p and tail stretch in L_p-norm estimation of the location parameter; namely, that although the optimal value of p or nature of the weighting function is known for certain distributions e.g. the uniform and the normal, the problem, in general, is to find some workable relationship for the case when no prior information is available on the nature of the distribution of the population. This relationship can really only be found by using a simulation study to examine the performance of certain relationships over a range of distributions.

Note that the case of negative weightings will not be considered - the case where the optimal L-estimator would have tail elements with negative weights would correspond to the optimal L_p-norm case with $p < 1$. Sarhan (1954) states, regarding the existence of a distribution for which such a weighting would be optimal: *"The author does not know any example at this time."* The case with negative weights in the middle and large positive weights in the tails presumably has no parallel in the L_p-norm case. Although theoretical symmetric distributions do exist for which such a weighting function would be optimal (which would have kurtosis greater than 1.0 and less than about 1.8 e.g. Sarhan's "U-shaped" distribution), they would be rare in practice.

Three distributions for which information about the optimal weighting distribution is known, are the uniform, normal and Laplace. The optimal parameters of the beta function for these three distributions are as follows:

(i) Uniform Distribution - beta weighting function with $p = q = -\infty$ with a theoretical kurtosis of 1.8.

(ii) Normal Distribution - beta weighting function with $p = q = 1$ with a theoretical kurtosis of 3.0.

(iii) Large kurtosis Distribution (e.g. Laplace) - beta weighting function with $p = q \to \infty$ (as $n \to \infty$ in the case of Laplace) with a large theoretical kurtosis (6.0 in the case of Laplace).

These three distributions give us definite guide lines as to the form of an adaptive L-estimator which utilizes the beta function to relate tail stretch and distribution of the weights. As before, kurtosis is used to measure tail stretch.

From the above the following relationship between sample kurtosis and $p = q$ is proposed:

$$p = q = 1 + \log_c\left(\frac{\hat{k}-1}{2}\right) \tag{1.4.1}$$

where $\hat{k}(b_2)$ is the sample kurtosis (see Appendix A) of the data set, and where c is a constant, the determination of which is discussed below. It is seen that the value of p is less than 1 for $\hat{k}$ less than 3, equal to 1 when $\hat{k}$ equals 3 and greater than 1 when $\hat{k}$ is greater than 3. Once the value of $p = q$ has been calculated the beta function is calculated at the points $\frac{1}{n+1}$ through $\frac{n}{n+1}$.

These weights are then scaled so as to sum to unity and assigned to the ranked data set.

A simulation study was carried out to establish a value of c for which (1.4.1) would perform well. A variety of statistical distributions were used to generate the raw data sets. All were symmetric and were chosen to cover a wide range of kurtosis. (For details, refer to Part III, Chapter 1). The sample size n (for each iteration of the simulation) used was chosen as 10, 30 and 50; and 100 iterations were preformed for each sample size and each distribution. The sample mean square error (MSE) was then computed over the 100 iterations.

Values of c around 1.2 were found to yield minimum sample MSE for the distributions with high and low kurtosis, and values of c around 2.6 were best for those distributions with kurtosis close to that corresponding to a normal distribution. Table 1.4.1 gives the results for $c = 1.2$(i) and $c = 2.6$(ii).

TABLE 1.4.1
SAMPLE MEAN SQUARE ERROR OF THE LOCATION PARAMETER ESTIMATES

Distribution	Kurtosis	Sample Size 10 (i)	10 (ii)	30 (i)	30 (ii)	50 (i)	50 (ii)
Uniform	1.8	0.802	0.920	0.053	0.174	0.024	0.116
Normal	3.0	0.939	0.891	0.349	0.301	0.219	0.159
Con.Normal	3.5	0.978	0.918	0.349	0.290	0.224	0.182
Con.Normal	4.0	0.936	0.875	0.308	0.287	0.202	0.173
Con.Normal	4.5	0.855	0.767	0.262	0.273	0.179	0.186
Con.Normal	5.0	0.739	0.813	0.174	0.213	0.123	0.157
Con.Normal	5.5	0.624	0.689	0.110	0.159	0.095	0.132
Laplace	6.0	0.941	0.988	0.281	0.299	0.147	0.173
Cauchy	-	0.129	0.661	0.022	0.050	0.011	0.019

On the basis of these results the following relationship between the beta parameter p and kurtosis was proposed.

$$p = q = \log_{1.2}\left(\frac{\hat{k}-1}{2}\right); \qquad \hat{k} > 4.0 \text{ or } \hat{k} < 2.0$$

$$p = q = \log_{2.6}\left(\frac{\hat{k}-1}{2}\right); \qquad 2.0 \le \hat{k} \le 4.0 \qquad (1.4.2)$$

This estimator with weights calculated using (1.4.1) and c calculated using (1.4.2), was then evaluated by calculating the sample MSE over the range of distributions considered above for 500 iterations. The simulated results for this estimator are given below:

TABLE 1.4.2
SAMPLE MEAN SQUARE ERROR OF THE
LOCATION PARAMETER ESTIMATES

Distribution	Kurtosis	Sample Size 10	30	50
Uniform	1.8	0.821	0.170	0.056
Normal	3.0	0.889	0.313	0.180
Con.Normal	3.5	1.112	0.293	0.185
Con.Normal	4.0	0.983	0.307	0.165
Con.Normal	4.5	0.864	0.293	0.143
Con.Normal	5.0	0.839	0.227	0.127
Con.Normal	5.5	0.703	0.163	0.091
Laplace	6.0	0.870	0.236	0.122
Cauchy	-	0.138	0.023	0.010

Since a rule which does not provide a continuous functional relationship between c and $\hat{k}$ (only a piecewise continuous relationship) may seem inappropriate, an attempt was made to derive such a function. Table 1.4.3 below gives the c for which sample MSE was a minimum for a 100 iterations for each of the set of distributions considered. A grid of c from 1.2 to 3.0 in increments of 0.2 was used.

TABLE 1.4.3

Distribution	Kurtosis	Sample Size 10	30	50
Uniform	1.8	1.2	1.2	1.2
Normal	3.0	3.0	3.0	3.0
Con.Normal	3.5	2.6	3.0	3.0
Con.Normal	4.0	3.0	2.8	2.8
Con.Normal	4.5	3.0	2.6	2.6
Con.Normal	5.0	1.2	1.2	1.2
Con.Normal	5.5	1.2	1.2	1.2
Laplace	6.0	1.4	1.2	1.2
Cauchy	-	1.2	1.2	1.2

Note that:

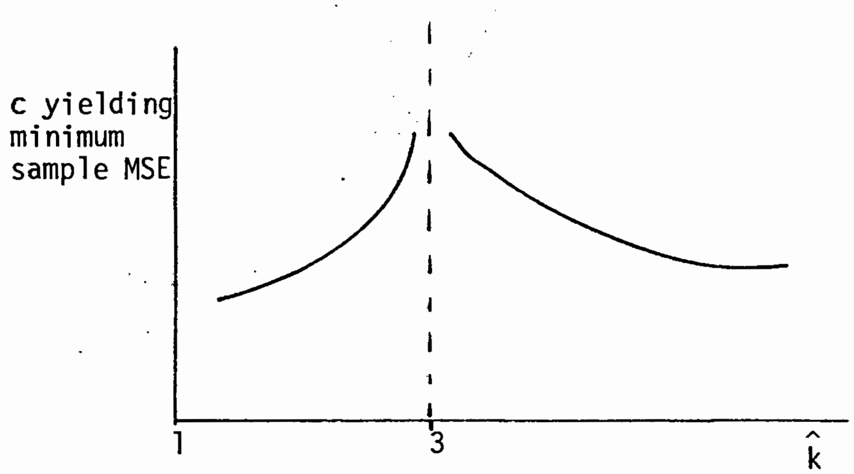
(i) There does not appear to be any marked difference in the c yielding minimum sample MSE for the different sample sizes of each distribution. Thus it seems that c may be expressed independently of n.

(ii) In theory the optimal value of c for the normal is ∞ yielding $p = q = 1$ in the formula (1.4.1) relating sample element weights to kurtosis, but in practice no improvement in MSE in the third decimal place was achieved for c larger than 2.6 and up to 10.0.

(iii) The decrease in the optimal value of c for a platykurtic distribution appears sharper than the decrease for a leptokurtic distribution with the same absolute deviation in sample kurtosis from 3.0.

[The reader is referred to Table 1.5.10 for an indication of how sample kurtosis varies with population kurtosis for the distributions considered here.]

The relationship between c and sample kurtosis yielding minimum MSE is thus in the following form:



After some trials and adjustments the following functional relationship was proposed:

$$
\begin{aligned}
c &= 1.05 && ; \quad \hat{k} \leq 1.75 \\
c &= 0.4 + \frac{1}{(3-\hat{k})^2} && ; \quad 1.75 < \hat{k} \leq 3 \\
c &= 1 + \frac{1}{(\hat{k}-3)^2} && ; \quad \hat{k} > 3
\end{aligned}
\tag{1.4.3}
$$

Using such a relationship the following results were obtained. (Table 1.4.4 below) for the sample MSE of the estimator of location using formula (1.4.3) in the same way as for Table 1.4.2.

TABLE 1.4.4

Distribution	Kurtosis	Sample Size		
		10	30	50
Uniform	1.8	0.804	0.150	0.041
Normal	3.0	0.887	0.309	0.178
Con.Normal	3.5	1.106	0.295	0.184
Con.Normal	4.0	0.992	0.311	0.167
Con.Normal	4.5	0.870	0.298	0.145
Con.Normal	5.0	0.853	0.234	0.132
Con.Normal	5.5	0.726	0.171	0.101
Laplace	6.0	0.891	0.240	0.128
Cauchy	-	0.143	0.022	0.010

It is seen that such a formulation does not on average represent an improvement over the original adaptive scheme (Table 1.4.2). Its large sample performance is however comparable and for this reason, and the superior appeal of the formulation, it is used in the comparative simulation study below (section 1.5). Although there is no doubt that a function exists for calculating a c which would represent an improvement over this for all sample sizes, it is conjectured that the effort involved in finding such an improvement would not be adequately rewarded.

1.4.3 Asymptotic Variance of proposed L-estimator.

Crowe and Siddiqui (1967) examined the asymptotic distribution of their proposed L-estimators (see Part I) for a range of distributions. Although they proposed weighting distribution functions which took a linear form, their theoretical exposition of the distributional properties of such estimators was more general with the weighting function $B(t)$ taking the following form:

$$\frac{dB(t)}{dt} = b(t) = \tfrac{1}{2}(s+1)^{-s-1}\,(\tfrac{1}{2}+p-t)^{s}\ , \quad s \geq 0$$
$$\tfrac{1}{2} \leq t \leq \tfrac{1}{2} + p \qquad (1.4.4)$$

where p is the proportion truncated in the right tail. In the case where no points are truncated $(p = \tfrac{1}{2})$ (1.4.4) reduces to:

$$b(t) = 2^{s}(s+1)(1-t)^{s}\ , \qquad \tfrac{1}{2} \leq t \leq 1 \qquad (1.4.5)$$

As the proposed L-estimating procedure does not exclude points, only functions with $p = \tfrac{1}{2}$ will be considered.

As an introduction to the methodology the derivation of the asymptotic distribution of the L-estimator using (1.4.5) will be sketched and then extended to the more general weighting function used in the L-estimator proposed in section

1.4.2. The case of the uniform distribution on $[-\frac{1}{2},\frac{1}{2}]$ is used to illustrate the procedures throughout because of the ease of algebraic manipulation. Owing to the symmetry of the distribution however we need only consider the right hand half.

This has distribution function:

$$f(-z) = f(z) = \begin{cases} 1 & ; \quad 0 \le z \le \frac{1}{2}, \\ 0 & ; \quad z > \frac{1}{2}, \end{cases}$$

and cumulative distribution function:

$$F(z) = 1 - F(-z) = \begin{cases} \frac{1}{2} + z & ; \quad 0 \le z \le \frac{1}{2}, \\ 1 & ; \quad z > \frac{1}{2}. \end{cases}$$

Define $\zeta(t) = F^{-1}(t)$

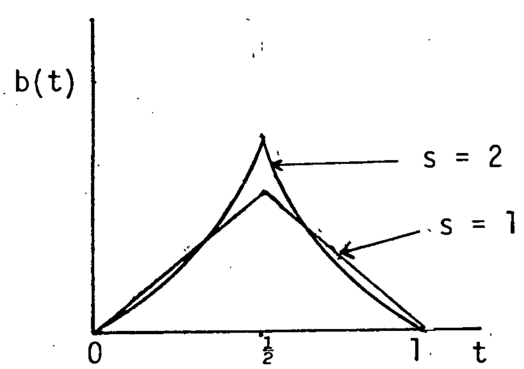
Thus:

$$1 - \zeta(1-t) = \zeta(t) = t - \tfrac{1}{2} \quad ; \quad \tfrac{1}{2} \le t < 1 .$$

We consider firstly the weighting function $B(t)$ with

$$\frac{dB(t)}{dt} = b(t) = 2^s(s+1)(1-t)^s \quad ; \quad \tfrac{1}{2} \le t \le 1, \quad s \ge 0$$

(Beta (1, s+1) distribution which is monotonically decreasing over $[\frac{1}{2}, 1]$, reflected about $\frac{1}{2}$.)



Note that the weightings decrease as one moves from the central values to the tails.

b(t) is continuous and symmetric about $t = \frac{1}{2}$ and as before we consider only the right hand interval $[\frac{1}{2}, 1]$.

Crowe and Siddiqui demonstrate that for such a weighting distribution the asymptotic variance:

$$\sigma^2 = 2 \int_0^x D\,(z) f(z)\, dz,$$

where

$$x = \zeta(1) \quad \text{and,}$$

$$D(z) = \int_0^z b\,(F(y))\, dy.$$

For the case in question (i.e. uniform) therefore:

$$D(z) = \int_0^z b\,(F(y))\, dy$$

$$= \begin{cases} \displaystyle\int_0^z 2^s\,(s+1)(\tfrac{1}{2}-y)^s\, dy & 0 \leq z \leq \tfrac{1}{2} \\ (= 2^s(-(\tfrac{1}{2}-z)^{s+1} + (\tfrac{1}{2})^{s+1}) & \\ 0 & \text{otherwise} \end{cases}$$

Now:

$$x = \zeta(1) = \tfrac{1}{2},$$

so $$\sigma^2 = 2 \int_0^{\frac{1}{2}} D^2(z)\, dz$$

$$= 2 \int_0^{\frac{1}{2}} \left[2^{2s}\, (\tfrac{1}{2}-z)^{2s+2} - 2(\tfrac{1}{2}-z)^{s+1} (\tfrac{1}{2})^{s+1} + (\tfrac{1}{2})^{2s+2} \right] dz$$

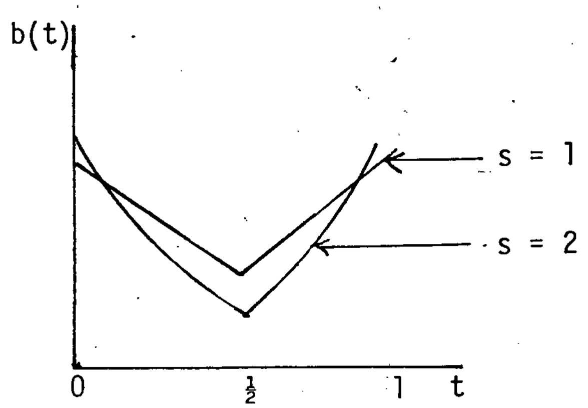
$$= 2\,{\Big[}_{0}^{\frac{1}{2}}\, 2^{2s}\left(-\frac{(\frac{1}{2}-z)^{2s+3}}{2s+3} + \frac{(\frac{1}{2})^{s}(\frac{1}{2}-z)^{s+2}}{s+2} + (\tfrac{1}{2})^{2s+2}\right)\Big]$$

$$= 2\left[2^{2s}\left((\tfrac{1}{2})^{2s+3} + \frac{(\frac{1}{2})^{2s+3}}{2s+3} - \frac{(\frac{1}{2})^{2s+3}}{s+2}\right)\right]$$

$$= \left[(\tfrac{1}{2})^{2} + \frac{(\frac{1}{2})^{2}}{2s+3} - \frac{(\frac{1}{2})}{s+2}\right]$$

$$= \frac{2s^{2} + 4s + 2}{4(2s+3)(s+2)}$$

The asymptotic variance of this form of estimator is thus at a minimum when $s = 0$ (flat weighting function). Since $B(t)$ (the anti-derivative of $b(t)$) and $D(z)$ are functions of bounded variation, this estimator will be asymptotically normally distributed (see Crowe and Siddiqui pp 366-376). Such a weighting function is clearly restrictive, so we consider one of the same family in which the outlying values are weighted more than the central ones (Beta $(s+1,1)$ on $[\frac{1}{2},1]$, reflected about $\frac{1}{2}$.)



We have:

$$b(t) = \frac{2^s}{2^{s+1}-1}(s+1)t^s \qquad ; \qquad \tfrac{1}{2} \le t \le 1$$
$$s \ge 0$$

$$D(z) = \int_0^z 2^s(s+1)(\tfrac{1}{2}+y)^s\, dy \qquad ; \qquad 0 \le z \le \tfrac{1}{2}$$

$$= 2^s((\tfrac{1}{2}+z)^{s+1} - (\tfrac{1}{2})^{s+1})$$

$$\sigma^2 = 2\int_0^{\frac{1}{2}} D^2(z)\, dz$$

$$= 2\int_0^{\frac{1}{2}} \frac{2^s}{2^{s+1}-1}\left[(\tfrac{1}{2}+z)^{2s+2} - (\tfrac{1}{2})^s(\tfrac{1}{2}+z)^{s+1} + (\tfrac{1}{2})^{2s+2}\right]$$

$$= \frac{2^{s+1}}{2^{s+1}-1}\ \frac{1}{2s+3}\left[\frac{-(\frac{1}{2})^s}{s+2}\ \frac{-(\frac{1}{2})^{2s+3}}{2s+3} + \frac{(\frac{1}{2})^{2s+2}}{s+2}\right]$$
$$+ (\tfrac{1}{2})^{2s+3}$$

$$= \frac{2^{s+1}}{2^{s+1}-1}\left[\frac{(\frac{1}{2})^s}{s+2}\left(\frac{2^{s+2}-1}{2^{2s+2}}\right) + \frac{1}{2s+3} + (\tfrac{1}{2})^{2s+3}\left(\frac{2s+2}{2s+3}\right)\right]$$

Since each of these terms decreases as s increases the minimum variance is obtained as $s \to \infty$ (midrange).

The proposed estimator (assuming $s > -1$) has the more general weighting function with:

$$b(t) = \frac{(t(1-t))^s}{(\Gamma(s+1))^2}(\Gamma(2s+2)) \qquad ; \qquad \tfrac{1}{2} \le t \le 1$$

Thus:

$$D(z) = \int_0^z \frac{(\Gamma(2s+2))}{(\Gamma(s+1))^2}((\tfrac{1}{2}+y)(\tfrac{1}{2}-y))^s\, dy \ ; \quad 0 \le z \le \tfrac{1}{2}$$

$$= \frac{(\Gamma(s+2))}{(\Gamma(2s+2))^2}\sum_{r=0}^{s}\binom{s}{r}(-1)^r\frac{z^{2r+1}}{2r+1}(\tfrac{1}{4})^{s-r}$$

This problem thus becomes algebraically complicated although in general it will be numerically tractable. Asymptotic normality is derived from the bounded $B(t)$ and $D(z)$.

It is conjectured that the problem of finding the asymptotic variance of the proposed estimator for a certain s will in general be numerically tractable for any known distribution. (See Crowe and Siddiqui's examples for the weighting function $2^s(s+1)(1-t)^s$ - pp 369-376.) In a later chapter the problem of finding the distribution for the finite case is tackled, essentially as a means of establishing the distribution of the L_p-estimator, because of the similarity between the two estimators.

1.5 A COMPARATIVE SIMULATION OF LOCATION PARAMETER ESTIMATORS FOR SYMMETRIC DISTRIBUTIONS

1.5.1 Introduction

In order to test the performance of the two new adaptive estimators proposed in this part, as against some well known alternatives, a simulation was conducted. The model simulated and estimated was the one dimensional model:

$$\underline{x} = \theta\underline{1} + \underline{e}$$

with $\underline{x} = \begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{pmatrix}$ a random unordered sample of n from a known distribution with $E(\underline{x}) = \theta\underline{1}$,

$$\underline{e} = \begin{pmatrix} e_1 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{pmatrix} \quad \text{a random unordered sample of } n \text{ from a known distribution with}$$

$$E(\underline{e}) = \underline{0}$$

$$E(\underline{e}\underline{e}') = \sigma^2 I_n$$

1.5.2 Design of the Simulation

A variety of statistical distributions were used to generate the raw data sets. All were symmetric and were chosen to cover a wide range of kurtosis. The sample sizes n (for each iteration of the simulation) used were chosen as 10, 30 and 50; and 500 iterations were performed for each sample size and each distribution. The basic criterion for evaluation of each estimator was sample MSE (the sample mean of the 500 squared deviations of the calculated values from the true value). Efficiencies (relative to the best estimator for each n and distribution) then can be calculated and the estimator may be evaluated according to its "best-worst-performance" or minimax efficiency.

Several distinct methods were used in the estimation of the location parameter θ all of which, except those discussed above have been discussed in Part I.

A. Least-Absolute deviation ($\hat{\theta}$ = sample median).

B. The adaptive L_p– method (1.3.3, Part II)

C. Ordinary Least squares ($\hat{\theta} = \overline{X}$).

D. Huber method* (a = 1.5).

E. Hampel method* (a = 1.7, b = 3.4 and c = 8.5).

F. Andrews method* (d = 2.1).

G. Trimmed mean (using Hogg's criterion (3.3.1, Part I)).

H. The adaptive L-method (Formulation 1.4.3, Part II).

The algorithm used for D, E and F was that due to Huber (1975) with the median used as a starting value and all the programs were written in UNIVAC (1100 series) DOUBLE PRECISION ASCII FORTRAN. Data sets were simulated from:

(a) Uniform distribution (kurtosis = 1.8).

(b) Normal distribution (kurtosis = 3.0).

(c) Contaminated Normal distributions (kurtosis = 3.5, 4.0 4.5, 5.0 and 5.5).

(d) Laplace distribution (kurtosis = 6.0).

(e) Cauchy distribution (kurtosis undefined).

(For details of the above distributions refer to Part III.)

Parameters were chosen so that each distribution (apart from the Cauchy) had $\theta = 0$ and $\sigma^2 = 9$. [The Cauchy distribution with parameters α and β has no moments but is symmetric about α which was chosen to be zero. β was determined by specifying that the 95th percentile of the Cauchy distribution had to coincide with the 95th percentile of the Normal (0,9) distribution.]

---

*D, E, and F were all applied using the scale invariant form due to Hampel (Hogg (1974)), refer to Section 2.1, Part I.

1.5.3 Experimental Results

As outlined above the aim of the simulation is to gain insight into the use of the new proposed L- and L_p-estimators of θ *vis-a-vis* the use of some of the more conventional robust estimators.

All the estimators considered, with the possible exception of B, are well known to be unbiased. Harvey (1978) has argued that the L_p-norm estimates are unbiased if the first moment exists and the underlying distribution is symmetric.

By considering sample MSE any conjecture on this point is avoided as this statistic gives us a joint measure of bias and variance.

The performance of the two proposed estimators, B and H, will be discussed separately. Tables 1.5.3, 1.5.6 and 1.5.9 exclude estimator H, which allows estimator B to be examined relative to the other estimators, excluding H. In considering estimator H separately it is not necessary to construct new efficiency tables because estimator B is nowhere 100% efficient. [Table 1.5.10 shows the average p used in estimator B.]

(i) Estimator B (excluding H).

Examination of the appropriate tables reveals that for a sample size of 10, B is, at worst, 48% efficient, whereas D is, at worst, 57% efficient, and F, at

TABLE 1.5.1

Sample Mean Square Error of the Location Parameter Estimates

N = 10 (500 iterations)

Distbn.	Kurtosis	A	B	C	D	E	F	G	H
Uniform	1.8	2.134	0.810	0.904	1.200	1.154	1.093	0.891	0.804
Normal	3.0	1.168	0.865	0.798	0.862	0.857	0.834	0.876	0.887
Con. Normal	3.5	1.172	1.076	1.044	1.016	1.014	1.016	1.102	1.106
Con. Normal	4.0	1.051	0.954	0.938	0.917	0.921	0.916	0.966	0.992
Con. Normal	4.5	0.792	0.851	0.884	0.753	0.770	0.782	0.859	0.870
Con. Normal	5.0	0.616	0.827	0.910	0.693	0.716	0.737	0.840	0.853
Con. Normal	5.5	0.336	0.691	0.783	0.524	0.560	0.600	0.676	0.726
Laplace	6.0	0.710	0.869	1.005	0.769	0.783	0.800	0.859	0.891
Cauchy	undefined	0.068	0.141	10.270	0.120	0.107	0.122	0.116	0.143
Average Sample Mean Square Error		0.894	0.787	1.948	0.762	0.765	0.767	0.798	0.808

TABLE 1.5.2

Efficiency of estimates (based on M.S.E.)

N = 10

Distbn.	Kurtosis	A	B	C	D	E	F	G	H
Uniform	1.8	38	99	89	67	70	74	90	100
Normal	3.0	68	92	100	93	93	96	91	90
Con. Normal	3.5	87	94	97	100	100	100	92	92
Con. Normal	4.0	87	96	98	100	99	100	95	92
Con. Normal	4.5	95	88	85	100	98	96	88	87
Con. Normal	5.0	100	74	68	89	86	84	73	72
Con. Normal	5.5	100	49	43	64	60	56	50	46
Laplace	6.0	100	82	71	92	91	89	83	80
Cauchy	-	100	48	1	57	64	56	59	48
Minimum Efficiency		38	48	1	57	60	56	50	46

TABLE 1.5.3

Efficiency of Estimates (based on M.S.E.) but excluding H.

N = 10

Distbn.	Kurtosis	A	B	C	D	E	F	G
Uniform	1.8	38	100	90	68	70	74	91
Normal	3.0	68	92	100	93	93	96	91
Con. Normal	3.5	87	94	97	100	100	100	92
Con. Normal	4.0	87	96	98	100	99	100	95
Con. Normal	4.5	95	88	85	100	98	96	88
Con. Normal	5.0	100	74	68	89	86	84	73
Con. Normal	5.5	100	49	43	64	60	56	50
Laplace	6.0	100	82	71	92	91	89	83
Cauchy	-	100	48	1	57	64	56	59
Minimum Efficiency		38	48	1	57	60	56	50

TABLE 1.5.4

Sample Mean Square Error of the Location Parameter Estimates
N = 30 (500 iterations)

Distbn.	Kurtosis	A	B	C	D	E	F	G	H
Uniform	1.8	0.798	0.200	0.312	0.355	0.333	0.337	0.254	0.150
Normal	3.0	0.457	0.313	0.297	0.318	0.312	0.304	0.314	0.309
Con. Normal	3.5	0.396	0.303	0.294	0.294	0.295	0.289	0.291	0.295
Con. Normal	4.0	0.354	0.310	0.304	0.296	0.297	0.294	0.315	0.311
Con. Normal	4.5	0.283	0.297	0.331	0.275	0.283	0.290	0.300	0.298
Con. Normal	5.0	0.205	0.227	0.285	0.214	0.225	0.236	0.236	0.234
Con. Normal	5.5	0.112	0.160	0.284	0.169	0.179	0.197	0.165	0.171
Laplace	6.0	0.183	0.227	0.304	0.230	0.237	0.243	0.236	0.240
Cauchy	undefined	0.021	0.024	2.590	0.033	0.029	0.032	0.023	0.022
Average Sample Mean Square Error		0.313	0.229	0.556	0.243	0.243	0.247	0.237	0.226

TABLE 1.5.5

Efficiency of estimates (based on M.S.E.)

N = 30

Distbn.	Kurtosis	A	B	C	D	E	F	G	H
Uniform	1.8	19	75	48	42	45	45	59	100
Normal	3.0	65	95	100	93	95	98	95	96
Con. Normal	3.5	73	95	98	98	98	100	99	98
Con. Normal	4.0	83	95	97	99	99	100	93	95
Con. Normal	4.5	97	93	83	100	97	95	92	92
Con. Normal	5.0	100	90	72	96	91	87	87	88
Con. Normal	5.5	100	70	39	66	63	57	68	65
Laplace	6.0	100	81	60	80	77	75	78	76
Cauchy	-	100	88	1	64	72	66	91	95
Minimum Efficiency		19	70	1	42	45	45	59	65

TABLE 1.5.6

Efficiency of Estimates (based on M.S.E.) but excluding H.

N = 30

Distbn.	Kurtosis	A	B	C	D	E	F	G
Uniform	1.8	25	100	64	56	60	59	79
Normal	3.0	65	95	100	93	95	98	95
Con. Normal	3.5	73	95	98	98	98	100	99
Con. Normal	4.0	83	95	97	99	99	100	93
Con. Normal	4.5	97	93	83	100	97	95	92
Con. Normal	5.0	100	90	72	96	91	87	87
Con. Normal	5.5	100	70	39	66	63	57	68
Laplace	6.0	100	81	60	80	77	75	78
Cauchy	-	100	88	1	64	72	66	91
Minimum Efficiency		25	70	1	56	60	57	68

TABLE 1.5.7

Sample Mean Square Error of the Location Parameter Estimates

N = 50 (500 iterations)

Distbn.	Kurtosis	A	B	C	D	E	F	G	H
Uniform	1.8	0.487	0.096	0.173	0.185	0.176	0.184	0.125	0.041
Normal	3.0	0.286	0.180	0.175	0.182	0.179	0.178	0.183	0.178
Con. Normal	3.5	0.241	0.183	0.185	0.179	0.179	0.179	0.186	0.184
Con. Normal	4.0	0.196	0.166	0.183	0.164	0.166	0.168	0.170	0.167
Con. Normal	4.5	0.148	0.144	0.177	0.136	0.141	0.148	0.148	0.145
Con. Normal	5.0	0.110	0.122	0.173	0.118	0.126	0.135	0.135	0.132
Con. Normal	5.5	0.075	0.095	0.211	0.108	0.114	0.132	0.086	0.101
Laplace	6.0	0.104	0.120	0.183	0.133	0.137	0.142	0.126	0.128
Cauchy	undefined	0.011	0.011	1.618	0.016	0.019	0.017	0.011	0.010
Average Sample Mean Square Error		0.184	0.124	0.342	0.136	0.137	0.143	0.130	0.121

TABLE 1.5.8

Efficiency of estimates (based on M.S.E.)

N = 50

Distbn.	Kurtosis	A	B	C	D	E	F	G	H
Uniform	1.8	8	43	24	22	23	22	33	100
Normal	3.0	61	97	100	96	98	98	96	98
Con. Normal	3.5	74	98	97	100	100	100	96	97
Con. Normal	4.0	84	99	90	100	99	98	96	98
Con. Normal	4.5	92	94	77	100	96	92	92	94
Con. Normal	5.0	100	90	64	93	87	81	81	83
Con. Normal	5.5	100	79	36	69	66	57	87	74
Laplace	6.0	100	87	57	78	76	73	83	81
Cauchy	-	91	91	1	63	53	59	91	100
Minimum Efficiency		8	43	1	22	23	22	33	74

TABLE 1.5.9

Efficiency of Estimates (based on M.S.E.) but excluding H.

N = 50

Distbn.	Kurtosis	A	B	C	D	E	F	G
Uniform	1.8	20	100	55	52	55	52	77
Normal	3.0	61	97	100	96	98	98	96
Con. Normal	3.5	74	98	97	100	100	100	96
Con. Normal	4.0	84	99	90	100	99	98	96
Con. Normal	4.5	92	94	77	100	96	92	92
Con. Normal	5.0	100	90	64	93	87	81	81
Con. Normal	5.5	100	79	36	69	66	57	87
Laplace	6.0	100	87	57	78	76	73	83
Cauchy	-	100	100	1	69	58	65	100
Minimum Efficiency		20	79	1	52	55	52	74

TABLE 1.5.10

Average p used in estimator B

With average sample kurtosis in italics
(500 iterations)

Distbn.	Kurtosis	Sample Size		
		10	30	50
Uniform	1.8	3.97 *2.21*	3.61 *1.92*	3.73 *1.85*
Normal	3.0	2.65 *2.94*	2.17 *3.01*	2.11 *3.02*
Con. Normal	3.5	2.47 *3.23*	1.96 *3.41*	1.91 *3.42*
Con. Normal	4.0	2.20 *3.54*	1.82 *3.76*	1.73 *3.89*
Con. Normal	4.5	2.16 *3.77*	1.68 *4.18*	1.59 *4.25*
Con. Normal	5.0	1.90 *4.24*	1.57 *4.66*	1.47 *4.83*
Con. Normal	5.5	1.81 *4.56*	1.42 *5.38*	1.36 *5.48*
Laplace	6.0	2.07 *3.98*	1.56 *4.88*	1.46 *5.25*
Cauchy	–	1.46 *7.33*	1.10 *6.91*	1.04 *26.46*

worst, 50% efficient. D and F thus marginally outperform B in terms of a maximin efficiency criterion for n equal 10. Using a average sample MSE criterion, B is outperformed by D, E and F. For sample sizes of 30 and 50, B is the best performer in terms of a maximin efficiency criterion and an average sample MSE criterion.

In addition B is the best global (across sample size) performer in terms of average sample MSE. It is seen that the relative superiority of B increases as the sample size increases. Tukey has commented (Princeton study) that adaptive statistics "*come into their own for fairly large samples, probably beyond 50*" and although the point is taken, it is seen that B's performance for n is 30 is still excellent and for n is 10, at worst, mediocre.

(ii) Estimator H (excluding B)

Estimator H stands out particularly for its good performance for the uniform distribution especially in large samples (in fact apart from the uniform distribution, the similarity with B is startling for n equal 50).

For a sample size of 10 it performs similarly to, but slightly worse than, B with a minimax efficiency of 46% and an average MSE of 0.808. For n equal to 30 and 50 it yields both the minimum average sample MSE value and the minimax efficiency (excluding B).

1.5.4 Conclusions for the simulation study

It is seen that both B and H do exceptionally well in this study in comparison to the class of estimators studied. They perform adequately for n equal 10 and extremely well for sample size 30 and 50. Given the fact that B and H perform comparably, and since H is much easier to compute, H will obviously be preferable if only limited computational facilities are available.

It should be noted in this study that all the non-normal distributions considered, except one, (the uniform) had kurtosis greater than 3.0. Thus possibly disproportionate weight has been given in this study to long tailed distributions. (This was in fact a criticism of the Princeton study - see Wegman and Carroll (1977).) If this had not been the case, one could assume from the above results that the relative performance of B and H would have been even better (since B and H were better than any alternatives for the uniform distribution in all three sample sizes examined.)

It can be argued therefore, on the grounds of the results obtained, that for the distributions considered, the adaptive L_p-norm or L-estimation methods are superior to blanket use of any of the alternative schemes for the estimation of the location parameter except in the case when the samples are very small.

1.5.5 Skewness and kurtosis of Sample Estimates for the 500 iterations

The tables provided are adequate <u>empirical</u> pointers (at least for the case n is 50) to the distributional properties of estimators B and H. (Tables 1.5.11 - 13.)

For the smaller sample sizes (10 and 30) it is probably unwise to draw any definitive distributional conclusions from the results given, but it is worth noting that neither B nor H exhibit deviations from normality which differ significantly from the deviations exhibited by C ($\overline{X}$).

Concentrating on the case n is 50, it is again seen that estimator B and H (apart from the uniform where it performs best) do not exhibit greater deviations on average from normality than does estimator C. As normality is in general assumed for $\overline{X}$ in applied statistical studies, it is felt that certainly no greater error would be made by making the same assumption for either B or H. If one could assume asymptotic normality, then confidence intervals could be constructed using the method of Crowe and Siddiqui, viz. given a class of distribution from which the data might possibly have been sampled, the variance of the estimator is obtained by "dividing an estimate of the distribution variance (such as the sample variance)" by the product of n and the guaranteed efficiency (minimax efficiency for the class of

TABLE 1.5.11

Kurtosis and Skewness (in italics) of Estimators

N = 10 (500 iterations)

Distbn.	Kurtosis	A	B	C	D	E	F	G	H
Uniform	1.8	2.63 *0.00*	4.10 *-0.01*	3.19 *-0.02*	3.33 *0.10*	3.58 *0.15*	3.61 *0.14*	3.88 *-0.04*	4.43 *-0.07*
Normal	3.0	3.16 *0.10*	3.14 *-0.03*	3.10 *0.00*	3.17 *0.03*	3.19 *0.03*	3.21 *0.06*	3.05 *-0.00*	3.19 *0.02*
Con. Normal	3.5	3.85 *-0.28*	3.52 *-0.05*	3.32 *0.00*	3.49 *-0.08*	3.50 *-0.07*	3.42 *-0.06*	3.54 *-0.05*	3.54 *-0.02*
Con. Normal	4.0	3.57 *0.11*	3.81 *0.13*	3.52 *0.09*	3.58 *0.07*	3.57 *0.07*	3.57 *0.08*	3.81 *0.10*	4.09 *0.16*
Con. Normal	4.5	3.58 *0.28*	3.59 *-0.08*	3.15 *-0.08*	3.26 *0.01*	3.22 *-0.03*	3.25 *-0.05*	3.47 *-0.02*	3.76 *-0.11*
Con. Normal	5.0	4.02 *0.03*	3.60 *0.10*	3.21 *0.05*	3.42 *-0.01*	3.40 *-0.03*	3.39 *-0.06*	3.77 *-0.09*	3.53 *0.08*
Con. Normal	5.5	4.39 *0.23*	4.50 *0.06*	3.19 *-0.01*	3.60 *-0.06*	3.46 *-0.06*	3.41 *-0.04*	4.44 *0.01*	4.61 *-0.06*
Laplace	6.0	4.14 *-0.29*	3.75 *-0.19*	3.05 *-0.11*	3.40 *-0.19*	3.43 *-0.18*	3.36 *-0.19*	3.69 *-0.26*	3.59 *-0.17*
Cauchy	-	3.86 *0.02*	17.94 *1.79*	15.83 *-0.40*	4.82 *0.05*	5.47 *0.07*	6.85 *0.33*	16.08 *1.80*	13.59 *1.33*

TABLE 1.5.12

Kurtosis and Skewness (in italics) of Estimators

N = 30 (500 iterations)

Distbn.	Kurtosis	A	B	C	D	E	F	G	H
Uniform	1.8	2.77 *-0.03*	3.48 *-0.02*	2.77 *-0.09*	2.95 *-0.05*	2.88 *-0.06*	2.79 *-0.08*	3.45 *-0.06*	5.30 *-0.11*
Normal	3.0	3.14 *0.10*	3.22 *0.23*	3.05 *0.15*	3.29 *0.24*	3.23 *0.22*	3.15 *0.20*	3.32 *0.19*	3.28 *0.20*
Con. Normal	3.5	2.61 *-0.04*	2.80 *-0.08*	2.75 *-0.06*	2.64 *-0.05*	2.66 *-0.05*	2.69 *-0.03*	2.72 *-0.07*	2.98 *-0.13*
Con. Normal	4.0	3.25 *0.01*	3.92 *0.08*	2.74 *0.10*	2.88 *0.12*	2.85 *0.11*	2.80 *0.11*	2.94 *0.07*	2.86 *0.08*
Con. Normal	4.5	2.69 *0.05*	3.20 *-0.13*	3.07 *-0.11*	3.11 *-0.04*	3.06 *-0.04*	3.04 *-0.06*	3.09 *-0.10*	3.12 *-0.15*
Con. Normal	5.0	2.78 *-0.02*	3.14 *-0.35*	2.92 *-0.25*	2.74 *-0.29*	2.77 *-0.30*	2.78 *-0.28*	3.37 *-0.39*	3.16 *-0.36*
Con. Normal	5.5	3.04 *-0.05*	3.96 *-0.30*	3.10 *-0.02*	3.47 *-0.01*	3.46 *0.00*	3.44 *0.01*	4.35 *-0.31*	3.96 *-0.25*
Laplace	6.0	3.18 *0.05*	3.62 *0.09*	3.18 *-0.06*	3.28 *-0.00*	3.29 *-0.03*	3.34 *-0.02*	3.72 *-0.05*	3.65 *0.03*
Cauchy	-	3.21 *0.16*	3.27 *0.21*	7.50 *-0.52*	3.37 *0.09*	3.23 *0.07*	3.15 *0.11*	3.75 *0.30*	3.49 *0.24*

TABLE 1.5.13

Kurtosis and Skewness (in italics) of Estimators

N = 50 (500 iterations)

Distbn.	Kurtosis	A	B	C	D	E	F	G	H
Uniform	1.8	2.78 *0.06*	3.37 *0.23*	2.99 *0.11*	3.16 *0.18*	3.05 *0.14*	3.02 *0.12*	3.79 *0.31*	9.96 *1.30*
Normal	3.0	3.00 *-0.02*	2.96 *0.09*	2.80 *0.06*	2.78 *0.07*	2.79 *0.08*	2.78 *0.07*	2.97 *0.11*	2.81 *0.08*
Con. Normal	3.5	2.86 *-0.09*	2.74 *0.07*	2.84 *0.08*	2.62 *0.02*	2.65 *0.02*	2.70 *0.04*	2.82 *0.09*	2.80 *0.08*
Con. Normal	4.0	2.96 *0.08*	3.05 *0.18*	3.00 *0.20*	3.14 *0.19*	3.12 *0.20*	3.06 *0.21*	3.02 *0.15*	3.05 *0.18*
Con. Normal	4.5	3.20 *-0.01*	3.09 *-0.06*	2.89 *0.02*	3.01 *-0.01*	3.01 *0.00*	2.98 *0.03*	3.11 *-0.04*	3.14 *-0.04*
Con. Normal	5.0	3.11 *0.21*	3.44 *0.22*	3.03 *0.09*	3.09 *0.01*	3.12 *0.03*	3.10 *0.07*	3.55 *0.16*	3.50 *0.16*
Con. Normal	5.5	3.19 *-0.01*	2.80 *0.04*	2.39 *0.04*	2.60 *0.05*	2.65 *-0.01*	2.60 *0.01*	3.14 *0.02*	2.91 *0.05*
Laplace	6.0	3.55 *0.18*	3.63 *0.05*	3.65 *0.07*	3.52 *-0.01*	3.51 *0.02*	3.56 *0.08*	3.45 *-0.03*	3.57 *0.01*
Cauchy	-	2.82 *-0.08*	2.97 *-0.08*	5.90 *-0.10*	2.92 *-0.10*	2.77 *-0.13*	2.78 *-0.15*	2.81 *-0.08*	2.83 *-0.13*

distributions for as wide a class of estimators as possible). "Approximate confidence intervals are then available from the asymptotic normality of the estimators."

CHAPTER 2

RELATIONSHIP BETWEEN L_p- AND L-ESTIMATORS IN FINITE SAMPLES

2.1 INTRODUCTION

The problem of determining the distribution of L_p-norm estimates of β parameters in the regression case, or θ in the location parameter case has not been considered to any great extent.

As was made clear in section 1.5 the motivation for the study of L-estimation was primarily its similarity to L_p-norm estimation which presented a way of at least approximating the distribution of the L_p-estimates. It was noted that an exact relationship between L_p-estimates and L-estimates in the case of the estimation of θ is provided when $p = 1$ (L-estimator the sample median), $p = 2$ (L-estimator the sum of the order statistics weighted by $\frac{1}{n}$) and $p = \infty$ (L-estimator the midrange). It is clear that for values of p between 1 and ∞ (excluding 2), the L_p-estimator will be some non-linear combination of the order statistics, weighted in such a way that the distribution of the estimator is symmetrical.

2.2 DERIVATION OF THE ADAPTABLE L-ESTIMATOR

The problem of relating the two estimators was first considered with a sample size of three, because in this case it is clear that a symmetrically weighted function of the order statistics can be constructed to coincide exactly with the L_p-norm estimator for the case of $p = 1.0$, 2.0 and ∞.

Working with sample sizes greater than 3 will raise questions as to the form of the weighting function (whether linearly or geometrically declining, for example). However consideration of the problem for sample size equal 3 will yield many valuable insights and test the feasibility of its generalization.

Given observations X_1, X_2, X_3 from a symmetrical distribution $f(X)$, define

Y_1, Y_2, Y_3 to be the ordered values of X.

We consider the distribution of:

$$u = kY_1 + (1-2k)Y_2 + kY_3 \;; \qquad 0 \leq k \leq \tfrac{1}{2}.$$

$k = 0$, u is the median,

when $k = \frac{1}{3}$, u is $\bar{X}$,

$k = \frac{1}{2}$, u is the midrange,

u is thus a symmetrical unbiased estimator of the $E(X)$.

Defining

$v = Y_2$

$w = Y_3$

we consider:

Case (i): X distributed as $U(0,1)$ [uniform distribution on [0,1]]

The mapping from $(Y_1, Y_2, Y_3) \to (u,v,w)$ is the mapping from the subsection of the unit cube enclosed by

$$\begin{array}{lcl} y_1 = 0 & & w = 1 \\ y_2 = y_1 & & v = w \\ y_2 = y_3 & \text{to that enclosed by} & u - (1-2k)v - kw = 0 \\ y_3 = 1 & & u - (1-k)v - kw = 0, \end{array}$$

the Jacobian of the transformation is $\frac{1}{k}$.

Therefore the joint density of u, v and w,

$$f(u,v,w) = \frac{3!}{k}$$

<u>Case (ii)</u>: X distributed as a Laplace with parameter λ . The mapping from $(Y_1, Y_2, Y_3) \to (u,v,w)$ is the mapping from the unbounded (positive and negative) region

$$\begin{array}{lcl} y_2 = y_3 & & v = w \\ y_2 = y_1 & \text{to} & u - (1-k)v - kw = 0 \end{array}$$

We establish that $f(u,v,w) = \frac{3!}{k\lambda} e^{-\lambda(|v|+|w|+|\frac{u-kw-(1-2k)v}{k}|)}$

<u>Case (iii)</u>: X distributed as a Normal distribution with parameters μ and σ. The mapping is as in (ii) and,

$$f(u,v,w) = \frac{3!}{k(2\pi\sigma)^{\frac{3}{2}}} e^{-\frac{1}{2}\left((\frac{v-\mu}{\sigma})^2 + (\frac{w-\mu}{\sigma})^2 + \left(\left(\frac{u-kw-(1-2k)v}{k} - \mu\right)/\sigma\right)^2 \right)}$$

To derive the distribution of u the variables v and w must be "integrated out".

For case (i) the problem splits into 2 major cases:

(a) $u \geq w(1-k)$

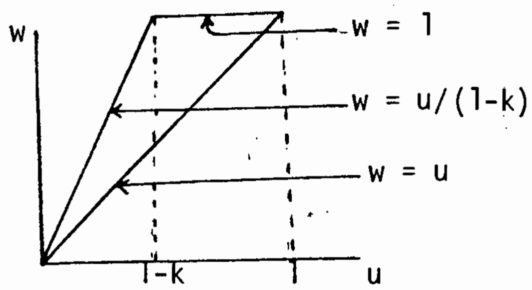


FIGURE 2.1

For this case we consider (see Figure 2.1)

(a_1) $1 \geq u \geq 1-k$, and (a_2) $0 < u \leq 1-k$

For (a_1),

$$f(u) = \int_u^1 \int_{\frac{u-kw}{1-k}}^{w} \frac{3!}{k} \, dv dw \tag{2.1}$$

For (a_2),

$$f(u) = \int_u^{\frac{u}{1-k}} \int_{\frac{u-kw}{1-k}}^{w} \frac{3!}{k} \, dv dw \tag{2.2}$$

(b) $u < w(1-k)$

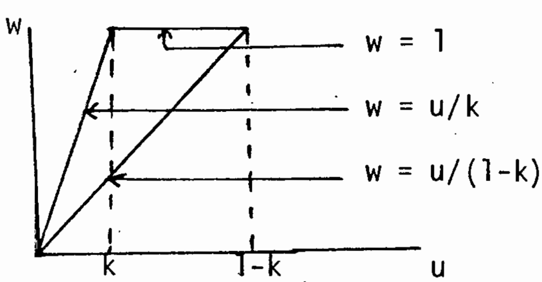


FIGURE 2.2

For this case we consider

(b_1) $k \leq u \leq 1-k$, and (b_2) $0 \leq u < k$.

For (b_1),

$$f(u) = \int_{\frac{u}{1-k}}^{1} \int_{\frac{u-kw}{1-k}}^{\frac{u-kw}{1-2k}} \frac{3!}{k}\, dvdw \tag{2.3}$$

For (b_2),

$$f(u) = \int_{\frac{u}{1-k}}^{\frac{u}{k}} \int_{\frac{u-kw}{1-k}}^{\frac{u-kw}{1-2k}} \frac{3!}{k}\, dvdw \tag{2.4}$$

For case (ii) the problem splits into 2 major cases as before:

(a) $u \geq w(1-k)$

Considering both negative and positive sections

$$f(u) = \int_{u}^{\frac{u}{1-k}} \int_{\frac{u-kw}{1-k}}^{w} \left[\frac{3!}{\lambda k}\left(e^{-\lambda(v+w+(\frac{u-kw-(1-2k)v}{k}))} + e^{\lambda(v+w+(\frac{u-kw-(1-2k)v}{k}))}\right)\right] dvdw$$

$$-\infty < u < \infty$$

(b) $u < w(1-k)$

$$f(u) = \int_{\frac{u}{1-k}}^{\infty} \int_{\frac{u-kw}{1-k}}^{\infty} \left[\frac{3!}{\lambda k}\left(e^{-\lambda(v+w+(\frac{u-kw-(1-2k)v}{k}))} + e^{\lambda(v+w+(\frac{u-kw-(1-2k)v}{k}))}\right)\right] dvdw$$

$$-\infty < u < \infty .$$

Case (iii) splits up as case (ii); for (a) it will be

$$f(u) = \int_{u}^{\frac{u}{1-k}} \int_{\frac{u-kw}{1-k}}^{w} \left[\frac{3!}{k(2\pi\sigma)^{\frac{3}{2}}}\left(e^{-\frac{1}{2}((\frac{v-\mu}{\sigma})^2+(\frac{w-\mu}{\sigma})^2+((\frac{u-kw-(1-2k)v}{k}-\mu)/\sigma)^2}\right)\right] dvdw$$

$$-\infty < u < \infty$$

Only case (i) (the uniform distribution) was considered in detail because of its relative algebraic tractability.

The integrals (2.1) through to (2.4) are evaluated for the relevant regions so that:

(I) For $0 \leq u < k$, (2.4) yields

$$f(u) = \frac{3(1-2k)}{k(1-k)^3} \cdot u^2,$$

and, (2.2) yields

$$f(u) = \frac{3k}{(1-k)^3} \cdot u^2$$

(II) For $k \leq u \leq 1-k$, (2.3) yields

$$\frac{6}{(1-k)(1-2k)} \cdot \left(u - \frac{k}{2} - \frac{(2-3k)}{2(1-k)^2} \cdot u^2\right)$$

and, (2.2) yields

$$\frac{6k}{(1-k)^3} \cdot u^2$$

(III) For $1-k < u < 1$, (2.1) yields

$$\frac{3(1-u)^2}{k(1-k)}$$

The density $f(u)$ is thus (in the case of the uniform distribution):

$$f(u) = \begin{cases} \dfrac{3}{k(1-k)}\, u^2 & , \quad 0 \leq u \leq k \quad , \\[2ex] \dfrac{3k}{1-k} & , \quad u = k \quad , \\[2ex] \dfrac{6}{(1-k)(1-2k)}\,(u(1-u)-\dfrac{k}{2}) & , \quad k < u < 1-k, \\[2ex] \dfrac{3k}{1-k} & , \quad u = 1-k \quad , \\[2ex] \dfrac{3}{k(1-k)}\,(1-u)^2 & , \quad 1-k < u < 1, \end{cases}$$

or alternatively in modulus form,

$$f(u) = \begin{cases} \dfrac{3}{k(1-k)}(\tfrac{1}{2}-|u-\tfrac{1}{2}|)^2 & , \quad 1 \geq 2|u-\tfrac{1}{2}| > 1-2k, \\[2ex] \dfrac{3k}{1-k} & , \quad 2|u-\tfrac{1}{2}| = 1-2k \quad , \\[2ex] \dfrac{6}{(1-k)(1-2k)} \cdot (u(1-u)-\dfrac{k}{2}), & \quad 0 \leq 2|u-\tfrac{1}{2}| < 1-2k. \end{cases}$$

Note that this density is continuous and differentiable over its domain except for the special case $k = 0$ when it is not differentiable at $u = \frac{1}{2}$.

For the boundary points we have:

At $u = k$, $f' = \dfrac{6}{1-k}$,

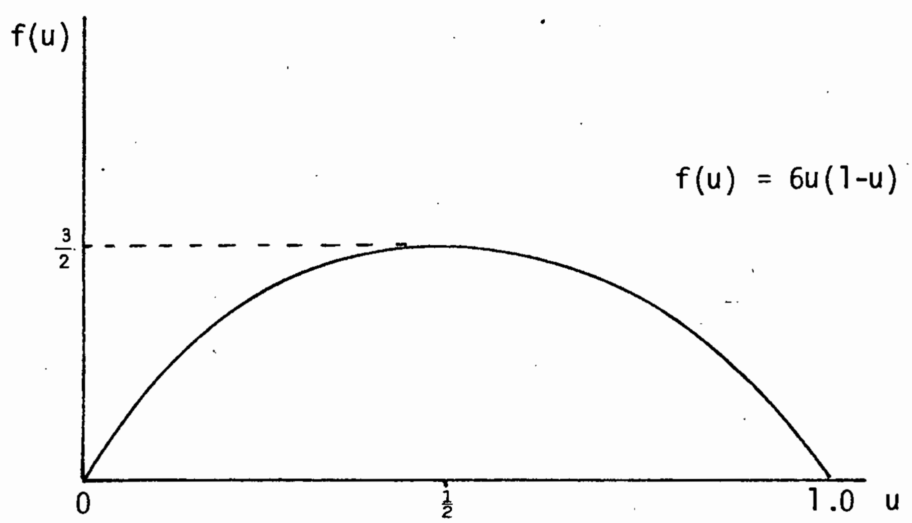
and at $u = 1-k$, $f' = \dfrac{-6}{1-k}$.

The cumulative distribution function of u; $F(u)$ is evaluated as:

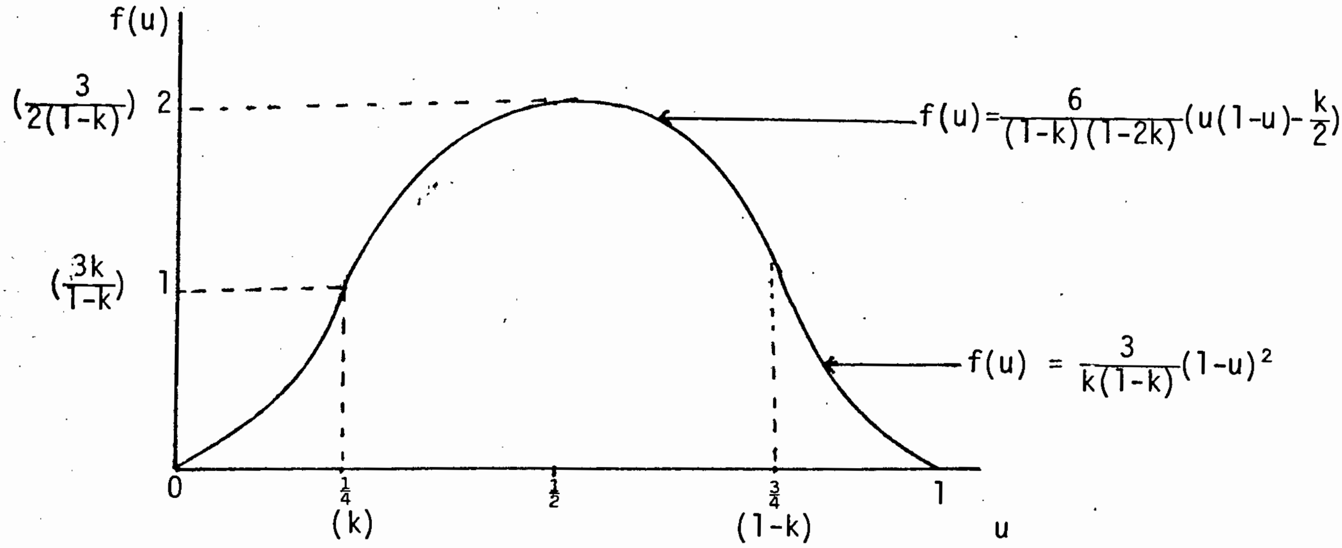
$$F(u) = \begin{cases} \dfrac{u^3}{k(1-k)} & , \quad 0 \le u < k \quad , \\[2ex] \dfrac{k^2}{(1-k)(1-2k)} + \dfrac{6}{(1-k)(1-2k)}\left(\dfrac{u^2}{2} - \dfrac{ku}{2} - \dfrac{u^3}{3}\right) & , \quad k \le u \le k-1 \; , \\[2ex] 1 - \dfrac{(1-u)^3}{k(1-k)} & , \quad 1-k < u \le 1 \; . \end{cases}$$

2.2.1 Shape of the density function for various values of k

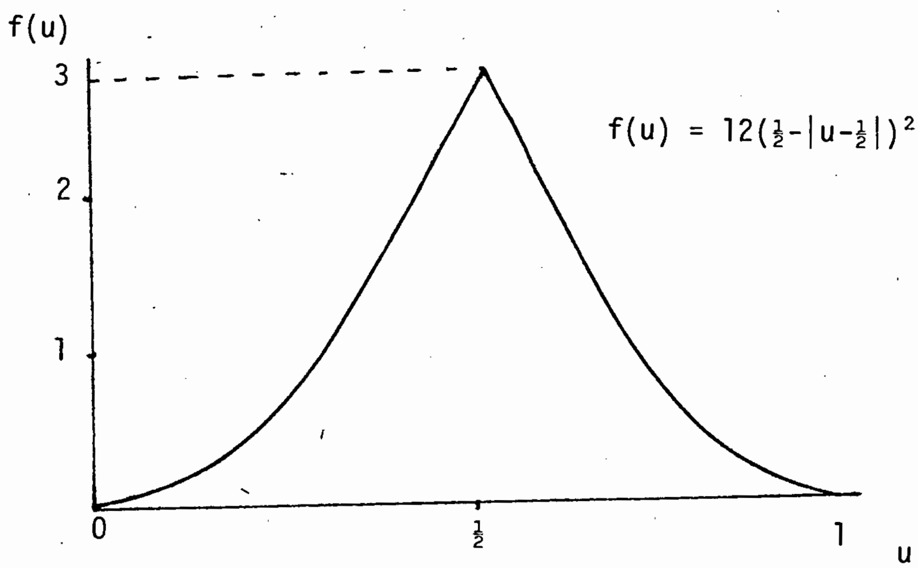
(i) $k = 0$ (the median)



(ii) $k = \frac{1}{4}$ $\quad (0 < k < \frac{1}{2})$



(iii) $k = \frac{1}{2}$ (the midrange)



Note that this family of curves is -

(i) symmetrical about $u = \frac{1}{2}$,

(ii) has ordinate $\frac{3k}{1-k}$ at $u = k, 1-k$, and

(iii) reaches its maximum of $\frac{3}{2(1-k)}$ at $u = \frac{1}{2}$.

2.3 MOMENTS OF THE DISTRIBUTION

The r^{th} moment of the distribution about the origin is defined as:

$$\mu_r' = \frac{3}{k(1-k)}\left[\int_{1-k}^{1} x^r (1-x)^2 dx + \int_0^k x^{2+r} dx\right]^{(1)}$$

$$- \frac{3k}{(1-k)(1-2k)}\left[\int_k^{1-k} x^r dx\right]^{(2)}$$

$$+ \frac{6}{(1-k)(1-2k)}\left[\int_k^{1-k} x^{r+1}(1-x) dx\right]^{(3)}$$

Expanding in powers of k yields the following coefficients of k^{r+s}:

For square bracket (1),

$$\frac{(-1)^{4+r}+1}{3+r} \qquad , \quad s = 3$$

$$(-1)^{r+s+1}\left(1 - \frac{2}{r+2}\right) \qquad , \quad s = 2$$

$$(-1)^{r+s+1}\left(\binom{r+3}{r+s}\cdot\frac{1}{r+3} - \frac{2}{r+2}\binom{r+2}{r+s} + \frac{1}{r+1}\binom{r+1}{r+s}\right) \qquad , \quad -r < s \le 1$$

$$0 \qquad s = -r$$

$$0 \qquad \text{otherwise}$$

As $f(u)$ varies from the midrange to the median it is seen that the kurtosis decreases from $\frac{20}{7}$ to $\frac{15}{7}$.

2.4 THE RELATIONSHIP BETWEEN k AND p

One important facet of this study was the way in which the above distribution can be used as a viable approximation to the distribution of the L_p-norm estimates of the location parameter of a uniform distribution with sample size 3.

A simulation exercise was carried out whereby for each of a grid of p values from 1.0 to 3.0 in increments of 0.01, 500 samples of size 3 from a uniform distribution were taken and the L_p-norm estimate of the location parameter of each sample calculated.

The value of k implied by each L_p-norm estimate of the location parameter was calculated for each sample. For any particular sample a plot of the implied weighting factor k against p gives a smooth monotonically increasing function. (The implied weighting factor k for a particular p is calculated by making k the subject of the formula in $u = ky_1 + ky_3 + (1-2k)y_2$ where u is the L_p- estimate of θ and y_i the i^{th} order statistic of the sample.) As expected, however, since the linear combination of order statistics is only an approximation, different samples give different weighting factors for the same p. The form of the average weighting factor as a function of p is as below:

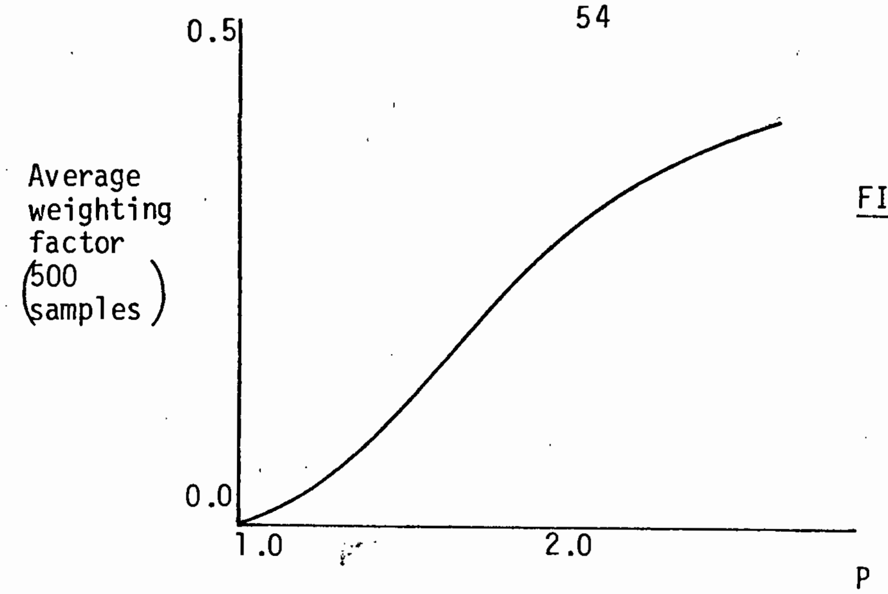


FIGURE 2.5

A plot of the standard deviation of the weighting factor for the 500 samples against p is of the following form, exhibiting the exact relationship for p = 1.0 and p = 2.0. As p moves away from these values the linear approximation becomes less accurate in the way shown.

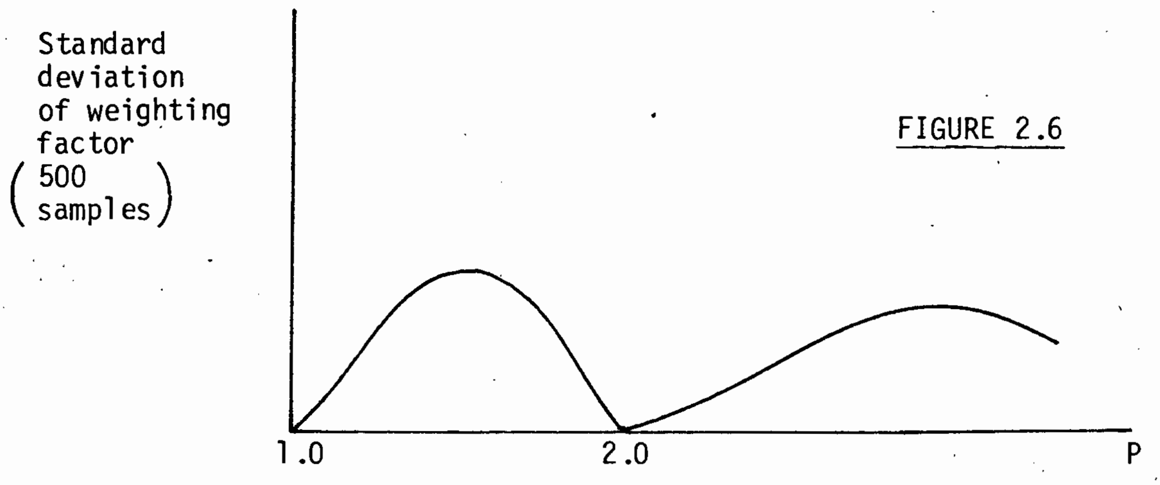


FIGURE 2.6

Unfortunately the issue is clouded to some extent by the fact that the algorithm producing the L_p-norm estimates (Fletcher and Powell (1968)) does not yield exact values but only iterates to a certain degree of accuracy. This makes it difficult to disentangle the problem of error in the algorithm from the accuracy of the weighted sum of order statistics as a proxy for the L_p-norm estimates.

Since we know that the L_p-norm estimates are unbiased (if the first moment exists and the underlying distribution is symmetric) particular interest is focussed on the variance of these estimates. If the variance structure of the L_p-norm estimator can be approximated, it would be an important step towards the construction of confidence intervals about θ for the L_p-estimator.

The simulation exercise above was used to examine the relationship between the variance structure of the two estimators in the following way. A plot of the estimated variance of the L_p-estimator against the average implied weighting factor for the 500 iterations was made for each p and it was found that this plot was almost identical to the plot of variance against k in Figure 2.3 with less than 4% deviation at each point. In fact, it is worth noting that the deviation between the two plots at k equal to 0 and $\frac{1}{3}$ (where they should coincide) was also of the same order. Since the deviation at these two points is due solely to error in the computational approximation of the

L_p-estimates, it is possible that the major part of the deviation between the two plots above is also due to the error in the L_p-estimation technique.

2.5 CONCLUSION

It seems feasible therefore, in the case of small samples, to use approximations to L_p-estimators in the form of linear functions of order statistics to give good estimates of variance for such estimators. The relationship between p and the form of the weights is probably best derived from a simulation. However, the derivation of the distribution of linear functions of order statistics for sample sizes greater than 3 is algebraically difficult and requires some subjective decisions about the distribution of weights. It is hypothesized however that certain weighting distributions will exist for all sample sizes which give estimators with very similar properties to those of the L_p-norm estimates. In addition there is tremendous scope for the examination of the finite distributional properties of weighted linear functions of order statistics in their own right, with the added attraction that the derivation of the distributional properties of these estimators appears tractable.

CHAPTER 3

ESTIMATION OF LOCATION FOR SKEWED DATA SETS

3.1 L_p-APPROACH

3.1.1 Introduction

This section considers the effect that skewed data sets have on the selection of the optimal p in the L_p-norm estimation of the location parameter. It is well known that there is a downward sloping relationship between tail-stretch and optimal p for symmetric distributions, and in Chapter 1 a specific form of this relationship was tested which showed it performed well *vis-a-vis* other conventional "robust estimators" for certain sample sizes and certain symmetric distributions. The relationship between skewness and optimal p in the estimation of the location parameter has not received much attention in the literature. Central to this lack of attention is the fact that, when skewed data is considered, no unique population value of the location parameter exists, and in the absence of *a priori* information on the suitability of a particular one, *vis-a-vis* the others, it is not clear which one should receive special attention.

Hogg in Stigler (1977) proposes an estimator based upon a measure of skewness $(\beta_1^{\frac{1}{2}})$ and suggests that kurtosis (β_2)

could also be incorporated. (The notation β_2 is used instead of the formerly used k to denote kurtosis to tie in with the notation of Johnson.) He states that his investigations lead him to believe that those based on skewness alone are better than those based on kurtosis alone. However a high value of $(\beta_1)^{\frac{1}{2}}$ is sufficient for a high value of β_2 since $\beta_2 \geq \beta_1 + 1$ (Johnson (1949)). So in some ways the form of the estimators might be closely related for skewed data sets. For symmetric distributions ranging from small to high kurtosis an adaptive estimator based on skewness alone would imply the same estimation scheme, and would presumably perform poorly relative to others which incorporate the degree of tail stretch.

3.1.2 Simulation Study

In order to gain insights into the influence of skewness and kurtosis, a simulation was designed to look at a large enough cross section of skewnesses and kurtoses to make their influence on the selection of p apparent.

It was assumed in this study that the expected value of the distribution represented the required measure of central tendency.

Use was made of the suite of programs written by Hill (1976) and Hill et al (1976) which make use of the Johnson S_U-S_B set of distributions (Johnson (1949)). A grid of skewnesses and kurtoses was examined with skewness ranging from -2.0 to +2.0 in intervals of 1.0, and kurtosis from 2.5 to 6.0 in intervals of 0.5 and 12.0 to 30.0 in intervals of 6.0, with the restriction that $\beta_2 \geq \beta_1 + 1$. (In the simulations β_2 was chosen so that $\beta_2 \geq \beta_1 + 2$.) The theoretical mean and variance were set at 0.0 and 9.0 respectively. Sample sizes of 10, 30 and 50 were examined. 500 iterations for each true skewness and kurtosis combination were performed for the sample size of 10, and 200 iterations for the sample sizes of 30 and 50. For each sample size and distribution the L_p-norm estimate of location was calculated for p in the range 1.0 to 3.0 at intervals of 0.1. The squared error was calculated for each iteration and the sample mean of the squared errors (MSE) over the total number of iterations was calculated. The values of p for which the MSE was a minimum for each case was observed and listed in Tables 3.1.1, 3.1.2 and 3.1.3. If two adjacent values of p gave very close values (differing in the third decimal for samples of size 10 and in the fourth decimal for samples of size 30 and 50) for the MSE, the average of the two p values was taken. In any study of this type, where the aim is to establish the relationship between skewness and kurtosis and optimal p, it is important to be aware of how close the <u>sample estimates</u> of these parameters are to the true (theoretical) values. For each iteration of each distribution and sample size, a sample estimate of skewness and kurtosis was calculated. The average of these statistics over the total number of iterations for each sample size are

TABLE 3.1.1

VALUES OF P FOR WHICH SAMPLE MEAN SQUARE ERROR IS A MINIMUM

N = 10 (500 samples)					
True Kurtosis	True Skewness				
	-2.0	-1.0	0.0	1.0	2.0
2.5		2.40	2.30	2.40	
3.0		2.25	2.00	2.25	
3.5		2.10	1.80	2.10	
4.0		1.95	1.65	1.95	
4.5		1.85	1.55	1.85	
5.0		1.77	1.50	1.80	
5.5		1.70	1.45	1.70	
6.0	1.90	1.50	1.45	1.65	1.85
12.0	1.65	1.35	1.20	1.35	1.65
18.0	1.40	1.15	1.15	1.20	1.45
24.0	1.35	1.10	1.10	1.10	1.35
30.0	1.15	1.05	1.00	1.05	1.15

TABLE 3.1.2

VALUES OF P FOR WHICH SAMPLE MEAN SQUARE ERROR IS A MINIMUM

N = 30 (200 samples)					
True Kurtosis	True Skewness				
	-2.0	-1.0	0.0	1.0	2.0
2.5		2.15	2.30	2.15	
3.0		2.10	2.05	2.10	
3.5		2.05	2.00	2.05	
4.0		2.00	1.85	2.00	
4.5		2.00	1.70	2.00	
5.0		1.95	1.60	1.95	
5.5		1.90	1.55	1.90	
6.0	2.00	1.85	1.50	1.85	1.95
12.0	1.85	1.55	1.35	1.50	1.85
18.0	1.70	1.40	1.30	1.40	1.70
24.0	1.60	1.30	1.25	1.35	1.60
30.0	1.45	1.30	1.25	1.30	1.40

TABLE 3.1.3

VALUES OF P FOR WHICH SAMPLE MEAN SQUARE ERROR IS A MINIMUM

N = 50 (200 samples)					
True Kurtosis	True Skewness				
	-2.0	-1.0	0.0	1.0	2.0
2.5		2.10	2.30	2.10	
3.0		2.05	2.00	2.05	
3.5		2.00	1.80	2.05	
4.0		1.95	1.70	2.00	
4.5		1.90	1.60	1.95	
5.0		1.90	1.55	1.90	
5.5		1.85	1.50	1.85	
6.0	1.95	1.80	1.50	1.85	2.00
12.0	1.85	1.55	1.35	1.55	1.85
18.0	1.75	1.40	1.30	1.45	1.75
24.0	1.65	1.35	1.25	1.40	1.65
30.0	1.55	1.30	1.25	1.35	1.60

listed in Tables 3.1.4, 3.1.5 and 3.1.6.

The estimate of β_2 (b_2) was calculated using:

$$b_2 = 3 + \frac{k_4}{k_2^2}$$

where k_i is an unbiased estimate of the i^{th} cumulant. Similarly the estimate of $\sqrt{\beta_1}$ used was:

$$b_1^{\frac{1}{2}} = \frac{k_3}{k_2^{\frac{3}{2}}}$$

On referring to Tables 3.1.4, 3.1.5 and 3.1.6 it will be noticed that for symmetric distributions with true kurtosis less than or equal to 3.0, on average the kurtosis is over-estimated, while for distributions with true kurtosis in excess of 3.0, on average the kurtosis is under-estimated with this being enchanced in small samples. For the case of skewed data a similar situation exists, except that the changes do not form such consistent patterns. Note, especially for large skewness in small samples, that there is a change in the pattern of skewness and kurtosis as one moves from the S_B to the S_U distribution (the boundary is the log normal line defined by the parametric equations:

$$\beta_1 = (w-1)(w+2)^2$$

$$\beta_2 = w^4 + 2w^3 + 3w^2 - 3,$$

(see Appendix B)and is indicated in Tables 3.1.4, 3.1.5 and 3.1.6 by the broken line). In practice, one obviously only has sample estimates of β_2 and $\beta_1^{\frac{1}{2}}$, and so cognizance must

TABLE 3.1.4

AVERAGE SAMPLE ESTIMATES OF SKEWNESS (IN ITALICS) AND KURTOSIS

N = 10 (500 samples)					
True Kurtosis	True Skewness -2.0	-1.0	0.0	1.0	2.0
2.5		*-1.19* 3.85	*-0.03* 2.91	*1.15* 3.71	
3.0		*-1.01* 3.66	*-0.03* * 3.21	*0.97* 3.53	
3.5		*-0.87* 3.62	*-0.04* 3.42	*0.82* 3.51	
4.0		*-0.76* 3.63	*-0.04* 3.57	*0.71* 3.53	**
4.5		*-0.68* 3.65	*-0.04* 3.69	*0.61* 3.56	
5.0		*-0.62* 3.70	*-0.04* 3.80	*0.55* 3.61	
5.5		*-0.57* 3.79	*-0.04* 3.88	*0.50* 3.71	
6.0	*-1.92* 6.65	*-0.54* 3.87	*-0.04* 3.96	*0.45* 3.80	*1.88* 6.48
			***		
12.0	*-0.93* 4.39	*-0.35* 4.41	*-0.05* 4.45	*0.26* 4.37	*0.85* 4.25
18.0	*-0.66* 4.58	*-0.29* 4.65	*-0.05* 4.68	*0.20* 4.63	*0.57* 4.51
24.0	*-0.55* 4.75	*-0.26* 4.80	*-0.05* 4.82	*0.17* 4.79	*0.46* 4.70
30.0	*-0.48* 4.87	*-0.24* 4.91	*-0.04* 4.93	*0.15* 4.90	*0.39* 4.84

* Normal

** S_B

*** S_U

TABLE 3.1.5

AVERAGE SAMPLE ESTIMATES OF SKEWNESS (IN ITALICS) AND KURTOSIS

N = 30 (200 samples)					
True Kurtosis	True Skewness -2.0	-1.0	0	1.0	2.0
2.5		*-1.09* 2.97	*-0.01* 2.69	*1.09* 3.06	
3.0		*-1.05* 3.39	*-0.02* * 3.12	*1.04* 3.42	
3.5		*-1.01* 3.74	*-0.03* 3.48	*0.98* 3.71	
4.0		*-0.96* 4.01	*-0.03* 3.77	*0.92* 3.91	**
4.5		*-0.90* 4.18	*-0.04* 4.02	*0.85* 4.03	
5.0		*-0.86* 4.31	*-0.05* 4.23	*0.79* 4.13	
5.5		*-0.82* 4.49	*-0.05* 4.42	*0.75* 4.29	
6.0	*-2.14* 7.46	*-0.79* 4.65	*-0.05* 4.58	*** *0.71* 4.43	*2.14* 7.60
12.0	*-1.41* 6.28	*-0.62* 5.84	*-0.08* 5.79	*0.47* 5.64	*1.31* 5.86
18.0	*-1.13* 6.60	*-0.54* 6.47	*-0.09* 6.42	*0.37* 6.29	*0.99* 6.22
24.0	*-0.99* 6.95	*-0.50* 6.88	*-0.09* 6.84	*0.31* 6.73	*0.83* 6.62
30.0	*-0.91* 7.25	*-0.47* 7.19	*-0.10* 7.15	*0.27* 7.06	*0.72* 6.94

* Normal

** S_B

*** S_U

TABLE 3.1.6

AVERAGE SAMPLE ESTIMATES OF SKEWNESS (IN ITALICS) AND KURTOSIS

N = 50 (200 samples)					
True Kurtosis	True Skewness -2.0	-1.0	0.0	1.0	2.0
2.5		*-1.03* 2.68	*0.01* 2.58	*1.06* 2.77	
3.0		*-1.01* 3.14	* *0.02* 3.05	*1.04* 3.23	
3.5		*-0.98* 3.55	*0.02* 3.46	*1.01* 3.64	
4.0		*-0.94* 3.86	*0.02* 3.81	*0.97* 3.96	**
4.5		*-0.90* 4.08	*0.03* 4.11	*0.93* 4.21	
5.0		*-0.85* 4.25	*0.03* 4.37	*0.89* 4.41	
5.5		*-0.82* 4.48	*** *0.03* 4.61	*0.86* 4.63	
6.0	*-2.06* 6.68	*-0.78* 4.69	*0.03* 4.81	*0.83* 4.82	*2.09* 6.93
12.0	*-1.46* 6.63	*-0.59* 6.30	*0.02* 6.31	*0.63* 6.28	*1.49* 6.79
18.0	*-1.17* 7.28	*-0.51* 7.13	*0.01* 7.12	*0.54* 7.08	*1.21* 7.23
24.0	*-1.03* 7.79	*-0.46* 7.69	*0.01* 7.67	*0.48* 7.63	*1.06* 7.69
30.0	*-0.94* 8.19	*-0.43* 8.12	*0.01* 8.09	*0.44* 8.05	*0.96* 8.07

* Normal   *** S_U
** S_B

be taken of the way the sample estimates perform for various theoretical values and different sample sizes.

3.1.3 Empirical relationship between optimal p and skewness and kurtosis

In section 1.5 it was shown that use of the relationship:

$$p = 1 + \frac{9}{(b_2)^2}$$

gave good results over certain distributions and sample sizes with symmetric distributions $(\beta_1^{\frac{1}{2}} = 0)$.

From examination of Tables 3.1.1, 3.1.2 and 3.1.3 for the case of $\beta_1^{\frac{1}{2}} = 0$, it is clear that such a rule would work adequately with these distributions. In fact, it seems that there is a sharper trade-off in smaller than larger samples so that improvement (for this family of distributions) could probably be made using:

$$p = 1 + \left(\frac{3}{b_2}\right)^q \tag{3.1.1}$$

with q a function of sample size. Examination of the empirical results suggested that we might use;

$$q = \frac{4.0}{N^{\frac{1}{4}}} ,$$

which gives q ranging from 2.25 for $N = 10$ to 1.5 for $N = 50$.

Plots of sample MSE against p portrayed further interesting features of the interrelationship between the sample and p. Firstly the above curves tended to be much steeper on both sides of the minimum for smaller sample size than for larger sample size. See Figure 3.1 below:

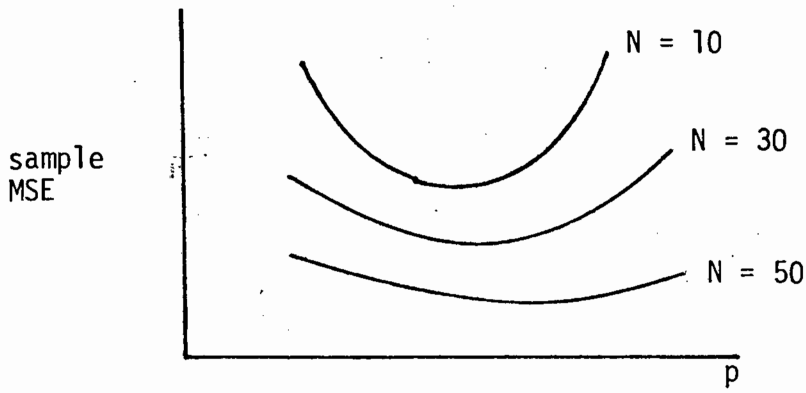


FIGURE 3.1

Therefore it is seen that global use of $q = 2$ *in (3.1.1) (i.e. Formula 1.3.3) will be an adequate proxy for the optimal* p *for the sample sizes considered, because loss of MSE in large samples (N = 50) will be small and* $q = 2$ *is near optimal in small samples.*

In addition to the size of sample effect on the shape of the curve, it was also evident that distributions with kurtosis close to the boundary $\beta_2 = 1 + \beta_1$ also yielded steeper curves for a given sample size. (See Figure 3.2 below.)

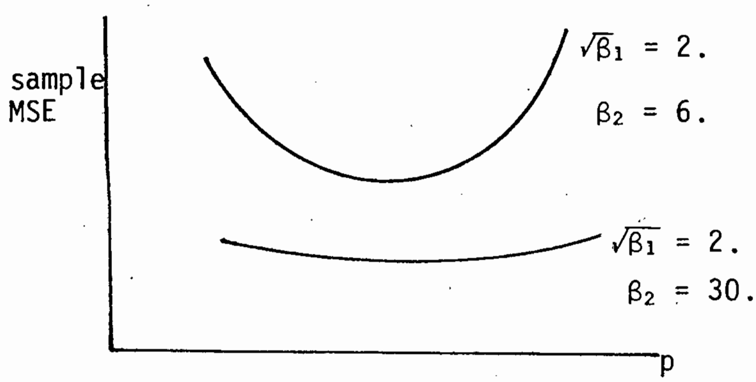


FIGURE 3.2

The results for the skewed data sets (Tables 3.1.1, 3.1.2 and 3.1.3 exhibit certain clearly defined effects. Namely:

(i) as skewness increases for any fixed kurtosis the optimal p gets closer to 2.0;

(ii) for any fixed skewness, increase in kurtosis yields a smaller optimal p.

(iii) As above, in the symmetric case, these changes are less marked in larger than in smaller data sets.

What does appear to be in evidence is that the excess of kurtosis over skewness squared plus one is a significant factor in the determination of an optimal p. That is, the excess of that amount by which kurtosis must necessarily (mathematically) exceed skewness.

It appears, essentially, that given increases of kurtosis in excess of skewness squared plus one give similar decreases in optimal p for given different skewnesses in the same sample size.

It appears that a relationship of the form:

$$p = 1 + \left(\frac{3}{b_2 - b_1}\right)^q \tag{3.1.2}$$

or

$$p = 1 + \left(\frac{3+b_1}{b_2}\right)^q \tag{3.1.3}$$

could prove workable.

The optimal q appears to be a function of sample size and sample kurtosis and possibly even sample skewness in the following way.

Larger sample size seems to effect q negatively in a similar way to the symmetrical case. Lower sample kurtosis distributions require higher values of q and to some extent a higher value of the absolute value of sample skewness required smaller q.

3.1.4 The use of L_p-estimation to establish higher moments of the underlying distribution

At this point, we consider an interesting application of L_p-estimation of skewed data sets to the more general problem of the estimation of the moments of a distribution. We consider, in the first case, the problem of calculating an estimate of the variance (V) of the underlying distribution from some random sample. If the data is from a normal distribution the maximum likelihood estimate of V is obtained by minimizing:

$$\sum_{i=1}^{n} |(x_i - \hat{\theta})^2 - V|^2 \tag{3.1.4}$$

yielding $\hat{V} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\theta})^2$

It might be conjectured that under deviations from normality, minimisation of

$$\sum_{i=1}^{n} |(x_i-\hat{\theta})^2-V|^p, \tag{3.1.5}$$

where $\hat{\theta}$ is obtained from minimisation of $\sum_{i=1}^{n}|x_i-\theta|^p$, and p is some function of kurtosis in each case, may yield superior estimates in terms of mean square error than blanket use of (3.1.4).

However, even if the underlying X is from a symmetric distribution the distribution of $(X-\theta)^2$ (and all such even powered transformations) will be non-symmetric and thus the problem of the selection of p will probably involve the skewness, as well as the kurtosis of the transformed data set. (It is worth noting here that use of the variance is an often used, but not unique, measure of dispersion. The performance of the L_p— approach with alternatives such as the mean absolute deviation has however still to be investigated.)

In the minimisation of (3.1.5), some success was achieved by using:

$$p = (3 + \frac{3}{b_2})^{\frac{1}{2}} \tag{3.1.6}$$

as a criterion for the selection of p in the second stage of the operation ($\hat{\theta}$ in (3.1.5) was still calculated using the estimator B of Chapter 1) when compared with the use of p equal to 2. It is noted that (3.1.6) is much less sensitive to changes in b_2 than the sample equivalent of

(1.4.1). Particularly encouraging results were achieved for the exponential distribution (which is non-symmetrical) and these results, along with results for the distributions studied in Chapter 1, are presented. The Cauchy distribution was excluded because the variance is not defined. The simulation study (100 iterations) conducted here only compared the proposed estimator with the L_2 estimator.

TABLE 3.1.7
SAMPLE MEAN SQUARE ERROR OF ESTIMATES
OF VARIANCE (100 ITERATIONS)

Distbn.	Kurtosis	n = 10		n = 30		n = 50	
		L_p	L_2	L_p	L_2	L_p	L_2
Uniform	1.8	8.416	8.164	2.225	2.064	1.541	1.431
Normal	3.0	19.795	20.036	6.294	6.155	3.569	2.971
Con.Normal	3.5	15.646	15.409	6.563	7.326	4.111	4.304
Con.Normal	4.0	21.188	22.646	7.731	7.617	5.257	5.719
Con.Normal	4.5	22.117	22.533	10.025	9.433	6.924	5.238
Con.Normal	5.0	25.142	31.670	12.137	13.190	7.777	7.354
Con.Normal	5.5	28.948	31.579	11.715	11.831	7.942	7.426
Laplace	6.0	51.386	68.489	10.075	13.041	8.332	8.907
Exponential	9.0	46.990	75.406	13.927	19.871	9.385	11.337

TABLE 3.1.8

EFFICIENCY OF ESTIMATORS (BASED ON MSE)

Distbn.	Kurtosis	n = 10		n = 30		n = 50	
		L_p	L_2	L_p	L_2	L_p	L_2
Uniform	1.8	97	100	93	100	93	100
Normal	3.0	100	99	98	100	83	100
Con.Normal	3.5	98	100	100	90	100	96
Con.Normal	4.0	100	94	99	100	100	92
Con.Normal	4.5	100	98	94	100	76	100
Con.Normal	5.0	100	79	100	92	95	100
Con.Normal	5.5	100	92	100	99	94	100
Laplace	6.0	100	75	100	77	100	94
Exponential	9.0	100	62	100	70	100	83

The above indicates that the estimator proposed seems to have comparative advantage over blanket use of $p = 2$ with small samples and certain distributions, viz. Laplace and Exponential.

CONCLUSIONS

It is seen above that improvement in variance estimation using L_p– methods is certainly less dramatic than in estimating the location parameter θ in symmetric distributions, primarily because the transformed data set is non-symmetric. It is however tentatively suggested that an improvement over the conventional methods for the estimation of odd moments, may, in general, be achieved by using the L_p– method (or a variation thereof) outlined above.

3.1.5 A comparative study utilising the data sets published by Stigler (1977)

3.1.5.1 Introduction and comments on the study

The data sets published by Stigler (and discussants) (1977) constituting 24 sets of observations by scientists of physical phenomena, and his comparison of a set of robust estimators of location, provides a valuable testing ground for any new location estimator.

Before describing the analysis, it is of some value to discuss some of the salient problems associated with the estimation of location with such data sets, some of which have been raised by the discussants of Stigler's paper. Firstly, since some of the data sets exhibit very significant skewness, the whole problem of the uniqueness of the location parameter has to be given consideration. It is not clear which population measure of location in such sets is most appropriate. A discussant of the paper (Eisenhart) demonstrates the considerable bias in some of the data sets; for example he remarks that data sets 17, 23 and 24 have "true values" at approximately the 32nd, 88th and between the 5th and 8th percentiles respectively. He goes on to say: "*True values such as these that lie 'out in wings' are clearly unsuitable for judging the relative merits of a group of estimators that are 'arguing' over which value in a central 'core' of a set of data 'best' summarizes the 'evidence' of the set as a whole.*" Hoaglin (discussant) says in this connection, "...... *when a data set involves a shift or a bias, it is not at all clear whether that bias should be charged against the performance of any estimator. To do so implies that the estimator should be able to see beyond the data to the 'true' value of the physical quantity* ......"

A second problem relates to that of comparing different estimators of sets of data with vastly differing variability, necessarily giving rise to sets of estimates with widely differing variances. Stigler uses a "robust" measure of variability for the j^{th} data set (s_j) - the average of the absolute deviations of the set of estimates obtained for that data set. His relative error is then calculated as the absolute deviation of the estimate divided by this measure of spread. Small absolute errors may thus be associated with large relative errors.

Given constraints imposed by the data at hand, it appears, however, that the method of comparison adopted gives an adequate portrayal of the relative performance of the estimators examined for a specific real life situation.

3.1.5.2 Results for the previously proposed L_p-estimators

This method of comparison (with the reservations outlined above) was thus applied when evaluating the L_p-estimators (3.1.3) as members of the class of robust estimators. The same values for the mean absolute deviation of the estimates for each data set were used; as Stigler says in connection with extrapolating the results to a new estimator: "*the fact that the new estimator does not contribute to the* s_j *should make little difference.*"

The adaptive estimators used draw from the work of section 3.1 on the use of L_p-norm estimation with skewed data. The more general version of the formula is used, viz:

$$p = 1 + \left(\frac{3}{b_2 - b_1}\right)^q \qquad (3.1.2)$$

where b_1 is a measure of squared skewness

b_2 is a measure of kurtosis

and where values of q varying from 0.5 to 2.0 are considered for comparison purposes.

The results are given in Table 3.1.10, with average relative errors for the small data sets (1-20) and large data sets (21-24) as defined by Stigler; standard deviations of the relative errors are given in parentheses. For comparison purposes Stigler's results are also given in Table 3.1.9.

It is seen that for all the values of q the estimators perform adequately. It is seen however that the value for data set 5 is inflating the results. Data set 5 has in fact a very low s_j of 0.078; thus although the relative error for the $q = 1$ estimator (for example) in Table 3.1.10 is 1.297, the value of the mean square error is 1×10^{-2} which, relative to the true value (8.798), is not important enough to distort the results as it has. It is seen that after data set 5 is excluded, the estimator pro-

TABLE 3.1.9

MEAN RELATIVE ERROR FOR STIGLER'S SELECTED ESTIMATORS
(STANDARD DEVIATIONS IN PARENTHESES)

	Small Samples	Large Samples
Mean	0.931 (.20)	0.924 (.19)
Median	1.149 (.28)	1.152 (.18)
Edgeworth	1.018 (.08)	0.945 (.07)
Outmean	1.038 (.58)	0.774 (.50)
10% Trim	0.916 (.20)	0.944 (.06)
15% Trim	0.938 (.10)	0.991 (.04)
25% Trim	1.039 (.08)	1.073 (.12)
Huber P15	0.922 (.20)	0.985 (.05)
Andrews AMT	0.966 (.14)	1.032 (.13)
Tukey Biweight	1.023 (.13)	1.097 (.17)
Hogg T1	1.014 (.07)	1.084 (.13)

TABLE 3.1.10

RELATIVE ERROR FOR STIGLER'S DATA SETS (FORMULA 3.1.2)

(VALUES OF p USED IN ITALICS)

	Values of q			
Data Set	0.5	1.0	1.5	2.0
1	0.836 *1.845*	0.868 *1.884*	0.900 *1.845*	0.933 *1.799*
2	1.003 *1.691*	0.998 *1.477*	1.010 *1.330*	1.033 *1.228*
3	0.146 *1.912*	0.013 *1.831*	0.105 *1.758*	0.214 *1.691*
4	0.685 *2.082*	0.638 *2.170*	0.588 *2.266*	0.536 *2.369*
5	1.104 *2.088*	1.297 *2.184*	1.499 *2.288*	1.711 *2.402*
6	0.755 *2.028*	0.743 *2.056*	0.730 *2.086*	0.718 *2.116*
7	0.991 *1.855*	0.992 *1.732*	0.992 *1.626*	0.993 *1.535*
8	0.977 *1.813*	0.987 *1.661*	0.992 *1.538*	0.994 *1.438*
9	1.192 *1.817*	1.093 *1.668*	1.040 *1.545*	0.998 *1.446*
10	0.968 *2.182*	0.957 *2.398*	0.946 *2.652*	0.937 *2.953*
11	0.981 *1.813*	0.995 *1.660*	1.009 *1.537*	1.022 *1.436*
12	0.929 *2.065*	0.920 *2.135*	0.912 *2.209*	0.903 *2.289*
13	1.031 *2.261*	1.043 *2.591*	1.055 *3.007*	1.068 *3.531*
14	0.939 *1.878*	0.955 *1.771*	0.965 *1.677*	0.973 *1.594*
15	1.013 *2.273*	1.015 *2.621*	1.017 *3.063*	1.019 *3.627*
16	1.058 *2.026*	1.061 *2.053*	1.064 *2.081*	1.067 *2.109*
17	0.985 *1.821*	1.024 *1.673*	1.067 *1.552*	1.114 *1.453*
18	0.847 *2.112*	0.813 *2.237*	0.783 *2.375*	0.755 *2.529*
19	1.092 *1.978*	1.087 *1.957*	1.083 *1.936*	1.079 *1.915*
20	0.928 *2.080*	0.922 *2.166*	0.915 *2.258*	0.909 *2.358*
21	0.950 *1.669*	1.053 *1.447*	1.127 *1.299*	1.204 *1.200*
22	0.708 *2.020*	0.695 *2.040*	0.682 *2.060*	0.669 *2.081*
23	1.011 *1.535*	0.995 *1.287*	1.002 *1.153*	1.013 *1.082*
24	1.006 *1.948*	1.005 *1.898*	1.003 *1.852*	1.002 *1.807*
Mean for set 1-20	0.923(0.217)*	0.921(0.254)	0.934(0.262)	0.948(0.279)
Mean for set 1-20 with set 5 excluded	0.913(0.218)	0.901(0.245)	0.904(0.234)	0.908(0.223)
Mean for set 21-24	0.918(0.143)	0.937(0.163)	0.953(0.190)	0.972(0.222)

* Standard deviations in parentheses

TABLE 3.1.11

RELATIVE ERROR FOR STIGLER'S DATA SETS (FORMULA 3.1.1)

(VALUES OF p USED IN ITALICS)

Data Set	Values of q				
	4.0/N$^{\frac{1}{4}}$	0.5	1.0	1.5	2.0
1	1.033 *1.670*	0.863 *1.902*	0.922 *1.814*	0.981 *1.734*	1.040 *1.662*
2	1.072 *1.142*	0.998 *1.609*	1.005 *1.371*	1.034 *1.226*	1.075 *1.137*
3	1.147 *1.269*	0.177 *1.713*	0.532 *1.509*	0.857 *1.363*	1.185 *1.259*
4	0.829 *1.818*	0.758 *1.948*	0.785 *1.898*	0.811 *1.851*	0.835 *1.806*
5	0.326 *1.548*	0.575 *1.851*	0.244 *1.725*	0.084 *1.617*	0.412 *1.525*
6	0.898 *1.686*	0.808 *1.904*	0.844 *1.818*	0.877 *1.739*	0.906 *1.668*
7	0.993 *1.529*	0.991 *1.843*	0.992 *1.711*	0.992 *1.600*	0.993 *1.506*
8	0.994 *1.431*	0.979 *1.798*	0.988 *1.637*	0.992 *1.508*	0.994 *1.406*
9	0.930 *1.056*	1.006 *1.466*	0.949 *1.217*	0.935 *1.101*	0.929 *1.047*
10	0.940 *2.840*	0.969 *2.175*	0.958 *2.380*	0.947 *2.621*	0.938 *2.905*
11	1.016 *1.477*	0.981 *1.812*	0.995 *1.659*	1.009 *1.535*	1.022 *1.434*
12	0.977 *1.718*	0.947 *1.916*	0.958 *1.840*	0.968 *1.769*	0.979 *1.705*
13	1.058 *3.106*	1.029 *2.218*	1.039 *2.483*	1.050 *2.805*	1.060 *3.198*
14	0.993 *1.287*	0.961 *1.719*	0.979 *1.516*	0.989 *1.371*	0.994 *1.267*
15	1.019 *3.488*	1.013 *2.272*	1.015 *2.619*	1.017 *3.060*	1.019 *3.621*
16	1.034 *1.822*	1.050 *1.949*	1.044 *1.901*	1.039 *1.856*	1.033 *1.812*
17	1.113 *1.455*	0.988 *1.806*	1.031 *1.650*	1.079 *1.524*	1.131 *1.423*
18	2.449 *2.449*	0.848 *2.107*	0.816 *2.225*	0.787 *2.356*	0.759 *2.501*
19	1.061 *1.825*	1.085 *1.946*	1.075 *1.894*	1.065 *1.846*	1.056 *1.800*
20	0.913 *2.301*	0.928 *2.079*	0.922 *2.165*	0.915 *2.258*	0.909 *2.357*
21	1.131 *1.293*	0.954 *1.661*	1.059 *1.437*	1.133 *1.289*	1.214 *1.191*
22	0.813 *1.859*	0.755 *1.948*	0.787 *1.898*	0.818 *1.852*	0.847 *1.807*
23	1.015 *1.037*	0.955 *1.308*	1.011 *1.095*	1.015 *1.029*	1.015 *1.009*
24	1.004 *1.873*	1.006 *1.948*	1.005 *1.898*	1.003 *1.851*	1.002 *1.807*
Mean for Sets 1-20	0.968 (0.169)	0.899 (0.206)	0.905 (0.199)	0.921 (0.213)	0.963 (0.162)
Mean for Sets 21-24	0.991 (0.132)	0.918 (0.111)	0.966 (0.121)	0.992 (0.130)	1.020 (0.150)

TABLE 3.1.12

Sample Skewness (in italics) and Sample Kurtosis

Data Set		Data Set	
1	*0.575* 3.688	13	*0.371* 2.023
2	*-1.344* 8.094	14	*-1.384* 5.809
3	*1.512* 5.895	15	*0.045* 1.853
4	*0.882* 3.341	16	*0.693* 3.329
5	*1.267* 4.139	17	*0.398* 4.615
6	*0.911* 3.669	18	*0.150* 2.449
7	*0.342* 4.217	19	*-0.468* 3.354
8	*0.417* 4.709	20	*0.032* 2.575
9	*-3.052* 13.810	21	*-0.403* 6.867
10	*0.164* 2.174	22	*0.673* 3.339
11	*0.113* 4.555	23	*-4.598* 31.615
12	*-0.965* 3.573	24	*-0.019* 3.340

posed does well for the small sample size data sets.

The large sample data sets 23 and 24 (50 to 100 observations) have been shown to possess considerable bias, as discussed above, and it therefore appears that it is difficult to form useful comparisons. One thing which is startling is that the sample mean does so well with sample set 21 which has a sample kurtosis of 6.867 (see Table 3.1.12) and a low sample skewness (-0.403). Its very significant improvement over the median, which is normally far superior with such a long tailed data set, contradicts any simulation results the author has at hand.

The data was also analysed using L_p-norm estimation with the formula:

$$p = 1 + \left(\frac{3}{b_2}\right)^q \qquad (3.1.1)$$

that is, no adjustment made for skewness, with q varying from 0.5 to 2.0 as before. These results are presented in Table 3.1.11. In addition, the L_p-norm estimate with p yielded by q calculated from the formula:

$$q = \frac{4.0}{N_{\frac{1}{4}}}, \quad \text{(see 3.1.3)},$$

is provided which gives an unique, easily computable value of q. It is seen again how sensitive the results are to set 5. If the results from data set 5 are excluded these results will look rather worse (apart from the q = 2.0 case). In fact the improvement over formula (3.1.1) by formula (3.1.2) is not as great as expected. It is hypothesized that formula (3.1.2) will perform markedly better than formula (3.1.1) when $\beta_2 \simeq \beta_1 + 1$ (note that β_2 will always be greater than $\beta_1 + 1$). If this is not the case, then the two formulae will not differ that significantly - this is borne out by the simulation of (3.1.5). Such cases, when kurtosis is close to squared skewness plus one, do not predominate in this study.

3.1.5.3 Conclusions

It is apparent, from examination of Tables 3.1.10 and 3.1.11, that for the smaller values of q ($q = 1.0$) the adaptive L_p-estimator, using either of the two proposed formulae, performs extremely well relative to the other estimators. As q increases the performance of both estimators worsens and for these data sets there is no obvious trade off between optimal p and tail length. Since this contradicts both intuitive reasoning and simulated results, it is probably wise to take serious note of the various criticisms voiced against this study as a means of testing robust estimators.

Overall, even given the limitations and problems associated with this study, it is seen that of the "new fangled estimators" (Eisenhart (discussant)) the L_p-norm approach outlined deserves serious consideration.

3.2 L-ESTIMATION WHEN THE UNDERLYING DISTRIBUTION IS SKEWED

3.2.1 Introduction

As discussed below, when considering skewed distributions one is faced with the problem of deciding which of a range of population location parameters it is most desirable to estimate. In general it was noted that of these the $E(X)$ is usually estimated.

It was noted in section 3.2, Part I, that Sarhan (1954) had considered finding the BLSS for estimating the location parameter of certain asymmetric distributions. No general rule was however evident from this study regarding which tail of the L-estimator should receive greater weight if the distribution was asymmetric. For example he found that the Beta (2,3) distribution (skewed to the left) had an asymmetric BLSS with less weight given to the left tail while the exponential distribution (skewed to the right) had $\overline{X}$ as the BLSS.

3.2.2 The simulation study

A simulation study was set up to test whether a-symmetrically weighted functions of the order statistics could yield better estimates for the $E(X)$ than $\overline{X}$ for asymmetric distributions. Distributions with skewnesses of -1 and 1 and kurtoses of 4 and 24 as well as with skewness of 12 and 2 and kurtoses of 6 and 24 were simulated using the suite of programs written by Hill et al (1976) and Hill (1976) utilising the S_U-S_B distribution system of Johnson (1949). Sample sizes of 10, 30 and 50 were selected and 2000 iterations were used for each distribution and each sample size.

It was shown (section 1.4) that using $p = q = 1$ in the beta function yielded a symmetric distribution of weights. Deviations away from this symmetrical weighting function were considered in the following two forms:

(i) $q = 1$; $p > 1$ or $p < 1$ i.e. x^{p-1}

(ii) $p = 1$; $q > 1$ or $q < 1$ i.e. $(1-x)^{q-1}$

It was found that minimum sample MSE was reached with the absolute value of $p-1$ (or $q-1$) being less or equal to 0.4. The sample MSE for a grid of $p-1$ (or $q-1$) over [-0.4, 0.4] in increments of 0.1 are given in the Tables 3.2.1 to 3.2.3 below. As it would appear not to add anything of significance, for the case with kurtosis equal to 24.0 only the first weighting function was used.

The reader is referred to Tables 3.1.4 to 3.1.6 for an indication of how sample skewness and kurtosis varies with the theoretical values for the S_U-S_B family.

REMARKS

(i) For this family of distributions it is seen that functions which weight the order statistics in the

TABLE 3.2.1

SAMPLE M.S.E. OVER 2000 ITERATIONS

N = 10

Skewness	Kurtosis	*Values of a in x^a							
		-.4	-.3	-.2	-.1	0	.1	.2	.3
-1.0	4.0	-	-	1.1625	0.9711	0.8683	0.8419	0.8797	-
-1.0	24.0	-	-	1.1411	0.9726	0.8850	0.8673	0.9086	-
+1.0	4.0	-	-	0.8986	0.8602	0.8812	0.9569	1.0820	-
+1.0	24.0	-	-	1.0132	0.9166	0.8885	0.9209	1.0059	-
-2.0	6.0	-	-	1.3063	1.0403	0.8651	0.7652	0.7260	0.7344
-2.0	24.0	-	-	1.2106	1.0025	0.8819	0.8361	0.8527	-
+2.0	6.0	0.7248	0.7180	0.7433	0.8015	0.8927	1.0163	1.1711	1.3555
+2.0	24.0	-	-	0.9346	0.8817	0.8892	0.9514	1.0618	-
		*Values of b in $(1-x)^b$							
		-.4	-.3	-.2	-.1	0	.1	.2	.3
-1.0	4.0	-	-	0.8671	0.8379	0.8683	0.9536	1.0884	-
+1.0	4.0	-	-	1.5150	0.9726	0.8812	0.8651	0.9122	-
-2.0	6.0	0.6913	0.6850	0.7115	0.7715	0.8651	0.9917	1.1501	-
2.0	6.0	-	1.3201	1.0619	0.8927	0.7971	0.7608	0.7710	-

* $a = p-1$
$b = q-1$

TABLE 3.2.2

SAMPLE M.S.E. OVER 2000 ITERATIONS

N = 30

Skewness	Kurtosis	*Values of a in x^a				
		-.2	-.1	0	.1	.2
-1.0	4.0	0.6569	0.4068	0.3103	0.3326	0.4439
-1.0	24.0	0.6181	0.3973	0.3117	0.3290	0.4231
1.0	4.0	0.4523	0.3359	0.3113	0.3664	0.4894
1.0	24.0	0.5282	0.3655	0.3142	0.3511	0.4565
-2.0	6.0	0.7084	0.4299	0.3079	0.2986	0.3667
-2.0	24.0	0.6674	0.4146	0.3106	0.3166	0.4041
2.0	6.0	0.3035	0.2847	0.3062	0.3669	0.4649
2.0	24.0	0.4718	0.3454	0.3151	0.3631	0.4744
		*Values of b in $(1-x)^b$				
		-.2	-.1	0	.1	.2
-1.0	4.0	0.4279	0.3230	0.3103	0.3773	0.5120
1.0	4.0	0.6282	0.3933	0.3113	0.3470	0.4760
-2.0	6.0	0.2958	0.2815	0.3079	0.3741	0.4781
2.0	6.0	0.6863	0.4189	0.3062	0.3046	0.3789

* a = p-1
b = q-1

TABLE 3.2.3

SAMPLE M.S.E. OVER 2000 ITERATIONS

N = 50

Skewness	Kurtosis	*Values of a in $x^a(1-x)^0$				
		-.2	-.1	0	.1	.2
-1.0	4.0	0.5506	0.2741	0.1785	0.2167	0.3510
-1.0	24.0	0.5166	0.2703	0.1836	0.2118	0.3212
1.0	4.0	0.3607	0.2190	0.1810	0.2310	0.3540
1.0	24.0	0.4339	0.2439	0.1842	0.2224	0.3333
-2.0	6.0	0.5768	0.2841	0.1767	0.1959	0.2968
-2.0	24.0	0.5608	0.2839	0.1827	0.2058	0.3145
2.0	6.0	0.2187	0.1796	0.1832	0.2278	0.3115
2.0	24.0	0.3808	0.2270	0.1841	0.2280	0.3389
		*Values of b in $x^0(1-x)^b$				
		-.2	-.1	0	.1	.2
-1.0	4.0	0.3374	0.2060	0.1785	0.2387	0.3718
1.0	4.0	0.5252	0.2633	0.1810	0.2312	0.3761
-2.0	6.0	0.2066	0.1702	0.1767	0.2248	0.3122
2.0	6.0	0.5672	0.2835	0.1832	0.2077	0.3126

* a = p-1
b = q-1

opposite direction to the skewness would appear to give estimates of the location parameter which have smaller sample MSE than $\overline{X}$ (a = 0 or b = 0 as the case may be).

(ii) Sample size appears to play a large role in the selection of some optimal function to determine the weighting distribution. It is seen that smaller sample sizes require much larger adjustment (i.e. larger p-1 or q-1) than larger sample sizes. This is in some way strange because the smaller sample sizes had lower average sample skewness - see Tables 3.1.4 to 3.1.6.

(iii) The optimal left hand tail weighting for the positively skewed distribution appear larger than the optimal right hand tail weighting for the negatively skewed distributions for the weighting function $x^a (1-x)^0$ and vice versa for $x^0(1-x)^b$.

(iv) Increases of kurtosis for a fixed skewness did not appear to have a very marked effect over the "a" which would minimise sample MSE but if anything it tended to pull the optimal estimation towards that of least squares (i.e. when a = 0).

3.2.3 Conclusion

The most important conclusion to be drawn from this study is that although this simulation indicates that improvement in sample MSE is <u>possible</u> by adjusting for skewness, when estimating the $E(X)$ the improvement is very small. Given the ambiguity of a location measure with skewed data sets it is doubtful whether an adjustment for skewness is really worthwhile for this estimator.

CONCLUSION TO PART II

In conclusion it is worth noting that the studies compiled above indicate quite clearly that there are definite advantages in using the adaptive L_p-estimator and L-estimator of location proposed over conventional measures, and indeed over the more sophisticated robust estimators which have been more recently proposed.

On reflection of the various results obtained above it does appear that for the estimation of the location parameter for symmetrical distributions, the use of L_p-estimation may not offer equivalent advantage to those of L-estimation. In the first place the empirical results indicate that L-estimators can be constructed which are as good (in the MSE sense) as L_p-estimators. In addition it is pertinent to remark that in terms of computer time the adaptive L-estimator used between 10% and 30% of the time required by the L_p- method (note that for large p $(p > 3.0)$ and p close to 1.0 L_p takes rather more time than when p is close to 2.0.) Finally, the distributional properties of such estimators are attainable when the underlying distribution is known in the finite and asymptotic cases.

PART III

INTRODUCTION

As pointed out in Part I the first section of this part of the work which covers research into L_p- estimation in the regression model was carried out before the work of Part II. It is, however, covered at this point because in that way the thesis follows a more natural progression. In addition to the work on L_p estimation a proposal is made for the extension of the method of Lloyd (see Part I) to the regression case as well as a tentative suggestion for the application of L-estimation to the regression model.

where $\underline{y}$ is an $n \times 1$ vector of observable random variables;

X is an $n \times m$ matrix of the known regressor variables;

$\underline{\beta}$ is an $m \times 1$ vector of unknown parameters, and

$\underline{e}$ is an $n \times 1$ unobservable error vector.

The ordinary least squares estimator (OLS) of $\underline{\beta}$ is obtained by minimizing the sum of the squares of the errors and is given by:

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}.$$

If it is assumed that $E(\underline{e}) = 0$ and $E(\underline{e}\underline{e}') = \sigma^2 I$, then it follows (Gauss-Markov theorem) that $\hat{\underline{\beta}}$ is the best (minimum variance) linear (linear function of the $\underline{y}$) unbiased estimator (BLUE) of $\underline{\beta}$. This approach can be easily generalized so as to minimize the sum of any power of the errors and is known as the L_p - norm. More formally, given a sample of size n on both X and $\underline{y}$, under fairly general conditions (Kiountouzis (1971)) a "best" estimate of $\underline{\beta}$, $\tilde{\underline{\beta}}$ the particular value of $\underline{b}$ which minimizes

$$\sum_{i=1}^{n} |y_i - X_i \underline{b}|^p$$

where $\underline{b} = b_0, \ldots, b_{m-1}$ and this estimator is known as the L_p-norm estimator. As stressed before, the fact that the least squares estimator $(p = 2)$ is the BLUE does not detract from interest in values of $p \neq 2$ since the estimates obtained using values other than 2 are not linear estimators and hence can have lower variance than those obtained using least squares.

Besides the least squares case of $p = 2$, two other values of p have been of particular interest in the literature, namely $p = 1$ and $p = \infty$. The case of $p = 1$ (that is, minimization of the sum of the absolute errors) corresponds to some of the earliest approaches to curve fitting (e.g. Fourier (1824) and Edgeworth (1888)). Wagner (1959) showed that this problem could be formulated as a linear programming problem by considering the unrestricted error term to be the difference between two non-negative variables. Letting $e_i = u_i - v_i$ where $u_i, v_i \geq 0$ and $e_i = y_i - X_i \underline{b}$ the problem becomes:

this follows

$$\text{Minimize } \sum_{i=1}^{n} (u_i + v_i)$$

$$\text{subject to } y_i = \sum_{j=0}^{m-1} b_j x_{ij} + u_i - v_i \quad (i = 1,2,\ldots,n)$$

$$b_j \text{ unrestricted in sign; } u_i, v_i \geq 0, \quad (i = 1,2,\ldots,n).$$

Since Wagner's paper, considerable interest has been shown in the case $p = 1$, and numerous researchers have produced results indicating that the L_1—norm is more suitable than the traditional L_2—norm in certain circumstances, especially when the error distribution has long tails (e.g. Blattberg and Sargent (1971), Kiountouzis (1973), Harter (1977)).

The case $p = \infty$ corresponds to minimization of the maximum error and has become known as Chebychev estimation (in addition to L_∞ estimation). Wagner (1959) showed that the L_∞ estimation problem could be formulated as a linear programming problem by letting $D = \underset{i}{\text{Max}}\{|e_i|\}$. The problem can then be stated as follows:

$$\begin{aligned}
&\text{Minimize} \quad D \\
&\text{subject to} \quad \left.\begin{aligned} &D \geq e_i; \quad D \geq y_i - X_i\underline{b} \\ &D \geq -e_i; \quad D \geq -y_i + X_i\underline{b} \end{aligned}\right\} i = 1,\ldots,n \\
&\qquad\qquad\quad D \geq 0; \ b_j \ \text{unrestricted in sign}
\end{aligned}$$

Interest in Chebychev estimation has been limited (e.g. Appa and Smith (1973)) although Harter (1977) has indicated that it could be superior to both L_1– and L_2– estimation when the (sample) kurtosis of the error term is less than 2.25.

There is no theoretical reason why values of p other than 1,2 and ∞ should not be considered. Forsythe (1972) has suggested that $p = 1.5$ might be a good compromise value as it provides estimates which are substantially better than least squares when the error distribution has long tails, and is not very bad when the errors have a Normal distribution (when least squares is most appropriate).

In this chapter values of $p = 1.00, 1.25, 1.50, 1.75, 2.00$, and ∞ are examined. It should be noted that the cases of $p = 1,2$, and ∞ provide exact solutions, while the other values of p give rise to a nonlinear programming problem, whose solution can only be found to a given level of convergence. The method used to obtain the L_p– estimates for $p = 1.25, 1.50$, and 1.75 was that described by Fletcher and Powell (1963) and Forsythe (1972), which is part of the IBM Scientific Subroutine Package (1968).

1.3 DESIGN OF THE SIMULATION STUDY

The simulation model examined was of the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad , \quad i = 1,2,\ldots,n$$

and had the following specifications.

The sample size n (for each iteration) was chosen as 25, and 500 iterations were performed for each of the error distributions examined. β_0, β_1 and β_2 were arbitrarily selected as 10, 8, and -6 respectively. The 25 values of X_1 (fixed for all iterations) were chosen randomly from a uniform distribution on the range (0, 40). The 25 fixed values of X_2 were also randomly selected from a uniform (0, 40) distribution and it was checked that X_1 and X_2 were uncorrelated (that is, $|\rho_{X_1,X_2}| < 0.01$). Finally, for each iteration, 25 error terms were randomly generated such that:

$$E(e_i) = 0 \quad \text{and} \quad Var(e_i) = 9 \quad \text{for} \quad i = 1,2,\ldots,25.$$

Thus the y_i were generated using the model:

$$y_i = 10 + 8\, x_{1i} - 6\, x_{2i} + e_i \quad , \quad i = 1,2,\ldots,25$$

and the various L_p estimates of β_0, β_1 and β_2 were obtained.

A variety of statistical distributions were used to generate the random errors. These distributions were all symmetric and were chosen so as to cover a fairly wide range of kurtosis. The distributions used are detailed below:

(i) <u>Uniform Distribution</u>: (Kurtosis = 1.8)

$$f_X(x) = \frac{1}{\beta-\alpha} \quad , \quad \alpha < x < \beta$$

In order to have mean zero and standard deviation 3, α and β must be selected as $\alpha = -\beta$ and $\beta = \sqrt{3\sigma^2} = \sqrt{27}$.

(ii) Normal Distribution: (Kurtosis = 3)

Obviously $\mu = 0$ and $\sigma^2 = 9$ were the parameters selected.

(iii) Contaminated Normal Distribution: (Kurtosis = 3.5, 4.0, 4.5, 5.0, 5.5)

$$f_X(x) = w \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} + (1-w)\frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}$$

$$-\infty < x < \infty$$

$$0 \leq w \leq 1 .$$

For this study, w was selected as $\frac{1}{2}$ (that is equal weight was given to both distributions) and $\mu_1 = \mu_2 = 0$ (ensuring symmetry). This family of contaminated normal distributions was used to obtain error distributions with the required kurtosis coefficients by choosing σ_1^2 and σ_2^2 appropriately, while still ensuring that the overall variance remained 9. For further details of the properties of the contaminated normal distribution the reader is referred to Johnson and Kotz (1970).

(iv) Laplace (or Double Exponential) Distribution: (Kurtosis = 6)

$$f_X(x) = \frac{1}{2\beta} \exp\left(-\frac{|x-\alpha|}{\beta}\right) , \quad -\infty < x < \infty \quad .$$

In order to have mean zero and variance 9, $\alpha = 0$ and $\beta = \sqrt{\sigma^2/2} = \sqrt{4.5}$ were selected.

(v) <u>Cauchy Distribution</u>: (Kurtosis undefined)

$$f_X(x) = \frac{1}{\pi\beta\{1+[(x-\alpha)/\beta]^2\}} \qquad -\infty < x < \infty$$

The Cauchy distribution has no moments and hence no mean, variance or kurtosis. However it is symmetric about the median α and hence this was chosen to be zero. β was determined by specifying that the 95^{th} percentile of the Cauchy distribution had to coincide with the 95^{th} percentile of the Normal (0,9) distribution used in (ii). It should be noted that the kurtosis of the Cauchy distribution (although undefined) can be thought of as infinite.

Before concluding this section, mention must be made of the manner in which the random observations from the error distributions were generated. For the Uniform, Laplace and Cauchy distributions the inverse cumulative distribution functions are easily obtained and hence these errors can be easily generated using uniform (0,1) random numbers. These uniform (0,1) random numbers were obtained using shuffled (Maclaren and Marsaglia (1965)) values of the uniform random numbers given by the UNIVAC routine RANDU. The normal and contaminated normal errors were generated using the same shuffled uniform random numbers as above and the Box-Müller (1958) transformation.

1.4 EXPERIMENTAL RESULTS

The aim of the empirical investigation (simulation) was to gain insight into the properties of the L_p— estimators of the coefficients of a regression model, and to establish guidelines for a "suitable" choice of p. These findings are presented below.

1.4.1 Unbiasedness

It is stressed that the simulation study was only performed for symmetric error distributions. It is well known that the least squares estimates $(p = 2)$ are unbiased, and so far studies using symmetric error distributions have produced no evidence of bias in the estimates obtained using other values of p (Forsythe (1972), Kiountouzis (1973) and Harvey (1978)). This more extensive study confirms these findings.

Means of the 500 sample estimates of the regression coefficients were computed and compared to the true parameter values for values of p = 1.00, 1.25, 1.50, 1.75, 2.00, and ∞. This was repeated for each of the error distributions examined and the results are presented in Table 1.1.1 below. Under the hypothesis of unbiasedness of the estimates, an overestimate of the true coefficient is just as likely as an underestimate. Using the normal approximation to the binomial distribution, it follows that for a sample of 500 estimates the 95% confidence interval for the number of estimates falling above the true parameter value is (228,272). For all choices of p and for all error distributions examined, not one count fell outside the above limits. Moreover, examination of Table 1.1.1 reveals that all of the means of the sample estimates are "close" to the true parameter values and therefore it is concluded that the L_p norm estimators are unbiased for all $p \geq 1$ when the error distribution is symmetric.

1.4.2 Efficiency of the Individual estimates

The sampling variances (based on the 500 individual sample estimates) of each of the three regression coefficients were computed for each choice of p, and for all the error distributions considered and are presented in Table 1.1.2 below.

On examination of Table 1.1.2 it is apparent that for $p = 2$ the sample variances are approximately the same for all error distributions with the exception of the Cauchy. This is not surprising since the true covariance matrix of the L_2 estimate is given by $\sigma^2(X'X)^{-1}$ which is independent of the error distribution, provided $E(\underline{e}) = \underline{0}$ and $E(\underline{e}\underline{e}') = \sigma^2 I$ (Gauss-Markov theorem). The Cauchy distribution would obviously provide an exception since σ^2 is undefined (infinite) and hence $E(\underline{e}\underline{e}') \neq \sigma^2 I$. Further examination of the table reveals that for $p \neq 2$ the sample variances are not constant but vary according to the error distribution. In addition, for any given error distribution these variances are dependent on the choice of p. In fact, it should be noted that as the kurtosis of the error distribution increases, a choice of $p < 2$ results in estimates with smaller sample variances than those obtained using least squares. This is especially noticeable when examining the results obtained for the Cauchy distribution.

As yet it has not been possible to derive a theoretical expression for the covariance matrix when $p \neq 2$ but it is conjectured that the true covariance matrix for these L_p-estimators will be of the form $(\sigma^2(X'X)^{-1})^{-\ell}$ where ℓ is a function of p and the kurtosis of the error distribution.

1.4.3 Generalised Variance and the Choice of p

The empirical generalized variance of the regression coefficients is defined as the determinant of the empirical covariance matrix of these estimates, and can be considered as a univariate summary of the information present in the sample covariance matrix. Since the estimates have been shown to be unbiased for all p, it is reasonable to base the choice of p on the generalized variance. Clearly, from a practical point of view, it is desirable that the generalized variance be as small as possible.

The generalized variances of the 500 sample estimates are presented in Table 1.1.3 below for all values of p and for all error distributions considered. The chosen symmetric errors distributions had coefficients of kurtosis varying from 1.8 (short tailed) through 3 (medium tailed) to ∞ (long tailed; Cauchy). A plot of the kurtosis of the error distribution against the p which gives the smallest empirical generalized variance is presented in Figure 1. (It should

TABLE 1.1.1

Comparison of Average Values of the Estimated Regression Coefficients with Population Values ($n = 25$; $\sigma^2 = 9$)

Distbn.	Kurt.	β	True Value	p 1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	β_0	10	9.98	9.99	10.00	10.00	10.00	9.95
		β_1	8	8.00	8.01	8.01	8.01	8.01	8.00
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-5.99
Normal	3.0	β_0	10	9.98	9.99	9.99	10.00	10.00	9.99
		β_1	8	7.99	8.00	8.00	8.00	8.00	8.00
		β_2	-6	-5.99	-5.99	-5.99	-5.99	-6.00	-6.00
Contam. Normal	3.5	β_0	10	9.99	10.00	10.01	10.02	10.03	10.01
		β_1	8	8.00	8.00	8.00	8.00	8.00	8.00
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.00
Contam. Normal	4.0	β_0	10	10,04	10.05	10.06	10.05	10.04	10.06
		β_1	8	8.00	8.00	8.00	8.00	8.00	8.00
		β_2	-6	-6.01	-6.00	-6.00	-6.01	-6.01	-5.99
Contam. Normal	4.5	β_0	10	9.96	9.96	9.95	9.96	9.96	10.01
		β_1	8	8.01	8.01	8.01	8.01	8.01	7.99
		β_2	-6	-5.99	-6.00	-6.00	-6.00	-6.00	-6.01
Contam. Normal	5.0	β_0	10	9.99	10.00	9.99	9.97	9.97	10.10
		β_1	8	8.00	8.00	8.00	8.00	8.00	8.00
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.01
Contam. Normal	5.5	β_0	10	9.98	9.99	9.99	10.01	10.02	10.13
		β_1	8	8.00	8.00	8.00	8.00	7.99	7.97
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.00
Laplace	6.0	β_0	10	10.04	10.05	10.03	10.02	10.01	9.89
		β_1	8	8.00	8.00	8.00	8.00	8.00	8.02
		β_2	-6	-6.00	-6.01	-6.00	-6,00	-6,00	-5.99
Cauchy	–	β_0	10	10.01	10.03	10.07	10.25	10.39	9.74
		β_1	8	8.00	8.00	7.97	7.98	7.97	8.04
		β_2	-6	-6.00	-6.00	-5.98	-6.01	-6.02	-6.04

TABLE 1.1.2

Empirical Variances of the Individual Regression Estimates ($n = 25$; $\sigma^2 = 9$)

		Values of p for $\tilde{\beta}_0$						Values of p for $\tilde{\beta}_1$						Values of p for $\tilde{\beta}_2$					
Distbn.	Kurt.	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	5.17	3.47	2.74	2.32	2.05	1.58	0.056	0.040	0.032	0.028	0.024	0.018	0.051	0.037	0.030	0.025	0.022	0.015
Normal	3.0	3.56	2.82	2.53	2.43	2.43	7.16	0.038	0.029	0.025	0.024	0.024	0.076	0.036	0.027	0.024	0.023	0.022	0.060
Con.Norm	3.5	3.25	2.70	2.44	2.35	2.36	6.99	0.036	0.027	0.024	0.023	0.023	0.085	0.032	0.026	0.024	0.023	0.023	0.070
Con.Norm	4.0	2.34	1.83	1.76	1.82	1.96	9.02	0.028	0.022	0.021	0.022	0.024	0.111	0.024	0.020	0.019	0.020	0.022	0.090
Con.Norm	4.5	2.25	1.95	1.92	2.03	2.25	10.23	0.028	0.024	0.023	0.024	0.026	0.110	0.021	0.018	0.018	0.019	0.020	0.094
Con.Norm	5.0	1.82	1.74	1.88	2.13	2.45	12.22	0.022	0.021	0.022	0.024	0.028	0.139	0.020	0.018	0.019	0.021	0.023	0.117
Con.Norm	5.5	1.26	1.27	1.49	1.83	2.25	13.65	0.012	0.012	0.015	0.018	0.023	0.160	0.014	0.014	0.017	0.021	0.026	0.143
Laplace	6.0	1.86	1.55	1.57	1.73	1.98	12.33	0.021	0.018	0.018	0.020	0.023	0.127	0.018	0.016	0.016	0.017	0.020	0.112
Cauchy	-	0.20	0.30	1.52	6.51	22.56	1128	0.002	0.003	0.093	0.195	0.744	22.32	0.002	0.003	0.170	0.196	0.695	7.62

be noted that the two points with a coordinate value of infinity are not shown.)

TABLE 1.1.3

Generalized Variance of Regression Estimates ($n = 25$; $\sigma^2 = 9$)

		p					
Distbn.	Kurt.	1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	24.55	8.85	4.52	2.77	1.87	0.63
Normal	3.0	8.35	3.93	3.64	2.21	2.09	39.00
Con.Normal	3.5	6.08	3.24	2.37	2.10	2.20	77.79
Con.Normal	4.0	2.86	1.56	1.35	1.48	1.82	150.90
Con.Normal	4.5	2.14	1.31	1.17	1.33	1.73	193.00
Con.Normal	5.0	1.19	0.93	1.10	1.53	2.30	321.00
Con.Normal	5.5	0.31	0.30	0.53	1.07	2.12	591.00
Laplace	6.0	1.06	0.63	0.66	0.90	1.36	286.95
Cauchy	∞	0.0018	0.0051	10.40	416.30	195000	$5.8\times10^{+7}$

To obtain the true generalized variances, values in the above table should be multiplied by 10^{-4}, except for the Cauchy row where the actual figures are recorded.

Assuming minimum generalized variance to be a suitable criterion for choosing p, this figure clearly shows that the "best" p depends on the kurtosis of the underlying error distribution. The longer tailed the distribution (that is, the larger the kurtosis) the smaller the "optimal" p, with $p = 1$ providing minimum empirical generalized variance when the kurtosis is "infinite". Also, the shorter the tail of the distribution, the larger the "optimal" p, with $p = \infty$ being most appropriate when the errors have a uniform distribution (lowest kurtosis of all the distributions considered).

From consideration of Figure 1 it is proposed that the functional relationship:

$$p = \frac{9}{k^2} + 1 \qquad (1.4.1)$$

(where k is the kurtosis of the error distribution) be used in practice to determine a suitable p, provided the error distribution is symmetric. In particular it should be noted that for the Normal distribution ($k = 3$), (1.4.1) suggests the use of $p = 2$.

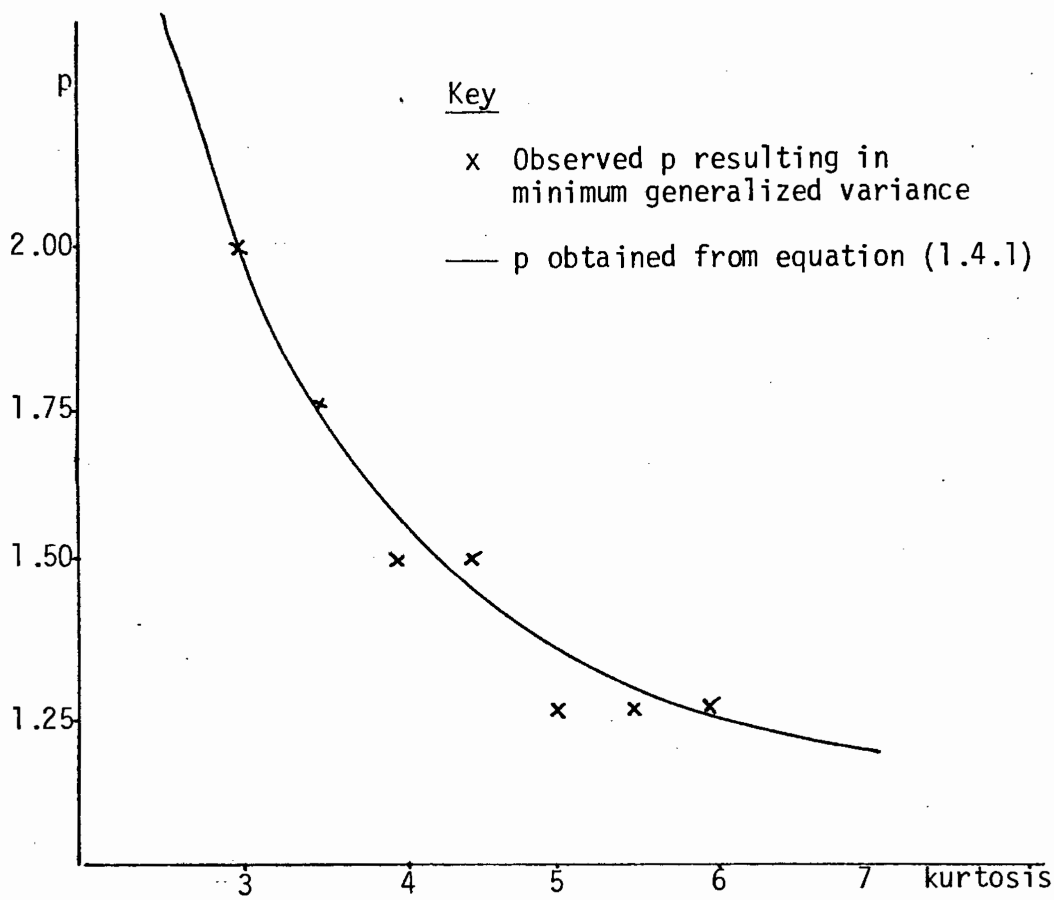


Figure 1

TABLE 1.1.4
Efficiency (based on generalized variance) of
Regression Estimates ($n = 25$; $\sigma^2 = 9$); ($\times 10^{-2}$)

Distbn.	Kurt.	p = 1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	0.026	0.071	0.140	0.227	0.337	1.000
Normal	3.0	0.250	0.532	0.792	0.946	1.000	0.054
Con.Normal	3.5	0.345	0.648	0.886	1.000	0.995	0.027
Con.Normal	4.0	0.472	0.865	1.000	0.912	0.742	0.009
Con.Normal	4.5	0.547	0.893	1.000	0.880	0.676	0.006
Con.Normal	5.0	0.782	1.000	0.845	0.603	0.404	0.003
Con.Normal	5.5	0.968	1.000	0.566	0.280	0.142	0.001
Laplace	6.0	0.594	1.000	0.955	0.700	0.463	0.002
Cauchy	∞	1.000	0.353	0.000	0.000	0.000	0.000

In the light of these results it is interesting to observe how the criterion suggested by Harter (1977) (that is, $k < 2.2$ use $p = \infty$; $2.2 < k < 3.8$ use $p = 2$; $k > 3.8$ use $p = 1$) performs in practice. The results presented in Table 1.1.4 indicate immediately that Harter's method provides a substantial improvement over the universal use of least squares (or any other p for that matter) regardless of the error distribution. However, considerable improvements can still be obtained over Harter's method by using equation (1.4.1). For example, if the kurtosis is 4.0 then use of Harter's method ($p = 1$) results in estimates which are only 47% efficient relative to those obtained using $p = 1.50$, the closest of the p's examined to the "optimum" p of 1.56.

The same type of analysis can be performed on the estimates of the variances of the individual regression coefficients and these results are presented in Table 1.1.5. Examination of this table yields conclusions analogous to those obtained from consideration of the generalized variance, namely that equation (1.4.1) provides vastly superior results to the universal use of $p = 2$ (or any other value of p) and a substantial improvement over Harter's method.

1.4.5 Further Simulation Studies

The results presented in Sections 1.4.1 to 1.4.4 above were obtained using a simulation model with n (the sample size) = 25, and σ^2 (the variance) = 9.

In order to examine the general validity of the above results it is obviously necessary to extend the simulation study. This has been done and tables analogous to Tables 1.1.1 to 1.1.5 are produced below as follows.

Tables 1.2.1 to 1.2.5 present results for the case where n remains 25 but σ^2 is 1 (instead of 9) while Tables 1.3.1 to 1.3.5 presents similar results for $\sigma^2 = 100$ (n still 25). Tables 1.4.1 to 1.4.5 and Tables 1.5.1 to 1.5.5 consider the cases where $\sigma^2 = 9$ (as before) but the sample sizes (n) are 10 and 50 respectively.

Examination of these results reveals that for each of the cases examined, the results obtained are essentially the same as those discussed in Sections 1.4.1 to 1.4.4 above. It is therefore concluded that the results and conclusions presented in these Sections (1.4.1 to 1.4.4) are valid for all

TABLE 1.1.5

Efficiency (based on the Empirical Variances) of the Individual Regression Estimates ($n = 25$; $\sigma^2 = 9$) ; ($\times 10^{-2}$)

		Values of p for $\tilde{\beta}_0$						Values of p for $\tilde{\beta}_1$						Values of p for $\tilde{\beta}_2$					
Distbn.	Kurt.	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	0.31	0.46	0.58	0.68	0.77	1.00	0.32	0.45	0.56	0.64	0.75	1.00	0.29	0.41	0.50	0.60	0.68	1.00
Normal	3.0	0.68	0.86	0.96	1.00	1.00	0.34	0.63	0.83	0.96	1.00	1.00	0.32	0.61	0.81	0.92	0.96	1.00	0.37
Con.Normal	3.5	0.72	0.87	0.96	1.00	1.00	0.34	0.64	0.85	0.96	1.00	1.00	0.27	0.72	0.88	0.96	1.00	1.00	0.33
Con.Normal	4.0	0.75	0.96	1.00	0.97	0.90	0.20	0.75	0.95	1.00	0.95	0.88	0.19	0.79	0.95	1.00	0.95	0.86	0.21
Con.Normal	4.5	0.85	0.98	1.00	0.95	0.85	0.19	0.82	0.96	1.00	0.96	0.88	0.21	0.86	1.00	1.00	0.95	0.90	0.19
Con.Normal	5.0	0.96	1.00	0.93	0.82	0.71	0.14	0.95	1.00	0.95	0.88	0.75	0.15	0.90	1.00	0.95	0.86	0.78	0.15
Con.Normal	5.5	1.00	0.99	0.85	0.69	0.56	0.09	1.00	1.00	0.80	0.67	0.52	0.08	1.00	1.00	0.82	0.67	0.54	0.10
Laplace	6.0	0.83	1.00	0.99	0.90	0.78	0.13	0.86	1.00	1.00	0.90	0.78	0.14	0.89	1.00	1.00	0.94	0.80	0.14
Cauchy	-	1.00	0.67	0.13	0.03	0.01	0.00	1.00	0.67	0.02	0.01	0.00	0.00	1.00	0.67	0.01	0.01	0.00	0.00

TABLE 1.2.1

Comparison of Average Values of the Beta Estimates with True Values ($n = 25$; $\sigma^2 = 1$)

Distn.	Kurt.	β	True Value	p 1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	β_0	10	10.00	10.00	10.00	10.00	10.01	10.01
		β_1	8	8.00	8.00	8.00	8.00	8.00	8.00
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.00
Normal	3.0	β_0	10	10.00	9.99	9.99	9.99	9.99	9.97
		β_1	8	8.00	8.00	8.00	8.00	8.00	8.00
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.00
Contam. Normal	3.5	β_0	10	10.00	9.99	9.98	9.98	9.98	9.95
		β_1	8	8.00	8.00	8.00	8.00	8.00	8.01
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.00
Contam. Normal	4.0	β_0	10	10.00	10.02	10.02	10.03	10.03	10.10
		β_1	8	8.00	8.00	8.00	8.00	8.00	7.99
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.01
Contam. Normal	4.5	β_0	10	10.01	10.02	10.03	10.03	10.04	10.01
		β_1	8	8.00	8.00	8.00	8.00	8.00	8.00
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.00
Contam. Normal	5.0	β_0	10	10.01	10.01	10.01	10.01	10.01	10.04
		β_1	8	8.00	8.00	8.00	8.00	8.00	8.00
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.00
Contam. Normal	5.5	β_0	10	10.02	10.02	10.02	10.03	10.03	10.12
		β_1	8	8.00	8.00	8.00	8.00	8.00	7.99
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.00
Laplace	6.0	β_0	10	9.99	9.99	9.99	10.00	10.00	10.08
		β_1	8	8.00	8.00	8.00	8.00	8.00	7.99
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.00
Cauchy	∞	β_0	10	9.99	9.99	9.93	9.65	9.37	2.57
		β_1	8	8.00	8.00	7.99	7.98	7.98	8.07
		β_2	-6	-6.00	-6.00	-5.98	-5.94	-5.95	-5.93

TABLE 1.2.2

Empirical Variances of the Individual Regression Estimates ($n = 25$; $\sigma^2 = 1$)

		β_0						β_1						β_2					
Distbn.	Kurt.	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞
Uniform	4.8	55	40	32	27	24	15	61	45	35	30	27	17	58	43	33	28	24	16
Normal	3.0	38	29	26	25	24	71	45	34	30	28	27	80	36	28	25	24	24	63
Con.Norm	3.5	36	28	26	25	25	84	37	30	27	27	27	93	36	28	25	24	24	82
Con.Norm	4.0	31	26	25	25	26	112	32	27	25	26	27	121	30	24	23	24	25	116
Con.Norm	4.5	25	21	21	22	24	119	29	25	24	26	28	125	26	21	20	21	23	129
Con.Norm	5.0	20	18	19	22	25	131	21	19	20	23	27	153	17	17	18	20	23	131
Con.Norm	5.5	13	13	15	18	22	152	16	16	19	23	28	187	11	12	14	17	22	147
Laplace	6.0	23	21	22	24	28	147	23	20	21	23	27	165	24	21	21	24	28	137
Cauchy	[illegible]	3	4	143	2859	18578	1595955	3	6	488	7052	35124	12824	2	4	1395	4050	9435	36706

To obtain the true variances values in the β_0 columns should be multiplied by 10^{-2}, and those in the β_1 and β_2 columns by 10^{-4}.

TABLE 1.2.3

Generalized Variance of Regression Estimates (n = 25; σ^2 = 1)

Distbn.	Kurt.	p: 1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	36.0	14.0	7.0	4.4	3.0	0.9
Normal	3.0	10.7	5.1	3.6	3.0	2.9	63.5
Con. Normal	3.5	8.3	4.2	3.2	2.9	3.0	100.8
Con. Normal	4.0	5.3	3.1	2.7	2.8	3.4	237.0
Con. Normal	4.5	3.2	1.9	1.7	2.0	2.6	306.9
Con. Normal	5.0	1.4	1.0	1.2	1.8	2.9	504.9
Con. Normal	5.5	0.4	0.4	0.7	1.3	2.5	721.5
Laplace	6.0	1.8	1.3	1.5	2.1	3.1	561.4
Cauchy	∞	0	0	1.6×10^{-3}	2×10^1	9×10^2	7×10^4

*To obtain the true generalized variances, values in the above table must be multiplied by 10^{-7}, except for the "Cauchy" row, where the actual figures are recorded.

TABLE 1.2.4

Efficiency (based on Generalized Variance) of Regression Estimates (n = 25; σ^2 = 1); ($\times10^{-2}$)

Distbn.	Kurt.	p: 1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	0.03	0.06	0.13	0.20	0.30	1.00
Normal	3.0	0.27	0.57	0.81	0.97	1.00	0.05
Con.Normal	3.5	0.35	0.69	0.91	1.00	0.97	0.03
Con.Normal	4.0	0.51	0.87	1.00	0.96	0.79	0.01
Con.Normal	4.5	0.53	0.89	1.00	0.85	0.65	0.01
Con.Normal	5.0	0.71	1.00	0.83	0.56	0.34	0.00
Con.Normal	5.5	1.00	1.00	0.57	0.31	0.16	0.00
Laplace	6.0	0.72	1.00	0.87	0.62	0.42	0.00
Cauchy	∞	1.00	1.00	0.00	0.00	0.00	0.00

TABLE 1.2.5

Efficiency (based on the Empirical Variances) of the Individual Regression Estimates
(n = 25; $\sigma^2 = 1$); ($\times 10^{-2}$)

		β_0						β_1						β_2					
Distbn.	Kurt.	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	0.27	0.38	0.47	0.56	0.63	1.00	0.28	0.38	0.49	0.57	0.63	1.00	0.28	0.37	0.48	0.57	0.67	1.00
Normal	3.0	0.63	0.83	0.92	0.96	1.00	0.34	0.60	0.79	0.90	0.96	1.00	0.34	0.67	0.86	0.96	1.00	1.00	0.38
Con.Normal	3.5	0.69	0.89	0.96	1.00	1.00	0.30	0.73	0.90	1.00	1.00	1.00	0.29	0.67	0.86	0.96	1.00	1.00	0.29
Con.Normal	4.0	0.81	0.96	1.00	1.00	0.96	0.22	0.78	0.93	1.00	0.96	0.93	0.21	0.77	0.96	1.00	0.96	0.92	0.20
Con.Normal	4.5	0.84	1.00	1.00	0.95	0.88	0.18	0.83	0.96	1.00	0.92	0.86	0.19	0.77	0.95	1.00	0.95	0.87	0.16
Con.Normal	5.0	0.90	1.00	0.95	0.82	0.72	0.14	0.90	1.00	0.95	0.83	0.70	0.12	1.00	1.00	0.94	0.85	0.74	0.13
Con.Normal	5.5	1.00	1.00	0.87	0.72	0.59	0.09	1.00	1.00	0.84	0.70	0.57	0.09	1.00	0.92	0.79	0.65	0.50	0.07
Laplace	6.0	0.91	1.00	0.95	0.88	0.75	0.14	0.87	1.00	0.95	0.87	0.74	0.12	0.88	1.00	1.00	0.88	0.75	0.15
Cauchy	∞	1.00	0.75	0.02	0.00	0.00	0.00	1.00	0.50	0.01	0.00	0.00	0.00	1.00	0.50	0.00	0.00	0.00	0.01

TABLE 1.3.1

Comparison of Average Values of the Beta Estimates with True Values ($n = 25$; $\sigma^2 = 100$)

Distbn.	Kurt.	β	True Value	p: 1.00	1.25	1.50	1.75	2.00	∞
		β_0	10	10.27	10.14	10.16	10.16	10.17	10.07
Uniform	1.8	β_1	8	8.00	8.02	8.01	8.01	8.00	7.99
		β_2	-6	-6.02	-6.01	-6.01	-6.01	-6.00	-5.99
		β_0	10	10.49	10.39	10.36	10.33	10.32	10.74
Normal	3.0	β_1	8	7.97	7.96	7.98	7.98	7.99	7.97
		β_2	-6	-6.04	-6.03	-6.04	-6.03	-6.04	-6.09
Contam.		β_0	10	9.48	9.45	9.49	9.52	9.51	9.23
Normal	3.5	β_1	8	8.03	8.03	8.03	8.03	8.03	8.09
		β_2	-6	-5.95	-5.96	-5.96	-5.97	-5.97	-6.00
Contam.		β_0	10	9.84	9.82	9.81	9.79	9.77	9.53
Normal	4.0	β_1	8	8.02	8.02	8.01	8.01	8.01	7.99
		β_2	-6	-6.00	-6.00	-6.00	-5.99	-5.99	-5.90
Contam.		β_0	10	9.95	10.03	10.02	10.00	9.98	9.43
Normal	4.5	β_1	8	8.01	8.00	8.00	8.00	8.00	8.06
		β_2	-6	-6.03	-6.03	-6.02	-6.02	-6.02	-6.00
Contam.		β_0	10	10.24	10.21	10.20	10.19	10.19	9.85
Normal	5.0	β_1	8	7.97	7.98	7.98	7.98	7.98	8.01
		β_2	-6	-6.03	-6.03	-6.03	-6.03	-6.03	-6.00
Contam.		β_0	10	10.10	10.12	10.13	10.13	10.12	9.60
Normal	5.5	β_1	8	8.02	8.02	8.02	8.02	8.02	8.07
		β_2	-6	-6.02	-6.03	-6.03	-6.04	-6.04	-6.05
		β_0	10	9.91	9.88	9.90	9.89	9.86	9.67
Laplace	6.0	β_1	8	8.01	8.02	8.02	8.02	8.02	7.97
		β_2	-6	-6.00	-6.01	-6.02	-6.02	-6.03	-6.02
		β_0	10	10.01	9.97	8.03	-2.21	-18.95	-148.0
Cauchy		β_1	8	8.00	8.01	8.19	9.16	10.67	9.24
		β_2	-6	-6.00	-6.01	-5.97	-5.96	-5.99	-4.66

TABLE 1.3.2*

Empirical Variances of the Individual Regression Estimates ($n = 25$; $\sigma^2 = 100$)

		β_0						β_1						β_2					
Distbn.	Kurt.	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	54.6	37.1	29.6	25.2	22.2	16.4	0.64	0.45	0.36	0.30	0.26	0.19	0.61	0.43	0.34	0.29	0.25	0.16
Normal	3.0	40.2	31.7	27.6	26.0	25.7	75.7	0.49	0.39	0.34	0.31	0.30	0.86	0.38	0.29	0.26	0.24	0.24	0.65
Con.Norm.	3.5	36.3	29.4	27.0	26.2	26.4	91.0	0.39	0.31	0.28	0.27	0.27	0.94	0.36	0.30	0.27	0.27	0.27	0.92
Con.Norm.	4.0	26.5	20.9	19.6	19.9	21.1	105.2	0.34	0.27	0.25	0.26	0.27	1.21	0.26	0.21	0.19	0.19	0.20	1.06
Con.Norm.	4.5	23.9	20.6	21.1	23.0	25.9	127.9	0.28	0.24	0.25	0.27	0.30	1.34	0.24	0.19	0.20	0.21	0.23	1.05
Con.Norm.	5.0	20.0	18.4	19.8	22.7	26.6	132.8	0.23	0.21	0.22	0.25	0.29	1.33	0.20	0.18	0.20	0.23	0.27	1.38
Con.Norm.	5.5	15.0	14.8	17.4	21.4	26.5	173.0	0.16	0.17	0.20	0.25	0.31	1.80	0.16	0.16	0.18	0.22	0.26	1.50
Laplace	6.0	20.9	18.1	18.1	19.8	22.9	146.5	0.24	0.21	0.22	0.25	0.29	1.62	0.21	0.18	0.19	0.21	0.24	1.36
Cauchy	∞	2.5	4.0	3044	∞	∞	∞	0.03	0.05	23.0	713.8	3437	1388	0.02	0.05	4.1	94.2	451	2145

*Any variance greater than 10 000 has been recorded as ∞.

TABLE 1.3.3

Generalized Variance of Regression Estimates ($n = 25$; $\sigma^2 = 100$)

Distbn.	Kurt.	p 1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	4.173	1.498	0.759	0.452	0.298	0.072
Normal	3.0	1.178	0.581	0.393	0.324	0.303	5.81
Con.Normal	3.5	0.726	0.385	0.294	0.275	0.287	11.20
Con.Normal	4.0	0.433	0.226	0.182	0.186	0.219	20.73
Con.Normal	4.5	0.277	0.170	0.174	0.217	0.296	28.56
Con.Normal	5.0	0.137	0.105	0.128	0.187	0.297	39.91
Con.Normal	5.5	0.057	0.058	0.097	0.187	0.358	79.29
Laplace	6.0	0.145	0.094	0.100	0.140	0.227	69.10
Cauchy	∞	0.0003	0.0019	8.909	∞	∞	∞

To obtain the true generalized variances, values in the above table must be multiplied by 10^{-2}, except for the "Cauchy" row, where the actual figures are recorded.

TABLE 1.3.4

Efficiency (based on Generalized Variance) of Regression Estimates ($n = 25$; $\sigma^2 = 100$); ($\times 10^{-2}$)

Distbn.	Kurt.	p 1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	0.02	0.05	0.09	0.16	0.24	1.00
Normal	3.0	0.26	0.52	0.77	0.94	1.00	0.05
Con.Normal	3.5	0.38	0.71	0.94	1.00	0.96	0.02
Con.Normal	4.0	0.42	0.81	1.00	0.98	0.83	0.01
Con.Normal	4.5	0.61	1.00	0.98	0.78	0.57	0.01
Con.Normal	5.0	0.77	1.00	0.82	0.56	0.35	0.00
Con.Normal	5.5	1.00	0.98	0.59	0.30	0.16	0.00
Laplace	6.0	0.65	1.00	0.94	0.67	0.41	0.00
Cauchy	∞	1.00	0.16	0.00	0.00	0.00	0.00

TABLE 1.3.5

Efficiency (based on the Empirical Variances) of the Individual Regression Estimates
($n = 25$; $\sigma^2 = 100$); ($\times 10^{-2}$)

		β_0						β_1						β_2					
Distbn.	Kurt.	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	0.30	0.44	0.55	0.65	0.74	1.00	0.30	0.42	0.53	0.63	0.73	1.00	0.26	0.37	0.47	0.55	0.64	1.00
Normal	3.0	0.64	0.81	0.93	0.99	1.00	0.34	0.61	0.77	0.88	0.97	1.00	0.35	0.63	0.83	0.92	1.00	1.00	0.37
Con.Norm.	3.5	0.72	0.89	0.97	1.00	0.99	0.29	0.69	0.87	0.96	1.00	1.00	0.29	0.75	0.90	1.00	1.00	1.00	0.29
Con.Norm.	4.0	0.74	0.94	1.00	0.98	0.93	0.19	0.74	0.93	1.00	0.96	0.93	0.21	0.73	0.90	1.00	1.00	0.95	0.18
Con.Norm.	4.5	0.86	1.00	0.98	0.90	0.80	0.16	0.86	1.00	0.96	0.89	0.80	0.18	0.79	1.00	0.95	0.90	0.83	0.18
Con.Norm.	5.0	0.92	1.00	0.93	0.81	0.69	0.14	0.91	1.00	0.95	0.84	0.72	0.16	0.90	1.00	0.90	0.78	0.67	0.13
Con.Norm.	5.5	0.99	1.00	0.85	0.69	0.56	0.09	1.00	0.94	0.80	0.64	0.52	0.09	1.00	1.00	0.89	0.73	0.62	0.11
Laplace	6.0	0.89	1.00	1.00	0.91	0.79	0.12	0.88	1.00	0.95	0.84	0.72	0.13	0.86	1.00	0.95	0.86	0.75	0.13
Cauchy	[illegible]	1.00	0.63	0.00	0.00	0.00	0.00	1.00	0.60	0.00	0.00	0.00	0.00	1.00	0.40	0.00	0.00	0.00	0.00

TABLE 1.4.1

Comparison of Average Values of the Beta Estimates with True Values (n = 10; σ^2 = 9)

Distbn.	Kurt.	β	True Value	p: 1.00	1.25	1.50	1.75	2.00	∞
		β_0	10	9.88	9.96	9.97	9.97	9.97	9.95
Uniform	1.8	β_1	8	8.01	8.00	8.00	8.01	8.01	8.01
		β_2	-6	-6.00	-6.01	-6.01	-6.01	-6.01	-6.01
		β_0	10	10.16	10.14	10.12	10.11	10.10	10.10
Normal	3.0	β_1	8	8.02	8.02	8.02	8.02	8.02	8.02
		β_2	-6	-6.03	-6.02	-6.02	-6.02	-6.02	-6.02
Contam.		β_0	10	10.00	10.02	9.99	9.96	9.96	9.95
Normal	3.5	β_1	8	8.00	8.00	8.00	8.00	8.00	7.99
		β_2	-6	-5.99	-5.99	-5.99	-5.99	-5.99	-5.98
Contam.		β_0	10	9.98	9.94	9.92	9.91	9.91	9.82
Normal	4.0	β_1	8	8.01	8.00	8.00	8.00	8.00	7.99
		β_2	-6	-5.99	-5.99	-5.99	-5.99	-5.99	-5.98
Contam.		β_0	10	9.94	9.92	9.92	9.91	9.90	9.78
Normal	4.5	β_1	8	8.00	8.00	8.00	8.00	8.00	8.01
		β_2	-6	-5.99	-5.99	-5.99	-5.99	-5.99	-5.98
Contam.		β_0	10	9.95	9.97	9.98	9.98	9.98	9.99
Normal	5.0	β_1	8	8.01	8.00	8.00	8.00	8.00	7.99
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.01
Contam.		β_0	10	9.98	9.98	9.95	9.92	9.89	9.80
Normal	5.5	β_1	8	7.99	7.99	7.99	8.00	8.00	8.01
		β_2	-6	-5.99	-5.99	-5.99	-5.99	-5.98	-5.98
		β_0	10	9.93	9.94	9.94	9.95	9.96	10.03
Laplace	6.0	β_1	8	8.02	8.01	8.01	8.01	8.01	8.00
		β_2	-6	-6.01	-6.01	-6.01	-6.01	-6.01	-6.01
		β_0	10	10.04	9.78	8.33	7.89	5.51	19.21
Cauchy	[illegible]	β_1	8	8.00	7.86	7.25	6.73	6.64	4.96
		β_2	-6	-6.00	-5.90	-5.39	-4.92	-4.55	-4.05

TABLE 1.4.2

Empirical Variances of the Individual Regression Estimates ($n = 10$; $\sigma^2 = 9$)

Distbn.	Kurt.	β_0 1.00	1.25	1.50	1.75	2.00	∞	β_1 1.00	1.25	1.50	1.75	2.00	∞	β_2 1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	7.95	6.09	5.23	4.80	4.56	5.33	0.087	0.067	0.056	0.050	0.047	0.054	0.095	0.075	0.063	0.057	0.053	0.052
Normal	3.0	6.75	5.71	5.26	5.11	5.09	8.98	0.070	0.057	0.052	0.050	0.050	0.094	0.082	0.064	0.058	0.056	0.056	0.088
Con.Norm.	3.5	7.06	5.82	5.21	4.97	4.93	8.34	0.075	0.059	0.052	0.049	0.049	0.085	0.070	0.057	0.051	0.050	0.050	0.084
Con.Norm.	4.0	5.69	4.67	4.45	4.53	4.72	9.51	0.061	0.051	0.050	0.051	0.054	0.109	0.070	0.054	0.050	0.049	0.051	0.085
Con.Norm.	4.5	4.99	4.62	4.62	4.81	5.08	10.71	0.051	0.044	0.042	0.043	0.046	0.095	0.064	0.055	0.054	0.055	0.058	0.107
Con.Norm.	5.0	5.57	4.92	4.80	4.98	5.28	10.81	0.061	0.052	0.051	0.053	0.056	0.111	0.063	0.054	0.051	0.052	0.054	0.090
Con.Norm.	5.5	4.57	3.92	3.91	4.18	4.56	10.79	0.046	0.041	0.042	0.046	0.050	0.117	0.048	0.044	0.044	0.047	0.052	0.115
Laplace	6.0	4.39	3.73	3.67	3.87	4.13	9.17	0.049	0.044	0.044	0.046	0.049	0.103	0.049	0.042	0.042	0.045	0.050	0.102
Cauchy	—	1.40	49.99	1797	4162	7740	74984	0.011	9.59	323.9	820.8	1259	6070	0.013	4.97	200.2	556.4	878.7	1533

TABLE 1.4.3*

Generalized Variance of Regression Estimates ($n = 10$; $\sigma^2 = 9$)

Distbn.	Kurt.	p 1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	1.453	0.626	0.372	0.268	0.215	0.278
Normal	3.0	0.807	0.411	0.303	0.270	0.266	1.463
Con.Normal	3.5	0.644	0.327	0.240	0.212	0.211	1.229
Con.Normal	4.0	0.479	0.253	0.213	0.221	0.249	1.999
Con.Normal	4.5	0.323	0.201	0.185	0.200	0.233	2.185
Con.Normal	5.0	0.328	0.203	0.184	0.206	0.249	2.470
Con.Normal	5.5	0.184	0.122	0.131	0.170	0.232	3.454
Laplace	6.0	0.184	0.130	0.136	0.168	0.220	2.753
Cauchy	∞	0.000029	1.564	9864	∞	∞	∞

*To obtain the true generalized variances, values in the above table should be multiplied by 10^{-2}, except for the "Cauchy" row, where the actual figures are recorded.

TABLE 1.4.4

Efficiency (based on Generalized Variance) of Regression Estimates ($n = 10$; $\sigma^2 = 9$); ($\times 10^{-2}$)

Distbn.	Kurt.	p 1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	0.15	0.34	0.58	0.80	1.00	0.77
Normal	3.0	0.33	0.65	0.88	0.99	1.00	0.18
Con.Normal	3.5	0.33	0.65	0.88	1.00	1.00	0.17
Con.Normal	4.0	0.44	0.84	1.00	0.96	0.86	0.11
Con.Normal	4.5	0.57	0.92	1.00	0.93	0.79	0.08
Con.Normal	5.0	0.56	0.91	1.00	0.89	0.74	0.07
Con.Normal	5.5	0.66	1.00	0.93	0.72	0.53	0.04
Laplace	6.0	0.71	1.00	0.96	0.77	0.59	0.05
Cauchy	∞	1.00	0.00	0.00	0.00	0.00	0.00

TABLE 1.4.5

Efficiency (based on the Empirical Variances) of the Individual Regression Estimates
($n = 10$; $\sigma^2 = 9$); ($\times 10^{-2}$)

Distbn.	Kurt.	β_0						β_1						β_2					
		1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	0.57	0.75	0.87	0.95	1.00	0.86	0.54	0.70	0.84	0.94	1.00	0.87	0.55	0.69	0.83	0.91	0.98	1.00
Normal	3.0	0.75	0.89	0.97	1.00	1.00	0.57	0.71	0.88	0.96	1.00	1.00	0.53	0.68	0.88	0.97	1.00	1.00	0.64
Con.Normal	3.5	0.70	0.85	0.95	0.99	1.00	0.59	0.65	0.83	0.94	1.00	1.00	0.58	0.71	0.88	0.98	1.00	1.00	0.60
Con.Normal	4.0	0.79	0.95	1.00	0.98	0.94	0.47	0.82	0.98	1.00	0.98	0.93	0.46	0.70	0.91	0.98	1.00	0.96	0.58
Con.Normal	4.5	0.93	1.00	1.00	0.96	0.91	0.43	0.82	0.95	1.00	0.98	0.91	0.44	0.84	0.98	1.00	0.98	0.93	0.50
Con.Normal	5.0	0.86	0.98	1.00	0.96	0.91	0.44	0.84	0.98	1.00	0.96	0.91	0.46	0.81	0.94	1.00	0.98	0.94	0.57
Con.Normal	5.5	0.86	1.00	1.00	0.94	0.86	0.36	0.89	1.00	0.98	0.89	0.82	0.35	0.92	1.00	1.00	0.94	0.85	0.38
Laplace	6.0	0.84	0.98	1.00	0.95	0.89	0.40	0.90	1.00	1.00	0.96	0.90	0.43	0.86	1.00	1.00	0.93	0.84	0.41
Cauchy	∞	1.00	0.03	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00

TABLE 1.5.1

Comparison of Average Values of the Beta Estimates with True Values ($n = 50$; $\sigma^2 = 9$)

Distbn.	Kurt.	β	True Value	p					
				1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	β_0	10	10.02	10.03	10.03	10.02	10.02	10.00
		β_1	8	7.99	7.99	7.99	7.99	7.99	8.00
		β_2	-6	-5.99	-5.99	-5.99	-5.99	-6.00	-6.00
Normal	3.0	β_0	10	9.88	9.88	9.89	9.98	9.90	10.07
		β_1	8	8.00	8.00	8.00	8.00	8.00	7.99
		β_2	-6	-5.99	-5.99	-5.99	-5.99	-5.99	-6.00
Contam. Normal	3.5	β_0	10	9.99	10.01	10.02	10.04	10.05	10.04
		β_1	8	7.99	8.00	8.00	8.00	8.00	8.00
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.00
Contam. Normal	4.0	β_0	10	9.90	9.93	9.94	9.95	9.96	9.89
		β_1	8	8.01	8.00	8.00	8.00	8.00	8.03
		β_2	-6	-5.99	-5.99	-6.00	-6.00	-6.00	-6.01
Contam. Normal	4.5	β_0	10	9.98	9.96	9.95	9.94	9.93	9.92
		β_1	8	8.00	8.00	8.00	8.00	8.00	8.01
		β_2	-6	-6.00	-5.99	-5.99	-5.99	-5.99	-6.00
Contam. Normal	5.0	β_0	10	10.02	10.01	10.00	9.99	9.98	9.85
		β_1	8	8.00	8.00	8.00	8.00	8.00	8.00
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-5.97
Contam. Normal	5.5	β_0	10	9.97	9.98	9.97	9.98	9.98	10.09
		β_1	8	8.00	8.00	8.00	8.00	8.00	7.98
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-5.99	-6.00
Laplace	6.0	β_0	10	9.99	10.01	10.00	10.01	10.01	10.08
		β_1	8	8.01	8.01	8.01	8.01	8.01	8.02
		β_2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.01
Cauchy		β_0	10	10.00	13.61	18.13	90.54	257.3	2272.8
		β_1	8	8.00	7.81	7.37	2.08	-9.23	7.95
		β_2	-6	-6.00	-6.24	-6.21	-8.39	-14.08	-6.02

TABLE 1.5.2

Empirical Variances of the Individual Regression Estimates ($n = 50$; $\sigma^2 = 9$)

		β_0						β_1						β_2					
Distbn.	Kurt.	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	2.92	2.10	1.62	1.32	1.12	0.50	0.037	0.025	0.019	0.015	0.013	0.005	0.032	0.023	0.018	0.015	0.013	0.005
Normal	3.0	1.89	1.46	1.28	1.18	1.15	5.24	0.021	0.017	0.015	0.014	0.013	0.059	0.020	0.016	0.014	0.014	0.014	0.061
Con.Norm.	3.5	1.52	1.19	1.11	1.10	1.14	7.08	0.018	0.014	0.013	0.013	0.013	0.074	0.019	0.015	0.014	0.013	0.014	0.088
Con.Norm.	4.0	1.45	1.24	1.18	1.19	1.25	8.88	0.016	0.013	0.013	0.013	0.013	0.091	0.014	0.012	0.011	0.012	0.013	0.093
Con.Norm.	4.5	1.12	0.99	1.00	1.09	1.22	10.34	0.013	0.011	0.011	0.011	0.013	0.109	0.014	0.013	0.013	0.014	0.015	0.117
Con.Norm.	5.0	0.89	0.82	0.89	1.03	1.24	11.75	0.011	0.009	0.010	0.011	0.013	0.130	0.009	0.009	0.010	0.011	0.013	0.136
Con.Norm.	5.5	0.56	0.62	0.77	1.01	1.33	13.16	0.006	0.006	0.008	0.011	0.014	0.151	0.006	0.006	0.007	0.010	0.013	0.144
Laplace	6.0	0.80	0.75	0.82	0.97	1.19	14.42	0.010	0.010	0.010	0.012	0.015	0.177	0.010	0.008	0.009	0.010	0.012	0.131
Cauchy	[illegible]	0.083	6548	∞	∞	∞	∞	0.001	19.3	205.7	∞	∞	11.07	0.001	27.6	20.5	2911	∞	53.88

Any variance greater than 10 000 has been recorded as ∞ in the above table.

TABLE 1.5.3*

Generalized Variance of Regression Estimates ($n = 50$; $\sigma^2 = 9$)

Distbn.	Kurt.	p = 1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	5.630	1.877	0.838	0.446	0.272	0.025
Normal	3.0	1.305	0.625	0.428	0.348	0.322	31.83
Con.Norm.	3.5	0.836	0.419	0.324	0.306	0.328	75.21
Con.Norm.	4.0	0.448	0.279	0.247	0.265	0.321	136.17
Con.Norm.	4.5	0.322	0.218	0.220	0.269	0.368	220.40
Con.Norm.	5.0	0.137	0.106	0.131	0.203	0.343	299.38
Con.Norm.	5.5	0.028	0.034	0.064	0.145	0.325	443.29
Laplace	6.0	0.121	0.096	0.119	0.183	0.316	511.29
Cauchy	–	1×10^{-8}	0.0235	0.966	37491	∞	∞

*To obtain the true generalised variances, values in the above table should be multiplied by 10^{-4} except for the "Cauchy" row, where the actual figures are recorded.

TABLE 1.5.4

Efficiency (based on Generalized Variance) of Regression Estimates ($n = 50$; $\sigma^2 = 9$); ($\times10^{-2}$)

Distbn.	Kurt.	p = 1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	0.00	0.01	0.03	0.06	0.09	1.00
Normal	3.0	0.25	0.52	0.75	0.93	1.00	0.01
Con.Norm.	3.5	0.37	0.73	0.94	1.00	0.93	0.00
Con.Norm.	4.0	0.55	0.89	1.00	0.93	0.77	0.00
Con.Norm.	4.5	0.68	1.00	0.99	0.81	0.59	0.00
Con.Norm.	5.0	0.77	1.00	0.81	0.52	0.31	0.00
Con.Norm.	5.5	1.00	0.82	0.44	0.19	0.09	0.00
Laplace	6.0	0.79	1.00	0.81	0.52	0.30	0.00
Cauchy	–	1.00	0.00	0.00	0.00	0.00	0.00

TABLE 1.5.5

Efficiency (based on the Empirical Variances) of the Individual Regression Estimates
($n = 50$; $\sigma^2 = 9$); ($\times 10^{-2}$)

		β_0						β_1						β_2					
Distbn.	Kurt.	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞
Uniform	1.8	0.17	0.24	0.31	0.38	0.45	1.00	0.14	0.20	0.26	0.33	0.38	1.00	0.16	0.22	0.28	0.33	0.38	1.00
Normal	3.0	0.61	0.79	0.90	0.97	1.00	0.22	0.62	0.76	0.87	0.93	1.00	0.22	0.70	0.88	1.00	1.00	1.00	0.23
Con.Norm.	3.5	0.72	0.92	0.99	1.00	0.96	0.16	0.72	0.93	1.00	1.00	1.00	0.18	0.68	0.87	0.93	1.00	0.93	0.15
Con.Norm.	4.0	0.81	0.95	1.00	0.99	0.94	0.13	0.81	1.00	1.00	1.00	1.00	0.14	0.79	0.92	1.00	0.92	0.85	0.12
Con.Norm.	4.5	0.88	1.00	0.99	0.91	0.81	0.10	0.85	1.00	1.00	1.00	0.85	0.10	0.93	1.00	1.00	0.93	0.87	0.11
Con.Norm.	5.0	0.92	1.00	0.92	0.80	0.66	0.07	0.82	1.00	0.90	0.82	0.69	0.07	1.00	1.00	0.90	0.82	0.69	0.07
Con.Norm.	5.5	1.00	0.90	0.73	0.55	0.42	0.04	1.00	1.00	0.75	0.55	0.43	0.04	1.00	1.00	0.86	0.60	0.46	0.04
Laplace	6.0	0.94	1.00	0.91	0.77	0.63	0.05	1.00	1.00	1.00	0.83	0.67	0.06	0.80	1.00	0.89	0.80	0.67	0.06
Cauchy	[illegible]	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00

sample sizes and variances considered.

To examine the effect of the sample size on the relative efficiency (based on generalized variance) of the least squares estimates (relative to the generalized variance of the optimum p). Table 1.1.6 below was constructed.

TABLE 1.1.6
Relative Efficiency (based on Generalized Variance) of Least Squares ($\times 10^{-2}$)

Distbn.	Kurtosis	n = 10	n = 25	n = 50
Uniform	1.8	1.00	0.34	0.09
Normal	3.0	1.00	1.00	1.00
Con.Normal	3.5	1.00	1.00	0.93
Con.Normal	4.0	0.86	0.74	0.77
Con.Normal	4.5	0.79	0.68	0.59
Con.Normal	5.0	0.74	0.40	0.31
Con.Normal	5.5	0.53	0.14	0.09
Laplace	6.0	0.59	0.46	0.30

Examination of the above table shows that for all distributions considered, the relative efficiency of the least squares estimates is greater for $n = 10$ than for $n = 50$. It is clear that as the sample size increases least squares becomes progressively less efficient if the error distribution is non-normal. It appears therefore that the larger the sample size, the more critical is the choice of p.

1.4.6 Empirical Distribution of the L_p- Estimate for the Regression Model

Tables 1.1.7 and 1.1.8 give estimates of skewness and kurtosis for the regression coefficients for the study with sample size $n = 25$ and $\sigma^2 = 9$. It is seen that, apart from the case of the Cauchy, the skewness and kurtosis estimates do not exhibit excessive deviations from normality. In fact, of the class of distributions (excluding the Cauchy) studied, the $p = \infty$ estimator for the uniform distribution has the largest deviations in skewness and kurtosis from those associated with normality. As this estimator is the maximum likelihood estimator for the regression model with disturbances following a uniform distribution, the estimator is asymptotically normal, but clearly an assumption of normality in the finite sample sizes studied is unreasonable. However, it is worth noting that the existence of normality (asymptotically) is often extended to the finite case although in practice the costs of such an assumption are probably small.

The L_p– estimators for other distributions do not appear to exhibit significant deviation from normality and if one is to make the assumption of normality in the case $p = 2$ for these distributions, it is quite tenable, on the basis of these results, to extend this assumption to other values of p. For the case of the Cauchy, it is clear that if one were to use a value of p which was close to optimal for that distribution an assumption of normality would be reasonable.

TABLE 1.1.8

KURTOSIS OF BETA ESTIMATES FOR VARIOUS p

			β_0						β_1						β_2					
Distbn.	Skew	Kurt	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞
Uniform	0	1.8	2.57	2.51	2.44	2.45	2.47	6.20	2.73	2.78	2.65	2.62	2.61	4.91	2.70	2.59	2.69	2.78	2.85	5.57
Normal	0	3.0	2.99	3.02	3.05	3.06	3.04	3.10	3.00	3.04	3.12	3.13	3.11	3.09	2.98	3.96	3.07	3.12	3.12	2.87
Con.Normal	0	3.5	3.15	2.86	2.81	2.83	2.87	3.60	3.15	3.00	3.01	3.10	3.19	3.24	3.05	3.09	3.05	2.99	2.94	3.29
Con.Normal	0	4.0	3.18	2.97	2.90	2.88	2.86	3.21	3.41	3.21	3.18	3.15	3.08	2.84	3.36	3.35	3.18	3.17	3.17	2.92
Con.Normal	0	4.5	2.98	2.88	2.77	2.65	2.60	3.08	3.20	3.04	2.94	2.91	2.88	2.90	3.62	3.57	3.47	3.36	3.20	3.14
Con.Normal	0	5.0	3.58	4.13	4.31	4.26	4.10	2.67	3.44	3.49	3.42	3.32	3.21	3.00	3.44	3.38	3.49	3.42	3.31	3.28
Con.Normal	0	5.5	3.25	3.26	3.33	3.33	3.31	3.04	3.28	3.21	2.99	2.81	2.67	2.91	3.97	3.66	3.39	3.22	3.07	3.01
Laplace	0	6.0	4.66	3.56	3.42	3.30	3.25	3.75	3.45	3.18	3.02	2.95	2.89	3.49	3.43	3.52	3.56	3.53	3.50	3.23
Cauchy	0	–	4.42	4.47	20.84	19.31	30.54	89.88	4.45	4.43	175.97	203.42	240.32	418.16	6.37	6.51	196.21	97.63	96.13	52.51

TABLE 1.1.7

SKEWNESS OF BETA ESTIMATES FOR VARIOUS p

			β_0						β_1						β_2					
Distbn.	Skew	Kurt	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞	1.00	1.25	1.50	1.75	2.00	∞
Uniform	0	1.8	0.07	-0.06	-0.11	-0.16	-0.19	0.52	-0.04	-0.011	-0.00	0.04	0.067	0.35	0.114	0.110	0.13	0.15	0.15	0.37
Normal	0	3.0	-0.13	0.02	-0.02	-0.05	-0.04	-0.05	0.10	0.13	0.18	0.20	0.20	0.08	0.04	0.04	0.07	0.10	0.13	-0.17
Con.Normal	0	3.5	-0.00	0.01	-0.02	-0.05	-0.04	0.16	0.04	-0.01	0.03	0.07	0.09	-0.08	-0.05	-0.05	-0.01	0.02	0.02	0.16
Con.Normal	0	4.0	0.10	0.07	-0.08	0.09	0.10	0.07	0.07	0.04	0.06	0.09	0.08	-0.01	-0.15	0.21	0.10	0.07	0.06	0.09
Con.Normal	0	4.5	-0.28	0.19	-0.11	-0.03	0.03	0.06	0.26	0.15	0.08	0.03	0.00	-0.22	0.32	0.22	0.13	0.09	0.06	-0.01
Con.Normal	0	5.0	-0.25	-0.33	-0.28	-0.23	-0.19	0.03	0.10	0.11	0.07	0.06	0.06	0.10	0.08	0.03	0.01	0.02	0.02	-0.20
Con.Normal	0	5.5	-0.12	-0.01	-0.01	-0.00	0.01	0.19	-0.19	-0.19	-0.14	-0.10	-0.08	-0.05	-0.09	-0.19	-0.23	-0.26	0.30	-0.01
Laplace	0	6.0	-0.14	0.04	-0.02	-0.08	-0.14	-0.06	-0.04	-0.09	-0.05	-0.01	0.03	0.22	-0.05	-0.02	0.02	0.05	0.08	-0.18
Cauchy	0	[illegible]	-0.24	0.12	-1.20	1.62	2.04	-5.42	0.35	0.05	-11.30	10.82	12.12	19.42	-0.42	-0.20	13.02	-5.98	-5.80	-2.52

In conclusion, this study suggests that if one were to use a p close to that which is optimal for the distribution (something which appears possible using formula 1.4.1 with sample kurtosis substituted for population kurtosis), an assumption that the estimates of the $\underline{\beta}$ vector in the regression model are normally distributed is not unreasonable. In addition, the assumption of normality for the estimate of the $\underline{\beta}$ vector in OLS is no more reasonable than an assumption of normality for the L_p- estimates with p calculated from formula 1.4.1.

1.5 CONCLUSIONS

The empirical results of the simulation study suggest the following main conclusions. Firstly, the L_p- estimates of the coefficients in the regression model are unbiased for all $p \geq 1$ (when the error distribution is symmetric). Secondly, a suitable p can be chosen by using the formula:

$$p = \frac{9}{k^2} + 1 \quad .$$

The advantage of such a formula is that it will preclude any ambiguity in the choice of the L_p- norm. Thirdly, the choice of a suitable p becomes more important as the sample size increases.

Finally it is noted that no large deviation from normality is evident in the regression estimates when a suitable p is chosen (except for the extreme case of the Cauchy distribution).

CHAPTER 2

L_p - NORM ESTIMATION AND THE CHOICE OF p : A PRACTICAL APPROACH

2.1 INTRODUCTION

In practical applications it is unlikely that the true kurtosis of the error distribution will be known. In this chapter a similar simulation to the one above is undertaken except that a sample estimate of kurtosis is used as a proxy for the true kurtosis. If the experimental data contains outlier points, then the sample kurtosis will be large and hence use of the proposed equation (1.4.1) will result in a low value of p, causing less weight to be given to these outlier observations. In Chapter 1 it was suggested that in the case of symmetric error distributions p be selected according to the functional relationship:

$$p = \frac{9}{k^2} + 1$$

(where k is the kurtosis of the error distribution).

In general, of course, the kurtosis of the error distribution is unknown and must be estimated from the data. Harter (1977) has proposed that an OLS regression be performed on the sample data and that the residuals from this regression be used to estimate the kurtosis of the error distribution.

In this chapter it is suggested that the true kurtosis can be replaced in the above formula by a sample estimate based on an ordinary least squares fit.The value of p obtained can then be used to determine the L_p- estimates of the regression coefficients. This procedure is compared to OLS and Harter's adaptive procedure. Alternatives to the suggestion of using OLS to obtain estimates of the residuals are also considered. On the basis of a simulation study a final proposal is made for the case where no prior information is available on the distribution of the errors.

2.2 DESIGN OF THE SIMULATION STUDY

A simulation study was performed to examine whether use of a sample estimate of the true kurtosis, rather than the theoretical kurtosis, results in similar comparative advantages over OLS (where the sample estimate of the kurtosis is calculated from the residuals of an initial OLS regression).

The simulation model examined is of the form :

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \qquad i = 1,2,\ldots,n$$

and had similar specifications to that used in Chapter 1 with three sample sizes (n) being chosen as 10,25 and 50. As before 500 iterations were performed for each of the distributions examined. A new set of values for the independent variables was used in this study, randomly selected on the closed interval [0,10] and as with Chapter 1 a set with $|\rho_{X_1,X_2}| < 0.01$ was chosen for each sample size. The errors were drawn from the same set of distributions used in Chapter 1 with the population variance being set at 9.0. These distributions were all symmetric and were chosen so as to cover a fairly wide range of kurtosis. For details of the properties of the distributions considered the reader is referred to Chapter 1.

A problem was encountered in the simulation for the Cauchy distribution, where 1 or 2 data sets for each sample size did not yield feasible results (estimates of β_1 or β_2 greater than 50% from true parameter value), thereby tending to swamp any comparisons between the different methods. This was overcome by transforming the set of uniformly distributed numbers on [0,1] used to generate this distribution to a set uniformly distributed on [0.001, 0.999]. This distribution is denoted by Cauchy†.

For each iteration an OLS regression was performed. An estimate of the kurtosis of the residuals from this regression was then calculated as:

$$b_2 = 3 + \frac{k_4}{k_2^2}*$$

where k_2 and k_4 are unbiased estimates of the second and fourth cumulants respectively (Kendall and Stuart Vol. II (1966)). Using b_2 as an estimate of the kurtosis ($\hat{k}$ say) p was calculated as:

---

* See Appendix A

$$p = \frac{9}{(\hat{k})^2} + 1 \qquad (2.1)$$

The various L_p– estimates of β_0, β_1 and β_2 were then obtained using the method described by Fletcher and Powell (1963) which is part of the IBM Scientific Subroutine Package (1968).

2.3 EXPERIMENTAL RESULTS

2.3.1 Comparison with Ordinary Least Squares

The results are summarised in Tables 2.1, 2.2 and 2.3 for the three different sample sizes used. (As in Chapter 1 the value of the empirical generalized variance of the regression coefficients is used as a yard-stick for comparison purposes.)

TABLE 2.1* (sample size n = 10)

Distribution	Kurtosis	Generalized variance of OLS estimates	Generalized variance of L_p– estimates	Efficiency of OLS relative to L_p– estimates
Uniform	1.8	3.92	3.88	99
Normal	3.0	4.43	5.30	120
Con.Normal	3.5	4.84	5.25	108
Con.Normal	4.0	4.89	5.45	111
Con.Normal	4.5	4.34	4.31	99
Con.Normal	5.0	4.92	5.06	103
Con.Normal	5.5	3.99	3.78	95
Laplace	6.0	5.06	4.73	93
Cauchy†	-	7.75	8.23×10^{-6}	0

*To obtain the actual estimates of the generalized variances, values in the above table should be multiplied by 10^{-3}, except for the "Cauchy" row where the actual figures are recorded.

TABLE 2.2* (Sample size n = 25)

Distribution	Kurtosis	Generalized variance of OLS estimates	Generalized variance of L_p- estimates	Efficiency of OLS relative to L_p- estimates
Uniform	1.8	4.61	2.50	54
Normal	3.0	4.38	4.73	108
Con.Normal	3.5	4.32	4.82	112
Con.Normal	4.0	3.74	3.86	103
Con.Normal	4.5	4.15	3.62	87
Con.Normal	5.0	4.50	2.88	64
Con.Normal	5.5	4.45	1.91	43
Laplace	6.0	4.67	2.90	62
Cauchy†	-	2.2×10^{-1}	7.8×10^{-7}	0

*To obtain the actual estimates of the generalized variances, values in the above table should be multiplied by 10^{-4}, except for the "Cauchy" row where the actual figures are recorded.

TABLE 2.3* (Sample size $n = 50$)

Distribution	Kurtosis	Generalized variance of OLS estimates	Generalized variance of L_p estimates	Efficiency of OLS relative to L_p estimates
Uniform	1.8	6.99	1.99	28
Normal	3.0	7.17	7.72	108
Con.Normal	3.5	6.83	6.61	97
Con.Normal	4.0	7.16	6.56	92
Con.Normal	4.5	6.31	4.79	76
Con.Normal	5.0	8.06	3.01	37
Con.Normal	5.5	7.79	1.72	22
Laplace	6.0	7.02	2.73	39
Cauchy†	-	6.49×10^{-2}	4.00×10^{-8}	0

*To obtain the actual estimates of the generalized variances, values of the above table should be multiplied by 10^{-5}, except for the "Cauchy" row where the actual figures are recorded.

It becomes apparent from examination of these tables that the advantages of the L_p– estimation procedure proposed becomes more pronounced as one increases sample size from 10 to 50. For the nine distributions examined, the OLS estimates are superior (in terms of generalized variance) in four cases when $n = 10$, three cases when $n = 25$ and only one case when $n = 50$. Across the sample sizes considered it is seen that the efficiency of the L_p– estimates relative to the OLS estimate is at worst 83%. On the other hand, the efficiency of the OLS estimates relative to the L_p– estimates is often much lower, the magnitude of this decrease in efficiency increasing with sample size. The efficiencies for sample sizes of 10, 25 and 50 are, for example, 99%, 54% and 28% for the Uniform and 93%, 62% and 39% for the Laplace distributions. For the distributions and sample sizes considered it can, therefore, be argued that the L_p— estimation method is superior to the OLS method (at least in terms of the generalized variance) and that this superiority increases with sample size.

2.3.2 Comparison with Harter's method

It is also of interest to contrast the practical performance of the general L_p method proposed in this paper with that of Harter's adaptive procedure (using the sample kurtosis). Harter (1977) has suggested that the p be selected as a function of kurtosis according to the scheme:

$$\begin{aligned} p &= 1 ; & \hat{k} &> 3.8, \\ p &= 2 ; & 2.2 < \hat{k} &\leq 3.8, \\ p &= \infty ; & \hat{k} &\leq 2.2 \end{aligned}$$

The results of a simulation study using the same model and sample sizes as before are presented in Tables 2.4, 2.5 and 2.6.

TABLE 2.4* (sample size n = 10)

Distribution	Kurtosis	Generalized Variance of Harter's estimate	Generalized Variance of L_p estimates	Efficiency of Harter's estimates relative to the L_p estimates
Uniform	1.8	5.92	3.88	66
Normal	3.0	7.04	5.30	75
Con.Normal	3.5	7.75	5.25	68
Con.Normal	4.0	6.63	5.45	82
Con.Normal	4.5	5.46	4.31	79
Con.Normal	5.0	6.10	5.06	83
Con.Normal	5.5	4.53	3.78	83
Laplace	6.0	6.19	4.73	76
Cauchy†	-	7.19×10^{-7}	8.23×10^{-6}	1140

*To obtain the actual estimates of the generalized variances, values in the above table should be multiplied by 10^{-3} except for the "Cauchy" row where the actual figures are recorded.

TABLE 2.5* (sample size n = 25)

Distribution	Kurtosis	Generalized variance of Harter's estimate	Generalized variance of L_p estimates	Efficiency of Harter's estimates relative to the L_p estimates
Uniform	1.8	3.32	2.50	75
Normal	3.0	7.55	4.73	63
Con.Normal	3.5	6.27	4.82	77
Con.Normal	4.0	5.52	3.86	70
Con.Normal	4.5	4.95	3.62	73
Con.Normal	5.0	3.34	2.88	86
Con.Normal	5.5	1.87	1.91	102
Laplace	6.0	3.42	2.90	85
Cauchy†	-	4.9×10^{-7}	7.8×10^{-7}	159

*To obtain the actual estimates of the generalized variances values in the above tables should be multiplied by 10^{-4}, except for the "Cauchy" row where the actual figures are recorded.

TABLE 2.6* (sample size n = 50)

Distribution	Kurtosis	Generalized variance of Harter's estimate	Generalized variance of L_p estimates	Efficiency of Harter's estimates relative to the L_p estimates
Uniform	1.6	1.28	1.99	155
Normal	3.0	9.79	7.72	79
Con.Normal	3.5	8.28	6.61	80
Con.Normal	4.0	8.69	6.56	75
Con.Normal	4.5	6.26	4.79	77
Con.Normal	5.0	3.18	3.01	95
Con.Normal	5.5	1.43	1.72	120
Laplace	6.0	3.27	2.73	83
Cauchy†	-	3.00×10^{-8}	4.00×10^{-8}	133

*To obtain the actual estimates of the generalized variances, values in the above table should be multiplied by 10^{-5} expect for the "Cauchy" row where the actual figures are recorded.

The L_p– estimates are seen to be superior to the estimates using Harter's method except in the following cases:

(a) the Cauchy distribution, for all sample sizes. The Cauchy is however rather an extreme case and not of considerable interest.

(b) The Uniform distribution when $n = 50$. This result may imply that a larger p should be used for distributions with kurtosis smaller than that of the normal when n is large.

In both (a) and (b) it is worth noting that the difference in efficiency between the L_p– estimates and Harter's estimates is small when compared with the efficiency of the OLS estimates.

(c) The contaminated normal (kurtosis 5.5) when n is 25 and 50.

The difference in efficiency when n is 25 indicates no practical difference between the methods. When n is 50 the difference is presumably related in some way to the special characteristics of this specific distribution and not tail stretch as L_p outperforms Harter in the case of the Laplace distribution and the Contaminated Normal (kurtosis 5.0) for each sample size.

Thus, for the distributions considered, it can be argued that apart from the Cauchy, the L_p– estimation method yields on average superior estimates to those of Harter's method, as regards the generalized variance of the regression estimates.

2.4 THE PROBLEM OF THE ESTIMATION OF THE ERROR DISTRIBUTION

The method for estimating the regression coefficients as described above hinges on the calculation of the kurtosis of the error distribution; the procedure adopted being that due to Harter (1977) where an OLS regression is performed initially and the kurtosis of the residuals used as a proxy for the kurtosis of the error distribution.

Application of OLS for estimation of the regression coefficients can result in poor estimates of the regression coefficients should "outliers" or "inliers" be present in the dependent variable. The chances of obtaining such "outliers" or "inliers" in a finite sample from symmetrical distributions with tails longer or shorter than those of the normal are high; and application of OLS criterion in these cases will yield estimates with high mean square error (MSE) compared with the estimator with minimum MSE. In such cases the appropriateness of the residuals from the OLS fitted line is in doubt.

Essentially then, the problem is a circular one. Good (MSE criterion) $\hat{\beta}$'s yield good residuals but if one needs the residuals to get the $\hat{\beta}$'s what does one do without any prior information? If one knew for example, that on the average one's data would be normally distributed one could then use OLS to get one's initial residual set and if one knew it would be leptokurtic then one could use L_1—regression to get one's initial residuals.

Given no prior information, it makes sense to use a criterion which is identical to that used for the β's to obtain the residuals, i.e. select a p using the kurtosis of the residuals obtained from an L_p— regression with p estimated from some initial set of residuals using the same p selection criterion throughout. One immediately has the problem of how to calculate these initial residuals but it is hypothesized that this 2 step procedure will be insensitive to the calculation of the estimates of the initial residuals. Harter, in comments on Hogg (1974), has already stated that there is some justification for assuming that the final estimate will not be unduly influenced by the estimation procedure for the residuals on which the choice of p is based, when a one step procedure is used. The applicability of this statement will, of course, depend on the criterion for the choice of p and is probably truer in the case of the Harter criterion than in the case of the p proposed in Chapter 1. It was therefore considered of value to investigate this problem, and to use the L_p– procedure adopted in Section 2.2 to obtain estimates using residuals from three distinct initial fits, viz:

(i) an OLS fit,

(ii) an L_1 – fit,

(iii) an L_p – fit where p is obtained from the Kurtosis of a prior OLS fit using formula (2.1)

A simulation study was performed with identical specifications to those of Section 2.2. As in that study, the value of the empirical generalized variance of the regression coefficients is used as a yardstick for comparison purposes. The results are summarized in the following three tables.

TABLE 2.7* (sample size n = 10)

Distribution	Kurtosis	Generalized variance of L_p– estimates under (i)	Generalized variance of L_p– estimates under (ii)	Generalized variance of L_p– estimates under (iii)
Uniform	1.8	3.88	6.72	4.13
Normal	3.0	5.30	6.12	5.87
Con.Normal	3.5	5.25	5.82	5.75
Con.Normal	4.0	5.45	5.16	5.92
Con.Normal	4.5	4.31	3.88	4.54
Con.Normal	5.0	5.06	3.82	5.48
Con.Normal	5.5	3.78	2.49	3.96
Laplace	6.0	4.73	3.87	5.03
Cauchy†	-	8.23×10^{-6}	3.56×10^{-7}	8.44×10^{-7}

*To obtain the actual estimates of the generalized variance, values in the above table must be multiplied by 10^{-3}, except in the "Cauchy" row where the actual figures are recorded.

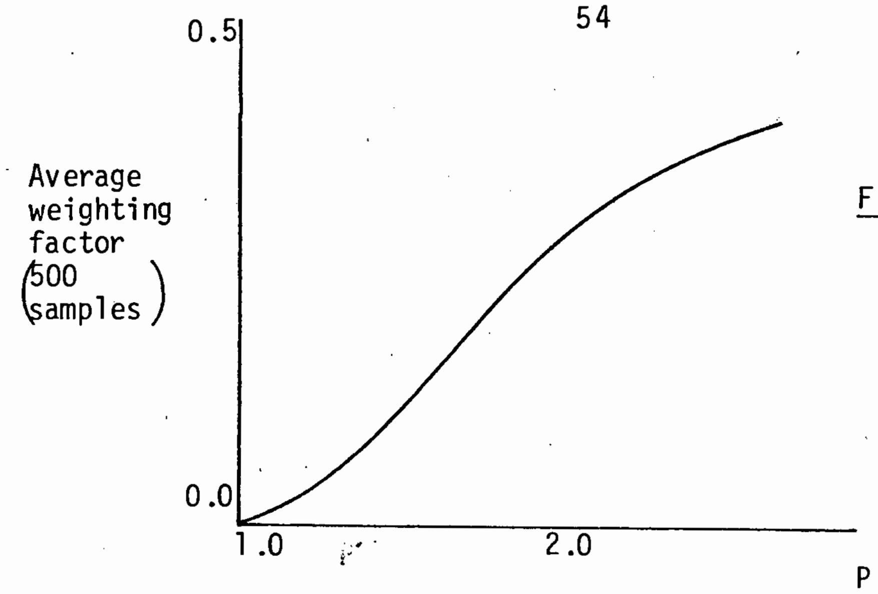


FIGURE 2.5

A plot of the standard deviation of the weighting factor for the 500 samples against p is of the following form, exhibiting the exact relationship for $p = 1.0$ and $p = 2.0$. As p moves away from these values the linear approximation becomes less accurate in the way shown.

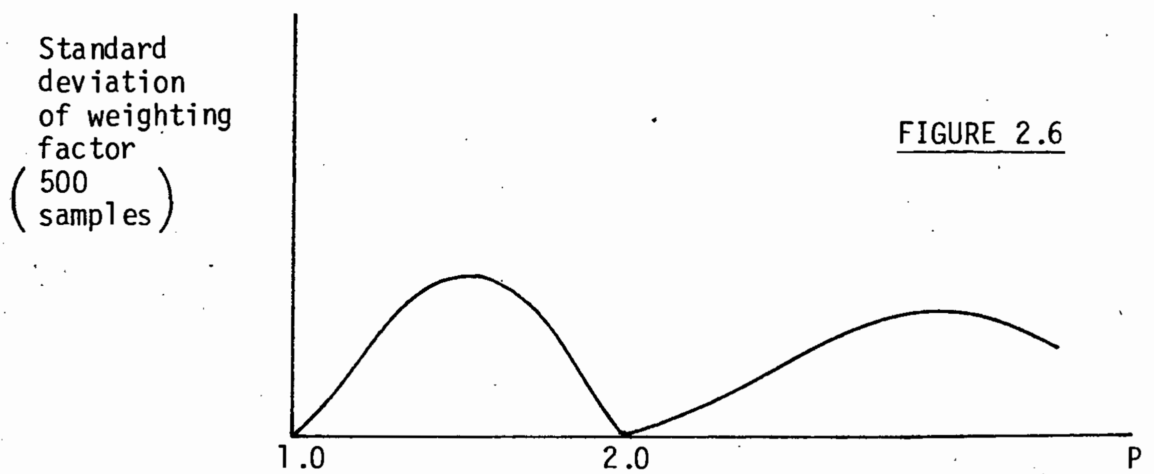


FIGURE 2.6

*To obtain the actual estimates of the generalized variance, values in the above table must be multiplied by 10^{-5} except in the "Cauchy" row where the actual figures are recorded.

One notices that the results using the L_p– estimation procedures under the conditions (i) and (ii) (henceforth known as L_p (i) and L_p(ii)) bear some relation to the properties of the straight-forward L_2– and L_1– estimates. L_1 and L_p(ii) perform better with leptokurtic distributions and L_2 and L_p(i) do better with near-normal distributions. L_p(iii) gives much better results than L_p(ii) for near-normal distribution with sample sizes 25 and 50 and is not much worse (in terms of efficiency) in the cases when L_p(ii) performs best.

For the case of a sample size of 10, L_p(i) performs best for distributions with kurtosis equal to or less than 3.5 and L_p(ii) best for higher kurtosis distributions. If one excludes the case of the Cauchy then a maximin criterion would suggest the use of L_p(i). However, its performance may be considered so bad in the case of the Cauchy that L_p(iii) may even offer the most acceptable alternative. In summary, it is noted that the applicability of the various alternative methods considered does depend to some extent on sample size. In terms of a maximin relative efficiency criterion L_p(iii) would be considered superior for sample sizes of 25 and 50. For the case of a sample size equal to 10 a clear cut choice is more difficult but again the use of L_p(iii) presents no serious disadvantages.

2.5 CONCLUSIONS

In the first section of this chapter it was found that a suitable p could be chosen using the formula

$$p = \frac{9}{\hat{k}^2} + 1 ,$$

where $\hat{k}$ was an estimate of the kurtosis of the error distribution. Based on the empirical generalized variance of the estimates, the results obtained using this formula were found to be generally superior to those which used either ordinary least squares or Harter's adaptive procedure. The procedure giving rise to the most appropriate set of residuals from which $\hat{k}$ is estimated was then considered. On the basis of this study it was proposed that an initial OLS fit to obtain a set of residuals, followed by an L_p-fit to obtain a further set of residuals, and finally another L_p-fit to obtain the coefficients, would give the best performance, at least in terms of a maximin efficiency criterion. When there is no prior information about the distribution of the errors it is therefore believed that the above method will provide good regression estimates for a wide variety of unknown error distributions.

CHAPTER 3

PERFORMANCE OF A GENERALIZED ALGORITHM FOR L_p-NORM REGRESSION ESTIMATES

3.1 INTRODUCTION

In Chapter 2 a method was suggested whereby p is selected as a function of the sample estimate of the kurtosis of the residuals from some initial fit. Among the methods which have been proposed to obtain these L_p regression fits are those due to Fletcher and Powell (1963) and a one dimensional solution by Sposito, Kennedy and Gentle (1977) based on an extension of Schlossmacher (1973).

In this chapter the algorithm of Sposito, Kennedy and Gentle is extended to the m dimensional case and a comparison is made between the method (hence known as WLS (weighted least squares)) and that of Fletcher and Powell (hence known as FP).

3.2 THE PROBLEM

The algorithm is designed to obtain estimates for the β parameters in the model:

$$\underline{y}' = \sum_{j=0}^{m-1} \beta_j \underline{x}_j + \underline{e} \tag{3.2.1}$$

for some m and some error distribution $\underline{e}$. This involves minimizing the L_p-norm.

$$= \sum_{i=1}^{n} |y_i - \sum_{j=o}^{m-1} b_j x_{ij}|^p \qquad (3.2.2)$$

for n observations on $\underline{y}$ and the x's and for some p.

The case $p = 2$ is the ordinary least squares case. For values of $p = 1$ and $p = \infty$ exact linear programming solutions for the b_j may be found.

In the field of research into robust regression, values of p not equal to those above have been shown to be important; in particular $1 < p \leq 3$.

The algorithm of Sposito et al considers minimizing:

$$I = \sum_{i=1}^{n} W_i R_i^2$$

where the R_i are the residuals and W_i are weighting factors. Clearly if the W_i are put equal to 1.0 the least squares solution is obtained.

Using the iterative process of Schlossmacher:

$$I(k+1) = \sum_{i=1}^{n} \frac{1}{|R(k)_i|^{2-p}} (R(k+1)_i)^2.$$

If $|R(k)_i - R(k+1)_i| \approx 0$ for $i = 1, \ldots, n$

then $I(k+1) \approx \sum_{i=1}^{n} |R(k+1)_i|^p$,

and the scheme has converged yielding $\tilde{\beta}$'s which minimize 3.2.2 to some predetermined level of accuracy. The suggestion of Schlossmacher regarding observations close (i.e. less than some predetermined error parameter) to the fitted line has been implemented viz, that these are deleted but may be re-introduced if the residual increases at a later stage.

3.3 CONTROL RETURNS

Control is returned to the main program if

(i) the scheme converges (the Sposito el al convergence criterion was used)

(ii) the norm increases from one iteration to the next. (Note that the suggestion of Porter and Winstanley (1979) has been implemented which tests for two successive increases in the norm when observations have been deleted, before control is returned.)

(iii) The number of iterations exceeds 50.

~~(iv) The correlation matrix of the x's is singular.~~

3.4 COMPARISON BETWEEN WLS AND FP

In order to compare these two methods a simulation study was conducted. The regression model examined was of the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \quad i = 1,\ldots,n \qquad (3.4.1)$$

and had the following specifications:

(i) 3 sample sizes were considered; n = 10; 25 and 50.

(ii) β_0, β_1 and β_2 were selected to be 10.0; 8.0 and -6.0 respectively.

(iii) The values of x_1 and x_2 were chosen randomly from a uniform distribution on the range (0; 10) and X_1 and X_2 were uncorrelated ($|\rho_{X_1X_2}| < 0{,}01$).

(iv) e_i were randomly drawn from a $N(0; 9)$ distribution.

(v) A range of p was used from p = 1.0 through p = 3.0 with interval 0.2.

(vi) For each sample size 100 data sets were generated using 3.4.1 (the x's being fixed in each iteration) and WLS and FP estimates computed for each data set and each p.

(vii) For the case p = 1.0 "exact" estimates were obtained using the method of Barrodale and Young (1966).

3.5 RESULTS OF THE STUDY

(i) It was found that WLS and FP used approximately the same CPU time. WLS tended to use more (15% on average) for p close to 1.0 and 3.0 and FP more (20% on average) for p close to 2.0.

(ii) The storage required by FP was about double that required by WLS.

(iii) <u>Performance when p = 1.0</u>

FP was found not to be viable as an estimation procedure when p = 1.0 as approximately 15% of cases for all sample sizes did not converge to feasible results (estimates of β_1 or β_2 greater than 80% from true parameter value). The results for the WLS were more satisfactory. 1 out of 100 estimates in the cases N = 10 and N = 25 not converging to feasible results (same criterion as above) and 2 out of 100 for N = 50 not yielding feasible results. When these particular cases were omitted for the WLS estimations the following results were obtained.

	β_0	β_1	β_2
N = 10			
Average absolute % error	1,351	0,471	0,319
Worst absolute % error	15,427	3,702	1,891
N = 25			
Average absolute % error	2,421	0,519	0,544
Worst absolute % error	16,111	3,450	4,785
N = 50			
Average absolute % error	1,073	0,188	0,177
Worst absolute % error	4,700	0,723	0,878

3.6 COMPARATIVE PERFORMANCE FOR $1 < p \leq 3,0$

For the cases $p = 1.2$ to $p = 3.0$ no exact results were available for comparison. The WLS and FP methods gave the same estimates (to 4 decimal places) for each iteration of each sample for p in the range 1.4 to 2.6. In the case p is 1·2, an average of 12% (across sample size) of the WLS coefficient sets did not equal the FP coefficient sets but in each case the difference between the two was less than 1%. No evidence was available which suggested that either of the two methods was more reliable for $p = 1.2$ but based on the superior WLS results for the case $p = 1.0$ it is rational to conjecture that WLS is best for $p = 1.2$. For the case $p = 2.8$ approximately 22% of WLS coefficient sets did not equal the FP estimates. In each of these cases WLS terminated due to an increase in the norm and simply yielded the OLS estimates, indicating that the WLS algorithm did not manage to move away from the initial (OLS) estimate, the norm increasing (by more than EPS) at the second iteration. Note that even though WLS did, on some occasions terminate in this way for $p = 1.0$ and $p = 1.2$ (but not for $1.4 \leq p \leq 2.6$) it never yielded OLS estimates in these cases. For $p = 3.0$ WLS yielded OLS estimates for each iteration of each sample size (in the way explained above). For values of $p = 2.8$ and $p = 3.0$ FP did not exhibit lack of convergence in any of the cases examined, the values of the Beta coefficients obtained moving consistently in the same direction as p moved from 1.0 through to the above values.

3.7 CONCLUSIONS

This chapter suggests in the first place that WLS is a useful algorithm for L_p-norm regression fits in the range $1.0 \leq p \leq 2.6$. However as it offers no dramatic time or storage advantages over the exact solution in the case $p = 1.0$ it is suggested that it is only used for

SAMPLE DATA SET (N = 10) WITH WLS AND FP COEFFICIENT ESTIMATES FOR VARIOUS P

X_1	X_2	Y
9,5879	8,0301	39,1870
1,9237	5,3014	-0,9137
5,7588	2,1639	41,3892
2,6606	9,8429	-30,1510
10,0000	1,2005	83,2386
5,2518	10,0000	-10,3460
6,3277	7,9704	9,8553
0,0000	4,5362	-17,2696
9,0393	2,3101	71,1160
0,2391	0,0000	12,8679

(Y simulated from model

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i; \quad i = 1,\ldots,10$

$e_i \sim N(0{,}0;\ 9{,}0)$

where $\beta_0 = 10{,}0$, $\beta_1 = 8{,}0$, $\beta_2 = -6{,}0$)

p		β_0	β_1	β_2
1,0	FP	10,9537	8,0561	-6,1442
	WLS	10,9582	7,9875	-6,3252
	*"Exact"	10,9581	7,9874	-6,3252
1,2	FP	11,0115	7,9988	-6,3322
	WLS	11,0115	7,9988	-6,3322
1,4	FP	11,1724	8,0168	-6,3287
	WLS	11,1724	8,0168	-6,3287
1,6	FP	11,3114	8,0168	-6,3215
	WLS	11,3114	8,0168	-6,3215
1,8	FP	11,4706	8,0050	-6,3159
	WLS	11,4706	8,0050	-6,3159

*"Exact" values for p = 1.0 were computed using the algorithm due to Barrodale and Young (1966) - see Appendix C.

p		β_0	β_1	β_2
2,0	FP	11,7480	7,9745	-6,3120
	WLS	11,7480	7,9745	-6,3120
2,2	FP	12,0134	7,9427	-6,3089
	WLS	12,0134	7,9427	-6,3089
2,4	FP	12,2382	7,9133	-6,3052
	WLS	12,2382	7,9133	-6,3052
2,6	FP	12,4228	7,8869	-6,3006
	WLS	12,4228	7,8869	-6,3006
2,8	FP	12,5737	7,8630	-6,2953
	WLS	11.7480	7,9745	-6,3120*
3,0	FP	12,6978	7,8416	-6,2896
	WLS	11,7480	7,9745	-6,3120*

* OLS estimates

```
      SUBROUTINE WLS(Z,B,OBJ,P,NVE,N,IFAULT,R,SD,RATE,NPO,IT)
C
C       DOUBLE PRECISION ASCII FORTRAN
C
C      EXTENSION OF SPOSITO,V.A. GENTLE,J.E. & KENNEDY,W.J.(1977)
C      ALGORITHM AS110
C      LP-NORM FIT OF A STRAIGHT LINE. APPL.STATIST. 26 114-118
C      DIMENSIONED TO HANDLE UP TO 10 X'S
C
      IMPLICIT REAL*8(A-H,O-Z)
      REAL*8 Z(N,11),ZC(10,10),XC(10,10),YC(10,1),R(1)
      REAL*8 V(3),JC(10),B(10,1),VMEAN(10),DMROW(10),A(10)
C
C        DEFINITIONS OF VARIABLES
C        N = SAMPLE SIZE
C        P = P IN LP-NORM
C        Z = AUGMENTED MATRIX (X,Y) OF OBSERVATIONS
C        ZC = AUGMENTED MOMENT MATRIX
C        XC = X MOMENT MATRIX
C        YC = YX MOMENT VECTOR
C        SD = LP-NORM
C        R = VECTOR OF RESIDUALS
C        RATE = RATE OF CHANGE OF NORM AT TIME OF CONTROL RETURN
C        IT = NUMBER OF ITERATIONS
C        NPO = NUMBER OF POINTS OMITTED
C        IFAULT =  0  ]  CONVERGENCE
C               =  1  ]  NORM INCREASED
C               =  2  ]  MAXIMUM ITERATIONS EXCEEDED
C               =  3  ]  XC NON -SINGULAR
C
C       INITIALIZATION OF PARAMETERS
C
      DATA EPS/1.0D-6/,MAXIT/50/
      V(1)=3
      IFAULT=0
      EPS2=2.*EPS
      SD=0.
      SD4=-10.
      WP=P-2.
C
C
      DO 1 I=1,N
1     R(I)=1.
      DO 11 IT=1,MAXIT
      NPO=0
C
C       INITIALIZATION OF ARRAYS
C
      SUMWT=0.
      DO 13 I=1,NVE
      VMEAN(I)=0.
      DO 13 J=1,NVE
13    ZC(J,I)=0.
C
```

```
C       COMPUTE MATRIX OF SUMS OF SQUARES AND CROSS-PRODUCTS
C       USING HERRAMAN ALGORITHM
C       SEE HERRAMAN,C.(1968) ALGORITHM AS12
C       SUMS OF SQUARES & CROSS PRODUCTS
C       APPL. STATIST. 17,289-292
C
      DO 14 NN=1,N
      DO 15 I=1,NVE
15    DMROW(I)=Z(NN,I)
      ABSRI=ABS(R(NN))
      IF(ABSRI.LE.EPS) GOTO 18
      WEIGHT=ABSRI**WP
      SUMWT=SUMWT + WEIGHT
      DIV=WEIGHT/SUMWT
      DO 16 I=1,NVE
      DMROW(I)=DMROW(I)-VMEAN(I)
      DI=DMROW(I)
      DO 17 J=1,I
      DIJ=DI*DMROW(J)*WEIGHT
      ZC(I,J)=ZC(I,J)+DIJ-DIJ*DIV
      ZC(J,I)=ZC(I,J)
17    CONTINUE
      VMEAN(I)=VMEAN(I)+DI*DIV
16    CONTINUE
      GOTO 14
18    NPO=NPO+1
14    CONTINUE
C
C
      NVE1=NVE-1
      DO 2 I=1,NVE1
      YC(I,1)=ZC(I,NVE)
      DO 2 J=1,NVE1
2     XC(I,J)=ZC(I,J)
C
C
C     COMPUTE REGRESSION COEFFICIENTS
C
C     UNIVAC MATHPACK DOUBLE PRECISION MATRIX INVERSION ROUTINE
C     (ANY SUITABLE ROUTINE MAY BE USED (NOTE XC DESTROYED HERE))
C
      CALL DGJR(XC,10,10,NVE1,NVE1,$98,JC,V)
C
      CALL DMULT(XC,YC,B,NVE1,NVE1,1,10,10,1)
C
      SXM=0.
      DO 21 J=1,NVE1
21    SXM=SXM+B(J,1)*VMEAN(J)
      DO 22 I=1,N
      R(I)=0.
      DO 22 J=1,NVE1
22    R(I)=R(I)+Z(I,J)*B(J,1)
      B(NVE,1)=VMEAN(NVE)-SXM
C
```

```
C         CALCULATE RESIDUALS AND TEST CONVERGENCE
C
       SD2=0.
       ISW=0
       DO 4 I=1,N
       RES=Z(I,NVE)-B(NVE,1)-R(I)
       ABSRI=ABS(RES)
       IF(ABS(ABSRI-ABS(R(I))).GT.EPS2)ISW=1
       SD2=SD2+ABSRI**P
       R(I)=RES
 4     CONTINUE
       RATE=ABS(SD2-SD)/SD2
       IF(ISW.EQ.0) GOTO 99
       IF(IT.EQ.1) GOTO 5
C
C      TEST FOR INCREASE IN NORM
C
C      SEE PORTER,M.A. & WINSTANLEY,D.J.(1979) AS R29 ON AS110
C      APPL. STATIST.,28,112-113
C
       SD3=SD2-SD
       IF(NPO.EQ.0)GOTO 41
       IF(SD3.GT.EPS.AND.SD4.GT.EPS)GOTO 7
       SD4=SD3
       GOTO 5
 41    IF(SD3.GT.EPS)GOTO 7
       SD4=-10.
C
 5     SD=SD2
       DO 51 J=1,NVE
 51    A(J)=B(J,1)
 11    CONTINUE
C
C   FAILED TO CONVERGE IN MAXIMUM ITERATIONS
C
       IFAULT=2
       GOTO 99
C
C   NORM INCREASED, RESTORE BETAS AND R, THEN STOP
C
 7     IFAULT=1
       DO 71 J=1,NVE
 71    B(J,1)=A(J)
       DO 81 I=1,N
       R(I)=0.
       DO 81 J=1,NVE1
 81    R(I)=R(I)+Z(I,J)*B(J,1)
       DO 8 I=1,N
 8     R(I)=Z(I,NVE)-B(NVE,1)-R(I)
       GOTO 99
C
C       XC NON-SINGULAR
C
 98    IFAULT=3
C
 99    RETURN
       END
```

```
      SUBROUTINE DMULT(A,B,C,N,K,M,IMAX,JMAX,KMAX)
C
C      ASCII FORTRAN DOUBLE PRECISION
C
      REAL*8 A(IMAX,JMAX),B(JMAX,KMAX),C(IMAX,KMAX)
C
C      MULTIPLIES MATRIX A BY MATRIX B TO YIELD MATRIX C
C
      DO 1 J=1,M
      DO 1 I=1,N
      C(I,J)=0.
      DO 1 KK=1,K
    1 C(I,J)=C(I,J)+A(I,KK)*B(KK,J)
      RETURN
      END
```

CHAPTER 4

L-ESTIMATION IN THE REGRESSION CASE

4.1 INTRODUCTION

As was discussed in Part II of this thesis, L-estimation has two possible applications in the context of location parameter estimation. The first is the possibility of relating L-estimators to L_p- estimators as one approach to determining distributional properties of L_p- estimators. This appeared possible for the location parameter case because, in the first place, the distribution of L - estimators was at least algebraically feasible in finite samples and relatively straightforward for the asymptotic case and, in the second place, one may construct L-estimators which are useful approximations to L_p- estimators. The second reason for studying L-estimators was the more obvious one of its use as a location parameter estimator in its own right.

The application of L-estimation to the regression model is however less straightforward as will become clear below. A tentative scheme is proposed on the basis of theoretical work done for the uniform distribution. Due to the difficult theoretical problems involved in extending L-estimation to the regression situation, the scheme is incomplete as far as practical applications are concerned. The theoretical results below indicate, however, that further work done in this area could be rewarding.

4.2 OPTIMAL L-ESTIMATION FOR THE ESTIMATION OF THE LOCATION PARAMETER

Lloyd (1952) has shown that by considering a random sample $X_1, X_2, \ldots, X_n$ from a known distribution (which depends on scale and location parameters alone) as an ordered sample $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, then generalized least squares applied to this dependent ordered sample yields an estimate of location which is at least as efficient as the Gauss-Markov BLUE estimator applied to the unordered sample. Equality in efficiency will be obtained when the row totals of the covariance matrix of the ordered sample are all equal. In general this will not be the case; two cases for which it is true are the normal and exponential distributions.

As an example of his methodology Lloyd examines the case of the uniform distribution. (The treatment for other distributions is exactly the same.)

The one dimensional model which Lloyd examines is thus:

$$\underline{x} = \underline{1}\theta + \underline{e}$$

where $\underline{x}$ is an $n \times 1$ vector of the form

$$\begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{pmatrix},$$

$\underline{e}$ is a vector of independently distributed error terms with :

$$E(\underline{e}) = \underline{0}$$

$$E(\underline{ee}') = \sigma^2 I$$

and each component of $\underline{e}$ has a uniform distribution. θ represents the location parameter. Transformation to an ordered (vector form) yields:

$$\underline{Y} = \underline{1}\theta + \underline{u}$$

where $\underline{Y} = \begin{pmatrix} X_{(1)} \\ \cdot \\ \cdot \\ \cdot \\ X_{(n)} \end{pmatrix}$

and $\underline{u}$ the vector of ordered disturbance terms.

$\therefore$ $E(\underline{uu}') = \sigma^2 \underline{V}$ ($\underline{V}$ known, positive definite).

As discussed in Part I, generalized least squares applied to the ordered sample yields

$$\hat{\theta} = (\underline{1}'V^{-1}\underline{1})^{-1}(\underline{1}'V^{-1}\underline{Y})$$

the best linear (in the ordered sample) unbiased estimator of θ.

4.2.1 Efficiency of ordinary least squares in relation to Lloyd's estimator

The most general measure of the efficiency of the OLS estimator, say $\hat{\theta}_{OLS}$, in relation to another estimator, say $\hat{\theta}^*$, is the ratio of the determinants of the covariance matrices:

$$\frac{|\text{Cov } \hat{\theta}^*|}{|\text{Cov } \hat{\theta}_{OLS}|} .$$

For this particular case this expression equals:

$$\frac{\sigma^2(\underline{1}'V^{-1}\underline{1})^{-1}}{\sigma^2(\underline{1}'\underline{1})^{-1}} \qquad (4.2.1.1)$$

Lloyd has shown that for the particular case of the uniform distribution the inverse of V may be written as:

$$\Omega = V^{-1} = \frac{(n+1)(n+2)}{12} \begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & & -1 & 2 \end{pmatrix} .$$

(4.2.1.1) can thus be written as :

$$\frac{(\underline{1}'\underline{1})}{(\underline{1}'\Omega\underline{1})}$$

$$= \frac{6n}{(n+1)(n+2)} .$$

For n = 2 the estimators are the same and therefore the efficiency of OLS is equal to 1.0. However for n > 2 the efficiency of OLS is less than 1.0 and decreases as n increases.

4.3 EXTENSION OF LLOYD'S METHOD TO THE REGRESSION CASE (OPTIMAL L-REGRESSION)

In the location parameter estimation case, there is a

1-1 relationship between the ranked vectors $\underline{Y}$ and $\underline{u}$. If one is to attempt to extend these procedures to the regression situation, one is faced with the problem that no such relationship holds between the vectors $\underline{Y}$ and $\underline{u}$. Since the ranking of the true error vector will always be unknown one is faced with the problem of establishing this ranking in some way. This problem will be dealt with below; in fact it is not unlike the circular problem of estimating the kurtosis of the residuals in L_p-norm regression (Chapter 2 above), where the residuals are a function of p, with p itself being a function of the kurtosis of these residuals.

4.3.1 The efficiency of OLS for the regression case

To motivate this study the efficiency of such a method, if the true ranking of the error vector was known, will be established for the uniform distribution for which an explicit form for Ω (the inverse of the covariance matrix of the ranked sample) has been derived (see above). Sarhan (1954) has also derived the explicit form of Ω for the exponential distribution but this case will not be dealt with because the exponential is asymmetric. Sarhan also treats the case of the Laplace but, although he evaluates Ω for small values of n, he gives no explicit formulation of Ω as a function of n and one may assume that there is a problem of algebraic tractability.

Before the efficiency of OLS *vis-a-vis* the proposed estimator is established we need the following theorem. The theorem draws from the ideas of Watson (1967) who was tackling a similar problem, that of establishing a lower bound for the efficiency of OLS in the case where the disturbances are autocorrelated.

Theorem 4.1 *The upper and lower bounds of the expression* $\frac{|X'X|}{|X'\Omega X|}$, *where* X *is an* $n\times k$ *matrix of rank* k *and* Ω *is an* $n\times n$ *positive definite matrix, are given by*

$$\frac{1}{\lambda_1\lambda_2\ldots\lambda_k} \quad \text{and} \quad \frac{1}{\lambda_{n-k-1}\ldots\lambda_{n-1}\lambda_n}$$

respectively, where λ_i ; $i=1,\ldots,n$ *are the ranked (smallest to largest) eigenvalues of* Ω.

*Proof** *In order to prove this theorem the notion of the* k^{th} *compound of an arbitrary matrix,* A *say, is introduced.*

If A *is an* $m\times n$ *matrix then the* k^{th} *compound of* A, *denoted by* $A^{(k)}$ *where* $k \leq m$ *and* $k \leq n$ *is the* $\binom{m}{k}\times\binom{n}{k}$ *matrix of all possible* $k\times k$ *minors of* A *arranged in lexicographical order. Clearly* $A^{(1)} = A$ *and*

$$A^{(m)} = |A| \text{ if } m = n.$$

It can be shown (MacDuffee (1946)) that

$$(AB)^{(k)} = (A)^{(k)}B^{(k)}$$

if the k^{th} *compounds are defined.*

---

*Professor D Nel (UOVS) has remarked that an alternative proof provided by Khatri, C.G. (1978) can be modified to obtain the above result.

$$= \frac{(|\sigma^2 X'\Omega X)^{-1}|}{|\sigma^2(X'X)^{-1}|}, \quad \text{where} \quad \Omega = V^{-1}$$

$$= \frac{|X'X|}{|X'\Omega X|}$$

and by our theorem,

$$\frac{1}{\lambda_n \lambda_{n-1} \cdots \lambda_{n-k+1}} \leq \frac{|X'X|}{|X'\Omega X|} \leq \frac{1}{\lambda_1 \lambda_2 \cdots \lambda_k}$$

4.3.2 Attainment of the bounds for the maximum and minimum efficiency of OLS

Considering our ratio of determinants

$$\frac{|X'X|}{|X'\Omega X|},$$

we let Q be orthogonal such that

$Q'\Omega Q = \Lambda$ (diagonal matrix of eigenvalues of Ω)

Letting $L = Q'X$

$X = QL$

$$\therefore \frac{|X'X|}{|X'\Omega X|} = \frac{|L'L|}{|L'\Lambda L|}$$

The upper bound will be attained when

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & 1 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}_{n\times k} \leftarrow k^{th} \text{ position}$$

(assume the elements of the diagonal of Λ are ranked - smallest to largest).

connection, "......, *when a data set involves a shift or a bias, it is not at all clear whether that bias should be charged against the performance of any estimator. To do so implies that the estimator should be able to see beyond the data to the 'true' value of the physical quantity ......*"

A second problem relates to that of comparing different estimators of sets of data with vastly differing variability, necessarily giving rise to sets of estimates with widely differing variances. Stigler uses a "robust" measure of variability for the j^{th} data set (s_j) - the average of the absolute deviations of the set of estimates obtained for that data set. His relative error is then calculated as the absolute deviation of the estimate divided by this measure of spread. Small absolute errors may thus be associated with large relative errors.

Given constraints imposed by the data at hand, it appears, however, that the method of comparison adopted gives an adequate portrayal of the relative performance of the estimators examined for a specific real life situation.

3.1.5.2 Results for the previously proposed L_p-estimators

This method of comparison (with the reservations outlined above) was thus applied when evaluating the L_p-estimators (3.1.3) as members of the class of robust estimators. The same values for the mean absolute deviation of the estimates for each data set were used; as Stigler says

4.3.4 Practical implementation of optimal L-regression

As was pointed out earlier there is a real problem regarding the ranking of the $\underline{y}$ to correspond with the true ranked disturbance term even if one knew the form of the Ω matrix. Two possible schemes are suggested for the practical implementation of L-regression:

(i) Initially perform an OLS (or L_1) fit and rank the y's in the order of the residuals from this fit and then perform an L-regression. In fact some sort of iterative scheme could probably be used after the initial fit with successive L-regressions being performed until a $\hat{\underline{\beta}}$ is obtained for which the residuals are ranked.

(ii) Perform L-regressions for all possible permutations of $\underline{y}$ and select the $\hat{\underline{\beta}}$ for which the residuals are ranked (if one exists).

It was stressed before that no explicit form of an Ω matrix has been derived for any symmetric distribution with kurtosis larger than 3.0. If one could be derived for the Laplace, for example, this method would lend itself well to an adaptive scheme. For example; perform an initial OLS (or L_1) fit - if the kurtosis of the residuals is below say 2.0 use the above uniform adjustment, if above say 4.0 use the Laplace adjustment, otherwise stick to OLS.

Such a study would of course demand a large simulation for a range of distributions before any definite views can be taken on its practical applicability. The theoretical results seem encouraging but the establishment of Ω matrices for distributions other than the uniform appears to be algebraically difficult.

4.4 EXTENSION OF L-ESTIMATION TO THE REGRESSION CASE USING THE METHOD OF PERCENTILE PLANES

An alternative extension of L-estimation to the regression case of Hogg (1979), outlined in Part I, is proposed here. Hogg has remarked that percentile planes defined by certain M-estimators represent regression estimators which are analogous to straightforward percentile L-estimators in the location parameter case. One obtains the ($100p^{th}$) percentile plane by using the ρ function:

$$\rho(x) = -(1-p)x, \quad x < 0$$
$$= px, \quad x \geq 0 .$$

Denoting the $\hat{\underline{\beta}}$ vector representing the estimate of the $100p^{th}$ percentile plane by $\hat{\underline{\beta}}_p$, it is suggested that a simple extension of the method of L-estimation with some adaptive weight distribution could be used for the regression situation. The estimator, $\hat{\beta}_L$ say, would then be in the form of:

$$\hat{\beta}_L = \sum_{p=1}^{99} w_p \hat{\underline{\beta}}_p$$

The w_p could be determined in a manner similar to the adaptive scheme of Part II where the weights are determined from a beta function in a way determined by the kurtosis of the sample. The same problem found in Chapter 2, namely finding the kurtosis of the residuals is encountered and proposed solutions to the problems for L_p – estimation, such as using the residuals from an initial robust L_1 – fit, could be implemented here.

CONCLUSION TO PART III

The robust methods for the estimation of the location parameter are more obvious and straightforward to implement than those for the estimation of $\underline{\beta}$ in the regression model. In addition, for the regression case, one has to consider a number of problems, such as multicollinearity, and model specification, which do not relate to the existence of wild points in the error vector. For example, the greater proportion of the work done in this section took the form of simulation studies and conclusions drawn from these studies are limited by the constraints necessarily imposed on these studies, such as the 2 dimensional regression model considered and the particular sample sizes studied.

It is, however, hoped that the work done in this area represents a contribution in a challenging area of theoretical and methodological research.

SUMMING UP AND THE DIRECTION OF FUTURE RESEARCH

The simulated results for the various procedures proposed for the estimation of θ, the location parameter, indicate that these procedures merit serious consideration in this field of robust estimation. The results for the estimation of $\underline{\beta}$ in the regression model are also good but the number of alternative estimators considered in the comparative studies was rather limited. The evaluation of the usefulness of robust procedures in the regression case is, however, more time consuming and one has to make a number of restrictive assumptions (such as the number of independent variables in the model) which are not necessary to make in the case of location parameter estimation. The bulk of one's research effort in such a field is thus often directed to the case of location parameter estimation. One direction of future research is thus to undertake a much more extensive comparative simulation study for the regression model similar to that undertaken for the location parameter in section 1.5, Part II.

As discussed above, alternative measures of tail stretch (other than kurtosis) have been proposed in the literature e.g. Hogg's Q-statistic. The inclusion of such alternatives in the simulation studies alongside kurtosis would certainly make the research more complete. However the results as they stood were so encouraging that it was thought unncessary to investigate alternative measures of tail stretch until other seemingly more important areas of research had been given attention. Quite clearly, though, it is an important area for future research.

Finally, a point which has been stressed before. Namely, that although the problem of obtaining the exact distributional properties of the L_p-norm estimator of the location parameter may well be intractable, the work of Part II implies that approximation by L-estimators is a feasible approach and worth pursuing. However as stated above, since the proposed L-estimator performs as well as the L_p-estimator for the location case and since so much more is known about the distributional properties of L-estimators, it may well be a better idea to direct one's attention towards L-estimators to the exclusion of L_p-estimation for the location case.

No obvious parallel to L_p-estimation exists for the regression case and the problem of deriving the distributional properties of the L_p-estimates of the $\underline{\beta}$ vector cannot be approximated in the same way as for the location case. Moreover, the distributional properties of the L_p-estimates of $\underline{\beta}$ may well increase in complexity as the dimension of the model increases. This area of research is certainly a very challenging one.

B I B L I O G R A P H Y

AITKEN, A.C. (1935). On least squares and linear combinations of observations. Proc. Roy. Soc. Edin., A, 55, 42.

ANDREWS, D.F., BICKEL, P.J., HAMPEL, R.F., HUBER, P.J., ROGERS, W.H., and TUKEY, J.W. (1972). Robust Estimates of Location Survey and Advances. New Jersey : Princeton University Press.

ANDREWS, D.F. (1974). A Robust Method for Multiple Linear Regression. Technometrics, 16, 523-531.

APPA, G., and SMITH, C. (1973). On L_1 and Chebychev Estimation. Math. Program, 5, 73-87.

ASHAR, V.G. and WALLACE, T.D. (1963). A Sampling Study of Minimum Absolute Deviations Estimators. Opns. Res., 11, 747-758.

BARNETT, V.D. (1966). Order statistics estimators of the location of the Cauchy distribution. J. Amer. Statist. Assoc., 61, 1206-18.

BARRODALE, I., and YOUNG, A. (1966). An Algorithm for Best L_1 and L_∞ Linear Approximations on a discrete set. Numerical Math. 8, 295-306.

BEATON, A.E. and TUKEY, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band - spectroscopic data. Technometrics, 16, 147-85.

BICKEL, P.J. (1965). On some Robust Estimates of Location. Ann. Math. Stats., 36, 847-58.

BLATTBERG, R., and SARGENT, T. (1971). Regression with Non-Gaussian Stable Distrubances. Econometrica, 39, 501-510.

BLOOMFIELD, P., and STEIGER, W. (1977). Least Absolute Deviations Curve Fitting. Technical Report No. 137, Ser. 2, Department of Statistics, Princeton University.

BOX, G.E.P. (1953). Non-Normality and Tests on Variances. Biometrika, 40, 318-335.

BOX, G.E.P., and MULLER, M.E. (1958). A note on the generation of random normal deviates. Ann. Math. Statist., 29, 610-611.

CHAN, L.K., and RHODIN, L.S. (1980). Robust Estimation of Location using optimally chosen sample quartiles. Technometrics, 22, 225-237.

CROW, E.L. and SIDDIQUI, M.M. (1967). Robust estimation of location. J. Amer. Statist. Assoc., 62, 353-389.

de WET, T., and van WYK, J.W.J. (1979a). Efficiency and Robustness of Hogg's adaptive Trimmed Means. Commun. Statist. - Theor. Meth., A8(2), 117-128.

de WET, T., and van WYK, J.W.J. (1979b). Large Sample Properties of Hogg's Adaptive Trimmed Means. South African Statist. J., 13, 53-70.

DENNIS, J.E., and WELSCH, R.E. (1976). Techniques for non-linear least squares and robust regression. 1976 Proc. Amer. Statist. Assoc. Statist. Comp. Section. Washington, D.C. : American Statistical Association, 83-87.

DUTTNER, R. (1977). Numerical Solution of Robust Regression Problems : Computational Aspects, a Comparison. J. Statist. Comp. Sim., 5, 207-238.

EDGEWORTH, F.Y. (1887). On observations relating to several Quantities. Hermathena, 6(13), 279-285.

EDGEWORTH, F.Y. (1888). On a new method of reducing observations relating to several quantities. Phil. Magazine, 25 (Ser 5), 185-191.

EDGEWORTH, F.Y. (1923). On the use of medians for reducing observations. Phil. Magazine, 66 (ser 6), 1074-1088.

FAIR, R.C. (1974). On the robust estimation of econometric models. Ann. Econ. Social Measurement, 3, 667-78.

FISHER, R.A. (1921). On the mathematical foundations of theoretical statistics. Phil. Trans., A, 222, 309.

FISHER, D.F. (1972). Classification, Selection and Testing procedures for Asymmetric Distributions. Unpublished Ph.D. thesis. Department of Statistics, University of Iowa.

FLETCHER, R., and POWELL, M.J.D. (1963). A rapidly convergent method for minimization. Computer J. 6, 163-168.

FORSYTHE, A.B. (1972). Robust Estimation of Straight Line Regression Coefficients by Minimizing Pth Power Deviations. Technometrics, 14, 159-166.

FOURIER, J.B.B. (1824). Solution d'une question particulière au calcul des inegalités, second extrait. Histoire de l'Académie de Sciences pour 1824, xlvii-lv. Reprinted in Oeuvres de Fourier (ed. Gaston Darboux) Vol 2, 325-328, Gauthier-Villars, Paris, 1890.

GASTWIRTH, J.L. (1966). On robust procedures. J. Amer. Statist. Assoc., 61, 929-948.

GASTWIRTH, J.L. and RUBIN, H.R. (1969). On Robust Linear Estimators. Ann. Math. Statist., 40, 24-39.

GASTWIRTH, J.L. and COHEN, M.L. (1970). Small sample behaviour of some robust linear estimators of location. J. Amer. Statist. Assoc., 65, 946-973.

GENTLE, J.E., KENNEDY, W.J. and SPOSITO, V.A. (1977). On Least Absolute Values Estimation. Commun. Statist - Theor. Meth., A6(9), 839-845.

GOVINDARAJULU, Z. (1966). Best linear estimates under symmetric consoring of the parameters of a double exponential population. J. Amer. Statist. Assoc., 61, 248-258.

HAMPEL, F. (1974). The influence curve and its role in robust estimation. J. Amer. Statist. Assoc., 69, 393-393.

HARTER, H.L. (1972). The Method of Least Squares and some Alternatives, ARL 72-019, Ad 75221, Aerospace Research Laboratories, U.S. Airforce, Wright-Patterson Air Force Base, Ohio.

HARTER, H.L. (1974). The Method of Least Squares and some Alternatives - Part I. Int. Statist. Rev., 42, 147-174.

HARTER, H.L. (1974). The Method of Least Sqaures and some Alternatives - Part II. Int. Statist. Rev., 42, 235-264.

HARTER, H.L. (1974). Comment on Hogg (1974) in J. Amer. Statist. Assoc., 69, 923-925.

HARTER, H.L. (1975). The Method of Least Squares and some Alternatives - Part III. Int. Statist. Rev., 43, 1-44.

HARTER, H.L. (1975). The Method of Least Squares and some Alternatives - Part IV. Int. Statist. Rev., 43, 125-190.

HARTER, H.L. (1975). The Method of Least Squares and some Alternatives. Addendum to Part IV. Int. Statist. Rev., 43, 273-278.

HARTER, H.L. (1975). The Method of Least Squares and some Alternatives - Part V. Int. Statist. Rev., 43, 269-272.

HARTER, H.L. (1977). Nonuniqueness of Least Absolute Values Regression. Commun. Statist. - Theor. Meth., A6(9), 829-838.

HARTER, H.L., MOORE, A.H. and CURRY, T.F. (1979). Adaptive robust estimation of Location and Scale parameters of Symmetric populations. Commun. Statist. - Theor. Meth., A8, 1473-1491.

HARVEY, A.C. (1978). On the unbiasedness of Robust Regression Estimators. Commun. Statist. - Theor. Meth., A7(8), 779-783.

HAVLICEK, J. (1968). Contemporary Agricultural Marketing. Editor: I. Dubow, University of Tennessee Press, Knoxville, USA.

HERRAMAN, C. (1968). Algorithm AS12. Sums of squares and products matrix. Appl. Statist., 17, 289-292.

HILL, I.D., HILL, R. and HOLDER, R.L. (1976). Algorithm AS99. Fitting Johnson curves by moments. Appl. Statist., 25, 180-189.

HILL, I.D. (1976). Algorithm AS100. Normal-Johnson and Johnson-Normal Transformations. Appl. Statist., 25, 190-192.

HODGES, J.L., Jr. and LEHMANN, E.L. (1963). Estimates of location based on rank tests. Ann. Math. Statist., 34, 598-611.

HOGG, R.V. (1967). Some Observations on Robust Estimation. J. Amer. Statist. Assoc., 62, 1179-86.

HOGG, R.V. and CRAIG, R.T. (1970). Introduction to Mathematical Statistics. Macmillan, London.

HOGG, R.V. (1972). More Light on the Kurtosis and Related Statistics. J. Amer. Statist. Assoc., 67, 422-24.

HOGG, R.V. (1974). Adaptive Robust Procedures : A Partial Review and Some Suggestions for Future Applications and Theory. J. Amer. Statist. Assoc., 69, 909-923.

HOGG, R.V. (1977). An Introduction to Robust Procedures. Commun. Statist. - Theor. Meth. A6(9), 789-794.

HOGG, R.V. (1979). Statistical Robustness : One view of its use in Applications Today. Amer. Statist., 33, 108-115.

HOLLAND, P.W. and WELSCH, R.E. (1977). Robust Regression using interactively Reweighted Least Squares. Commun. Statist. - Theor. Meth. A6(9), 813-827.

HUBER, P.J. (1964). Robust estimation of a location parameter. Ann. Math. Statist., 35, 73-101.

HUBER, P.J. (1970). Studentized Robust Estimates, in M.L. Puri, ed., Non parametric Techniques in Statistical Inference. London, New York : Cambridge University Press, 453-463.

HUBER, P.J. (1972). The 1972 WALD Lecture Robust Statistics : A Review. Ann. Math. Statist., 43, 1041-1067.

HUBER, P.J. (1973). The 1972 WALD Memorial Lectures, Robust Regression : Asymptotics, Conjectures and Monte Carlo. Ann. Statist., 5, 799-821.

HUBER, P.J. (1975). Robust methods of estimation of regression coefficients. Presented to 2nd Int. Summer School on Problems of Model Choice on Regres. Anal. at Rheinhardsbrum, G.D.R., Nov 8-18.

IBM Corporation, (1968). IBM System/360. Scientific Subroutine Package.

JAECKAL, L.A. (1971). Robust Estimates of Location : Symmetry and Asymmetric Contanination. Annals of Mathematical Statistics, 42, 1020-34.

JOHNSON, N.L. (1949). Systems of frequency curves generated by methods of translation. Biometrika, 35, 149-176.

JOHNSON, N.L. and KOTZ, S.K. (1970). Continuous Univariate Distributions - I and II. Houghton Mifflin Company, Boston.

JOHNSTON, J. (1972). Econometric Methods (2nd Edition). McGraw-Hill, New York.

KARST, O.J. (1958). Linear Curve Fitting Using Least Deviations. J. Amer. Statist. Assoc., 53, 118-132.

KENDALL, M. and STUART, A.J. (1966). The Advanced Theory of Statistics, Vols. I and II. Griffin.

KHATRI, C.G. (1978). Some Optimization Problems with applications to Canonical Correlations and Sphericity Tests. Journal of Multivariate Analysis, 8, 453-467.

KIOUNTOUZIS, E.A. (1971). Optimal L_p Approximation. Techniques and Data Analysis. Extrait du Bull. De La Soc. Mathematique de Grece Novelle Serie, Tome 12, Fasc. 1, 191-206.

KIOUNTOUZIS, E.A. (1972). Mathematical Programming and Best Linear L_p Approximation. Extrait Du Bull. De La Soc. Mathematique De Grece Novelle Serie, Tome 13, Fasc. 1, 46-57.

KIOUNTOUZIS, E.A. (1973). Linear Programming Techniques in Regression Analysis. Applied Statistics, 22, 69-73.

LLOYD, E.H. (1952). Least-squares estimation of location and scale parameters using order statistics. Biometrika, 39, 88-95.

MACDUFFEE (1946). The Theory of Matrices. New York, Chelsea.

MACLAREN, M.D. and MARSAGLIA, G. (1965). Uniform random number generators. J. Ass. Comput. Mach., 12, 83-89.

NARULA, S.C. and WELLINGTON, J.F. (1977). Prediction Linear Regression and Minimum Sum of Relative Errors. Technometrics, 19, 185-190.

PORTER, M.A. and WINSTANLEY, D.J. (1979). Remark AS R29. Remarks on AS110 : L_p Norm fit of a Straight Line. Appl. Statist., 28, 112-113.

*

RHODES, E.C. (1930). Reducing Observations by the Method of Minimum deviations. Phil. Magazine, 9 (ser. 7) 974-992.

RICE, J.R. (1964). The Approximation of Functions. Vol. 1. Addison-Wesley.

SADOVSKI, A.N. (1974). Algorithm AS 74. L_1 norm fit of a straight line. Appl. Statist., 23, 244-248.

SARHAN, A.E. (1954). Estimation of the mean and standard deviation by order statistics. Ann. Math. Statist., 25, 317-328.

SARHAN, A.E. (1955a). Estimation of the mean and standard deviation by order statistics. Ann. Math. Statist., 26, 505-511.

SARHAN, A.E. (1955b). Estimation of the mean and standard deviation by order statistics. Ann. Math. Statist., 26, 576-592.

*PRESCOTT, P. (1978). The Robustness of Adaptive Trimmed Means. Paper presented at the European Meeting of Statisticians at Oslo.

APPENDIX A

A NOTE ON THE MEASUREMENT OF SKEWNESS AND KURTOSIS

For some random variable X with cumulative distribution function $F(X)$ we define in the normal way:

$$\mu_r = E(X-E(X))^r$$

$$\kappa_r = \left|\left(\frac{d}{dr}\right)^r \left[\text{Log } M_X(t)\right]\right|_{t=0}$$

where $M_X(t) = E(e^{Xt})$

μ_r is known as the r^{th} population moment about the mean and κ_r as the r^{th} cumulant.

$$\text{Define } \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{\kappa_4}{(\kappa_2)^{\frac{3}{2}}} \tag{A.1}$$

$$\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{\kappa_3}{(\kappa_2)^{\frac{3}{2}}} \tag{A.2}$$

β_2 and $\sqrt{\beta_1}$ are known as the coefficients of skewness and kurtosis respectively.

The two estimators of skewness and kurtosis used throughout this thesis are:

$$b_2 = 3 + \frac{k_4}{k_2^2} \tag{A.3}$$

$$\sqrt{b_1} = \frac{k_3}{k_2^{\frac{3}{2}}} \tag{A.4}$$

where k_r is known as the r^{th} k-statistic and k_r is an unbiased estimator of κ_r.

More explicitly:

$$k_2 = \frac{n}{(n-1)} m_2$$

$$k_3 = \frac{n^2}{(n-1)(n-2)} m_3$$

$$k_4 = \frac{n^2}{(n-1)(n-2)(n-3)} \{(n+1)m_4 - 3(n-1)m_2^2\}$$

where $m_r = \sum_{i=1}^{n} \frac{(x_i-\bar{X})^r}{n}$

Thus the numerator terms in formulas (A.3) and (A.4) (k_4 and k_3) are unbiased estimators of κ_4 and κ_3 but this is not true of the denominator terms since

$$E(k_2^2) \neq \kappa_2^2$$

$$E(k_2^{\frac{3}{2}}) \neq \kappa_2^{\frac{3}{2}}$$

An unbiased estimator of $\kappa_2^2 \left[(k_2^*)^2 = \frac{n-1}{n+1}\left(k_2^2 - \frac{k_4}{n}\right)\right]$ can be devised but k_4 and k_2^2 $(k_2^*)^2$ are only independent if X has a normal distribution. In that case:

$$E\left(\frac{k_4}{(k_2^*)^2}\right) = \frac{E(k_4)}{E(k_2^*)^2} = \frac{\kappa_4}{\kappa_2^2} = 0$$

and

$$E\left(\frac{k_4}{(k_2)^2}\right) = 0$$

so both are unbiased and there is no clear advantage in using k_2^*. In general however k_4 and k_2^2 are not independent and b_2 will not be an unbiased estimator of β_2.

Common alternative estimators for β_2 and $\sqrt{\beta_1}$ are the ratios of sample moments

$$\frac{m_4}{(m_2)^2} \quad \text{and} \quad \frac{m_3}{(m_2)^{\frac{3}{2}}} \quad \text{respectively} \qquad (A.5)$$

Results are tiven below of a study for which the average MSE of the 2 estimators of kurtosis cited above was calculated for a range of distributions and sample sizes. 500 samples of size 10, 30, 50 were simulated and the kurtosis using the two formulas calculated for each sample along with the mean square error of this estimate.

AVERAGE MSE OF ESTIMATED KURTOSIS
A.3 in bold face
A.5 in italics

	True kurtosis	Sample Size		
		10	30	50
Uniform	1.8	0.302 *0.977*	0.073 *0.103*	0.038 *0.047*
Normal	3.0	0.830 *1.544*	0.682 *0.940*	0.357 *0.413*
Con.Normal	3.5	1.574 *2.242*	1.093 *1.320*	0.713 *0.806*
Con.Normal	4.0	2.341 *2.746*	1.372 *1.525*	1.035 *1.096*
Con.Normal	4.5	3.545 *3.421*	1.900 *2.337*	1.344 *1.467*
Con.Normal	5.0	4.720 *4.786*	2.380 *2.611*	1.672 *1.746*
Con.Normal	5.5	5.813 *5.621*	3.025 *3.384*	2.214 *2.434*
Laplace	6.0	10.012 *7.549*	6.247 *6.359*	5.674 *6.049*

The table indicates that over the distributions studied, as an estimator of kurtosis (A.3) has clear cut advantages over (A.5).

REFERENCES

KENDALL, M. and STUART, A.J. (1966). The Advanced Theory of Statistics, Vol. II, Griffin.

APPENDIX B

This appendix describes the S_U-S_B system of distribution porposed by Johnson (1949) and follows the preamble to the article of Hill, Hill and Holder (1976).

Johnson (1949) has described a system of frequency curves which cover all feasible combination of skewness ($(\beta_1)^{\frac{1}{2}}$) and kurtosis (β_2). The system is broken into three important types viz. S_B (or bounded system), S_U (or unbounded system) and S_L (or lognormal system). The regions of (β_1, β_2) which give rise to these systems are shown in the figure below. Note that attainable distributions comprise those with the restriction $\beta_2 > \beta_1 + 1$.

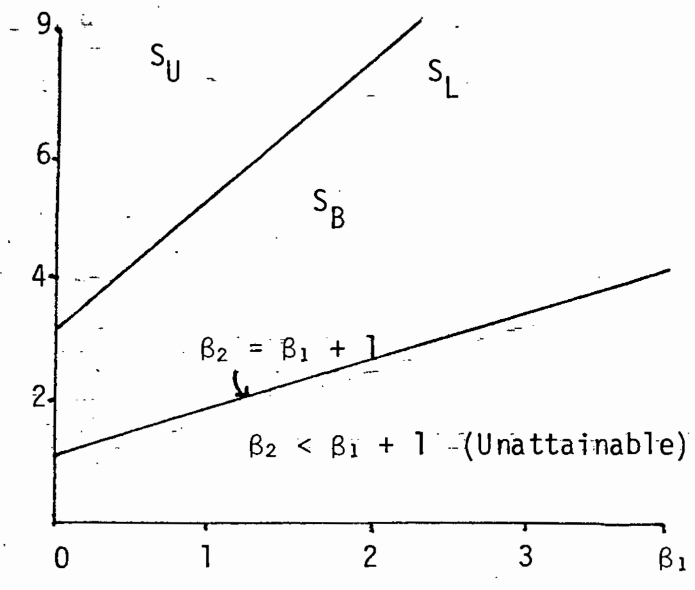


FIGURE B.1

For any (β_1, β_2) combination the parameters of the system may be set so as to generate distributionswith any mean and variance.

The distributions are generated using the following transformations:

1) S_B : $z = \gamma + \delta \ell n((x-\xi)/(\xi+\lambda-x))$, $\xi \le x < \xi + \lambda$,

2) S_U : $z = \gamma + \delta \sinh^{-1}((x-\xi)/\lambda$ (B.1)

3) S_L : $z = \gamma + \delta \ell n(x-\xi)$, $\xi < x$,

where z is a standardized normal variable in each case.

The parameters γ, δ, λ and ξ are then derived by matching the first four moments of x with the desired distribution moments in the way outlined below.

Hill et al also included the case of the so called S_T distribution, the case of the S_B curve on the $\beta_2 = \beta_1 + 1$ boundary. (In the section of this thesis that utilises this system of distributions the condition $\beta_2 \geq \beta_1 + 2$ was, however, adhered to, so the family of S_T distributions was never used.)

Numerical Method (Hill et al)

As is illustrated in Figure B.1 the S_L curves lie on a line in the (β_1, β_2) plane.

Letting $w = \exp(\delta^{-2})$

the β_2 value for the S_L curve is found by solving

$$\beta_1 = (w-1)(w+2)^2$$

for w, and then evaluating

$$\beta_2 = w^4 + 2w^3 + 3w^2 - 3.$$

If the required value of β_2 is less than this calculated value then an S_B curve is appropriate, if greater than this calculated value an S_U curve is appropriate.

1) S_L curves:

Using the w value calculated above the values of the parameters are computed using,

$$\delta = (\ln w)^{-\frac{1}{2}}$$

$$\xi = (\text{sign}(\mu_3))\mu_1' - \exp((1/2\delta - \gamma)/\delta)$$

$$\gamma = \tfrac{1}{2}\,\delta\,\ln(w(w-1)/\mu_2),$$

$$\lambda = \text{sign}(\mu_3).$$

2) S_U curves:

For the case $\beta_1 = 0$

$$w = ((2\beta_2 - 2)^{\frac{1}{2}} - 1)^{\frac{1}{2}}$$

$$\delta = (\ln w)^{-\frac{1}{2}}$$

$$\gamma = 0$$

For $\beta_1 \neq 0$

$$w_1 = ((2\beta_2 - 2.8\beta_1 - 2)^{\frac{1}{2}} - 1)^{\frac{1}{2}}$$

is taken as a first estimate of w and w, δ and γ found by Johnson's iterative method (Elderton and Johnson (1969)) ξ and λ are then computed from

$$\mu_2 = \tfrac{1}{2}\lambda^2(w-1)(w\cosh(-2\,\text{sign}(\mu_3\,\gamma/\delta) + 1)$$

$$\mu_1 = \xi - \lambda\,w^{\frac{1}{2}}\sinh(-\text{sign}(\mu_3)\gamma/\delta)$$

3) S_B curves:

In the words of Hill et al - Approaching the S_T boundary, $\delta \to 0$; approaching the S_L boundary δ tends to the same value as for an S_L curve. A first approximation to δ can then be found by interpolating between these two values. The interpolation is made by assuming the shape of the function to be the same at the required β_1 value as it is between the same two δ values when $\beta_1 = 0$. This is well approximated by

$$\delta = (0.626\beta_2 - 0.408)/(3.0-\beta_2)^{0.479} \qquad \text{if} \quad \beta_2 \geq 1.8$$

and by $\delta = 0.8(\beta_2-1)$ otherwise.

For a given β_1 and first approximation to δ, a first approximation to γ is found using formulae due to Draper (1951). Evaluation of the first six moments at the given δ and γ values using Draper's (1952) form of Goodwin's (1949) integral, then enables a two-dimensional Newton-Raphson process to converge on the required values. Since the first six moments are evaluated at each stage, when the required δ and γ have been found, the first two moments are available to determine λ and ξ.

When the appropriate values of γ, δ, ξ and λ have been calculated the relevant distribution is simulated using (B.1). The routine used was that due to Hill (1976).

REFERENCES

(from Hill et al (1976))

DRAPER, J. (1951). Properties of distributions resulting from certain simple transformations of the normal distribution. M.Sc. thesis, University of London.

DRAPER, J. (1952). Properties of distributions resulting from certain simple transformations of the normal distribution. _Biometrika_, 39, 290-301.

ELDERTON, W.P. and JOHNSON, N.L. (1969). _Systems of Frequency Curves_. Cambridge: University Press.

JOHNSON, N.L. (1949). Systems of frequency curves generated by methods of translation. _Biometrika_, 36, 149-176.

ORD, J.K. (1972). _Families of Frequency Distributions_. London: Griffin.

PEARSON, E.S. and HARTLEY, H.O. (1972). _Biometrika Tables for Statisticians_, Vol. 2. Cambridge: University Press.

APPENDIX C

Program listings for the L_p-routines used in this thesis are given below.

All programs were written in UNIVAC 1100 DOUBLE PRECISION ASCII FORTRAN.

They are respectively (with their daughter subroutines)

a) Subroutine MSPE - calculates L_p-norm ($1 < p < \infty$) estimates of the $\underline{\beta}$ vector in the regression model.

b) Subroutine MMAE - calculates L_∞-estimates of the $\underline{\beta}$ vector in the regression model.

c) Subroutine MABSE - calculates L_1-estimates of the $\underline{\beta}$ vector in the regression model.

a)

```
      SUBROUTINE MSPE(X,Y,M,N,OBJ,SUMSQ,BETA,P,
     +NRS,NC,XI,X1,X2,X3,H,G,S2RES)
C
C      FLETCHER & POWELL(1963)** ROUTINE TO CALCULATE LP-NORM
C      ESTIMATES FOR THE REGRESSION MODEL IN THE PARTCULAR
C      CASE WHEN THERE ARE THREE EXPLANATORY VARIABLES
C      **FLETCHER,R. AND POWELL,M.J.D.(1963)
C      A RAPIDLY-CONVERGENT METHOD FOR MINIMIZATION
C      COMPUTER J. 6,163-168
C
C
C      INPUT:
C      1) Y - VECTOR OF OBSERVATIONS ON DEPENDENT VARIABLE
C      2) X - NXM MARIX OF OBSERVATION ON THE INDEPENDENT VARIABLES
C         NOTE M=3 HERE
C      3) P - VALUE OF P IN LP-NORM
C
C      OUTPUT:
C      1) BETA - VECTOR OF LP-NORM REGRESSION ESTIMATES
C      2) OBJ - SUM OF THE ABSOLUTE VALUES OF THE RESIDUALS
C      3) SUMSQ - SUM OF SQUARES OF THE RESIDUALS
C      4) S2RES - VECTOR OF RESIDUALS
C
C  *******DOUBLE PRECISION*********
C
      IMPLICIT REAL*8(A-H,O-Z)
      DIMENSION X(NRS,NC),Y(NRS),XI(NC),X1(NRS),X2(NRS),X3(NRS),H(NRS),
     + G(NC),S2RES(50),BETA(NC)
      LIMIT = 20
      EST = OBJ**P
      EPS = 0.0001
      DO 100 J=1,M
      XI(J) = BETA(J)
100   CONTINUE
      DO 110 I=1,N
      X1(I) = X(I,1)
      X2(I) = X(I,2)
      X3(I) = X(I,3)
110   CONTINUE
      CALL FMFP(M,XI,F,G,EST,EPS,LIMIT,IER,H,N,P,Y,X1,X2,X3)
      DO 120 J=1,M
      BETA(J) = XI(J)
120   CONTINUE
      OBJ = 0.0
      SUMSQ = 0.0
      DO 140 I=1,N
      ERR = 0.0
      DO 130 J=1,M
      ERR = ERR + BETA(J)*X(I,J)
130   CONTINUE
      ERR = ERR - Y(I)
      S2RES(I)=ERR
      OBJ = OBJ + ABS(ERR)
      SUMSQ = SUMSQ + ERR**2
140   CONTINUE
      RETURN
      END
```

```
      SUBROUTINE FMFP(N,X,F,G,EST,EPS,LIMIT,IER,H,NOBS,P,Y,X1,X2,X3)
      IMPLICIT REAL*8(A-H,O-Z)
      DIMENSION H(1),X(1),G(1),Y(1),X1(1),X2(1),X3(1)
      CALL FUNCT (NOBS,N,P,Y,X1,X2,X3,X,F,G)
      IER=0
      KOUNT=0
      N2=N+N
      N3=N2+N
      N31=N3+1
    1 K=N31
      DO 4 J=1,N
      H(K)=1.00
      NJ=N-J
      IF( NJ )5,5,2
    2 DO 3 L=1,NJ
      KL=K+L
    3  H(KL)=0.00
    4 K=KL+1
    5 KOUNT=KOUNT+1
      OLDF=F
      DO 9 J=1,N
      K=N+J
      H(K)=G(J)
      K=K+N
      H(K)=X(J)
      K=J+N3
      T=0.
      DO 8 L=1,N
      T=T-G(L)*H(K)
      IF(L-J)6,7,7
    6 K = K+N-L
      GO TO 8
    7 K=K+1
    8 CONTINUE
    9 H(J)=T
      DY=0.00
      HNRM = 0.0
      GNRM=0.00
      DO 10 J=1,N
      HNRM=HNRM+ABS(H(J))
      GNRM=GNRM+ABS(G(J))
   10 DY=DY+H(J)*G(J)
      IF(DY)11,51,51
   11 IF(HNRM/GNRM-EPS)51,51,12
   12 FY=F
      ALFA=2.*(EST-F)/DY
      AMBDA=1.
      IF(ALFA)15,15,13
   13 IF(ALFA-AMBDA)14,15,15
   14 AMBDA=ALFA
   15 ALFA=0.
   16 FX=FY
      DX=DY
      DO 17 I=1,N
   17 X(I)=X(I)+AMBDA*H(I)
```

```
      CALL FUNCT (NOBS,N,P,Y,X1,X2,X3,X,F,G)
      FY=F
      DY=0.
      DO 18 I=1,N
   18 DY=DY+G(I)*H(I)
      IF(DY)19,36,22
   19 IF(FY-FX)20,22,22
   20 AMBDA=AMBDA+ALFA
      ALFA=AMBDA
      IF(HNRM*AMBDA-1.E10)16,16,21
   21 IER=2
      RETURN
   22 T=0.
   23 IF(AMBDA)24,36,24
   24 Z=3.*(FX-FY)/AMBDA+DX+DY
      ALFA=AMAX1(ABS(Z),ABS(DX),ABS(DY))
      DALFA=Z/ALFA
      DALFA=DALFA*DALFA-DX/ALFA*DY/ALFA
      IF(DALFA)51,25,25
   25 W=ALFA*SQRT(DALFA)
      ALFA=(DY+W-Z)*AMBDA/(DY+2.00*W-DX)
      DO 26 I=1,N
   26 X(I)=X(I)+(T-ALFA)*H(I)
      CALL FUNCT (NOBS,N,P,Y,X1,X2,X3,X,F,G)
      IF(F-FX)27,27,28
   27 IF(F-FY)36,36,28
   28 DALFA=0.
      DO 29 I=1,N
   29 DALFA=DALFA+G(I)*H(I)
      IF(DALFA)30,33,33
   30 IF(F-FX)32,31,33
   31 IF(DX-DALFA)32,36,32
   32 FX=F
      DX=DALFA
      T=ALFA
      AMBDA=ALFA
      GO TO23
   33 IF(FY-F)35,34,35
   34 IF(DY-DALFA)35,36,35
   35 FY=F
      DY=DALFA
      AMBDA=AMBDA-ALFA
      GO TO 22
   36 DO 37 J=1,N
      K=N+J
      H(K)=G(J)-H(K)
      K=N+K
   37 H(K)=X(J)-H(K)
      IF(OLDF-F+EPS)51,38,38
   38 IER=0
      IF(KOUNT-N)42,39,39
```

```
39 T=0.
   Z=0.
   DO 40 J=1,N
   K=N+J
   W=H(K)
   K=K+N
   T=T+ABS(H(K))
40 Z=Z+W*H(K)
   IF(HNRM-EPS)41,41,42
41 IF(T-EPS)56,56,42
42 IF(KOUNT-LIMIT)43,50,50
43 ALFA=0.
   DO 47 J=1,N
   K=J+N3
   W=0.
   DO 46 L=1,N
   KL=N+L
   W=W+H(KL)*H(K)
   IF(L-J)44,45,45
44 K=K+N-L
   GO TO 46
45 K=K+1
46 CONTINUE
   K=N+J
   ALFA=ALFA+W*H(K)
47 H(J)=W
   IF(Z*ALFA)48,1,48
48 K=N31
   DO 49 L=1,N
   KL=N2+L
   DO 49 J=L,N
   NJ=N2+J
   H(K)=H(K)+H(KL)*H(NJ)/Z-H(L)*H(J)/ALFA
49 K=K+1
   GO TO 5
50 IER=1
   RETURN
51 DO 52 J=1,N
   K=N2+J
52 X(J)=H(K)
   CALL FUNCT (NOBS,N,P,Y,X1,X2,X3,X,F,G)
   IF(GNRM-EPS)55,55,53
53 IF(IER)56,54,54
54 IER=-1
   GO TO 1
55 IER=0
56 RETURN
   END
```

```
      SUBROUTINE FUNCT(NOBS,N,P,Y,X1,X2,X3,X,F,G)
      IMPLICIT REAL*8(A-H,O-Z)
      DIMENSION X(1),G(1),Y(1),X1(1),X2(1),X3(1)
      DO 101 I=1,N
      G(I)=0.
  101 CONTINUE
      F=0.0
      DO 100 I=1,NOBS
      YX=Y(I)-X(1)*X1(I)-X(2)*X2(I)-X(3)*X3(I)
      F=F+DABS(YX)**P
      IF ( DABS(YX) .GT. 0.0 )  GO TO 400
      YXX= 0.0
      GO TO 410
400   YXX=(DABS(YX))**(P-2.0)
410   G(1)=G(1)+YX*YXX*P*(-1.0)
      G(2)=G(2)+X2(I)*YX*YXX*P*(-1.0)
      G(3)=G(3)+X3(I)*YX*YXX*P*(-1.0)
100   CONTINUE
      RETURN
      END
```

b)

```
      SUBROUTINE MMAE(X,Y,M,N,OBJ,SUMSQ,BETA,IK)
C
C   THIS ROUTINE EVALUATES THE CHEBYCHEV REGRESSION ESTIMATES
C   FOR THE GENERAL REGRESSION MODEL Y = XB + E
C   USING THE METHOD OF WAGNER(1959) :
C   WAGNER,H.M.(1959)LINEAR PROGRAMMING TECHNIQUES FOR
C   REGRESSION ANALYSIS. J.AMER.STAT.ASSOC. 54 , 206-212
C
C
C   FOR THE DETAILS OF THE SIMPLX SUBROUTINE THE READER
C   IS REFERRED TO THE MANUAL " SIMPDX/SIMPLX LINEAR PROGRAMMING SUBROUTINES
C   REFERENCE MANUAL FOR THE UNIVAC 1108, UNIVERSITY OF
C   WISCONSIN COMPUTER CENTRE, MAY 1970" FOR ANY FURTHER DETAILS.
C
C   INPUT:
C   1)Y - VECTOR OF N OBSERVATIONS ON THE DEPENDENT VARIABLE
C   2) X - NXM MATRIX OF OBSERVATIONS ON THE INDEPENDENT VARIABLES
C
C   OUTPUT:
C   1) BETA - L^ ESTIMATES OF BETA VECTOR
C   2) IK - IF.IK.EQ.1 THEN OPTIMAL SOLUTION HAS NOT BEEN FOUND
C      OTHERWISE SOLUTION IS OPTIMAL
C   3) OBJ - SUM OF ABSOLUTE VALUES OF RESIDUALS
C   4) SUMSQ - SUM OF SQUARES OF RESIDUALS
C
      DIMENSION A(100,11),RHS(100),COST(11),T(100),
     +IFIX(20),TOL(4),XX(102),JX(100),PI(100),E(102,102),ERR(4),IOUT(4),
     +YY(102),S(111),SOL(111)
      REAL*8 X(50,5),Y(50),BETA(5),OBJ,SUMSQ
      DATA T1,T2,T3/1H ,1H+,1H-/
      DO 70 I=1,11
      COST(I) = 0.0
      DO 80 J=1,100
      A(J,I) = 0.0
80    CONTINUE
70    CONTINUE
      DO 90 I=1,17
      IFIX(I) = 0
90    CONTINUE
      COST(1) = 1.0
      DO 100 I=1,N
      RHS(I) = Y(I)
      RHS(I+N) = Y(I) * (-1.0)
100   CONTINUE
      MM = N * 2
      NN = M*2 + 1
      DO 120 I=1,N
      A(I,1) = 1.0
      A(I+N,1) = 1.0
      DO 110 J=1,M
      JJ = 2*J + 1
      A(I,JJ-1) = X(I,J)
      A(I,JJ) = X(I,J) * (-1.0)
      A(I+N,JJ-1) = (-1.0) * X(I,J)
      A(I+N,JJ) = X(I,J)
110   CONTINUE
120   CONTINUE
```

```
      DO 130 I=1,MM
      T(I) = T3
130   CONTINUE
      IFIX(1) = 100
      IFIX(2) = 11
      IFIX(3) = MM
      IFIX(4) = NN
      IFIX(7) = MM
      IFIX(15) = 102
      IFIX(16) = NN
      IFIX(9) = 1
      CALL SIMPLX(A,T,RHS,COST,IFIX,TOL,OOBJ,XX,JX,PI,E,ERR,IOUT,YY,S)
      OBJ=OOBJ
      IF ( IOUT(1) .EQ. 1 )  GO TO 150
      WRITE(5,140) (IOUT(I),I=1,N)
140   FORMAT(1H1,' ERROR IN IN SIMPLX ROUTINE MMAE ',4I8)
      IK = IOUT(1)
      RETURN
150   DO 160 J=1,NN
      SOL(J) = 0.0
160   CONTINUE
      DO 170 I=1,MM
      JJ = JX(I)
      SOL(JJ) = XX(I)
170   CONTINUE
      DO 180 J=2,NN,2
      JJ = (J-2)/2 + 1
      BETA(JJ) = SOL(J) - SOL(J+1)
180   CONTINUE
      SUMSQ = 0.0
      OBJ = 0.0
      DO 200 I=1,N
      SUM = 0.0
      DO 190 J=1,M
      SUM = SUM + BETA(J)*X(I,J)
190   CONTINUE
      EI = Y(I) - SUM
      OBJ = OBJ + ABS(EI)
      SUMSQ = SUMSQ + EI**2
200   CONTINUE
      IK = 0
      RETURN
      DEBUG SUBCHK
      END
```

c)

```
      SUBROUTINE MABSE(X,Y,M,N,OBJ,SUMSQ,BETA,IK,NROWS,
     + NCOLS,NR,NC,Q,ALPHA,R)
C
C     THIS ROUTINE CALCULATES AN L1 REGRESSION ESTIMATE
C     FOR THE STANDARD REGRESSION MODEL Y=XB + E
C     USING THE METHOD OF :
C     BARRODALE,I. AND YOUNG,A.(1966)
C     AN ALGORITHM FOR BEST L1 AND L^ LINEAR APPROXIMATIONS
C     ON A DISCRETE SET
C     NUMERICAL MATH 8,  295-306
C
C
C     INPUT:
C     1) Y - VECTOR OF N OBSERVATIONS ON THE DEPENDENT VARIABLE
C     2) X - NXM MATRIX OF OBSERVATIONS ON THE INDEPENDENT VARIABLES
C
C     OUTPUT:
C     1) BETA - L1 ESTIMATES OF BETA VECTOR
C     2) IK - IF.IK.EQ.1 THEN OPTIMAL SOLUTION HAS NOT BEEN FOUND
C        OTHERWISE SOLUTION IS OPTIMAL
C     3) OBJ - SUM OF ABSOLUTE VALUES OF RESIDUALS
C     4) SUMSQ - SUM OF SQUARES OF RESIDUALS
C     5) R - VECTOR OF RESIDUALS
C
C  *******DOUBLE PRECISION**********
C
      IMPLICIT REAL*8(A-H,O-Z)
      REAL*8 Q(NROWS,NCOLS),X(50,5),Y(50),BETA(NC),ALPHA(60),R(50)
C
      DO 140 I=1,N
      Q(I+1,1) = Y(I)
      DO 130 J=1,M
      Q(I+1,J+1) = X(I,J)
130   CONTINUE
140   CONTINUE
```

```
      CALL MSMOD(M,N,Q,NROWS,NCOLS)
      IDUM = Q(N+2,M+3)+0.2
      IF ( IDUM .EQ. 1 )  GO TO 160
      WRITE(5,150)
150   FORMAT(1H1,' OPTIMAL SOLUTION NOT FOUND')
      IK = 1
      RETURN
160   DO 170 I=1,N
      ALPHA(I) = 0.0
170   CONTINUE
      DO 180 I=1,N
      INDIC = Q(I+1,M+3) + 0.5
      IF ( INDIC .LE. N )  GO TO 180
      INDIC = INDIC - N
      ALPHA(INDIC) = Q(I+1,1)
180   CONTINUE
      DO 190 I=1,M
      BETA(I) = ALPHA(I) - ALPHA(M+1)
190   CONTINUE
      OBJ = 0.0
      SUMSQ = 0.0
      DO 210 I=1,N
      ERR = Y(I)
      DO 200 J=1,M
      ERR = ERR - BETA(J)*X(I,J)
200   CONTINUE
      R(I)=ERR
      OBJ = OBJ + ABS(ERR)
      SUMSQ = SUMSQ + ERR*ERR
210   CONTINUE
      IK = 0
      RETURN
      DEBUG SUBCHK
      END
```

```
      SUBROUTINE MSMOD(M,N,Q,NROWS,NCOLS)
      IMPLICIT REAL*8(A-H,O-Z)
      REAL*8 Q(NROWS,NCOLS)
      INTEGER T,OUT
      MMM = M + 1
      DO 100 J=1,MMM
      Q(1,J+1) = 0.0
      Q(N+2,J+1) = N+J
100   CONTINUE
      Q(1,1) = 0.0
      Q(N+2,1) = 0.0
      DO 120 I=1,N
      Q(I+1,M+3) = I
      A = 0.0
      JDUM = 0
      IF ( Q(I+1,1) .LT. 0.0 )  JDUM = 1
      DO 110 J=0,M
      IF ( JDUM .EQ. 1 )  Q(I+1,J+1) = -1.0*Q(I+1,J+1)
      A = A - Q(I+1,J+1)
      Q(1,J+1) = Q(1,J+1) + Q(I+1,J+1)
110   CONTINUE
      IF ( JDUM .EQ. 1 )  Q(I+1,M+3) = -1.0*Q(I+1,M+3)
      Q(I+1,M+2) = A + Q(I+1,1)
      Q(1,M+2) = Q(1,M+2) + Q(I+1,M+2)
120   CONTINUE
      IT = -1.0
130   A = 10.0**(-4)
      ZZZZ = -1.0*2.0 - A
      T = 0
      IT = IT + 1
      Q(1,M+3) = IT
      MMM = M + 1
      DO 150 J=1,MMM
      Z = Q(1,J+1)
      IDUM = Q(N+2,J+1) + 0.2
      IF (IDUM .GT. N) GO TO 140
      IF ( (ZZZZ-Z) .LE. A )  GO TO 140
      IN = J
      T = 2
      A = ZZZZ - Z
140   IF ( Z .LE. A )  GO TO 150
      IN = J
      T = 1
      A = Z
```

```
150     CONTINUE
        IF ( T .NE. 0 )  GO TO 155
        Q(N+2,M+3) = 1
        GO TO 200
155     B = 10.0**9
        DO 160 I=1,N
        D = Q(I+1,IN+1)
        IF ( T .EQ. 2 )  D = -1.0*D
        IF ( D .LT. 10.0**(-5))  GO TO 160
        D = Q(I+1,1) / D
        IF ( D .GE. B )  GO TO 160
        B = D
        OUT = I
160     CONTINUE
        IF ( B .LT. 10.0**5 )  GO TO 165
        DO 411 I=1,27
        WRITE(5,410) (Q(I,J),J=1,6)
410     FORMAT(6F12.7)
411     CONTINUE
        Q(N+2,M+3) = 2.0
        GO TO 200
165     IF ( T .NE. 2 )  GO TO 167
        Q(N+2,IN+1) = -1.0*Q(N+2,IN+1)
        Q(1,IN+1) = -1.0*A
167     P = Q(OUT+1,IN+1)
        DO 180 I=0,N
        IF ( I .EQ. OUT )  GO TO 180
        D = Q(I+1,IN+1)/P
        MMM = M+1
        DO 170 J=0,MMM
        Q(I+1,J+1) = Q(I+1,J+1) - D*Q(OUT+1,J+1)
170     CONTINUE
        Q(I+1,IN+1) = -1.0*D
180     CONTINUE
        P = ABS(P)
        MMM = M+1
        DO 190 J=0,MMM
        Q(OUT+1,J+1) = Q(OUT+1,J+1) / P
190     CONTINUE
        Q(OUT+1,IN+1) = 1.0 / P
        I = Q(OUT+1,M+3) + 0.2
        Q(OUT+1,M+3) = Q(N+2,IN+1)
        Q(N+2,IN+1) = I
        GO TO 130
200     RETURN
        DEBUG SUBCHK
        END
```