1978

Aspects of linear regression estimation under the criterion of minimizing the maximum absolute residual

Michael Lawrence Hand
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Statistics and Probability Commons

Recommended Citation
Hand, Michael Lawrence, "Aspects of linear regression estimation under the criterion of minimizing the maximum absolute residual " (1978). Retrospective Theses and Dissertations. 6551.
https://lib.dr.iastate.edu/rtd/6551
INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoging at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

University Microfilms International
300 North Zeeb Road
Ann Arbor, Michigan 48106 USA
St. John's Road, Tyler's Green
High Wycombe, Bucks, England HP10 8HR
HAND, MICHAEL LAWRENCE
ASPECTS OF LINEAR REGRESSION ESTIMATION UNDER
THE CRITERION OF MINIMIZING THE MAXIMUM
ABSOLUTE RESIDUAL.

IOWA STATE UNIVERSITY, PH.D., 1978
Aspects of linear regression estimation under the criterion of minimizing the maximum absolute residual

by

Michael Lawrence Hand

A Dissertation Submitted to the Graduate Faculty in Partial Fulfillment of The Requirements for the Degree of DOCTOR OF PHILOSOPHY

Major: Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1978
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER I. INTRODUCTION</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Model and Alternative Estimation Criteria</td>
<td>1</td>
</tr>
<tr>
<td>Optimality Properties of Chebyshev Estimation</td>
<td>6</td>
</tr>
<tr>
<td>Possible Applications of Chebyshev Estimation</td>
<td>13</td>
</tr>
<tr>
<td>Overview of the Remaining Chapters</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER II. THE BASIC THEORY OF SOLUTION PROCEDURES FOR CHEBYSHEV ESTIMATION AND ITS NATURAL CONSEQUENCES</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Classical Exchange Algorithm for Chebyshev Approximation</td>
<td>19</td>
</tr>
<tr>
<td>The Linear Programming Formulation of the Chebyshev Problem</td>
<td>23</td>
</tr>
<tr>
<td>The Dual Formulation of the Chebyshev Problem</td>
<td>25</td>
</tr>
<tr>
<td>Properties of the Chebyshev Solution</td>
<td>28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER III. DEVELOPMENT OF EFFICIENT ALGORITHMS FOR CHEBYSHEV ESTIMATION</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Simplex Algorithm</td>
<td>36</td>
</tr>
<tr>
<td>The Construction of Effective Pivot Selection Rules</td>
<td>43</td>
</tr>
<tr>
<td>The Simplex Method Applied to the Chebyshev Problem</td>
<td>45</td>
</tr>
<tr>
<td>The Algorithm of Barrodale and Phillips</td>
<td>49</td>
</tr>
<tr>
<td>Using an Approximate Chebyshev Estimator to Accelerate Convergence in the Algorithm of Barrodale and Phillips</td>
<td>55</td>
</tr>
<tr>
<td>The Gradient Projection Pivot Selection Technique</td>
<td>65</td>
</tr>
<tr>
<td>The Chebyshev Solution of a Sample Problem</td>
<td>79</td>
</tr>
<tr>
<td>Chapter IV. Aspects of the Statistical Distribution of Chebyshev Estimators</td>
<td>Page</td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>A Survey of the Theory of M-Estimation</td>
<td>90</td>
</tr>
<tr>
<td>The Limiting Distribution of the Sample Midrange</td>
<td>95</td>
</tr>
<tr>
<td>The Results of Koenker and Bassett for $L_1$ Estimation</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter V. Applications of Chebyshev Estimation</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Chebyshev Criterion in Approximating Functions</td>
<td>106</td>
</tr>
<tr>
<td>The Efficiency of Chebyshev Estimation</td>
<td>109</td>
</tr>
<tr>
<td>Application of Chebyshev Estimation in Data Analysis</td>
<td>118</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter VI. Conclusions and Summary</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIBLIOGRAPHY</td>
<td>129</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>136</td>
</tr>
</tbody>
</table>
CHAPTER I. INTRODUCTION

The Model and Alternative Estimation Criteria

Consider the linear model

**Model 1.1.** \( y = X\beta + e \)

where

- \( y = \{y_i\}_{i=1}^{n} \) is a vector of \( n \) observations on a dependent variable
- \( X = \{x_{ij}\}_{j=1}^{p},_{i=1}^{n} \) is an \( nxp \) (\( n>p \)) fixed design matrix of \( n \) observations on each of \( p \) independent or explanatory variables
- \( \beta = \{\beta_j\}_{j=1}^{p} \) is a vector of \( p \) fixed, but unknown, parameters
- \( e = \{e_i\}_{i=1}^{n} \) is an \( n \)-vector (unknown) of disturbances or deviations from fit which are assumed to be independent and identically distributed according to some common distribution function \( F \).

The problem of linear regression of \( y \) upon \( X \) is to estimate the coefficient vector \( \beta \) such that the deviations of the observed \( y \) from the fit, \( X\beta \), are in some sense minimized. Since there is no unique criterion for minimizing a vector, we are generally satisfied to minimize some functional, typically some vector norm, of the residual vector \( (y-X\beta) \). Thus, given an appropriate vector norm \( ||.|| \), the solution
set of the linear regression problem is given by

\[ \{ \hat{\beta}^*: | |y - X\hat{\beta}^*| | \leq | |y - Xb| | \ \forall \ b \in \mathbb{R}^p \} \]

where $\mathbb{R}^p$ denotes Euclidean p-space.

Classically, the minimization criterion employed is the Euclidean norm which leads to the least squares estimation procedure of minimizing the sum of squared residuals. That is, we determine $\hat{\beta}$ such that

\[ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - Xb_i)^2 \ \forall \ b \in \mathbb{R}^p. \]

While the class of all vector norms is indeed limitless, an interesting subclass of alternatives is given by the so-called $L_p$ norms. Since we use $p$ throughout to denote the number of parameters or independent variables in a model, we shall henceforth refer to this class of norms as $L_q$ norms.

The $L_q$ norm of an $n$-vector $v$ is defined for $1 \leq q < \infty$ as

\[ L_q(v) = \{ \sum_{i=1}^{n} |v_i|^q \}^{1/q}. \]

Of particular interest is the limiting case as $q$ goes to infinity. This defines the $L_\infty$ norm as

\[ L_\infty(v) = \max_{1 \leq i \leq n} |v_i|. \]

The $L_\infty$ norm is also commonly known as the uniform or
Chebyshev norm. Commonly considered estimation criteria which are members of the class of $L_q$ norms are the $L_1$ criterion ($q=1$) of minimizing the sum of absolute residuals and, of course, the familiar $L_2$ ($q=2$) or least squares criterion.

The justification for studying alternative estimation criteria lies in the theory of errors. As emphasized by Rice and White (1964), the effectiveness of an $L_q$ norm in estimation depends essentially on the distribution of errors, i.e. - the distribution of the elements of $e$. A well-known (see e.g. Rice and White (1964), Harter (1974, 1975)), but useful, result which establishes the connection between $L_q$ estimation and a familiar subfamily of the exponential family of distributions is given by the following theorem.

**Theorem 1.1.** Under the assumptions of Model 1.1, if the common distribution of the elements of the error vector $e$ is of the exponential type with probability density function of the form

$$f(e_i) = c \exp\{-|e_i|^q\}, \quad i=1,2,\ldots,n, \quad 1<q<\infty$$

then the maximum-likelihood estimator $\hat{\beta}^*$ of $\beta$ in Model 1.1 is given as

$$\{\hat{\beta}^*: \sum_{i=1}^{n} |(y-X\hat{\beta}^*_i)|^q \leq \sum_{i=1}^{n} |(y-X\beta_i)|^q \forall \beta \in \mathbb{E}^p\}.$$
i.e., the $L_q$ estimator of $\beta$.

In view of this result, it is clear that $L_1$ estimation is optimal under a Laplace distribution of errors and least squares is optimal for normally distributed errors. Furthermore, as will be proven in the following section, the $L_\infty$ estimator is maximum-likelihood in the case of uniformly distributed errors.

It is instructive to consider a special case of Model 1.1. Consider the following simple location model.

**Model 1.2.** $y = \alpha + e$ with all the same assumptions as Model 1.1.

Note that Model 1.2 is just a special case of Model 1.1 with $p=1$, $\beta_1=\alpha$, and $x_{i1}=1$ ($i=1,2,\ldots,n$). Let us denote the $L_q$ estimator of $\alpha$ as $\alpha^*_q$. Rice and White (1964) have tabled the asymptotic variance of $\alpha^*_q$ as a function of the sample size, $n$, for $q=1,2,\infty$ and for several distributions of the error vector $e$. This table is reproduced as Table 1.1.

In viewing Table 1.1, it is apparent that $L_1$ estimation is most effective for heavy-tailed distributions of error, $L_\infty$ estimation is most efficient for error distributions with sharply defined extremes, while least squares is optimal for the normal distribution and performs well for error distributions with reasonably light tails.

For the remainder of the thesis, we shall be concerned
with the $L_\infty$ or Chebyshev estimation criterion of minimizing the maximum absolute residual. The remaining two sections of this chapter will deal with optimality properties of Chebyshev estimation and possible applications of the Chebyshev criterion, respectively.

Table 1.1. Asymptotic variance of $L_q$ estimators of location for $q = 1, 2, \infty$

<table>
<thead>
<tr>
<th>Error 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>-</td>
<td>$\frac{1}{6n}$</td>
<td>$\frac{(4-\pi)}{4n}$</td>
</tr>
<tr>
<td>Normal</td>
<td>$\frac{\pi}{2n}$</td>
<td>$\frac{1}{n}$</td>
<td>$\frac{\pi^2}{12 \log n}$</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>
Optimality Properties of Chebyshev Estimation

The motivation for considering Chebyshev estimation is possibly best done in view of its optimality properties when the error distribution is uniform with expectation zero. Hence we shall begin by proving the following theorem which extends the result of Theorem 1.1 to the limiting case $q \to \infty$.

**Theorem 1.2.** Under the assumptions of Model 1.1, let $e_1, e_2, \ldots, e_n$ be independent identically distributed uniformly on some symmetric interval $[-\varepsilon, \varepsilon]$ where $\varepsilon$ is fixed, but unknown. Then the Chebyshev or $L_\infty$ estimator \( \hat{\beta} \) of the unknown parameter vector \( \beta \) is the maximum-likelihood estimator of \( \beta \).

**Proof.** It is an immediate consequence of the distribution of \( \varepsilon \) and the assumption that \( X \) and \( \beta \) are fixed, that

\[ y_i \sim U[(X\beta)_i - \varepsilon, (X\beta)_i + \varepsilon) \]  

\[ \forall \ i = 1, 2, \ldots, n. \]

That is

\[ f(y_i; \beta, \varepsilon) = \frac{1}{2\varepsilon} \text{ if } (X\beta)_i - \varepsilon \leq y_i \leq (X\beta)_i + \varepsilon \]

\[ = 0 \quad \text{otherwise.} \]

And, since the errors are assumed to be independent, we may express the joint likelihood of \((\beta, \varepsilon)\) given the sample
$y_1, y_2, \ldots, y_n$ as

$$L(\beta, \varepsilon; y) = \prod_{i=1}^{n} f(y_i; \beta, \varepsilon)$$

$$= (2\varepsilon)^{-n} \prod_{i=1}^{n} \psi[(y-x_0)_{i}, \varepsilon]$$

(1.1)

where

$$\psi(t_1, t_2) = \begin{cases} 1 & \text{if } |t_1| \leq t_2 \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that the likelihood function $L(\beta, \varepsilon; y)$ is maximized by minimizing $\varepsilon$ subject to the constraint that $\psi[(y-x_0)_{i}, \varepsilon]=1$ for all $i=1,2,\ldots,n$. That is, $L(\beta, \varepsilon; y)$ is maximized by minimizing $\varepsilon$ subject to the constraint that the maximum absolute residual is less than or equal to $\varepsilon$. Thus, the joint maximum-likelihood estimator $(\tilde{\beta}, \tilde{\varepsilon})$ is given by

$$\left\{ (\beta, \varepsilon) : \max_{1 \leq i \leq n} |(y-x_{\beta})_{i}| \right\}$$

$$\leq \max_{1 \leq i \leq n} \left| (y-x_{\beta})_{i} \right| \forall \beta \in \mathbb{E}, \tilde{\varepsilon} = \max_{1 \leq i \leq n} \left| (y-x_{\tilde{\beta}})_{i} \right|.$$ 

But this is exactly the Chebyshev criterion of minimizing the maximum absolute residual. Thus the result is proved.

There does not in general exist any closed-form solution for $\beta$ in Model 1.1 under the Chebyshev criterion. Thus the Chebyshev estimator must be found iteratively. This not only presents computational difficulties for
those who wish to apply the Chebyshev criterion in practice, but also provides severe stumbling blocks to the development of a distribution theory. These difficulties, which are common to all alternative norm problems, represent a major reason for the favored role of least squares estimation in regression analysis. However, in the case of the simple location model, Model 1.2, a closed-form solution for the Chebyshev estimator is guaranteed to exist.

It is necessary, before proceeding, to establish that for Model 1.2 a closed-form solution always exists and is given by

$$\tilde{a} = \frac{Y(1) + Y(n)}{2}$$  \hspace{1cm} (1.2)

where $Y(1)$ denotes the $r^{th}$ smallest value of $\{y_1, y_2, \ldots, y_n\}$. This is accomplished in the following lemma.

Lemma 1.1. Under the assumptions of Model 1.2, the Chebyshev estimator of location is given by (1.2).

Proof. Consider the equivalent problem of minimizing

$$\max_{1 \leq i \leq n} |y_i - (\tilde{a} + \delta)|$$

with respect to $\delta \in \mathbb{R}$. 


max_{1 \leq i \leq n} |y_i - (\tilde{\alpha} + \delta)| \geq \max\{ |y_{(1)} - \tilde{\alpha} - \delta|, |y_{(n)} - \tilde{\alpha} - \delta| \}

= \max\{ \max\{ |y_{(1)} - y_{(n)}|/2 - \delta|, |y_{(n)} - y_{(1)}|/2 - \delta| \} \}

= \max\{ \frac{|y_{(n)} - y_{(1)}|}{2} - \delta, \frac{|y_{(n)} - y_{(1)}|}{2} + \delta \}

= \max_{1 \leq i \leq n} |y_i - \tilde{\alpha}| + |\delta|

\geq \max_{1 \leq i \leq n} |y_i - \tilde{\alpha}| \text{ with equality iff } \delta = 0.

Thus \( \max_{1 \leq i \leq n} |y_i - \alpha| \) is minimized, with respect to \( \alpha \), for \( \alpha = \tilde{\alpha} = \frac{y_{(1)} + y_{(n)}}{2} \).

Thus, it is established that the Chebyshev estimator of location is given by the sample midrange (1.2). The following lemma establishes sufficient conditions on the distribution of errors for the Chebyshev estimator of \( \alpha \) to be unbiased.

Lemma 1.2. Under Model 1.2, assume \( e_1, e_2, \ldots, e_n \) are independent identically distributed according to a distribution \( F \) which is symmetric about zero and such that \( E[e_1] < \infty \). Then \( E[\tilde{\alpha}] = \alpha \), i.e. \( \tilde{\alpha} \) is an unbiased estimator of \( \alpha \).
Proof. First of all, note that the existence of $E[e_1]$ is sufficient to guarantee the existence of $E[e_{(n)}]$ for all $n$ (see e.g. David (1970)). Then

$$E[\hat{a}] = \frac{1}{2}\{E[y_{(1)}] + E[y_{(n)}]\} \]
$$

$$= \frac{1}{2}\{\alpha + E[e_{(1)}] + \alpha + E[e_{(n)}]\} \]
$$

$$= \alpha + \frac{1}{2}\{-E[e_{(n)}] + E[e_{(n)}]\} \]
$$

$$= \alpha \]
$$

where the identity $E[e_{(1)}] = -E[e_{(n)}]$ follows from the symmetry assumption (see e.g. David (1970)).

**Corollary 1.1.** Under Model 1.2, if $e_1, e_2, ..., e_n$ are independent identically distributed as Uniform $[-\varepsilon, \varepsilon]$. Then $\hat{a}$, the sample midrange, is an unbiased estimator of $\alpha$.

The next result we wish to prove is that $\hat{a}$ is the minimum variance unbiased estimator (MVUE) of $\alpha$. Before doing this we state, without proof, two classical theorems in minimum variance unbiased estimation.

**Theorem 1.3.** (Factorization Theorem of Sufficiency)

Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from the density $f(x; \theta)$, where the parameter $\theta$ may be vector valued. A statistic $T = t(X_1, X_2, ..., X_n)$ is sufficient for $\theta$ if and only if the joint density of $X_1, X_2, ..., X_n$, which is
\[ 
\prod_{i=1}^{n} f(x_i; \theta), \text{ factors as} 
\]
\[
\frac{f_{x_1, x_2, \ldots, x_n}(x_1, x_2, \ldots, x_n; \theta)}{\prod_{i=1}^{n} f(x_i; \theta)} = g(t(x_1, x_2, \ldots, x_n; \theta) \cdot h(x_1, x_2, \ldots, x_n) 
\]
\[= g(t; \theta) h(x_1, x_2, \ldots, x_n) \]
where \( h(x_1, x_2, \ldots, x_n) \) is nonnegative and does not depend upon \( \theta \) and \( g(t(x_1, x_2, \ldots, x_n; \theta) \) is nonnegative and depends on \( x_1, x_2, \ldots, x_n \) only through the function \( t(x_1, x_2, \ldots, x_n) \).

**Theorem 1.4.** (Lehmann–Scheffé) Let \( X_1, X_2, \ldots, X_n \) be a random sample of size \( n \) from a density \( f(x; \theta) \). If \( S = s(X_1, X_2, \ldots, X_n) \) is a complete sufficient statistic for \( \theta \) and if \( T = t(S) \), a function of \( S \), is an unbiased estimator of \( \tau(\theta) \), then \( T \) is the unique minimum variance unbiased estimator of \( \tau(\theta) \).

**Theorem 1.5.** For Model 1.2, the simple location model, if \( e_1, e_2, \ldots, e_n \) are independent identically distributed as Uniform \([-\varepsilon, \varepsilon]\), the sample midrange \( [y_{(1)} + y_{(n)}]/2 \) is a minimum variance unbiased estimator of the location parameter \( \alpha \).

**Proof:** From (1.1), the joint probability density function of \( y_1, y_2, \ldots, y_n \) for fixed \( \alpha \) is given by
\[
f(y; \alpha, \varepsilon) = (2\varepsilon)^{-n} \prod_{i=1}^{n} \psi(y_i - \alpha, \varepsilon) 
\]
where
\[
\psi(t_1, t_2) = \begin{cases} 
1 & \text{if } |t_1| \leq t_2 \\
0 & \text{otherwise}.
\end{cases}
\]

But \(f(y; \alpha, \varepsilon)\) can be re-expressed as
\[
f(y; \alpha, \varepsilon) = (2\varepsilon)^{-n} \phi[y_{(1)}-(\alpha-\varepsilon)] \phi[(\alpha+\varepsilon)-y_{(n)}]
\]
where
\[
\phi(t) = \begin{cases} 
1 & \text{if } t \geq 0 \\
0 & \text{otherwise},
\end{cases}
\]
since \(\prod_{i=1}^{n} \psi(y_i-\alpha; \varepsilon) = 1\) if and only if \(\alpha-\varepsilon \leq y_i \leq \alpha+\varepsilon\)
for all \(i=1, 2, \ldots, n\); that is, if and only if \(\alpha-\varepsilon \leq y_{(1)}\) and
\(y_{(n)} \leq \alpha+\varepsilon\). Thus, using the factorization of \(f(y; \alpha, \varepsilon)\)
given in (1.3), it follows from Theorem 1.3 that \((y_{(1)}, y_{(n)})\)
are jointly sufficient for \((\alpha, \varepsilon)\). The completeness of
\((y_{(1)}, y_{(n)})\) can be verified by taking partial derivatives
of the equation
\[
\int_{\alpha-\varepsilon}^{\alpha+\varepsilon} \int_{\alpha-\varepsilon}^{y_{(n)}} g(y_1, y_n)(y_n-y_1)^{n-2} dy_1 dy_n = 0 \quad \forall (\alpha, \varepsilon)
\]
with respect to \(\alpha\) and \(\varepsilon\) to obtain the necessary implication
that \(g(y_{(1)}, y_{(n)}) \equiv 0\). That is, the only unbiased estimator
of zero that is a function of \((y_{(1)}, y_{(n)})\) is a statistic that
is identically zero with probability one. Furthermore, by
Corollary 1.1, the sample midrange is an unbiased estimator
of \(\alpha\) and is a function of the joint complete sufficient
statistic for \((\alpha, \varepsilon)\). Therefore, by Theorem 1.4, the sample
midrange \( \frac{y(1) + y(n)}{2} \) is the minimum variance unbiased estimator of the location parameter \( \alpha \) when the error distribution is uniform.

In this section, we have established two key optimality properties of Chebyshev estimation under the assumption of a uniform error distribution, symmetric about zero. First, that the Chebyshev estimator is maximum-likelihood and second, that the Chebyshev estimator is a minimum variance unbiased estimator for the location model. While this has not been an exhaustive treatment of the statistical properties of Chebyshev estimation, its intent is to motivate the study of Chebyshev estimation. Further results on the distribution of Chebyshev estimates are treated in Chapter IV.

Possible Applications of Chebyshev Estimation

In view of the optimality of the Chebyshev estimator under the assumption of independent identically distributed uniform errors, it is clear that Chebyshev estimation is a natural candidate whenever this situation is believed to exist. Uniformity of errors can be detected in data analytic situations through a half-normal plot of the least squares residuals. That is, if one plots the ordered absolute values of the least squares residuals versus
the corresponding expected order statistics from a standard normal parent, uniformity manifests itself in this curve as a sigmoid shape, as opposed to the straight line which is expected under normality.

A key situation where uniformity of errors is known to be an appropriate assumption is in the approximation of a tabled function. Suppose that a function \( y = f(x) \) has been computed or measured quite accurately and that rounded values have been tabulated. Then the only source of error in the tabulated values is due to roundoff. Roundoff error is known to be uniformly distributed between -0.5 and +0.5, in units of the last digit retained in the tabulated values. Thus, the Chebyshev criterion would be appropriate in the fitting of a function to approximate the tabled function. Even under more general conditions, the Chebyshev norm is the accepted norm of tabling. Even when there are possibly other sources of error which are not necessarily uniform, computer approximations of standard mathematical functions are routinely assessed in terms of their maximum error.

In many industrial applications where errors are controlled in some way, the Chebyshev criterion can be efficient. If errors are small relative to the "signal", uniformity is often a reasonable assumption. 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 errors with limited range in the accepted population. Depending on the flatness of the untruncated error distribution and upon the degree of truncation, Chebyshev estimation is again potentially very efficient. This contention will be supported by the results of a Monte Carlo study of relative efficiency to be reported in Chapter V.

Overview of the Remaining Chapters

To this point we have attempted to introduce Chebyshev estimation and to provide some insight as to its properties and potential usefulness. In the chapters which remain, we shall present algorithms for the solution of the Chebyshev problem, distributional results and characterizations of the Chebyshev solution, as well as a more extensive coverage of applications.

Chapter II will first deal lightly with the classical exchange algorithms of approximation theory. We shall then proceed to develop the linear programming formulation of the Chebyshev problem and to prove several properties of Chebyshev solutions which arise naturally out of the linear programming context.

The third chapter will discuss the algorithm of Barrodale
and Phillips (1974 and 1975). This algorithm is based on linear programming, but takes advantage of the special structure of the Chebyshev problem to produce considerable gains in efficiency over more general linear programming procedures. Chapter III will proceed on the basic assumption that the algorithm of Barrodale and Phillips (1974 and 1975) is the most efficient known. Two major types of modifications to their algorithm are proposed. The first type of modification involves the use of initial approximations to the Chebyshev solution to obtain a scaling and centering of the dependent variable $y$ which improves the rate of convergence of Barrodale and Phillips' algorithm. While the solution procedure is not changed significantly by this modification, a simulation study will demonstrate that considerable improvement in the rate of convergence is frequently obtained. The second type of modification involves the examination of alternative variable selection criteria which perform well in spite of the degeneracy which exists in the initial phase of Barrodale and Phillips' algorithm. In particular, we propose a variable selection technique which is based on a nonlinear optimization technique known as "gradient projection." Again, the improvements which can be expected in rate of convergence are demonstrated by a Monte Carlo study.
The question of "what is the exact distribution of the Chebyshev estimator?" remains and will continue to remain, making the derivation of exact tests of significance intractable. Of course, a major source of the difficulty is due to the nonexistence of a closed-form solution for the Chebyshev estimator. Chapter IV will present what is known about the distribution of the Chebyshev estimator. We are able to treat this theory rather thoroughly for the case of the simple location model (Model 1.2), but for the general linear model (Model 1.1) we must be satisfied to merely identify the nature of the distribution. Characterizations of the Chebyshev solution are derived which seem to have potential in the discovery of an asymptotic sampling theory; much in analogy to the work of Koenker and Bassett (1978) with respect to the $L_1$ estimator.

Chapter V will deal more extensively with applications of Chebyshev estimation, including a discussion of its use in tabling problems and for truncated error distributions as well as in standard data analytic problems where residual analysis indicates uniformity of errors. The efficiency of Chebyshev estimation relative to least squares for several different error distributions is examined via simulation study.

The purpose of Chapter VI is to present and discuss some of the most important unanswered questions which
remain in Chebyshev estimation. These are, most notably, the lack of significance tests and such questions as model selection and interpretation of coefficient estimates.
It was mentioned several times in the previous chapter that, in general, no closed-form solution exists for the Chebyshev problem. Thus, the solution must be obtained iteratively. Therefore, it is only natural that development of algorithms for Chebyshev estimation is a problem of considerable interest. The Chebyshev norm has long held a central position in classical approximation theory and the early work in algorithm development came primarily from researchers in this field. The principal objective of this chapter will be to provide, along with some measure of historical development, the necessary groundwork for the discussion of algorithm enhancements which are proposed in Chapter III.

The Classical Exchange Algorithm for Chebyshev Approximation

We begin by giving a brief survey of the classical theory of Chebyshev approximation adapted from Osborne and Watson (1967). The key assumption of the classical theory is that (with respect to Model 1.1) the design matrix, \( X \), satisfies the Haar condition.
Definition 2.1. A matrix $X$ with $n$ rows and $p$ columns is said to satisfy a Haar condition if any $pxp$ submatrix of $X$ is nonsingular.

Consider a slight rewriting of the model equations of Model 1.1

$$X\beta = y - e \quad (2.1)$$

Any set of $p+1$ equations from (2.1) will be termed a reference. Any reference will have corresponding a sub-matrix and two sub-vectors which we denote as $X_p$, $y_p$, and $e_p$ respectively. By the Haar condition on $X$, $X_p$ must have rank $p$ and therefore there exists a unique vector $\lambda$ (up to a scalar multiplier), each of whose components is different from zero, satisfying

$$\lambda'X_p = 0 \quad (2.2)$$

where $0$ denotes a null vector of appropriate dimension.

In (2.1) $\beta$ is called a reference vector if

$$\text{sgn}((e_p)_i) = \delta \cdot \text{sgn}(\lambda_i) \quad \forall \ i = 1,2,...,p+1$$

where $\delta$ is $+1$ and $\text{sgn}$ is defined as

$$\text{sgn}(t) = \begin{cases} 
1 & \text{if} \ t>0 \\
0 & \text{if} \ t=0 \\
-1 & \text{if} \ t<0.
\end{cases}$$

Define a vector $\eta$ as $\eta_i = \text{sgn}(\lambda_i) \quad \forall \ i = 1,2,...,p+1$. Then
the matrix \((X_p | n)\) is nonsingular so that the vector \(\hat{\beta}\) is uniquely defined by the equations

\[ X_p \hat{\beta} = Y_p - \varepsilon n. \]  \hspace{1cm} (2.3)

In this case, \(\hat{\beta}\) is termed a levelled reference vector and \(\varepsilon\) is termed a reference deviation.

It is interesting to note that in multiplying (2.3) through by \(\lambda'\) we obtain

\[ \lambda' Y_p = \varepsilon \sum_{i=1}^{p+1} |\lambda_i| \]

and that for any reference vector

\[ 0 = \lambda' Y_p + \sum_{i=1}^{p+1} |\lambda_i| \cdot |(e_p)_i| \]

so that \(|\varepsilon|\) lies between the least and the greatest of the \(|(e_p)_i|\).

The whole point of the foregoing development is that in the presence of a Haar condition on \(X\), the Chebyshev approximation can be expressed as the unique solution to Equation (2.3), a system of \(p+1\) equations in \(p+1\) unknowns, for an appropriate choice of \(p+1\) equations from (2.1).

This result was first obtained by De la Vallée-Poussin (1911) and is summarized in the following.
Theorem 2.1. Assuming that the coefficient matrix $X$ satisfies a Haar condition, the Chebyshev solution to Equation (2.1) is a levelled reference vector for some reference. Furthermore, $\max_{1 \leq i \leq n} |e_i| = |\varepsilon|$ where $\varepsilon$ is the corresponding reference deviation for this reference.

The following theorem, the Exchange Theorem, proved by Stiefel (1959) provides the necessary basis for what is known as the Stiefel exchange algorithm.

Theorem 2.2. Given any reference (set of $p+1$ equations from (2.1)) and a corresponding reference vector, it is possible to add to the reference any other equation and to drop an appropriate equation from the reference so that the given vector is also a reference vector for the new reference.

In computational practice, the chosen vector would be the levelled reference vector for the given reference, and the equation to be added would be that associated with the component of $\varepsilon$ with maximum modulus. If this equation is already in the reference then the calculation is complete. It was shown by Stiefel (1959) that, under this procedure, the reference deviation rises monotonically. Since there are but a finite number of references, the procedure is guaranteed to converge to the Chebyshev solution in a finite number of iterations.

It should be understood that the entire development
of the classical theory is done under the assumption of a Haar condition on the coefficient matrix X. In practice, particularly in the statistical context, this assumption is unrealistic and, as emphasized by Osborne and Watson (1967), completely unnecessary. Although the linear programming formulation of the $L_\infty$ program is fundamentally equivalent to the classical theory, it requires none of the restrictive assumptions.

The Linear Programming Formulation of the Chebyshev Problem

It is well-known that the Chebyshev problem can be formulated and solved as a linear programming problem. This has been shown by many authors (see e.g. Kelley (1958), Wagner (1959), Stiefel (1960), Appa and Smith (1973), and Sposito (1976)). Even though the linear programming formulation of the Chebyshev problem is equivalent to the classical theory, it represents a significant advance in the development of Chebyshev algorithms. Not only does the linear programming framework provide an improvement in the computational aspect of the problem, but it also greatly simplifies the theory involved. Thus, the statement and proof of many of the properties and characterizations of Chebyshev estimation have been deferred until this development is complete.
Let us again consider Model 1.1 as introduced in Chapter I. Recall that the objective is to determine \((\bar{g}, \bar{e})\) such that

\[
\bar{e} = \max_{1 \leq i \leq n} |(y - X\bar{g})_i| \leq \max_{1 \leq i \leq n} |(y - X\bar{b})_i| \quad \forall \in \mathbb{R}^p. \tag{2.4}
\]

Objective (2.4) can easily be re-expressed as

-\(\min\)ize \(e\)
  subject to \(\max_{1 \leq i \leq n} |(y - X\bar{g})_i| \leq e\)

or

-\(\min\)ize \(e\)
  subject to \(|(y - X\bar{g})_i| \leq e\) \quad \forall \ i = 1, 2, \ldots, n

or

-\(\min\)ize \(e\)
  subject to \((y - X\bar{g})_i \leq e\) \quad \forall \ i = 1, 2, \ldots, n\tag{2.5}
  \((y - X\bar{g})_i \geq -e\)

Rearrangement of the constraints in (2.5) leads to what is easily recognized to be a linear programming problem in primal form, that is, a problem of minimization of a function subject to linear equality or inequality constraints. We present the problem here as

**Problem 2.1.** Minimize \(e\)
subject to \(-(X\bar{g})_i + e \geq -y_i \quad \forall i = 1, 2, \ldots, n.\)

\((X\bar{g})_i + e \geq y_i\)

\(\beta_1, \ldots, \beta_p, e\) unrestricted
The Dual Formulation of the Chebyshev Problem

It is a well-known result of linear duality theory, that every well-formulated linear programming problem possesses an equivalent dual problem. While a full exposition of duality theory will not be attempted here, we shall instead present a very illuminating geometric example given by Luenberger (1969, page 9).

**Example 2.1.** Duality principles are based on the geometric relation illustrated in Figure 2.1. The shortest distance from a point to a convex set is equal to the maximum of the distances from the point to a hyperplane separating the point from the convex set.

---

*Figure 2.1. Duality*
Before proceeding to derive the dual form of Problem 2.1, we ought to justify our interest in so doing. We may notice that Problem 2.1 involves 2n constraints in p+1 variables. The corresponding dual problem will necessarily have p+1 constraints in 2n variables. While there are no strong theoretical results to cite, extensive computational experience (see e.g., Kuhn and Quandt (1963)) suggests that it is generally more efficient to solve the problem with the least number of constraints. It is a condition on Model 1.1 that the number of observations, n, is greater than the number of explanatory variables, p. In practice 2n is generally much greater than p+1 and thus we prefer to solve the Chebyshev dual problem.

It is convenient, at this point, to introduce shortened notation for some of the expressions of Problem 2.1. Define

\[ k' = (0,0,\ldots,0,1), \text{ a } (p+1)-\text{vector}, \]
\[ \beta' = (\beta_1,\beta_2,\ldots,\beta_p,\epsilon), \]
\[ c' = (-y_1,-y_2,\ldots,-y_n,y_1,y_2,\ldots,y_n), \]
\[ 1' = (1,1,\ldots,1), \text{ an } n-\text{vector}, \]
\[ A' = \begin{bmatrix} -X & 1 \\ X & 1 \end{bmatrix}. \]

In this notation Problem 2.1 becomes

\[
\begin{align*}
\text{minimize} & \quad k'\beta \\
\text{subject to} & \quad A'\beta \geq c, \quad \beta \text{ unrestricted.}
\end{align*}
\]
Sposito (1975) gives rules for obtaining the dual problem from the primal problem. The rules are to transpose the constraint matrix $A'$, change minimization to maximization, interchange $k$ and $c$, and reverse the inequality sign in the constraints. Sposito (1975) also shows that if the $i^{th}$ variable, $\theta_i$, is unrestricted in sign then the $i^{th}$ dual constraint is an equality. Applying these rules to (2.6) we obtain the Chebyshev dual problem,

$$\text{maximize } c'd$$
$$\text{subject to } Ad = k$$
$$d > 0 \quad (2.7)$$

where $d$ is a $2n$-vector denoting the dual variables. It is advantageous to think of $d$ as partitioned as $d' = (t', s') = (t_1, t_2, ..., t_n, s_1, s_2, ..., s_n)$. Reverting back to the longer, but more descriptive, notation we have

Problem 2.2. maximize $-y't + y's$
subject to $-X't + X's = 0$

$$\sum_{i=1}^{n} (t_i + s_i) = 1$$
$$t_i, s_i \geq 0 \quad \forall \ i = 1, 2, ..., n.$$

We have previously mentioned the equivalence of solving the primal and dual problems. Equivalence is used in the sense that, given the solution to the dual problem, one can easily obtain the solution to the primal. In theory, there
is a simple transformation from one solution to the other. In practice, not even this small amount of additional calculation is necessary since the primal solution can be read directly from the optimal dual tableau.

Properties of the Chebyshev Solution

Numerous properties of the Chebyshev solution are stated and proved quite naturally in the linear programming framework. We begin by first stating the Existence Theorem of linear programming, which we shall use to prove that an optimal solution for the Chebyshev problem always exists.

Theorem 2.4. (Existence Theorem of Linear Programming)
If Problem 2.1 and 2.2 each have at least one feasible solution (i.e. - at least one solution which satisfies the constraint equations) \( \theta^* \) and \( d^* \), respectively, then each problem possesses an optimal solution. Furthermore, the optima are equal.

Theorem 2.5. An optimal solution to the Chebyshev estimation problem always exists.

Proof. Consider the vector \( \theta^* = \left[ \varepsilon^* \right] \) where \( \varepsilon^* = \max_{1 \leq i \leq n} |y_i| \). \( \theta^* \) satisfies the constraints of Problem 2.1. Furthermore,
\( d^* = [t^*] \) where \( t^{**} = s^{**} = (0,0,...,0,1/2) \), is a feasible solution for Problem 2.2. Therefore, by Theorem 2.4, an optimal solution to the Chebyshev problem always exists.

The next result provides a very useful interpretation of the optimal dual solution.

**Theorem 2.6.** Let \((\tilde{\beta}, \tilde{\varepsilon})\) be the optimal solution to Problem 2.1. Consider the dual problem

\[
\begin{align*}
\text{maximize} & \quad ( -y', x') (t', s')' \\
\text{subject to} & \quad \begin{bmatrix} -X' & X' \\ 1' & 1' \end{bmatrix} \begin{bmatrix} t' \\ s' \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
& \quad t', s' > 0.
\end{align*}
\tag{2.8}
\]

Denote the optimal solution of (2.8) as \([\tilde{t} \tilde{s}']\). If \(\tilde{t}_i > 0\) (or \(\tilde{s}_i > 0\)), then the maximum deviation \(\tilde{\varepsilon} = \max_{1 \leq i \leq n} |(y-X\tilde{\beta})_i|\) occurs at the \(i^{th}\) sample point. Furthermore, the \(i^{th}\) observation, \(y_i\), will lie below (above) the fitted surface \(X\tilde{\beta}\).

**Proof.** By the principle of complementary slackness (see e.g. Cooper and Steinberg (1974), pp. 174-177) the cross product between the dual solution vector and the primal constraint must be the zero vector.
That is, \( (\tilde{e}', \tilde{s}') \begin{bmatrix} -y + x_0^0 - \frac{1}{\tilde{e}} \\ y - x_0^0 - \frac{1}{\tilde{e}} \end{bmatrix} = 0 \)

or componentwise,

\[
\tilde{e}_i (-y + x_0^0 - \frac{1}{\tilde{e}}) = 0 \quad \forall \quad i = 1, 2, \ldots, n. \tag{2.9}
\]

\[
\tilde{s}_i (y - x_0^0 - \frac{1}{\tilde{e}}) = 0
\]

In (2.9), if \( \tilde{e}_i > 0 \), for some fixed \( i \), then \( (-y + x_0^0 - \frac{1}{\tilde{e}}) \) must be equal to zero. But this implies that

\[
(X_0^0 - y)_i = \tilde{e} > 0. \tag{2.10}
\]

It is immediate from (2.10) that the deviation at the \( i \)th sample point is equal to the maximum deviation, \( \tilde{e} \), and that \( y_i \) lies below \( (X_0^0)_i \). The corresponding result for \( \tilde{s}_i > 0 \) follows similarly.

**Theorem 2.7.** For Model 1.1, there exists an optimal Chebyshev hyperplane, \( X_0^0 \), which is vertically equidistant, at a distance \( \tilde{e} \), from at least \( p+1 \) of the \( y_i \)'s, provided that the design matrix \( X \) has at least \( p \) distinct rows.

**Proof.** While the result holds as long as \( X \) has \( p \) distinct rows, the proof will be provided only for the case where the Chebyshev solution is unique.

Consider, Problem 2.2, the Chebyshev dual problem. By the linear programming formulation, there exists a
basic optimal solution; that is, an optimal solution with at most \( p+1 \) of the dual variables strictly greater than zero. At the same time, the assumption of uniqueness implies that none of the basic variables may be in the solution at a zero level, otherwise alternate optimal solutions exist. Therefore, it must be the case that exactly \( p+1 \) of the components of \( \tilde{\mathbf{g}} \) are strictly greater than zero. Thus, by Theorem 2.6, the maximum deviation, \( \tilde{\varepsilon} \), occurs for at least \( p+1 \) of the \( y_i \)'s.

The foregoing result, Theorem 2.7, is the equivalent of the classical characterization of De La Vallée-Poussin (1911) given in Theorem 2.1. The very simple idea that runs through these characterization results is that, as an alternative to solving Problem 2.1 for the optimal values of \( (\bar{\mathbf{g}}, \bar{\varepsilon}) \) we may, equivalently, solve the dual problem of determining the points at which the maximum deviation is to occur and, for each of these points, whether it is to lie above or below the fitted hyperplane, \( X\bar{\mathbf{g}} \).

Definition 2.2. The convex hull of \( n \) points in Euclidean \( p \)-space is defined as the smallest convex set in \( \mathbb{E}^p \) which contains each of the \( n \) points.
Corollary 2.1. With respect to Model 1.1, the p+1 rows of
the design matrix, X, which correspond to the observation
points that determine the optimal Chebyshev hyperplane,
$X_{\tilde{\beta}}$, are vertices of the convex hull of the n rows of X in
$\mathbb{E}^p$.

Proof. Consider the two hyperplanes parallel to $X_{\tilde{\beta}}$, one at a
vertical distance of $\tilde{\varepsilon}$ and the other at a vertical distance
of $-\tilde{\varepsilon}$. The optimality of $(\tilde{\beta}, \tilde{\varepsilon})$ implies that the constraints
of Problem 2.1 are binding; that is, $(X_{\tilde{\beta}})_{i} - \tilde{\varepsilon} < y_{i} < (X_{\tilde{\beta}})_{i} + \tilde{\varepsilon}$.
But, the p+1 points which determine $X_{\tilde{\beta}}$ lie on $X_{\tilde{\beta}} - \tilde{\varepsilon}$ and
$X_{\tilde{\beta}} + \tilde{\varepsilon}$, and thus, are vertices of the convex hull.

An extremely important consequence of Corollary 2.1,
one which must be borne in mind especially by statisticians,
is that the Chebyshev estimator is completely determined by
the convex hull of the observations. That is, the Cheby­
shev estimator is determined by the most extremal observa­
tions. At first glance, this may appear a very undesirable
property. For those who would apply an estimation criterion
indiscriminantly without knowing something about the
distribution of errors, the Chebyshev criterion can be
extremely inefficient.

As an example of the effects of extremal observations
upon the Chebyshev estimator, consider the following example
adapted from Appa and Smith (1973). Consider nine collinear
points and a single outlying observation configured as in Figure 2.2. If one was interested in fitting the linear model \( y = \beta_0 + \beta_1 x \), examination of the data would certainly indicate an estimated slope parameter of about one. But under the Chebyshev criterion and in the presence of an outlier the estimate \( \hat{\beta}_1 \) would be more on the order of minus one, completely counter to what one would expect.

The point is not that Chebyshev estimation is not useful. The optimality results obtained in Chapter I indicate the use of Chebyshev estimation under uniformity of errors and a simulation study presented in Chapter V will demonstrate the efficiency of Chebyshev estimation for other finite-range distributions. The point is that the Chebyshev estimator is extremely sensitive to outlying observations and errors should be checked for approximate uniformity before applying the Chebyshev estimation.

It is appropriate that we say something about conditions under which the Chebyshev estimator, \( \hat{\beta} \), is unique. In the theory of least squares, there is a very simple condition for the uniqueness of the least squares estimator, \( \hat{\beta} \). \( \hat{\beta} \) is unique, provided the design matrix, \( X \), has full column rank, \( p \). And, even when \( \hat{\beta} \) is not unique, the optimal least squares hyperplane \( X \hat{\beta} \) is always unique. The corresponding conditions for Chebyshev estimation are not so nice.
Figure 2.2. The effect of an outlying observation on the Chebyshev estimator
The weakest known condition sufficient for the uniqueness of the Chebyshev estimator, \( \hat{\beta} \), is that the design matrix, \( X \), satisfy a Haar condition. This is an extremely stringent condition not often satisfied by designed experiments. But, the difficulties arising are primarily theoretical, not practical. In practice, the Haar condition, is far from necessary. That is, unique solutions are obtained for a large number of problems which do not satisfy the Haar condition on \( X \). Furthermore, any linear programming based solution procedure for Chebyshev estimation provides, as a by-product, an indication of whether or not the solution obtained is unique. Thus, in practice, once the Chebyshev solution is obtained it is simple to determine if alternate optimal solutions exist.

In this chapter, we have discussed the linear programming formulation of the Chebyshev problem and properties which are natural consequences of this formulation. In Chapter III, the computational details of linear programming based Chebyshev algorithms will be covered. In addition, the efficient algorithm of Barrodale and Phillips (1974 and 1975) will be presented, as well as several proposed modifications.
CHAPTER III. DEVELOPMENT OF EFFICIENT ALGORITHMS FOR CHEBYSHEV ESTIMATION

All the algorithms and modifications thereof to be proposed in this chapter are based on the simplex algorithm of linear programming developed by Dantzig (1951). In order to effectively convey the concepts involved in the development of Chebyshev algorithms, it is unavoidable that we briefly discuss the simplex method. An excellent brief description of the simplex method is given by Kuhn and Quandt (1963). We shall follow this description quite closely. For more thorough discussions see Dantzig (1951), Hadley (1962), Dantzig (1963), Cooper and Steinberg (1974), and Sposito (1975).

The Simplex Algorithm

Consider the linear programming problem

**Problem 3.1.** maximize $c'x$

subject to $Ax \leq b$

$x > 0$

where the definitions and assumptions of the problem elements are described in the following exposition.

For the purposes of this exposition, with no loss of generality, the objective of a linear program is to determine
the maximum value of a linear form (the objective function) over a convex polyhedral set (the set of feasible vectors or feasible region). If a maximum is attained for some vector in the feasible region and if the feasible region has extreme points (basic feasible vectors) then the maximum is attained at an extreme point. An extreme point, which we shall also call a basic feasible vector, is simply a vertex of the feasible region. For a linear program with at least one extreme point, the simplex method examines a finite sequence of extreme points, passing from vertex to vertex along edges of the polyhedron while never decreasing the objective function. The algorithm terminates at the maximum or with the information that the objective function is unbounded.

The central feature of the algorithm is an operation (pivoting) which transforms an algebraic description of the set of feasible vectors with a specified extreme point into another algebraic description of the set specifying either the same or a neighboring extreme point. In order to introduce appropriate notation, assume that the feasible region is described as the solution set of a system of m linear inequalities in the positive orthant of an n-dimensional space. It is then possible to choose, as an initial extreme point, coordinates $x_1, x_2, \ldots, x_n$ such that the objective
function is \( z = c_1 x_1 + c_2 x_2 + \ldots + c_n x_n + t \), and the
constraints are satisfied as

\[
\begin{align*}
-a_{11} x_1 - a_{12} x_2 - \ldots - a_{1n} x_n + b_1 & \geq 0 \\
-\ldots & \ldots \\
-a_{m1} x_1 - a_{m2} x_2 - \ldots - a_{mn} x_n + b_m & \geq 0
\end{align*}
\]

where the constants \( b_1, b_2, \ldots, b_m \) are nonnegative. The
description is abbreviated conveniently in the condensed
simplex tableau as in Figure 3.1.

```
<table>
<thead>
<tr>
<th>-x_1</th>
<th>...</th>
<th>-x_n</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>a_{11}</td>
<td>...</td>
<td>a_{1n}</td>
<td>b_1</td>
</tr>
</tbody>
</table>
| ... | ... | ... | ... | ...
| a_{m1} | ... | a_{mn} | b_m | = v_m |
| -c_1 | ... | -c_n | t  | = z  |
```

Figure 3.1. Condensed simplex tableau

Notice that the variables \( x_j \) (\( j=1,2,\ldots,n \)) and \( v_i \) =
\(-a_{i1} x_1 - a_{i2} x_2 - \ldots - a_{in} x_n + b_i \) (\( i=1,2,\ldots,m \)) are restricted
to be nonnegative.

The pivot operation of the simplex algorithm effects an
exchange of the roles of some basic variable, \( v_i \), and some
nonbasic variable, \( x_j \), in this tableau. If the variables to
be exchanged are $v_r$ and $x_s$, then this is accomplished by solving for $x_s$ in the equation

$$v_r = -a_{rl}x_l - \ldots - a_{rs}x_s - \ldots - a_{rn}x_n + b_r$$

and substituting the result in the remaining equations. This produces the new tableau of Figure 3.2.

<table>
<thead>
<tr>
<th>$-x_1$</th>
<th>$-v_r$</th>
<th>$-x_n$</th>
<th>l</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{ll}$</td>
<td>$a_{ls}$</td>
<td>$a_{ln}$</td>
<td>$b'_1$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$a_{rl}$</td>
<td>$a_{rs}$</td>
<td>$a_{rn}$</td>
<td>$b'_r$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$a_{ml}$</td>
<td>$a_{ms}$</td>
<td>$a_{mn}$</td>
<td>$b'_m$</td>
</tr>
<tr>
<td>$-c'_1$</td>
<td>$-c'_s$</td>
<td>$-c'_n$</td>
<td>$b'$</td>
</tr>
</tbody>
</table>

Figure 3.2. Transformed simplex tableau

The rules for calculating the entries in the new tableau follow:

$$\begin{align*}
    a'_rs &= 1/a_{rs} \\
    a'_{rj} &= a_{rj}/a_{rs} \quad \text{for } j \neq s \\
    b'_r &= b_r/a_{rs}
\end{align*}$$

(3.1) (3.2)
These rules have been organized into four types applying (3.1) to the pivot element \( a_{rs} \); (3.2) to the pivot row; (3.3) to the pivot column; and (3.4) to elements which are in neither the pivot row nor the pivot column.

The transformed tableau can be interpreted exactly as the original. The updated coordinates are \( x_1', ..., v_r', ..., x_n \). The objective function becomes

\[
    z = c_1'x_1 + ... + c_r'v_r + ... + c_n'x_n + t' .
\]

And, under the new coordinate system, the feasible region is described by the \( m \) inequalities

\[
    -a_{11}'x_1 - ... - a_{is}'v_r - ... - a_{ln}'x_n + b_1' \geq 0 \\
    ... ... ... ... \\
    a_{m1}'x_1 - ... - a_{ms}'v_r - ... - a_{mn}'x_n + b_m' \geq 0.
\]
If the pivot element, $a_{rs}$, is chosen so that the updated right hand sides $b'_1, b'_2, \ldots, b'_m$ are nonnegative, then the origin $x_1 = \ldots = v_r = \ldots = x_n = 0$ (with respect to the transformed coordinate system) is an extreme point of the feasible region. If the pivot is such that $t' > t$, then the objective function has not been decreased by the pivot operation. Selection of a pivot element satisfying both conditions and performing the pivot operation corresponds to a single iteration of the simplex algorithm.

All that remains to complete the description of the algorithm is a procedure for selecting a pivot element, $a_{rs}$. While there is in general no unique choice of a pivot element, there do exist some very simple conditions which are necessary and sufficient to insure that feasibility is maintained and that the objective function does not decrease.

To maintain feasibility in the updated tableau, all that is required is that the right hand sides $b'_1, b'_2, \ldots, b'_m$ remain nonnegative. That is,

$$
\begin{align*}
b_i' &= \begin{cases} 
  b_i / a_{is} & i = r \\
  b_i - a_{is} b_r / a_{rs} & i \neq r 
\end{cases} 
\end{align*}
$$

must be greater than or equal to zero for $i = 1, 2, \ldots, m$. If $b_r = 0$, the program is said to be degenerate. In this case, any $a_{rij} \neq 0$ in the row of $b_r$ may serve as a pivot element. We might also notice from (3.4) that, in this
case, the objective function is not increased. If \( b_i > 0 \)
for all \( i = 1, 2, \ldots, m \), it is clear from (3.5) that the pivot
element must be positive and that
\[
\frac{b_r}{a_{rs}} < \frac{b_i}{a_{is}} \quad \text{for all } i \text{ with } a_{is} > 0. \tag{3.6}
\]

This condition (3.6), leads to a rule which, given the
pivot column, determines the pivot row, essentially uniquely.
This rule is often referred to as the minimum ratio
rule. In the case that ties occur for the minimum ratio,
any method for breaking ties is acceptable.

To satisfy the requirement that the objective function
not decrease, it is sufficient (from (3.4)) that
\[
t' = t + \frac{c_s b_r}{a_{rs}} \geq t
\]
or
\[
c_s b_r / a_{rs} \geq 0 \tag{3.7}
\]
In the case of degeneracy, there are no restrictions other
than \( a_{rs} \neq 0 \). But, in the typical case of strictly positive
right hand sides and positive pivot, it is necessary that
\[
-c_s \leq 0. \tag{3.8}
\]
If there are no negative marginal costs in the current
tableau - that is, if \( -c_j \geq 0 \) for \( j = 1, 2, \ldots, n \) - then the
The current tableau is optimal.

The Construction of Effective Pivot Selection Rules

In general, there may be, for each step of the simplex algorithm, a number of legal pivot choices. The proliferation of possible pivot choices leads to a wide variation in the number of iterations required to solve a given linear program. Thus one ought to be interested in deriving criteria for choosing pivot elements which will cause the iterative procedure to converge rapidly. Possibly the most natural criterion is to maximize the increase in the objective function at each iteration.

We must emphasize that, given any candidate pivot column, the pivot row must be selected by the minimum ratio rule if feasibility is to be maintained. Assuming that some rule (any rule) has been devised for breaking ties among ratios, the pivot row is then determined as a function of the pivot column. Thus, the selection of a pivot element essentially reduces to the selection of a pivot column. The notation \( r(s) \) will be used, whenever it is needed for clarity, to denote the selection of the \( r \text{th} \) row by the minimum ratio rule applied in the \( s \text{th} \) column.

Two pivot selection rules which are in common use are known as the maximum increase rule and the standard
simplex rule. The maximum increase rule chooses a pivot \( a_{r(s),s} \) such that the objective function is increased as much as possible at each iteration. The maximum increase rule is given by

\[
\left\{ a_{r(s),s} : \frac{c^b r}{a_{r(s),s}} > \frac{c^b i}{a_{i(j),j}}, \forall j \neq -c_j < 0 \right\} 
\]

(3.9)

The standard simplex pivot selection rule is given by

\[
\left\{ a_{r(s),s} : -c_s \leq -c_j, \forall j \neq -c_j < 0 \right\} 
\]

(3.10)

That is, we select the column with the most negative cost coefficient. While this rule does not in general increase the objective function as rapidly as (3.9), it requires considerably less calculation at each iteration. The standard simplex rule was recommended by Dantzig (1951) and studies as well as computational experience with the simplex method indicate that (3.10) may be more efficient than (3.9) in terms of total computation time for general applications (see e.g. Kuhn and Quandt (1963)).

Note that neither of the two pivot selection rules we have described can effect an increase in the objective function if any of the right hand sides \( b_1, b_2, \ldots, b_m \) are zero. In this case, the selection of a pivot element requires much more thought. The rules described by (3.9) and (3.10) can still be applied to obtain a new feasible
extreme point, but there is no guarantee that a pivot so chosen moves in a direction approaching the maximum. In the Chebyshev problem, initially p of the right hand sides are zero. Thus the problem of degenerate tableaux is one we shall give further consideration in the context of the Chebyshev problem.

The Simplex Method Applied to the Chebyshev Problem

In Chapter II, we discussed the linear programming formulation of the Chebyshev problem. Earlier in this chapter, the groundwork was laid for the description of computational procedures for the solution of the Chebyshev problem via linear programming. The goal of the current section is to treat the computational details associated with the Chebyshev problem.

Consider the Chebyshev dual problem, Problem 2.2. Problem 2.2 can be expressed as a condensed simplex tableau, just as Problem 3.1 was expressed in Figure 3.1. The condensed tableau for Problem 2.2 is expressed here, in matrix notation, as Figure 3.3.

There are two major points to be made at this stage of the development. First of all, it is necessary to identify an initial feasible extreme point. Also, we may
Figure 3.3. Condensed simplex tableau for Chebyshev dual problem (matrix notation)

<table>
<thead>
<tr>
<th>t'</th>
<th>s'</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>-X'(pxn)</td>
<td>X'(pxn)</td>
<td>0 (px1)</td>
</tr>
<tr>
<td>1'(lxn)</td>
<td>1'(lxn)</td>
<td>1</td>
</tr>
<tr>
<td>Y'(lxn)</td>
<td>-Y'(lxn)</td>
<td>0</td>
</tr>
</tbody>
</table>

notice that all but one of the right hand sides in Figure 3.3 are zero. This indicates a high degree of degeneracy in the initial tableau. While this degeneracy will prove not to pose a serious difficulty in obtaining a solution via the standard simplex algorithm, it is possible that recognition of this degeneracy can lead to solution procedures which are more efficient.

The description of the simplex method is based on the assumption that one has identified an initial feasible extreme point and that the initial tableau is represented in terms of the corresponding basis. That is, before the simplex algorithm can be initiated it is necessary to either identify or create an identity matrix structure in the tableau. When no natural starting basis exists, it is
possible to build one by adjoining artificial variables and applying a two-phase solution procedure. We shall not discuss artificial variables in any generality, but rather just in the context of the Chebyshev problem. For more general treatments see Hadley (1962), Dantzig (1963), Cooper and Steinberg (1974), and Sposito (1975).

There is, in fact, no natural starting basis for the Chebyshev dual problem (Problem 2.2, Figure 3.3). We can create an artificial basis by adjoining, to the dual tableau (Figure 3.3), a \((p+1) \times (p+1)\) identity matrix whose columns are to correspond to \(p+1\) artificial variables. Let us denote the artificial variables by \(a_1, a_2, \ldots, a_{p+1}\). Then the dual tableau can be expressed in matrix notation as in Figure 3.4, where we have now added a column on the left to indicate which variables are in the current basis.

<table>
<thead>
<tr>
<th>B</th>
<th>(t')</th>
<th>(s')</th>
<th>(a')</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>(\mathbf{-X'}(pxn))</td>
<td>(\mathbf{X'}(pxn))</td>
<td>(\mathbf{I}(p+1)\times(p+1))</td>
<td>(0)</td>
</tr>
<tr>
<td>(l'\times(1xn))</td>
<td>(\mathbf{l'}(1xn))</td>
<td></td>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>(-)</td>
<td>(\mathbf{y'}(1xn))</td>
<td>(\mathbf{-y'}(1xn))</td>
<td>(0')</td>
<td>(0)</td>
</tr>
</tbody>
</table>

Figure 3.4. Simplex tableau with artificial variables adjoined
Starting with the tableau given in Figure 3.4, an initial feasible basis can be created by replacing each artificial variable in the basis by some legitimate variable from \textit{torsors}. This replacement is done by the appropriate pivoting operation. Once an artificial variable has left the basis, it is not allowed to reenter. Thus we obtain a two-phase procedure which builds an initial feasible basis in the first phase and then proceeds to solve the problem by the standard simplex algorithm. The solution $-\beta_1, \beta_2, \ldots, \beta_p, \tilde{c}$ to the primal problem is obtained from the optimal tableau as the final marginal costs associated with $a_1, a_2, \ldots, a_p, a_{p+1}$, respectively.

The method of adjoining artificial variables is a standard method in linear programming. Thus it is possible to obtain solutions to the Chebyshev problem using completely general procedures. We have yet to exploit any of the specific structure of the Chebyshev problem. The general algorithm can, in fact, be specialized to obtain considerable gains in efficiency. Barrodale and Phillips (1974 and 1975) constructed an efficient three-stage algorithm which makes use of the peculiarities of the Chebyshev problem. This algorithm will be the subject of the next section.
Barrodale and Phillips (1974 and 1975) describe a three-stage algorithm for the solution of the Chebyshev approximation problem which is very efficient, both in terms of computer storage and solution time. In their algorithm, they exploit the structure of the Chebyshev problem to achieve a gain in efficiency. They begin by noticing that, of the $2n + p+1$ columns representing $t$, $s$, and $a$ in the dual tableau (Figure 3.4), only $n$ of them need be retained in storage since the remaining columns can be easily reconstructed. It is immediate that the $p+1$ columns in the current basis can always be suppressed since they always take the form of a $(p+1) \times (p+1)$ identity matrix. Furthermore, the right hand side vector, $R$, is equal to the column corresponding to $a_{p+1}$ and $t_i$ can be reconstructed as $2a_{p+1}s_i$ for each $i = 1, 2, \ldots, n$. Thus, we need store only the columns of the tableau corresponding to $s_1, s_2, \ldots, s_n$. In retrospect, while the use of those relationships is certainly clever, unless the problem is extremely large, the savings in storage is not likely to be worth the additional time and inconvenience required to reconstruct suppressed columns. The algorithm can be described analogously for either the full or condensed
tableau. We prefer to deal in terms of the full tableau, as pictured in Figure 3.4.

In the description of Barrodale and Phillips' algorithm, it may be helpful to have a full representation of the tableau being employed. The full tableau is given in Figure 3.5. The algorithm can be divided into three consecutive stages, the first consisting of \( p \) simplex iterations, the second being just a single simplex iteration, with the third stage consisting of the remaining simplex iterations required to attain optimality.

In the first stage, only legitimate variables, \( d^j \), (recall \( d' = (s'|t') \)) are allowed to enter the basis, while only artificial variables, \( a_i \), are allowed to leave. Specifically, in each of the first \( p \) iterations, the variable to enter the basis is selected as that having the most negative marginal cost. This corresponds with the standard simplex rule (3.10), but the row selection is somewhat different. The pivot element is selected from the first \( p \) elements of the pivot column as that having the largest absolute value opposite an artificial variable, \( a_i \), in the basis. Note that there is no problem in selecting a negative pivot element since each of the right hand sides is zero. It doesn't really matter which artificial variable is removed at each iteration since, eventually, all of
them will be replaced. Thus the pivot is chosen as the maximum column pivot to provide a degree of numerical stability.

At the completion of the first stage we have done nothing more than establish an initial feasible extreme point. Because of the degeneracy in the tableau (zero right hand sides), there is no indication that we have selected the initial feasible extreme point efficiently with respect to progress toward the optimum. In fact, the pivot selection is governed primarily by numerical stability considerations rather than optimality considerations. Of course, the standard pivot selection rules, (3.9) and (3.10), simply are not meaningful in the presence of degeneracy. We underscore this point here, since we shall later propose two modifications which attempt to construct "good" initial feasible extreme points.

After the completion of stage one, it is possible to iterate the resulting tableau, by the standard simplex rule, (3.10), until optimality is attained. But, no increase in the objective function is possible as long as we pivot in rows with zero right hand sides. Thus, we would like to pivot on an element in the \((p+1)^{st}\) row. This would allow an increase in the objective function and would also resolve most, if not all, of the degeneracy by changing
<table>
<thead>
<tr>
<th>Basis</th>
<th>( t_1 )</th>
<th>( t_2 )</th>
<th>...</th>
<th>( t_n )</th>
<th>( s_1 )</th>
<th>( s_2 )</th>
<th>...</th>
<th>( s_n )</th>
<th>( a_1 )</th>
<th>( a_2 )</th>
<th>...</th>
<th>( a_p )</th>
<th>( a_{p+1} )</th>
<th>( R )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_1 )</td>
<td>(-x_{11})</td>
<td>(-x_{21})</td>
<td>...</td>
<td>(-x_{n1})</td>
<td>(x_{11})</td>
<td>(x_{21})</td>
<td>...</td>
<td>(x_{n1})</td>
<td>(1)</td>
<td>(0)</td>
<td>...</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>(-x_{12})</td>
<td>(-x_{22})</td>
<td>...</td>
<td>(-x_{n2})</td>
<td>(x_{12})</td>
<td>(x_{22})</td>
<td>...</td>
<td>(x_{n2})</td>
<td>(0)</td>
<td>(1)</td>
<td>...</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>...</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>...</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>...</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>...</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( a_p )</td>
<td>(-x_{1p})</td>
<td>(-x_{2p})</td>
<td>...</td>
<td>(-x_{np})</td>
<td>(x_{1p})</td>
<td>(x_{2p})</td>
<td>...</td>
<td>(x_{np})</td>
<td>(0)</td>
<td>(0)</td>
<td>...</td>
<td>(1)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>( a_{p+1} )</td>
<td>(1)</td>
<td>(1)</td>
<td>...</td>
<td>(1)</td>
<td>(1)</td>
<td>(1)</td>
<td>...</td>
<td>(1)</td>
<td>(0)</td>
<td>(0)</td>
<td>...</td>
<td>(0)</td>
<td>(1)</td>
<td>(1)</td>
</tr>
<tr>
<td>Marginal costs</td>
<td>(y_1)</td>
<td>(y_2)</td>
<td>...</td>
<td>(y_n)</td>
<td>(-y_1)</td>
<td>(-y_2)</td>
<td>...</td>
<td>(-y_n)</td>
<td>(0)</td>
<td>(0)</td>
<td>...</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
</tbody>
</table>

Figure 3.5. Full dual Chebyshev tableau
the right hand sides. Stage two consists of a single simplex iteration, pivoting in the \((p+1)^{st}\) row. The pivot column is selected as that with the most negative marginal cost.

In stage two the pivot row has been selected without regard to the minimum ratio rule. Thus, there is no guarantee that feasibility will be maintained after pivoting. In fact, feasibility is lost unless each of the first \(p\) elements in the pivot column is nonpositive. Due to the special structure of the Chebyshev problem, this can easily be arranged. For each of the first \(p\) rows of the tableau for which the pivotal column element is positive, add twice the row to the pivotal row (the \(p+1^{st}\) row) and change the sign of each element of the original row. This operation corresponds to replacing a variable \(s_j\) (or \(t_j\)) in the basis by its corresponding \(t_j\) (or \(s_j\)). The labels of the basis variables should be changed accordingly.

In the third stage, we simply apply the standard simplex pivot selection rule (3.10); that is, the pivot column is determined by the most negative marginal cost and the pivot column is determined by the minimum ratio rule. The algorithm remains in stage three, iterating until optimality is attained.

This completes the description of Barrodale and
Phillips' algorithm. In order to simplify the description, we have omitted some minor details which pertain to special cases. It is worthwhile to consider the interpretation of quantities in the intermediate tableaux. Each iteration produces a basic feasible solution, $b$, to the primal problem (Problem 2.1). This feasible solution can be found as the marginal costs corresponding to the columns containing the artificial variables. The remaining updated marginal costs have a nice interpretation in terms of residuals about the primal solution.

For convenience, let us define the notation $c(u)$.

$$c(u) = \text{the marginal cost corresponding to the column in the dual tableau labelled } u. \tag{3.11}$$

Thus, $c(t_i)$, with respect to the original tableau (Figure 3.4), is $y_i$. The result is that

$$c(a_{p+1}) - c(s_i) = c(t_i) - c(a_{p+1}) = r_i$$

$$= (y-Xb)_i \quad (\forall \ i = 1,2,...,n), \tag{3.12}$$

the $i^{th}$ residual of $y$ about the feasible hyperplane, $Xb$. It follows from (3.12) that the dual variable chosen to enter the basis at each iteration corresponds to selecting the point of largest deviation about the current feasible primal solution. In fact, stage three is equivalent to
the classical exchange algorithm described in Chapter II.

While the algorithm of Barrodale and Phillips is very efficient and converges quite rapidly, it is felt that improvements can be made, particularly with respect to the construction of an initial feasible basis in stage one. The following two sections will deal with modifications proposed to improve the algorithm.

**Using an Approximate Chebyshev Estimator to Accelerate Convergence in the Algorithm of Barrodale and Phillips**

In this section we shall show that if one has an approximation, \( \hat{\beta} \), to the optimal Chebyshev estimator, \( \beta^* \), the approximation can be used as a starting point for Barrodale and Phillips' algorithm and can thereby reduce the number of iterations required to obtain the optimal Chebyshev estimator. Although the least squares estimator is a natural initial approximation, we emphasize that the notation \( \hat{\beta} \) is used here to denote any initial approximation. After discussing the procedure for incorporating an initial approximation into the algorithm, we shall present the results of a simulation study which examines the effect of using the least squares estimator as a starting point.
The basis for the modification to be proposed is provided by the following theorem.

**Theorem 3.1.** Consider the general linear model, Model 1.1, and the estimation of $\hat{\beta}$ under the criterion of minimizing $||y-X\hat{\beta}||$ where $||.||$ is any vector norm. Let $\hat{\beta}$ be any p-vector taken as an estimate of $\beta$. Define $\hat{e}$ as $\hat{e} = y-X\hat{\beta}$, the residuals about the approximate solution. If $\hat{\alpha}$ is the minimum norm estimator (with respect to the same norm, $||.||$) of $\alpha$ in the model

$$\hat{e} = X\hat{\alpha} + \hat{e},$$

then the minimum norm estimator $\hat{\beta}$ of $\beta$ is given by

$$\hat{\beta} = \hat{\beta} + \hat{\alpha}.$$  

**Proof.** $\min_{\beta} ||y-X\beta|| = \min_{\alpha} ||y-X(\hat{\beta} + \alpha)||$

$$= \min_{\alpha} ||\hat{e} - X\alpha||$$

$$= ||\hat{e} - X\alpha||$$

$$= ||y-X\hat{\beta} - X\alpha||$$

$$= ||y-x(\hat{\beta} + \hat{\alpha})||.$$  

Thus, $\hat{\beta} = \hat{\beta} + \hat{\alpha}$. 
A consequence of Theorem 3.1 is that, in the formulation of Barrodale and Phillips' algorithm, we can take any initial approximation, \( \hat{\beta} \), to the Chebyshev estimator, replace the vector of dependent variables, \( y \), by the residuals about \( X^\delta \), solve the resulting modified problem, and then recover the solution to the original problem by adding the initial approximation, \( \hat{\beta} \), back on. Furthermore, the adjustment of the final solution for \( \hat{\beta} \) can be done automatically by inserting \( \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p \) in the initial tableau (Figure 3.5) as the marginal costs corresponding to \( a_1, a_2, \ldots, a_p \), respectively. Thus, what is suggested is that, if an initial approximation, \( \hat{\beta} \), to the estimator, \( \hat{\beta} \), can be obtained, the initial tableau (Figure 3.5) is modified as in Figure 3.6.

Within the framework we have set up for using an initial approximation, Barrodale and Phillips' unmodified algorithm can be viewed as a special case with \( \hat{\beta} = 0 \). In the discussion to follow, this situation will be referred to as using a null starting value.

While it should be clear, at this point, that it is possible to use an initial approximation as described, it may not be so obvious why such a procedure is likely to be beneficial. We have noticed that both the classical exchange algorithm and Barrodale and Phillips' algorithm
### Figure 3.6. Chebyshev dual tableau, modified for an initial approximation to $\hat{\beta}$

| Basis | $t_1$ | $t_2$ | ... | $t_n$ | $s_1$ | $s_2$ | ... | $s_n$ | $a_1$ | $a_2$ | ... | $a_p$ | $a_{p+1}$ | R |
|-------|-------|-------|-----|-------|-------|-------|-----|-------|-------|-------|-----|-------|-------|-------|-----|
| $a_1$ | $-x_{11}$ | $-x_{21}$ | ... | $-x_{n1}$ | $x_{11}$ | $x_{21}$ | ... | $x_{n1}$ | 1 | 0 | ... | 0 | 0 | 0 |
| $a_2$ | $-x_{12}$ | $-x_{22}$ | ... | $-x_{n2}$ | $x_{12}$ | $x_{22}$ | ... | $x_{n2}$ | 0 | 1 | ... | 0 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| $a_p$ | $-x_{1p}$ | $-x_{2p}$ | ... | $-x_{np}$ | $x_{1p}$ | $x_{2p}$ | ... | $x_{np}$ | 0 | 0 | ... | 1 | 0 | 0 |
| $a_{p+1}$ | 1 | 1 | ... | 1 | 1 | 1 | ... | 1 | 0 | 0 | ... | 0 | 1 | 1 |

| Marginal costs | $\hat{e}_1$ | $\hat{e}_2$ | ... | $\hat{e}_n$ | $-\hat{e}_1$ | $-\hat{e}_2$ | ... | $-\hat{e}_n$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | ... | $\hat{\beta}_p$ | 0 | 0 |
operate by attempting to identify the sample points at which the maximum residual occurs. Variables which exhibit the largest deviations about some intermediate approximation are selected for entry into the basis. When a variable which will not be present in the optimal basis is selected for entry we will refer to it as a selection error. If one observes the iteration process for a number of different problems, it is seen that a large number of selection errors are made in the early stages, but as the approximation is improved the selection error rate decreases. Thus it is reasonable to assume that if a good initial approximation is supplied, the number of selection errors, and hence iterations, are likely to be reduced. Essentially all that is being proposed is a scaling and centering of $y$ which is likely to have better convergence properties.

A small simulation study was conducted to assess the performance of Barrodale and Phillips' algorithm with a null starting value versus the least squares estimator as a starting value. In order to determine if the effect of the starting values depends in any way on the problem size, the study was conducted for a number of different values of $n$ (number of observations) and $p$ (number of parameters or explanatory variables). The conduct of the study is as follows.
For each fixed n and p, 250 random problems were generated for solution under each of two starting values; \( \hat{\beta} = 0 \) and \( \hat{\beta} = \{\text{the least squares estimator}\} \). For each problem a true solution, \( \beta' = (\beta_1', \beta_2', \ldots, \beta_p') \), was generated at random such that \( \beta_1', \beta_2', \ldots, \beta_p' \) are independent identically distributed as Uniform \((-3,3)\). The design matrix, \( X \), was also generated at random for each problem such that \( x_{il} = 1 \) for \( i = 1,2,\ldots,n \) and \( x_{ij} \) are independent and identically distributed as Uniform \((-10,10)\) for \( i = 1,2,\ldots,n \) and \( j = 2,3,\ldots,p \). Finally, the dependent variables were generated as \( y_i = (X\beta)'_i + e_i \) where \( e_i \) are independent identically distributed as Uniform \((-1,1)\) for \( i = 1,2,\ldots,n \). Each of the problems so generated was solved using both the null starting value and the least squares starting value. The problems were solved on an IBM 370/Model 158 using double precision arithmetic by the computer program published by Barrodale and Phillips (1975).

Tables 3.1, 3.2, and 3.3 summarize the results of the simulation study for the values \( p = 2, 5, \) and 10, respectively. Procedure I denotes the use of the null starting value, while Procedure II denotes the use of the least squares starting value. Quantities exhibited in the tables are the minimum, maximum, mean and median number of iterations required for the set of problems solved by each procedure for each problem size. Also included are comparison
figures which indicate the number of times Procedure II required fewer, the same, or more iterations for solution than Procedure I. The dominance ratio is defined as the fraction of the cases in which the given procedure required as few or fewer iterations for solution. Naturally, the procedure with the higher dominance ratio is to be preferred on the basis of iteration counts.

The total central processor time required to solve 250 problems under each procedure was recorded. The first set of time figures includes for Procedure II, the amount of time required to calculate the least squares starting value. The second set of time figures have the time required to calculate the least squares estimator subtracted out. That is, the second set of figures assume that the least squares estimator is known. Finally, CPU efficiency ratios are calculated. Each is simply the amount of CPU time required for Procedure I divided by the amount of CPU time required for Procedure II. Values less than one favor Procedure I. Values greater than one favor Procedure II.

The results of the study show very clearly that iterations are saved by using the least squares starting value. But, in terms of CPU time, the savings are offset by the additional time required to obtain the least squares estimator. However, considering the dominant role of
Table 3.1. Comparison summary for null (I) versus least squares (II) starting values (p=2)

<table>
<thead>
<tr>
<th>Observations (n)</th>
<th>Procedure</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>10</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
<td>I</td>
</tr>
<tr>
<td>Min Iteration</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Max Iteration</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Mean Iteration</td>
<td>4.48</td>
<td>4.01</td>
<td>5.18</td>
<td>4.82</td>
<td>5.84</td>
<td>5.14</td>
<td></td>
</tr>
<tr>
<td>Median Iteration</td>
<td>4.42</td>
<td>4.01</td>
<td>5.68</td>
<td>4.87</td>
<td>5.69</td>
<td>5.13</td>
<td></td>
</tr>
<tr>
<td>Compared to (I)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fewer Iterations</td>
<td>-</td>
<td>101</td>
<td>-</td>
<td>148</td>
<td>-</td>
<td>141</td>
<td></td>
</tr>
<tr>
<td>Same</td>
<td>-</td>
<td>131</td>
<td>-</td>
<td>75</td>
<td>-</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>More Iterations</td>
<td>-</td>
<td>18</td>
<td>-</td>
<td>27</td>
<td>-</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>Dominance Ratio</td>
<td>.56</td>
<td>.93</td>
<td>.41</td>
<td>.89</td>
<td>.44</td>
<td>.83</td>
<td></td>
</tr>
<tr>
<td>CPU Time (sec)</td>
<td>4.34</td>
<td>4.96</td>
<td>20.58</td>
<td>20.50</td>
<td>39.51</td>
<td>40.98</td>
<td></td>
</tr>
<tr>
<td>CPU Eff Ratio</td>
<td>-</td>
<td>.88</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>1.04</td>
<td></td>
</tr>
<tr>
<td>CPU Time, (\hat{\beta})</td>
<td>4.34</td>
<td>4.15</td>
<td>20.58</td>
<td>17.52</td>
<td>39.51</td>
<td>36.44</td>
<td></td>
</tr>
<tr>
<td>CPU Eff Ratio, (\hat{\beta})</td>
<td>-</td>
<td>1.05</td>
<td>-</td>
<td>1.18</td>
<td>-</td>
<td>1.08</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.2. Comparison summary for null (I) versus least squares (II) starting values (p=5)

<table>
<thead>
<tr>
<th>Observations (n)</th>
<th>15</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedure</td>
<td>I</td>
<td>II</td>
<td>I</td>
</tr>
<tr>
<td>Min Iteration</td>
<td>6</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Max Iteration</td>
<td>14</td>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>Mean Iteration</td>
<td>10.05</td>
<td>8.94</td>
<td>13.35</td>
</tr>
<tr>
<td>Median Iteration</td>
<td>10.02</td>
<td>8.82</td>
<td>13.20</td>
</tr>
<tr>
<td>Compared to (I)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fewer Iterations</td>
<td>-</td>
<td>152</td>
<td>-</td>
</tr>
<tr>
<td>Same</td>
<td>-</td>
<td>61</td>
<td>-</td>
</tr>
<tr>
<td>More Iterations</td>
<td>-</td>
<td>37</td>
<td>-</td>
</tr>
<tr>
<td>Dominance Ratio</td>
<td>.40</td>
<td>.85</td>
<td>.26</td>
</tr>
<tr>
<td>CPU Time</td>
<td>17.44</td>
<td>19.31</td>
<td>61.60</td>
</tr>
<tr>
<td>CPU Eff Ratio</td>
<td>-</td>
<td>.90</td>
<td>-</td>
</tr>
<tr>
<td>CPU Time, ( \hat{\theta} )</td>
<td>17.44</td>
<td>15.61</td>
<td>61.60</td>
</tr>
<tr>
<td>CPU Eff Ratio, ( \hat{\theta} )</td>
<td>-</td>
<td>1.12</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.3. Comparison summary for null (I) versus least squares (II) starting values (p=10)

<table>
<thead>
<tr>
<th>Observations (n) Procedure</th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
</tr>
<tr>
<td>Min Iteration</td>
<td>12</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>Max Iteration</td>
<td>26</td>
<td>26</td>
<td>38</td>
</tr>
<tr>
<td>Mean Iteration</td>
<td>18.34</td>
<td>16.41</td>
<td>25.58</td>
</tr>
<tr>
<td>Median Iteration</td>
<td>18.19</td>
<td>16.16</td>
<td>25.46</td>
</tr>
<tr>
<td>Compared to (I)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fewer Iterations</td>
<td>-</td>
<td>179</td>
<td>-</td>
</tr>
<tr>
<td>Same</td>
<td>-</td>
<td>30</td>
<td>-</td>
</tr>
<tr>
<td>More Iterations</td>
<td>-</td>
<td>41</td>
<td>-</td>
</tr>
<tr>
<td>Dominance Ratio</td>
<td>.28</td>
<td>.84</td>
<td>.23</td>
</tr>
<tr>
<td>CPU Time</td>
<td>53.82</td>
<td>67.28</td>
<td>170.31</td>
</tr>
<tr>
<td>CPU Eff Ratio</td>
<td>-</td>
<td>.80</td>
<td>-</td>
</tr>
<tr>
<td>CPU Time, ( \hat{\beta} )</td>
<td>53.82</td>
<td>50.50</td>
<td>170.31</td>
</tr>
<tr>
<td>CPU Eff Ratio, ( \hat{\beta} )</td>
<td>-</td>
<td>1.07</td>
<td>-</td>
</tr>
</tbody>
</table>
least squares estimation in regression estimation, it would not be unusual to assume that as a matter of routine one would wish to have the least squares estimator for comparison purposes. Thus, if the least squares estimator were required, it could be used to obtain the Chebyshev estimator more efficiently. In any event, the workability of the principle is confirmed. As was previously emphasized, any initial approximation to the Chebyshev estimator may be used.

The Gradient Projection Pivot Selection Technique

In an attempt to discover efficient algorithms for Chebyshev estimation, it is natural to examine some of the classical techniques of nonlinear optimization. A large class of techniques, known as gradient methods, would seem to have some potential. Gradient methods are based on the well-known fact that, in the optimization of some function $f(x)$, the direction of maximum increase at a point $x_0$ is given by $\nabla f(x_0)$, the gradient vector. The problem is, when $x$ is constrained to lie within some feasible region, $\Omega$, the direction of the gradient vector may not be feasible. Thus, it is necessary to determine a direction which not only improves the value of $f(x)$, but one which also remains in the feasible region. This is generally
accomplished by projecting, in some sense, the gradient vector upon the boundary of the feasible region. Excellent general treatments of the concepts of gradient methods are given by Hadley (1964), Künzi and Krelle (1966), and Cooper and Steinberg (1970). While these treatments allow a very general objective function and feasible region, we shall need only to deal with the case of a linear objective function subject to linear constraints.

Consider the Chebyshev primal problem given by Problem 2.1.

\[
\begin{align*}
\text{minimize} \quad & (0,0,...,0,1) \begin{bmatrix} \beta \\ \varepsilon \end{bmatrix} \\
\text{subject to} \quad & \begin{bmatrix} -X & 1 \\ X & 1 \end{bmatrix} \begin{bmatrix} \beta \\ \varepsilon \end{bmatrix} \geq \begin{bmatrix} -Y \\ Y \end{bmatrix}
\end{align*}
\]  

(3.14)

Before proceeding, consider a very simple example problem.

Example 3.1. Consider the linear model

\[
Y = \beta_1 X + \varepsilon
\]  

(3.15)

with

\[
Y = \begin{bmatrix} 1 \\ 3 \\ 6 \end{bmatrix}, \quad X = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad n=3, \quad p=1.
\]
Problem 2.1 becomes

\[ \begin{align*} 
\text{minimize} & \quad \varepsilon \\
\text{subject to} & \quad \beta_1 + \varepsilon \geq 1 \\
& \quad 2\beta_1 + \varepsilon \geq 3 \\
& \quad 3\beta_1 + \varepsilon \geq 6 \\
& \quad -\beta_1 + \varepsilon \geq -1 \\
& \quad -2\beta_1 + \varepsilon \geq -3 \\
& \quad -3\beta_1 + \varepsilon \geq -6 
\end{align*} \] (3.16)

Figure 3.7 presents a graphical representation of the feasible region defined as the intersection of the six half-planes identified by the inequality constraints of (3.16).

In Example 3.1, the feasible region is convex and open-ended at the top. This will always be the case. It is clear that any value of \( \beta_1 \) is feasible provided that \( \varepsilon \) is chosen large enough. We now leave the example to consider a more general case.

Let us define, for ease of exposition, the matrix notation

\[ \begin{align*} 
k' &= (0,0,\ldots,0,1), \quad \text{a (p+1) vector,} \\
\theta' &= (\beta_1, \beta_2, \ldots, \beta_p, \varepsilon), \\
c' &= (-y_1, -y_2, \ldots, -y_n, y_1, y_2, \ldots, y_n), \\
A &= \begin{bmatrix} -x & 1 \\ x & 1 \end{bmatrix} 
\end{align*} \]
Figure 3.7. Feasible region for Chebyshev primal problem of Example 3.1
Where \( \mathbf{1} \) denotes a vector each of whose components is one.

The objective is, given any feasible solution \( \mathbf{\theta}_0 \), determine a direction vector, \( \mathbf{\delta}_0 \), and scalar magnitude, \( \lambda > 0 \), such that \( \mathbf{\theta}_1 = \mathbf{\theta}_0 + \lambda \mathbf{\delta}_0 \) satisfies

\[
\mathbf{k}'\mathbf{\theta}_1 \leq \mathbf{k}'\mathbf{\theta}_0 \quad \text{and} \quad \mathbf{A}\mathbf{\theta}_1 \geq \mathbf{c}.
\]  

(3.18)

Naturally, if it is possible, we would like to move in the direction of minus the gradient, or \( \mathbf{\delta}_0 = -\mathbf{k} \), to achieve the maximum decrease in \( \mathbf{k}'\mathbf{\theta} \). But, it is clear from Figure 3.7 that once the boundary of the feasible region is reached, the negative gradient, \( -\mathbf{k}' = (0,0,...,0,-1) \), points straight down or out of the feasible region. Thus it becomes necessary to project the negative gradient upon the boundary of the feasible region to obtain a feasible direction. There are a number of feasible directions which could be taken in the boundary of the feasible region.

We shall consider the orthogonal projection of \( -\mathbf{k} \) upon the boundary of the feasible region as proposed by Rosen (1960).

At a feasible solution, \( \mathbf{\theta}_0 \), on the boundary of the feasible region, the boundary is defined by the equations in the system \( \mathbf{A}\mathbf{\theta}_0 \geq \mathbf{c} \) which are satisfied as equalities or are binding. Let us denote the \( i^{th} \) row of the matrix \( \mathbf{A} \) as \( \mathbf{a}_i \). Then, if the \( i^{th} \) constraint of \( \mathbf{A}\mathbf{\theta}_0 \geq \mathbf{c} \) is binding at \( \mathbf{\theta}_0 \), we may write

\[
\mathbf{a}_i'\mathbf{\theta}_0 = \mathbf{c}_i.
\]  

(3.19)
The orthogonal projection of the negative gradient upon the boundary of the feasible region at \( \theta_0 \) is given by

\[
\delta_0 = (I_{p+1} - M_0 (M_0^t M_0)^{-1} M_0^t) (-k)
\]

where \( I_{p+1} \) denotes a \((p+1) \times (p+1)\) identity matrix, \( M_0 \) denotes a \((p+1) \times r\) matrix whose columns are the rows of \( A \) which correspond to the \( r \) constraints which are binding at the point \( \theta_0 \). That is,

\[
M_0 = (a_{i1} | a_{i2} | \ldots | a_{ir})
\]

where

\[
a_{it}^t \theta_0 = c_{it} \quad t = 1, 2, \ldots, r < p+1.
\]

\((M_0^t M_0)^{-1}\) is defined as any generalized inverse of the \((p+1) \times (p+1)\) matrix \((M_0^t M_0)\).

The expression (3.20) can be easily recognized as the residual vector resulting from the least squares regression of \(-k\) upon the columns of \( M_0 \). This observation provides a convenient procedure for calculating the direction \( \delta_0 \), since numerous subroutines are available to perform least squares fitting. Thus, we have determined the direction we wish to move in. The next issue is the determination of \( \lambda \), or how far we can move in the given direction, \( \delta_0 \).

Once the direction, \( \delta_0 \), is determined, the only factor restricting the choice of \( \lambda \) is that feasibility be
maintained. That is, we wish to choose $\lambda$ as large as possible such that

$$\mathbf{A}(\bar{\theta}_0 + \lambda \bar{\delta}_0) \geq \mathbf{c}. \tag{3.22}$$

Thus, choose $\lambda$ as

$$\lambda = \min \left\{ \frac{c_i - a_i' \bar{\theta}_0}{a_i' \bar{\delta}_0}; \frac{a_i' \delta_0}{a_i' \delta_0} < 0; \frac{a_i' \delta_0}{a_i' \delta_0} > c_i \forall i \right\}; \tag{3.23}$$

that is, as the maximum move, $\lambda$, such that none of the constraints in (3.22) are violated.

Thus we have described an iterative procedure. A single iteration from a feasible solution, $\bar{\theta}_0$, to an improved feasible solution $\bar{\theta}_1$ is defined by $\bar{\theta}_1 = \bar{\theta}_0 + \lambda \bar{\delta}_0$ where $\bar{\delta}_0$ is given by (3.20) and $\lambda$ is given by (3.23). It must be noted that this procedure can be applied so long as the number of binding constraints $r$ is less than $p+1$. When $p+1$ constraints are binding, we are at a vertex of the feasible region which is simply a point in $\mathbb{R}^{p+1}$.

The projection of the gradient upon a vertex will always result in the null vector, $0$. Once we are at a vertex of the feasible region, the rules for obtaining an improved feasible solution must be modified. The modified rules are completely equivalent to the classical exchange algorithm or stage three of Barrodale and Phillips' algorithm.

A possible algorithm for the solution of the Chebyshev
problem would be to operate on the primal problem (3.14) according to Rosen's gradient projection technique until $p+1$ constraints are binding, then use the resulting approximation $\hat{\theta}_0 = (\hat{\beta}_0, \hat{\epsilon}_0)$ as a starting value (as described in the previous section) to initiate Barrodale and Phillips' algorithm. While this idea is certainly workable, a slightly closer examination of how the gradient projection technique operates, will reveal a more efficient procedure for integrating these concepts into a Chebyshev algorithm.

At each stage of gradient projection, we begin with $r$ constraints binding. Since we choose to move in a feasible direction which is along the boundary of the feasible region and move as far as possible in that direction without violating any constraints, at the end of each iteration, the original $r$ constraints remain binding as well as one additional constraint. Hence, what is really occurring at each step of gradient projection is that a new primal constraint is selected to become binding. But this is equivalent to selecting the corresponding dual variable for entry into the basis of the dual tableau (see Figure 3.5).

With respect to Barrodale and Phillips' algorithm for the solution of the Chebyshev dual problem, a stage one column selection rule is obtained by running a primal gradient projection procedure in parallel, simply noting
which new primal constraint becomes binding at each iteration, and entering the corresponding dual variable into the dual basis. The technique appears rather involved, but the computations can be organized so that the algorithm is surprisingly efficient. A computer program has been written to implement this algorithm. The results of a simulation study of the algorithm will be reported later in this section.

We have described an algorithm which, although it is somewhat intricate, assures us that we are moving in the direction of the optimum in the Chebyshev dual problem; in spite of the fact that the objective function can not increase as long as the tableau is degenerate. Henceforth, we shall refer simply to the gradient projection pivot selection technique for the Chebyshev dual problem.

One other minor modification was made to Barrodale and Phillips' algorithm to accommodate gradient projection pivot selection. Barrodale and Phillips (1974 and 1975) employed the standard simplex pivot selection rule, (3.10), in stage three of their algorithm. It was observed that much of the benefit of the gradient projection was lost in stage three under the standard simplex rule. A number of dual variables which would eventually appear in the optimal dual basis were correctly selected by gradient projection, but the correct variables were being replaced
in the basis and would ultimately have to reenter before optimality could be attained. This problem was largely remedied by the incorporation of the more rigorous maximum increase pivot selection rule, (3.9), in stage three.

Lastly, it was found that a carefully chosen starting value, $\theta_0$, was fairly important to the efficient operation of the gradient projection pivot selection rule. The least squares estimator and the maximum least squares residual were found to perform satisfactorily for this purpose.

In summary, we arrive at an algorithm which we shall refer to as the gradient projection algorithm. The algorithm is built around the algorithm of Barrodale and Phillips with a least squares starting value, gradient projection pivot selection in stage one, and maximum increase pivot selection in stage three.

A small simulation study was conducted to assess the performance of the gradient projection algorithm relative to that of Barrodale and Phillips' algorithm. In order to determine if the relative performance of these algorithms varied with problem size parameters, the study was conducted for a number of different values of $n$ (number of observations) and $p$ (number of parameters or explanatory variables). The conduct of the study was as follows.

For each fixed combination of $n$ and $p$, 50 random
problems were generated for solution by each of the algorithms. For each problem, the true solution was fixed as \( \hat{\beta}' = (1,1,\ldots,1) \). The design matrix, \( X \), was also generated at random for each problem such that \( x_{il} = 1 \) for \( i = 1,2,\ldots,n \) and \( x_{ij} \) are independent and identically distributed as Uniform \((-5,5)\) for \( i = 1,2,\ldots,n \) and \( j = 2,3,\ldots,p \). Finally, the dependent variables were generated as \( y_{il} = (X\hat{\beta})_i + e_i \) where the \( e_i \) are independent and identically distributed as Uniform \((-1/2, 1/2)\) for \( i = 1,2,\ldots,n \). Each problem was solved by each of the two algorithms.

Tables 3.4, 3.5, and 3.6 summarize the results of the simulation study for the values \( p = 2, 5, \) and \( 10, \) respectively. Procedure I denotes Barrodale and Phillips' algorithm, while Procedure II denotes the gradient projection algorithm. Quantities exhibited in the tables are the minimum, maximum, mean, and median number of iterations required for the set of problems solved by each procedure for each problem size. Also included are comparison figures which indicate the number of times Procedure II required fewer, the same, or more iterations for solution than Procedure I. The dominance ratio is defined as the fraction of the cases in which the given procedure required as few or fewer iterations for solution. Naturally, the procedure with the higher dominance ratio is to be preferred on the basis of iteration counts.
Table 3.4. Comparison summary for standard (I) versus gradient projection (II) pivot selection techniques (p=2)

<table>
<thead>
<tr>
<th>Observations (n)</th>
<th>Procedure</th>
<th>10</th>
<th>II</th>
<th>50</th>
<th>II</th>
<th>100</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Min Iteration</td>
<td></td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Max Iteration</td>
<td></td>
<td>6</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Mean Iteration</td>
<td></td>
<td>4.48</td>
<td>3.62</td>
<td>5.88</td>
<td>3.82</td>
<td>6.04</td>
<td>3.86</td>
</tr>
<tr>
<td>Median Iteration</td>
<td></td>
<td>4.43</td>
<td>3.72</td>
<td>5.82</td>
<td>3.83</td>
<td>5.96</td>
<td>3.88</td>
</tr>
<tr>
<td>Compared to (I)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fewer Iterations</td>
<td></td>
<td>-</td>
<td>31</td>
<td>-</td>
<td>48</td>
<td>-</td>
<td>45</td>
</tr>
<tr>
<td>Same</td>
<td></td>
<td>-</td>
<td>18</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>More Iterations</td>
<td></td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Dominance Ratio</td>
<td></td>
<td>.38</td>
<td>.98</td>
<td>.04</td>
<td>1.00</td>
<td>.10</td>
<td>.96</td>
</tr>
</tbody>
</table>
Table 3.5. Comparison summary for standard (I) versus gradient projection (II) pivot selection techniques (p=5)

<table>
<thead>
<tr>
<th>Observations (n)</th>
<th>15</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>II</td>
<td>II</td>
</tr>
<tr>
<td>Min Iteration</td>
<td>7</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Max Iteration</td>
<td>14</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>Mean Iteration</td>
<td>10.24</td>
<td>13.08</td>
<td>14.60</td>
</tr>
<tr>
<td></td>
<td>7.00</td>
<td>8.64</td>
<td>10.35</td>
</tr>
<tr>
<td>Median Iteration</td>
<td>10.38</td>
<td>13.00</td>
<td>14.46</td>
</tr>
<tr>
<td></td>
<td>6.97</td>
<td>8.33</td>
<td>10.35</td>
</tr>
<tr>
<td>Compared to (I)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fewer Iterations</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>46</td>
<td>48</td>
<td>42</td>
</tr>
<tr>
<td>Same</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>More Iterations</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Dominance Ratio</td>
<td>.08</td>
<td>.04</td>
<td>.16</td>
</tr>
</tbody>
</table>
Table 3.6. Comparison summary for standard (I) versus gradient projection (II) pivot selection techniques (p=10)

<table>
<thead>
<tr>
<th>Observations (n)</th>
<th>Procedure</th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>Min Iteration</td>
<td>14</td>
<td>12</td>
<td>19</td>
<td>13</td>
</tr>
<tr>
<td>Max Iteration</td>
<td>26</td>
<td>23</td>
<td>36</td>
<td>27</td>
</tr>
<tr>
<td>Mean Iteration</td>
<td>18.72</td>
<td>15.04</td>
<td>25.20</td>
<td>19.90</td>
</tr>
<tr>
<td>Median Iteration</td>
<td>18.15</td>
<td>14.39</td>
<td>24.42</td>
<td>19.31</td>
</tr>
</tbody>
</table>

Compared to (I)
- Fewer Iterations - 44 - 44 - 41
- Same - 1 - 2 - 4
- More Iterations - 5 - 4 - 5
- Dominance Ratio 0.12 0.90 0.12 0.92 0.18 0.90
No comparison of the algorithms was attempted on the basis of total CPU time required. The differences between the programs with respect to amount of computer storage required, form of tableau operated upon, and level of refinement of the respective computer codes would have rendered any such comparison meaningless.

The results of the study show very clearly that iterations are saved by using the gradient projection algorithm, with a 20-30 percent reduction in the number of iterations required for gradient projection uniformly over all problem sizes tested. It should be borne in mind that gradient projection requires more computational effort per iteration than does the algorithm of Barrodale and Phillips. Thus the results are not totally conclusive. Yet it would seem that a worthwhile step has been made in the integration of the best features of two different optimization techniques for the solution of the Chebyshev estimation problem.

The Chebyshev Solution of a Sample Problem

In this section we shall illustrate the algorithms that have been discussed in this chapter. First we shall examine a very small problem with five observations on one independent variable. The sequence of simplex tableaux
required to obtain an optimal solution will be displayed for Barrodale and Phillips' algorithm with both a null starting value and a least squares starting value. Then we shall wish to make a comparison of Barrodale and Phillips' algorithm versus the gradient projection algorithm. It is difficult at best to see the difference in these two algorithms for small problems. So, rather than present the whole sequence of tableaux for some small problem, we have instead chosen to solve a much larger problem and display iteration logs which show how the bases change with each iteration along with the optima.

Consider the data

\[
\begin{align*}
    \mathbf{y} &= \begin{bmatrix} 0 \\ 2 \\ 6 \\ 7 \\ 6 \end{bmatrix}, \\
    \mathbf{x} &= \begin{bmatrix} 3 \\ 2 \\ 0 \\ -2 \\ -3 \end{bmatrix}
\end{align*}
\]

and the model

\[ Y = \beta_0 + \beta_1 X + \varepsilon \]

The least squares estimator of \((\beta_0, \beta_1, \varepsilon)\) is given by

\[ (\hat{\beta}_0, \hat{\beta}_1, \hat{\varepsilon}) = (4.200, -1.077, 1.8), \]

while the Chebyshev estimator of \((\beta_0, \beta_1, \varepsilon)\) is given by
(\beta_0, \beta_1, \epsilon) = (4.500, -1.000, 1.5).

The simplex tableaux for Barrodale and Phillips' algorithm under a null starting value and a least squares starting value are given in Figure 3.8 and Figure 3.9, respectively.

It is interesting to observe the differences in the initial tableau for each procedure. Notice that the starting values for the parameter estimates can be read as the marginal costs corresponding to \( a_1 \) and \( a_2 \). The marginal costs corresponding to the legitimate variables are simply the residuals about the starting values. Barrodale and Phillips' algorithm required only three iterations to attain optimality under the least squares starting value as opposed to four under the null starting value. It is also apparent from Figure 3.9 that the scaling and centering of \( y \) given by the least squares residuals isolates the critical dual variables quite rapidly. In the initial tableau of Figure 3.9, the three most negative marginal costs correspond to the dual variables \( s_3, t_5, \) and \( t_1 \); the dual variables which exist in the optimal basis. As was mentioned previously, it is difficult to observe any difference between Barrodale and Phillips' algorithm and the gradient projection procedure for small problems. In fact, the sequence of simplex tableaux for the sample problem under the gradient projection algorithm is exactly the same as for Barrodale and Phillips'
Figure 3.8. Simplex tableaux for sample problem under Barrodale and Phillips' algorithm initiated with a null starting value
<table>
<thead>
<tr>
<th>0.000</th>
<th>1.000</th>
<th>4.500</th>
<th>0.000</th>
<th>1.000</th>
<th>0.000</th>
<th>0.000</th>
<th>0.000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>1.000</td>
<td>4.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.000</td>
<td>1.000</td>
<td>4.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.000</td>
<td>1.000</td>
<td>4.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
**Figure 3.9. Simplex tableaux for sample problem under Barrodale and Phillips’ algorithm initiated with a least squares starting value**

**ITERATION 0**

<table>
<thead>
<tr>
<th>BASIS</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>S1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>-1.000</td>
<td>-1.000</td>
<td>-1.000</td>
<td>-1.000</td>
<td>-1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>A2</td>
<td>-3.000</td>
<td>-2.000</td>
<td>0.0</td>
<td>2.000</td>
<td>3.000</td>
<td>3.000</td>
</tr>
<tr>
<td>A3</td>
<td>-1.000</td>
<td>-1.000</td>
<td>-1.000</td>
<td>-1.000</td>
<td>-1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>MC</td>
<td>-0.969</td>
<td>-0.046</td>
<td>1.800</td>
<td>0.646</td>
<td>-1.431</td>
<td>0.969</td>
</tr>
</tbody>
</table>

**ITERATION 1**

<table>
<thead>
<tr>
<th>BASIS</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>S1</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>-1.000</td>
<td>-1.000</td>
<td>-1.000</td>
<td>-1.000</td>
<td>-1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>A2</td>
<td>-3.000</td>
<td>-2.000</td>
<td>0.0</td>
<td>2.000</td>
<td>3.000</td>
<td>3.000</td>
</tr>
<tr>
<td>A3</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.000</td>
<td>0.0</td>
</tr>
<tr>
<td>MC</td>
<td>-2.769</td>
<td>-1.846</td>
<td>0.0</td>
<td>-1.154</td>
<td>-3.231</td>
<td>2.769</td>
</tr>
</tbody>
</table>

**ITERATION 2**

<table>
<thead>
<tr>
<th>BASIS</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>S1</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>-2.000</td>
<td>-1.667</td>
<td>-1.000</td>
<td>-0.333</td>
<td>0.0</td>
<td>2.000</td>
</tr>
<tr>
<td>T5</td>
<td>-1.000</td>
<td>-0.667</td>
<td>0.0</td>
<td>0.667</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>A3</td>
<td>4.000</td>
<td>3.333</td>
<td>2.000</td>
<td>0.667</td>
<td>0.0</td>
<td>-2.000</td>
</tr>
<tr>
<td>MC</td>
<td>-6.000</td>
<td>-4.000</td>
<td>0.0</td>
<td>1.000</td>
<td>0.0</td>
<td>6.000</td>
</tr>
</tbody>
</table>

**ITERATION 3**

<table>
<thead>
<tr>
<th>BASIS</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>S1</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.000</td>
<td>0.0</td>
<td>1.000</td>
</tr>
<tr>
<td>T5</td>
<td>0.0</td>
<td>0.167</td>
<td>0.500</td>
<td>0.833</td>
<td>1.000</td>
<td>0.500</td>
</tr>
<tr>
<td>T1</td>
<td>1.000</td>
<td>0.833</td>
<td>0.500</td>
<td>0.167</td>
<td>0.0</td>
<td>-0.500</td>
</tr>
<tr>
<td>MC</td>
<td>0.0</td>
<td>1.000</td>
<td>3.000</td>
<td>2.000</td>
<td>0.0</td>
<td>3.000</td>
</tr>
<tr>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>S5</td>
<td>A1</td>
<td>A2</td>
<td>A3</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>2.000</td>
<td>0.0</td>
<td>-2.000</td>
<td>-3.000</td>
<td>0.0</td>
<td>1.000</td>
<td>0.0</td>
</tr>
<tr>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.0</td>
<td>0.0</td>
<td>1.000</td>
</tr>
<tr>
<td>0.046</td>
<td>-1.800</td>
<td>-0.646</td>
<td>1.431</td>
<td>4.290</td>
<td>-1.077</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>2.000</td>
<td>0.0</td>
<td>-2.000</td>
<td>-3.000</td>
<td>0.0</td>
<td>1.000</td>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-1.000</td>
<td>0.0</td>
<td>1.000</td>
</tr>
<tr>
<td>1.846</td>
<td>0.0</td>
<td>1.154</td>
<td>3.231</td>
<td>6.000</td>
<td>-1.077</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.667</td>
<td>1.000</td>
<td>0.333</td>
<td>0.0</td>
<td>1.000</td>
<td>0.333</td>
<td>0.0</td>
</tr>
<tr>
<td>0.667</td>
<td>0.0</td>
<td>-0.667</td>
<td>-1.000</td>
<td>0.0</td>
<td>0.333</td>
<td>0.0</td>
</tr>
<tr>
<td>-1.333</td>
<td>0.0</td>
<td>1.333</td>
<td>2.000</td>
<td>-1.000</td>
<td>-0.667</td>
<td>1.000</td>
</tr>
<tr>
<td>4.000</td>
<td>0.0</td>
<td>-1.000</td>
<td>0.0</td>
<td>6.000</td>
<td>-0.000</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.500</td>
<td>0.000</td>
<td>0.500</td>
</tr>
<tr>
<td>0.333</td>
<td>0.0</td>
<td>-0.333</td>
<td>-0.500</td>
<td>-0.250</td>
<td>0.167</td>
<td>0.250</td>
</tr>
<tr>
<td>-0.333</td>
<td>0.0</td>
<td>0.333</td>
<td>0.500</td>
<td>-0.250</td>
<td>-0.167</td>
<td>0.250</td>
</tr>
<tr>
<td>2.000</td>
<td>0.0</td>
<td>1.000</td>
<td>3.000</td>
<td>4.500</td>
<td>-1.000</td>
<td>1.500</td>
</tr>
</tbody>
</table>
algorithm with a least squares start, given by Figure 3.9.

In order to demonstrate, in an example, the effectiveness of the gradient projection algorithm we have simply selected one of the problems generated at random for the simulation study of the previous section and displayed iteration logs produced by the two algorithms. The problem selected for the example is one involving 100 observations on each of five explanatory variables. The iteration log produced by each algorithm is simply a list of the dual variables in the basis at each iteration along with the value of the dual objective function. A considerable amount of information is available from these logs. First of all we can monitor the increase of the objective function. Also it is possible to determine when selection errors are made; that is, when a vector which will not be present in the optimal basis is entered into the current basis. Of course, we are particularly interested in the stage one iterations since the modifications proposed for the gradient projection algorithm were designed specifically to improve the generation of an initial feasible extreme point in stage one.

The iteration log for Barrodale and Phillips' algorithm is given by Figure 3.10, while the iteration log for the gradient projection algorithm is presented as Figure 3.11. Under Barrodale and Phillips' algorithm the problem required
<table>
<thead>
<tr>
<th>Iteration</th>
<th>Basis</th>
<th>Optimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$a_1$</td>
<td>$a_2$</td>
</tr>
<tr>
<td>1</td>
<td>$a_1$</td>
<td>$s_{58}$</td>
</tr>
<tr>
<td>2</td>
<td>$a_1$</td>
<td>$s_{58}$</td>
</tr>
<tr>
<td>3</td>
<td>$a_1$</td>
<td>$s_{58}$</td>
</tr>
<tr>
<td>4</td>
<td>$a_1$</td>
<td>$s_{58}$</td>
</tr>
<tr>
<td>5</td>
<td>$s_{9}$</td>
<td>$s_{58}$</td>
</tr>
<tr>
<td>6</td>
<td>$s_{9}$</td>
<td>$t_{58}$</td>
</tr>
<tr>
<td>7</td>
<td>$s_{9}$</td>
<td>$t_{26}$</td>
</tr>
<tr>
<td>8</td>
<td>$s_{9}$</td>
<td>$t_{26}$</td>
</tr>
<tr>
<td>9</td>
<td>$s_{61}$</td>
<td>$t_{26}$</td>
</tr>
<tr>
<td>10</td>
<td>$s_{61}$</td>
<td>$t_{26}$</td>
</tr>
<tr>
<td>11</td>
<td>$s_{61}$</td>
<td>$t_{26}$</td>
</tr>
<tr>
<td>12</td>
<td>$s_{61}$</td>
<td>$t_{26}$</td>
</tr>
<tr>
<td>13</td>
<td>$s_{61}$</td>
<td>$t_{26}$</td>
</tr>
<tr>
<td>14</td>
<td>$s_{61}$</td>
<td>$t_{26}$</td>
</tr>
<tr>
<td>15</td>
<td>$s_{61}$</td>
<td>$t_{26}$</td>
</tr>
<tr>
<td>16</td>
<td>$s_{61}$</td>
<td>$s_{55}$</td>
</tr>
<tr>
<td>17</td>
<td>$s_{61}$</td>
<td>$s_{55}$</td>
</tr>
</tbody>
</table>

Figure 3.10. Iteration log for a random Chebyshev problem solved by Barrodale and Phillips' algorithm
<table>
<thead>
<tr>
<th>Iteration</th>
<th>Basis</th>
<th>Optimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$a_1$</td>
<td>$a_2$</td>
</tr>
<tr>
<td>1</td>
<td>$a_1$</td>
<td>$a_2$</td>
</tr>
<tr>
<td>2</td>
<td>$a_1$</td>
<td>$a_2$</td>
</tr>
<tr>
<td>3</td>
<td>$s_{44}$</td>
<td>$a_2$</td>
</tr>
<tr>
<td>4</td>
<td>$s_{44}$</td>
<td>$t_{26}$</td>
</tr>
<tr>
<td>5</td>
<td>$s_{44}$</td>
<td>$t_{26}$</td>
</tr>
<tr>
<td>6</td>
<td>$s_{44}$</td>
<td>$s_{26}$</td>
</tr>
<tr>
<td>7</td>
<td>$s_{44}$</td>
<td>$t_{60}$</td>
</tr>
<tr>
<td>8</td>
<td>$s_{44}$</td>
<td>$t_{60}$</td>
</tr>
<tr>
<td>9</td>
<td>$s_{44}$</td>
<td>$t_{7}$</td>
</tr>
</tbody>
</table>

Figure 3.11. Iteration log for a random Chebyshev problem solved by the gradient projection algorithm
seventeen iterations to attain optimality as compared to only nine for the gradient projection algorithm. But even more striking is the comparative performances in stage one. In stage one, the gradient projection algorithm correctly selected four of the dual variables which were present in the eventual optimal basis, while the algorithm of Barrodale and Phillips was able to select none.
CHAPTER IV. ASPECTS OF THE STATISTICAL DISTRIBUTION OF CHEBYSHEV ESTIMATORS

In this chapter we shall discuss certain distributional properties of Chebyshev estimators. In the theory of least squares estimation, we are fortunate indeed to have a simple closed-form solution for \( \hat{\beta} \), the estimator of \( \beta \) in Model 1.1. In fact, \( \hat{\beta} = C'y \) is a linear function of \( y \), where \( C = (X'X)^{-1}X' \). Furthermore, any linear function of normally distributed random variables is again normally distributed. Hence, by assuming normally distributed errors, the whole distribution theory of least squares begins to unfold with sums of squares being distributed as Chi-square and ratios thereof following an F-distribution. Thus under the assumption of normality, we can very readily obtain tests of significance and confidence intervals for least squares estimates.

For Chebyshev estimation, in contrast to least squares, there is no closed-form solution for \( \tilde{\beta} \), the Chebyshev estimator of \( \beta \) in Model 1.1, except for the special case of the simple location model, Model 1.2. This means that we must attempt to derive the distribution of the solution to a linear programming problem. No matter what assumptions are made about the distribution of the error vector \( e \), and hence the dependent variable vector \( y \), it is not possible to translate these assumptions simply into a statement
about the distribution of $\xi$, since $\xi$ is not a simple function of $y$.

While not wishing to become deeply involved with the principles of functional analysis, it is worth mentioning the fundamental concepts from which many of the nice properties of least squares derive. Recall the definition of an $L_q$ norm of an $n$-dimensional vector $v$,

$$L_q(v) = \left( \sum_{i=1}^{n} |v_i|^q \right)^{1/q}, \quad 1 \leq q < \infty. \quad (4.1)$$

A normed vector space is simply a vector space having a norm defined upon it. We shall refer to $n$-dimensional vector space with an $L_q$ norm as $L_q$ space. It is a well-known result in functional analysis (see e.g. Royden (1968), p. 121) that, for $q$ strictly between one and infinity, the dual space of $L_q$ space is (isometrically isomorphic to) $L_r$ space where $r = q/(q-1)$. A consequence of this result is that the $L_2$ space is self-dual; that is, $r=q$ for $q=2$. The $L_2$ space of least squares estimation is the only of the $L_q$ spaces which is self-dual.

A fact closely related to this is that the $L_2$ space is the only space of the $L_q$ spaces whose norm can be defined as an inner product. Such a space is referred to as a Hilbert space. A key notion which exists in Hilbert space, but not more generally in normed vector spaces, is that
of orthogonality. Two elements of a vector space are said to be orthogonal if their inner product is zero. Hilbert spaces, equipped with their inner products, possess a wealth of structural properties which generalize many geometrical concepts; for example, the orthogonal decomposition of a vector with respect to the column space of a matrix $X$. The last example is, of course, just a functional analytic description of least squares estimation. It is the rich structure of Hilbert space which leads to a nice distribution theory for least squares that is so sadly lacking for estimation under other $L_q$ criteria.

It is clear that we should not expect to obtain the exact sampling distribution of the Chebyshev estimator. This does not rule out the possibility of developing an asymptotic theory. It may be possible to obtain a limiting distribution for the Chebyshev estimator as the sample size approaches infinity which could serve as an approximation for large, but finite, samples. There are two very natural directions to explore in searching for an asymptotic theory. Huber (1964) in a paper on robust estimation of location defined what he called M-estimators of location and proved their asymptotic normality under fairly general regularity conditions. He extended his results to the multiparameter case, which includes multiple regression, in Huber (1967). Huber's
results are applicable to $L_q$ estimation and could possibly lead to the asymptotic distribution in the limiting case of Chebyshev estimation. Koenker and Bassett (1978) provide a proof of the asymptotic normality of the $L_1$, or least sum of absolute error estimator. Their technique of proof seems to have great potential as a direction for obtaining the asymptotic distribution of the Chebyshev estimator.

A Survey of the Theory of M-Estimation

Given a set of $n$ observations $y_1, y_2, \ldots, y_n$, consider an estimator $\hat{T}_n$ of a parameter $\theta$ defined by

$$\min_{t} \sum_{i=1}^{n} \rho(y_i; t)$$

or by an implicit equation of the form

$$\sum_{i=1}^{n} \psi(y_i; T_n) = 0$$

where $\rho$ is an arbitrary function and

$$\psi(y_i; \theta) = \frac{\partial}{\partial \theta} \rho(y_i; \theta) .$$

The estimator $\hat{T}_n$, defined by (4.2) or (4.3) is called an M-estimator or maximum-likelihood type estimator. The class of M-estimators is easily seen to contain all ordinary maximum-likelihood estimators simply by taking $y_1, y_2, \ldots, y_n$ to be a
random sample from a common distribution with density \( f(y;\theta) \) and \( \rho(y;\theta) = -\log f(y;\theta) \). Furthermore, all \( L_q \) estimators of location (\( \alpha \) in Model 1.2) are expressible as M-estimators by taking \( \rho(y;\tilde{\alpha}) = |y-\tilde{\alpha}|^q \). Then (4.2) becomes

\[
\min_{\tilde{\alpha}} \sum_{i=1}^{n} |y_i - \tilde{\alpha}|^q
\]

which is equivalent to minimizing the \( L_q \) norm of the residual vector.

Huber (1964) showed, under fairly general conditions involving the continuity and convexity of \( \rho \) and \( \psi \), that M-estimators of location are asymptotically normal. The regularity conditions for limiting normality are satisfied for all \( L_q \) estimators of location where \( q \) is strictly between one and infinity. In Huber (1967), the normality results were extended to include the case of multiple regression. Thus, in the search for the asymptotic distribution of \( L_q \) estimators, only the boundary cases, \( L_1 \) estimation and Chebyshev estimation remain. Koenker and Bassett (1978), prove that the \( L_1 \) estimator has a limiting normal distribution, by methods completely apart from those of Huber. While one would hope, for the sake of simplicity, that Chebyshev estimators are asymptotically normal, it appears that such a result is not forthcoming. One can see why asymptotic normality will not hold in general simply by examining the Chebyshev estimator.
of location, the midrange.

The Limiting Distribution of the Sample Midrange

Consider the location model, Model 1.2

\[ y = \alpha + \varepsilon. \] (4.6)

The Chebyshev estimator of \( \alpha \) is given by the sample midrange

\[ \tilde{\alpha} = \frac{y_{(1)} + y_{(n)}}{2}. \] (4.7)

Since there is a closed-form solution for \( \tilde{\alpha} \), we can examine the distribution of it in a fairly straightforward way, applying some of the asymptotic theory of order statistics.

Let us assume that \( \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \) are independent identically distributed with common probability density function \( f \) which is symmetric about zero. Let \( \varepsilon_{(1)} \leq \varepsilon_{(2)} \leq \ldots \leq \varepsilon_{(n)} \) denote the ordered values of \( \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \). Then

\[ d \varepsilon_{(1)} = -\varepsilon_{(n)}. \] (4.8)

That is, for a symmetric parent distribution, the distribution of the sample minimum is the mirror image of the distribution of the sample maximum. It is also well-known that as the sample size \( n \) tends toward infinity that the sample minimum and sample maximum are asymptotically independently distributed (see e.g. David (1970), p. 211). Suppose the limiting distribution of the sample maximum from a parent distribution,
with probability density function, \( f \), exists and let us denote the cumulative distribution function of the limiting distribution by \( A \). Then, in view of the symmetry result, (4.8), and the asymptotic independence of the extremes, the sample midrange is asymptotically distributed like \( \bar{D}/2 \) where \( D \) is distributed as the difference of two independent identically distributed random variables with cumulative distribution function \( A \). That is, \( D \) has a cumulative distribution function which is expressible as the convolution of \( A \) with the cumulative distribution function corresponding to the limiting distribution of \(-e_{(n)}\).

All that is really needed to complete the description of the asymptotic distribution of \( \bar{a} \), the sample midrange, is a discussion of the limiting distribution of the sample maximum, \( e_{(n)} \). The distribution of the sample maximum in finite samples of size \( n \) is given by

\[
A^{(n)}(e) = [F(e)]^n. \quad (4.9)
\]

While a limiting distribution for (4.9) is not guaranteed to exist for all parent distributions, \( F \), if a limiting distribution does exist it is necessarily one of just three types. That is, for suitably chosen constants \( a > 0 \) and \( b_n \), \([F(a_n e + b_n)]^n \) converges to one of the following types of limiting distribution:
For details of this result and further references see David (1970).

It is clear that no convolution of limiting distributions of the types given in (4.10), (4.11), (4.12) results in a normal distribution. The fact that $L_q$ estimators of location are asymptotically normal for $1 < q < \infty$ while the Chebyshev estimator is not is very much analogous to a result in order statistics that certain well-behaved functions of the central order statistics, or quantiles, possess an asymptotic normal distribution, while the corresponding functions of the extremes do not.

Admittedly, the foregoing discussion has been rather general and lacking in specifics. The only point to be made is that the limiting distribution of Chebyshev estimators is not normal and that this is due to the fact that the Chebyshev estimator is determined wholly by the extremal observations. Further details on the distribution of extreme values and
functions thereof can be found in Gumbel (1944), Gumbel (1958), Gumbel et al. (1965), and David (1970). In order to fix some of the ideas presented in this section we shall consider the following concrete example.

**Example 4.1.** Consider the location model, Model 1.2. Let $e_1, e_2, \ldots, e_n$ be independent and identically distributed as Uniform $[-\varepsilon, \varepsilon]$. Then $y_1, y_2, \ldots, y_n$ are independent and identically distributed as Uniform $[\alpha-\varepsilon, \alpha+\varepsilon]$. The exact density of the sample midrange of the $y_1, y_2, \ldots, y_n$ is given by

$$f_n(a) = \frac{n(|e| a)^{n-1}}{2\varepsilon^n} \quad \alpha-\varepsilon \leq a \leq \alpha+\varepsilon$$

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

The first two moments of the sample midrange, $\tilde{a}$, are given by

$$E(\tilde{a}) = \alpha \quad (4.13)$$

$$\text{Var}(\tilde{a}) = \frac{2\varepsilon^2}{(n+1)(n+2)} \quad (4.14)$$

A small simulation study was done to confirm that the distribution of the sample midrange in Example 4.1 does not converge toward normality. For each of five sample
sizes $n = 10, 25, 50, 100, 500$; 100 samples of size $n$ were drawn from a Uniform $[-1/2, 1/2]$ population and standardized midranges were calculated as

$$T_0 = \frac{(y(1) + y(n)) \cdot \sqrt{(n+1)(n+2)/2}}{\sqrt{n+1}(n+2)/2}.$$  

(4.15)

The expression (4.15) results from the standardization of the sample midrange by subtracting its expected value (4.13), and dividing by the square root of the variance (4.14). For each sample size, $n$, a chi-square goodness-of-fit test was done against the standard normal distribution, using eleven equiprobable classes. The results of the study are summarized in Table 4.1.

Table 4.1 presents, for each sample size, the resulting Chi-square statistic from each goodness-of-fit test as well as the corresponding critical level. The critical level is simply the probability of observing a Chi-square value as large as that observed, under the null hypothesis that the midrange is normally distributed. We observe that the critical level is a decreasing function of $n$. That is, the study gives clear evidence that the distribution of the sample midrange in Example 4.1 converges away from normality.
Table 4.1. Summary of goodness-of-fit tests for normality of the sample midrange from samples of size n

<table>
<thead>
<tr>
<th>Sample size (n)</th>
<th>Chi-square</th>
<th>Critical level</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>13.52</td>
<td>0.1960</td>
</tr>
<tr>
<td>25</td>
<td>15.72</td>
<td>0.1079</td>
</tr>
<tr>
<td>50</td>
<td>19.24</td>
<td>0.0373</td>
</tr>
<tr>
<td>100</td>
<td>22.10</td>
<td>0.0146</td>
</tr>
<tr>
<td>500</td>
<td>29.14</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

The Results of Koenker and Bassett for $L_1$ Estimation

Rosenberg and Carlson (ca. 1972) studied, via Monte Carlo methods, a normal approximation for the limiting distribution of the $L_1$ estimator. Their sampling results agreed very closely with the approximation and led to the conjecture that the $L_1$ estimator possesses an asymptotically normal distribution. Koenker and Bassett (1978) proved this conjecture. Their techniques of proof seem to be very much applicable to the similar problem of deriving the asymptotic distribution of the Chebyshev estimator. The problem for Chebyshev estimation is somewhat more complicated since we know we are looking for some nonnormal limiting distribution, possibly some multivariate generalization of an extreme value.
distribution. We shall describe the proof of Koenker and Bassett (1978), commenting as we go on the analogous results for Chebyshev estimation.

The proof centers around a well-known characterization of the L₁ estimator. Koenker and Bassett first define

$$\eta_1 = \{1, 2, \ldots, n\}$$

and let $\mathcal{H}_1$ denote the set of p-element subsets of $\eta_1$. Then elements $h_1 \in \mathcal{H}_1$ have relative complement $\overline{h}_1 = \eta_1 - h_1$, and serve to partition the rows of $y$ and $X$. That is, $y(h_1)$ denotes the p-vector whose elements are given by $\{y_i : i \in h_1\}$ and $X(h_1)$ denotes an $(n-p) \times p$ matrix whose rows are given by $\{x_i : i \in \overline{h}_1\}$ where $x_i$ denotes the i-th row of the design matrix $X$. $1_p$ denotes a p-vector each of whose elements is one. Finally, they define

$$H_1 = \{h_1 \in \mathcal{H}_1 | \text{rank } X(h_1) = p\}.$$ 

Then some fundamental properties of $\hat{\beta}^*$, the elements of the L₁ solution set, $B^*$, are derived.

**Lemma 4.1.** If $X$ is of full rank $p$, then the solution set $B^*$ to the L₁ problem has at least one element of the form,

$$\hat{\beta}^* = X(h_1)^{-1}y(h_1)$$

for some $h_1 \in H_1$. 

Lemma 4.1 simply states the well-known property of $L_1$ estimation that there exists an optimal $L_1$ hyperplane which interpolates (passes through) at least $p$ observation points (see Appa and Smith (1973)). Thus the $L_1$ solution can be characterized by the $p$ observation points which it interpolates.

The corresponding characterization for the Chebyshev estimator is given by Theorem 2.1 and Theorem 2.7; that is, there exists an optimal Chebyshev hyperplane which is equidistant from $p+1$ observation points. This characterization can be expressed in a form analogous to Lemma 4.1.

Define

$$A = \begin{bmatrix} -X & \frac{1}{n} \\ X & \frac{1}{n} \end{bmatrix}, \quad c = \begin{bmatrix} y \\ -y \end{bmatrix}, \quad \theta = \begin{bmatrix} \beta \\ \epsilon \end{bmatrix}. \quad (4.16)$$

Also define

$$\eta_2 = \{1, 2, \ldots, 2n\}$$

and let $\mathcal{H}_2$ denote the set of all $(p+1)$-element subsets of $\eta_2$. Then elements $h_2 \in \mathcal{H}_2$ have relative complement $\overline{h}_2 = \eta_2 - h_2$, and serve to partition the rows of $c$ and $A$. Finally, define

$$H_2 = \{h_2 \in \mathcal{H}_2 \mid \text{rank } A(h_2) = p+1 \text{ and } i \in h \not\in (i+n) \forall h \}. \quad (4.17)$$

Let $\bar{\eta}$ denote the elements of the Chebyshev solution set, $\bar{\eta}$. Then the analogue of Lemma 4.1 for the Chebyshev problem is
given by Lemma 4.2.

**Lemma 4.2.** If \( A \) (in (4.16)) has rank \( p+1 \) then the solution set \( \bar{\theta} \) of the Chebyshev problem has at least one element of the form

\[
\bar{\theta} = A^{-1}(h_2)\zeta(h_2)
\]

for some \( h_2 \in H_2 \).

Koenker and Bassett (1978) next give a necessary and sufficient condition for, given \( h \in H \), \( \bar{\beta}^* = X(h_1)^{-1}y(h_1) \) to be a unique optimal \( L_1 \) estimator.

**Lemma 4.3.** If \( F \), the cumulative distribution function of the errors, is continuous then, with probability one, \( \bar{\beta}^* = X(h_1)^{-1}y(h_1) \) is the unique optimal \( L_1 \) estimator if and only if

\[
-l'_p < \frac{1}{p} \sum_{i \in H_1} \text{sgn}(y_i - x_i^\prime \bar{\beta}^*) x_i^\prime X(h_1)^{-1} < l'_p
\]

where the inequalities are interpreted componentwise.

A corresponding condition for the Chebyshev problem is given by Lemma 4.4.

**Lemma 4.4.** If \( F \), the cumulative distribution function of the errors, is continuous then, with probability one, \( \bar{\theta} = A(h_2)\zeta(h_2) \) is the unique optimal Chebyshev estimator if
and only if
\[ k' A(h_2)^{-1} > 0_{p+1} \] (4.17)
\[ c(h_2) - A(h_2) A(h_2)^{-1} c(h_2) > 0_{2n-p-1} \]
where \( k' = (0, 0, \ldots, 0, 1) \) and \( 0_{p+1} = (0, 0, \ldots, 0) \) are \((p+1)\)-vectors and the system of inequalities is interpreted componentwise.

The conditions given by (4.17) simply say that each component of the optimal dual solution vector (see Problem 2.2) is strictly positive, while the marginal costs corresponding to nonbasic vectors are nonnegative.

Koenker and Bassett (1978) then consider the probability element
\[ g^*(\delta) d\delta_1, \ldots, d\delta_p = \Pr\{\delta < \sqrt{n} (\beta_n^*(\gamma, X) - \bar{\beta}) < \delta + d\delta\} \] (4.18)
where \( d\delta' = (d\delta_1, \ldots, d\delta_p) \) and \( g^*(\delta) \) is the density of \( \sqrt{n} (\beta_n^* - \bar{\beta}) \). They then show that \( g^*_n(\delta) \) converges to a specified Gaussian density.

The proof of the convergence of (4.18) is critically dependent upon the characterization and uniqueness conditions of Lemma 4.1 and Lemma 4.3 which allow (4.18) to be reexpressed as
\[ g^*_n(\delta) d\delta_1, \ldots, d\delta_p = \sum_{h \in H_1} \Pr\{\delta < \sqrt{n} X(h_1)^{-1} e(h_1) < \delta + d\delta\} \]
\[ \cdot \Pr\{-l_{p-1} < \sum_{i \in H_1} \text{sgn}[e(h_1)] \cdot x_i' \hat{\delta}/\sqrt{n} x_i' X(h_1)^{-1} l_{p-1}\} \]
where \( e = y - X\hat{\theta} \). The remainder of the proof is not of interest to us since it does not apply to the Chebyshev problem.

Once again we give the analogue of (4.18) for the Chebyshev estimator.

\[
\tilde{g}_n(\delta) d\delta_1, \ldots, d\delta_{p+1} = \Pr\{ \delta < n(\hat{\theta}_n(y, X) - \theta) < \delta + d\delta \} \tag{4.19}
\]

Notice that in (4.18) a normalizing constant of \( \sqrt{n} \) is used, while we conjecture that \( n \) is the appropriate normalizing constant for the Chebyshev estimator. We can reexpress (4.19) as

\[
\tilde{g}_n(\delta) d\delta_1, \ldots, d\delta_{p+1} = \sum_{h_2 \in H_2} \Pr\{ \delta < nA(h_2)^{-1}e(h_2) < \delta + d\delta \} \cdot \Pr\{ k'A(h_2)^{-1} > 0 \text{ p+1, } c(h_2) - A(h_2)A(h_2)^{-1}c(h_2) > 0 \}^{2n-p-1} \tag{4.20}
\]

The results of Koenker and Bassett (1978), the analogous results for Chebyshev estimation, and (4.20) appear to provide a good starting point for continuing research on the limiting distribution of the Chebyshev estimator.
CHAPTER V. APPLICATIONS OF CHEBYSHEV ESTIMATION

In Chapter I we discussed possible applications of Chebyshev estimation. We shall expand on that in this chapter. Of course, with respect to Model 1.1, if the error vector $e$ follows a uniform distribution, the Chebyshev estimator is natural. But, as will be seen, the Chebyshev estimator is preferable to least squares for a number of low-kurtosis error distributions. We shall also demonstrate how Chebyshev estimation might be used in data analytic situations.

The Chebyshev Criterion in Approximating Functions

Consider some $y = \phi(t_1, t_2, \ldots, t_m)$ which is either tabled or some function which can be given by an explicit formula, but for practical or theoretical reasons is difficult to compute. We might wish to approximate $y$ as a function of several dependent variables $x_1, x_2, \ldots, x_p$ where some of the $x_1, x_2, \ldots, x_p$ may just be some of the $t_1, t_2, \ldots, t_m$. It is easy to visualize a linear approximation of $y$ as in Model 1.1, but we may wish to consider a more general approximating function or we may wish to approximate $y$ continuously, as opposed to the discrete linear approximation given by Model 1.1. In any of these cases, the Chebyshev criterion is a
favorite for the fitting of approximating functions, for reasons that have nothing to do with the distribution of errors.

Certainly, if the tabled values of $y$ contain small roundoff errors one would consider Chebyshev approximation, but in general there is a much more significant source of error which we will call modeling error. Modeling error is nonrandom and reflects the inability of the approximating function to accurately represent the true function $y$. We emphasize that the character of the errors is not random, simply unknown. Under such circumstances a natural most conservative criterion of fit is given by the Chebyshev criterion. Minimizing the maximum error insures that the approximation obtained is uniformly good over the entire range of independent variables $x_1, x_2, \ldots, x_p$ considered. Other criteria of fit necessarily yield approximations which are somewhat better than the Chebyshev estimator in some regions at the cost of being less accurate in others.

Philosophical issues aside, the Chebyshev criterion is the accepted criterion in the approximation of functions. Statisticians have long been dependent upon tables of statistical functions. With the increasing use of computers to perform routine statistical analyses, approximation of statistical functions is an attractive alternative to storage
of tables or more cumbersome direct calculations. Thus, the Chebyshev criterion ought to receive increasing use by statisticians in the solution of what are basically non-statistical, but still very important, problems.

Coming back to modeling error, one may raise a largely philosophical question. The optimality properties of various methods of estimation are derived under the tacit assumption that the model is known, when, in fact the model is essentially never known. Thus, besides sampling error which is generally random with some distribution there is, confounded with sampling error, some degree of modeling error which is not random. For purposes of pure fitting no assumptions about the errors are needed, but for purposes of interpreting an analysis the assumption of randomness is crucial. It is suggested that if sampling error is small relative to modeling error, the Chebyshev criterion is perhaps more appropriate for fitting, since we are then essentially back to the problem of approximation as opposed to estimation. By using the Chebyshev criterion in this situation it may appear that the fit obtained is not as good, but we are not deluded into believing that predictions made from the model will be better than they really are.
The Efficiency of Chebyshev Estimation

Just as one rarely observes a perfect normal distribution (some would say never), it is unlikely that one would be confronted with a perfectly uniform distribution of errors. Thus, it is natural to ask for what types of error distributions is Chebyshev estimation more efficient than least squares estimation, in the sense that the mean square error of the estimates of $\beta_1, \beta_2, \ldots, \beta_p$ are smaller. We shall consider the results of a small simulation study to determine the efficiency of Chebyshev estimation relative to least squares for several different error distributions. We shall also consider the results of Rider (1957) who studied the sample midrange as an estimator of location for symmetric distributions.

In order to compare the efficiency of Chebyshev estimation relative to least squares estimation for estimating $\beta$ in Model 1.1, a simulation study was conducted. The factors that were controlled in the experiment were the distribution of the errors $e_1, e_2, \ldots, e_n$ and the problem size parameters; $n$, the number of observations and $p$, the number of explanatory variables. The levels of $n$ considered were \{25, 50, 100\}, while the levels of $p$ considered were \{2, 4, 6\}. Four different error distributions were considered:
Uniform

\[ f_1(t) = \frac{1}{2\sqrt{3}} \text{ for } -\sqrt{3} < t < \sqrt{3} \]  
\[ = 0 \text{ otherwise} \] \hspace{1cm} (5.1)

Truncated Normal \( f_2(t) = \frac{1}{\phi(1)-\phi(-1)} \left( \frac{1}{2\sqrt{2\pi}} \right) \left( \frac{1}{\phi(1)-\phi(-1)} \right) \text{ for } -2 < t < 2 \]  
\[ = 0 \text{ otherwise} \] \hspace{1cm} (5.2)

Triangle

\[ f_3(t) = \frac{\sqrt{6} - |t|}{6} \text{ for } -\sqrt{6} < t < \sqrt{6} \]  
\[ = 0 \text{ otherwise} \] \hspace{1cm} (5.3)

Normal

\[ f_4(t) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{t^2}{2} \right\} \text{ for } -\infty < t < \infty . \]  
\hspace{1cm} (5.4)

Fifty problems were generated, at random, for each of the 36 combinations of \( n, p, \) and error distribution. For each fixed combination of \( n, p, \) and \( f \) let \( k = 1, 2, \ldots, 50 \) index the 50 problem replicates. Denote the \( j^{th} \) \((j=1,2,\ldots,p)\) parameter in the \( k^{th} \) \((k=1,2,\ldots,50)\) problem as \( \beta_{jk} \).
Similarly, denote the corresponding Chebyshev and least squares estimators as $\tilde{\beta}_{jk}$ and $\hat{\beta}_{jk}$, respectively. Then for each fixed combination of $n$, $p$, and $f$ we can calculate

$$\left(\tilde{\beta}_{jk} - \beta_{jk}\right)^2 \quad j = 1, 2, \ldots, p; \quad k = 1, 2, \ldots, 50$$  \hspace{1cm} (5.5)

and the mean square errors

$$\tilde{M}_j = (50)^{-1} \sum_{k=1}^{50} \left(\tilde{\beta}_{jk} - \beta_{jk}\right)^2 \quad j = 1, 2, \ldots, p.$$  \hspace{1cm} (5.6)

$$\hat{M}_j = (50)^{-1} \sum_{k=1}^{50} \left(\hat{\beta}_{jk} - \beta_{jk}\right)^2$$

From each of the mean square errors $\tilde{M}_j$, $\hat{M}_j$, we can calculate a relative efficiency of Chebyshev estimation relative to least squares for each coefficient $\beta_1, \beta_2, \ldots, \beta_p$ as

$$R_j = \frac{\hat{M}_j}{\tilde{M}_j}, \quad j = 1, 2, \ldots, p.$$  \hspace{1cm} (5.7)

And finally, in order to get a single number for each fixed $n$, $p$, and $f$, we average the $p$ relative efficiencies as

$$A = p^{-1} \sum_{j=1}^{p} R_j.$$  \hspace{1cm} (5.8)

We shall refer to this quantity as the average relative efficiency of Chebyshev estimation relative to least squares.
for the given combination of n, p, and f or simply, the average relative efficiency of Chebyshev estimation. The average relative efficiency of Chebyshev estimation is tabled in Table 5.1 for each of the 36 combinations of n, p, and error distribution.

Before discussing the results of the study, we shall clear up a few details of its conduct heretofore omitted. For each problem in the study, an error vector e was generated according to the appropriate density from (5.1)-(5.4) and a random design matrix X = \{x_{ij}\} was constructed such that x_{il} = 1 for i = 1, 2, ..., n and the remaining x_{ij} were independent identically distributed as Uniform (-5,5). The problem

\[ e = Xa + \delta \]

was then solved both by the Chebyshev criterion and by least squares to obtain the solutions \( \hat{a} \) and \( \hat{\beta} \), respectively. It is a straightforward consequence of Theorem 3.1 that

\[ (\hat{a} - \bar{a}) = \bar{a} \]

for all \( \delta \in \mathbb{E} \)

\[ (\hat{\beta} - \bar{\beta}) = \bar{\beta} \]

where \( \beta \), \( \bar{\beta} \), \( \hat{\beta} \) denote the true parameter value, the Chebyshev estimator, and the least squares estimator, respectively, for Model 1.1. That is, the errors in the estimators \( \hat{a} \), \( \hat{\beta} \) do
not depend on the true value of $\beta$. This fact was used to reduce computations considerably.

Table 5.1 displays the average relative efficiencies of Chebyshev estimation relative to least squares for the problem sizes indicated and for each of the four error distributions, (5.1)-(5.4) where these average relative efficiencies are computed according to (5.5)-(5.8). Entries greater than one indicate the superiority of Chebyshev estimation, while values less than one indicate the superiority of least squares. The results indicate that Chebyshev estimation is preferred for uniform and truncated normal error distributions. For these two error distributions, the relative efficiency is an increasing function of $n$ for fixed $p$ and a decreasing function of $p$ for each $n$. The least squares estimator is preferred for triangle and normal distributions of error. For normally distributed errors, the relative efficiency is a decreasing function of $n$ for fixed $p$ and an increasing function of $p$ for fixed $n$. We may note that trends with respect to $n$ and $p$ which are exhibited for the other three error distributions are not so clear for the triangle distribution. This, we conjecture, is in analogy to the location case where the table of Rice and White (1964) (see Table 1.1) indicates that the asymptotic relative efficiencies ought to be
Table 5.1. Average relative efficiencies of Chebyshev estimation relative to least squares for several different problem sizes and for each of four error distributions

<table>
<thead>
<tr>
<th>Observations (n)</th>
<th>Number of Explanatory Variables</th>
<th>2</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>UN\textsuperscript{a}</td>
<td>2.9257</td>
<td>1.0814</td>
<td>0.9430</td>
</tr>
<tr>
<td></td>
<td>TN\textsuperscript{b}</td>
<td>1.3468</td>
<td>0.9003</td>
<td>0.8189</td>
</tr>
<tr>
<td></td>
<td>TR\textsuperscript{c}</td>
<td>0.5574</td>
<td>0.6074</td>
<td>0.6310</td>
</tr>
<tr>
<td></td>
<td>NR\textsuperscript{d}</td>
<td>0.2578</td>
<td>0.4674</td>
<td>0.4772</td>
</tr>
<tr>
<td>50</td>
<td>UN</td>
<td>4.4594</td>
<td>2.0901</td>
<td>1.6965</td>
</tr>
<tr>
<td></td>
<td>TN</td>
<td>1.8417</td>
<td>1.3709</td>
<td>1.0853</td>
</tr>
<tr>
<td></td>
<td>TR</td>
<td>0.4491</td>
<td>0.4411</td>
<td>0.5475</td>
</tr>
<tr>
<td></td>
<td>NR</td>
<td>0.2296</td>
<td>0.2946</td>
<td>0.3152</td>
</tr>
<tr>
<td>100</td>
<td>UN</td>
<td>6.6898</td>
<td>2.7497</td>
<td>2.9570</td>
</tr>
<tr>
<td></td>
<td>TN</td>
<td>4.2354</td>
<td>1.6801</td>
<td>1.5957</td>
</tr>
<tr>
<td></td>
<td>TR</td>
<td>0.7998</td>
<td>0.5638</td>
<td>0.5387</td>
</tr>
<tr>
<td></td>
<td>NR</td>
<td>0.0885</td>
<td>0.1636</td>
<td>0.2061</td>
</tr>
</tbody>
</table>

\textsuperscript{a} UN specifies Uniform error distribution (see (5.1)).

\textsuperscript{b} TN specifies Truncated Normal error distribution (see (5.2)).

\textsuperscript{c} TR specifies Triangle error distribution (see (5.3)).

\textsuperscript{d} NR specifies Normal error distribution (see (5.4)).
infinity for the uniform and truncated normal distributions, zero for the normal distribution and some constant between zero and one for the triangle distribution.

The fact that Chebyshev estimation is preferred for the truncated normal distributions confirms our claims about the applicability of Chebyshev estimation for truncated error distributions. Of course, the goodness of the Chebyshev estimator necessarily depends upon the flatness of the peak of the distribution as well as the degree of truncation.

We shall continue our discussion of the efficiency of Chebyshev estimation relative to least squares by considering a study done by Rider (1957). Rider (1957) compares the sample midrange to the sample mean as an estimator of location for five different symmetric, low-kurtosis error distributions. It was found that the efficiency of the midrange relative to the mean was an increasing function of the kurtosis (standardized fourth central moment) of the error distribution. The midrange was found to be more efficient than the mean for each of the error distributions studied. The error distributions considered by Rider were

\[ f(t) = \frac{\cos(t)}{2}, \quad -\frac{\pi}{2} \leq t \leq \frac{\pi}{2} \]  

(5.9)

kurtosis = 2.19
Parabolic \( f(t) = \frac{3(1-t^2)}{4}, \quad -1 \leq t \leq 1 \) \hspace{1cm} (5.10)

kurtosis = 2.14

Uniform \( f(t) = \frac{1}{2}, \quad -1 \leq t \leq 1 \) \hspace{1cm} (5.11)

kurtosis = 1.8

Inverted Parabolic \( f(t) = \frac{2}{t^2}, \quad -1 \leq t \leq 1 \) \hspace{1cm} (5.12)

kurtosis = 1.19

Dichotomous \( f(t) = \frac{1}{2}, \quad t = \pm 1 \) \hspace{1cm} (5.13)

kurtosis = 1.00.

Table 5.2 presents the relative efficiency of the midrange to the mean as an estimator of location for several small sample sizes for each of the five error distributions. These relative efficiencies were derived analytically and not as the result of a simulation experiment. Clear trends are present in the table with relative efficiency being a decreasing function of kurtosis and an increasing function of \( n \) for the error distributions studied.

The results of Rider (1957) suggest that the sample kurtosis be used as a guideline in deciding whether or not the use of Chebyshev is indicated for a given sample.
Table 5.2. Relative efficiency of the sample midrange relative to the sample mean as an estimator of location for each of five error distributions

<table>
<thead>
<tr>
<th>Sample Size (n)</th>
<th>Population</th>
<th>Cosine</th>
<th>Parabolic</th>
<th>Uniform</th>
<th>Inverted Parabolic</th>
<th>Dichotomous</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$K=2.19$</td>
<td>$K=2.14$</td>
<td>$K=1.8$</td>
<td>$K=1.19$</td>
<td>$K=1$</td>
</tr>
<tr>
<td>2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>1.016</td>
<td>1.027</td>
<td>1.111</td>
<td>1.316</td>
<td>1.333</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.030</td>
<td>1.052</td>
<td>1.250</td>
<td>1.898</td>
<td>2.000</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.038</td>
<td>1.073</td>
<td>1.401</td>
<td>2.825</td>
<td>3.200</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.046</td>
<td>-</td>
<td>1.555</td>
<td>-</td>
<td>5.333</td>
<td></td>
</tr>
</tbody>
</table>

$K$ denotes kurtosis.
A sample estimate of kurtosis is given by

\[ g = \frac{n \sum_{i=1}^{n} (y_i - \bar{y})^4}{n \sum_{i=1}^{n} (y_i - \bar{y})^2}^{2} \] (5.14)

We have seen from Table 1.1 and Table 5.1 that least squares is more efficient than Chebyshev estimation for a triangle distribution (5.3) of errors which has kurtosis equal to 2.4. On the other hand, Table 5.2 demonstrates that Chebyshev estimation is more efficient for the cosine distribution (5.9) of errors which has kurtosis equal to 2.19. Thus it would seem a reasonable rule to employ Chebyshev estimation whenever the sample kurtosis is less than some preassigned value, say 2.2.

Application of Chebyshev Estimation in Data Analysis

In this section we shall examine the mechanics of Chebyshev estimation in data analytic situations for a hypothetical data set. Consider the data set given by Table 5.3 and the model

\[ y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon \] (5.15)

The data in Table 5.3 are simply two "observations" at each of 51 equally-spaced \( x_1 \) values from -12.5 to +12.5
Table 5.3. A hypothetical data set with a uniform error distribution

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>x₀</th>
<th>x₁</th>
<th>x₁²</th>
</tr>
</thead>
<tbody>
<tr>
<td>-21.549</td>
<td>1</td>
<td>-5.5</td>
<td>30.25</td>
<td></td>
</tr>
<tr>
<td>9.983</td>
<td>1</td>
<td>4.0</td>
<td>16.00</td>
<td></td>
</tr>
<tr>
<td>12.154</td>
<td>1</td>
<td>9.0</td>
<td>81.00</td>
<td></td>
</tr>
<tr>
<td>-49.050</td>
<td>1</td>
<td>-10.0</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td>11.797</td>
<td>1</td>
<td>9.5</td>
<td>90.25</td>
<td></td>
</tr>
<tr>
<td>9.298</td>
<td>1</td>
<td>3.5</td>
<td>12.25</td>
<td></td>
</tr>
<tr>
<td>-2.047</td>
<td>1</td>
<td>-1.0</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>12.309</td>
<td>1</td>
<td>8.5</td>
<td>72.25</td>
<td></td>
</tr>
<tr>
<td>-9.647</td>
<td>1</td>
<td>-3.0</td>
<td>9.00</td>
<td></td>
</tr>
<tr>
<td>-27.077</td>
<td>1</td>
<td>-6.5</td>
<td>42.25</td>
<td></td>
</tr>
<tr>
<td>-67.861</td>
<td>1</td>
<td>-12.5</td>
<td>156.25</td>
<td></td>
</tr>
<tr>
<td>12.502</td>
<td>1</td>
<td>6.5</td>
<td>42.25</td>
<td></td>
</tr>
<tr>
<td>12.743</td>
<td>1</td>
<td>7.5</td>
<td>56.25</td>
<td></td>
</tr>
<tr>
<td>-52.355</td>
<td>1</td>
<td>-10.5</td>
<td>110.25</td>
<td></td>
</tr>
<tr>
<td>-18.800</td>
<td>1</td>
<td>-5.0</td>
<td>25.00</td>
<td></td>
</tr>
<tr>
<td>-0.668</td>
<td>1</td>
<td>-0.5</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>-3.698</td>
<td>1</td>
<td>-1.5</td>
<td>2.25</td>
<td></td>
</tr>
<tr>
<td>-29.600</td>
<td>1</td>
<td>-7.0</td>
<td>49.00</td>
<td></td>
</tr>
<tr>
<td>-36.014</td>
<td>1</td>
<td>-8.0</td>
<td>64.00</td>
<td></td>
</tr>
<tr>
<td>10.447</td>
<td>1</td>
<td>4.5</td>
<td>20.25</td>
<td></td>
</tr>
<tr>
<td>-55.975</td>
<td>1</td>
<td>-11.0</td>
<td>121.00</td>
<td></td>
</tr>
<tr>
<td>-26.716</td>
<td>1</td>
<td>-6.5</td>
<td>42.25</td>
<td></td>
</tr>
<tr>
<td>8.095</td>
<td>1</td>
<td>12.5</td>
<td>156.25</td>
<td></td>
</tr>
<tr>
<td>10.285</td>
<td>1</td>
<td>4.0</td>
<td>16.00</td>
<td></td>
</tr>
<tr>
<td>-6.073</td>
<td>1</td>
<td>-2.0</td>
<td>4.00</td>
<td></td>
</tr>
<tr>
<td>-33.091</td>
<td>1</td>
<td>-7.5</td>
<td>56.25</td>
<td></td>
</tr>
<tr>
<td>8.167</td>
<td>1</td>
<td>12.5</td>
<td>156.25</td>
<td></td>
</tr>
<tr>
<td>9.091</td>
<td>1</td>
<td>12.0</td>
<td>144.00</td>
<td></td>
</tr>
<tr>
<td>11.266</td>
<td>1</td>
<td>10.5</td>
<td>110.25</td>
<td></td>
</tr>
<tr>
<td>4.803</td>
<td>1</td>
<td>1.5</td>
<td>2.25</td>
<td></td>
</tr>
<tr>
<td>-67.356</td>
<td>1</td>
<td>-12.5</td>
<td>156.25</td>
<td></td>
</tr>
<tr>
<td>-55.788</td>
<td>1</td>
<td>-11.0</td>
<td>121.00</td>
<td></td>
</tr>
<tr>
<td>7.803</td>
<td>1</td>
<td>2.5</td>
<td>6.25</td>
<td></td>
</tr>
<tr>
<td>8.769</td>
<td>1</td>
<td>3.0</td>
<td>9.00</td>
<td></td>
</tr>
<tr>
<td>3.483</td>
<td>1</td>
<td>1.0</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>9.066</td>
<td>1</td>
<td>11.5</td>
<td>132.25</td>
<td></td>
</tr>
<tr>
<td>-24.695</td>
<td>1</td>
<td>-6.0</td>
<td>36.00</td>
<td></td>
</tr>
<tr>
<td>11.196</td>
<td>1</td>
<td>5.5</td>
<td>30.25</td>
<td></td>
</tr>
<tr>
<td>-41.671</td>
<td>1</td>
<td>-9.0</td>
<td>81.00</td>
<td></td>
</tr>
<tr>
<td>10.370</td>
<td>1</td>
<td>10.5</td>
<td>110.25</td>
<td></td>
</tr>
<tr>
<td>1.998</td>
<td>1</td>
<td>0.5</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>12.620</td>
<td>1</td>
<td>6.0</td>
<td>36.00</td>
<td></td>
</tr>
<tr>
<td>$y$</td>
<td>$x_0$</td>
<td>$x_1$</td>
<td>$x_1^2$</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>-41.644</td>
<td>1</td>
<td>-9.0</td>
<td>81.00</td>
<td></td>
</tr>
<tr>
<td>1.987</td>
<td>1</td>
<td>0.5</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>-8.282</td>
<td>1</td>
<td>-2.5</td>
<td>6.25</td>
<td></td>
</tr>
<tr>
<td>-10.339</td>
<td>1</td>
<td>-3.0</td>
<td>9.00</td>
<td></td>
</tr>
<tr>
<td>-32.116</td>
<td>1</td>
<td>-7.5</td>
<td>56.25</td>
<td></td>
</tr>
<tr>
<td>12.540</td>
<td>1</td>
<td>9.5</td>
<td>90.25</td>
<td></td>
</tr>
<tr>
<td>12.867</td>
<td>1</td>
<td>9.0</td>
<td>81.00</td>
<td></td>
</tr>
<tr>
<td>-64.463</td>
<td>1</td>
<td>-12.0</td>
<td>144.00</td>
<td></td>
</tr>
<tr>
<td>11.852</td>
<td>1</td>
<td>8.0</td>
<td>64.00</td>
<td></td>
</tr>
<tr>
<td>13.278</td>
<td>1</td>
<td>7.5</td>
<td>56.25</td>
<td></td>
</tr>
<tr>
<td>-44.846</td>
<td>1</td>
<td>-9.5</td>
<td>90.25</td>
<td></td>
</tr>
<tr>
<td>7.041</td>
<td>1</td>
<td>2.0</td>
<td>4.00</td>
<td></td>
</tr>
<tr>
<td>-36.532</td>
<td>1</td>
<td>-8.0</td>
<td>64.00</td>
<td></td>
</tr>
<tr>
<td>6.646</td>
<td>1</td>
<td>-11.5</td>
<td>132.25</td>
<td></td>
</tr>
<tr>
<td>-60.720</td>
<td>1</td>
<td>-11.5</td>
<td>132.25</td>
<td></td>
</tr>
<tr>
<td>9.503</td>
<td>1</td>
<td>11.0</td>
<td>121.00</td>
<td></td>
</tr>
<tr>
<td>-19.789</td>
<td>1</td>
<td>-5.0</td>
<td>25.00</td>
<td></td>
</tr>
<tr>
<td>12.255</td>
<td>1</td>
<td>10.0</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td>10.405</td>
<td>1</td>
<td>5.0</td>
<td>25.00</td>
<td></td>
</tr>
<tr>
<td>9.837</td>
<td>1</td>
<td>4.5</td>
<td>20.25</td>
<td></td>
</tr>
<tr>
<td>-48.172</td>
<td>1</td>
<td>-10.0</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td>-38.107</td>
<td>1</td>
<td>-8.5</td>
<td>72.25</td>
<td></td>
</tr>
<tr>
<td>11.657</td>
<td>1</td>
<td>7.0</td>
<td>49.00</td>
<td></td>
</tr>
<tr>
<td>11.482</td>
<td>1</td>
<td>6.5</td>
<td>42.25</td>
<td></td>
</tr>
<tr>
<td>-2.965</td>
<td>1</td>
<td>-1.5</td>
<td>2.25</td>
<td></td>
</tr>
<tr>
<td>11.612</td>
<td>1</td>
<td>8.0</td>
<td>64.00</td>
<td></td>
</tr>
<tr>
<td>-64.739</td>
<td>1</td>
<td>-12.0</td>
<td>144.00</td>
<td></td>
</tr>
<tr>
<td>2.048</td>
<td>1</td>
<td>0.0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>10.534</td>
<td>1</td>
<td>11.5</td>
<td>132.25</td>
<td></td>
</tr>
<tr>
<td>-13.179</td>
<td>1</td>
<td>-4.0</td>
<td>16.00</td>
<td></td>
</tr>
<tr>
<td>-46.555</td>
<td>1</td>
<td>-9.5</td>
<td>90.25</td>
<td></td>
</tr>
<tr>
<td>-17.541</td>
<td>1</td>
<td>-4.5</td>
<td>20.25</td>
<td></td>
</tr>
<tr>
<td>12.448</td>
<td>1</td>
<td>10.0</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td>-20.498</td>
<td>1</td>
<td>-5.5</td>
<td>30.25</td>
<td></td>
</tr>
<tr>
<td>-10.868</td>
<td>1</td>
<td>-3.5</td>
<td>12.25</td>
<td></td>
</tr>
<tr>
<td>-15.237</td>
<td>1</td>
<td>-4.0</td>
<td>16.00</td>
<td></td>
</tr>
<tr>
<td>12.288</td>
<td>1</td>
<td>5.0</td>
<td>25.00</td>
<td></td>
</tr>
<tr>
<td>-30.878</td>
<td>1</td>
<td>-7.0</td>
<td>49.00</td>
<td></td>
</tr>
<tr>
<td>-17.645</td>
<td>1</td>
<td>-4.5</td>
<td>20.25</td>
<td></td>
</tr>
<tr>
<td>6.284</td>
<td>1</td>
<td>1.5</td>
<td>2.25</td>
<td></td>
</tr>
<tr>
<td>-1.018</td>
<td>1</td>
<td>-1.0</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>-37.788</td>
<td>1</td>
<td>-8.5</td>
<td>72.25</td>
<td></td>
</tr>
<tr>
<td>7.564</td>
<td>1</td>
<td>12.0</td>
<td>144.00</td>
<td></td>
</tr>
</tbody>
</table>
Table 5.3 (Continued)

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>x₀</th>
<th>x₁</th>
<th>x₁²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-4.559</td>
<td>1</td>
<td>-2.0</td>
<td>4.00</td>
</tr>
<tr>
<td></td>
<td>-61.149</td>
<td>1</td>
<td>-11.5</td>
<td>132.25</td>
</tr>
<tr>
<td></td>
<td>-22.982</td>
<td>1</td>
<td>-6.0</td>
<td>36.00</td>
</tr>
<tr>
<td></td>
<td>13.669</td>
<td>1</td>
<td>8.5</td>
<td>72.25</td>
</tr>
<tr>
<td></td>
<td>-1.759</td>
<td>1</td>
<td>-0.5</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>10.529</td>
<td>1</td>
<td>3.5</td>
<td>12.25</td>
</tr>
<tr>
<td></td>
<td>7.059</td>
<td>1</td>
<td>3.0</td>
<td>9.00</td>
</tr>
<tr>
<td></td>
<td>-9.020</td>
<td>1</td>
<td>-2.5</td>
<td>6.25</td>
</tr>
<tr>
<td></td>
<td>-13.245</td>
<td>1</td>
<td>-3.5</td>
<td>12.25</td>
</tr>
<tr>
<td></td>
<td>5.231</td>
<td>1</td>
<td>1.0</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>7.692</td>
<td>1</td>
<td>2.0</td>
<td>4.00</td>
</tr>
<tr>
<td></td>
<td>-0.300</td>
<td>1</td>
<td>0.0</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>13.122</td>
<td>1</td>
<td>5.5</td>
<td>30.25</td>
</tr>
<tr>
<td></td>
<td>10.618</td>
<td>1</td>
<td>6.0</td>
<td>36.00</td>
</tr>
<tr>
<td></td>
<td>11.003</td>
<td>1</td>
<td>7.0</td>
<td>49.00</td>
</tr>
<tr>
<td></td>
<td>-50.977</td>
<td>1</td>
<td>-10.5</td>
<td>110.25</td>
</tr>
</tbody>
</table>

by 0.5. Also included in the model are $x_1^2$ and $x_0=1$, an intercept term. A true solution

$$(\beta_0, \beta_1, \beta_2) = (1.2, 3.0, -0.2) \quad (5.16)$$

was chosen and the $y_i$ were generated as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + e_i$$

where the $e_i$ are independent and identically distributed as Uniform (-1.5, 1.5).

The model, (5.15), was first fit using least squares. The least squares estimator is given by

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (1.0856, 3.0217, -0.1987). \quad (5.17)$$
Figure 5.1. Half-normal plot of least squares residuals for a sample problem
In examining the least squares residuals, we may first notice that the sample kurtosis of the residuals is 1.9274. The maximum absolute least squares residual is

$$\hat{e} = 1.5677.$$ 

Figure 5.1 gives a half-normal plot of the least squares residuals. A half-normal plot is simply a plot of the ordered absolute least squares residuals versus the expected absolute standard normal order statistics. A uniform error distribution manifests itself as a somewhat sigmoid shaped curve which is convex for small values of the abscissa and concave for larger values of the abscissa. While the concavity at the upper end of the curve is easily distinguishable, there is not a nicely defined convex pattern at the lower end of the curve. But, in view of the sample kurtosis of 1.9274, the application of Chebyshev estimation is suggested.

The Chebyshev estimator of the parameters in the model (5.15) is given by

$$(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2) = (1.1487, 3.0097, -0.1993).$$

The maximum absolute Chebyshev residual is

$$\tilde{e} = 1.4484.$$
While we have no explicit formula for the variance of the Chebyshev estimator, we would expect, by interpolating in Table 5.1, that the variance of the Chebyshev estimator is roughly one-fourth that of the least squares estimator.
CHAPTER VI. CONCLUSIONS AND SUMMARY

In the preceding chapters we have attempted to motivate the study of alternatives to least squares in regression estimation, in particular, the criterion of minimizing the maximum absolute residual. Chapters II and III make up a significant portion of the thesis in dealing with algorithms for Chebyshev estimation. Chapter IV has collected some pertinent results concerning the limiting distribution of the Chebyshev estimator. Lastly, we have tried to give some discussion and make recommendations on the use of Chebyshev estimation. The motivating force has been to present Chebyshev estimation as a viable alternative to least squares in the appropriate situations.

As in almost any research, there are significant questions which remain unanswered. The most crucial of unanswered questions is that of the sampling distribution of the Chebyshev estimator. This is a seemingly intractable problem. There does seem to be some hope for the discovery of an asymptotic distribution theory for Chebyshev estimation, particularly by the methods employed by Koenker and Bassett (1978) in deriving the limiting distribution of the $L_1$ estimator.

Even the development of an approximation to the distribution of the Chebyshev estimator would be a significant
contribution to the theory of Chebyshev estimation. Rosen­berg and Carlson (ca. 1972) studied a normal approximation to the distribution of the $L_1$ estimator. They approximated the distribution of the $L_1$ estimator of $\hat{\beta}$ as a multivariate normal distribution with mean vector $\hat{\beta}$ and variance-covariance matrix $\omega^2(X'X)^{-1}$ where $\omega^2$ denotes the variance of the sample median from a sample of size $n$ from the given error distribution. Their simulation study showed that the normal approximation performed very well and eventually Koenker and Bassett (1978) proved that the approximation of Rosenberg and Carlson (ca. 1972), in fact, gave the asymptotic distribution of the $L_1$ estimator.

The development of a distribution theory for Chebyshev estimation - whether it is exact, asymptotic, or approximate - is essential to the resolution of many other unanswered questions. Once the distribution theory is treated, it will be possible to derive the variance of the Chebyshev estimator, tests of significance for the parameter estimates, and confidence regions. It would also be interesting to attempt to develop an approximate analysis of variance for Chebyshev estimation. We must emphasize that any such procedure could be approximate, at best, since the concept of orthogonal decomposition does not exist for Chebyshev estimation.

Another topic that deserves considerable attention is
the development of model building techniques. For example, given a set of $m$ explanatory variables, how does one choose the best subset of size $p$? What if one doesn't have a particular $p$ in mind? How many explanatory variables should be used? Given several candidate models we wish to consider, what is a reasonable criterion for choosing one model over the others? In least squares there are obvious criteria for ruling on these questions, but not so for Chebyshev estimation. These are all very serious issues to the statistical practitioner.

We have examined several aspects of Chebyshev estimation. While it would appear that the mechanics of the estimation procedure are well-understood, the theoretical implications of its use are not. With all the theoretical difficulties one encounters in considering alternatives, it is no small wonder that least squares is considered the regression criterion. One can hardly justify using alternative procedures whose properties are unknown.

More research is needed on the properties of alternative norm estimators. Once these estimators are well-understood, there is no question that they will have numerous applications. The analysis of residuals is a technique that has been known for years (see Anscombe (1961) and Anscombe and Tukey (1963)). We feel that residual analysis is not employed often enough.
in practice, primarily because there are not enough viable alternatives to least squares when normality assumptions are violated. That is, we feel that blind application of least squares is practiced, not because it is always thought to be the right thing to do, but for lack of alternatives. Therefore, alternative norm estimation, and in particular Chebyshev estimation, should be the worthwhile subject of significant future research.


Rosenberg, Barr. ca. 1972. Some results concerning the exact sampling distribution of least absolute residuals regression estimates. Multilithed paper. School of Business Administration, University of California, Berkeley.

Rosenberg, Barr and Daryl Carlson. ca. 1972. A simple approximation to the sampling distribution of least absolute residuals regression estimates. Multilithed paper. School of Business Administration, University of California, Berkeley.


ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my major professor, Professor Vincent A. Sposito. Professor Sposito has provided valuable insight and guidance with respect to every aspect of my graduate education. It has been an honor and a pleasure to work under a man who gives so freely of himself and his time. I am grateful to Professor Sposito, not only for his contributions to my education, but for his friendship which I value deeply.

I also wish to express my thanks to the remaining members of my graduate committee; Dr. David F. Cox, Dr. James E. Gentle, Dr. David A. Harville, and Dr. Clair G. Maple.

I gratefully acknowledge the financial support of the Statistical Laboratory at Iowa State University and, in particular, I wish to thank my immediate supervisors, Dr. William J. Kennedy and Dr. David F. Cox. In learning any discipline, there is no substitute for experience and I feel privileged to have had the opportunity to work as part of the Statistical Laboratory.

I would also like to thank my family, particularly my parents, for the encouragement and support they have given me over the years.

Finally, I express my deepest appreciation to my wife
Katherine, who has lived with me through good times and bad. Only through her love, patience, and understanding has the completion of this work been possible.