RECENT STUDIES ON $L_p$-NORM ESTIMATION

by

Hans Nyquist

AKADEMISK AVHANDLING

som med tillstånd av rektorsämbetet vid
Umeå universitet för erhållande av filosofie
doktorsexamen framlägges till offentlig
gränskning i Hörsal D, Samhällsvetarhuset,
fredagen den 12 december 1980 kl 10.15.
**TABLE OF CONTENTS**

**ABSTRACT**

**ACKNOWLEDGEMENTS**

**CHAPTER I. OVERVIEW OF THE PROBLEM**

**CHAPTER II. A BACKGROUND TO $L_p$-NORM ESTIMATION**

- 2.1 An introduction to the art of combining observations
- 2.2 Linear models
- 2.3 Stable distributions and their domains of attraction
- 2.4 Examples of models with fat-tailed or contaminated distributed random variables
- 2.5 Algorithms for computations of $L_1$- and $L_\infty$-norm estimates
- 2.6 Some Monte Carlo studies
- 2.7 Conclusions

**CHAPTER III. SOME BASIC PROPERTIES OF $L_p$-NORM ESTIMATORS**

- 3.1 Definition of the $L_p$-norm estimators
- 3.2 The existence of $L_p$-norm estimates
- 3.3 Characterizations of $L_p$-norm estimates
- 3.4 Uniqueness of $L_p$-norm estimates, $1 < p < \infty$
- 3.5 Uniqueness of $L_1$-norm estimates
- 3.6 Geometric interpretations
- 3.7 A general notion of orthogonality
- 3.8 Relations to M- and L-estimators
- 3.9 Equivalence of $L_p$-norm and maximum likelihood estimators $L_p$
- 3.10 Computation of $L_p$-norm estimates, $1 < p < \infty$, $p \neq 2$
- 3.11 Conclusions

**CHAPTER IV. SOME SAMPLING PROPERTIES OF $L_p$-NORM ESTIMATORS**

- 4.1 Introduction
- 4.2 The sampling distribution of $L_p$-norm estimators
- 4.3 Stochastic regressors
- 4.4 Heteroscedasticity
- 4.5 Linearly dependent residuals
- 4.6 The optimal $L_p$-norm estimator
- 4.7 Estimation of the variance
- 4.8 Geometrical interpretations
- 4.9 Conclusions
RECENT STUDIES ON $L_p$-NORM ESTIMATION

by

Hans Nyquist
Institute of Statistics
University of Umeå
S-901 87 UMEÅ
Sweden

ABSTRACT

When estimating the parameters in a linear regression model, the method of least squares ($L_2$-norm estimator) is often used. When the residuals are independent and identically normally distributed, the least squares estimator is BLUE as well as equivalent to the maximum likelihood estimator. However, the least squares estimator is known to be sensitive to departures from the assumption of normally distributed residuals. In a variety of applications there are theoretical as well as empirical evidences that the residuals display distributional properties different from those of normal distributions. It is therefore desirable to develop alternative estimators. Members of the class of $L_p$-norm estimators have here been proposed as alternatives to the least squares estimator.

In this monograph, questions concerning the existence, uniqueness and asymptotic distributions of $L_p$-norm estimators are discussed. It is seen that an $L_p$-norm estimate will always exist and it will be unique for $1 < p < \infty$. For $p = 1$ a necessary and sufficient condition on the regressors for unique estimation is given. The question of uniqueness in large samples is also discussed. The asymptotic distribution of $L_p$-norm estimators is shown to be normal, for sufficiently small $p$. When selecting an $L_p$-norm estimator, a procedure based on the asymptotic variance is proposed.

Finally, the possibilities to construct $L_p$-norm based estimators for estimating multicollinear regression models and models with serially dependent residuals are discussed. $L_p$-norm based methods for estimating interdependent systems are also considered.

Key words and phrases: $L_p$-norm estimation, robustness, linear regression models, fat-tailed distributed residuals.
ACKNOWLEDGEMENTS

Most of this study was carried out at the Institute of Statistics, University of Umeå.

I wish to express my sincere gratitude to Professor Uno Zackrisson, Head of the Institute, and Docent Anders Westlund for their guidance in preparation of this thesis. Their constructive criticism, ideas and advice added new, important dimensions to my research.

I was fortunate to find colleagues willing to discuss parts of this study with me. In this respect, I am especially grateful to Professor Lester D. Taylor, Department of Economics, University of Arizona, Docent Lennart Bondesson, Institute of Mathematical Statistics, University of Umeå and FK Göran Arnoldsson, Institute of Statistics, University of Umeå.

A special thanks to Berith Melander, an excellent typist, who unselfishly gave her time in order to help me meet deadlines; an activity in which Maj-Britt Backlund and Åsa Ågren also participated.

It is gratefully acknowledged that this work was supported by grants from the Swedish Research Council for the Humanities and the Social Sciences.

Umeå

November 1980          Hans Nyquist
CHAPTER I. OVERVIEW OF THE PROBLEM

The use of stochastic models has a long history. For example, it is known that scientists during the 18th century made use of stochastic models for describing the motions of celestial bodies. Stochastic models usually comprise a set of unknown parameters to be estimated. Therefore the problem of estimating parameters in stochastic models also has a long history. The first estimating techniques were based on different types of averaging empirical observations or functions of empirical observations.

A completely new methodology in estimating unknown parameters in stochastic models was introduced by Boscovich in 1757. He proposed that the parameters should be estimated according to the minimum of a function of the measurement errors. The function Boscovich proposed was the sum of the absolute measurement errors. This method is still employed under the name of MAD (minimum absolute deviations), LAE (least absolute error), least first power or $L_1$-norm estimator. However, that estimating method was computationally very complicated. Only relatively simple models could be estimated, and it took a long time to carry through all computations.

In the works of Legendre from 1805 and Gauss from 1809, another function to be minimized was proposed, viz. the sum of the squares of the measurement errors. By that, the computationally simple method of least squares (or $L_2$-norm estimator) was born. After the proposals of Legendre and Gauss, the method of least squares has been one of the most popular estimating techniques. The popularity of the least squares estimator is presumably due to the easy computation and the fact that when the residuals are independent and identically normally distributed, the least squares estimator of a linear regression model is BLUE (best linear unbiased estimator) as well as equivalent to the maximum likelihood estimator, implying the inference to be easily performed.
However, the least squares estimates are known to be sensitive to departures from the assumption of normally distributed residuals. In a variety of applications there are theoretical as well as empirical evidences that the residuals display distributional characteristics different from those of normal distributions. It is therefore desirable to develop alternative estimators.

Since there is an infinite number of ways to construct estimators, we have to restrict our attention to some class of estimators. Boscovich introduced the idea of, in some sense, minimizing the residual vector. The same idea is fundamental in Legendre's and Gauss' method of least squares. However there is no unique criterion for minimizing a vector. Therefore we have to minimize some functional of the residual vector. To add more mathematical structure to the estimator, we restrict the functional to be a vector norm. Thus, given an appropriate vector norm $||\cdot||$, the estimates are chosen as to minimize $||\hat{e}||$, where $\hat{e}$ is the observed residual vector.

While the class of all vector norms is limitless, an interesting subclass is given by the so-called $L_p$-norms. The $L_p$-norm of the vector of observed residuals is defined as

$$||\hat{e}|| = \begin{cases} \left\{ \sum_{i=1}^{n} |\hat{e}_i|^p \right\}^{1/p} & \text{for } 1 \leq p < \infty \\ \max|\hat{e}_i| & \text{for } p=\infty \end{cases}$$

An estimator minimizing a $L_p$-norm of the residual vector, will be called an $L_p$-norm estimator. Thus, Boscovich's proposal corresponds to $p=1$ and the least squares estimator to $p=2$ (therefore we frequently use the names $L_1$- and $L_2$-norm estimators, respectively).

With the computer routines that are available today, there should not be any serious problems in computing $L_p$-norm estimates.

The interest in the $L_1$-norm estimator has its origin in the results
of Laplace (1818) and Edgeworth (1887a and b). There it is shown that the \( L_1 \)-norm estimator is preferable to the least squares, when estimating a simple linear regression model with fat-tailed or contaminated distributed residuals. Results from several Monte Carlo studies, performed more recently, indicate that the same conclusions are to be drawn, even for more complicated models.

The interest in the class of \( L_p \)-norm estimators is that the least squares as well as the \( L_1 \)-norm estimators are members of that class. Furthermore, when estimating a linear regression model and the residual distribution belongs to the class of double exponential distributions, (which is a family of the exponential class of distributions), there is an \( L_p \)-norm estimator that is equivalent to the maximum likelihood estimator. For other distributions, there are Monte Carlo simulations that suggest that there is an \( L_p \)-norm estimator preferable both to the \( L_1 \)- and to the \( L_2 \)-norm estimator.

The purpose of the present monograph is the investigation of \( L_p \)-norm estimators of linear regression models. In particular, we are interested in giving theorems on the existence, uniqueness and asymptotic distributions of \( L_p \)-norm estimators. We will also give geometrical interpretations of \( L_p \)-norm estimation. Mostly, we will restrict our attention to \( 1 < p < \infty \) and only a few results on the \( L_\infty \)-norm estimator will be given.

Chapter 2 contains a background to \( L_p \)-norm estimation. Specially, a historical background is given and the linear regression model is defined. Since alternatives to the least squares estimator are of interest when residual distributions are non-normal, some examples of models with non-normal residual distributions are given. For that and for the various theorems on the asymptotic distribution of \( L_p \)-norm estimators, we need some concepts from the theory of stable distributions. Those concepts are defined in Chapter 2. Furthermore, the Monte Carlo studies referred
Some basic properties of \( L_p \)-norm estimators are given in Chapter 3. Thus, the existence and uniqueness problems are discussed. Some characterizations of \( L_p \)-norm estimates are also given. The results presented so far in Chapter 3, are given a geometrical interpretation.

In Huber (1968 and 1977) the class of maximum likelihood type estimators (M-estimators) and the class of estimators that are linear combinations of order statistics (L-estimators) were defined. In Chapter 3 relations between, on the first hand, \( L_p \)-norm estimators and, on the other hand, M-, L- and maximum likelihood estimators are given.

Chapter 4 deals with distributional properties of \( L_p \)-norm estimators. Specially we investigate the "classical" case of fixed regressors and independent and equally distributed residuals. Departures from these assumptions are discussed and cases with stochastic regressors, heteroscedastic and linearly dependent residuals are studied. Some geometrical interpretations of \( L_p \)-norm estimation in cases of heteroscedastic and linearly dependent residuals are also given.

The optimal \( L_p \)-norm estimator is defined as that \( L_p \)-norm estimator that possesses the smallest asymptotic variance. The concept of the optimal \( L_p \)-norm estimator and the problem of choosing \( L_p \)-norm estimator is also discussed in Chapter 4.

Based on more heuristic discussions than exact deduction, an outline on how more complex \( L_p \)-norm based estimation methods may be defined is given in Chapter 5. Thus, \( L_p \)-norm based methods for estimating multicolinear regression models and regression models with serially dependent residuals are proposed. We will also discuss \( L_p \)-norm based estimators of interdependent systems.
CHAPTER II. A BACKGROUND TO $L_p$-NORM ESTIMATION

2.1 An introduction to the art of combining observations

The problem of estimating unknown parameters in models has a long history. During the 18th century, for example, there was a magnificent pleiad of mathematicians, from Euler and the Bernoullis to d'Alembert, Lagrange and Laplace, working with models that comprised unknown parameters. These models were often deduced from the Newtonian mechanics. Descriptions and predictions of motions of celestial bodies in a solar system are examples of problems scientists were interested in. Measuring the size of the earth was another problem scientists showed a great interest. Models for that problem may be deduced from geometrical considerations. However, the models will still comprise unknown quantities.

Unknown parameters in models are often estimated from empirical observations. It was early evident, however, that empirical observations often were affected by a small random error. One of the first formulations in probabilistic terms of random errors in empirical observations is in de Moivre's 'Doctrine of chances' from 1738. After that the evolution of the theory of errors is due to especially Laplace and Gauss.

It was early well known that averages tended to "smooth out" small random errors. Therefore, the first estimating methods were constructed due to the Principle of Arithmetic Mean. If $y$ is known to be proportional to $x$, i.e. $y = \beta x + \varepsilon$ the arithmetic mean of the "observed ratios"

$$b_i = \frac{y_i}{x_i},$$

that is

$$\bar{b} = \frac{1}{n} \sum_{i=1}^{n} b_i = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i}$$

is the estimate of $\beta$ due to the Principle of Arithmetic Mean. Similarly, in the case of the two-parameter line $y = \alpha + \beta x + \varepsilon$, $\beta$ was estimated by the arithmetic mean of either all of the "observed slopes"

$$b_{ij} = \frac{(y_j - y_i)}{(x_j - x_i)} \quad i=1,2,...,n-1; \quad j=i+1,i+2,...,n$$
Figure 2.1 Estimation of a straight line due to the Principle of Arithmetic Mean. All possible lines through two of the four observational points are drawn. The "average" intercept and "average" slope constitute the estimated line.

or only of those "observed slopes" that corresponded to large differences in abscissa. Thus, the slope is estimated by the average slope of all straight lines through two observational points, provided the points are not too close. The intercept \( a \) is estimated in the same manner. See Figure 2.1.

When estimating the simple model \( y = \beta x + \epsilon \) and the observational points are believed to be affected by measurement errors, the slopes \( b_i \) are also affected by measurement errors. Further, if the errors in the observational points are in the \( y \)-direction only and the magnitude of the errors are independent of \( x \), the errors in an observed slope \( b_i \) are inversely proportional to the magnitude of \( x_i \). Thus, when estimating \( \beta \), Cotes (1722) suggested that in place of the simple arithmetic mean of \( b_i \), one should take the weighted arithmetic mean with weights proportional to the magnitude of \( x \). For positive \( x_i \), \( \beta \) is then estimated by
\[
\begin{align*}
\sum_{i=1}^{n} \frac{x_i}{y_i} &= \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \bar{y} \cdot \bar{x},
\end{align*}
\]

the rate of the means, which is the solution of the equation
\[
\sum_{i=1}^{n} (y_i - b x_i) = 0.
\]

In other words, if a set of observational points are to be fitted by a line of the form \( y = \beta x \), then Cotes' modification of the traditional procedure implies that the estimate of \( \beta \) is determined by the condition of zero sum of residuals. This may be regarded as an extension of the Principle of Arithmetic Mean, since the arithmetic mean \( \bar{y} \) of a set of numbers \( y_1, y_2, \ldots, y_n \) is the solution of the equation
\[
\sum_{i=1}^{n} (y_i - \bar{y}) = 0.
\]

The requirement of zero sum of residuals alone is not sufficient to determine an estimate of the parameters in a model with more than one parameter. Consequently, in 1748 Euler and Mayer independently devised the so called Method of Averages. Due to this method, the observational points are to be subdivided into as many subsets as there are parameters to estimate. The subdivision being in terms of the values of (one of) the independent variable(-s), and then applying the conditions of zero sum of residuals to the points of each subset. Thus, to estimate the linear model \( y = a + \beta x + \epsilon \), one first divide the observational points into two subsets according to the values of \( x \) and a specified number \( x_0 \). The estimates of \( a \) and \( b \) respectively, are then the solution of the equation system
\[
\begin{align*}
\sum_{x \leq x_0} (y_i - (a + bx_i)) &= 0, \\
\sum_{x \geq x_0} (y_i - (a + bx_i)) &= 0,
\end{align*}
\]
The observations are divided into two subsets. The estimated line is determined by the averages of the observations in each subset.

which is equivalent to

\[
\begin{align*}
\overline{y}' - a - bx' &= 0 \\
\overline{y}'' - a - bx'' &= 0
\end{align*}
\]

where \(\overline{x}', \overline{y}', \overline{x}''\) and \(\overline{y}''\) are the averages of \(x\) and \(y\) in the first and the second subset, respectively. See Figure 2.2.

When the Method of Averages is applied for estimating more than one parameter, there is some arbitrariness and room for subjective choice in the formation of subgroups. The estimates of the parameters are then affected with corresponding arbitrariness. A pure "objective" estimation method was proposed by Boscovich in 1757. Boscovich's principle was that \(a\) and \(\beta\) in the two parameter model \(y = a + \beta x + \varepsilon\) should be estimated by \(a\) and \(\beta\) respectively, according to the conditions

\[
\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \quad (2.1)
\]

\[
\sum_{i=1}^{n} |y_i - a - bx_i| = \text{minimum} \quad (2.2)
\]
Figure 2.3 Estimation of a straight line according to the conditions (2.1) and (2.2).

Condition (2.1) states that the estimated line shall pass through the centroid \((x, y)\) of the observational points, where 
\[ x = n^{-1} \sum_{i=1}^{n} x_i \text{ and } \]
\[ y = n^{-1} \sum_{i=1}^{n} y_i. \]
Replacing \(a\) in (2.2) by its value implied by (2.1), it is seen that the two conditions require that \(b\) shall satisfy
\[
\sum_{i=1}^{n} |(y_i - \bar{y}) - b(x_i - \bar{x})| = \text{minimum.} \tag{2.4}
\]

The parameter estimates due to Boscovich's method is then obtained from (2.3) and (2.4). See Figure 2.3.

Laplace recognized the great value of Boscovich's principle and employed it in his memoir from 1786 and in Mécanique Céleste, published in 1799. Laplace gives also an extension of the technique to the case of observational points of unequal weight, expressing conditions (2.1) and (2.2) as weighted sums.

In the papers from 1887, Edgeworth proposed that the condition (2.1) could be dropped and the model should be estimated only according to condition
For more general models, this method was generalized in an immediate way, thus, the parameter estimates were to minimize the sum of absolute observed measurement errors \( \sum_{i=1}^{n} |\hat{\epsilon}_i| \) was to be minimized. This method is still employed under the name MAD (minimum absolute deviation), or LAE (least absolute error). Since the sum of the first power of \( |\hat{\epsilon}_i| \) is to be minimized, the estimating technique is also called the method of least first power. Further, it appears that the sum of first powers of \( |\hat{\epsilon}_i| \) agrees with the \( L_1 \)-norm in mathematical analysis. Therefore, the estimator is sometimes called the \( L_1 \)-norm estimator. Here we will use the last notation. In Figure 2.4 an example of \( L_1 \)-norm estimation is shown.

The \( L_1 \)-norm estimator was, however, computationally very complicated. Only relatively simple models could be estimated, and it took a long time to carry through all the tedious computations.

At the very end of the 18th century, Gauss tried to determine the orbits of the at that time newly discovered planets Ceres and Pallas. That problem is reported in Gauss' work from 1809. Gauss proposed that the unknown parameters should be estimated according to the minimum of the squared observed measurement errors, i.e. \( \sum_{i=1}^{n} \hat{\epsilon}_i^2 \) was to be minimized.
Figure 2.5 Least squares (L$_2$-norm) estimation of a straight line. The estimated line minimizes the sum of squared observed measurement errors, \[ \sum e_i^2. \]

This method is known under the name of the method of least squares, or the L$_2$-norm estimator.

Independently of Gauss, the method of least squares was proposed by Legendre. Gauss asserted that he had applied the method since 1795. It was first published, however, by Legendre in 1805.

When the method of least squares was applied, the computations were considerably simplified. The choice of method was rather ad hoc, and Gauss writes in his work from 1820:

"If you object that this is arbitrary, we readily agree. The question with which we are concerned is vague in its very nature, it can only be made precise by pretty arbitrary principles. Determining a magnitude by observation can justly be compared to a game in which there is a danger of loss but no hope of gain ... . But if we do so, to what can we compare an error which has actually been made? That question is not clear, and its answer depends in part of convention. Evidently the loss in the game can't be compared directly to the error which has been committed,
for then a positive error would represent a loss, and a negative error a gain. The magnitude of the loss must on the contrary be evaluated by a function of the error whose value is always positive. Among the infinite number of functions satisfying these conditions, it seems natural to choose the simplest, which is, beyond contradiction, the square of the error."

In the well-known Théorie Analytique des Probabilités from 1818, Laplace considered the problem of estimating the parameter $\beta$ in the simple regression model $y_i = \beta x_i + \epsilon_i$. He then investigates the distribution of the $L_1$-norm estimate $b$ of $\beta$. Due to the assumption that the observational errors $\epsilon_i$ all have the same probability density $f$, and that $f$ is symmetric about zero, Laplace derived the density of $\hat{\epsilon}_i = y_i - bx_i$. He also showed that when $n$ increases, the density of $b$ approaches the normal density with mean zero and variance $\left(4f^2(0) \sum_{i=1}^{n} x_i^{-1}\right)^{-1}$. In the special case where all $x_i$'s are 1, this agrees with standard results for the sample median.

Laplace also derived the asymptotic distribution of the least squares estimator. It was then possible to compare the estimators on the basis of the variances of their asymptotic distributions. The conclusions to draw were that the $L_1$-norm estimator is preferable when

$$\left\{2f(0)\right\}^{-2} < \int_{-\infty}^{\infty} \epsilon^2 f(\epsilon) \, d\epsilon$$

Thus, in the case of normally distributed residuals, the least squares estimator is preferable. Further, when the variance in the residual distribution does not exist, one may choose the $L_1$-norm estimator.

With this work, Laplace was probably the first to compare two estimators in probabilistic terms. The probability theory was thus so developed in the beginning of the 19th century, that estimation was not only simple curve fitting. It was possible to choose an estimator due to a special assumption of the distribution of the measurement errors.
A valuable contribution to the theory of measurement errors and estimation was given by Edgeworth in two papers published in 1887. Edgeworth there discussed the possibility that the measurement errors consist of different errors with different and partly unknown probability distributions; contaminated residual distributions. Edgeworth noted that if this was the case, then the value of the density function of the measurement error is not inversely proportional to the square of the magnitude of the measurement error. This implies that the method of least squares is not suitable for that situation. As an alternative to the method of least squares, Edgeworth proposed the method of least first power.

So far, we have discussed the problem of estimating the parameters in a given model (or function) so that the estimated model, as good as possible, will follow a set of given observations. A closely related problem is that of estimating parameters in a relatively simple function, so that the estimated function, as good as possible, follows a given, more complicated function. In the middle of the 19th century, Chebychev worked with studies of mechanical systems in connection with steam engines. He was then concerned with the theory of approximation of functions. In a paper from 1857 he proposed that the parameters should be estimated so that the greatest absolute difference between the given function and the estimated function is as small as possible. Of course, it is possible to apply this method when estimating the parameters in a model. In our notations we then have to minimize \( \max_i |e_i| \). This estimation method is known under the name Chebychev estimation, or \( L_\infty \)-norm estimation. See Figure 2.6.

The discussion on the use of \( L_1^- \), \( L_2^- \) and \( L_\infty^- \)-norms respectively, in statistical estimation has been continued during the 20th century. The following sections are devoted to some of the more important contributions on that discussion. In the main we will concentrate our attention on the estimation of linear models. In Section 2.2 we therefore introduce \( L_1^- \), \( L_2^- \) and \( L_\infty^- \)-norms in the estimation of linear models.
Due to the modern probability theory, we know that the normal distribution is not the only limit distribution. An important class of limit distributions is the class of stable distributions. For non-normal stable distributions the variance does not exist. Then the variance of the least squares estimator of the parameters in a linear model with non-normal stable distributed residuals, does not exist. Due to Laplace's analytical treatment, we then conclude that the $L_1$-norm estimator is preferable when estimating the simple model $y = \beta x + \varepsilon$ and $\varepsilon$ is non-normal stable distributed. In Section 2.2 it is more heuristically argued that $L_1$-norm based methods may be preferable when estimating the general linear model and the residual distribution is fat-tailed. The class of stable distributions is also of considerable importance in the distribution theory of the estimators under study. Therefore, stable distributions and their properties are reviewed in Section 2.3. In Section 2.4 some examples of models with fat-tailed distributions are given.

One of the drawbacks of $L_1$- and $L_\infty$-norm estimation was the difficulty in the computation. In Section 2.5 we give some ideas of how the computations are simplified due to the theory of linear programming. When the computation of $L_1$- and $L_\infty$-norm estimates was almost as simple as the computation of $L_2$-norm estimates, several Monte Carlo studies were performed to compare, especially the $L_1$- and $L_2$-norm based estimators. Some comments and results from Monte Carlo studies are given in Section 2.6. In Section 2.7 some concluding remarks from this chapter are given.
2.2 Linear models

An important class of the models that are applied today is the linear models. A linear model can be written as

\[ y = X\beta + \epsilon, \]  

(2.5)

where \( y \) is an \( n \)-dimensional vector with (endogenous) observations, \( X \) is an \( n \times m \) matrix with (exogenous) observations, \( \beta \) is an \( m \)-dimensional vector with (possibly) unknown parameters and \( \epsilon \) is an \( n \)-dimensional vector with unobservable residuals. The columns in \( X \), i.e. the predictor variables, are denoted by \( X_1, X_2, \ldots, X_m \) with corresponding elements \( x_{11}, x_{21}, \ldots, x_{nm} \). The residuals are caused by stochastic variations in the underlying population, measurement errors, effects due to idealizations in the functional form of the model, or erroneously excluded predictor variables. The model may consist of one or more relations. In the case of more than one relation we distinguish between vector regression, recursive and interdependent systems. However, it is always possible to write each relation in the form as (2.5). For interdependent systems, we demand that the system is identified. A relation of the form (2.5) is then a relation in the reduced form of the model.

Our problem is to estimate the parameter vector \( \beta \). An \( L_1 \)-norm (least absolute) estimate is given by a vector \( b \) that minimizes

\[ \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{m} b_j x_{ij} \right|, \]  

(2.6)

an \( L_2 \)-norm (least squares) estimate of \( \beta \) is given by a vector \( b \) that minimizes

\[ \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{m} b_j x_{ij} \right)^2, \]  

(2.7)

and an \( L_\infty \)-norm (Chebychev) estimate of \( \beta \) is given by a vector \( b \) that minimizes

\[ \max_{i} \left| y_i - \sum_{j=1}^{m} b_j x_{ij} \right|. \]  

(2.8)
A simple case of the model (2.5) occurs when there is only one predictor variable and the value of that variable is always 1. The model then reduces to $y = \beta + \epsilon$, where $\beta$ is a location parameter in the probability distribution of $y$. The $L_1$-norm estimate of $\beta$ equals the sample median of $y$. Further, the $L_2$-norm estimate of $\beta$ equals the sample mean of $y$ and the $L_\infty$-norm estimate equals the sample midrange.

The method of least squares is both wellknown and often applied by the most statisticians. The properties of the method have been the subject in many studies and are rather wellknown. The assumptions demanded for least squares estimation are also wellknown.

$$E(\epsilon \epsilon') = \sigma^2 I,$$

where $I$ is the (nxn) identity matrix and $\sigma^2$ is the variance of the residual distribution, and

$$E(x_j \epsilon) = 0 \quad j=1,2,\ldots,m.$$

The vector $b$ minimizing (2.7) is explicitly given by

$$b = (X'X)^{-1} X'y.$$

The least squares estimates are thus rather simple to calculate, as Gauss already has pointed out. It is also seen that $b$ is a linear function of $y$. This is important when deducing other characteristics of the estimator, e.g. the expected value

$$E(b) = \beta$$

and variance

$$V(b) = \sigma^2 (X'X)^{-1}.$$

The linearity of the estimator also implies that it is rather easy to deduce the probability distribution of $b$, when the residual distribution is known. This simplifies the inference.
Due to the central limit theorem, the sum of a great number of independent stochastic variables, all with finite variances, is approximately normally distributed. It is then often argued that it is plausible to assume normally distributed residuals. When the residuals are normally distributed, it follows from the Gauss-Markov theorem that in the unirelational case, the least squares estimator is the best linear unbiased estimator (BLUE), with "best" here meaning the minimum variance. We also know that when the residuals are normally distributed, the least squares estimator is equivalent to the maximum likelihood estimator.

However, Edgeworth's arguments referred to earlier are still relevant, i.e. if residuals are distributed according to a contaminated distribution, the method of least squares is not suitable any longer. We also know a more general form of the central limit theorem, see e.g. Gnedenko and Kolmogorov (1954). Due to this general theorem, we know that under fairly general assumptions the sum of a great number of independent stochastic variables follows some stable distribution. The normal distribution is a special case of the class of stable distributions, and is the only stable distribution with finite variance. In Section 2.4 several examples of stochastic variables possessing infinite variance are given. It is immediate that residuals distributed according to a non-normal stable distribution do not satisfy the assumptions for least squares estimation.

As we already have noted, the $L_1$- and $L_2$-norm estimates of a location parameter correspond to the sample median and the sample mean, respectively. It is also wellknown that the mean is sensitive to and the median robust against great measurement errors, which are likely to occur when the distribution is fat-tailed or skew. It is possible to generalize these arguments to more complicated models. Consider now the least squares estimator. To each observation a weight is assigned. The weight is proportional to the square of the magnitude of the residual. Then, a very big weight is assigned to an observation with a great residual and con-
sequently, the estimate is affected by a big error. If great residuals appear with great probability, then there is also a great probability for great errors in the estimate. On the other hand, when estimating in $L_1$-norm, to each observation a weight, proportional to the magnitude of the residual, is assigned. For great residuals, this $L_1$-weight is much smaller than the corresponding $L_2$-weight. This implies that the $L_1$-norm estimates are less affected by observations with great residuals.

Thus, it is believed that the $L_1$-norm estimator is preferable when the residual distribution is fat-tailed. On the other hand, when the residuals are limited in magnitude (for example uniformly distributed) it is expected that an estimator more sensitive to large residuals (such as the $L_\infty$-estimator) would be preferable. For the simple location model $y_i = \beta + \varepsilon_i$, Rice and White (1964) illustrated this by computing the asymptotic variances of estimators of $\beta$. See Table 2.1.

<table>
<thead>
<tr>
<th>Residual distribution</th>
<th>$L_1$-norm</th>
<th>$L_2$-norm</th>
<th>$L_\infty$-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>$\frac{1}{4n}$</td>
<td>$\frac{1}{12n}$</td>
<td>$\frac{1}{2n^2}$</td>
</tr>
<tr>
<td>Triangle</td>
<td>$\frac{\pi}{2n}$</td>
<td>$\frac{1}{6n}$</td>
<td>$\frac{(4-\pi)}{4n}$</td>
</tr>
<tr>
<td>Normal</td>
<td>$\frac{\pi^2}{2n}$</td>
<td>$\frac{1}{n}$</td>
<td>$\frac{\pi^2}{12\ln n}$</td>
</tr>
<tr>
<td>Laplace</td>
<td>$\frac{1}{2n}$</td>
<td>$\frac{2}{n}$</td>
<td>$\frac{\pi^2}{12}$</td>
</tr>
<tr>
<td>Cauchy</td>
<td>$\frac{\pi^2}{4n}$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

Table 2.1 Asymptotic variances of the $L_1$, $L_2$, and $L_\infty$-norm estimators of $\beta$ in the simple location model $y_i = \beta + \varepsilon_i$, as is given in Rice and White (1964).

In this case it is seen that the $L_1$-norm estimator is, as expected, most effective for fat-tailed residual distributions, while the $L_\infty$-norm estimator is most effective for residual distributions with sharply
defined extremes. The least squares estimator is optimal for the normal distribution and performs well for distributions with reasonably light tails.

Applications with distributions without tails are not discussed in the literature as much as applications involving fat-tailed distributions. However, Hand (1978) gives two important examples. The first example concerns tabled functions. Suppose that a function \( y = f(x) \) has been measured quite accurately, and that rounded values have been tabulated. Since the uniform distribution is appropriate for round off errors, the \( L_\infty \)-norm estimator would be useful in the fitting of a function to approximate the tabled function.

For the second example in Hand (1978), suppose in some application, units must meet certain specifications within a given tolerance or else they are rejected. This procedure would produce a truncated distribution of residuals with limited range in the accepted population. If the distribution is sufficiently flat in the acceptance interval, the \( L_\infty \)-norm estimator would again be an appropriate estimator.

We have already noted that \( L_1 \)-norm estimation led to computational difficulties during the 18th century. In Section 2.5 we see that nowadays there are efficient algorithms available to solve the minimization problem (2.6). Measured in computer time, the \( L_1 \)-norm estimation is comparable with least squares estimation.

Quite early it was recognized that the \( L_1 \)-norm estimator did not always lead to unique estimates. An early paper on that subject is Turner (1887). For example, the \( L_1 \)-norm estimate of \( \beta \) in the simple model \( y = \beta + \epsilon \), is not uniquely determined when the number of observations is even and the two observations on \( y \) "in the middle" are unequal. The problem of uniqueness in \( L_1 \)-norm estimation is still not entirely solved. However, in the next chapter a discussion on conditions for unique \( L_1 \)-norm estimation will be found.
Another drawback of the $L_1$- and $L_\infty$-norm estimators is that their sampling theory is almost unknown. Interesting results appear in Rosenberg and Carlsson (1970) concerning the $L_1$-estimator. One important result there is given within a theorem on the existence of moments and the consistency of the $L_1$-norm estimator. It is shown that the conditions for finite variances in $L_1$-norm estimation are weaker than the corresponding conditions for $L_2$-norm estimation. This means that it is quite possible that the $L_1$-norm estimator will have finite variance, although the residual distribution has infinite variance. Some results concerning the mechanisms of $L_1$-norm estimation appear also in Taylor (1974).

2.3 Stable distributions and their domains of attraction

In this section we will define and give some properties of the class of stable distributions. The notation $L(U) = L(V)$ will be used to indicate that the two random variables $U$ and $V$ have the same distribution. The notation $U_n \Rightarrow V$ is used to indicate that the sequence $\{U_n\}_{n=1}^\infty$ of random variables converges in distribution to the random variable $V$.

A random variable $X$ is said to have a stable distribution if for every integer $n > 0$ and independent random variables $X_1, \ldots, X_n$ with the same distribution as $X$, there are constants $a_n > 0$ and $b_n$ such that

$$L(X_1 + \ldots + X_n) = L(a_n X + b_n).$$

It is immediately seen that if $X \sim N(0,1)$ we can choose $a_n = \sqrt{n}$ and $b_n = 0$ to satisfy the definition. Thus, the standardized normal distribution is one member of the class of stable distributions. A generalization to the general normal distribution is immediate.

As one might expect from the definition, the stable distributions are important when investigating the convergence in distribution of normed sums $(X_1 + \ldots + X_n)/a_n - b_n$ of independent and identically distributed
random variables \( X_1, \ldots, X_n \). We state the relationship between the stable distributions and the central limit problem in the next theorem.

**THEOREM 2.1 (A CENTRAL LIMIT THEOREM)** \( X \) is the limit of a normed sum
\[
S_n = \frac{X_1 + \ldots + X_n}{a_n} - b_n
\]
of independent and identically distributed random variables \( \{X_i\}_{i=1}^{\infty} \) if and only if \( X \) is stable distributed.

The proof of Theorem 2.1 is nowadays rather standard and can be found in advanced textbooks in probability theory, e.g. Gnedenko and Kolmogorov (1954), Breiman (1968) and Feller (1971).

It is possible to show that the constants \( a_n \) are of the form \( n^{1/\alpha} \) with \( 0 < \alpha \leq 2 \). We omit the proof of this statement and refer the interested reader to Feller (1971). The constant \( \alpha \) is called the characteristic exponent of the distribution.

The characteristic exponent plays an important role in the analysis of stable distributions. In particular, it is central in the computation of the characteristic function of a stable distribution. This computation is a little tricky and tedious. Therefore we refer the interested reader to Feller (1971) for a proof and here we give only the results.

**THEOREM 2.2 (THE CHARACTERISTIC FUNCTIONS OF STABLE DISTRIBUTIONS)** The characteristic function \( \varphi(t) \) of a stable distribution is given by
\[
\log \varphi(t) = i t \gamma - \delta |t|^\alpha \{1 + \beta \frac{t}{|t|} \omega(t, \alpha)\}
\]
where
\[
\omega(t, \alpha) = \begin{cases} \tan \frac{\pi \alpha}{2} & \text{for } \alpha \neq 1 \\ 2 \pi \ln |t| & \text{for } \alpha = 1 \end{cases}
\]
and \( \alpha \in ]0,2] \) is the characteristic exponent, \( \beta \in [-1,1] \) is a skewness parameter (when \( \beta = 0 \) the distribution is symmetric), \( \gamma \) is a real location parameter and \( \delta \in ]0,\infty[ \) is a scale parameter.
To indicate that a random variable \( X \) is stable distributed with certain parameters, we use the notations \( X \sim S(\alpha, \beta, \gamma, \delta) \).

The stable distributions thus have a specific characteristic function, and the class of stable distributions is completely determined. For most values of the parameters \( \alpha, \beta, \gamma \), and \( \delta \), the analytical form of the frequency and/or distribution function is not known. However, when the characteristic exponent \( \alpha = 2 \), we have the normal distribution with expectation \( \gamma \) and variance \( 2\delta \). The parameter combination \( \alpha = 1 \) and \( \beta = 0 \) corresponds to the Cauchy distribution. In these cases and a few more, the form of the distribution function is known. In the other cases, series expansion techniques have yielded numerical approximations of the density functions. These expansions were first given in Bergström (1952), and have then been used in e.g. Bartels (1972).

It can be shown that every stable distribution with characteristic exponent \( \alpha < 2 \), has finite absolute moments \( E[|X|^h] \) of order \( h, h < \alpha \). All absolute moments of order \( h > \alpha \) do not exist for \( \alpha < 2 \). Thus, every non-normal stable distribution possesses and infinite variance, and only those with \( \alpha > 1 \) have a finite expectation. For the normal distribution, we know that all moments exist. Probability density functions for a normal and a non-normal symmetric stable distribution are shown in Figure 2.7.

![Figure 2.7](image-url) Probability density functions for a normal and a non-normal symmetric stable distribution.
The characteristic functions of symmetric stable distributions were mentioned by Cauchy, but it was not clear that they really corresponded to probability distributions. The general theory of stable distributions was initiated by Lévy (1924 and 1937).

We will now turn to the question of necessary and sufficient conditions on a distribution function $F$ such that a suitable normalized sum of independent random variables, with the common distribution function $F$, converge in distribution to a stable distributed random variable. Thus, we first define a concept of central importance, viz. the concept of domain of attraction.

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent and identically distributed random variables with the common distribution $F$. If there exist constants $a_n, b_n$ such that the sequence $\{S_n\}_{n=1}^{\infty}$ of partial sums $S_n = (X_1 + \ldots + X_n) / a_n - b_n$ converges in distribution to a stable distribution with characteristic exponent $\alpha$, then $F$ is said to belong to the domain of attraction, $D \in (\alpha)$, with characteristic exponent $\alpha$. We denote this by $F \in D(\alpha)$.

The problem is now to give necessary and sufficient conditions on the distribution function $F$ to be in a domain of attraction of a stable distribution with characteristic exponent $\alpha$. However, since the proof again is a little tricky and tedious, we only state the results. The proofs are found in e.g. Gnedenko and Kolmogorov (1954), Breiman (1968) and Feller (1971).

**Theorem 2.3 (Domain of Attraction)** Let $F$ be a distribution function, then

1° $F \in D(2)$ if and only if

$$\lim_{y \to \infty} \int \frac{y^2 f}{|x|} \frac{dF(x)}{y^2} = 0$$

for $|x| > y$.
2° \( F \in D(\alpha), \ 0 < \alpha < 2 \) if and only if there exist constants \( M^+ > 0, \ M^- > 0, \ M^+ + M^- > 0 \) such that

\[
\begin{align*}
a) \quad & \lim_{x \to \infty} \frac{F(x)}{1 - F(x)} = \frac{M^-}{M^+} \\
b) \quad & \text{for every } y > 0 \\
\quad & \lim_{x \to \infty} \frac{1 - F(xy)}{y^\alpha} = \frac{1}{1 - F(-x)} \\
\quad & \lim_{x \to \infty} \frac{F(-xy)}{F(-x)} = \frac{1}{y^\alpha}
\end{align*}
\]

The problem of domains of attraction was first analysed in Doeblin (1939 and 1940).

For distributions belonging to a domain of attraction, an interesting result on the existence of their moments appear in the next lemma. The proof of the lemma may be found in e.g. Feller (1971).

**Lemma 2.1** A distribution \( F \) belonging to a domain of attraction with characteristic exponent \( \alpha \) possesses absolute moments of all orders \( h < 2 \). If \( \alpha < 2 \) no moments of order \( h > \alpha \) exist.

The univariate family of stable distributions was extended to the multivariate case by Lévy (1937). The definition of multivariate stable distributions is an obvious extension of the univariate case. It is also easy to generalize Theorem 2.1 to obtain a multivariate central limit theorem for stable distributions. The characteristic function of a multivariate stable distribution was first computed by Lévy (1937) and Feldheim (1937). After a certain amount of algebra, it was possible for Ferguson (1955) and Press (1972) to give a simplified version of the characteristic function. Here we only state the simplified result and leave the proofs.
THEOREM 2.4 (THE CHARACTERISTIC FUNCTIONS OF MULTIVARIATE STABLE DISTRIBUTIONS)

The characteristic function of an $m$-variate stable distribution is given by

$$
\log \varphi(t_1, \ldots, t_m) = i\gamma(t_1, \ldots, t_m) - \delta(t_1, \ldots, t_m)\{1 + i\beta(t_1, \ldots, t_m)\omega(1, \alpha)\}
$$
or more simply

$$
\log \varphi(t) = i\gamma(t) - \delta(t)\{1 + i\beta(t)\omega(1, \alpha)\}
$$

where $\omega(t, \alpha)$ is defined as in Theorem 2.2, $\alpha \in [0, 2]$ and where for every scalar $s$:

$$
\begin{align*}
\beta(t_1 s, \ldots, t_m s) &= \frac{s}{|s|} \beta(t_1, \ldots, t_m), \quad \beta(t) \in [-1, 1], \\
\gamma(t_1 s, \ldots, t_m s) &= \gamma(t_1, \ldots, t_m)s - \\
&\quad - \delta(t_1, \ldots, t_m)\beta(t_1, \ldots, t_m)\frac{|s|}{|s|} [\omega(s, \alpha) - \omega(1, \alpha)], \\
\delta(t_1 s, \ldots, t_m s) &= |s|^\alpha \delta(t_1, \ldots, t_m), \quad \delta(t) \in [0, \infty[.
\end{align*}
$$

The distribution is multivariate symmetrically stable if and only if $\beta(t_1, \ldots, t_m) = 0$. To indicate that a random $m$-vector $X$ is stable distributed with certain parameters, we use the notation $X \sim S_m(\alpha, \beta, \gamma, \delta)$. Observe that $\beta, \gamma$ and $\delta$ are here functions rather than scalars.

In particular, it follows that

$$
\begin{align*}
\gamma(t) &= i\mathbf{t}^\prime \mathbf{y} \\
\delta(t) &= \frac{1}{2} \mathbf{t}^\prime \Sigma \mathbf{t}
\end{align*}
$$

satisfy the conditions on the functions $\gamma$ and $\delta$ for an $m$-vector $\mathbf{y}$ and a positive definite $(mxm)$-matrix $\Sigma$. Thus, for $\alpha = 2$ we obtain

$$
\log \varphi(t) = i\mathbf{t}^\prime \mathbf{y} - \frac{1}{2} \mathbf{t}^\prime \Sigma \mathbf{t}
$$

and the $m$-variate normal distribution with expectation vector $\mathbf{y}$ and covariance matrix $\Sigma$ is an $m$-variate stable distribution.
2.4 Examples of models with fat-tailed or contaminated
distributed random variables

2.4.1 Definitions

The notion of fat-tailed distributed random variables is rather vague. Intuitively, we mean that it is likely that an observation on a fat-tailed distributed random variable substantially differs from some location parameter. Here we will use the notion of fat-tailed distributions if the distribution possesses an infinite variance. We will use the notion contaminated if the density functions $f$ is of the form

$$f(y) = \lambda f_1(y) + (1-\lambda)f_2(y)$$

where $\lambda$ is near 1, $f_1$ is the density of the normal distribution and $f_2$ is the density of an arbitrary distribution.

As is seen, the properties of the $L_2$-norm estimator are poor, when residuals of great magnitude are likely to occur. It is precisely in those cases, that the $L_1$-norm estimator is expected to be superior. However, this observation is of limited importance if we can not show applications where residuals of great magnitude are likely to occur. Now, there are in fact, a great number of applications with random variables possessing a fat-tailed or contaminated distribution; much more than one first might expect. Due to this fact, the purpose of this section is not to give a complete list of all applied models with fat-tailed or contaminated distributions. The purpose is rather to give small but representative sample of such models. Thus, in Subsection 2.4.2 we will give a few examples of theories leading to non-normal stable distributions. In some applications where no theory to a definite distribution are available, empirical observations have been made and found to be fat-tailed. Subsection 2.4.3 discusses that topic. A discussion of contaminated distributions may be found in Subsection 2.4.4 and some concluding remarks in Subsection 2.4.5.
2.4.2 Non-normal stable distributions

As a non-normal stable distribution possesses an infinite variance, it is, with our notions, fat-tailed. It is easy to find applied models with stable distributions. Feller (1971, p. 173ff) mentions four interesting applications. We will now give a short description of the applications that Feller mentions.

First Feller discusses the Holtsmark distribution, which is the distribution of a gravity force in a certain direction at a randomly chosen point in a stellar system. The Holtsmark distribution is a symmetric stable distribution with the characteristic exponent \( \alpha = 1.5 \). Then it is clear that the variance of the Holtsmark distribution does not exist. The four dimensional counterpart to the Holtsmark distribution is a symmetric stable distribution with characteristic exponent \( \alpha = 4/3 \).

The next application Feller discusses, concerns Brownian motion. He starts from the notion of a one-dimensional diffusion process, i.e. the increments \( x_{s+t} - x_s \) for non-overlapping time intervals are independent and normally distributed with variance \( t \). If the process starts with \( x_0 = 0 \), there exists a point of time \( T_a \), where the process for the first time reaches the level \( a > 0 \). Then, \( T_a \) is stable distributed with a characteristic exponent \( \alpha = 0.5 \), i.e. neither the expectation nor the variance exist.

A two-dimensional Brownian motion is formed by a pair \((x_t, y_t)\) of independent one-dimensional Brownian motions. As a third application Feller discusses the point \((a, z_a)\) where a two-dimensional Brownian motion for the first time reaches the line \( x = a > 0 \). The hitting point \( z_a \) is then Cauchy distributed, i.e. symmetric stable distributed, with characteristic exponent \( \alpha = 1 \).

The fourth application of stable distributions, that Feller discusses concerns economic processes. Income distributions were among the first to
be discovered as being fat-tailed distributed within economic applications. After that, a lot of other types of economic variables have shown to be fat-tailed distributed. In the context of demand analysis, Arrow and Hoffenberg (1959) is one early example of that.

A valuable contribution on non-normal stable distributions in economic applications is also given by Bartels (1977). He argues that the conditions required for non-normal stable distributions are frequently satisfied by economic variables.

Stable distributions seem to be important even in hydrological applications. Boes and Salas-LaCruze (1973) showed that the stable distributions are central in studies of storage capacity of reservoirs, e.g. water reservoirs.

2.4.3 Empirical observations

Unless one or more of the terms in a sequence (of finite length) of observations on a stochastic variable takes infinite value, it is difficult to decide whether the underlying distribution possesses infinite variance. Since it is likely there will not be many observations from the tails of the underlying distribution, the chi-square goodness-of-fit test can never reveal an ill-fitting tail without a very large amount of data. Likewise, the observed cumulative distribution function will be close to zero or one in the tails, so that a Kolmogorov-Smirnov test will also fail.

Provided the distribution is stable, one method is to plot the observed cumulative distribution function on normal probability paper. Then if the plot deviates substantially from a straight line, the underlying distribution is not normal, and hence has infinite variance. However, if as well non-stable distribution alternatives are accepted, curvature in the plot on normal probability paper is not conclusive evidence of infinite variance.

Another method for testing infinite variance is to plot the observed
variance estimate $s^2_n$, based on the first $n$ observations, versus $n$, i.e.

$$s^2_n = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

where

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i$$

If the underlying distribution possesses finite variance, then the sequence $s^2_n$ should converge to a finite value. The non-convergence of $s^2_n$ does not however imply infinite variance. For example, the sequence $s^2_n$ is increasing over the observed time period if the variables $Y_i$ form a nonstationary sequence with increasing variance. Even for a stationary sequence with finite variance, convergence would be slowed by the presence of serial dependence.

This method has been used in Mandelbrot (1963) concerning changes of daily cotton spot prices, and in Mandelbrot (1967) concerning common stock price changes. For most of these sequences, the convergence is not clear-cut. However, for several sequences the variance shows no signs of diverging either. These observations indicate the difficulty in determining the value of the results.

A third method, originally proposed by Mandelbrot (1963), is to plot the estimate of $\log P(Y > u)$ against $\log u$. (In examining the lower tail, $\log P(Y > u)$ is plotted against $\log |u|$). Since a random variable $Y$ with a stable distribution satisfies the relation (see e.g. Feller, 1971)

$$\lim_{y \to \infty} y^{-\alpha}P(Y > u) = \text{constant}$$

the plot should be a straight line with slope $-\alpha$ for large $y$, if the underlying distribution is stable. Only stable distributions will have this property. A linear plot is then a strong indication that the distribution is stable, and if the slope is substantially greater than $-2$, the variance may be infinite.
This method has been used by Mandelbrot (1963 and 1967) concerning stock market data, and in Granger and Orr (1972) concerning U.S. treasury cash flows.

A weakness of all these methods is that they are all based on simple judgemental visual inspections of a graph, and not a reading on a test statistic vis-à-vis an associated critical interval.

More formal test statistics for testing normality against non-normal symmetric stable alternatives are the Shapiro-Wilk (SW) statistic and the studentized range (SR) statistic. Let $y_{(i)}$ be the $i$th order statistic from a sample of size $n$. Then the studentized range (see David et al., 1954) and the Shapiro-Wilk statistic (see Shapiro and Wilk, 1965) are defined as

$$SR = \frac{(y_{(n)} - y_{(1)})}{\left[\frac{1}{n-1} \sum_{i=1}^{n} (y_{(i)} - \bar{y})^2\right]^{1/2}}$$

$$SW = \frac{I(n/2)}{\sum_{i=1}^{n} a_{n-i+1}(y_{(n-i+1)} - y_{(i)})/\sum_{i=1}^{n} (y_{(i)} - \bar{y})^2}$$

where $\bar{y}$ is the sample mean, $I(n/2)$ is the greatest integer equal to or less than $n/2$, and the weights $a_{n-i+1}$ are the approximated coefficients suggested by Shapiro and Wilk (1965, p. 596). In Fama and Roll (1971) powers of these tests are reported. Fama and Roll then make the interesting observation that the power of the SR and SW tests, respectively, are better than the power of a number of other goodness-of-fit tests, viz. Sherman's (see Sherman, 1950), the Kolmogorov-Smirnov (see Kolmogorov, 1933), Cramér-von Mises (see Cramér, 1945), Anderson-Darling (see Anderson and Darling, 1954), Durbin's modified median, modified probability and modified Kolmogorov tests (see Durbin, 1961) and the chi-square test with 9 degrees of freedom.

In Saniga and Miles (1979) powers of some tests of normality against asymmetric alternatives are reported. Besides the SR and SW tests, tests based on the skewness and kurtosis statistics are investigated. Monte
Carlo experiments indicated that tests based on the skewness statistic and SW statistic seemed to be the most powerful for small sample sizes \(n<20\), \(\alpha<1.3\) and \(|\beta|<0.5\). For larger sample sizes \(n>50\), \(|b|<0.75\), the test based on the kurtosis statistic was generally preferred. When \(|\beta|\) was large, \((|\beta|>0.75)\), the test based on the skewness statistic was usually preferred for large sample sizes \(n>50\).

If it is known that the distribution is stable, an estimate of the characteristic exponent \(\alpha\) could be a guide in determining the existence of the variance. For symmetric stable distributions, an estimator of \(\alpha\) is proposed in Fama and Roll (1971). The estimate is obtained by first standardize the observed distribution for different scale parameters

\[
z(f) = 0.827 \frac{y(f) - y(1-f)}{y(0.72) - y(0.28)}
\]

where \(y(f)\) indicates the \(f\)-th fractile in the observed sample. Given that \(Y\) is symmetric stable distributed with characteristic exponent \(\alpha\), \(z(f)\) is an estimator of the \(f\)-th fractile in a standardized symmetric stable distribution with characteristic exponent \(\alpha\). It is then possible to find an estimate \(\hat{\alpha}_f\) of \(\alpha\). The estimate \(\hat{\alpha}_f\) depends on the fractile \(f\) chosen. It seems difficult to find an optimal value of \(f\). In Monte Carlo experiments reported in Fama and Roll (1971), values of \(f\) in the interval \([0.93, 0.99]\) were investigated. The results indicated that for \(\alpha > 1.9\), the best estimator in terms of bias as well as standard deviation, is \(\hat{\alpha}_{0.99}\). For \(\alpha < 1.9\), lower values of \(f\) were preferable. Thus, for \(\alpha\) near 1.0, \(\hat{\alpha}_{0.95}\) showed relatively good results.

All parameters in a stable distribution can be estimated with a method based on the empirical characteristic function. Given the observations \(y_i, i=1, \ldots, n\), the empirical characteristic function \(\hat{\phi}(t)\) is defined by

\[
\hat{\phi}(t) = \frac{1}{n} \sum_{i=1}^{n} \exp(it y_i)
\]
Note that $\{ \hat{\varphi}(t) | -\infty < t < \infty \}$ is a stochastic process, and for each $t$, $|\hat{\varphi}(t)|$ is bounded above by unity. Hence all moments of $\hat{\varphi}(t)$ exist. For a fixed $t$, $\hat{\varphi}(t)$ is the mean of independent and identically distributed random variables. By the law of large numbers, $\hat{\varphi}(t)$ is a consistent estimator of the characteristic function $\varphi(t)$.

Since two distributions are equal if and only if their respective characteristic functions are equal, the parameter estimates are the parameter values that minimize

$$\sup_{t} |\varphi(t) - \hat{\varphi}(t)|$$

or, for fixed $p \geq 1$,

$$\int_{-\infty}^{\infty} |\varphi(t) - \hat{\varphi}(t)|^p W(t) dt$$

where $W$ is a weighting function. Observe that (2.9) is the Chebychev norm ($L_{\infty}$) and that (2.10) is a weighted $L_p$-norm in a space of functions.

In Leitch and Paulson (1975) (2.10) is used with $p = 2$ and $W(t) = \exp(-s^2)$ to estimate the parameters in distributions of changes in stock prices. The results support earlier results, viz. the distributions are too fat-tailed to possess an existing variance. Further, estimates of the skewness parameter indicate that symmetry is rather the expectation, not the rule in these distributions. This asymmetry results support results from a study of Fieletz and Smith (1972).

Besides the articles already mentioned, empirical evidence in favour of fat-tailed distributions are given e.g. in Fama (1965,a) concerning stock-market prices, in Houthakker and Taylor (1966) concerning demand equations, in Roll (1968) concerning interest rate changes, in Fama (1965,b), in Blume (1968), in Sharpe (1971), in Arad (1980) and in Simkowitz and Beedles (1980) concerning portfolio and security theory, in Carlson (1975) concerning price expectations, and in Seastrand (1978) and Dijk and Kloek (1980) concerning income distributions.

We are forced to conclude that many of the studies are based on vague methods. However, the variety of studies makes it difficult to reject the hypothesis of non-normal stable distributions in some applications.
2.4.4 Contaminated distributions

In a paper from 1886, Newcombe notes "In practice, large errors are more frequent than this equation (the normal distribution) would indicate them to be." He suggests that the square exponent of the normal density function should be replaced by a less rapidly increasing function. "The management of such an exponent might, however, prove inconvenient, and I shall adopt a law of error founded on the very probable hypothesis that we are dealing with a mixture of observations having various measures of precision."

The idea that the observations are drawn from distributions with unequal scale parameters, is the same as that in the papers of Edgeworth from 1887 (in which he proposes the use of $L_1$-norm estimation).

After the 1880's there was essentially no progress on the idea of contaminated distributions until Tukey and the Statistical Research Group at Princeton in the 1940's began to work with robust estimation problems. A survey of this work was later published by Tukey (1960); compare also Tukey (1962).

In the contamination model (Gauss-error-model, Huber, 1964) the probability density function $f$ of the sampled population is assumed to be of the form

$$f(y) = \lambda f_1(y) + (1-\lambda)f_2(y)$$

where $\lambda$ is near 1, $f_1$ is the density function of (e.g.) the standard normal distribution and $f_2$ is the density of an arbitrary distribution. The interpretation is that an observation is "good" with probability $\lambda$ but may be anything with the small probability $1-\lambda$.

In Granger and Orr (1972) two examples of contaminated data are given: The first concerns a firm's weekly production series. The contamination can there be due to a strike, a sales promotion, or an annual vacation
shutdown. In daily interest rate changes, which is the second example, the result of a governmental policy action, such as a large open-market purchase, might be viewed as contamination.

An example from high energy physics is given in Henriksson (1972). Particles were emitted from a proton synchrotron into a spark chamber. The trajectory of a particle was then represented by a spark. However, spurious sparks due to irregularities in the spark chamber and the presence of cosmic particles may occur. Therefore, observations that do not correspond to the trajectory under study occur with a certain probability. Since a trajectory was assumed to be normally distributed about a "mean trajectory", the observations were contaminated distributed.

The contamination model has been extensively used in the context of robust estimation, see e.g. Hampel (1973) and Huber (1977).

In linear model theory the problem of contaminated distributed observations is known as the problem of heteroscedasticity. The methods designed to handle the heteroscedasticity problem are usually based on a weighting system in combination with the least squares estimator. However, when it is not known from which population an observation is drawn, it is difficult to assign such a weight system.

2.4.5 Conclusions

In this section the notions of fat-tailed distribution and contaminated distribution have been defined. Some examples of fat-tailed distributed and contaminated distributed variables have also been given. In some applications there exists a more or less detailed theory on the underlying generation mechanism of a random variable. In those cases it is often possible to deduce the probability distribution of the variable. Due to existing theories it is then possible to decide whether the random variable is fat-tailed distributed/contaminated distributed or not. Examples
of that are given in Subsections 2.4.2 and 2.4.4. In cases when it is impossible to deduce the distributional properties of a random variable under study, the decision may be based on empirical observations. Powers of the usually applied goodness-of-fit tests will in many cases, however, be poor. Therefore, more efficient methods have been developed. Examples of these methods and applications of them are given in Subsection 2.4.3.

The examples given indicate that the notions of fat-tailed distributions and contaminated distributions are of great importance in a great number of applications. It is in those non-normal cases the $L_1$-norm alternative is expected to be superior to the $L_2$-norm estimator. Therefore, the $L_1$-norm estimator is expected to be suitable in many applications.

2.5 Algorithms for computations of $L_1$- and $L_\infty$-norm estimates

One of the advantages of the method of least squares is that it is computationally simple. The computation is equivalent to the solution of a linear equation system, i.e. the equation system that is known as the normal equations. The computation of $L_1$-norm and $L_\infty$-norm estimates is today computationally not much more difficult than to compute least squares estimates. That was not the case in the 19th century. The computation of an $L_1$-norm estimate of models with one predictor variable was early known. In the 1820's Fourier was concerned with the computation of $L_1$-norm estimates in models with two predictor variables. The method of Fourier has much in common with the simplex method. Edgeworth treated the same problem in his work from 1887. The computation of $L_1$-norm estimates in simple regression models has also been studied by Rhodes (1930) and Karst (1950).

A computation algorithm for $L_1$-norm estimation of models more general than the two-variable case, was not known before Charnes, Cooper and Ferguson (1955), Wagner (1959) and Fisher (1961).
The computation of the $L_1$-norm estimate of a general linear model is equivalent to compute the solution of the optimization problem

$$\min_{i=1}^{n} \sum_{i=1}^{n} |\hat{c}_i|$$

subject to the restrictions

$$y_1 = \sum_{j=1}^{m} b_j x_{1j} + \hat{e}_1$$
$$\vdots$$
$$y_n = \sum_{j=1}^{m} b_j x_{nj} + \hat{e}_n,$$

where $\hat{e}_i$, $i=1,2,...,n$ are the estimated residuals.

When defining

$$\hat{e}_i = u_i - v_i \quad i=1,2,...,n,$$

where $u_i > 0$, $v_i > 0$

and $b_j = b_j' - b_j'' \quad j=1,2,...,m,$

where $b_j' > 0$, $b_j'' > 0$,

we can write the problem in the canonical form, i.e. we obtain the linear program

$$\min_{i=1}^{n} \sum_{i=1}^{n} u_i + \sum_{i=1}^{n} v_i$$

subject to

$$\sum_{j=1}^{m} (b_j' - b_j'')x_{1j} + u_1 - v_1 = y_1$$
$$\vdots$$
$$\sum_{j=1}^{m} (b_j' - b_j'')x_{nj} + u_n - v_n = y_n.$$

The computation of an $L_1$-norm estimate is then equivalent to solve a linear program. This formulation of the problem was first given by Barrodale and Young (1966). There are several suggestions how to improve the algo-
rithm of Barrodale and Young, e.g. Roberts and Ben-Israel (1969). Abdelmalek (1971) and Spyropoulos et al. (1973). Probably the best algorithm today is that of Barrodale and Roberts (1972), (see the discussion in Ekblom, 1974). When estimating simple regression models, a specialized version of the Barrodale-Roberts algorithm has been presented in Armstrong and Tam Kung (1978).

To compare computation times when estimating a linear model under $L_1$- and $L_2$-norm, respectively, we performed a simple experiment. Three models were considered, one model with intercept and one exogenous variable, one model with five and one model with ten exogenous variables. All exogenous variables were drawn independently from a uniform $[-10,10]$ distribution. Residuals were drawn from a standard normal distribution and the endogenous variables were computed when all parameters were set to one. The sample sizes considered were $n=10, 30$ and $100$. Only one replicate of each model was performed.

Computations for estimating the models were made on the CDC Cyber 172 at the University of Umeå using the Armstrong-Tam Kung (when applicable) and the Barrodale-Roberts-algorithm for $L_1$-norm estimation and the IMSL-routine LLSQF for the $L_2$-norm estimation. The obtained computation times in ms are presented in Table 2.2.

<table>
<thead>
<tr>
<th>number of exog. var.</th>
<th>number of observations</th>
<th>Computation time in ms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Armstrong-Tam Kung</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>44</td>
</tr>
<tr>
<td>(with intercept)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>252</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1330</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>611</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>4010</td>
</tr>
</tbody>
</table>

Table 2.2 Obtained computation times in ms when estimating different models with different number of observations.
When computing the $L_\infty$-norm estimate, we first define the variable $d$ as the absolute of the largest estimated residual,

$$d = \max_i |y_i - \sum_{j=1}^{m} b_j x_{ij}|$$

As in the computation of the $L_1$-norm estimate, we obtain a linear program

$$\min d$$

subject to the restrictions

$$\sum_{j=1}^{m} (b'_j - b''_j) x_{1j} - y_1 \leq d$$

$$- \sum_{j=1}^{m} (b'_j - b''_j) x_{nj} + y_1 \leq d$$

$$\vdots$$

$$\sum_{j=1}^{m} (b'_j - b''_j) x_{nj} - y_n \leq d$$

$$- \sum_{j=1}^{m} (b'_j - b''_j) x_{nj} + y_n \leq d$$

which can be solved in the usual manner. This formulation of the problem has been shown by many authors (see e.g. Kelley, 1958, Wagner, 1959, Stiefel, 1960, Appa and Smith, 1973 and Sposito, 1976). A very efficient algorithm for the solution of the problem has been described by Barrodale and Phillips (1974 and 1975). In Hand (1978) a modified version of the Barrodale-Phillips algorithm is presented. Experiments with the algorithms showed that the number of iterations is reduced 20-30 percent when using the modified version. However, the modified algorithm requires more computational effort per iteration, and the results in terms of total computation time were not totally conclusive.
2.6 Some Monte Carlo studies

2.6.1 On the use of Monte Carlo studies

In the last years, there has been a substantial increase in the number of simulation studies. The aim of the studies is often to find properties of estimators or significance tests, when an analytical treatment is difficult or impossible to realize. When studying estimators, the usual succession of tasks is represented by the following scheme:

1. A set of values of the elements in the matrix $X$ and the vector $\beta$ is determined.

2. The elements in the residual vector $\varepsilon$ are generated by a random number generator. The elements are then distributed according to a specified probability distribution.

3. The vector $y$ is computed by
   \[ y = X\beta + \varepsilon. \]

4. The fixed values on $\beta$ and the generated values on $\varepsilon$ are "forgotten" for a while. Due to the "observations" on $y$ and the known values on $X$, the vector $\beta$ is to be estimated by the estimators under study.

5. 2-4 are repeated $T$ times. Then, for each estimator a sample of $T$ estimates of $\beta$ is obtained. The sample now gives a picture of the sampling distributions of the estimators.

Based on the obtained sample, the estimators under study are now to be evaluated. The evaluating technique is of course dependent on the aim of the study. However, it is in common practice to estimate the bias of the estimator

\[
\text{bias } \hat{\beta}_j = \frac{1}{T} \sum_{k=1}^{T} \hat{\beta}_{jk} - \beta_j, \quad j=1,2,\ldots,m
\]

and the mean square error, $\text{MSE}$

\[
\text{MSE } \hat{\beta}_j = \frac{1}{T} \sum_{k=1}^{T} (\hat{\beta}_{jk} - \beta_j)^2, \quad j=1,2,\ldots,m,
\]
or the root mean square error, RMSE

$$\text{RMSE} = (\text{MSE}_j)^{1/2} \quad j=1,2,...,m,$$

where $\hat{\beta}_{jk}$ is the estimate of the parameter $\beta_j$ in the k-th replicate.

Note that if the estimator has infinite variance, MSE, and RMSE, are to increase in an unrestricted way for an increasing sample size. Furthermore, when the expected value of the estimator does not exist, the expression for bias does not converge. Since a Monte Carlo study always gives a sample of finite size, all measures are finite. This implies that if the variance and/or expected value does not exist, we estimate an infinite quantity with a finite one. If the purpose is to compare two estimators, the comparison will be dubious. The fact that the existence of the variances and/or expected values of the estimators in most cases are unknown, makes the situation even more complicated. In cases where the usual measures of precision may not exist it is better to compute more simple measures, e.g. to count the number each estimator produces the estimate that is closest to the true value.

When the purpose is to compare several estimators, it is also usual to estimate some summarizing measures of the qualities of each estimator. One such quality measure of an estimator is to count the number of parameters that in some meaning has been better estimated than by the other estimators under study. In some studies, the mean integrated square error, MISE, appears. MISE is a weighted sum of the MSE-values

$$\text{MISE} = \sum_{j=1}^{m} \gamma_j \text{MSE}_j$$

where

$$\gamma_j = \sum_{i=1}^{n} w_i x^2_{ij}$$

for some weighting system $w_i$, $i=1,2,...,n$. Besides that the MISE-measure is ad hoc due to the choice of weighting system, it is also affected with the problem concerning non existing variance, as we have already discussed.
When the distribution of the estimator is unknown, which is the most likely case in Monte Carlo studies, there is not at all a matter of course that quadratic differences between estimate and true parameter value is the only relevant basis for a comparison. In the quotation of Gauss in Section 2.1, for example, it is argued that every value of $\beta$ in

$$\sum_{k=1}^{T} |\hat{\beta}_k - \beta|^p$$

is to be considered as ad hoc.

Besides these difficulties when studying properties of estimators by Monte Carlo simulations, we must note that a simulation experiment only gives a sample of estimates. Inductively drawn conclusions based on samples always include a certain amount of uncertainty.

Thus, many objections can be raised against Monte Carlo simulations and against conclusions based on Monte Carlo simulations. However, one may not ignore the fact that simulation studies may illuminate a theory under progress.

2.6.2 Unirelational models

In a Monte Carlo simulation by Blattberg and Sargent (1971), a simple linear model without intercept was analysed. The residuals were generated according to symmetric stable distributions. The $L_1$-norm estimator was compared with the $L_2$-norm estimator and with the best linear unbiased estimator (BLUE) for each residual distribution. Best here means the least scale parameter in the distribution of the parameter estimator. For each residual distribution $n = 50$ observations were generated, and each experiment was repeated $T = 100$ times. The results indicate that for small values of the characteristic exponent $\alpha$, ($\alpha < 1.5$), the $L_1$-norm estimator was superior to the $L_2$-norm estimator and to the BLUE.

Smith and Hall (1972) reported a simulation study where the residuals were distributed according to the Laplace (or double exponential) distribution, i.e. with the frequency function
\[ f(\varepsilon_1) = \frac{1}{2\delta} \exp\{-|\varepsilon_1|/\delta\}. \]

The \( L_1 \)-norm estimator is then equivalent to the maximum likelihood estimator. The authors studied three models; two models with two predictor variables, where in one of the models, the predictor variables were highly correlated. The third model consisted of three predictor variables. The residuals were generated according to three different variance levels, i.e. three different values on \( \delta \). Thus, there were nine experiments, each consisting of \( n=20 \) observations. Each experiment was repeated \( T=50 \) times. The conclusions to draw were that the \( L_1 \)-norm estimator was at least marginally better than the \( L_2 \)-norm estimator.

Ekblom (1974) studied among other estimators, the \( L_1 \)-norm and \( L_2 \)-norm estimators. Two models were considered, one model with one, and one model with two predictor variables. Further, in each model there was an intercept. The residuals were generated according to different distributions. One residual distribution was chosen so that the \( L_2 \)-norm estimator was equivalent to the maximum likelihood estimator (i.e. \( N(0,1) \)) and one distribution so that the \( L_1 \)-norm estimator was a maximum likelihood estimator (i.e. Laplace distribution). In both of these residual distributions the variance and the expected value exist. A residual distribution where neither of these exist is the Cauchy distribution (i.e. a stable distribution with \( \alpha = 1.0 \)), which also was considered in the study. Further, three contaminated residual distributions were considered, namely \( N(0,1) + N(0,4) \), \( N(0,1) + N(0,9) \) and \( N(0,1) + N(4,1) \). Observe that the first two are symmetric and the third is skew. Another skew residual distribution under study was a non-central \( \chi^2(6) \). For each estimator and each model the relative MISE (i.e. MISE for the best estimator was divided by MISE for the other estimator), was computed and is reported in Table 2.3.
When the estimators were compared in this manner, the $L_2$-norm estimator was preferable when the residuals were pure normal. In all other cases except one, the $L_1$-norm estimator was superior.

An interesting Monte Carlo simulation was performed by Kadiyala and Murthy (1977). The authors there compared the $L_1$-norm and the maximum likelihood estimator. In the simulation a linear regression model with an intercept and two predictor variables was considered. Three sample sizes were under study, i.e. $n=20$, $n=50$ and $n=100$. Each experiment was repeated $T=100$ times. The residuals were generated from a Cauchy distribution. The maximum likelihood estimator often produced the best results. On the other hand, the $L_1$-norm estimator was slightly worse. Since the maximum likelihood estimator is expected to be sensitive to errors in the specification of the residual distribution, while the $L_1$-norm estimator is expected to be robust against such errors, it should be interesting to compare the estimators when the residual distribution is non-Cauchy.

### Table 2.3 Relative MISE for the $L_1$-norm and $L_2$-norm estimators for different residual distributions in the study by Ekblom (1974).

<table>
<thead>
<tr>
<th>Model</th>
<th>Estimator</th>
<th>Residual distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N(0,1)</td>
</tr>
<tr>
<td>$m=2$</td>
<td>$L_1$</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>$L_2$</td>
<td>1.00</td>
</tr>
<tr>
<td>$m=3$</td>
<td>$L_1$</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>$L_2$</td>
<td>1.00</td>
</tr>
</tbody>
</table>
2.6.2 Interdependent systems

When estimating the parameters in an interdependent system, the situation is more complicated. The parameters in the structural form could be estimated in one, two or three stages. In one-stage estimation, the parameters will be estimated relation for relation in the structural form under $L_1$- or $L_2$-norm. We call these estimators OLA (ordinary least absolute) and OLS (ordinary least squares), respectively. In two stage estimation the reduced form is first estimated relation for relation. The endogenous variables in the model are then estimated, due to the estimated reduced form. These variable estimates are used in the structural form, as instruments for the endogenous variables, and the parameters in the structural form are estimated relation for relation. The estimators in the two stages may either be based on $L_1$- or $L_2$-norm. We call the corresponding estimator 2SLA (2-stage least absolute) and 2SLS (2-stage least squares), respectively.

When estimating the parameters of the reduced form, an estimate based on any of the four methods mentioned above, may be transformed to an estimate of the reduced form parameters. It is also possible to estimate the reduced form parameters directly, by estimating the reduced form relation for relation under $L_1$- or $L_2$-norm. These two estimators are called LANR (least absolute no restrictions) and LSNR (least squares no restrictions), respectively.

Three stage estimators in other norms than $L_2$ have not been constructed. We can therefore exclude three-stage estimators in this discussion.

One of the first Monte Carlo simulations in order to compare $L_1$- and $L_2$-norm estimators of interdependent systems is Glahe and Hunt (1970). The model they were studying, was an overidentified two-relational model. The residuals were normally distributed. It was difficult to draw some general conclusions, but the $L_1$-norm estimators seemed to be as
good as, or even better than the $L_2$-norm estimators, although the residuals were normally distributed. In estimating the reduced form, LANR seemed to be superior to OLA and 2SLA, and in estimating the structural form, OLA seemed to be superior to 2SLA. LANR was better than LSNR in small sample sizes.

The same model and the same structural estimators were studied in Hunt, Dowling and Glahe (1974). Estimators of the reduced form were not under study. The residuals were distributed according to the Laplace distribution and to the normal distribution. The $L_1$-norm estimators produced all best estimates when the residuals were distributed according to the Laplace distribution, and 37.5 percent of the best estimates when the residuals were distributed according to the normal distribution.

In Nyquist and Westlund (1977) a three-relational interdependent system was under study. The residual distribution was stable with seven different levels on the characteristic exponent $\alpha$. Two sample sizes were considered, namely $n = 10$ and $n = 20$. Each experiment was repeated $T = 100$ times. Bias and RMSE for a parameter corresponding to one of the exogenous and a parameter corresponding to one of the endogenous variables are illustrated in Figure 2.8. The patterns for the other parameters are similar to the patterns of these two parameters. When the residuals are normally distributed, $\alpha = 2$, the $L_2$-norm estimators are preferable. When the residuals tend to be more non-normal, $\alpha$ decreases, the $L_1$-norm estimators are to be superior. The study also indicated that 2SLA was better than OLA, when the sample size was increasing.

Furthermore, predictions on the endogenous variables, based on each estimate, were generated. The predictions were then evaluated due to Ball's $Q^2$ (see Ball, 1963). The results are illustrated in Figure 2.9. When the residuals were normally distributed, 2SLS seemed to be the best alternative. For non-normal residual distributions, 2SLA was the best alternative, closely followed by LANR. In extremely non-normal cases, LANR seemed to be more robust than 2SLA.
a) Observed bias for a parameter corresp. to one of the endogenous variables.

b) Observed RMSE for a parameter corresp. to one of the endogenous variables.

c) Observed bias for a parameter corresp. to one of the exogenous variables.

d) Observed RMSE for a parameter corresp. to one of the exogenous variables.

Figure 2.8 Observed bias and RMSE of estimators of some parameters in the Monte Carlo experiment reported in Nyquist and Westlund (1977). The sample size is n = 20.
Figure 2.9 Prediction based on different estimators, evaluated due to Ball's $Q^2$. The data is from the Monte Carlo experiment reported in Nyquist and Westlund (1977). The sample size is $n = 20$.

2.6.4 Autoregressive processes

When analysing a stochastic process, techniques based on the sample autocorrelation sequence are often applied. In Nyquist (1980) effects on the sample autocorrelation sequence, when the probability distribution of the disturbance term is fat-tailed, are discussed. The discussion is based on a Monte Carlo study of a first order autoregressive process. $T = 200$ replicates of time series, each of length $n = 30$ for each model, was generated in the study.
When identifying the process, the results indicated that methods based on the sample autocorrelation sequence remain useful, when data are drawn from fat-tailed distributions. This is further supported in an unpublished lecture by Wise (1963) and in simulation studies reported in Granger and Orr (1972). An intuitive explanation is as follows. In Figure 2.10, a realization of a first order autoregressive process with symmetric stable distributed disturbances with $\alpha = 0.5$ is compared with a process with $\alpha = 2.0$. In the fat-tailed distribution case, a chock of considerable magnitude occurred at time $t = 11$. A smaller one, occurred at time $t = 28$. These two observations and the observations made at time points just after the shocks dominate the sums of squares and products in the computation of the sample autocorrelation coefficients. Since the observation at time $t = 12, \ldots, 20$ and $t = 29, 30$ are almost one half times the preceding value, the autoregression scheme appears very clear. Thus, the sample autocorrelation sequence shows the typical pattern of an autoregressive process.

In the familiar normally distribution case, the observations have a smaller magnitude. Relatively, the process fluctuates more and possesses no shocks of considerable magnitude, as in the fat-tailed distribution case.

In estimating the parameter in the process under study, an $L_1$-norm based alternative seemed to have a smaller MSE than the first term in the sample autocorrelation sequence.

The case when it is not possible to observe the process directly is discussed in Nyquist (1979). A simple linear regression model with a first order autoregressive residual process is there studied. The residual process is generated according to various members of the family of symmetric stable distributions. $n = 30$ observations and $T = 200$ replicates were used.
Figure 2.10 A realization of a first order autoregressive process with symmetric stable distributed innovations with $\alpha = 0.5$ and $\alpha = 2.0$. 
When estimating the structural parameters, the Monte Carlo experiment indicated that the $L_1$-norm estimator was preferable in non-normal distribution cases. However, when the residuals were serially dependent, the $L_1$-norm estimator looses its robustness properties. For a very heavy-tailed residual distribution and highly dependent residuals, the $L_1$-norm and the $L_2$-norm estimators perform, roughly, equally bad. Frequencies of best (closest to the true parameter) results for structural parameter estimates are reported in Table 2.4.

The usual Durbin-Watson test was applied to detect dependence in the residuals. It was observed that the test seemed not to be very affected by fat-tailed residual distributions. Further, the results do not indicate more difficulties in detecting serially dependent residuals, when the estimation methods are based on $L_1$-norm.

The generalized least squares and the $L_1$-norm estimators are compared in a study by Ashar and Wallace (1963). A first order autoregressive residual structure was assigned with a known value, 0.5, of the autoregressive parameter. The disturbance term in the autoregression process was independent and normally distributed with zero mean and unit variance.

The results indicate that the efficiency of the $L_1$-norm estimator is only about 20% of that of the generalized least squares. These results can, however, not be used as arguments against $L_1$-norm methods in the context of serially dependent residuals. The results are rather expected, since the residual distribution is not fat-tailed (least squares methods are preferable when the residuals are normally distributed) and the generalized least squares estimator makes use of the known dependence structure in the residuals, while the considered $L_1$-norm estimator does not.
Para-Estimator & $\alpha$
\hline
\text{Intercept} & 2.0 & 1.5 & 1.0 & 0.5 \\
L_1 & 76 & 122 & 174 & 199 \\
L_2 & 124 & 78 & 26 & 1 \\
slope & 84 & 122 & 178 & 198 \\
L_1 & 116 & 78 & 22 & 2 \\
\hline
\text{Intercept} & 2.0 & 1.5 & 1.0 & 0.5 \\
L_1 & 72 & 132 & 162 & 190 \\
L_2 & 128 & 68 & 38 & 10 \\
slope & 85 & 139 & 165 & 189 \\
L_1 & 115 & 61 & 35 & 11 \\
\hline
\text{Intercept} & 2.0 & 1.5 & 1.0 & 0.5 \\
L_1 & 91 & 96 & 105 & 127 \\
L_2 & 109 & 104 & 95 & 73 \\
slope & 86 & 97 & 101 & 121 \\
L_1 & 114 & 103 & 99 & 79 \\
\hline

a) Serially independent residuals

b) A first order autoregressive process in the residuals with parameter $\varphi = 0.5$.

c) A first order autoregressive process in the residuals with parameter $\varphi = 0.9$.

Table 2.4 Frequencies of best (closest to the true parameter) results for structural parameter estimates in Nyquist (1979).
2.6.5 Conclusions

We have now considered some Monte Carlo studies on different residual distributions. Although the conclusions from a single simulation study may be dubious, we have to note the almost unanimous conclusion from all studies: when the residual distribution tends to be fat-tailed, the $L_1$-norm estimators seem to be superior to the $L_2$-norm estimators. This general conclusion may be taken as an indication that it will be of great importance to develop the analytical properties of the $L_1$-norm based estimators.

2.7 Conclusions

In this chapter we have discussed some methods for estimating unknown parameters in models. The first estimating methods were based on different types of averaging. Methods based on the Principle of Arithmetic Mean, Cotes' extension of the Principle of Arithmetic Mean and the Method of Averages are examples of that. The idea of minimizing a certain function of observed residuals was introduced by Boscovich in 1757. The function Boscovich proposed was the sum of absolute of observed residuals. Later on, in 1820, Gauss argued that the choice of function is rather arbitrary. If the function used is the sum of squares of observed residuals, considerable simplifications in computations were made. Therefore, Gauss proposed the use of that function. The corresponding estimator is the wellknown and often applied least squares estimator, or, as we will call it, $L_2$-norm estimator. Still another estimator is defined by minimizing the greatest residual in absolute. That estimator, which was proposed by Chebychev in 1857, is called $L_\infty$-norm estimator.

Mostly this chapter deals with some earlier results concerning $L_1$-, $L_2$- and $L_\infty$-norm estimators of parameters in linear regression models.
Thus, the linear regression model was defined in Section 2.2. For the simple location parameter model, variances in the asymptotic distributions of the $L_1$, $L_2$, and $L_\infty$-norm estimators are known. In that case it is seen that the $L_1$-norm estimator is most effective for fat-tailed residual distributions, while the $L_\infty$-norm estimator is most effective for residual distributions with sharply defined extremes. The $L_2$-norm estimator is optimal for the normal distribution and performs well for distributions with reasonably light tails. On more heuristic grounds it is argued that the same results hold true even for more general regression models.

The notion of fat-tailed distributions was defined as a probability distribution whose variance does not exist. Therefore, all non-normal stable distributions are fat-tailed.

The class of stable distributions is important for two reasons. Firstly, since all but one member of that class are fat-tailed, the $L_1$-norm estimator is expected to be preferable when residuals follow a non-normal stable distribution. Secondly, properties of stable distributions and their domains of attraction are of crucial importance when deducing asymptotic distributions of estimators. Therefore, stable distributions and their domains of attraction were discussed in Section 2.3.

In Section 2.4 a small sample of applied models with fat-tailed or contaminated distributed random variables was given. The examples indicate that fat-tailed distributed and contaminated distributed variables appear in a great number of applications. Therefore, the $L_1$-norm estimator is expected to be suitable in many applications.

Before the 1950's the computation of $L_1$- and $L_\infty$-norm estimates of general linear regression models were very tedious. It seems that the difficulties in computation has been one of the main reasons why the use of $L_1$- and $L_\infty$-norm estimators has been rare. In the 1950's however, discoveries of significant importance for the computation of $L_1$- and $L_\infty$-norm estimators were made. At that time it was observed that it is possible
to formulate the minimization problems as linear programs. After that several improvements of the algorithms have been made. A very simple computation experiment performed in Section 2.5 showed that the most efficient algorithms for $L_1$-norm estimation are comparable in computation time with algorithms for least squares estimation. Therefore, the computation of $L_1$- and $L_\infty$-norm estimates is not any longer an obstacle to their use as estimators of linear regression models.

Since fat-tailed distributed residuals appear in many applications and it is expected that the $L_1$-norm estimator performs well in those cases, several simulation experiments have been performed to study its properties. References to some of these experiments were given in Section 2.6. An almost unanimous conclusion from all experiments was that when the residual distribution tends to be fat-tailed, the $L_1$-norm estimator seems to be superior to the least squares estimator. Thus, the $L_1$-norm estimator seems to behave as we had expected. Therefore, an analytical treatment of the $L_1$-norm estimator is motivated.

In the next chapter we will define the class of $L_p$-norm estimators. The $L_1$, $L_2$, and $L_\infty$-norm estimators are members of that class. We will discuss questions concerning their existence, uniqueness and computation of estimates. We will also give some geometrical interpretations and discuss relations between the class of $L_p$-norm estimators and other classes of estimators. Chapter 4 is devoted to a discussion of the asymptotic distribution of $L_p$-norm estimators.
CHAPTER III. SOME BASIC PROPERTIES OF $L_p$-NORM ESTIMATORS

3.1 Definition of the $L_p$-norm estimators

Consider again the linear model defined in (2.5). With the condition (2.2) Boscovich and Edgeworth introduced a fairly new methodology in estimating parameters in stochastic models. The observed residuals are measured by some metric and this quantity is minimized by the estimated model. For the $L_1$-norm estimator, the observed residuals are measured by $\Sigma |\hat{e}_i|$. The $L_2$- and $L_\infty$-norm estimators also rely on minimizing a measure of observed residuals. Thus, the observed residuals are measured by $\Sigma \hat{e}_i^2$ and $\max |\hat{e}_i|$ in the $L_2$- and $L_\infty$-norm, respectively.

Any metric can, of course, be used in measuring observed residuals. Therefore, each metric will give rise to an estimator. Since any norm can be deduced from a metric, each norm will also give rise to an estimator.

One class of norms is the class of $L_p$-norms. Thus, to each $L_p$-norm it corresponds an estimator. The corresponding estimator will be called $L_p$-norm estimator.

An $L_p$-norm estimate of the parameter vector $\beta$ in (2.5) is therefore a vector which minimizes

$$
\sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{m} x_{ij} b_j \right|^p \quad \text{for } 1 \leq p < \infty
$$

(3.1)

$$
\max_{i} \left| y_i - \sum_{j=1}^{m} x_{ij} b_j \right| \quad \text{for } p = \infty
$$

To each vector $\hat{\beta}$ minimizing (3.1), there corresponds a vector $\hat{y} = X\hat{\beta}$.

All vectors $\hat{y}$ defined in that manner, constitute a set, notated by $P_G(y)$.

Based on trivially algebraic arguments, it follows that if $X$ is of full rank, there corresponds a uniquely determined $\hat{\beta}$ to each vector
\[ \hat{y} \in \mathcal{P}_G(y) \] and that \( \hat{\beta} \) minimizes (3.1). Instead of studying properties of the vectors \( \hat{\beta} \) that minimize (3.1), we may study the properties of the vectors \( \hat{y} \in \mathcal{P}_G(y) \). Therefore, we will always assume that \( X \) is of full rank.

The columns in the matrix \( X, X_1, X_2, \ldots, X_m \) and the vectors \( y \) and \( \hat{y} \) are all vectors in an \( n \)-dimensional vector space. We denote that vector space by \( L_p \). If \( z \) is a vector in \( L_p \), we define the norm of \( z \), \( \| z \| \), by

\[
\| z \| = \begin{cases} 
\left( \sum_{i=1}^{n} |z_i|^p \right)^{1/p} & \text{for } 1 \leq p < \infty \\
\max_{i} |z_i| & \text{for } p = \infty.
\end{cases}
\]

Let \( G \) be the set of all vectors which are a linear combination of the vectors \( X_1, X_2, \ldots, X_m \), i.e.

\[ G = \{ g \in L_p | g = Xr; \quad r \in \mathbb{R}^m \}. \]

Then \( G \) is the linear subspace of \( L_p \), spanned by \( X_1, X_2, \ldots, X_m \).

For a more general setting of the problem, let \( G \) be a linear subspace and \( y \) a vector in a Banach space \( E \). In our notation, the vectors \( \hat{y} \in \mathcal{P}_G(y) \) are precisely those vectors which satisfy

\[
\min_{g \in G} \| y - g \| = \| y - \hat{y} \|
\]

and a formal definition of the set \( \mathcal{P}_G(y) \) is

\[ \mathcal{P}_G(y) = \{ \hat{y} \in G | \| y - \hat{y} \| = \inf_{g \in G} \| y - g \| \} \]

Let \( \overline{G} \) be the closure of \( G \). Since

\[ \mathcal{P}_G(y) = \begin{cases} 
\{ y \} & \text{if } y \in G \\
\emptyset & \text{if } y \in \overline{G} \setminus G
\end{cases} \]

for any linear subspace \( G \), we can restrict ourselves to consider the case \( y \in E \setminus \overline{G} \). To exclude trivial cases when the set \( E \setminus \overline{G} \) is empty, we will always assume that \( \overline{G} \neq E \). Note that for \( E = L_p \), \( G \) is finite dimensional and therefore \( \overline{G} = G \).
By a linear functional \( f \) defined on a vector space \( E \), we mean a linear mapping of \( E \) into the set of real numbers \( \mathbb{R} \); \( f: E \to \mathbb{R} \). Suppose \( f_1, f_2 \) and \( f \) are bounded linear functionals defined on a normed linear space \( E \). By defining
\[
(f_1 + f_2)(z) = f_1(z) + f_2(z)
\]
\[
(af)(z) = af(z)
\]
\( a \in \mathbb{R} \)
it is easy to verify that the class of such linear functionals is a linear vector space. By the conjugate space \( E^* \) of a linear vector space \( E \) we mean the vector space of all bounded linear functionals defined on \( E \) under the norm
\[
\| f \| = \sup_{z \neq 0} \frac{f(z)}{\| z \|}, \quad f \in E^*
\]
From now on, \( E^* \) denotes the conjugate space of the linear vector space \( E \). Similarly, \( L_p^* \) denotes the conjugate space of \( L_p \).

It is immediately seen that if \( f \in L_p^* \) and \( f(\delta^i) = k_i \), where \( \delta^i \) is the vector defined by
\[
\delta^i_j = \begin{cases} 
1 & \text{if } i=j \\
0 & \text{if } i \neq j 
\end{cases}
\]
we have that
\[
f(z) = \sum_{i=1}^{n} k_i z_i 
\]
Equation (3.2)
We also know (see e.g. Wloka, 1971, p. 76 and p. 80) that \( L_p \) is an \( n \)-dimensional Banach space with norm
\[
\| f \| = \begin{cases} 
\max_i |k_i| & \text{for } p = 1 \\
\left( \sum_{i=1}^{n} |k_i|^q \right)^{1/q} & \text{for } 1 < p < \infty \text{ and } \frac{1}{p} + \frac{1}{q} = 1 \\
\sum_{i=1}^{n} |k_i| & \text{for } p = \infty 
\end{cases}
\]
Equation (3.3)
There are now some problems we will examine. The first problem is whether there exists any vectors \( y \) in \( \mathbb{R}_g^*(y) \), or if the set \( \mathbb{R}_g^*(y) \) is empty. In
Section 3.2 it is shown, that when the subspace $G$ is finite dimensional, the set $P^G_C(y)$ is non-empty. In particular, this implies that there always exists an $L_p$-norm estimate. When we know that $P^G_C(y)$ is non-empty, it is meaningful to describe the vector(-s) in $P^G_C(y)$. Such a description is given in Section 3.3. The next problem to examine is whether there exist more than one vector in $P^G_C(y)$. In Section 3.4 it is proved that for the spaces $L_p$, $1 < p < \infty$, there exists exactly one vector in $P^G_C(y)$, i.e. the $L_p$-norm estimate is unique. The situation is more complicated in the space $L_1$, i.e., in some cases there may be more than one element in $P^G_C(y)$. The uniqueness problem in $L_1$-norm estimation is discussed in Section 3.5. In Section 3.6 the formal results of Section 3.2-2.5 are given geometric interpretations.

For the Hilbert space $L_2$ the notion of orthogonality has a specific meaning. In Section 3.7 this notion is generalized to suit our discussion of operations in Banach spaces.

In Section 3.8 it is shown that the $L_p$-norm estimators belong to the class of maximum likelihood type (M-) estimators, as defined in Huber (1972 and 1977). It is also known that the $L_1^-$, $L_2^-$ and $L_\infty^-$-norm estimators may be considered as combinations of order statistics (L-estimators).

We know that only for $p=2$ the $L_p^-$-norm estimator is linear. This implies that it is only the $L_2^-$-norm estimator that may be BLUE (i.e. best linear unbiased estimator). As we know, there exists a case when the $L_2^-$-norm estimator is BLUE, viz. when the residuals are normally distributed. On the other hand, the $L_p^-$-norm estimator may be equivalent to the maximum likelihood estimator for any $p$. This is exemplified in Section 3.9.

The last problem we will examine in this chapter concerns the computation of $L_p^-$-norm estimators. Thus, in Section 3.10 an algorithm for computation of $L_p^-$-norm estimates is presented. The chapter ends in Section 3.11 with some concluding remarks.
3.2 The existence of $L_p$-norm estimates

In this section we will prove the existence of an $L_p$-norm estimate, i.e. the set $H_G(y)$ is non-empty.

**THEOREM 3.1 (EXISTENCE THEOREM)** Let $G$ be a finite-dimensional linear subspace of $E$ and let $y \in E \setminus G$. Then the set $H_G(y)$ is non-empty.

**PROOF** Let $g_0$ be an arbitrary vector in $G$ and consider the set

$$H = \{g \in G | \|g - y\| \leq \|g_0 - y\|\}$$

Since $H$ is a closed and bounded set, it follows from the Heine-Borel theorem that $H$ is a compact set.

From the definition of an infimum, it follows that we may find a sequence of vectors $\{h_k\}$ in $H$, such that

$$\lim_{k \to \infty} \|y - h_k\| = \inf_{g \in G} \|y - g\|$$

Since $H$ is a compact set, we may assume that the sequence $\{h_k\}$ converges to a point $h \in H$. To complete the proof, we have only to show that $h \in H_G(y)$.

By the triangle inequality we have

$$\|y - h\| \leq \|y - h_k\| + \|h_k - h\|$$

The left member of this inequality is independent of $k$, and the right member approaches $\inf_{g \in G} \|y - g\|$ as $k \to \infty$. Therefore $\|y - h\| \leq \inf_{g \in G} \|y - g\|$. Since $h \in H$, $\|y - h\| \geq \inf_{g \in G} \|y - g\|$, and hence $\|y - h\| = \inf_{g \in G} \|y - g\|$, i.e. $h \in H_G(y)$.

Q.E.D.

The proof of this Theorem may be found in any standard text in approximation theory, e.g. Rivlin (1969) and Cheney (1966). Note that the Theorem is valid for infinite dimensional vector spaces $E$, but the assumption of a finite dimension of the subspace $G$ is essential. The Existence theorem implies for $E = L_p$ that at least one vector $\hat{y} \in H_G(y)$, i.e. there exists at least one $L_p$-norm estimate to each data set.
3.3 Characterizations of \( L_p \)-norm estimates

The theorems in this section relate the \( L_p \)-norm estimate and the existence of a linear functional in \( L_p^* \) with certain characteristics.

**THEOREM 3.2 (CHARACTERIZATION THEOREM)** Let \( E \) be a normed linear space, \( G \) a linear subspace of \( E \), \( y \in E \setminus \overline{G} \) and \( \hat{y} \in G \). Then, \( \hat{y} \in \mathcal{H}_G(y) \) if and only if there exists a linear functional \( f \in E^* \) such that

\[
\begin{align*}
\| f \| &= 1 \\
f(g) &= 0 \quad \forall g \in G \\
f(y - \hat{y}) &= \| y - \hat{y} \|
\end{align*}
\]

**PROOF** Assume that \( \hat{y} \in \mathcal{H}_G(y) \). Since \( y \in E \setminus \overline{G} \) we have that \( \| y - \hat{y} \| > 0 \). By a well-known corollary of Hahn-Banach's theorem (see e.g. Bachman and Narici, 1966, Theorem 12.3) there exists a linear functional \( f \in E^* \) satisfying (3.4), (3.5) and (3.6). On the other hand, assume there exists a linear functional \( f \in E^* \) satisfying (3.4), (3.5) and (3.6). For any \( g \in G \), we then have that

\[
\| y - \hat{y} \| = f(y - \hat{y}) = |f(y-g)| \leq \| f \| \| y - g \| = \| y - g \|
\]

that is \( \hat{y} \in \mathcal{H}_G(y) \).

Q.E.D.

**COROLLARY 3.1** Let \( E \) be a normed linear space, \( G \) a linear subspace of \( E \), \( y \in E \setminus \overline{G} \) and \( M \subseteq \overline{G} \). Then we have that \( M \subseteq \mathcal{H}_G(y) \) if and only if there exists a linear functional \( f \in E^* \) satisfying (3.4), (3.5) and

\[
f(y - \hat{y}) = \| y - \hat{y} \| \quad \forall \hat{y} \in M
\]

**PROOF** The Corollary follows immediately from the Characterisation Theorem and from the equality \( \| y - \hat{y}_1 \| = \| y - \hat{y}_2 \| \) which trivially is satisfied for all \( \hat{y}_1, \hat{y}_2 \in \mathcal{H}_G(y) \).

q.e.d.

The first part of the Characterization Theorem follows directly from Hahn-Banach's Theorem. From Hahn-Banach's Theorem it also follows (see...
e.g. Banach-Narici, 1966 Theorem 12.2) that for any vector \( y' \) in a normed linear space \( E \), there exists a linear functional \( f \in E^* \) such that
\[
\|f\| = 1 \quad \text{and} \quad f(y') = \|y'\| \quad (3.7)
\]
If \( y' \in E \) is a vector such that there exists exactly one linear functional \( f \in E^* \) with the property (3.7), \( y' \) will be called normal and we denote the corresponding linear functional by \( f_{y'} \). When \( y - \hat{y} \) is a normal vector, we then have the following Corollary of the Characterization Theorem.

**COROLLARY 3.2** Let \( E \) be a normed linear space, \( G \) a linear subspace of \( E \), \( y \in E \setminus \overline{G} \) and \( \hat{y} \in G \). Assume that \( y - \hat{y} \) is normal. Then \( \hat{y} \in \mathcal{H}_G(y) \) if and only if
\[
f_{y - \hat{y}}(g) = 0 \quad \forall g \in G \quad (3.8)
\]

**PROOF** Assume that \( \hat{y} \in \mathcal{H}_G(y) \). By the Characterization Theorem there exists a linear functional \( f \in E^* \) satisfying (3.4), (3.5) and (3.6). Since \( y - \hat{y} \) is normal, we have by (3.4) and (3.6) that \( f = f_{y - \hat{y}} \) and therefore (3.8) follows from (3.5). On the other hand, assume there exists a linear functional \( f_{y - \hat{y}} \) satisfying (3.8). Then \( f_{y - \hat{y}} \) must satisfy (3.4), (3.5) and (3.6) and by the Characterization Theorem \( \hat{y} \in \mathcal{H}_G(y) \).

q.e.d.

A normed linear space \( E \) is said to be strictly convex, if \( \|y + z\| = \|y\| + \|z\| \), \( y, z \in E \) implies the existence of some vector \( u \in E \) such that \( y = \alpha u \) and \( z = \beta u \) where \( \alpha, \beta \in \mathbb{R}^+ \). The next two propositions relate the notions of strict convexity and normality for \( L_p \)-spaces.

**PROPOSITION 3.1** For \( 1 < p < \infty \), the \( L_p \)-spaces are strictly convex.

**PROOF** See e.g. Köthe (1960 § 25.2).

**PROPOSITION 3.2** All vectors in a strictly convex space are normal.

**PROOF** See e.g. Köthe (1960 § 26.3).

From these propositions it follows trivially that all vectors in the \( L_p \)-spaces, \( 1 < p < \infty \), are normal. For \( L_p \)-norm estimators, \( 1 < p < \infty \), we then have the following characterization.
THEOREM 3.3 (CHARACTERIZATION IN $L_p$, $1 < p < \infty$) Let $E = L_p$, $1 < p < \infty$ and let $G$ be a linear subspace of $E$. If $y \in E \setminus G$ and $\hat{y} \in G$ we have that $\hat{y} \in \text{IP}_G(y)$ if and only if

$$\sum_{i=1}^{n} g_i |y_i - \hat{y}_i|^{p-1} \text{sign}(y_i - \hat{y}_i) = 0 \quad \forall g \in G$$

If the partial derivatives with respect to $b_j$, $j=1,...,m$ of (3.1) is set to zero, a system of $m$ equations is obtained. The equations are after dividing by $p$

$$\sum_{i=1}^{n} x_{ij} |y_i - \hat{y}_i|^{p-1} \text{sign}(y_i - \hat{y}_i) = 0, \quad j=1,...,m \quad (3.9)$$

Since the vectors $X_1,...,X_m$ span $G$, all vectors $g \in G$ may be written as a linear combination of $X_1,...,X_m$. Therefore the first part in the last Theorem will follow by multiplying the equations in the system (3.9) by real constants and adding the equations. The second part is, however, more difficult to prove, and relies on a strict convexity property of the function (3.1). For the proof of the Theorem, we will therefore use the powerful Characterization Theorem. The proof will then be simple as well as elegant.

PROOF For the normal vector $y-\hat{y}$ we have that

$$f_{y-\hat{y}}(z) = \sum_{i=1}^{n} z_i \frac{|y_i - \hat{y}_i|^{p-1}}{\|y-\hat{y}\|^{p-1}} \text{sign}(y_i - \hat{y}_i), \quad z \in L_p$$

since

$$\|f_{y-\hat{y}}\| = \left[ \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|^{(p-1)}}{\|y-\hat{y}\|^{p-1}} \text{sign}(y_i - \hat{y}_i) \right]^{1/q} = \left[ \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|^{(p-1)q}}{\|y-\hat{y}\|^{(p-1)q}} \right]^{1/q} = \left[ \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|^p}{\|y-\hat{y}\|^p} \right]^{1/q} = \frac{\|y-\hat{y}\|^p}{\|y-\hat{y}\|^p} = 1.$$
where \(
\frac{1}{p} + \frac{1}{q} = 1 \Leftrightarrow p + q = pq \Leftrightarrow pq - q = p
\)

and

\[
f_{\hat{y}}(y - \hat{y}) = \sum_{i=1}^{n} \left( y_i - \hat{y}_i \right) \frac{|y_i - \hat{y}_i|^{p-1}}{\|y - \hat{y}\|^{p-1}} \text{sign}(y_i - \hat{y}_i) = \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|^p}{\|y - \hat{y}\|^{p-1}} = \frac{\|y - \hat{y}\|^p}{\|y - \hat{y}\|^{p-1}} = \|y - \hat{y}\|
\]

and the Theorem follows from Corollary 2.2.

Q.E.D.

In the \(L_1\)-space not all vectors are normal. It is therefore not possible to make a characterization as good as that in the case of \(L_p\)-space, \(1<p<\infty\). What we can do in the \(L_1\)-space is summarized in the next Theorem.

**THEOREM 3.4 (CHARACTERIZATION IN \(L_1\))** Let \(E = L_1\) and let \(G\) be a subspace of \(L_1\). If \(y \in E \setminus G\) and \(\hat{y} \in G\), then the following statements are equivalent.

1° \(\hat{y} \in \text{EP}_G(y)\).

2° There exists a linear functional \(f \in L_1^*\) as defined in (3.2), such that

\[
\|f\| = \max_{i} |k_i| = 1. \quad (3.10)
\]

\[
f(g) = \sum_{i=1}^{n} k_i g_i = 0, \quad \forall g \in G. \quad (3.11)
\]

\[
f(y - \hat{y}) = \sum_{i=1}^{n} k_i (y_i - \hat{y}_i) = \|y - \hat{y}\| = \sum_{i=1}^{n} |y_i - \hat{y}_i|. \quad (3.12)
\]

3° There exists a vector \(\alpha \in L_\infty\) which satisfies

\[
\|\alpha\| = \max_{i} |\alpha_i| = 1. \quad (3.13)
\]

\[
\sum_{i=1}^{n} \alpha_i g_i = 0, \quad \forall g \in G. \quad (3.14)
\]

\[
\alpha_i (y_i - \hat{y}_i) = |y_i - \hat{y}_i|, \quad i=1, \ldots, n. \quad (3.15)
\]
We have that
\[ \sum_{i=1}^{n} g_i \text{ sign}(y_i - \hat{y}_i) \leq \sum_{i \in I} |g_i|, \quad \forall g \in G, \tag{3.16} \]
where \( I \) is the set of exactly those indices \( i \) for which \( y_i = \hat{y}_i \).

**PROOF** The equivalence of 1° and 2° follows directly from the Characterization Theorem and from the representation of linear functionals in \( \mathbb{L}^* \) in (3.2) and (3.3).

Since \( \mathbb{L}^* \) and \( \mathbb{L}_\infty \) are isomorphic, statement 2° in the Theorem is equivalent to the existence of a vector \( \alpha \in \mathbb{L}_\infty \) satisfying (3.13), (3.14) and
\[ \sum_{i=1}^{n} \alpha_i (y_i - \hat{y}_i) = \| y - \hat{y} \| = \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.17} \]
We now show that (3.13) and (3.17) are equivalent to (3.13) and (3.15).

It follows immediately that (3.13) and (3.15) implicate (3.13) and (3.17).

Assume now that there is a vector \( \alpha \in \mathbb{L}_\infty \) satisfying (3.13) and (3.17) and that \( \alpha_i (y_i - \hat{y}_i) \neq |y_i - \hat{y}_i| \) for some index \( i \) belonging to an index set \( I_0 \). By (3.13) we have
\[ \alpha_i (y_i - \hat{y}_i) < |y_i - \hat{y}_i|, \quad \text{for } i \in I_0. \]
Otherwise,
\[ |y_i - \hat{y}_i| \leq \alpha_i (y_i - \hat{y}_i) \leq |\alpha_i (y_i - \hat{y}_i)| \leq |y_i - \hat{y}_i|, \quad \text{for } i \in I_0, \]
that is,
\[ \alpha_i (y_i - \hat{y}_i) = |y_i - \hat{y}_i|, \quad \text{for } i \in I_0, \]
which is incompatible with our assumption. Then we have
\[ \sum_{i=1}^{n} \alpha_i (y_i - \hat{y}_i) < \sum_{i=1}^{n} |y_i - \hat{y}_i|, \]
which is incompatible with (3.17), that is (3.15) follows. We now have established the equivalence \( 2° \Leftrightarrow 3° \).
It remains to show the equivalence $1^0 \Rightarrow 4^0$. Assuming $1^0$ to hold, then we know by the equivalence $1^0 \Rightarrow 3^0$ that there exists a vector $\alpha \in C_{\infty}$ which satisfies (3.13), (3.14) and (3.15). Furthermore, (3.15) implies that

$$\alpha_i = \text{sign}(y_i - \hat{y}_i), \quad \text{for } y_i \neq \hat{y}_i.$$ 

Then by (3.14) we have

$$\sum_{i=1}^{n} g_i \cdot \text{sign}(y_i - \hat{y}_i) + \sum_{i \in I} \alpha_i g_i = 0, \quad \forall g \in G.$$ 

By (3.13) we have $|\alpha_i| \leq 1$ and it immediately follows due to the triangle inequality that

$$|\sum_{i=1}^{n} g_i \cdot \text{sign}(y_i - \hat{y}_i)| = |\sum_{i \in I} \alpha_i g_i| \leq |\sum_{i \in I} g_i| \leq \sum_{i \in I} |g_i|.$$ 

Conversely, assume that $4^0$ holds. Then, for any $g \in G$,

$$\sum_{i=1}^{n} |y_i - \hat{y}_i| = \sum_{i=1}^{n} (y_i - \hat{y}_i) \cdot \text{sign}(y_i - \hat{y}_i) =$$

$$= \sum_{i=1}^{n} (y_i - g_i) \cdot \text{sign}(y_i - \hat{y}_i) + \sum_{i=1}^{n} (g_i - \hat{y}_i) \cdot \text{sign}(y_i - \hat{y}_i).$$

Since $g - \hat{y} \in G$, (3.16) implies

$$|\sum_{i=1}^{n} (g_i - \hat{y}_i) \cdot \text{sign}(y_i - \hat{y}_i)| \leq \sum_{i \in I^c} |g_i - \hat{y}_i|,$$

and by the triangle inequality

$$|\sum_{i=1}^{n} (y_i - g_i) \cdot \text{sign}(y_i - \hat{y}_i)| \leq \sum_{i \in I^c} |y_i - g_i|,$$

where $I^c$ is the complement of the index set $I$. This will imply

$$\sum_{i=1}^{n} |y_i - \hat{y}_i| \leq \sum_{i \in I^c} |y_i - g_i| + \sum_{i \in I} |g_i - \hat{y}_i|. $$

By definition, $y_i = \hat{y}_i$ in the index set $I$, which implies

$$\sum_{i \in I} |g_i - \hat{y}_i| = \sum_{i \in I} |y_i - g_i|.$$
and

$$\|y-\hat{y}\| = \sum_{i=1}^{n} |y_i - \hat{y}_i| \leq \sum_{i=1}^{n} |y_i - g_i| = \|y-g\|$$

Since $g$ is an arbitrary vector in $G$, we have that $\hat{y} \in \mathbb{H}_G(y)$, and 1° follows. All the equivalences in the Theorem are now proved.

The equivalence $1° \iff 4°$ in this Theorem was first proved in Kripke and Rivlin (1965).

3.4 Uniqueness of $L_p$-norm estimates, $1 < p < \infty$

In this section we will prove the uniqueness of $L_p$-norm estimates, when $1 < p < \infty$. As we know that there exists at least one $L_p$-norm estimate, this is the same as to prove that there is at most one $L_p$-norm estimate, or the set $\mathbb{H}_G(y)$ contains at most one element. In the proof we will make use of the notion of strictly convex spaces.

THEOREM 3.5 (UNIQUENESS IN $L_p$, $1 < p < \infty$) Let $E = L_p$, $1 < p < \infty$ and let $G$ be a linear subspace of $E$. For $y \in E \setminus G$ there exists exactly one element $\hat{y}$ in $\mathbb{H}_G(y)$.

PROOF Let $\hat{y}, \hat{y}' \in \mathbb{H}_G(y)$. If $\delta = \|y-\hat{y}\| = \|y-\hat{y}'\|$ we have that

$$\delta \leq \|y- \frac{1}{2} (\hat{y}+\hat{y}')\| \leq \frac{1}{2} \|y-\hat{y}\| + \frac{1}{2} \|y-\hat{y}'\| = \delta$$

and so, for some $z$, $y-\hat{y} = \alpha z$ and $y-\hat{y}' = \beta z$. But since $\|y-\hat{y}\| = \|y-\hat{y}'\|$, $\alpha = \beta$.

Q.E.D.
3.5 Uniqueness of $L_1$-norm estimates

3.5.1 Some introductory comments on the uniqueness problem

Quite early it was recognized that the $L_1$-norm estimator not always lead to unique estimates. An early paper on that subject is Turner (1887). For example, the $L_1$-norm estimate of $\hat{y}$ in the simple model $y = \beta + \epsilon$ is equivalent to the sample median. But when the number of observations is even, the sample median may not be uniquely determined. In this section we will, to some extent, discuss the uniqueness problem of $L_1$-estimates in linear models. Thus, we first give another two examples of non-uniqueness and a necessary and sufficient condition for unique $L_1$-norm estimation.

**EXAMPLE 3.1** Consider a simple linear model without intercept. With the observations

\[
\begin{array}{cccc}
  x & 2 & 2 & 4 & 4 \\
  y & 1.5 & 2 & 2 & 4 \\
\end{array}
\]

all $b$ in the interval $[0.75, 1]$ are $L_1$-norm estimates of the regression parameter. See figure 3.1.

**EXAMPLE 3.2** Consider a simple linear regression model with intercept and the data

\[
\begin{array}{cccc}
  x & 1 & 2 & 3 & 4 \\
  y & 1 & 1.5 & 2.5 & 4 \\
\end{array}
\]

Then, all regression lines in the shadowed area in Figure 3.2 are $L_1$-norm estimates.

We will now give a necessary and sufficient condition for unique estimation in $L_1$-norm. The necessity part of the proof was first pointed out by Kripke and Rivlin (1965) and the sufficiency part has been proved in Rivlin (1969).
Figure 3.1 Figure for Example 3.1.

Figure 3.2 Figure for Example 3.2.
THEOREM 3.6 (UNIQUENESS IN $L_1$) Let $G$ be a linear subspace of $L_1$ and let $y \in L_1 \setminus G$. Then $\hat{y}$ is the only element in $\mathcal{H}_G(y)$ if and only if
\[
\sum_{i=1}^{n} |g_{i} \text{sign}(y_i - \hat{y}_i)| < \sum_{i \in I} |g_i|, \quad \forall g \in G \setminus \{0\},
\]
where $I$ is the set of all indices $i$ such that $y_i = \hat{y}_i$.

PROOF The necessity part of the proof is a simple modification of the proof of the implication $4^0 \Rightarrow 1^0$ in Theorem 3.4.

Assume now that
\[
\sum_{i=1}^{n} |g_{oi} \text{sign}(y_i - \hat{y}_i)| = \sum_{i \in I} |g_{oi}|, \quad \text{for some } g_{oi} \in G \setminus \{0\},
\]
and let
\[
\hat{y}_t = \hat{y} + t g_0, \quad t \in \mathbb{R},
\]
Then $\hat{y}_t \in G$, and we would like to show that $\hat{y}_t \in \mathcal{H}_G(y)$ for some $t > 0$.

Since $|y_i - \hat{y}_i| > 0$ for all $i \in I^c$, where $I^c$ is the complement to $I$, there must exist a real number $\epsilon > 0$ such that
\[
\text{sign}(y_i - \hat{y}_i) = \text{sign}(y_i - \hat{y}_{ti}), \quad \text{for } i \in I^c \text{ and } 0 < |t| < \epsilon.
\]
For $|t| < \epsilon$, we then have
\[
\|y - \hat{y}_t\| - \|y - \hat{y}\| = \sum_{i=1}^{n} |y_i - \hat{y}_{ti}| - \sum_{i=1}^{n} |y_i - \hat{y}_i| =
\]
\[
= \sum_{i=1}^{n} (y_i - \hat{y}_{ti}) \text{sign}(y_i - \hat{y}_i) - \sum_{i=1}^{n} (y_i - \hat{y}_i) \text{sign}(y_i - \hat{y}_i) =
\]
\[
= \sum_{i=1}^{n} [(y_i - \hat{y}_{ti}) - (y_i - \hat{y}_i)] \text{sign}(y_i - \hat{y}_i) + \sum_{i \in I} (y_i - \hat{y}_{ti}) \text{sign}(y_i - \hat{y}_{ti}).
\]
But
\[
\sum_{i=1}^{n} [(y_i - \hat{y}_{ti}) - (y_i - \hat{y}_i)] \text{sign}(y_i - \hat{y}_i) =
\]
\[
= \sum_{i=1}^{n} (\hat{y}_i - \hat{y}_{ti}) \text{sign}(y_i - \hat{y}_i) = -t \sum_{i=1}^{n} g_{oi} \text{sign}(y_i - \hat{y}_i)
\]
Thus, for a sufficiently small $t$, $0 < |t| < \epsilon$, with a suitable sign, it follows that
\[ \| y - \hat{y}_t \| - \| y - \hat{y} \| = 0, \]
that is $\hat{y}_t \in \mathcal{P}_G(y)$ and $\hat{y}$ is not the only element in $\mathcal{P}_G(y)$. Q.E.D.

Let $G$ be the linear subspace spanned by the regressor variables $X_1, \ldots, X_m$. If there always exist unique estimates, irrespectively of $y$, we say that the model is $G$-unique. In the next subsection, necessary and sufficient conditions for $G$-uniqueness are given. When a model is not $G$-unique, the uniqueness question depends on $y$. This is discussed in Subsection 3.5.3. In Subsection 3.5.4 the theories are applied to the introductory examples.

If $y$ is random in a case of non-$G$-uniqueness, the question of unique $L_1$-norm estimates is, due to the results in Subsection 3.5.3, of stochastic nature. If the difference between two arbitrary $L_1$-norm estimates is expected to be small, the non-uniqueness should not cause any practical problem of selecting one as "the estimate". For some sequences of models, it happens that we are expecting unique estimates when the number of observations tends to infinity. We call that weak asymptotic uniqueness. In cases of weak asymptotic uniqueness, the non-uniqueness problem is less important when the number of observations is sufficiently large. A sufficient condition for weak asymptotic uniqueness is given in Subsection 3.5.5.

The section ends with some concluding remarks in Subsection 3.5.6.
3.5.2 Models that always have unique $L_1$-norm estimates

In this subsection we will give a necessary and sufficient condition on the linear subspace $G$ in $L_1$ that the $L_1$-norm estimates always will be unique. We refer to this case as the $G$-uniqueness case. Since the vectors $X_1, X_2, \ldots, X_m$ span $G$, this condition can easily be transformed to a condition on these vectors. In the proof of the theorem the following Lemma is needed.

**Lemma 3.1** Let $G$ be a linear subspace of $L_1$. Then the $L_1$-norm estimate $\hat{y}$ in $G$ of any vector $y \in L_1 \setminus G$ is unique if and only if there is no linear functional $f \in L_1^*$ defined by (3.2) and no vectors $y_1, y_2 \in L_1$, $y_1 - y_2 \in G \setminus \{0\}$, such that

$$||f|| = \max_i |k_i| = 1,$$  

(3.18)

$$f(g) = \sum_{i=1}^{n} k_i g_i = 0, \quad \forall g \in G,$$  

(3.19)

$$f(y_j) = \sum_{i=1}^{n} k_i y_{ji} = \sum_{i=1}^{n} |y_{ji}| = ||y_j||, \quad j=1,2.$$  

(3.20)

**Proof** Assume that there exists a linear functional $f \in L_1^*$ and vectors $y_1, y_2 \in L_1$, $y_1 - y_2 \in G \setminus \{0\}$ such that (3.18), (3.19) and (3.20) are satisfied. We would like to show that the vector $y_1 \in L_1 \setminus G$, and that the $L_1$-norm estimate of $y_1$ in $G$ is not unique. Let

$$\hat{y} = y_1 - y_2.$$  

Then

$$f(y_1 - \hat{y}) = f(y_2) = ||y_2|| = ||y_1 - \hat{y}||,$$  

(3.21)

that is $y_1 \notin L_1 \setminus G$. Otherwise, we have, following (3.19) and (3.20), $0 = f(y_1) = ||y_1||$ and according to (3.19) and (3.21), $0 = f(y_1 - \hat{y}) = ||y_2||$, which is in conflict with the hypothesis that $y_1 - y_2 \neq 0$. By Theorem 3.4 and (3.18), (3.19) and (3.21) it now follows that $\hat{y}$ is an $L_1$-norm
estimate in $G$ of $y$. But the hypothesis in Theorem 3.4 is also satisfied by (3.18), (3.19) and (3.20) so that 0 is also an $L_1$-norm estimate in $G$ of $y$. Since $\hat{y} \neq 0$, the $L_1$-norm estimate of $y$ is not unique.

For the proof of the second part of the Lemma, assume that there is a vector $z \in L_1 \setminus G$ which has no unique $L_1$-norm estimate in $G$. Then there exist two vectors $z_1, z_2 \in G$ such that $z_1 \neq z_2$, and $\|z - z_1\| = \|z - z_2\| = \inf_{g \in G} \|z - g\|$. We would like to show that there exists a linear functional $f \in L_*^1$ and vectors $y_1, y_2 \in L_1$, $y_1 - y_2 \in G \setminus \{0\}$ such that (3.18), (3.19) and (3.20) are satisfied.

According to Corollary 3.1 there exists a linear functional $f \in L_*^1$ satisfying (3.18), (3.19) and

$$f(z - \hat{z}_j) = \|z - \hat{z}_j\|, \quad j = 1, 2.$$ 

Let

$$y_1 = z - \hat{z}_1 \quad \text{and} \quad y_2 = z - \hat{z}_2.$$ 

It follows that $y_1, y_2 \in L_1$, $y_1 - y_2 = \hat{z}_2 - \hat{z}_1 \in G \setminus \{0\}$ which will immediately imply (3.20). q.e.d.

We are now able to state and prove a necessary and sufficient condition on $G$ that the $L_1$-norm estimates are always unique.

**THEOREM 3.7 (G-UNIQUENESS THEOREM)** Let $G$ be a linear subspace of $L_1$. Then the $L_1$-norm estimate $\hat{y}$ in $G$ of any vector $y \in L_1 \setminus G$ is unique if and only if there is no vector $v \in L_\infty$ and no vector $g_0 \in G \setminus \{0\}$, such that

$$v_i = \pm 1, \quad \text{when} \quad g_{0i} \neq 0 \quad (3.22)$$

$$\sum_{i=1}^{n} v_i g_i = 0, \quad \forall g \in G, \quad (3.23)$$

$$\|v\| = 1. \quad (3.24)$$
**PROOF** Assume that there exist vectors \( v \in L_\infty \) and \( g_0 \in G \setminus \{0\} \) such that (3.22), (3.23) and (3.24) are satisfied. Define a linear functional \( f \in L_1^* \) by

\[
f(z) = \sum_{i=1}^{n} z_i v_i, \quad z \in L_1.
\]

From (3.22) and (3.23) follow (3.18) and (3.19). We then have to show that (3.20) also follows. Therefore, let \( y_0 \) be the vector in \( L_1 \in G \) defined by

\[
y_{oi} = |g_{oi}| v_i, \quad i=1,2,\ldots,n.
\]

Then we have

\[
y_{oi} v_i = |g_{oi}| |v_i| = |g_{oi}| \geq 0, \quad i=1,2,\ldots,n
\]

and

\[
(y_{oi} - g_{oi}) v_i = |g_{oi}| - g_{oi} v_i = |g_{oi}| \pm |g_{oi}|, \quad i=1,2,\ldots,n.
\]

Since

\[
|y_{oi}| = |g_{oi}| |v_i| = |g_{oi}|, \quad i=1,2,\ldots,n
\]

we have

\[
f(y_0) = \sum_{i=1}^{n} y_{oi} v_i = \sum_{i=1}^{n} |g_{oi}| = \sum_{i=1}^{n} |y_{oi}| = ||y_0||
\]

and

\[
f(y_0 - g_0) = \sum_{i=1}^{n} (y_{oi} - g_{oi}) v_i = \sum_{i=1}^{n} |y_{oi} - g_{oi}| = ||y_0 - g_0||
\]

and (3.20) follows with \( y_1 = y_0 \) and \( y_2 = y_0 - g_0 \). Due to Lemma 3.1, the \( L_1 \)-norm estimate of \( y_0 \) can not be unique.

Conversely, assume that there exists a vector \( y \in L_1 \setminus G \) and two vectors \( \hat{y}_1, \hat{y}_2 \in G, \hat{y}_1 \neq \hat{y}_2 \), which are \( L_1 \)-norm estimates of \( y \). By the isomorphy between \( L_1^* \) and \( L_\infty \), Corollary 3.1 will imply the existence of a vector \( v \in L_\infty \) satisfying (3.10), (3.11) and

\[
\sum_{i=1}^{n} (y_i - \hat{y}_i) v_i = \sum_{i=1}^{n} |y_i - \hat{y}_i|, \quad j=1,2.
\]
i.e. there exists a vector \( v \in L_\infty \) satisfying (3.23) and (3.24). We have to show the existence of a vector \( g_0 \in G \setminus \{0\} \) such that (3.22) also is satisfied. Let

\[
g_0 = \hat{y}_1 - \hat{y}_2.
\]

Since \( \hat{y}_1 \neq \hat{y}_2 \) we have that \( g_0 \in G \setminus \{0\} \). Due to the equivalence \( 2^0 \Leftrightarrow 3^0 \) in Theorem 3.4 we have

\[
v_i(y_i - \hat{y}_{1i}) = |y_i - \hat{y}_{1i}|
\]

\[
v_i(y_i - \hat{y}_{2i}) = |y_i - \hat{y}_{2i}|
\]

that is

\[
|v_i| = 1 \quad \text{if} \quad y_i \neq \hat{y}_{1i}
\]

\[
|v_i| = 1 \quad \text{if} \quad y_i \neq \hat{y}_{2i}
\]

But if \( g_{0i} \neq 0 \) we have that \( \hat{y}_{1i} \neq \hat{y}_{2i} \) and at least one of \( \hat{y}_{1i} \) and \( \hat{y}_{2i} \) is not equal to \( y_i \), thus \(|v_i| = 1\).

Q.E.D.

The G-uniqueness Theorem indicates that if the \( L_1 \)-norm estimate in a subspace \( G \) of a vector \( y \in L_1 \setminus G \) is not unique, there will exist vectors \( v \in L_\infty \) and \( g_0 \in G \setminus \{0\} \) satisfying (3.22), (3.23) and (3.24). Since the vectors \( X_1, X_2, \ldots, X_m \) span \( G \), each vector \( g \in G \) is of the form

\[
g = \sum_{i=1}^{m} \alpha_i X_i
\]

for some constants \( \alpha_j, j=1,2,\ldots,m \). (3.23) is then seen to be equivalent to

\[
\sum_{i=1}^{n} \sum_{j=1}^{m} v_i \alpha_j x_{ji} = 0,
\]

Hence (3.23) is equivalent to

\[
\sum_{j=1}^{m} \sum_{i=1}^{n} \alpha_j v_i x_{ji} = 0, \quad \forall \alpha_j \in \mathbb{R},
\]

and

\[
\sum_{i=1}^{n} v_i x_{ji} = 0, \quad \forall X_j, \quad j=1,2,\ldots,m
\]  \hspace{1cm} (3.25)
When \( m = 1 \) all vectors \( g \in G \setminus \{0\} \) are of the form \( \alpha X \), where \( \alpha \) is a non-zero constant. In particular, this means that \( g_i \neq 0 \), if and only if \( x_i \neq 0 \), i.e. \( |v_i| = 1 \), if \( x_i \neq 0 \). When \( x_i = 0 \) it makes no difference whether \( |v_i| = 1 \), or \( |v_i| \neq 1 \).

A necessary condition for non-unique \( L_1 \)-norm estimation, when \( m = 1 \), is then that there exists a summation \( \sum_{i=1}^{n} \pm x_i = 0 \). When no such summation exists, the \( L_1 \)-norm estimation is unique according to the \( G \)-uniqueness Theorem. Unfortunately, the case is more complicated when \( m > 1 \), as is seen in (3.22) and (3.25).

### 3.5.5 Models that may have non-unique \( L_1 \)-norm estimates

In this section we will, in some more detail, investigate the case of non-\( G \)-uniqueness. From Example 3.3 given below, it is seen that unique \( L_1 \)-norm estimation may occur even in cases of non-\( G \)-uniqueness.

**Example 3.3** Consider a simple linear regression model with intercept,

\[
y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \epsilon_i, \quad i=1,2,3,4,
\]

and the data

| \( x_0 \) | 1 | 1 | 1 | 1 |
|---|---|---|---|
| \( x_1 \) | 1 | 2 | 3 | 4 |
| \( y \)   | 1 | 1.5 | 2.5 | 3 |

For the vectors \( v \in L \) and \( g_o \in G \setminus \{0\} \) defined by

\[
v = 1 \quad -1 \quad -1 \quad 1
\]

\[
g_o = 1 \quad 1 \quad 1 \quad 1
\]

it is seen that

\[
v_i = \pm 1 \quad \text{when} \quad g_{o_i} \neq 0
\]

\[
\sum_{i=1}^{n} v_i g_i = 0 \quad \forall g \in G
\]

since for any \( g \in G \) there exist constants \( a \) and \( b \) such that \( g = a x_o + b x_1 \), and
\[ \sum_{i=1}^{4} v \cdot g_i = \sum_{i=1}^{4} v_i (a x_i + b x_i) = a \sum_{i=1}^{4} v_i x_i + b \sum_{i=1}^{4} v_i \cdot x_i = \]
\[ = a(1 \cdot 1 + (-1) \cdot 1 + (-1) \cdot 1 + 1 \cdot 1) + b(1 \cdot 1 + (-1) \cdot 2 + (-1) \cdot 3 + 1 \cdot 4) = 0 \]

and
\[ \| v \| = \max(|1|, |-1|, |-1|, |1|) = 1 \]

Thus, the condition in the \( G \)-uniqueness Theorem is not satisfied, but the \( L_1 \)-norm estimates of the intercept and the slope are unique and equals \( 1/3 \) and \( 2/3 \), respectively. See Figure 3.3.

The Non-uniqueness Theorem given below indicates that if the \( L_1 \)-norm estimate in a subspace \( G \) of a vector \( y \in L_1 \setminus G \) is not unique, \( y \) must be contained within a certain subset of \( L_1 \), determined by \( G \) and the vector \( v \) described in the \( G \)-uniqueness Theorem. We first have to formulate and prove two preliminary lemmas.

In Lemma 3.2 we give a representation of the vectors \( \hat{y} \in H_G(y) \) in the case of non-unique estimates. The representation is in terms of the vector \( v \) as defined in the \( G \)-uniqueness Theorem.

![Figure 3.3](image-url)
**LEMMA 3.2** Let $G$ be a linear subspace of $L_1$ and assume that there exists a vector $v \in L_\infty$ such that

$$v_i = \pm 1 \quad \text{for } i = 1, \ldots, n$$

$$\sum_{i=1}^{n} v_i g_i = 0 \quad \forall g \in G$$

If the $L_1$-norm estimate $\hat{y}$ in $G$ of a vector $y \in L_1 \setminus G$ is not unique, we have that

$$\mathbb{P}_G(y) = \{ \hat{y} \in G | \hat{y}_i = y_i + \alpha_i v_i, \text{ for all constants } \alpha_i \text{ such that} \}$$

$$\sum_{i=1}^{n} \alpha_i = \| y - \hat{y} \|$$

### PROOF

Assume that $y$ can not be uniquely estimated. By Lemma 3.1 and the $G$-uniqueness Theorem we know that the vector $v \in L_\infty$ satisfies

$$\sum_{i=1}^{n} v_i (y_i - \hat{y}_i) = \| y - \hat{y} \|$$

Since it always is possible to represent the vector $y$ as

$$y_i = \hat{y}_i + \alpha_i v_i \quad i = 1, \ldots, n$$

(3.26)

for some constants $\alpha_i$, it follows

$$\sum_{i=1}^{n} v_i (y_i - \hat{y}_i) = \sum_{i=1}^{n} v_i \alpha_i v_i = \sum_{i=1}^{n} \alpha_i = \| y - \hat{y} \|$$

Thus, for different $\hat{y} \in \mathbb{P}_G(y)$, the representation of $y$ in (3.26) must be such that

$$\sum_{i=1}^{n} \alpha_i = \| y - \hat{y} \|$$

We also have that

$$v_i (y_i - \hat{y}_i) = |y_i - \hat{y}_i| \quad i = 1, \ldots, n$$

and thus
\[ v_i \alpha_i v_i = |\alpha_i v_i| \quad i=1,\ldots,n \]

implying \( \alpha_i \geq 0 \). But then we have that

\[ \hat{y}_i = y_i - \alpha_i v_i, \quad \sum_{i=1}^{n} \alpha_i = \| y - \hat{y} \|, \quad \alpha_i \geq 0 \quad i=1,\ldots,n \]

Since \( \| y - \hat{y} \| = \sum_{i=1}^{n} \alpha_i \) is the same for all vectors \( \hat{y} \) with this representation, the Lemma follows.

q.e.d.

**Lemma 3.3** Let \( G \) be a linear subspace of \( L_1 \) and assume that there exists a vector \( v \in L_\infty \) such that

\[ v_i = \pm 1 \quad \text{for } i=1,\ldots,n \]

\[ \sum_{i=1}^{n} v_i g_i = 0 \quad \forall g \in G \]

If there exist constants \( \alpha_i > 0 \) such that

\[ y_i = \hat{y}_i + \alpha_i v_i, \quad i=1,\ldots,n \]

where \( \hat{y} \) is an \( L_1 \)-norm estimate of \( y \) in \( G \), then \( \hat{y} \) is not unique.

**Proof** From the equivalence \( 1^0 \Leftrightarrow 3^0 \) in the Characterization in \( L_1 \)-Theorem, it follows that

\[ \sum_{i=1}^{n} g_i \text{sign}(y_i - \hat{y}_i) = \sum_{i=1}^{n} g_i \text{sign}(\hat{y}_i + \alpha_i v_i - \hat{y}_i) = \]

\[ = \sum_{i=1}^{n} g_i \text{sign}(\alpha_i v_i) = \sum_{i=1}^{n} g_i v_i = 0 \quad \forall g \in G \]

Since \( \alpha_i > 0 \) and \( v_i = \pm 1 \) for all indices \( i \), it follows that the set \( I \) of all indices \( i \) such that \( y_i \neq \hat{y}_i \) is empty. Therefore,

\[ \sum_{I} |g_i| = \sum_{\emptyset} |g_i| = 0, \quad \forall g \in G \]

and by the Uniqueness Theorem in \( L_1 \), the Lemma is proved.

q.e.d.
**THEOREM 3.8 (NON-UNIQUENESS)** Let $G$ be a linear subspace of $L_1$, and assume that there exists a vector $v \in L_\infty$ such that

$$v_i = \pm 1 \quad \text{for } i = 1, \ldots, n$$

$$\sum_{i=1}^{n} v_i g_i = 0 \quad \forall g \in G$$

Then the $L_1$-norm estimate $\hat{y}$ in $G$ of a vector $y \in L_1 \setminus G$ is not unique if and only if $y \in H$,

$$H = \{ y \in L_1 | y_i = \hat{y}_i + \alpha_i v_i, \text{ for some } \hat{y} \in \mathcal{P}_G(y) \text{ and constants } \alpha_i > 0 \}$$

**PROOF** From Lemma 3.2 we have that

$$\mathcal{P}_G(y) = \{ \hat{y} \in G | \hat{y}_i = y_i + \alpha_i v_i, \text{ for all constants } \alpha_i \text{ such that} \}$$

$$\sum_{i=1}^{n} \alpha_i = \| y - \hat{y} \| \alpha_i \geq 0, \ i = 1, \ldots, n \}$$

In particular, we can choose the constants $\alpha_i$ as

$$\alpha_i = \| y - \hat{y} \| / n > 0$$

and thus $y \in H$.

Conversely, if $y \in H$, it follows directly from Lemma 3.3 that $y$ can not be uniquely estimated.

Q.E.D.

### 3.5.4 The examples reexamined

In this section we will reexamine the examples introduced in the previous subsections and by means of the G-uniqueness and Non-uniqueness Theorems explain why we sometimes will and sometimes will not obtain unique estimates.

**EXAMPLE 3.1 (CONTINUED)** Since the vector $v = (-1, 1, -1, 1)$ satisfies (3.22), (3.23) and (3.24) in the G-uniqueness Theorem, it is possible to obtain a non-unique estimate. One estimate is $\hat{y} = (1.6, 1.6, 3.2, 3.2)$. 
From the table below we see that there are constants \( a_i > 0 \), i.e. \( y \in H \). Then, by the Non-uniqueness Theorem it is clear that the estimation is not unique. See Table 3.1.

<table>
<thead>
<tr>
<th>i</th>
<th>x</th>
<th>y</th>
<th>v</th>
<th>xv</th>
<th>( \hat{y} )</th>
<th>( y - \hat{y} )</th>
<th>( a = \frac{y - \hat{y}}{v} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1.5</td>
<td>-1</td>
<td>-2</td>
<td>1.6</td>
<td>-0.1</td>
<td>0.1 &gt; 0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1.6</td>
<td>0.4</td>
<td>0.4 &gt; 0</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>2</td>
<td>-1</td>
<td>-4</td>
<td>3.2</td>
<td>-1.2</td>
<td>1.2 &gt; 0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>3.2</td>
<td>0.8</td>
<td>0.8 &gt; 0</td>
</tr>
</tbody>
</table>

Table 3.1 The elements of the vectors \( x, y \) and \( v \) and the computation of the constants \( a_i \) for Example 3.1.

**EXAMPLES 3.2 AND 3.3 (CONTINUED)** In the Examples 3.2 and 3.3 we applied a simple linear regression model with intercept, \( y_i = \beta_0 x_{oi} + \beta_1 x_{1i} + \epsilon_i \), \( i = 1, 2, 3, 4 \), and the data

\[
\begin{align*}
  x_0 & = 1 1 1 1 \\
  x_1 & = 1 2 3 4 \\
\end{align*}
\]

We have already showed that the vector \( v \in L_\infty \) defined by \( v = (1, -1, -1, 1) \) satisfies (3.22), (3.23) and (3.24) in the G-uniqueness Theorem, and that it is possible to obtain non-unique estimates.

In Example 3.2 we select the estimate \( \hat{y} = (0.6, 1.7, 2.8, 3.9) \). Then there are constants \( a_i > 0 \) and thus \( y \in H \), that is \( y \) cannot be uniquely estimated. For the computation of the constants \( a_i \), see Table 3.2. Furthermore, form Lemma 3.2 it follows that the set of all estimates, \( H_G(y) \), is determined by

\[
H_G(y) = \{ \hat{y} = b_0 x_0 + b_1 x_1 | \hat{y}_1 = 1 + a_1, \hat{y}_2 = 1.5 - a_2, \hat{y}_3 = 2.5 - a_3, \hat{y}_4 = 4 + a_4, a_1 + a_2 + a_3 + a_4 = 1, a_1 \geq 0, a_2 \geq 0, a_4 \geq 0; b_0, b_1 \in \mathbb{R} \}
\]
In Example 3.3, on the other hand, \( \hat{y} = (1, \frac{5}{3}, \frac{7}{3}, 3) \). Thus, only \( \alpha_2 \) is strictly greater than 0 (see Table 3.3), and \( y \notin \mathcal{H} \). By the Non-uniqueness Theorem the estimation is unique. See Table 3.3.

### Table 3.2
The elements of the vectors \( x_0, x_1, y \) and \( v \) and the computation of the constants \( \alpha_i \) for Example 3.2.

<table>
<thead>
<tr>
<th>i</th>
<th>( x_0 )</th>
<th>( x_1 )</th>
<th>( y )</th>
<th>( v )</th>
<th>( vx_1 )</th>
<th>( vx_2 )</th>
<th>( \hat{y} )</th>
<th>( y - \hat{y} )</th>
<th>( \alpha = \frac{y - \hat{y}}{v} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.6</td>
<td>0.4</td>
<td>0.4 &gt; 0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1.5</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>1.7</td>
<td>-0.2</td>
<td>0.2 &gt; 0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2.5</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
<td>2.8</td>
<td>-0.3</td>
<td>0.3 &gt; 0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3.9</td>
<td>0.1</td>
<td>0.1 &gt; 0</td>
</tr>
</tbody>
</table>

### Table 3.3
The computation of the constants \( \alpha_i \) for Example 3.3.

<table>
<thead>
<tr>
<th>i</th>
<th>( y )</th>
<th>( \hat{y} )</th>
<th>( y - \hat{y} )</th>
<th>( v )</th>
<th>( \alpha = \frac{y - \hat{y}}{v} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0 X 0</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>( \frac{5}{3} )</td>
<td>-( \frac{1}{6} )</td>
<td>-1</td>
<td>( \frac{1}{6} ) &gt; 0</td>
</tr>
<tr>
<td>3</td>
<td>2.5</td>
<td>( \frac{7}{3} )</td>
<td>( \frac{1}{6} )</td>
<td>-1</td>
<td>-( \frac{1}{6} ) X 0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0 X 0</td>
</tr>
</tbody>
</table>

### 3.5.5 Weak asymptotic uniqueness

If the set \( \mathcal{H}_G(y) \) of \( L_1 \)-norm estimates of a vector \( y \) is, in somewhat pragmatic terms, small, or the estimate is practically unique, the non-uniqueness problem will be less important. We therefore introduce the notion of weak asymptotic uniqueness. A sequence of models is said to be weakly asymptotically unique if, for any element \( \hat{y} \in \mathcal{H}_G(y) \), the expectation \( E[\hat{y}] \) converges to a vector \( y \) when \( n \) tends to infinity.
EXAMPLE 3.4 Consider the sequence of models \( \{ y_i = \beta + \varepsilon_i, i=1, \ldots, 2n \} \), i.e. simple location models with an even number of observations. We know that the models are not G-unique and that \( \mathbb{P}_G(y) \) equals the interval \([y_n, y_{n+1}]\), where \( y_n \) and \( y_{n+1} \) are the nth and (n+1)st ordered value of \( y_i \), respectively. For a continuous residual distribution, \( y_n \) and \( y_{n+1} \) are equal with probability zero, i.e. the estimation is unique with probability zero. However, for an increasing number of observations, the expectations \( \mathbb{E}[y_n] \) and \( \mathbb{E}[y_{n+1}] \) converges to the median in the distribution of \( y \) when \( n \to \infty \). Therefore, we expect the set of estimates of the simple location model with a continuous residual distribution and an even number of observations to be arbitrarily small for a sufficiently large number of observations. In other words, the sequence of simple location models with a continuous residual distribution and an even number of observations is weakly asymptotically unique.

EXAMPLE 3.5 Consider again the sequence of location models \( \{ y_i = \beta + \varepsilon_i, i=1, \ldots, 2n \} \) and let the residual distribution be defined by the probability density function

\[
f(\varepsilon) = \begin{cases} 
\frac{1}{2} & \text{if } \varepsilon \in [-2, -1) \cup [1, 2] \\ 
0 & \text{else}
\end{cases}
\]

The median in this distribution is not uniquely determined. Any \( \varepsilon \in [-1,1] \) satisfies \( F(\varepsilon) = \frac{1}{2} \), where \( F \) is the distribution function. Note that \( f(\varepsilon) = 0 \) for any \( \varepsilon \in [-1,1] \).

For an even number of observations, we have that the expectations \( \mathbb{E}[y_n] \) and \( \mathbb{E}[y_{n+1}] \) do not converge. The sequence of simple location models with this residual distribution and an even number of observations is therefore not weakly asymptotically unique.

In the Examples 3.4 and 3.5 the existence of a uniquely determined
median in the residual distribution played an important role. In the next theorem, we will show more generally that the existence of a uniquely determined median in the residual distribution is sufficient for weakly asymptotically unique estimation.

**THEOREM 3.9 (WEAK ASYMPTOTIC UNIQUENESS)** Let \( G \) be the \( m \)-dimensional linear subspace of \( L_1 \) spanned by the \( n \)-dimensional vectors \( X_1, \ldots, X_m \).

Assume that \( \lim_{n \to \infty} n^{-1}X'X \) is a positive definite matrix, where \( X \) is the matrix defined by \( X = (X_1, \ldots, X_m) \). Let \( \beta \) be an \( m \)-dimensional vector and define the \( n \)-dimensional vector \( y \) by \( y = X\beta + \epsilon \), where \( \epsilon \) is an \( n \)-dimensional random vector. If the components \( \epsilon_i \) are independent and equally distributed random variables with a uniquely defined median in zero and a probability density function \( f \) such that \( f(\epsilon) > 0 \) for all \( \epsilon \) in a neighbourhood of zero, then \( y \) is weakly asymptotically uniquely estimated in \( G \).

**PROOF** Define the function \( S_n \) by

\[
S_n(b) = \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{m} X_i b_j \right| - \sum_{i=1}^{n} |\epsilon_i|
\]

Since the term \( \sum |\epsilon_i| \) is independent of \( b \), \( S_n \) will attain its minimum for the \( L_1 \)-norm estimate \( \hat{\gamma} = \sum_{j=1}^{m} j \hat{\beta}_j \). We will show that \( n^{-1}S_n \) converges in probability to a function \( S \) which has a unique minimum. Let therefore the random variables \( V_i \) and \( W_i \) be defined by

\[
V_i = \begin{cases} 
1 & \text{if } \epsilon_i \geq 0 \\
0 & \text{else}
\end{cases}
\]

and

\[
W_i = \begin{cases} 
1 & \text{if } y_i \geq \sum_{j=1}^{m} X_i b_j \\
0 & \text{else}
\end{cases}
\]
Then, we have

\[ S_n(b) = 2 \sum_{i=1}^{n} (W_i - \bar{V}_i) \varepsilon_i - 2 \sum_{i=1}^{n} W_i \sum_{j=1}^{m} X_{ij}(b_{ji} - \beta_{ji}) + \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}(b_{ji} - \beta_{ji}) \]

Since

\[ E[(W_i - \bar{V}_i) \varepsilon_i] = \int_{-\infty}^{\infty} t f(t) \, dt \sum_{j=1}^{m} X_{ij}(b_{ji} - \beta_{ji}) \]

we have by Chebyshev's inequality

\[ S(b) = \lim_{n \to \infty} n^{-1} S_n(b) = 2 \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{\infty} t f(t) \, dt - \sum_{i=1}^{m} \sum_{j=1}^{m} X_{ij}(b_{ji} - \beta_{ji}) + \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}(b_{ji} - \beta_{ji}) \]

But then we have

\[ \frac{\partial S}{\partial b_k} = -2 \lim_{n \to \infty} \sum_{i=1}^{n} \int_{-\infty}^{\infty} f(t) \, dt X_{ki} + \lim_{n \to \infty} \sum_{i=1}^{n} X_{ki} \sum_{j=1}^{m} X_{ij}(b_{ji} - \beta_{ji}) \]

for \( k = 1, \ldots, m \)

Since the vectors \( X_1, \ldots, X_m \) linearly independent, \( \frac{\partial S}{\partial b_k} = 0 \) for \( k = 1, \ldots, m \) if and only if

\[ \int_{-\infty}^{\infty} f(t) \, dt = \frac{1}{2} \quad i=1, \ldots, n \]

But, since the residual distribution has a unique median in zero, \( \frac{\partial S}{\partial b_k} = 0 \) for \( k = 1, \ldots, m \) if and only if

\[ \sum_{j=1}^{m} X_{ij}(b_{ji} - \beta_{ji}) = 0 \quad i=1, \ldots, n \]
Thus, $b = \beta$ is the only extreme point. To see that $b = \beta$ is a minimum, we compute the second derivative.

$$\frac{\partial^2 S}{\partial b_k \partial b_l} = 2 \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} f(\sum_{i} X_{ij} (b_j - \beta_j)) X_{ik} X_{il}$$

Since $\lim_{n \to \infty} n^{-1} X'X$ is positive definite, $\frac{\partial^2 S}{\partial b_k \partial b_l}$ is positive.

Thus, $S$ has a unique minimum

Q.E.D.

The idea in this proof is much the same as in the proof of consistency in Amemiya (1979).

3.5.6 Some concluding remarks

From the Existence Theorem in Section 3.2 we know that at least one $L_1$-norm estimate always will exist. The question whether the estimate is unique or not, is more complicated. In this section a necessary and sufficient condition that a model always has unique $L_1$-norm estimates (G-uniqueness) is first proved. It is then shown, that when the G-uniqueness condition is not satisfied, there may exist data sets leading to unique $L_1$-norm estimates. In the non-uniqueness theorem, the collection of data sets leading to non-unique $L_1$-norm estimates is described. However, the analysis is here rather complicated, and a restrictive case of non-G-uniqueness has to be considered. It is therefore desirable to prove a stronger version of the non-uniqueness theorem, including all cases of non-G-uniqueness.

However, the determination of non-unique estimates in practice seems to be difficult. For example, for the determination of G-uniqueness in a simple linear regression model without intercept, we have to check $2^n$ sums when the number of observations is $n$. For more complicated models, this control routine is not less laborious. The Non-uniqueness Theorem, however, indicates that it may be possible to construct a general and
more simple routine for checking non-uniqueness, when an estimate is already available. Thus, more research will be needed for constructing such a checking routine, and implementing it into the computation algorithms.

Finally, we introduced the notion of weak asymptotic uniqueness. An interpretation of this notion is that the expectation \( E[\hat{y}] \) of any \( L_1 \)-norm estimate \( \hat{y} \in H_0(y) \), converges when the number of observations is increasing. The Weak Asymptotic Uniqueness Theorem indicates that conditions for weak asymptotic uniqueness in linear regression models are rather mild. The conditions are satisfied for a large class of regression models.

3.6 Geometric interpretations

It is often instructive to make geometric interpretations of analytical problems. Unfortunately, it is impossible to imagine spaces with more than three dimensions. In order to make the interpretations as easy as possible we consider only the case with two dimensions, i.e. we have only two observations. The realism in the example is then yielded to the benefit of the perspicuousness. A two-dimensional vector space is illustrated in Figure 3.4. We consider the case of only one predictor variable with the observations \( x_1 \) and \( x_2 \). The subspace \( G \) then consists of all vectors on the straight line through the origin and parallel with vector \( x = (x_1, x_2)' \). With the observations \( y_1 \) and \( y_2 \) on \( y \) it is also possible to point out that vector in the \( L_p \)-space. Our problem is to choose a vector \( \hat{y} \in G \), such that the vector \( y - \hat{y} \) is as short as possible, i.e. \( \hat{y} \) is to minimize \( \|y - \hat{y}\| \).

Consider now all vectors that are on a certain distance, say \( r \) from the vector \( y \). In \( L_2 \) then, we are considering the set \( \{z \in L_2 | \sum_{i=1}^{n} (y_i - z_i)^2 = r \} \). See Figure 3.5.
Figure 3.4 The vector space $L_p$.

Figure 3.5 The space $L_2$ and the subset of all vectors with the distance $r$ from the vector $y$. 
This set constitutes a circle with the center in \( y \) and with radius \( r \). By continuously increasing the radius \( r \) in the circle, for a sufficiently large \( r \) the circle touches the line \( G \). It is clear that the same conclusion holds even in higher dimensions. In higher dimensions \( G \) is a hyperplane and the set of vectors of a constant distance from the vector \( y \) is a hyperball. For a sufficiently large radius of the hyperball, it touches the hyperplane \( G \). Since the tangent point represents that vector in \( G \), which is closest to \( y \), the circle (resp hyperball) touches \( G \) precisely in \( \hat{y} \). It is then clear that there always exists an \( L_2 \)-norm estimate. Since the intersection is in just one point, it follows that the \( L_2 \)-norm estimate is uniquely determined.

The cases \( p=1 \) and \( p=\infty \) are illustrated in Figure 3.6 and 3.7, respectively. In both cases the set of all vectors of a constant distance from the vector \( y \) is a square with center in \( y \). For \( p=1 \) the diagonals in the square are parallel with the coordinate axes and are of length \( 2r \). For \( p=\infty \) the sides are parallel with the coordinate axes and are of length \( 2r \). By continuously increasing \( r \), the square is to touch the line \( G \) for a sufficiently large \( r \). Then the existence of \( L_1 \)- and \( L_\infty \)-norm estimates are guaranteed. On the other hand, if \( G \) is parallel to one of the sides of the square, they intersect in more than one point, i.e. the estimates are not unique.

![Figure 3.6](image1.png)  
**Figure 3.6** The space \( L_1 \) and the set of all vectors with the distance \( r \) from the vector \( y \).

![Figure 3.7](image2.png)  
**Figure 3.7** The space \( L_\infty \) and the set of all vectors with the distance \( r \) from the vector \( y \).
For $1 < p < \infty$, the corresponding arguments concerning the existence of $L_p$-norm estimates hold. Further, for $1 < p < \infty$ the set $\{z \in L_p | \|y-z\| \leq r\}$ is strictly convex, see Figures 3.8 and 3.9. This means that if two points in the set are connected with a straight line, all points on the line are interior points in the set. The strict convexity is thus a guarantee for unique points of intersection, i.e. all $L_p$-norm estimates are unique when $1 < p < \infty$.

![Figure 3.8](image1.jpg)  
**Figure 3.8** The space $L_p$ and the set of all vectors with the distance $r$ from the vector $y$, when $1 < p < 2$.  

![Figure 3.9](image2.jpg)  
**Figure 3.9** The space $L_p$ and the set of all vectors with the distance $r$ from the vector $y$, when $2 < p < \infty$.

### 3.7 A general notion of orthogonality

In this section we will give a definition of orthogonality in general normed linear spaces and examine how the notion of orthogonality is related to $L_p$-norm estimation. A vector $y$ in a normed linear space $E$ is said to be orthogonal to a vector $z \in E$ if

$$\|y + \alpha z\| \geq \|y\|$$

for all scalars $\alpha$. 
If two vectors, \( y \) and \( z \), are orthogonal, we will denote this by \( y \perp z \).

This definition of orthogonality was originally given by Birkhoff (1935). It is an extension of the usual definition of orthogonality, since in a pre-Hilbert space \( H \), we have that \( y \perp z \) if and only if \( (y,z) = 0 \), where \( (\ , \ ) \) is the inner product. If \( (y,z) \neq 0 \), we have for the scalar \( \alpha = \frac{(y,z)}{(z,z)} \) that

\[
\|y + \alpha z\|^2 = (y - \frac{(y,z)}{(z,z)} z, y - \frac{(y,z)}{(z,z)} z) = \\
(y,y) - 2 \frac{(y,z)^2}{(z,z)} \frac{(y,z)}{(z,z)} = \\
(y,y) - \frac{(y,z)^2}{(z,z)} < (y,y) = \|y\|^2
\]

and \( y \) is not orthogonal to \( z \). Conversely, if \( (y,z) = 0 \), we have for any scalar \( \alpha \), that

\[
\|y + \alpha z\|^2 = (y+\alpha z, y+\alpha z) = \|y\|^2 + \alpha^2 \|z\|^2 \geq \|y\|^2
\]

that is \( y \perp z \).

A vector \( y \) in a normed linear space \( E \) is said to be orthogonal to a set \( M \subseteq E \), denoted by \( y \perp M \) if

\[ y \perp z, \quad \forall z \in M \]

The relation between orthogonality and \( L_p \)-norm estimation is given in the next Theorem.

**THEOREM 3.10 (ORTHOGONALITY)** Let \( E \) be a normed linear space, \( G \) a linear subspace of \( E \), \( y \in E \setminus \hat{G} \) and \( \hat{y} \in G \). Then \( \hat{y} \in \text{Proj}_G(y) \) if and only if \( y - \hat{y} \perp G \).

**PROOF** We have that \( \hat{y} \in \text{Proj}_G(y) \) if and only if \( \hat{y} \in G \) minimizes the functional \( \phi \) defined on \( G \) by

\[
\phi(g) = \phi_{G,y}(g) = \|y - g\| \quad g \in G
\]

or, the element \( 0 \in G \) minimizes the functional \( \psi \) defined on \( G \) by

\[
\psi(g) = \phi(\hat{y} + g) = \|y - \hat{y} + g\|
\]
But this is equivalent to the functional \( X_g \) defined on the set of all scalars by

\[
X_g(\alpha) = \psi(\alpha g) = \| y - \hat{y} + \alpha g \| \quad \alpha \text{ is a scalar}
\]

which is minimized by \( \alpha = 0 \). From the definition of orthogonality, we have that \( y - \hat{y} \perp \hat{y} \) if and only if

\[
\| y - \hat{y} + \alpha g \| \geq \| y - \hat{y} \| \quad \text{for all } g \in G \text{ and all scalars } \alpha.
\]

This condition is obviously satisfied by (3.27) and thus \( y - \hat{y} \perp \hat{y} \).

\[ Q.E.D. \]

From the Orthogonality Theorem it immediately follows that \( y - \hat{y} \perp \hat{y} \) since \( \hat{y} \in G \). We further have the following Corollary.

**COROLLARY 3.3** \( z \perp G \) if and only if \( 0 \in H_p(z) \).

**PROOF** Assume that \( z \perp G \). Then we have that \( z - 0 \perp G \) and by the Orthogonality Theorem \( 0 \in H_p(z) \). Conversely, assume that \( 0 \in H_p(z) \). From the Orthogonality Theorem it follows that \( z - 0 \perp G \) and \( z \perp G \).

\[ q.e.d. \]

Observe that Corollary 3.3 implies that the \( L_p \)-norm estimate of \( y - \hat{y} \) is the zero vector 0.

### 3.8 Relations to M- and L-estimators

In this section we will define two classes of estimators. The classes are, with the notation in Huber (1972 and 1977), maximum likelihood type estimators (M-estimators) and linear combinations of order statistics (L-estimators). We will also determine the relations between \( L_p \)-norm estimators and M- and L-estimators, respectively.

As before, consider the linear regression model

\[
y = X\beta + \varepsilon
\]
For the maximum likelihood type estimators (M-estimators) let $\rho$ be a real valued function with argument vector $\beta$ and with derivative $\psi = \rho'$. The M-estimator of $\beta$ is the vector $b_M$ that minimizes

$$
\sum_{i=1}^{n} \rho \left( y_i - \sum_{j=1}^{m} x_{ij} b_j \right)
$$

that is $b_M$ is a solution of the equations

$$
\sum_{i=1}^{n} x_{ik} \psi \left( y_i - \sum_{j=1}^{m} x_{ij} b_j \right) = 0.
$$

If we choose

$$
\psi(x) = f'(x)/f(x),
$$

where $f$ is the probability density function of the residuals, we have the maximum likelihood estimator of the parameter vector $\beta$. Further, if we choose

$$
\rho(x) = |x|^p
$$

it is clear from (3.1) that the $L_p$-norm estimators are M-estimators. Thus, the maximum likelihood estimators as well as the $L_p$-norm estimators are subsets in the same class of estimators. In Section 3.9 we will examine the intersection of these subsets, that is we will give a class of probability distributions with the property that a maximum likelihood and an $L_p$-norm estimator coincide.

Consider now the simple location model

$$
y = \beta + \epsilon
$$

and let $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ be the ordered sample. Linear combinations of order statistics ($L$-estimators) are then of the form

$$
b_L = \sum_{i=1}^{n} a_i y_{(i)}
$$

where the weights $a_i$ are generated by
\[ a_i = \frac{i/n}{J(t) \, dt} \quad (i-1)/n \]

for some function \( J \) satisfying \( \int_0^1 J(t) \, dt = 1 \).

It is immediately seen that the \( L_1 \)-norm estimator (median) is the \( L \)-estimator corresponding to the function \( J \) with unit mass in \( t = \frac{1}{2} \). In the same manner, the \( L_\infty \)-norm estimator is the \( L \)-estimator corresponding to the function \( J = \frac{1}{2}(J_0 + J_1) \) where \( J_0 \) and \( J_1 \) has unit mass in \( t=0 \) and \( t=1 \), respectively. The \( L \)-estimator corresponding to the constant function \( J=1 \) is clearly equivalent to the \( L_2 \)-norm estimator.

These notations for the simple location model are generalized to multiple linear models in Koenker and Bassett (1978). With the same arguments as above, the \( L_1 \), \( L_\infty \) and \( L_\infty \)-norm estimators belong to the class of \( L \)-estimators. However, for \( 1 < p < 2 \) or \( 2 < p < \infty \), the \( L_p \)-norm estimators and the \( L \)-estimators do not coincide.

### 3.9 Equivalence of the \( L_p \)-norm and maximum likelihood estimators

The notes in this section have been reported by Turner (1960). Consider the probability density function

\[ f(y_i) = \frac{\gamma}{2 \delta \Gamma\left(\frac{1}{\gamma}\right)} \exp\left(-\frac{|y_i - \beta|^\gamma}{\delta^\gamma}\right) \]

for \(-\infty < y_i < \infty\), \(-\infty < \beta < \infty\), \( \delta > 0 \) and \( \gamma > 0 \). Here \( \Gamma \) stands for the gamma function. This density function is symmetric around \( \beta \). For \( \gamma = 1 \) we have the Laplace (double exponential) distribution

\[ f(y_i) = \frac{1}{2 \delta} \exp\left(-\frac{|y_i - \beta|}{\delta}\right), \quad -\infty < y_i < \infty. \]

If \( \gamma = 2 \) we have the normal distribution
\[ f(y_i) = \frac{1}{\delta/2\pi} \exp\left(\frac{-(y_i - \beta)^2}{\delta^2}\right), \quad -\infty < y_i < \infty, \]

with expectation \( \beta \) and variance \( \delta^2/2 \). If \( \gamma \) increases unbounded, the limit distribution is the uniform distribution

\[ f(y_i) = \frac{1}{2\delta}, \quad \beta - \delta \leq y_i \leq \beta + \delta. \]

Assume now that we for a known \( \gamma > 0 \) make \( n \) observations on \( y \). The log-likelihood function is then

\[ l(\beta, \delta, y) = \ln \gamma^n - \ln 2\delta \Gamma \left(\frac{1}{\gamma}\right)^n - \frac{1}{\delta^\gamma} \sum_{i=1}^{n} |y_i - \beta|^{\gamma}. \]

Clearly, it is only the last term in the loglikelihood function that depends on \( \beta \). Since it appears with a negative sign, the maximum likelihood estimate of \( \beta \) is that \( \hat{\beta} \) which minimizes

\[ \sum_{i=1}^{n} |y_i - \beta|^{\gamma}. \]

In the case of uniform distribution (\( \gamma = \infty \)), the maximum likelihood estimate of \( \beta \) is that \( \hat{\beta} \) which minimizes

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} |y_i - \beta|, \]

i.e. that \( \hat{\beta} \) which minimizes

\[ \max_{i} |y_i - \beta|. \]

Then it is clear that the maximum likelihood estimator of \( \beta \) is the same as the \( L_p \)-norm estimator of \( \beta \) with \( p = \gamma \).

Further, the maximum likelihood estimate of \( \delta \) is for \( \gamma = 1 \) the mean of the absolute deviations, \( \hat{\delta} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{\beta}| \), for \( \gamma = 2 \), the square root of two times the sample standard deviation, \( \hat{\delta} = \left(2 \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta})^2\right)^{1/2} \), and for \( \gamma = \infty \), the maximum likelihood estimate of \( \delta \) equals half of the sample range, \( \hat{\delta} = (y_{\text{max}} - y_{\text{min}})/2 \).
It is now easy to generalize Turner's observations to more complicated models. Thus, the maximum likelihood estimator and the $L_p$-norm estimator are equivalent for $p = \gamma$, when the residuals $\epsilon_i$, are independent and equally distributed with density function

$$f(\epsilon_i) = \frac{\gamma}{2 \delta \Gamma \left( \frac{1}{\gamma} \right)} \exp\left\{ -\frac{|\epsilon_i|^\gamma}{\delta^\gamma} \right\} .$$

3.10 Computation of $L_p$-norm estimates, $1 < p < \infty$, $p \neq 2$

In Section 2.5 we discussed some algorithms for computation of $L_1$- and $L_\infty$-norm estimates. The computation of $L_2$-norm estimates is equivalent to solve a set of linear equations, viz. the normal equations. Since algorithms for the solution of systems of linear equations are well-known, we can leave the $L_2$-norm case without any further comments. To complete the discussion of the computation of $L_p$-norm estimates, we will in this section outline some methods applicable in the case $1 < p < \infty$.

From Theorem 3.3, Characterization in $L_p$, $1 < p < \infty$, it follows that $\hat{\beta}$ is an $L_p$-norm estimate if and only if

$$T_j(\hat{\beta}) = \sum_{i=1}^{n} x_{ij} |y_i - \sum_{k=1}^{m} \hat{\beta}_k x_{ik}|^{p-1} \text{sign}(y_i - \sum_{k=1}^{m} \hat{\beta}_k x_{ik}) = 0, \quad j=1, \ldots, m$$

or, in matrix notation

$$T(\hat{\beta}) = X'Q(\hat{\beta}) = 0$$

where $Q(b)$ is an $(nx1)$-vector with elements $q_i = |y_i - \sum_{k=1}^{m} b_k x_{ik}|^{p-1} \text{sign}(y_i - \sum_{k=1}^{m} b_k x_{ik})$. Thus, the problem of computing $L_p$-norm estimates is equivalent to solving a system of non-linear equations. There are several techniques for solving such systems. The algorithms belonging to the class of gradient methods are often intuitively simple. It appears also that
it is often relatively easy to write computer programs for those algorithms.

We will now describe an iterative procedure based on gradient methods for the computation of an $L_p$-norm estimate. Let therefore $\{b^r\}_{r=1}^\infty$ be the sequence of estimates generated from the iterative method, given an initial guess $b^0$ of $\hat{\beta}$. A first order Taylor expansion of $T_j$ around $b^r$ is given by

$$T_j(b) \approx \sum_i^n x_{ij}y_i - \sum_k^m b_k^r x_{ik}^p \text{sign}(y_i - \sum_k^m b_k^r x_{ik}) -$$

$$- (p-1) \sum_k^m (\hat{b}_k - b_k^r) \sum_i^n x_{ij}^p |y_i - \sum_k^m b_k x_{ik}|^{p-2}, \quad j=1, \ldots, m,$$

or in matrix notations

$$T(\hat{b}) \approx X'Q(b^r) - (p-1)X'W(b^r)X(\hat{b} - b^r)$$

where $W(b)$ is a $(nxn)$-diagonal matrix with elements $w_{ii} = |y_i - \sum_k^m b_k^r x_{ik}|^{p-2}$.

Since $T(\hat{b}) = 0$, it follows that

$$(p-1)X'W(b^r)X(\hat{b} - b^r) \approx X'Q(b^r).$$

Therefore, the next iterative value $b^{r+1}$ is defined by

$$b^{r+1} = b^r + \frac{1}{p-1} (X'W(b^r)X)^{-1}X'Q(b^r)$$

For $p > 2$ the functions $T_j(b)$, $j=1, \ldots, m$ are concave-convex and therefore the sequence $\{b^r\}_{r=1}^\infty$ will converge to the $L_p$-norm estimate $\hat{b}$. When $1 < p < 2$, starting values $b^0$ far from $\hat{b}$ may cause a divergence of the sequence $\{b^r\}_{r=1}^\infty$. It is therefore convenient to define the iterative procedure by

$$b^{r+1} = b^r + c_r \frac{1}{p-1} (X'W(b^r)X)^{-1}X'Q(b^r)$$

where $c_r$ is a constant, $0 < c_r \leq 1$ to assume that

$S(b^{r+1}) < S(b^r)$.
where

$$S(b) = \sum_{i=1}^{n} |y_i - \sum_{j=1}^{m} b_{i,j}|^p$$

For more formal proofs of the convergency of the iterative procedure, and optimal choices of the constant $c_\epsilon$, we refer to Ortega and Rheinboldt (1970), Gallant (1971) and Lawrence (1975).

If the starting value $b^0$ is close to the estimate $\hat{\beta}$, the convergency will be faster. Therefore, for $p$ near 2, the $L_2$-norm estimate may be a suitable starting value, while the $L_1$-norm estimate may be used for smaller values of $p$. In cases of large $p$, the $L_\infty$-norm estimate is recommended as the starting value.

Another approach to solve a system of non-linear equations is to apply the Newton-Raphson iterative procedure. The Newton-Raphson procedure is based on a second order Taylor expansion of $T(\hat{\beta})$ around $b^r$. Formally we have

$$T(\hat{\beta}) \approx T(b^r) + T'(b^r)(\hat{\beta} - b^r) + \frac{1}{2}(\hat{\beta} - b^r)'T''(b^r)(\hat{\beta} - b^r)$$

Given a starting value $b^0$, the next value in the sequence $b^{r+1}$ is defined as the solution of

$$T(b^r) + T'(b^r)(b^{r+1} - b^r) + \frac{1}{2}(b^{r+1} - b^r)'T''(b^r)(b^{r+1} - b^r) = 0.$$ 

However, in many practical cases, the additional computational burden is not justified by a smaller computational effort totally.

### 3.11 Conclusions

In this chapter we have defined the class of $L_p$-norm estimators. The $L_1$-, $L_2$- and $L_\infty$-norm estimators are members of that class. For a fixed $p$, $1 \leq p \leq \infty$, the $L_p$-norm estimator is defined as to minimize the $L_p$-norm ("the length") of the observed residual vector.
Mostly this chapter is devoted to a discussion of some basic results for $L_p$-norm estimators. Thus, questions concerning existence, characterization, uniqueness and computation of estimates are illuminated.

The results from Section 3.2 show that there always exists at least one $L_p$-norm estimate. When we know that there exist estimates, it is meaningful to describe them. Therefore characterizations of $L_p$-norm estimates are given in Section 3.3. Particularly for $1 < p < \infty$, it is possible to give a most powerful characterization of the estimates, useful for the computation and for the derivation of distributional properties. The characterization of $L_1$-norm estimates is more complex and finds most of its applications in the discussion of uniqueness.

The question of uniqueness is divided into two parts, viz. one simple part containing the cases $1 < p < \infty$ and one difficult part containing the case $p = 1$. The simple case has a simple answer: when $1 < p < \infty$ the estimates are unique. For $p = 1$ the answer is more complex and still not entirely developed. In Section 3.5, however, a survey over the recent results on the uniqueness problem is given. We introduced the notion of G-uniqueness when the $L_1$-norm estimate always is unique. The most powerful result in Section 3.5 is a necessary and sufficient condition on the design matrix $X$ for G-uniqueness. When the matrix $X$ does not satisfy the condition for G-uniqueness, the question of unique estimates depends on $y$. In that case, events of unique and non-unique estimates appear with a certain probability, determined by the matrix $X$ and the probability distribution of $y$.

When an $L_1$-norm estimate is not unique, there exists an uncountable infinite number of estimates. Therefore, it is meaningless to talk about "the number" of estimates. If, for a sequence of models, the expectation $E[\hat{y}]$ of any $L_1$-norm estimate $\hat{y} \in \mathcal{P}_G(y)$, converges when the number of observations is increasing, we say that the sequence of estimates is weakly asymptotically unique. The idea behind this notion is that under
the conditions for weak asymptotic uniqueness, the problem of non-unique $L_1$-norm estimates is not of practical importance, when the number of observations is sufficiently large. It is shown that when the residuals are independent and equally distributed with a uniquely determined median in zero and a positive density function in a neighbourhood of zero, sequences of estimates from linear regression models are weakly asymptotically unique.

The existence and uniqueness problems are geometrically illustrated in Section 3.6. From these geometric considerations it also follows that the orthogonality concept, defined on the Hilbert space $L_2$, can more generally be defined on all Banach spaces $L_p$. With this generalized orthogonality concept it was shown in Section 3.7 that the vector of estimated residuals is orthogonal to all linear combinations of the regressor variables, in particular to the estimated dependent variable $\hat{y}$.

In Huber (1972 and 1977) the class of maximum likelihood type (M-) estimators and the class of estimators that are linear combinations of order statistics (L-estimators) were introduced. The results of Section 3.8 show that the class of $L_p$-norm estimators is a subclass of the class of M-estimators. Further, it was shown that the $L_1$, $L_2$ and $L_\infty$-norm estimators are members of the class of L-estimators. For some residual distributions, the $L_p$-norm estimators are equivalent to maximum likelihood estimators. The class of residual distributions that give rise to this equivalence was given in Section 3.9.

The last topic discussed in this chapter is the computation of $L_p$-norm estimates, when $1 < p < \infty$. Thus, in Section 3.10 a suitable algorithm for the computation was given.

In the next chapter, a more specific property of $L_p$-norm estimators will be discussed, viz. the asymptotic distribution.
CHAPTER IV. SOME SAMPLING PROPERTIES OF $L_p$-NORM ESTIMATORS

4.1 Introduction

In Chapter 3 we defined the class of $L_p$-norm estimators and proved some elementary properties for the members in that class. In particular, we proved existency, and for $1 < p < \infty$, uniqueness of the estimates. Necessary and sufficient conditions for unique $L_1$-norm estimates were also given. One of the most striking results was that in sufficiently large samples, the set of $L_1$-norm estimates could in mean be made arbitrary small for a large class of linear regression models. Therefore, problems of non-unique $L_1$-norm estimates should not, seen from a practical view, cause any serious problems in sufficiently large samples.

Further on, algorithms for computation of $L_p$-norm estimators are known and has been summarized in Sections 2.5 and 3.10.

In this chapter we will continue the derivation of properties of $L_p$-norm estimators. All results derived here concern the sample distributions of $L_p$-norm estimators of the parameter vector $\beta$ in the linear regression model (2.5). Section 4.2 deals with the "classical regression case", i.e. fixed non-stochastic regressor variables and independent and identically distributed residuals. It turns out that for $p$ sufficiently small, $\sqrt{n}(\hat{\beta}_{pn} - \beta)$ is asymptotically normally distributed, where $\hat{\beta}_{pn}$ is the $L_p$-norm estimator of the parameter vector $\beta$. The variance in the asymptotic distribution of $\sqrt{n}(\hat{\beta}_{pn} - \beta)$ is $\omega_p^2 Q^{-1}$ where $Q = \lim_{n \to \infty} n^{-1} x'x$ and $\omega_p^2$ is a constant related to some moments of the residual distribution. For $p = 1$, $\omega_1^2$ equals the asymptotic variance of the sample median, when a sample is drawn from the residual distribution. For $1 < p < \infty$ we have that $\omega_p^2 = \mathbb{E}[|\epsilon_1|^{2p-2}] / \{(p-1)^2 \mathbb{E}[|\epsilon_1|^{p-2}]\}$.

Sections 4.3, 4.4 and 4.5 are devoted to some departures from the "classical" assumptions. Thus, the case of stochastic regressors is treated.
in Section 4.3. When the regressors are stochastic, we have to add a "non-correlation" assumption, viz. \( E[X_{1j}\mid \varepsilon_1]^{P^{-1}\text{sign}(\varepsilon_1)} = 0, j=1,\ldots,m \). This is automatically satisfied when the regressors and residuals are independent. The converse is, however, not generally true, i.e. if the "non-correlation" assumption is satisfied, regressors and residuals are not necessarily independent.

In the case of stochastic regressors, it appears that \( \sqrt{n}(\hat{\beta}_p - \beta) \) is asymptotically normally distributed for sufficiently small \( p \). When regressors and residuals are independent, the variance in the asymptotic distribution is seen to be reduced to a formula, similar to that of the classical case.

The residuals are here said to be heteroscedastic, if they are equally distributed up to a scale parameter. In Section 4.4 it is shown that the \( L_p \)-norm estimators are asymptotically normally distributed in the heteroscedastic case. However, in the heteroscedastic case, the regression model under consideration may also be transformed to a model with equally distributed residuals. After the transformation, the results from Sections 4.2 and 4.3, respectively, may be applied.

Linearly dependent residuals are defined in Section 4.5 as the case when a linear transformation of the residuals gives independent and equally distributed residuals. Thus, ARMA processes and residuals with a multivariate stable distribution are special cases of that. In two Monte Carlo simulations, summarized in Section 4.5, the results indicate that the sample distribution of the \( L_1 \)-norm estimator is extremely sensitive to departures from the assumption of independent residuals. However, if the model is transformed to a model with independent and equally distributed residuals, the results from Sections 4.2 and 4.3, respectively, may be applied.

When the regressors are fixed and when the regressors are stochastic and independent of the residuals, the asymptotic variance of the \( L_p \)-norm
estimators depends on the residual distribution only through $\omega_p^2$. For a fixed residual distribution, $\omega_p^2$ may be viewed as a function of $p$. The optimal $L_p$-norm estimator is defined as that $L_p$-norm estimator which has the least $\omega_p^2$. In Section 4.6, $\omega_p^2$ is computed for members of the class of double exponential distributions (the class of distributions where $L_p$-norm and maximum likelihood estimators are equivalent). $\omega_p^2$ is also computed for members of a class of contaminated distributions. The problem of estimating $\omega_p^2$ is discussed in Section 4.7.

In Section 4.8 $L_p$-norm estimation in cases of heteroscedastic and linearly dependent residuals are viewed geometrically. In Section 4.9 some conclusions are given.

In this chapter, $S_{pn}$ will be used to denote the function defined by

$$S_{pn}(b) = \sum_{i=1}^{n} |y_i - \sum_{j=1}^{m} b_j x_{ij}|^p = \|y - Xb\|^p, \quad 1 \leq p < \infty$$

where the vector $y$ and the matrix $X$ refer to the linear model (2.5).

4.2 The sampling distribution of $L_p$-norm estimators

In this section we will give a theorem on the asymptotic distribution of the $L_p$-norm estimator. We will restrict our attention to a linear regression model under somewhat classical assumptions, viz. we assume non-stochastic regressors and independent and identically distributed residuals. We will also study conditions for unbiased $L_p$-norm estimation. The results in Theorem 4.1 will be based on the following set of assumptions:

A1. The model under consideration is that given in (2.5).

A2. The elements $x_{ij}$ in $X$ are constants and $Q = \lim_{n \to \infty} X'X$ is a positive definite matrix.

A3. The residuals $\varepsilon_i$ are independent and identically distributed with common distribution function $F_\varepsilon$. 
A4. For \( p = 1 \), \( F_\varepsilon \) has its median in zero and has a positive derivative in zero.

A5. For \( p > 1 \), \( E[|\varepsilon_1|^{p-1}\text{sign}(\varepsilon_1)] = 0 \). Furthermore, the expectations \( E[|\varepsilon_1|^{p-2}] \) and \( E[|\varepsilon_1|^{2p-2}] \) exist, and for a constant \( c \) the stochastic variables \( \tilde{T}_{i}^{jk} \), defined by

\[
\tilde{T}_{i}^{jk} = x_{ik} \varepsilon_i - cx_{ik}/\sqrt{n} |x_{ik}|^{p-1}\text{sign}(\varepsilon_i - cx_{ik}/\sqrt{n}), \quad j,k = 1, \ldots, m,
\]

satisfy the Lindeberg conditions, and

\[
\frac{1}{n} \sum_{i=1}^{n} (\tilde{T}_{i}^{jk} - E[\tilde{T}_{i}^{jk}]) \in \mathcal{N}(0, V[\tilde{T}_{i}^{jk}])^{1/2}
\]

converges to a normally distributed stochastic variable \( U_j \), with expectation zero and variance one, with

\[
\lim_{n \to \infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[\tilde{T}_{i}^{jk}] = - (p-1)E[|\varepsilon_1|^{p-2}]g_{jk} c
\]

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} V[\tilde{T}_{i}^{jk}] = E[|\varepsilon_1|^{2p-2}]g_{jj}
\]

\[
\text{Cov}[U_j, U_k] = \frac{q_{jk}}{\sqrt{q_{jj}q_{kk}}}
\]

and where \( q_{jk} \) are elements in the matrix \( Q \) defined in Assumption A2.

**THEOREM 4.1 (ASYMPTOTIC DISTRIBUTION)** Let \( p > 1 \) be a constant and let Assumptions A1 - A5 be satisfied. Then \( \sqrt{n}(\hat{\beta}_{pn} - \beta) \) is asymptotically \( m \)-variate normally distributed with mean zero and covariance matrix \( \omega_p^2 Q^{-1} \), where \( \hat{\beta}_{pn} \) is the \( L_p \)-norm estimator and

\[
\omega_p^2 = \begin{cases} 
\{2F_{\varepsilon}(0)\}^{-2} & \text{for } p = 1 \\
E[|\varepsilon_1|^{2p-2}]/\{(p-1)E[|\varepsilon_1|^{p-2}]\}^2 & \text{for } p > 1
\end{cases}
\]
The case \( p = 1 \) is proved in Bassett and Koenker (1978). Therefore we only consider the case \( p > 1 \). The partial derivative of \( S_{pn} \) with respect to its argument \( b_k \) is

\[
\frac{3S_{pn}(b)}{\partial b_k} = -p \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ik} \left| \varepsilon_i + \sum_{j=1}^{m} (\beta_j - b_j) x_{ij} \right| |p^{-1} \text{sign}(\varepsilon_i + \sum_{j=1}^{m} (\beta_j - b_j) x_{ij})|
\]

Since \( S_{pn} \) is strictly convex for \( p > 1 \), \( S_{pn} \) has a unique minimum, \( \hat{\beta}_{pn} \), and the partial derivatives \( \frac{3S_{pn}(b)}{\partial b_k} \) are negative iff all \( b_k < \hat{\beta}_{pn} \), where \( \hat{\beta}_{pnk} \), \( k=1, \ldots, m \) are the components in \( \hat{\beta}_{pn} \). With the same arguments it follows that \( S_{pn}(b)/\partial b_k \) is positive iff all \( b_k > \hat{\beta}_{pn} \).

Denoting the distribution function of \( \sqrt{n}(\hat{\beta}_{pn} - \beta) \) by \( F_n \), straightforward computations give

\[
F_n(c) = P(\sqrt{n}(\hat{\beta}_{pn} - \beta) < c_k, \ k=1, \ldots, m) =
\]

\[
= P(\hat{\beta}_{pn} - c_k/\sqrt{n} < \beta_{pn}, \ k=1, \ldots, m) =
\]

\[
= P(3S_{pn}(c_k/\sqrt{n} + \beta_{pn})/\partial b_k > 0, \ k=1, \ldots, m) =
\]

\[
= P(-p \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ik} \left| \varepsilon_i - \sum_{j=1}^{m} c_j x_{ij} \right| /\sqrt{n}|p^{-1} \text{sign}(\varepsilon_i - \sum_{j=1}^{m} c_j x_{ij} /\sqrt{n}) > 0, \ k=1, \ldots, m)
\]

Let \( V_n \) be the \( n \)-vector of transformed residuals with components

\[
v_{ni} = |\varepsilon_i - \sum_{j=1}^{m} c_j x_{ij} /\sqrt{n}|p^{-1} \text{sign}(\varepsilon_i - \sum_{j=1}^{m} c_j x_{ij} /\sqrt{n}) \quad i=1, \ldots, n
\]

and let \( U_n \) be the \( m \)-vector defined by

\[
U_n = X'V_n
\]

Denoting the components in \( U_n \) by \( u_{nk} \) we have

\[
F_n(c) = P(u_{nk} < 0, \ k=1, \ldots, m)
\]
From Assumption A5 it is clear that
\[ \frac{1}{p-1} \sum_{i=1}^{n} (p-1) E[|\varepsilon_1|^{p-2}] Q_c \]
will converge in distribution to a normally distributed random vector with expectation zero and covariance matrix \( QE[|\varepsilon_1|^{2p-2}] \). But then
\[ \frac{1}{p-1} \sum_{i=1}^{n} (p-1) Q^{-1} U_n \]
is an asymptotic \( m \)-variate normally distributed random vector with expectation \( c \) and covariance matrix \( \frac{2}{p} Q^{-1} \) and the proof is complete.

Q.E.D.

We will now discuss the applicability of the Asymptotic Distribution Theorem. Consider therefore Assumption A5. From the Lapunov Central Limit Theorem it follows that a normed sum of independent stochastic variables converge to a normally distributed random variable with expectation zero and variance one, if, for \( \delta > 0 \), moments of order \( 2+\delta \) exist and if
\[ \lim_{n \to \infty} \left\{ \sum_{i=1}^{n} b_i \right\}^{1/(2+\delta)} \]
where \( b_i \) and \( c_i \neq 0 \) denote the absolute central moment of order \( 2+\delta \), and variance of the terms, respectively.

Note that if \( F \in D(\alpha) \), moments of order \( h < \alpha \) exist when \( \alpha < 2 \). The moments of order \( 2+\delta \) of \( T_{ik}^{j} \) may be written as
\[ E[|T_{ik}^{j}|^{2+\delta}] = E[|x_{ij}| \varepsilon_i - cx_{ik}/\sqrt{n}]^{P-1} \text{sign}(\varepsilon_i - cx_{ik}/\sqrt{n})^{2p-2+\delta(p-1)} \]
and it is seen that they exist if \( 2p-2+\delta(p-1) < \alpha \). Thus, it is necessary to choose \( p < (\alpha+2)/2 \) when \( F \in D(\alpha), \alpha < 2 \). This condition is also necessary to guarantee the existence of \( E[|\varepsilon_1|^{2p-2}] \). On the other hand, when \( p < 2 \), the expectation \( E[|\varepsilon_1|^{P-2}] \) will be a negative moment and exists only if \( P(\varepsilon_1=0) = 0 \), i.e. only if distribution function is continuous in zero.
To verify the expressions for \( \lim \frac{1}{\sqrt{n}} \sum E[T_{ij}^k] \) and \( \operatorname{Cov}[U_j, U_k] \), we make use of a Taylor expansion for \( cx_{ik}/\sqrt{n} \) near zero. Thus,

\[
\lim \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[T_{ij}^k] = \lim \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[x_{ij} | \varepsilon_i - cx_{ik}/\sqrt{n}]^{p-2} \text{sign}(\varepsilon_i - cx_{ik}/\sqrt{n}) = \\
= \lim \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[x_{ij} | \varepsilon_i]^{p-1} \text{sign}(\varepsilon_i) - (p-1)x_{ij} \frac{cx_{ik}}{\sqrt{n}} |\varepsilon_i|^{p-2} + \frac{1}{\sqrt{n}} 0(1) = \\
= -(p-1)E[|\varepsilon_i|^{p-2}] \lim \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij} x_{ik} c = -(p+1)E[|\varepsilon_i|^{p-2}] q_{jk} c
\]

since \( E[|\varepsilon_i|^{p-2}] = 0 \) by Assumption A4.

In the same manner we obtain

\[
\lim \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V[T_{ij}^k] = \lim \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V[x_{ij} | \varepsilon_i - cx_{ik}/\sqrt{n}]^{p-1} \text{sign}(\varepsilon_i - cx_{ik}/\sqrt{n}) = \\
= \lim \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V[x_{ij} | \varepsilon_i]^{p-1} \text{sign}(\varepsilon_i) + 0(1) = \\
= V[|\varepsilon_i|^{p-1} \text{sign}(\varepsilon_i)] \lim \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{ij}^2 = \\
= (E[|\varepsilon_i|^{2p-2}] - E^2[|\varepsilon_i|^{p-1} \text{sign}(\varepsilon_i)])q_{jj} = \\
= E[|\varepsilon_i|^{2p-2}]q_{jj}
\]

By independence of the residuals, it follows

\[
\lim \operatorname{Cov}[U_i, U_j] = \lim \frac{1}{n} \sum_{i=1}^{n} \operatorname{Cov}[x_{ij} | \varepsilon_i - cx_{ik}/\sqrt{n}]^{p-1} \text{sign}(\varepsilon_i - cx_{ik}/\sqrt{n})/ \\
= \{E[|\varepsilon_i|^{2p-2}]q_{jj}\}^{1/2} \cdot x_{ik} |\varepsilon_i - cx_{ik}/\sqrt{n}|^{p-1} \text{sign}(\varepsilon_i - cx_{ik}/\sqrt{n})/ \\
= \{E[|\varepsilon_i|^{2p-2}]q_{kk}\}^{1/2}
\]
Therefore, when Assumption A1 - A4 are satisfied and $F_\varepsilon \in D(\alpha)$, $F_\varepsilon$ is continuous in zero and $p < (\alpha+2)/2$, we have that $\sqrt{n}(\hat{\beta}_p^{\varepsilon})$ is asymptotically normally distributed with mean zero and covariance matrix $\omega_p^2 Q^{-1}$, where $\hat{\beta}_p^{\varepsilon}$ is the $L_p$-norm estimator and

$$
\omega_p^2 = \begin{cases} 
1/(2F'(0))^2 & \text{for } p = 1 \\
E[|\varepsilon_i|^{2p-2}]/(p-1)E[|\varepsilon_i|^{p-2}]^2 & \text{for } p > 1
\end{cases}
$$

Under the conditions in the Asymptotic Distribution Theorem, it follows that the $L_p$-norm estimator is asymptotically unbiased. The sufficient condition for asymptotically unbiased $L_p$-norm estimators is a certain location condition on the residual distribution, viz. for $p = 1$, the median equals zero and for $1 < p < \infty$, $E[|\varepsilon_i|^{p-1}\text{sign}(\varepsilon_i)] = 0$. In particular, these conditions are satisfied for distributions that are symmetric around zero.

Several conditions for unbiased $L_p$-norm estimators have been proposed. Harvey (1976) showed that if the residual distribution is symmetric around zero and the expectation $E[\hat{\beta}_p^{\varepsilon}]$ exists, then the $L_p$-norm estimators are unbiased. A condition valid even for asymmetric residual distributions was given by Hogan (1976). He showed that the $L_p$-norm estimator is unbiased if for any $\delta \in \mathbb{R}^n$,

$$
E_\delta[\varepsilon] = E[\varepsilon|\varepsilon = \alpha \delta \text{ for some } \alpha \in \mathbb{R}] = 0
$$
and if the expectation $E[\beta_{pn}]$ exists. An interpretation of Hogans condition is that the expectation of the residual vector should be zero in all linear subspaces. However, it is often difficult to decide whether the condition is satisfied for a given residual distribution. Furthermore, the condition is not valid for distributions belonging to $D(\alpha)$ with $\alpha \leq 1$.

Using Harveys proposition, it follows that $L_p$-norm estimators are unbiased for sufficiently small $p$, when the residual distribution is symmetric. We state this result in the next proposition.

**Proposition 4.1** Let $p \geq 1$ be a constant and let Assumptions A1–A5 be satisfied. Then, for a residual distribution, symmetric around zero, the $L_p$-norm estimator is unbiased.

At the first sight, one might expect that the unbiasedness condition should imply a corresponding sampling property for the residuals, i.e. the unbiasedness condition should imply

$$\sum_{i=1}^{n} |\hat{e}_i|^{p-1} \text{sign}(\hat{e}_i) = 0 \quad (4.1)$$

However, that is not the case, which is seen from the Characterization Theorems 3.3 and 3.4. Nevertheless, there are cases when the residuals satisfy (4.1), viz. when $1 < p < \infty$ and there is an intercept in the model. That does not follow from the unbiasedness condition, but from the Characterization Theorem 3.3.

**Corollary 4.1** If there is an intercept in the linear regression model (2.5), the $L_p$-norm estimated residuals satisfy (4.1), when $1 < p < \infty$.

From Corollary 4.1 it follows that the distribution of the estimated residuals of a model with intercept is located according to (4.1). Particularly, it implies that the mean of the least squares residuals equals zero, which is a well-known fact.
4.3 Stochastic regressors

In the Asymptotic Distribution Theorem, a somewhat classical case was considered, viz. the case of non-stochastic regressors and independent and identically distributed residuals. We will now discuss some departures from these classical assumptions. The first of the departures we will investigate is the case of stochastic regressors, which is the topic of this section.

Consider therefore the linear regression model (2.5). We still assume that the residuals are independent and identically distributed. The elements $x_{ij}$ in the matrix $X$ are, however, assumed to be stochastic. Since there always exists an $L_p$-norm estimate (Existence Theorem) and that the estimate is unique for $1 < p < \infty$ (Uniqueness in $L_p$, $1 < p < \infty$ Theorem), the $L_p$-norm estimator is well defined for all realizations of $X$, when $1 < p < \infty$.

The following assumptions are necessary in order to prove statements given in Theorem 4.2.

B1. The model under consideration is that given in (2.5).

B2. The rows $x_{i \cdot}$ in $X$ are independent and identically distributed stochastic vectors and $Q = E[x_{i \cdot} x_{i \cdot}']$ is a positive definite matrix.

B3. The residuals $\varepsilon_{i \cdot}$ are independent and identically distributed with a common distribution function $F_{\varepsilon}$.

B4. $E[x_{1j} | \varepsilon_1 |^{p-1} \text{sign}(\varepsilon_1)] = 0, j = 1, \ldots, m$.

B5. For $p > 1$, $E[|\varepsilon_1|^{p-1} \text{sign}(\varepsilon_1)] = 0$. Furthermore, the expectations $E[x_{1j} x_{1k} | \varepsilon_1 |^{p-2}]$ and $E[x_{1j}^2 | \varepsilon_1 |^{2p-2}]$ exist and the stochastic variables $T_{jk}^{ij}$ defined by

$$T_{jk}^{ij} = x_{ij} | \varepsilon_i - cx_{ik} / \sqrt{n} |^{p-1} \text{sign}(\varepsilon_i - cx_{ik} / \sqrt{n})$$

satisfy the Lindeberg conditions, and
\[ \sum_{i=1}^{n} (T_{ik}^{jk} - E[T_{ik}]) \approx \left\{ \sum_{i=1}^{n} V[T_{ik}^{jk}] \right\}^{1/2} \]

converges to a normally distributed stochastic variable \( U_j \), with expectation zero and variance one, with

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} E[T_{ik}^{jk}] = -(p-1)cE[x_{1j}x_{1k}^2 | \varepsilon_1 |^{p-2}] \]

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} V[T_{ik}^{jk}] = E[x_{1j}^2 | \varepsilon_1 |^{2p-2}] \]

\[ \text{Cov}[U_j, U_k] = \frac{E[x_{1j}x_{1k}^2 | \varepsilon_1 |^{p-2}]}{(E[x_{1j}^2 | \varepsilon_1 |^{2p-2}]E[x_{1k}^2 | \varepsilon_1 |^{2p-2}])^{1/2}} \]

**THEOREM 4.2 (ASYMPTOTIC DISTRIBUTION WITH STOCHASTIC REGRESSORS)** Let \( p > 1 \) be a constant and let Assumptions B1 - B5 be satisfied.

Then \( \sqrt{n}(\hat{\beta} - \beta) \) is asymptotically m-variate normally distributed with mean zero and covariance \( \frac{1}{(p-1)^2} \) \( G^{-1} H G^{-1} \) where \( G \) and \( H \) are m×m-matrices with elements \( \{g_{ij}\} = E[x_{1j}x_{1k}^2 | \varepsilon_1 |^{p-2}] \) and \( \{h_{ij}\} = E[x_{1j}^2 | \varepsilon_1 |^{2p-2}] \), respectively.

**PROOF** With the notations as in the proof of the Asymptotic Distribution Theorem, we still have

\[ F_n(c) = P(U_{nk} < 0, k=1, \ldots, m) \]

Under Assumption B5 it follows that

\[ U_n + (p-1)Gc \]

where \( G \) is the (m×m)-matrix with elements \( \{g_{ij}\} = E[x_{1j}^2 | \varepsilon_1 |^{p-2}] \), will converge in distribution to a normally distributed stochastic vector with expectation zero and covariance matrix \( H \), with elements \( \{h_{ij}\} = E[x_{1j}^2 | \varepsilon_1 |^{2p-2}] \). But then \( \frac{1}{(p-1)} G^{-1} U_n \) is an asymptotically m-variate
normally distributed random vector with expectation \( \mathbf{c} \) and covariance matrix \( \frac{1}{(p-1)^2} \mathbf{G}^{-1} \mathbf{H} \mathbf{G}^{-1} \), and the theorem follows. Q.E.D.

It turns out that when the regressor variables are stochastic, we have to add the "non-correlation" assumption \( E[X_{ij} | \epsilon_1 | \epsilon_1^p \text{sign}(\epsilon_1)] = 0 \).

When the regressor variables and the residuals are independent, the asymptotic covariance matrix of \( \sqrt{n}(\hat{\beta} - \beta) \) reduces to \( \omega_p^2 Q^{-1} \), where \( \omega_p^2 \) is as in Section 4.2, and \( Q \) is as in Assumption B2.

Note that we assumed the regressor variables to obey finite variances and the observations on them are independent. The assumption of independent observations is in many applications not realistic. Consider for example an econometric application where the regressor variables show certain trends. However, the variables used in econometric models are often built up by a first order (and higher) differences, implying that trend effects are, to some extent, reduced. Thus, if the assumption of independent observations on the regressors is not satisfied, it is often possible to transform the observations as to satisfy the independence assumption.

A theorem on asymptotic distribution of \( L_p \)-norm estimators when regressors are stochastic has earlier been given by Ronner (1977). However, the assumptions in that theorem are more restrictive in that the regressors and residuals are assumed to be independent, the residual distribution is assumed to be symmetric and obey a finite variance.

### 4.4 Heteroscedasticity

Heteroscedasticity is in some textbooks defined as the case where the variance of the residuals is not constant. Such a definition is not valid if we allow the variance of the residuals to be infinite; e.g. the residual distribution is fat-tailed. Therefore we have to use a more
general definition of heteroscedasticity. Since there always will exist some kind of scale parameter, we will base our definition on that.

Assume that the residuals in the linear regression model (2.5) are distributed with distribution functions \( F_{\epsilon_i} \), \( i = 1, \ldots, n \), and that there exist a distribution function \( F_\epsilon \) and constants \( \delta_i > 0 \), \( i = 1, \ldots, n \), such that

\[
F_{\epsilon_i}(\frac{1}{\delta_i} x) = F_\epsilon(x) \quad i = 1, \ldots, n
\]

(4.2)

The residuals are still assumed to satisfy the conditions concerning the expectations in the Asymptotic Distribution Theorems 4.1 and 4.2. We will refer to this case as the heteroscedasticity case.

**THEOREM 4.3 (ASYMPTOTIC DISTRIBUTION IN THE HETEROSCEDASTIC CASE)**

Let \( p > 1 \) be a constant and denote the corresponding \( L_p \)-norm estimator with \( \hat{\beta}_{pn} \). Consider the linear regression model (2.5) with heteroscedastic residuals defined by (4.2). For fixed elements in \( X \) according to Assumption A2, assume that stochastic variables with distribution function \( F_\epsilon \) satisfies Assumption A5. Then, \( \sqrt{n}(\hat{\beta}_{pn} - \beta) \) is asymptotically normally distributed with mean zero and covariance matrix \( \frac{2}{p} R^{-1} PR^{-1} \), where \( \omega_p^2 = \text{E}[|\epsilon|^{2p-2}]/((p-1)\text{E}[|\epsilon|^{p-2}])^2 \), \( R = \lim_{n \to \infty} n^{-1} X' S X \), \( P = \lim_{n \to \infty} n^{-1} X' T X \) and \( S \) and \( T \) are \( mxn \) diagonal matrices with elements \( S = \text{diag} \left\{ \frac{1}{\delta_i^{2p-2}} \right\} \), respectively.

For stochastic elements in \( X \) according to Assumption B2, assume that stochastic variables with density function \( F_\epsilon \) satisfies Assumption B4 and B5. Then, \( \sqrt{n}(\hat{\beta}_{pn} - \beta) \) is asymptotically normally distributed with mean zero and covariance matrix \( \frac{2}{p} G^{-1} \text{H} G^{-1} \), where \( \tau_p^2 = \frac{e}{(p-1)^2 d^2} \),

\[
d = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \frac{1}{\delta_i^{2p-2}}, \quad e = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \frac{1}{\delta_i^{2p-2}} \quad \text{and} \quad G \quad \text{and} \quad H \quad \text{are} \quad mxm-\text{matrices with elements} \quad \{g_{ij}\} = \text{E}[x_{ij} x_{ij} | \epsilon|^{p-2}] \quad \text{and} \quad \{h_{ij}\} = \text{E}[x_{ij} x_{ij} | \epsilon|^{2p-2}],
\]

respectively.
We first observe that

\[ x_{ij} | \varepsilon_i - x_{ik} / \sqrt{n} |^{p-1} \text{sign}(\varepsilon_i - x_{ik} / \sqrt{n}) = \]

\[ = \frac{x_{ij}}{\hat{\delta}^p | \hat{\delta}^i |} | \hat{\delta}^i \varepsilon_i - \hat{\delta}^i x_{ik} / \sqrt{n} |^{p-1} \text{sign}(\hat{\delta}^i \varepsilon_i - \hat{\delta}^i x_{ik} / \sqrt{n}) \]

By assumption, \( \hat{\delta}^i \varepsilon_i \), \( i=1,2,\ldots \), are independent and identically distributed stochastic variables, satisfying Assumption A5 in the case of fixed regressors and Assumptions B4 and B5 in the case of stochastic regressors. Therefore, the Asymptotic Distribution Theorems 4.1 and 4.2 are applicable and the result is immediately obtained.

Q.E.D.

In the heteroscedastic case, it is also possible to transform the linear regression model so that the residuals are equally distributed and hence satisfy all assumptions in the Asymptotic Distribution Theorems 4.1 and 4.2. Define therefore the \( nxn \) diagonal matrix \( D \) by

\[ D = \text{diag}\{ \delta_1, \ldots, \delta_n \} . \]  

\[ (4.3) \]

The desired transformation is

\[ Dy = DX\hat{\beta} + D\varepsilon . \]

With the notations \( y^* = Dy \), \( X^* = DX \) and \( \varepsilon^* = D\varepsilon \) we obtain

\[ y^* = X^*\hat{\beta} + \varepsilon^* \]  

\[ (4.4) \]

where the residuals in \( \varepsilon^* \) are independent and equally distributed. The results in the Asymptotic Distribution Theorems 4.1 and 4.2 respectively, are now applicable and we immediately obtain the following Corollary.

COROLLARY 4.2 Consider the linear regression model \( (4.4) \) and let \( p \geq 1 \) be a constant. For fixed elements in \( X \) according to Assumption A2, let Assumptions A3-A5 be satisfied for the elements in the vector \( \varepsilon^* \). Then

\[ \sqrt{n}(\hat{\beta}_{pn} - \beta) \] is asymptotically normally distributed with mean zero and covariance matrix \( \frac{2}{w_{p}} (Q^*)^{-1} \), where \( \hat{\beta}_{pn} \) is the \( L_{p} \)-norm estimator,
\[ \omega_{W,p} = \begin{cases} \left( \frac{\delta_1^2}{2F' (0)} \right)^2 & \text{for } p = 1 \\ \left( \frac{\delta_1^2 E[|\varepsilon_1|^{2p-2}]}{(p-1) E[|\varepsilon_1|^{p-2}]} \right)^2 & \text{for } p > 1 \end{cases} \]

and \( Q_* = \lim_{n \to \infty} n^{-1} X'D'DX \).

For stochastic elements in \( X \) according to Assumption B2, let \( p > 1 \) and let Assumptions B3 - B5 be satisfied for the elements in the vector \( \varepsilon_* \).

Then \( \sqrt{n}(\hat{\beta} - \beta) \) is asymptotically normally distributed with mean zero and covariance matrix \( \tau_{X,p}^{-2} G^{-1} \Gamma^{-1} G^{-1} \), where

\[ \tau_{W,p}^2 = \frac{\delta_1^2}{(p-1)^2 r} \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \delta_i^2 \]

and \( G \) and \( H \) are \( m \times m \) matrices with elements \( \{g_{ij}\} = E[x_i x_j | \varepsilon_i |^{p-2}] \), and \( \{h_{ij}\} = E[x_i^2 x_j^2 | \varepsilon_i |^{2p-2}] \), respectively.

### 4.5 Linearly dependent residuals

In this section, we will discuss the case of linearly dependent residuals in the linear regression model (2.5). By linearly dependent residuals we mean that there exists a non-singular \( (n \times n) \)-matrix \( M \) such that

\[ \varepsilon_* = M \varepsilon \quad (4.5) \]

is a vector of independent and equally distributed residuals.

At first sight, the notion of linear dependence may seem too restrictive to be of practical importance. However, when the residuals possess an ARMA-process, it may be shown that they are linearly dependent. Another interesting case of linear dependence appears when the residual vector has a multivariate symmetric stable distribution.

As an example of linearly dependent residuals, we will now in some more detail discuss the case of a first order autoregressive residual process with increments following a symmetric stable distribution. Assume therefore that

\[ \varepsilon_i = \rho \varepsilon_{i-1} + \delta_i, \quad i = 1, \ldots, n \]
where \( |\rho| < 1 \) is a constant, \( \delta_1, i=1, \ldots, n \) are independent and identically distributed random variables, and \( \varepsilon_i, i=1, \ldots, n \) are the residuals in the linear regression model (2.5). Assume further that the common distribution of \( \delta_i \) is symmetric stable, with characteristic exponent \( \alpha \), and that \( \varepsilon_0 = 0 \).

To characterize the simultaneous distribution of the residual vector \( \varepsilon \), we are looking for its characteristic function. It follows that

\[
\varphi_\varepsilon (s) = E[\exp(is\varepsilon)] =
\]

\[
= E[\exp(i(s_1 \varepsilon_1 + \ldots + s_n \varepsilon_n))] =
\]

\[
= E[\exp(i(s_1 \delta_1 + s_2 (\rho \delta_1 + \delta_2) + \ldots + s_n (\rho^{n-1} \delta_1 + \ldots + \rho \delta_{n-1} + \delta_n)))] =
\]

\[
= E[\exp(i(s_1 + s_2 \rho + s_3 \rho^2 + \ldots + s_n \rho^{n-1}) \delta_1)] \ldots =
\]

\[
= E[\exp(i(s_{n-1} + s_n \rho) \delta_{n-1})] E[\exp(is_n \delta_n)] =
\]

\[
= \varphi_{\delta_1} (s_1 + s_2 \rho + s_3 \rho^2 + \ldots + s_n \rho^{n-1}) \ldots \varphi_{\delta_{n-1}} (s_{n-1} + s_n \rho) \varphi_{\delta_n} (s_n)
\]

where the last equality follows from the assumption that the random variables \( \delta_i, i=1, \ldots, n \) are independent. Now, from Theorem 2.2, The Characteristic Functions of Stable Distributions, we have that

\[
\ln \varphi_{\delta_i} (s) = -|s|^\alpha, \quad i=1, \ldots, n
\]

if we assume that \( \delta_i, i=1, \ldots, n \) are symmetric about zero and the scale parameter equals one. Since the scale parameter only is a multiplicative constant, there is no restriction to assume it to equals one. Putting this together, we obtain

\[
\ln \varphi_\varepsilon (s) = \ln \varphi_{\delta_1} (s_1 + s_2 \rho + s_3 \rho^2 + \ldots + s_n \rho^{n-1}) + \ldots +
\]

\[
+ \ln \varphi_{\delta_{n-1}} (s_{n-1} + s_n \rho) + \ln \varphi_{\delta_n} (s_n) =
\]

\[
= -|s_1 + s_2 \rho + s_3 \rho^2 + \ldots + s_n \rho^{n-1}|^\alpha + \ldots - |s_{n-1} + s_n \rho|^\alpha - |s_n|\alpha
\]

(4.6)

The characteristic functions of the marginal distributions \( \varphi_{\varepsilon_i} (s_i) \) are now seen to be
\[ \ln \varphi_{\varepsilon_i}(s_i) = - s_i \rho^{i-1} |\alpha| - s_i \rho^{i-2} |\alpha| - \cdots - s_i \rho |\alpha| - s_i |\alpha| = - (|\rho|^{i-1} + \cdots + |\rho|^{i-2} + \cdots + |\rho| + 1) |s_i|^{\alpha} \]

and it follows from Theorem 2.2, The Characteristic Functions of Stable Distributions, that the marginal distributions of the residual vector \( \varepsilon \) are symmetric stable with characteristic exponent \( \alpha \) and scale parameter \( |\rho|^{i-1} |\alpha| + |\rho|^{i-2} |\alpha| + \cdots + |\rho| + 1 \). The result that the marginal distributions are stable is not surprising since \( \varepsilon_i \) is a linear combination of independent and identically distributed stable variables.

Formula (4.6) is the characteristic function of \( \varepsilon \). To see that (4.6) is equivalent to the expression in Theorem 2.4, The Characteristic Functions of Multivariate Stable Distributions, we note that

\[ \ln \varphi_{\varepsilon}(s) = - ((s_1 + s_2 \rho + s_3 \rho^2 + \cdots + s_n \rho^{n-1})^2)^{\alpha/2} - \cdots - ((s_n + s_n \rho)^2)^{\alpha/2} - ((s_n)^2)^{\alpha/2} \]

But the quadratic expressions in the parenthesis may be written in matrix form as

\[ (s_1 + s_2 \rho + s_3 \rho^2 + \cdots + s_n \rho^{n-1})^2 = s' \Delta(1) s \]

\[ \vdots \]

\[ (s_{n-1} + s_n \rho)^2 = s' \Delta(n-1) s \]

\[ (s_n)^2 = s' \Delta(n) s \]

where \( \Delta(i) \) is a symmetric (\( n \times n \)) matrix where the elements in the first i-1 rows and columns equals zero and

\[ \Delta(i) = \begin{bmatrix}
0 & \ldots & 0 & 0 & \ldots & 0 \\
\vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\
0 & 1 & 1\rho & 1\rho^2 & \ldots & 1\rho^{n-i} \\
0 & \frac{1}{\rho} & \frac{1}{\rho^2} & \frac{1}{\rho^3} & \ldots & \frac{1}{\rho^{n-i}} \\
0 & \frac{1}{\rho^2} & \frac{1}{\rho^3} & \frac{1}{\rho^4} & \ldots & \frac{1}{\rho^{n-i+1}} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
0 & \frac{1}{\rho^{n-i}} & \frac{1}{\rho^{n-i+1}} & \frac{1}{\rho^{n-i+2}} & \ldots & \rho^{2(n-i)} 
\end{bmatrix} \]
Therefore, the characteristic function of $\epsilon$ may be written as

$$\ln \varphi_\epsilon(s) = - \frac{1}{2} \sum_{i=1}^{n} (s'A(i)s)^{\alpha/2}$$

Finally, for a constant $c$, we have that

$$\frac{1}{n} \sum_{i=1}^{n} (c(s'A(i)s))^{\alpha/2} = \frac{1}{n} \sum_{i=1}^{n} (c^2(s'A(i)s))^{\alpha/2} = |c|^\alpha \sum_{i=1}^{n} (s'A(i)s)^{\alpha/2}$$

Thus, $\epsilon$ has a multivariate symmetrical stable distribution with characteristic exponent $\alpha$ and a characteristic function given by (4.6) or equivalently (4.7)

To see that the elements in the residual vector $\epsilon$ are linearly dependent, we define the $(nxn)$-matrix $M$ as

$$M = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ -\rho & 1 & 0 & \ldots & 0 & 0 \\ 0 & -\rho & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -\rho & 1 \end{bmatrix}$$

and show that the elements in $\epsilon^*$ defined in (4.5) are independent.

Using (4.6) the characteristic function of the distribution of $\epsilon^*$ is given by

$$\varphi_{\epsilon^*}(s) = \varphi_{\epsilon M}(s) = \varphi_{\epsilon}(M's) =$$

$$= \exp(-|s_1-\rho s_2+(s_2-\rho s_3)s_3^2+\ldots+\rho^{n-1}s_n|^\alpha -$$

$$- \ldots - |s_{n-1}-\rho s_n|^\alpha - |s_n|^\alpha) =$$

$$= \exp(-|s_1|^\alpha \ldots - |s_{n-1}|^\alpha - |s_n|^\alpha) =$$

$$= \exp(-|s_1|^\alpha) \ldots \exp(-|s_{n-1}|^\alpha) \exp(-|s_n|^\alpha)$$

that is, it is possible to write the characteristic function of the distribution of $\epsilon^*$ as the product of the characteristic functions of the
distributions of $\varepsilon^*_1, i=1,\ldots,n$. Therefore $\varepsilon^*_1, i=1,\ldots,n$ are independent. Furthermore, the characteristic functions of $\varepsilon^*_1, i=1,\ldots,n$, are all equal, implying that the elements in $\varepsilon^*$ are identically distributed.

Note that it is possible to obtain the same result by using (4.7) and the equality

$$(M's)'\Delta(i)(M's) = s'M\Delta(i)Ms = \frac{s^2}{i}$$

In the case of independent residuals, we know from the Asymptotic Distribution Theorems, that the $L_1$-norm estimator is asymptotically normally distributed for a large class of distributions. One case of dependent residuals was studied in the Monte Carlo experiment reported in Nyquist (1979) and summarized in Subsection 2.6.4. In the experiment, the residuals followed a first order autoregressive process with stable increments. The length of the time series was $T=30$.

As is seen in Table 4.1, the observed mean squared errors of the $L_1$-norm estimator were increasing when the degree of dependence was increasing. For very fat-tailed distributed and dependent residuals, the estimator showed an extremely high mean square error. It is difficult to decide whether that is caused by a divergency or an extremely slow convergency. In any way, in applications with time series of a short or moderate length, the $L_1$-norm estimator appears to be unsatisfactory, when the residuals are dependent and fat-tailed. Further on, there is no reason to believe that any other $L_p$-norm estimator should behave better than the $L_1$-norm estimator, in the context of dependent and, simultaneously, fat-tailed distributed residuals.

When the residuals are linearly dependent, we may multiply the regression model (2.5) from left by the matrix $M$ defined in (4.5). Thus,

$$My = MX\beta + Me$$
Table 4.1  Observed mean squared errors of the $L_1$-norm estimator of the model $y_t = \beta_0 + \beta_1 X_t + \epsilon_t + \epsilon_t = \rho \epsilon_{t-1} + \delta_t$ and $\delta_t$ has a symmetric stable distribution with characteristic exponent $\alpha$, $t=1, \ldots, 30$. The data is from the Monte Carlo experiment reported in Nyquist (1979).

With the notations $y^* = My$, $X^* = MX$ and $\epsilon^* = Me$ we obtain

$$y^* = X^* \beta + \epsilon^*$$

(4.8)

where the residuals in $\epsilon^*$ are independent and equally distributed. The results in the Asymptotic Distribution Theorems 4.1 and 4.2, respectively, are now applicable. We state the results in the following Corollary.

**COROLLARY 4.3** Consider the linear regression model (4.8) and let $p \geq 1$ be a constant. For fixed elements in $X$ according to Assumption A2, let Assumptions A3–A5 be satisfied for the elements in the vector $\epsilon^*$. Then, $\sqrt{n}(\hat{\beta}_{pn} - \beta)$ is asymptotically normally distributed with mean zero and covariance matrix $\omega^2 G_{p,n}(Q^*)^{-1}$, where $\hat{\beta}_{pn}$ is the $L_p$-norm estimator.
\[
\omega_{G,p}^2 = \begin{cases} 
\frac{1}{(2F'(0))^2} & \text{for } p = 1 \\
\frac{E[|\epsilon^*|^2 p - 2]}{(p-1)E[|\epsilon^*_1|^{p-2}]} & \text{for } p > 1
\end{cases}
\]

and \( Q^* = \lim_{n \to \infty} n^{-1}X'M'MX \).

For stochastic elements in \( X \) according to Assumption B2, let \( p > 1 \) and let Assumptions B3 - B5 be satisfied for the elements in the vector \( \epsilon^* \). Then \( \sqrt{n}(\hat{\beta}_{pn} - \beta) \) is asymptotically normally distributed with mean zero and covariance matrix \( \frac{1}{(p-1)^2} (G^*)^{-1}H^*(G^*)^{-1} \), where \( G^* \) and \( H^* \) are matrices with elements \( \{g^*_{ij}\} = E[x^*_i x^*_j | \epsilon^*_1]^{p-2} \) and \( \{h^*_{ij}\} = E[x^*_i x^*_j | \epsilon^*_1]^{2p-2} \), respectively, and \( x^*_{ij} \) and \( \epsilon^*_{ij} \) are elements in the matrix \( X^* = MX \) and the vector \( \epsilon^* = M\epsilon \), respectively.

The Corollary tells us that for linearly dependent residuals, the \( L_p \)-norm estimator may be suitable after a transformation of the model. To study the effects of the transformation a Monte Carlo experiment was performed. The study is reported in Coursey and Nyquist (1980). We will now give a short summary of some of the results in that study.

The design of the experiment was similar to that in Nyquist (1979). Thus, the model had one regressor variable and an intercept. The residuals were generated according to a first order autoregressive process with symmetrically stable distributed increments. \( N = 100 \) replicates were performed, each with a time series length of \( T = 15 \) and \( 30 \).

In table 4.2 the observed variances of the \( L_1 \)-norm estimator of the original as well as of the transformed model are shown. When estimating the original model, the results are similar to those reported in Nyquist...
(1979), i.e. the observed variances are extremely high for dependent fat-tailed distributed residuals. A notable difference in observed variances between the two studies appears in the case $\alpha = 0.5$. In Nyquist (1979), the observed variances of the OLA estimator of $\beta_0$ and $\beta_1$ were $1.76 \cdot 10^{10}$ and $3.46 \cdot 10^8$, respectively, for $\rho = 0.5$, while in Coursey and Nyquist (1980) the observed variances were 371 and 23.5, respectively, for $\rho = -0.5$ and $T = 30$. It is difficult to decide whether that difference depends on the difference in the autoregressive parameter $\delta$, on other differences in the models under consideration or on a sampling error due to the relatively small number of replicates in the experiments.

When estimating the transformed model, considerably smaller variances are observed and the GLA-estimator performs well. Comparing the results from the two sample sizes, it appears that the observed variances are smaller in all cases but one, for the larger sample size. For the OLA estimator, this may be taken as an indicator of a convergency. If there is a convergency, it seems to be slow.
Table 4.2 Observed variances of \( L_1 \)-norm based estimators of the model
\[
y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad \varepsilon_t = \rho \varepsilon_{t-1} + \delta_t
\]
and \( \delta_t \) has a symmetric stable distribution with characteristic exponent \( \alpha \), \( t=1,\ldots,T \). OLA refers to \( L_1 \)-norm estimation of the original model, while GLA refers to \( L_1 \)-norm estimation of the transformed model as defined in Corollary 4.4. The data is from the Monte Carlo experiment reported in Coursey and Nyquist (1980).

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>Parameter</th>
<th>Estimator</th>
<th>( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \beta_0 )</td>
<td>OLA and GLA</td>
<td>2.0</td>
</tr>
<tr>
<td>0</td>
<td>OLA</td>
<td>0.70</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>GLA</td>
<td>0.090</td>
<td>0.093</td>
</tr>
<tr>
<td>-0.5</td>
<td>OLA</td>
<td>0.82</td>
<td>1.44</td>
</tr>
<tr>
<td></td>
<td>GLA</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>-0.9</td>
<td>OLA</td>
<td>0.086</td>
<td>0.157</td>
</tr>
<tr>
<td></td>
<td>GLA</td>
<td>0.042</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>( \beta_1 )</td>
<td>OLA and GLA</td>
<td>7.56</td>
</tr>
<tr>
<td></td>
<td>OLA</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>GLA</td>
<td>0.026</td>
<td>0.029</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>Parameter</th>
<th>Estimator</th>
<th>( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \beta_0 )</td>
<td>OLA and GLA</td>
<td>2.0</td>
</tr>
<tr>
<td>0</td>
<td>OLA</td>
<td>0.36</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>GLA</td>
<td>0.011</td>
<td>0.014</td>
</tr>
<tr>
<td>-0.5</td>
<td>OLA</td>
<td>0.36</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>GLA</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td>-0.9</td>
<td>OLA</td>
<td>0.011</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>GLA</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>( \beta_1 )</td>
<td>OLA and GLA</td>
<td>2.47</td>
</tr>
<tr>
<td></td>
<td>OLA</td>
<td>1.4</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>GLA</td>
<td>0.078</td>
<td>2.09</td>
</tr>
<tr>
<td></td>
<td>GLA</td>
<td>0.004</td>
<td>0.004</td>
</tr>
</tbody>
</table>
4.6 The optimal Lp-norm estimator

In the last sections, the asymptotic distributions of Lp-norm estimators have been investigated for several sets of assumptions. Roughly, we have that for p sufficiently small, \( \sqrt{n}(\hat{\beta}_{pn} - \beta) \) is asymptotically normally distributed, where \( \hat{\beta}_{pn} \) is the Lp-norm estimator. When the residuals are independent and equally distributed and, in the case of stochastic regressors, where residuals and regressors are independent, the asymptotic variance of \( \sqrt{n}(\hat{\beta}_{pn} - \beta) \) equals \( \omega^2_p (E[Q])^{-1} \), where \( Q = \lim_{n \to \infty} n^{-1}X'X \) and \( \omega^2_p \) is related to the residual distribution. For a residual distribution \( F \in D(\alpha) \) with for p=1 positive derivative in zero, we have that

\[
\omega^2_p = \begin{cases} 
\frac{(2F'(0))^{-2}}{E[|\varepsilon_1|^{2p-2}]/((p-1)^2 E^2[|\varepsilon_1|^{p-2}])} & \text{for } p=1 \text{ and fixed regressors} \\
\text{for } 1 < p < \infty 
\end{cases}
\]

provided that the expectations exist. In cases with heteroscedastic and/or linearly dependent residuals, it is possible to transform the model as to satisfy the assumptions of equal distributions and independence. Thus, the asymptotic variance depends on p only through \( \omega^2_p \). In this section we will, for a fixed residual distribution, consider \( \omega^2_p \) as a function of p. If \( \omega^2_p \) has a global minimum, we define the optimal Lp-norm estimator as that Lp-norm estimator which corresponds to that p for which \( \omega^2_p \) attains its global minimum. The aim of this section is just to present the ideas and not to give a complete analytical treatment of the optimal Lp-norm estimator. The very few distributions considered will serve as illustrations to our discussion. Therefore they will not constitute a complete list of possible and/or important residual distributions.

If the residual distribution is known up to a scale parameter \( \delta \), it is immediate that \( \omega^2_p = \delta^2 \omega^2_{*,p} \), where \( \omega^2_{*,p} \) corresponds to the scale parameter \( \delta = 1 \). Therefore, the optimal Lp-norm estimator will be the same irrespective the scale parameter.
Let us first consider the class of distributions, where there is an $L_p$-norm estimator equivalent to the maximum likelihood estimator. This class is defined in Section 3.9. The frequency function is given by

$$f(u) = \begin{cases} 
\frac{\gamma}{2\delta \Gamma \left(\frac{1}{\gamma}\right)} \exp\left(-\frac{|u|^\gamma}{\delta}\right) & \text{for } 1 \leq \gamma < \infty \\
\frac{1}{2\delta}; \quad |u| \leq \delta & \text{for } \gamma = \infty
\end{cases}$$

There is no restriction to assume the scale parameter $\delta$ equal to one.

We will now compute $\omega_p^2$ as a function of $p$ for members of that class of distributions. Obviously we have

$$f(0) = \begin{cases} 
\frac{\gamma}{2\Gamma \left(\frac{1}{\gamma}\right)} & \text{for } 1 \leq \gamma < \infty \\
\frac{1}{2} & \text{for } \gamma = \infty
\end{cases}$$

For $1 \leq \gamma < \infty$ we have that

$$E[|u|^{2p-2}] = \int_{-\infty}^{\infty} |u|^{2p-2} \frac{\gamma}{2\Gamma \left(\frac{1}{\gamma}\right)} e^{-|u|^\gamma} du = \frac{\gamma}{\Gamma \left(\frac{1}{\gamma}\right)} \int_{0}^{\infty} u^{2p-2} e^{-u^\gamma} du$$

Now, let $t = u^{\gamma}$, which implies that $u = t^{1/\gamma}$ and $du = \frac{1}{\gamma} t^{1/\gamma - 1} dt$. This change of variables gives

$$E[|u|^{2p-2}] = \frac{\gamma}{\Gamma \left(\frac{1}{\gamma}\right)} \int_{0}^{\infty} (t^{1/\gamma})^{2p-2} e^{-t^{1/\gamma}} \frac{1}{\gamma} t^{1/\gamma - 1} dt = \frac{1}{\Gamma \left(\frac{1}{\gamma}\right)} \int_{0}^{\infty} t^{2p-1} e^{-t} dt = \frac{\Gamma(2p-1)}{\Gamma \left(\frac{1}{\gamma}\right)}$$
In the same way we obtain

\[ E[|u|^{p-2}] = \frac{\Gamma(\frac{p-1}{\gamma})}{\Gamma(\frac{1}{\gamma})} \]

Putting this together, we find that

\[ \omega_p^2 = \begin{cases} \frac{\Gamma(\frac{1}{\gamma})^2}{\Gamma(\frac{p-1}{y})/(\gamma)} & \text{for } p=1 \\ \frac{\Gamma(\frac{2p-1}{y})/\Gamma(\frac{1}{\gamma})/((p-1)\Gamma(\frac{p-1}{y}))^2} & \text{for } 1 < p < \infty \end{cases} \]

and \( 1 \leq \gamma < \infty \). For the uniform distribution, i.e. \( \gamma=\infty \), we have for \( p > 1 \)

\[ E[|u|^{2p-2}] = \int_{-1}^{1} |u|^{2p-2} \frac{1}{2} \, du = \left[ \frac{1}{2p-1} u^{2p-1} \right]_0^1 = \frac{1}{2p-1} \]

and

\[ E[|u|^{p-2}] = \int_{-1}^{1} |u|^{p-2} \frac{1}{2} \, du = \left[ \frac{1}{p-1} u^{p-1} \right]_0^1 = \frac{1}{p-1} \]

so

\[ \omega_p^2 = \frac{1}{2p-1} \quad \text{for } 1 \leq p < \infty \]

Graphs of \( \omega_p^2 \) in the cases \( \gamma=1 \) (Laplace distribution), \( \gamma=2 \) (normal distribution) and \( \gamma=\infty \) (uniform distribution) are shown in Figure 4.1.

For the Laplace distribution, \( \omega_p^2 \) is strictly increasing in \([1,\infty[\) and thus attains its minimum when \( p=1 \). That is exactly what we might expect, since the \( L_1 \)-norm estimator is equivalent to the maximum likelihood estimator when residuals are Laplace distributed. Thus, for Laplace distributed residuals, the \( L_1 \)-norm estimator is the optimal \( L_p \)-norm estimator.

When residuals are uniformly distributed, \( \omega_p^2 \) is strictly decreasing in \([1,\infty[\). Therefore, the larger \( p \), the smaller asymptotic variance of the \( L_p \)-norm estimator. Recall that in this case, the \( L_\infty \)-norm estimator is equivalent to the maximum likelihood estimator. The \( L_\infty \)-norm estimator is therefore the optimal \( L_p \)-norm estimator in this case.
Figure 4.1 Graphs of $\omega_p^2$ as a function of $p$ for the uniform [-2,2], normal (0,1) and Laplace (double exponential) distribution.
From the graphs in Figure 4.1 it also appears that for the normal distribution, $\omega_p^2$ is U-shaped with a minimum at $p=2$. That is not surprising, since the least squares estimator is equivalent to the maximum likelihood estimator in that case. For normally distributed residuals, the least squares estimator is thus the optimal $L_p^\pi$-norm estimator.

The class of contaminated distributions was defined in Subsection 2.4.4. For members in that class, the probability density function is assumed to be of the form

$$f(u) = \lambda f_1(u) + (1-\lambda) f_2(u)$$

where $\lambda$ is a constant in the interval $[0,1]$, $f_1$ is the density of (e.g.) the standard normal distribution and $f_2$ is the density of an arbitrary distribution. Here we will assume that $f_1$ is the density of the $N(0,1)$ and $f_2$ is the density of the $N(0,\sigma^2)$.

Therefore,

$$\omega_1^2 = \frac{1}{4(\lambda\frac{1}{\sqrt{2\pi}} + (1-\lambda) \frac{1}{\sigma\sqrt{2\pi}})^2} = \frac{\pi\sigma^2}{2(\lambda\sigma + (1-\lambda))^2}$$

and for $1 < p < \infty$

$$\omega_p^2 = \frac{\lambda + (1-\lambda)\sigma^{2p-2}}{(\lambda + (1-\lambda)\sigma^{p-2})^2} \omega_p^*$$

where $\omega_p^*$ is the asymptotic variance of the $L_p^\pi$-norm estimator when the residuals are standard normally distributed. Thus,

$$\omega_p^* = \frac{\Gamma\left(\frac{2p-1}{2}\right) \sqrt{\pi} \sqrt{2}}{(p-1)^2 \left(\Gamma\left(\frac{p-1}{2}\right)\right)^2}$$

and

$$\omega_p^2 = \frac{\lambda + (1-\lambda)\sigma^{2p-2}}{(\lambda + (1-\lambda)\sigma^{p-2})^2} \frac{\Gamma\left(\frac{2p-1}{2}\right) \sqrt{2\pi}}{(p-1)^2 \left(\Gamma\left(\frac{p-1}{2}\right)\right)^2}$$

for $1 < p < \infty$. 

Graphs of the function $\omega_p^2$ for some members of this class of contaminated distributions are shown in Figure 4.2. It appears that the optimal $L_p$-norm estimator corresponds to a smaller $p$ when $\sigma^2$ (the variance of the "bad" observations) and when $1 - \lambda$ is increasing (the probability of making a "bad" observation).

It should be very interesting to compute the asymptotic variances of the $L_p$-norm estimators when the residuals are stable distributed. However, since the frequency function is known only for a few members, it is difficult to compute $\omega_p^2$ in general. For the normal distribution we have already derived $\omega_p^2$. We will now derive $\omega_p^2$ for the Cauchy distribution ($\alpha=1$). For $p=1$, we have

$$\omega_1^2 = \frac{\pi^2}{4}$$

and, for $1 < p < 1.5$, we have

$$E[|u|^{2p-2}] = \int_{-\infty}^{\infty} |u|^{2p-2} \frac{1}{\pi(1+u^2)} \, du = \frac{2}{\pi} \int_{0}^{\infty} u^{2p-2} \frac{1}{1+u^2} \, du = \frac{1}{\sin \left( \frac{(2p-1)\pi}{2} \right)}$$

$$E[|u|^{p-2}] = \int_{-\infty}^{\infty} |u|^{p-2} \frac{1}{\pi(1+u^2)} \, du = \frac{2}{\pi} \int_{0}^{\infty} u^{p-2} \frac{1}{1+u^2} \, du = \frac{1}{\sin \left( \frac{(p-1)\pi}{2} \right)}$$

so

$$\omega_p^2 = \frac{1}{\sin \left( \frac{(2p-1)\pi}{2} \right)} \left( \frac{(p-1)}{\sin \left( \frac{(p-1)\pi}{2} \right)} \right)^2 = \frac{\sin^2 \left( \frac{(p-1)\pi}{2} \right)}{(p-1)^2 \sin \left( \frac{(2p-1)\pi}{2} \right)}$$

for $1 < p < 1.5$.
a) $\sigma = 2$.

b) $\sigma = 5$. 

---

The graphs show the relationship between $\omega^2, P, \lambda$ for different values of $\sigma$. The curves represent various values of $\lambda$ for $\sigma = 2$ and $\sigma = 5$. The graphs are used to illustrate the behavior of the system under different parameters.
The graph of $\omega_p^2$ in the case of Cauchy distributed residuals is shown in Figure 4.3. $\omega_p^2$ exists only for $1 < p < 1.5$ and has minimum in $p=1$. Therefore, the $L_p$-norm estimator is the optimal $L_p$-norm estimator.

When the residuals are Cauchy distributed, there is no $L_p$-norm estimator that is equivalent to the maximum likelihood estimator. In the Monte Carlo study reported in Kadyiala and Murthy (1977) and summarized in Subsection 2.6.2, the $L_1$-norm estimator and the maximum likelihood estimator
was compared in the case of Cauchy distributed residuals. In that study Kadyiala and Murthy found that the maximum likelihood estimator was slightly better.

From the examples we now have considered, it is clear that where there is an $L_p$-norm estimator, equivalent to the maximum likelihood estimator, the asymptotic variance of the optimal $L_p$-norm estimator attains the Cramér-Rao lower bound. Whether the asymptotic variance of the optimal $L_p$-norm estimator attains or is near the Cramér-Rao lower bound for other distributions is still an open question. However, the question is of most
interest, because if that is the case for a larger class of distributions, the optimal $L_p$-norm estimator will have a relatively small variance, when the residual distribution belongs to that class. Note that a minimum variance unbiased estimator does not always exist. E.g. in the Cauchy distribution case, there does not exist any sufficient statistic, implying that the Cramér-Rao lower bound may not be attained for any estimator.

Finally, we have to stress that all discussions in this section have been based on a known residual distribution. Recall also that $p$ is assumed fix in the Distribution Theorems. Therefore, the theory will have dubious interpretations if the choice of $p$ is based on an estimate of the residual distribution. When the residual distribution is a member of a well-defined class of distributions, $p$ may be chosen by a minimax procedure. We will now illustrate this idea by an example. Assume therefore that the residuals are either normally or Cauchy distributed. The first step in the proposed procedure is to determine the optimal $L_p$-norm estimator for each member in the class of possible distributions. In the case of normally distributed residuals, we know that the $L_2$-norm estimator is optimal. When the residuals are Cauchy distributed, the $L_1$-norm estimator is optimal. In the second step, the asymptotic variances of each optimal $L_p$-norm estimator is computed for each possible residual distribution. For our example, see table 4.3.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$L_1$</th>
<th>$L_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cauchy</td>
<td>2.47</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$N(0,1)$</td>
<td>1.27</td>
<td>1.00</td>
</tr>
<tr>
<td>max</td>
<td>2.47</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

Table 4.3 Choice of estimator when residuals are either Cauchy or normally distributed.
In the third step, the maximum of the variances for each estimator is determined. When applying the $L_1$-norm estimator, the worst that could happen is that the residuals are Cauchy distributed. In that case, the asymptotic variance equals $2.47$. For the least squares estimator, the variance does not exist in the Cauchy case. The estimator with the least maximum is now chosen. Thus, when residuals are either Cauchy or normally distributed, the $L_1$-norm estimator should be applied.

When there is no information available concerning the residual distribution, the class of possible distributions is $\mathcal{D} = \{D(\alpha), 0 < \alpha \leq 2\}$. For any $L_p$-norm estimator with $p > 1$, there exists a distribution belonging to $\mathcal{D}$ such that the asymptotic variance will be infinite. On the other hand, the $L_1$-norm estimator will always have finite asymptotic variance (and furthermore be asymptotically normally distributed). Therefore, due to the minimax procedure, the $L_1$-norm estimator should be applied, when no information concerning the residual distribution is available.

4.7 Estimation of the variance

From the Distribution Theorem, we know that for $p$ sufficiently small, $\sqrt{n}(\hat{\beta}_{pn} - \beta)$ is asymptotically normally distributed, where $\hat{\beta}_{pn}$ is the $L_p$-norm estimator of $\beta$ in the linear regression model (2.5). When the residuals are independent and equally distributed and, in the case of stochastic regressors, residuals and regressors are independent, the asymptotic variance of $\sqrt{n}(\hat{\beta}_{pn} - \beta)$ equals $\omega_p^2(E[Q])^{-1}$, where $Q = \lim_{n \to \infty} n^{-1}X'X$ and $\omega_p^2$ is related to the residual distribution. In this section we will give some notes on the estimation of $\omega_p^2(E[Q])^{-1}$.

The matrix $E[Q]$ can be consistently estimated by means of the sample
moment matrix \( n^{-1}x'x \). The estimation of \( \omega^2_p \) is more complicated.

For \( p=1 \) an estimator of \( \omega^2_1 = (2f(0))^2 \) has been suggested in Ronner (1977). The idea behind that estimator is that \( f(0) \) may be consistently estimated by the number of residuals \( \varepsilon_i \) in a neighbourhood of zero and that the observed residuals \( \hat{\varepsilon}_i = y_i - \sum_{j=1}^{m} \hat{\beta}_{n} x_{ij} \) converges in probability to \( \varepsilon_i \). Thus, letting \( \frac{n}{n} \) denote the number of observed residuals \( \hat{\varepsilon}_i \) in the interval \([-n^{-1/4}, n^{-1/4}] \), it may be shown that

\[
\omega^2_p = (2n^{3/4})^{-1} n \frac{N_n}{S_D}
\]

where

\[
S_N = \sum_{i=1}^{n} |y_i - \sum_{j=1}^{m} \hat{\beta}_{n} x_{ij}|^{2p-2}
\]

and

\[
S_D = \sum_{i=1}^{n} |y_i - \sum_{j=1}^{m} \hat{\beta}_{n} x_{ij}|^{p-2}
\]

The distribution properties of these estimators are still unknown and we do not know whether the estimators perform well in practice. However, when inferences are to be drawn in cases where \( \omega^2_p \) is unknown, it is of great importance that \( \omega^2_p \) may be estimated and that the distribution properties of the estimator are known. Therefore, the construction of an estimator of \( \omega^2_p \) and deduction of its distribution properties is, for the moment, one of the most important areas for further research in \( L_p \)-norm estimation.
4.8 Geometric interpretations

In Section 3.6 we gave some geometric interpretations of $L_p$-norm estimation. We will now continue by giving some more geometric interpretations of the notions in the preceding sections. Since we have to restrict our figures to two dimensions, we can only study the unrealistic case of two observations. However, generalizations to $n$ dimensions are immediate.

To fix our notations, let us first look at a realization of our data. See Figure 4.4 for an illustration of the $L_p$ space for $1 < p < 2$. The $L_p$-norm estimate $\hat{y}$ of $y$ is that vector in the subspace $G$ that is closest to $y$.

Let us first discuss the case of heteroscedastic residuals. One way to treat the problem is to transform the model;

$$ Dy = DX\beta + D\varepsilon $$

where $D$ is a diagonal $(mxm)$-matrix defined by (4.3). After the transformation an $L_p$-norm estimator is applied. Geometrically, we have the situation as shown in Figure 4.4 after the transformation. To see the situation before transformation, we multiply the vectors in Figure 4.4 by $D^{-1} = \text{diag}(1/\delta_1, 1/\delta_2)$. The result is shown in Figure 4.5. Those observations with a small scale parameter in the residual distribution

---

**Figure 4.4** The vector space $L_p$ for $1 < p < 2$. 
will have a small \( \delta \)-coefficient in the D-matrix. Therefore, the "L\( p \) unit ball" will be stretched out in directions corresponding to observations with "high precision". The "L\( p \) unit ball" will be more like an "L\( p \) egg" as is indicated in Figure 4.5.

In the case of linearly dependent residuals, the transformed model

\[
My = MX\beta + Me
\]

where \( M \) is an (nxn)-matrix defined by (4.5), was L\( p \)-norm estimated. Thus, after the transformation, we have a case similar to that shown in Figure 4.4. Again, to see the situation before the transformation, we transform the vectors in Figure 4.4 by \( M^{-1} \). This is shown in Figure 4.6. Since \( M \), and therefore \( M^{-1} \), have off-diagonal elements, the "L\( p \) unit ball" will be stretched out in directions not equal to the directions of the coordinate axises.

Some cases of "L\( p \) unit circles" are shown in Figure 4.7.

**Figure 4.5** L\( p \)-norm estimation in the case of heteroscedastic residuals.
Figure 4.6 \( L_p \)-norm estimation in the case of linearly dependent residuals.

\[ e_1 \]

\[ e_2 \]

\[ y_2 \]

\[ y_1 \]

\[ x_2 \]

\[ x_1 \]

\[ G = M^{-1}G^* \]

\[ y = M^{-1}y^* \]

\[ \epsilon = M^{-1}\epsilon^* \]

\[ \beta M^{-1}x^* \]

\[ x = M^{-1}x^* \]

a) normal case

b) cases of heteroscedastisity

c) cases of linearly dependent residuals

d) a case of a general linear transformation

Figure 4.7 "L_1 unit circles" in the case of independent and identically distributed residuals (the normal case) and different cases when linear transformations lead to the normal case.
4.9 Conclusions

In this chapter we have primarily been interested in the sampling distribution of $L_p$-norm estimators of linear regression models. Mostly our results concern the asymptotic distribution.

The "classical regression case", i.e. fixed non-stochastic regressor variables and independent and identically distributed residuals was studied in Section 4.2. There is was shown that for $p$ sufficiently small, $\sqrt{n}(\hat{\beta}_{pn} - \beta)$ is asymptotically normally distributed, where $\hat{\beta}_{pn}$ is the $L_p$-norm estimator of the parameter vector $\beta$. If the residual distribution is symmetric, the estimator was shown to be unbiased, for sufficiently small $p$.

The variance in the asymptotic distribution of $\sqrt{n}(\hat{\beta}_{pn} - \beta)$ in the "classical regression case" was shown to be $\omega_p^2 Q^{-1}$, where $Q = \lim_{n \to \infty} n^{-1}X'X$ and $\omega_p^2$ is a constant related to the residual distribution. Particularly, for $p = 1$, $\omega_p^2$ equals the asymptotic variance of the sample median, when a sample is drawn from the residual distribution. For $1 < p < \infty$ we have that $\omega_p^2 = E[|e_1|^{2p-2}]/\{(p-1)^2 E[|e_1|^{p-2}]\}$.

The case of stochastic regressors was studied in Section 4.3. In that case we have to add the "non-correlation" condition $E[X_{ij} | e_1 |^{p-1} \text{sign}(e_1)] = 0$, for regressor variables $X_{ij}$, $j=1,...,m$. For $p$ sufficiently small, $\sqrt{n}(\hat{\beta}_{pn} - \beta)$ is still asymptotically normally distributed. Particularly, when regressor variables and residuals are independent, the variance in the asymptotic distribution reduces to $\omega_p^2(E[Q])^{-1}$, where $\omega_p^2$ and $Q$ are defined as above.

The cases of heteroscedastic and linearly dependent residuals were studied in Sections 4.4 and 4.5, respectively. It was shown that $\sqrt{n}(\hat{\beta}_{pn} - \beta)$ still is asymptotically normally distributed for sufficiently small $p$, in the heteroscedastic case. It was also shown that the model may be linearly transformed as to satisfy the conditions in Sections 4.2 and 4.3, respectively.
Results from Monte Carlo simulations indicate that the sample distribution of the $L_p$-norm estimator is extremely sensitive to departures from the assumption of independent residuals. Either the distribution of $\sqrt{n}(\hat{\beta} - \beta)$ will not converge to a normal distribution, or the convergency will be very slow. However, if the residuals are linearly dependent, it is possible to linearly transform the model as to satisfy the assumptions in Sections 4.2 and 4.3, respectively. Thus, the parameter estimator for the transformed model will be asymptotically normally distributed for sufficiently small $p$.

When the regressors are fixed or stochastic and independent of the residuals, the asymptotic variance of the $L_p$-norm estimator depends on the residual distribution only through the constant $\omega_p^2$. In Section 4.6, $\omega_p^2$ was viewed as a function of $p$ for a fixed residual distribution. The optimal $L_p$-norm estimator for that distribution was defined as that $L_p$-norm estimator which has the least $\omega_p^2$.

For some distributions under study, it was found that the asymptotic variance of the optimal $L_p$-norm estimator equals or is near the Cramér-Rao lower bound. Whether this is true for a larger class of distributions is still an open question. However, the question is very important, since if that is the case, the optimal $L_p$-norm estimator will have a relatively small variance, when the residual distribution belongs to that class.

In the Distribution Theorems, $p$ is assumed fix. Therefore, the choice of $p$ may not be based on an estimate of the residual distribution. However, when the residual distribution is a member of a well-defined class of distributions, $p$ may be chosen by a minimax procedure described in Section 4.6.

If the residual distribution is known up to a scale parameter $\delta$, it is immediate that $\omega_p^2 = \delta^2 \omega_p^2$, where $\omega_p^2$ corresponds to the scale parameter $\delta = 1$. Therefore, the optimal $L_p$-norm estimator will be the same,
irrespective of the scale parameter. Thus, when the residual distribution is known up to a scale parameter, the optimal $L_p$-norm estimator may be determined, but the asymptotic variance of any $L_p$-norm estimator will be unknown. In this case, we have to construct an estimator $\hat{\omega}_p^2$ of $\omega_p^2$ and derive its distribution. In Section 4.7 some notes on the estimation of $\omega_p^2$ was given.

In Section 4.8, the $L_p$-norm estimation was given some geometrical interpretations.
5.1 An introduction to the use of $L_p$-norm estimators

We have so far been concerned with the $L_p$-norm estimators of linear regression models and some of their basic properties. The more important properties shown are the existence (Section 3.2), for $1 < p < \infty$ uniqueness (Section 3.4), and for $p = 1$ necessary and sufficient conditions for uniqueness (Section 3.5). We have also derived an asymptotic distribution theory under somewhat classical assumptions, i.e. fixed regressors and independent and identically distributed residuals (Section 4.2). The asymptotic distribution theory was then extended to cases of stochastic regressors (Section 4.3), and to cases where linear transformations of the residuals are independent and equally distributed (Sections 4.4 and 4.5). These results should be sufficient for a proper analysis in a large class of applications. Still there is a large, presumably larger, class of applications where the results developed are not sufficient. However, in these cases the presented results and ideas perform a basis for extensions and for further research. It is the aim of this chapter to present a small number of examples in statistical analysis where the results obtained easily can be applied and extended. The choice of cases discussed in this chapter is more or less subjective. Nevertheless, some of the cases discussed are important in some disciplines.

We will neither give a complete list of possible applications and extensions of $L_p$-norm estimators, nor will we extensively derive properties of the discussed estimators. To derive these properties, a lot of details involved in special applications should bring us far away beyond the frames of this discussion. The derivations needed are also usually in terms of matrix jugglery and will seldom provide insight into what is actually happening. For these reasons, we avoid the conventional approach to the problems and will use heuristic thinking rather than incomprehensible matrix
manipulations. By doing so, we will outline some of the areas for further research of $L_p$-norm estimators and indicate how the tools presented in the earlier chapters may be used.

Due to the Asymptotic Distribution Theorem, all $L_p$-norm estimators will asymptotically be affected by multicollinearity in a similar way. Thus, the presence of multicollinearity will increase the asymptotic variance of any $L_p$-norm estimator with the same relative amount. To some extent, the multicollinearity problem may be viewed as an information problem, viz. the information in the observations is not sufficient to properly estimate all parameters. In some cases, however, there is more information available, which if it could be used, should improve the precision of the estimator. In the context of multicollinearity, biased estimators which minimize the mean squared error have also been proposed. These problems are reviewed in the next section as the first application of $L_p$-norm estimators. We will there show how extra information may be incorporated into $L_p$-norm estimators and how to construct biased mean squared error minimizing $L_p$-norm estimators. These results are of most interest when constructing effective estimators in the presence of fat-tailed distributed residuals and at the same time, multicollinear data.

One of the most common problems when estimating econometric models is the presence of serially dependent residuals. If the dependence has a known linear structure, we know from Section 4.5 how to handle the problem. It is often the case that the dependence structure is assumed to be linear, but almost equally often, there are parameters, unknown to the researcher, involved in the structure. In Section 5.3 an interactive procedure for estimating the unknown parameters in a given dependence structure, and using the estimated structure for estimating the regression parameter, is given. This $L_p$-norm based iterative estimator has been studied by means of simulations, and a summary of that experiment is given. The
results from the study indicate that this estimator may constitute a good basis for the construction of effective estimators in the presence of fat-tailed distributed and, simultaneously, linearly dependent residuals.

The definition of $L_p$-norm estimators and all their properties as shown in Chapters 3 and 4 concern only uni-relational models. An analysis of multi-relational systems is much more complex and laborious. However, some guidelines to the construction of $L_p$-norm based estimators for the estimation of structural forms of interdependent systems are given in Section 5.4. The estimation may be performed in one, two or three stages. The one-stage and two-stage $L_p$-norm estimators are both immediate extensions of the least squares based OLS and 2SLS estimators. For the case of $p=1$, these estimators have already been studied in simulation studies, as reported in Subsection 2.6.3. Based on conclusions from these simulation experiments it is expected that the $L_1$-norm based estimators will give robust estimates when residuals are fat-tailed distributed. Further on, based on the ideas in Section 4.5 it is possible to make use of the dependence structure of the residuals, and hence define a three-stage $L_p$-norm based analogue to the 3SLS estimator.

In Section 5.5 we will give a brief summary of the conclusions to draw in this chapter.

5.2 Multicollinearity

In this section we will discuss the multicollinearity problem. Consider therefore again the linear regression model (2.5). It is well known that when the design matrix $X$ is illconditioned, the least squares estimator of $\beta$ may be very poor in terms of variances. That follows from the fact that the variance of the least squares estimator equals $\sigma^2(X'X)^{-1}$, where $\sigma^2$ is the residual variance, and the matrix $(X'X)^{-1}$ will contain very
large numbers when \( X \) is illconditioned. From Section 4.2 we know that the asymptotic variance of the general \( L_p \)-norm estimator equals \( \omega_p^2(X'X)^{-1} \). Therefore, we may not expect that any other \( L_p \)-norm estimator performs better than least squares, in the context of multicollinearity.

In many situations, however, the statistician has some a priori information concerning the parameters. This knowledge can be based on subjective considerations, see e.g. Raiffa and Schlaifer (1961), it can be based on theoretical considerations, see e.g. Theil and Goldberger (1961) or it can be based on experience from other sets of observations, see e.g. Anderson and Battiste (1975). We will call information based on something else than the given sample as extra information.

It is possible to view the extra information on regression parameters as certain restrictions on one or more elements in the parameter vector \( \beta \), viz.

\[
r = R \beta + \delta
\]

(5.1)

where \( R \) is a \( k \times m \)-dimensional matrix of known constants. Estimates of the linear combinations \( R\beta \) are contained in the \( k \)-dimensional vector \( r \) and \( \delta \) is a \( k \)-dimensional vector of residuals.

We will now show how the extra information may be incorporated in an \( L_p \)-norm estimation procedure. Assume therefore that the model residuals \( e \) and the extra information residuals \( \delta \) are independent and, up to a scale parameter, equally distributed. Assume also that the components in the \( \delta \)-vector are independent and located as to satisfy

\[
E[|\delta_i|^{p-1}\text{sign}(\delta_i)] = 0 \quad i=1,\ldots,k
\]

(5.2)

Define the constants \( \omega_p^2 \) as in the Asymptotic Distribution Theorem, i.e.

\[
\omega_p^2 = \begin{cases} 
1/(2F'(0))^2 & \text{for } p = 1 \\
E[|\varepsilon_1|^{2p-2}]/((p-1)^2E[|\varepsilon_1|^{p-2}]) & \text{for } 1 < p < \infty
\end{cases}
\]
and define the $k \times k$-dimensional diagonal matrix $\rho = \text{diag}(\rho_1, \ldots, \rho_k)$ by

$$\rho_j = \begin{cases} 2G'(0) & \text{for } p = 1 \\ \{(p-1)E[|\delta_j|^{p-2}]/E^{1/2}[|\delta_j|^{2p-2}] \} & \text{for } 1 < p < \infty \end{cases}$$

for $j = 1, \ldots, k$

where $F$ and $G$ are the distribution functions for $\epsilon_1$ and $\delta_1$, respectively. Combining the equations (2.5) and (5.1) and applying the ideas in Section 4.4 and Corollary 4.2 we arrive at

$$\begin{pmatrix} \frac{1}{\omega} y \\ \frac{1}{\omega} \epsilon \\ \rho \rho' r \end{pmatrix} = \begin{pmatrix} \frac{1}{\omega} X \\ \frac{1}{\omega} \rho' r \end{pmatrix} \beta + \begin{pmatrix} \frac{1}{\omega} \epsilon \\ \rho \delta \end{pmatrix}$$

(5.3)

All residuals in (5.3) are now independent and homeoscedastic and the parameter vector $\beta$ may be $L_p$-norm estimated in the usual manner, but using the combined vectors and matrices, respectively.

For the case $p = 2$, it is possible to derive an explicit expression of the estimator, viz.

$$b_{EI} = \left(\frac{1}{\omega} X'X + R'\rho' \rho R\right)^{-1} \left(\frac{1}{\omega} X'y + R'\rho' \rho r\right)$$

(5.4)

For further properties of this estimator in the case $p = 2$, we refer to Rao (1965), Theil (1963) and Battiste (1967). In the general case, it follows from Corollary 4.2 that for sufficiently small $p$, the asymptotic variance of the estimator is given by

$$V(b_{EI}) = \left(\frac{1}{\omega} X'X + R'\rho' \rho R\right)^{-1} \omega^2 (X'X)^{-1}$$

Since the matrix $R'\rho' \rho R$ is positive semidefinite, it follows that the elements in $V(b_{EI})$ are less than, or equal to the elements in the usual variance matrix $\frac{1}{\omega^2} (X'X)^{-1}$. Note that the estimator $b_{EI}$ is still asymptotically unbiased, due to the assumption (5.2).

Due to the extra information, it has been possible to increase the estimation precision. Since the estimator is based on $L_p$-norms, it is possible to choose a sufficiently small $p$, and thus we are able to handle
the combined problem if multicollinearity and fat-tailed residual distributions. Of course it is possible to make more general assumptions concerning the residual vectors $\varepsilon$ and $\delta$. For example, we may allow that their elements are linearly dependent and heteroscedastic such that there exists an $(n+r)\times(n+r)$-dimensional matrix $\Omega$ with the property that

$$\Omega \left( \begin{array}{c} \varepsilon \\ \delta \end{array} \right)$$

is an $(n+r)$-vector of independent and equally distributed stochastic variables. However, we feel that an investigation of that formulation is beyond the frames of this discussion.

Before we proceed, note that the elements $\rho_i$ in the $\rho$-matrix serve as inverse weights of the extra information. Thus, for a large value on $\rho_i$, we consider the prior estimate $r_i$ as highly uncertain and do not make much attention of it. As a limit when $\rho_i \to \infty$, we will ignore the prior estimate $r_i$. On the other hand, for a small value on $\rho_i$, we consider $r_i$ as a very good estimate and give it a high weight in the estimation. Therefore, the sample information plays a less important role. As a limit when $\rho_i \to 0$, we will ignore the sample information, when estimating the $i$-th row in $R$.

In cases of illconditioned design matrices, it has sometimes been argued that we have to exclude one or more variables from the model. This is, however, equivalent to certain restrictions on the parameters. In the terminology of this discussion, it is the same as an a priori estimate of zero for the parameters of the excluded variables, and in the estimation procedure, no sample information is used to estimate those parameters. For example, if the first two variables are to be excluded, we define

$$r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \end{pmatrix}, \quad \rho = \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix}$$

and let $\rho_1$ and $\rho_2$ tend to zero. The "method of excluded variables" may thus be considered as a special case of the "method of extra information".
The same arguments may be used to show that the "method of excluded principle components" as described by e.g. Massy (1965) and Greenberg (1975) and the "method of excluded latent roots" as described by e.g. Hawkins (1973) and Webster, Gunst and Mason (1974) also are special cases of the "method of extra information". The only difference from the "method of excluded variables" is that we exclude some linear combinations of variables instead of the variables themselves. The excluded linear combinations appear as rows in the matrix R.

Let us now consider the ridge estimator, originally described by Hoerl and Kennard (1970a and b) and defined by

\[ b_{RR} = (X'X + T)^{-1} X'y \]

where \( T \) is a mxm-dimensional matrix of constants and \( b_{RR} \) is the ridge estimator of \( \beta \). By letting \( r \) in (5.1) be the zero vector it follows from (5.4) that the ridge estimator is a special case of the extra information estimator, namely when \( T = \omega^2 R\rho'R \) and \( p = 2 \). Thus, an ridge estimator, based on \( L_p \)-norms, may be defined as an \( L_p \)-norm estimator of (5.3) with \( T = \omega^2 R\rho'R \) and \( r = 0 \).

Finally, we have to note that all estimators based on extra information are asymptotically unbiased as long as the condition (5.2) is satisfied, i.e. the extra information is unbiased. In applications this is, however, not always the case. For example, if \( \beta_1 \neq 0 \) and we exclude the corresponding variable \( x_1 \), we will estimate \( \beta_1 \) biased with zero. Further, it is not difficult to show that if \( x_2 \) is a variable in the model that is correlated with \( x_1 \), the corresponding parameter to \( x_2 \) will be estimated with bias.

5.3 Serially dependent residuals

The case of linearly dependent residuals was discussed in Section 4.5. There it was shown that a suitable linear transformation, the model
may be properly estimated. Thus, referring to the linear regression model (2.5) with linearly dependent residuals, there exists an \((nxn)\)-matrix \(M\) such that
\[
e^* = Me
\]
is a vector of independent and equally distributed residuals. After pre-multiplying the model with \(M\), we obtain
\[
My = MX\beta + Mc
\]
or, with \(y^* = My\) and \(X^* = MX\),
\[
y^* = X^*\beta + e^*
\]
and \(\beta\) is estimated by estimating this transformed model.

For the special case when the residuals are generated by a first order autoregressive process
\[
\epsilon_t = \rho\epsilon_{t-1} + \delta_t
\]
the matrix \(M\) is defined by
\[
M = \begin{bmatrix}
(1-\rho^\alpha)^{1/\alpha} & 0 & 0 & \ldots & 0 & 0 \\
-\rho & 1 & 0 & \ldots & 0 & 0 \\
0 & -\rho & 1 & \ldots & 0 & 0 \\
\vdots & & & & & \\
0 & 0 & 0 & \ldots & 1 & 0 \\
0 & 0 & 0 & \ldots & -\rho & 1
\end{bmatrix}
\]
Thus, \(M\) is a matrix of one parameter, the autoregressive parameter \(\rho\).

Presumably, we do not know \(\rho\). However, within the framework of the \(L_2\)-norm estimator, several iterative methods for estimating the \(\beta\) vector as well as \(\rho\) has been proposed. One idea is to replace \(\rho\) in \(M\) with an estimate of \(\rho\). A method to produce a consistent estimate of \(\rho\) is to first estimate \(\beta\) as if the residuals were independent. This yields a consistent estimate \(\hat{\beta}\) of \(\beta\). After computing the "observed" residuals \(\hat{\epsilon} = y - X\hat{\beta}\), the model
\[ \hat{\varepsilon}_t = \rho \hat{\varepsilon}_{t-1} + \rho_t \]
is estimated by regressing \( \hat{\varepsilon}_t \) on \( \hat{\varepsilon}_{t-1} \). The estimate of \( \rho \), which is consistent, is substituted into \( M \), and \( \beta \) is reestimated after transformation. This method is originally due to Cochrane and Orcutt (1949).

It is now easy to apply the Cochrane-Orcutt scheme to the case when the model is estimated in \( L_p \)-norms. The resulting estimator is expected to be of most interest in the presence of fat-tailed distributed and simultaneously serially dependent residuals.

When estimation is based on \( L_1 \)-norm, some results appear in Courtage and Nyquist (1980), where a Monte Carlo experiment was performed. The model under study was a simple linear regression model with one regressor variable and one intercept. The number of observations was \( n = 15, 30 \) and 100. 200 replicates were performed. Observed mean square errors for the case \( n = 30 \) are reported in Table 5.1. The immediate conclusion to draw is that the mean square errors have remarkably decreased due to the use of the transformation, although the transformation is based on an estimate. Based on the results from the Monte Carlo study, this method seems to be a promising alternative when residuals are fat-tailed distributed and, simultaneously, serially dependent. Therefore, further research in this area is motivated.
<table>
<thead>
<tr>
<th>φ</th>
<th>Parameter</th>
<th>Estimator</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.0</td>
</tr>
<tr>
<td>0</td>
<td>β₀</td>
<td>OLA</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLAU</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLAK</td>
<td>0.36</td>
</tr>
<tr>
<td>-0.5</td>
<td>β₁</td>
<td>OLA</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLAU</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLAK</td>
<td>0.011</td>
</tr>
<tr>
<td>-0.9</td>
<td>β₀</td>
<td>OLA</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLAU</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLAK</td>
<td>0.17</td>
</tr>
<tr>
<td>-0.9</td>
<td>β₁</td>
<td>OLA</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLAU</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLAK</td>
<td>0.005</td>
</tr>
<tr>
<td>-0.9</td>
<td>β₀</td>
<td>OLA</td>
<td>2.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLAU</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLAK</td>
<td>0.15</td>
</tr>
<tr>
<td>-0.9</td>
<td>β₁</td>
<td>OLA</td>
<td>0.078</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLAU</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GLAK</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Table 5.1 Observed mean square errors of $L_1$-norm based estimators of the parameters in $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \varepsilon_t = \rho \varepsilon_{t-1} + \delta_t, \varepsilon_t = \rho \varepsilon_{t-1} + \delta_t, t=1,\ldots,30$ and $\delta_t$ have a symmetric stable distribution with characteristic exponent $\alpha$. OLA is the usual $L_1$-norm estimator, GLAU is the $L_1$-norm estimator after a transformation based on an estimate of $\rho$, and GLAK is the $L_1$-norm estimator after a transformation based on a known value of $\rho$. The observations are from the Monte Carlo experiment reported in Coursey and Nyquist (1980).
5.4 Estimation of interdependent systems

We will now leave the discussion of estimating the linear regression model (2.5) for a while and concentrate upon estimation of interdependent systems. Let us therefore consider a set of $k$ simultaneous relations, written in matrix form as

$$Y = Y^\top + XA + \varepsilon$$

where $Y$ is a $(n \times k)$-matrix of $n$ observations on each of the $k$ endogenous variables, $X$ is a $(n \times m)$-matrix of $n$ observations on each of the $m$ predetermined variables, $\varepsilon$ is a $(n \times k)$-matrix of equally distributed and row-wise independent residuals and $T$ and $A$ are $(k \times k)$- and $(m \times k)$-matrices of unknown parameters to be estimated. The diagonal elements in $T$ are assumed to be zero. We will refer to this representation of the system as the structural form. The subject of this section is to introduce some ideas on how $L_p$-norms can be used to estimate the parameter matrices $T$ and $A$.

Probably the simplest way to estimate the parameters in the system (5.5) is to consider the relations one by one, estimating them as though they were $k$ independent unirelational models. The $L_p$-norm estimator is then applied to each of the $k$ relations. When least squares are applied, this estimator is usually called ordinary least squares, OLS. As an analogue, when an $L_p$-norm estimator is applied we call the resulting estimator ordinary least power, OLP. Thus, OLS is a special case of OLP. Another special case of OLP we have already discussed (see Subsection 2.6.3) is ordinary least absolute, OLA, which corresponds to $p = 1$ in the $L_p$-norm estimation.

Since we may assume interdependences in the system, the endogenous variables and the residuals will be dependent and thus not meet the non-correlation condition $E[Y_{ij} \varepsilon_{i',j'}] = 0$ for all $i,j,i'$ and $j'$. Therefore the OLS estimator will be biased and inconsistent. From the Stochas-
tic Regression Theorem, we know that the corresponding condition for the general $L_p$-norm estimator is $E[Y_{ij}|e_{i',j'}|^p\text{sign}(e_{i',j'})]=0$ for all $i, j, i', j'$. By the interdependency argument, the condition for the general $L_p$-norm estimator is not met. Thus, the OLP estimator is assumed to be biased and inconsistent for all $p$.

In the Monte Carlo study reported in Nyquist and Westlund (1977) and summarized in Subsection 2.6.3, a significant bias term was observed for the OLS and OLA estimators, supporting our hypotheses. See Figures 2.8 a and c for an illustration.

When the system under study is exactly identified or overidentified, two-stage and three-stage methods are often applied. Let us now see how the two-stage least squares technique, 2SLS, can be generalized to a two-stage estimator based on $L_p$-norms. Therefore, we introduce the reduced form

$$Y = X\Pi + V$$

(5.6)

where $\Pi = \Delta(1-\Gamma)^{-1}$ is an $(m\times k)$-matrix of parameters and $V = \varepsilon(1-\Gamma)^{-1}$ is an $(n\times k)$-matrix of residuals. Observe that the reduced form residuals in $V$ are linear combinations of the structural form residuals in $\varepsilon$.

The idea of 2SLS is well-known. In the first stage we construct a modified regressor $\hat{Y}$ that is independent of $\varepsilon$. This is done by estimating the reduced form (5.6) relation by relation. Since the predetermined variables are fixed or assumed to be uncorrelated with the residuals (i.e., assumed to satisfy $E[X_{ij}\varepsilon_{i',j'}]=0$ or equivalently $E[X_{ij}V_{i',j'}]=0$) the estimation procedure is consistent. The modified regressor $\hat{Y}$ is now computed from

$$\hat{Y} = X\hat{\Pi}$$
where $\hat{\varPi}$ is the estimate of $\varPi$. The assumption of zero covariance between predetermined variables $X$ and residuals $\varepsilon$, implies also that any linear function of the predetermined variables (in particular the linear combination $\hat{Y}$) and $\varepsilon$ will have zero covariance. Since the coefficients in $\hat{\varPi}$ are not constants, but are estimates, depending on $\varepsilon$, $\hat{Y}$ will be correlated with $\varepsilon$. This reservation holds true for a small number of observations $n$, but disappears when $n$ tends to infinity.

In the second stage $\hat{Y}$ is substituted for $Y$ in the right hand of the structural form (5.5), which gives

$$Y = \hat{Y} + X\Delta + \varepsilon^*$$

The parameters are now estimated by applying OLS. This estimation procedure is also consistent, since $\hat{Y}$ is uncorrelated with the adjusted residuals $\varepsilon^*$.

The generalization to a two-stage technique based on $L_p$-norms is now immediate: just replace the least squares estimations by $L_p$-norm estimations. We call the resulting estimator a two-stage least power estimator, 2SLP. Thus, the 2SLS and the two-stage least absolute, 2SLA, which we discussed in Subsection 2.6.3 are special cases of 2SLP.

Now, the idea of 2SLS is that we in the first stage compute a regressor that is uncorrelated with $\varepsilon$. If we in the second stage use the OLP-estimator, we still have a sort of uncorrelation condition to meet, viz.

for $1 < p < \infty$, $E[Y_{ij}|\varepsilon_{i,j},|P^{-1}\text{sign}(\varepsilon_{i,j})] = 0$. For $p = 1$ there is still an open question how the corresponding condition is to be formulated. The purpose of the first stage must then be to produce a regressor $\hat{Y}$ that satisfies this uncorrelation condition. The question now is whether that is possible or not. For $1 < p < \infty$, the answer relies on the result in the Characterization in $L_p$, $1 < p < \infty$ Theorem. With our notations in this section, the Theorem implies that

$$\frac{1}{n} \sum_{i=1}^{n} \hat{Y}_{ij} |\hat{Y}_{ij}|^{p-1}\text{sign}(\hat{V}_{ij}) = 0, \quad j=1,\ldots,k \quad (5.7)$$
where \( \hat{V}_{ij} = Y_{ij} - \hat{Y}_{ij} \) is the observed reduced form residuals. For \( p \neq 2 \), this expression is non-linear, implying that \( \hat{Y} \) and \( e = V\Gamma \) will not, at least in small samples, meet the uncorrelation condition for the second stage. In large samples, however, an interpretation of (5.7) is that the residual vectors \( \hat{V}_{ij} \), when projected on the subspace spanned by \( X_1, \ldots, X_m \), will be projected on the zero vector. Therefore, linear combinations of \( \hat{V}_{ij} \) (such as structural form residuals) will, in large samples, be projected on the zero vector, or the uncorrelation condition will be satisfied for \( \hat{Y} \) and structural form residuals. Note that this result holds true only if the same \( L_p \)-norm estimator is used in both stages.

Summarizing, we have that the 2SLP-estimator must be based on the same \( L_p \)-norm estimator in the first as well as the second estimation stage. A small sample bias is expected, but the estimator will be consistent. Referring again to Nyquist and Westlund (1977) and the summarized results in Subsection 2.6.3, a small bias-term was observed for the 2SLA-estimator. See Figures 2.8.a and 2.8.c. It was also observed that the 2SLA-estimator is fairly robust in terms of root mean squared error, against fat-tailed residual distribution, a property that is not shared by the 2SLS-estimator. See Figures 2.8.b and 2.8.d. The asymptotic distribution and other detailed questions concerning the 2SLP-estimator are still open. Since it seems that the 2SLP-estimator, and especially the 2SLA-estimator, is a significant step towards robust estimation of simultaneous models, we look forward to the solution of the more detailed questions.

The estimators we have discussed so far are "single-relation" estimators in the sense that the estimators operate on each relation separately. We will now complete this section with a discussion of "system" methods of estimation, i.e. methods that estimate all relations jointly.

As system methods take account of dependencies between relations, they are generally more (asymptotically) efficient than the single-relation
procedures. The increased efficiency is of course the motivation for considering joint estimation procedures. Let us first give a short description of the logic behind the three stage least square, 3SLS, estimator. Consider therefore again the system (5.5). When all relations have been estimated by 2SLS, the residual matrix \( \hat{\varepsilon} \) is estimated by

\[
\hat{\varepsilon} = Y(I - \hat{\Gamma}) - X \hat{\Delta}
\]

where \( \hat{\Gamma} \) and \( \hat{\Delta} \) are the 2SLS-estimates of \( \Gamma \) and \( \Delta \), respectively. \( \hat{\varepsilon} \) is now used to compute an estimate, \( \hat{\Sigma} \), of the covariance matrix of the residuals in the structural form. Thus,

\[
\hat{\Sigma}_{ij} = \frac{1}{n} \sum_{r=1}^{n} \hat{\varepsilon}_r \hat{\varepsilon}_j \quad i,j=1,\ldots,k \tag{5.8}
\]

where \( \hat{\Sigma}_{ij} \) is the \( i,j \)-element in the \( (k \times k) \)-matrix \( \hat{\Sigma} \). In the last step of 3SLS, \( \hat{\Sigma} \) is used in a GLS estimation of the structural form. More formally, the structural form (5.5) may be written as

\[
y_j = y_j y_j + \delta_j + \varepsilon_j = Z_j \beta_j + \varepsilon_j \quad j=1,\ldots,k
\]

where

\[
Z_j = (Y_j | X_j)
\]

and

\[
\beta_j = \left( \begin{array}{c} \gamma_j \\ \delta_j \end{array} \right)
\]

This set of relations may now be written as

\[
y* = Z* \alpha + \varepsilon*
\]

where

\[
y* = \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix}
\]

\[
z* = \begin{pmatrix} Z_1 & 0 \\ \vdots & \vdots \\ 0 & Z_k \end{pmatrix}
\]
In the final step of the 3SLS-estimator, (5.9) is GLS-estimated using 
\( \hat{\Sigma} M \) as an estimate of the covariance matrix of \( \epsilon^* \). The symbol \( M \) stands 
for the Kronecker product. However, there exists a \((k \times k)\)-matrix \( M \) with 
the property
\[
M(\hat{\Sigma} M) M' = I. \tag{5.10}
\]
Therefore, the final step in the 3SLS-estimation, may also be defined as 
a least square estimation of the model (5.9) premultiplied by \( M \):
\[
MY^* = MZ^* \alpha + Me^* \tag{5.11}
\]
It is now possible to generalize the 3SLS-estimator to a three stage least 
power, 3SLP, estimator. The two first stages coincide with the 2SLP-esti-
mator. The dependence structure in \( \epsilon^* \) is then estimated and, using the 
ideas in Section 4.5 and the decomposition of \( \hat{\Sigma} \) as in (5.10), the final 
stage is to \( L_p \)-norm estimate (5.11). Unfortunately, the properties of a 
transformation based on the estimator (5.8), are still unknown. If the 
dependence between the relations is linear, which is the case if the re-
siduals are symmetric multivariate stable distributed, the transformation 
should, of course, be linear. However, for a characteristic exponent \( \alpha < 2 \), 
the expression in (5.8) does not converge, and is thus expected to be a 
poor estimator of the dependence structure. The question how the structure 
efficiently will be estimated is beyond the frames of this discussion, and 
is therefore left open. Other questions we leave open are those concerning 
the asymptotic distribution of the 3SLP-estimator.

We feel that the 3SLP-estimator (and especially that based on \( L_1 \)-norms) 
is of considerable importance when constructing robust system estimators. 
Therefore, the more detailed development of this estimator is one of the 
most important tasks in this area.
5.5 Conclusions

Based more on heuristic discussions than on exact deduction, we have outlined how properties of $L_p$-norm estimators derived in Chapters 3 and 4 may be extended and used to define more complex $L_p$-norm based methods. The first two topics discussed concerned two common problems in stochastic model building, viz. multicollinearity and serially dependent residuals.

In some cases extra information on the parameters, e.g. information that may not be referred to the present sample, is available. In Section 5.2 it was shown how this extra information may be incorporated into an $L_p$-norm based estimation procedure. It was also shown how $L_p$-norm based biased, mean squared error minimizing estimators may be constructed.

$L_p$-norm based estimators of regression models with dependent residuals were discussed in Section 5.3. An iterative procedure to first estimate the residual dependence structure and then, using these estimates, to estimate the regression model, was proposed. A Monte Carlo simulation to study properties of the proposed iterative procedure was presented. The simulation results indicated a considerable improvement of the estimation precision over methods not taking account of the dependence structure.

From Chapter 4 we know that $L_p$-norm estimators with sufficiently small $p$, perform well in the context of fat-tailed distributed residuals. With the ideas presented in Sections 5.2 and 5.3 it is now possible to discuss "combined problems", i.e. the presence of fat-tailed distributed residuals in multicollinear models or in models with linearly dependent residuals. The extension to the "triple problem" linearly dependent fat-tailed distributed residuals in multicollinear models is immediate.

In Section 5.4 we discussed the possibilities of estimating interdependent systems with $L_p$-norms. We showed how one-, two- and three stage
estimators may be constructed. They are all, more or less, extensions of
the familiar one-, two- and three stage least squares estimators. One-
and two stage $L_1$-norm estimators have been studied in a Monte Carlo si-
mulation, which was briefly reviewed in Subsection 2.6.3. The simulation
results indicate that the $L_1$-norm based estimators of the structural form
perform well in the context of fat-tailed residuals. For the three stage
estimator, the problem of how the dependence structure of the structural
form residuals are to be efficiently estimated, when residuals are fat-
tailed distributed, remains. Once this problem is solved, we have defi-
ned a system estimator that is expected to perform well when residuals
are fat-tailed distributed.

These three examples of applications of $L_p$-norm estimators do not
constitute a complete list of possible applications. They may rather be
viewed as the beginning of a list with a size which is beyond the authors
apprehension. $L_p$-norm based methods may be developed in almost any si-
tuation of estimation. The principle is always the same, viz. minimizing
the $L_p$-norm of the observed residual vector. The results reported in
Chapters 3 and 4 are some of the tools necessary to develop their pro-
perties.

The $L_p$-norm based methods and all their properties are fairly abstract
concepts. However, in connection with real world models and real world
data many of the concepts will have a "real world interpretation". There-
fore, practical use of $L_p$-norm estimators, when applicable, will make the
investigator familiar with interpretations of our abstract concepts and,
presumably, spread light over still darkened areas. With this chapter,
we hope we have spread some light over the applicability of $L_p$-norm based
methods.
Consider the linear regression model

\[ y = X \beta + \epsilon, \]

where \( y \) is an \( n \)-dimensional vector with observations on an endogenous variable, \( X \) is an \( n \times m \) matrix with observations on \( m \) exogenous variables, \( \beta \) is an \( m \)-dimensional vector with unknown parameters and \( \epsilon \) is an \( n \)-dimensional vector with unobservable residuals. The matrix \( X \) is assumed to be of full rank \( = m \). When estimating the parameter vector \( \beta \), the least squares (L_2-norm) estimator is often used. When residuals are independent and identically normally distributed, the least squares estimator is BLUE as well as equivalent to the maximum likelihood estimator. However, the least squares estimator is known to be sensitive to departures from the assumption of normally distributed residuals. In a variety of applications there are theoretical as well as empirical evidences that the residuals display distributional characteristics different from those of normal distributions. It is therefore desirable to develop alternative estimators. Members of the class of L_p-norm estimators have here been proposed as alternatives to the least squares estimator.

For the linear regression model, an L_p-norm estimate of the parameter vector \( \beta \) is a vector which minimizes

\[
S_{pn}(b) = \sum_{i=1}^{n} |y_i - \sum_{j=1}^{m} x_{ij} b_j|^p \quad \text{for } 1 \leq p < \infty
\]

\[
S_{\infty}(b) = \max_i |y_i - \sum_{j=1}^{n} x_{ij} b_j| \quad \text{for } p = \infty.
\]
The need for an investigation of the $L_p$-norm estimators has been inspired by various Monte Carlo simulations, indicating that an $L_p$-norm estimator with $p < 2$ may be superior to the least squares estimator, when the residual distribution possesses an infinite variance.

In this monograph we have discussed questions concerning the existence, uniqueness and asymptotic distributions of $L_p$-norm estimators. Thus, we have proved that there always exists an $L_p$-norm estimate and that estimate is, for $1 < p < \infty$, unique. For $p = 1$, we have proved a necessary and sufficient condition on the matrix $X$ for unique estimation. We have also proved a condition on the residual distribution that $\lim_{n \to \infty} n^{-1} E[S_n(b)]$ has a unique minimum.

The asymptotic distribution of the $L_p$-norm estimator $\hat{\beta}_{pn}$ of the parameter vector $\beta$ was then discussed. In the case of fixed regressors and independent and identically distributed residuals, we found that $\sqrt{n}(\hat{\beta}_{pn} - \beta)$ is, for sufficiently small $p$, asymptotically normally distributed with mean zero and variance $\omega_p^{-1}$, where

$$\omega_p^2 = \begin{cases} \{2 F_\varepsilon(0)\}^{-2} & \text{for } p = 1 \\ \frac{E[|\varepsilon_1|^{2p-2}]/(p-1)E[|\varepsilon_1|^{p-2}]^2}{(p-1)E[|\varepsilon_1|^{p-2}]} & \text{for } 1 < p < \infty \end{cases}$$

$F_\varepsilon$ is the distribution function of the residual distribution and $Q = \lim_{n \to \infty} n^{-1} X'X$. Cases of stochastic regressors, heteroscedastic and linearly dependent residuals are also discussed.

The optimal $L_p$-norm estimator was defined as that $L_p$-norm estimator that possesses the smallest asymptotic variance. The concept of the optimal $L_p$-norm estimator and the problem of choosing $L_p$-norm estimator has been discussed and illustrated with some examples.
Finally, we discussed the possibilities to construct $L_p$-norm based estimators for estimating multicollinear regression models and regression models with serially dependent residuals. We also discussed $L_p$-norm based estimators of interdependent systems.

One of the earlier drawbacks of $L_p$-norm estimators (different from the least squares) was the difficulty in the computation. Another drawback was the lack of a distribution theory. However, nowadays there are efficient computer routines for the computation of $L_p$-norm estimates. There is also a distribution theory in progress and much is known concerning the asymptotic distribution.

A main conclusion from this monograph is that there is a large number of models with residuals, distributed such that there is an $L_p$-norm estimator superior to the least squares. In particular, when residuals are fat-tailed or contaminated distributed, alternatives with $p < 2$ are recommended. When the actual residual distribution is known, it is possible to derive the optimal $L_p$-norm estimator, i.e. the $L_p$-norm estimator with the smallest asymptotic variance. When the residual distribution is only known to belong to a class of distributions, a minimax procedure may be used to choose estimator. With this procedure, the precision of the estimators will often be considerably increased, and in a few cases slightly decreased, compared to a standard use of the least squares estimator. Therefore, the members of the class of $L_p$-norm estimators with $p \neq 2$ are viable alternatives to the least squares estimator and we recommend their use when the residual distribution shows characteristics different from the normal.
REFERENCES


Boscovich, R.J. (1757), De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa, *Bononieni Scientiarum et Artium Instituto atque Academia Commentarii*, 4, pp 353-396.


Chebychev, P.L. (1857), Les questions de minima qui se rattachent a la représentation approximative des fonctions, printed in MM.A. Markoff and N. Souin (Eds), Œuvres de P.L. Tchebychef, St. Petersburg, 1907, pp 273-321.


Edgeworth, F.Y. (1887b), On observations relating to several quantities, Hermathena, 6, pp 279-285.


Euler, L. (1748), Pièce qui a remporté le prix de l'Académie royale des sciences en 1748, sur les inégalités du mouvement de Saturne et de Jupiter. (Published in Paris 1749).


Harvey, A.C. (1976), On the unbiassness of L -norm estimators and M-estimators, University of Kent at Canterbury.


Huber, P.J. (1968), Robust estimation. Presented at the European Meeting 1968 on Statistics, Economics and Management Science in Amsterdam. Also in Mathematical Centre Tracts 27, Mathematical Centre at Amsterdam, Holland.


Kolmogorov, A.N. (1933), Sulla determinazione empirica di una legge di distribuzione, Giornale dell’Istituto Italiano degli Attuari, 4.


Sposito, V.A. (1976), Minimizing the maximum absolute deviation, SIGMAP, 20, pp 51-53.


