## 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 photoing 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.

**Xerox University Microfilms**
300 North Zeeb Road
Ann Arbor, Michigan 48106

74-15,668

BASSETT, Gilbert Wallace, Jr., 1947-
SOME PROPERTIES OF THE LEAST ABSOLUTE ERROR ESTIMATOR.

The University of Michigan, Ph.D., 1973
Statistics

University Microfilms, A XEROX Company, Ann Arbor, Michigan

THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED.

# SOME PROPERTIES OF THE LEAST ABSOLUTE ERROR ESTIMATOR

by

Gilbert Wallace Bassett, Jr.

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
1973

Doctoral Committee

Professor Lester D. Taylor, Chairman
Professor Bruce M. Hill
Professor Saul H. Hymans
Associate Professor Michael E. Manove

DEDICATION

For My Parents

## ACKNOWLEDGMENTS

I have benefitted from discussions with all of the members of my dissertation committee. Valuable assistance and advice was also provided by Professor Robert M. Stern. My greatest debt, however, is owed to Professor Lester Taylor, who has been a constant source of encouragement. He has never failed to listen to problems and offer insightful suggestions.

Financial assistance by the Ford Foundation in the form of a Rackham Prize Fellowship is gratefully acknowledged. I also want to thank Pat Rapley for the speed and efficiency with which she typed the manuscript and the various drafts. Finally, I want to thank my wife and son for sharing the burdens of this study.

TABLE OF CONTENTS


| | | Page |
|---|---|---|
| DEDICATION | | ii |
| ACKNOWLEDGMENTS | | iii |
| LIST OF FIGURES | | vi |
| LIST OF TABLES | | vii |
| Chapter | | |
| I | THE LINEAR REGRESSION MODEL: ASSUMPTIONS AND BELIEFS | 1 |
| | 1. The Linear Regression Model | 3 |
| | 2. The Least Squares Estimator | 6 |
| | 3. A Preference Ordering for Least Squares | 9 |
| | 4. A Decision Problem | 11 |
| | 5. Assumptions on F(u) and P(F) | 14 |
| | 6. Intuitive Reasons for Studying L.A.E. | 23 |
| | 7. Summary | 25 |
| II | THE L.A.E. ESTIMATOR: DEFINITION AND PROPERTIES | 29 |
| | 1. Introduction | 29 |
| | 2. The L.A.E. Minimization Problem | 30 |
| | 3. Properties of $\beta^*, u^*$ | 33 |
| | 4. Computational and Mathematical Problems | 45 |
| III | THE L.A.E. LINEAR PROGRAMMING PROBLEM | 49 |
| | 1. A Linear Programming Problem | 50 |
| | 2. The Dual L.A.E. Problem | 55 |
| | 3. Properties of the L.A.E. Linear Programming Problem | 60 |
| | 4. Monte Carlo Results | 66 |
| | 5. Conclusion | 72 |
| IV | SOLUTION TO THE L.A.E. LINEAR PROGRAMMING PROBLEM | 75 |
| | 1. Linear Programming | 76 |
| | 2. The Solution to the Linear Programming Problem | 81 |

Chapter Page

3. Conditions for a Unique Solution . . . . 87
4. A Characterization of $\beta^*$ . . . . . 91
5. A Mathematical Result . . . . . 95
6. $\beta^*(Y)$ As A Linear Estimator . . . . . 98
7. Conclusion . . . . . 102

V THE L.A.E. ESTIMATOR: THE MEDIAN . . . . . 104

1. The Median . . . . . 105
2. Sampling Density of $(\beta^*-\beta)$ . . . . . 107
3. Properties of $g(\theta)$ . . . . . 110
4. Variance of $\beta^*$ . . . . . 112
5. The Student Density . . . . . 122
6. Conclusion . . . . . 127

VI THE SAMPLING DISTRIBUTION OF THE L.A.E. ESTIMATOR . . . . . 129

1. The L.A.E. Sampling Density . . . . . 129
2. Some Mathematical Problems . . . . . 134
3. Finite Properties of $g(\phi)$ . . . . . 137
4. Consistency of $\beta^*$ . . . . . 138
5. Summary . . . . . 143

VII SUMMARY AND SUGGESTIONS FOR FUTURE RESEARCH . . . . . 145

1. Summary of L.A.E. Properties . . . . . 145
2. Explanation of Monte Carlo Results . . . 150
3. The Relationship Between $\beta^*$ and X . . . 152
4. The Use of the L.A.E. Estimator . . . . 154

APPENDIX I . . . . . 158

APPENDIX II . . . . . 161

APPENDIX III . . . . . 167

REFERENCES . . . . . 170

LIST OF FIGURES


| Figure | | Page |
|---|---|---|
| 1 | Symmetric Stable Densities . . . . . | 20 |
| 2 | Comparison of L.A.E. and Least Squares Estimates . . . . . | 24 |
| 3 | The Sets $\hat{D}$ and D* . . . . . | 34 |
| 4 | Illustration of Property 1 . . . . . | 36 |
| 5 | An Example of a Unique Solution . . . . . | 36 |
| 6 | Illustration of Property 2 . . . . . | 39 |
| 7 | Illustration of Property 3 . . . . . | 40 |
| 8 | Illustration of Property 5 . . . . . | 44 |
| 9 | Geometry of the Dual L.A.E. Problem . . . . . | 57 |
| 10 | Illustration of Property 6 . . . . . | 64 |
| 11 | The Laplace and Normal Densities . . . . . | 117 |
| 12 | Densities Such That, $Var\hat{\beta} < Var\beta^*$ . . . . . | 120 |
| 13 | An Approximate Partitioning of $F_2^*$ . . . . . | 123 |
| 14 | Student Densities Such That, $Var\beta^* < Var\hat{\beta}$; T = 11 . . . . . | 124 |
| 15 | Student Densities Such That, $Var\beta^* < Var\hat{\beta}$; T = 41 . . . . . | 124 |
| 16 | Student Densities Such That, $Var\beta^* < Var\hat{\beta}$; T = 101 . . . . . | 124 |
| 17 | Student Densities . . . . . | 126 |
| 18 | An Example where $\beta^*$ is Inconsistent . . . . . | 140 |

## LIST OF TABLES


| Table | | Page |
|---|---|---|
| 1 | Taylor's Monte Carlo Experiments . . . . . | 70 |
| 2 | Rosenberg and Carlson's Monte Carlo Experiments . . . . . | 70 |

# CHAPTER I

## THE LINEAR REGRESSION MODEL: ASSUMPTIONS AND BELIEFS

In this study we analyze the properties of the estimator which minimizes the sum of absolute errors. We use the notation L.A.E., for least absolute error, to refer to this estimator. We are primarily interested in comparing the properties of this estimator with the widely used least squares estimator under alternative assumptions on the probability distribution of the data generating process in economic models.

The assumptions which are used throughout this study are presented in this chapter. The prior beliefs which motivate our study of the L.A.E. estimator are also described. We have adopted a set of assumptions which leads to the choice of the least squares estimator, except we allow for the possibility of a nonnormal probability distribution. We argue that for economic models the normality assumption is not necessarily justified so that it becomes necessary to examine the properties of alternative estimators under a range of probability distributions. The optimal properties associated with the least squares estimator which are conditional on the normality assumption therefore become only one input into the problem of choosing an estimator.[1] Within this framework, the choice of an estimator depends on, among other

factors, the prior beliefs concerning the probability distribution of the data-generating process. These considerations lead to the study of alternatives to the least squares estimator.[2]

We consider the L.A.E. estimator as an alternative to the least squares estimator for two reasons. First, the practical application of the L.A.E. estimation method is feasible in the sense that computational costs are not significantly greater than for least squares. Second, for the set of probability distributions which we argue is appropriate in economic models, the L.A.E. estimator, at least intuitively, appears to possess desirable properties. It is by studying the properties of the L.A.E. estimator that we attempt to give this intuition some quantitative content.

This chapter is divided into seven sections. In Section 1 the linear regression model is defined, and we introduce our assumptions on the probability distribution. In Section 2 the properties of the least squares estimator are discussed. These properties are used in Section 3 to construct a preference ordering with which to rank estimators. In Section 4 we briefly indicate the nature of the decision problem which arises when there is uncertainty concerning the probability distribution of the data generating process. Section 5 includes the theoretical and empirical reasons for our assumptions on the probability distribution. In view of these assumptions, Section 6 presents at an intuitive level the reasons for studying the L.A.E. estimator. The conclusions to

this chapter are summarized in Section 7, along with a preview of the remaining chapters in this study.

## 1. The Linear Regression Model

### Assumptions

Given T observations on a dependent variable $y_t$, and K explanatory variables $x_{tk}$, we assume the observations obey the linear relationship:

$$y_t = \sum_{k=1}^{K} x_{tk} \beta_k + u_t \qquad t=1,\ldots,T$$

where $u_t$ is an unobservable random error term. This assumption in matrix notation is written:

A1. $$Y = X\beta + u \quad .$$

The linear regression model is defined by A1 along with the following assumptions on $\beta$, X, and u. The problem is to estimate the parameter vector $\beta$, and the error term u.

For $\beta$ and X the following assumptions are used:

A2. $\beta$ is a completely unknown vector in $R^K$

A3. rank X = K < T

A4. X is fixed; or

A4a. (X,u) are statistically independent.

Assumption A2 implies no prior knowledge is used to estimate $\beta$. Assumption A3 excludes any cases in which the columns of X are linearly dependent, and thereby rules out perfect

multicollinearity. Assumption A4 implies the values of X can be controlled or predetermined, while A4a allows X to be random, but independent of u. We will use A4, although all of the analysis remains valid under A4a, conditional on the observed X.

For $u = (u_1,\ldots,u_T)$ we assume:

A5. $u_t$ : t=1,...,T, are independent, identically distributed random variables.

Finally, the standard assumption for F(u),[3] the probability distribution of u, is that F(u) is a normal distribution, or F(u) is contained in the set of probability distributions $\hat{F}$ which is defined by;

Properties of $\hat{F}$

1. $E(u) = 0$
2. $Var(u) = \sigma^2 < \infty$

where $E(\cdot)$ and $Var(\cdot)$ are the expectation and variance operators, respectively. We modify this assumption in two primary ways. First, we exclude asymmetric probability distributions from $\hat{F}$, and second, we include probability distributions with undefined expectations or infinite variance. This set of probability distributions is denoted by F* and it is defined by;

Properties of F*

1. F(u) has continuous density, f(u).
2. $f(u) = f(-u)$

3. $f(0) < \infty$ .

Properties 1 and 3 are included for mathematical convenience. We assume F(u) is in F*; the reasons for this assumption are described in Section 5 below.

For our purposes, the linear regression model is defined by assumptions A1 to A5 along with F(u) in F*.[4]

Comment

Assumptions A1 to A5 are used in this study so that the L.A.E. and least squares estimators can be compared under alternative assumptions on F(u), while maintaining the other assumptions which lead to the choice of the least squares estimator. Assumptions A1 to A5 simplify the comparison, and they are not intended to be an adequate description of the experimental situation in economics. For example, assumption A2 is inappropriate when economic theory implies restrictions on $\beta$, or when previous estimates of $\beta$ exist. Assumptions A4 or A4a may also be inappropriate since theory implies that economic variables are interdependent. Finally, economic data are usually autocorrelated and this violates assumption A5.

Most of the work in econometrics has focused on these problems as evidenced by the study of Bayes estimators, simultaneous equation systems, and autocorrelation. To my knowledge, the normality assumption has been maintained in these studies so that the resulting estimators are

generalizations of the ordinary least squares estimator which is obtained under assumptions A1 to A5.[5] In this study we ignore issues dealing with simultaneous equations, autocorrelation, and prior information on $\beta$, and expect that these can be eventually dealt with using the L.A.E. estimator in ways that are analogous to the generalization of the least squares estimator.

## 2. The Least Squares Estimator

### Definition and Properties

The least squares estimator, $\hat{\beta}$, is defined as the solution to:

$$\min_{b \in R^K} \ (Y-Xb)'(Y-Xb) \quad .$$

This minimization problem always has a unique solution, and yields a well-defined estimator. A simple consequence of the quadratic minimization problem is that the solution is a linear function of Y.

One of the well-known properties of $\hat{\beta}$ is:

Theorem 1 (Cramer-Rao). Given A1 to A5 and F(u) is the normal distribution implies $\hat{\beta}$ is the minimum variance estimator in the class of all regular[6] unbiased estimators.

Proof: See Theil (1971), p. 390.

We assume that this theorem along with the belief that F(u) is normal provides the primary justification for the widespread use of the least squares estimator.

As is also well-known, the hypothesis in Theorem 1 can be weakened to include F(u) in $\hat{F}$ with the conclusion then being strengthened to linear estimators. That is:

Theorem 2 (Gauss-Markov). Given A1 to A5 and F(u) in $\hat{F}$ implies $\hat{\beta}$ is the minimum variance estimator in the class of all linear unbiased estimators.

Proof: See Theil (1971), p. 119.

The preference orderings implied by the choice of the least squares estimator on the basis of Theorems 1 and 2 are described below. Before looking at this ordering we briefly examine the characteristics; unbiasedness, variance, and linearity.

Unbiasedness. To choose an estimator which, on the average, is the correct value of a parameter at least casually seems preferable to choosing an estimator which on the average is incorrect. This reasoning, however, does not lead to the exclusive consideration of the class of unbiased estimators, unless one is willing to ignore all of the other characteristics that might be relevant in choosing an estimator. This seems to be a restrictive constraint, and yet, to my knowledge, it is necessary for the choice of the least squares estimator.

Minimum Variance. The significance of the variance of alternative estimators is based upon the asymptotic normality of a large class of estimators for $\beta$. The dispersion of a normal random variable is a function of the variance so for large T the dispersion of estimators is a function of the asymptotic variance. This measure of dispersion loses its

significance when the number of observations is not large enough to justify the normal approximation (or when the finite sample variance does not converge to the variance of the asymptotic normal distribution). The measure of dispersion then makes a difference since the minimum variance (unbiased) estimator may not be minimum dispersion when dispersion is measured by, say, mean absolute deviation. In order to conform to the least squares approach, we adopt the variance as the measure of dispersion even though a mean absolute deviation measure would probably be more favorable for L.A.E.

Linearity. Two practical considerations dictate a preference for linear estimators. First, the computational costs of nonlinear relative to linear estimators are usually prohibitive. Second, the mathematical intractability of many nonlinear estimators makes the derivation of sampling properties difficult, if not impossible. For these two reasons, one may want to restrict attention to only linear estimators. The dilemma which arises in attempting to calculate the opportunity cost of this restriction (in terms of, say, increased variance) is that one has to face the computational and mathematical difficulties which originally led to the preference for linearity. The L.A.E. estimator provides a partial solution to this dilemma since the computational costs are not significantly different than for least squares. Also, while the L.A.E. estimator is mathematically more complicated than least squares, there are methods which can be used to obtain

sampling properties, and developing these methods is one of the aims of this study.

## 3. A Preference Ordering for Least Squares

In order to compare estimators, we need a preference ordering with which to rank estimators. We adopt a preference ordering implied by the choice of the least squares estimator under the normality assumption, and then given this ordering examine the L.A.E. and least squares estimators as F(u) varies in F*. Here we describe an ordering which follows trivially from the choice of the least squares estimator on the basis of the Cramer-Rao Theorem.

We associate with any estimator, $\tilde{\beta}$, the set of characteristics bias squared and variance. That is;

$$((E(\tilde{\beta}-\beta))^2, \quad \mathrm{Var}(\tilde{\beta})) \quad .$$

(When $\tilde{\beta}$ is a vector then $(E(\tilde{\beta}-\beta))^2$ is a vector and $\mathrm{Var}(\tilde{\beta})$ is the variance-covariance matrix.) Estimators with the same bias squared and variance are considered equivalent even if they differ on some other characteristics. To obtain a preference ordering first rank estimators according to bias squared; for example, every unbiased estimator is preferred to any biased estimator, irrespective of the corresponding variances. Second, for any given bias squared, rank estimators according to variance. For this ordering the preferred estimator corresponds to the minimum variance estimator in the set of all unbiased estimators. By Theorem 1 these

characteristics correspond to the least squares estimator when F(u) is the normal distribution.

The distinctive feature of this preference ordering is that it is lexicographic. That is, preferences arise by considering bias and variance in order; no tradeoffs between the two are allowed. An unbiased estimator is preferred to a biased estimator even if the bias is small and the variance of the biased estimator is small relative to the unbiased estimator. Also, we have used the term preference ordering instead of utility function because of the fairly well-known result that lexicographic preference orderings cannot be represented by utility functions (see, for example, Gorman (1972)).

In this study we adopt the above lexicographic preference ordering and thereby focus on the unbiasedness and variance of the L.A.E. estimator. Later we show that for F(u) in F* the L.A.E. estimator is unbiased so that the comparison of the L.A.E. and least squares estimators reduces to a comparison of their respective variances. This provides a simple framework in which assumptions on F(u) can be translated into a ranking of the L.A.E. and least squares estimator. By adopting the lexicographic ordering over unbiasedness and variance we can investigate, to the extent that this ordering actually reflects the process which leads to the choice of the least squares estimator, those F(u) which lead to the choice of the L.A.E. estimator.[7]

In adopting the above lexicographic ordering, we ignore the Gauss-Markov Theorem as a basis for choosing the least squares estimator. This conflicts with most discussions in which least squares is viewed as a desirable "nonparametric" estimator since the conclusion to the Gauss-Markov Theorem holds for all F(u) in $\hat{F}$. We think the use of "nonparametric" in these discussions is misleading. To investigate nonparmetric properties requires an analysis of the properties of estimators over some range of F(u). The conclusion to be drawn from the Gauss-Markov Theorem therefore is that in investigating alternatives to the least squares estimator we can ignore all other linear estimators. However, there may exist a nonlinear unbiased estimator with a much smaller variance than least squares for all but those F(u) in some small neighborhood of the normal distribution. This possibility is not ruled out by the Gauss-Markov Theorem and, other things equal (for example, computational costs), we think most would consider this a desirable (in terms of the above lexicographic preference ordering) nonparametric estimator.

## 4. A Decision Problem

We noted above that for F(u) in F* and the lexicographic preference ordering over unbiasedness and variance, the comparison of the L.A.E. and least squares estimators reduces to a comparison of their respective variances. We can therefore define,

$$G(F) = \frac{\text{Var } \beta^*}{\text{Var } \hat{\beta}}$$

where $\beta^*$ is the L.A.E. estimator, and G(F) indicates that the relative variances depend on F(u). Then for any given F(u);

$G(F) < 1$ implies $\beta^*$ is preferred to $\hat{\beta}$

$G(F) = 1$ implies $\beta^*$ is indifferent to $\hat{\beta}$

$G(F) > 1$ implies $\hat{\beta}$ is preferred to $\beta^*$.

This provides an ordering of $\beta^*$ and $\hat{\beta}$ conditional on F(u). We now outline the problem which arises when there is uncertainty concerning F(u).[8] This problem, along with the arguments presented in the next section, motivates our study of the L.A.E. estimator by providing a framework in which it becomes necessary to analyze estimators under a range of F(u).

We first introduce the function P(F). This function can be viewed as a way of representing prior beliefs on F(u), or as merely a device for analyzing G(F) for a range of F(u). In either case, P(F) has all the properties which define a probability distribution, and therefore G(F) is a random variable. This makes the choice of an estimator depend on P(F) as well as G(F).

An example in which the use of P(F) does not complicate the choice of an estimator is when;

$$G(F) > 1 \quad \text{for all } F(u) \text{ such that } P(F) > 0$$

and, in particular, this occurs when;

$$P(F \text{ is the normal distribution}) = 1 \quad .$$

Here the least squares estimator dominates the L.A.E. estimator. This can be viewed as the classical assumption since it leads to the choice of the least squares estimator.

The use of the above assumption demonstrates the asymmetric treatment of the unobservable parameters in the usual treatment of the linear regression model. Extensive rationalization for the normality assumption is usually presented so that investigators at least act as if $F(u)$ is the normal distribution with probability one. This is in contrast to the other unobservable parameter, $\beta$, which is usually assumed to be with equal probability anywhere in $R^K$. This diffuse prior on $\beta$ and strong prior on $F(u)$ conflicts with most economic models since there is usually prior knowledge concerning $\beta$, and as we argue in the next section, there is no reason to believe that $F(u)$ is a normal distribution with probability one.

In general, the choice between the L.A.E. and least squares estimator becomes more complicated when;

$G(F) < 1$ for some $F$ such that $P(F) > 0$ and,

$G(F) > 1$ for some other $F$ such that $P(F) > 0$.

Here neither estimator is dominant, and a new objective function has to be introduced with which to rank estimators. For example, one possibility is to rank estimators by minimizing the maximum variance over all $F$ such that $P(F) > 0$. That is, if we maximize $G(F)$ over all $F(u)$ such that $P(F) > 0$

and if the maximum $G(F) < 1$ then the L.A.E. estimator is preferred to the least squares estimator. Alternatively, one may want to minimize the expected value of $G(F)$. For a discrete $P(F)$ we have;

$$E(G(F)) = \sum_{P(F)>0} G(F)\ P(F)$$

and if $E(G(F)) < 1$ then the L.A.E. estimator is preferred to the least squares estimator.

Except for an example presented in Chapter V, we do not formally analyze the decision problem which arises when there is uncertainty concerning $F(u)$. At present, this analysis is impossible because the properties of the L.A.E. estimator are not sufficiently well-known. However, this decision problem motivates our study of the L.A.E. estimator since we assume;

$$P(F \text{ is the normal distribution}) \neq 1$$

and so the least squares estimator is not the (unconditionally) "best" estimator.

We now present the reasons for assuming $F(u)$ in $F^*$, and describe qualitatively our $P(F)$.

## 5. Assumptions on F(u) and P(F)

In restricting $F(u)$ to $F^*$ we have attempted to characterize a wide class of distributions which are consistent with the casual reasoning used in economic models. The restriction to symmetric distributions (Property 2 of $F^*$) is a

consequence of the following;

> Since many factors may be at work [in determining u]...many may be pulling in opposite directions, we should expect small values of u to occur more frequently than large values. We are thus led to think of u as a variable with probability distribution centered at zero.... (Johnston (1963, p. 6).)

This quote implies Property 2 of F*, and it does not lead to a consideration of the asymmetric distributions in $\hat{F}$. However, $\hat{F}$ only contains distributions with finite variance, and this is usually used along with a central limit theorem to obtain the normality assumption. For example, Johnston concludes the above quote with;

> ...and [u] has finite variance $\sigma_u^2$.... In view of the many factors involved an appeal to the Central Limit Theorem would further suggest a normal distribution for u.

In terms of P(F) this leads to the classical assumption. In this section we present theoretical and empirical evidence to show why it is necessary to study the properties of the L.A.E. and least squares estimators under alternative F(u) in F*. Qualitatively, our P(F) allows for the possibility that F(u) is some nonnormal member of F*, and special attention is given to those F(u) which are more-peaked and fatter-tailed than the normal.

## The Normality Assumption

As is evidenced by the quote from Johnston, the normality assumption arises by making assumptions on the unobservable error term which are sufficient for an appeal to the central limit theorem. Typically, the error term is taken to

be the standardized sum of a large number of independent and identically distributed random variables with zero mean and finite variance; these conditions being sufficient for the asymptotic normality of F(u).[9] These arguments lead us to include the possibility that F(u) is normal, but they do not lead to the exclusive consideration of the normal distribution.

The above central limit theorem implies that F(u) approaches, and is not identical to the normal distribution.[10] Therefore, even if one adopts all of the above assumptions on the unobservable factors which determine the unobservable error term one is led to consider F(u) which are in a neighborhood of the normal distribution (see Huber (1964) and Rice and White (1964)). Also, an appeal to a central limit theorem without the further assumption that u has finite variance does not necessarily lead to the normal distribution. That is, the normal distribution is not the only limiting distribution of a standardized sum of random variables. For economic experiments there does not appear to be any compelling reason to assume Var (u) $< \infty$, and this leads to a consideration of all limiting distributions of a sum of random variables (see discussion below). Finally, unlike some controlled experiments in the physical sciences, it is possible that the probability distribution in economic experiments, especially time series, is not determined by a central limit theorem. These factors lead us to consider the possibility that F(u) is not a normal distribution, and so we allow F(u) to be any distribution in F*.

In practice, it is likely that investigators do not literally believe the assumptions made on the unobservable factors which determine u. Instead, these assumptions are introduced as ex post rationalizations for the choice of least squares. As Box and Tiao (1962) point out;

> Classical statistical arguments lead us to treat assumptions as if they were in some way axiomatic and yet consideration will show that, in fact, they are conjectures which in practice may be expected to be more or less true.... If we assume normality, we can proceed with an 'objective' classical analysis...however...as seems to be inevitably the case in other problems as well as this one, our 'objectivity' is gained by pretending to knowledge we do not have.... (p. 419.)

It is also interesting to note that historically the normality assumption was used to rationalize the choice of the least squares estimator. We quote Gauss' reasons for using the normality assumption.

> The author of the present treatise, who in the year 1797 first investigated this problem according to the principles of the theory of probability, soon realized that it was impossible to determine the most probable value of the unknown quantity, unless the function representing the probability of the errors is known. But since it is not, there is no other recourse than to assume such a function in a hypothetical fashion. It seemed most natural to him to take the opposite approach and to look for that function which must be taken as a base in order that for the simplest of all cases a rule is obtained which is generally accepted as a good one, namely that the arithmetic mean of several observations of equal accuracy for one and the same quantity should be considered the most accurate value. This implied that the probability of an error x must be assumed proportional to an exponential expression of the form $e^{-hhzz}$, and that then just the same method which he had found by other considerations already a few years earlier, would become necessary in general. This method, which afterwards, in particular since 1801, he had almost daily opportunity to use in

> diverse astronomical computations, and which in the meantime also Legendre had happened upon, now is in general use under the name method of least squares. (See Huber (1972).)

This is in contrast to the usual presentation in which least squares follows as a consequence of the strong prior belief in normality.

## Symmetric Stable Distributions

The belief in a central limit theorem along with the symmetry of F(u) leads to the symmetric stable class of distributions. These distributions generalize the classical central limit theorem to include cases in which the necessary conditions for a normal distribution are violated (for example, when the variance of one of the factors which determine u is large). These distributions are defined by the log characteristic function:

$$\log \phi_u(t) = w(t,\alpha) \ [-\sigma|t|^{\alpha}]$$

$$w(t,\alpha) = \begin{cases} \tan \pi\alpha/2 & \text{for } \alpha \neq 1 \\ 2\pi \log |t| & \text{for } \alpha = 1 \end{cases}$$

where $\sigma > 0$ is a scale parameter, and $0 < \alpha \leq 2$ is the characteristic exponent. A discussion of these distributions can be found in Gnedenko and Kolmogorov (1954) or Feller (1966). Here we briefly indicate some properties.

Closed form expressions for the symmetric stable densities are known in only a few cases; $\alpha = 1$ is a Cauchy and $\alpha = 2$ is a normal distribution. Fama and Roll (1970) have

tabulated these densities for selected $\alpha$ by using Taylor series approximations. Some of these densities are plotted in Figure 1. The densities for $\alpha < 2$ are all more peaked and fatter-tailed than the normal. Also, if u is symmetric stable then for $\alpha < 2$;

$$E|u|^{\delta} < \infty \quad \text{iff } \delta < \alpha \quad .$$

That is, the normal is the only distribution with finite variance which can arise as the limiting distribution of a standardized sum of random variables. This is the reason why we do not put any restrictions on the convergence of moments in defining F*.

The infinite variance of the nonnormal stable distributions has resulted in some objections being raised to the plausibility of these distributions. These objections have centered around

a. Boundedness

b. Mixtures.

These objections are now briefly reviewed.

Boundedness (see Roll and Fama (1970)). This argument says that the processes generating economic data must have finite variance because the relevant variables are bounded. In response, we offer two reasons for considering infinite variance distributions. First, the boundedness of observations does not rule out the use of unbounded random variables since the probability of impossible events can be set

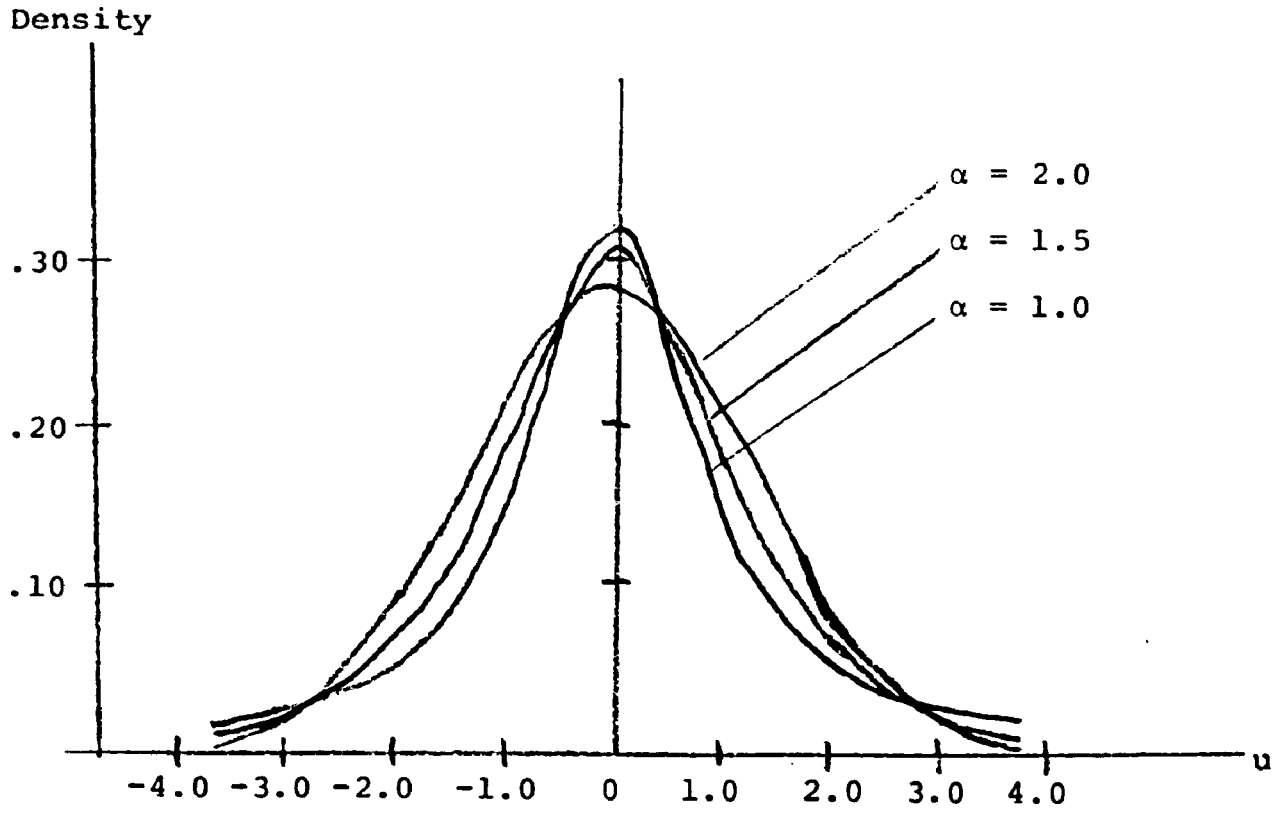



Fig. 1.--Symmetric Stable Densities

These are standardized symmetric stable densities. That is,

$$\log \phi_u(t) = -|t|^{\alpha}$$

where $\phi(t)$ is the characteristic function.
Source: Fama and Roll (1968, p. 824).

arbitrarily small. Second, random variables which are unbounded or have infinite variance while perhaps not a literal description of reality may work well (see the comment by Tsiang (1973) and reply by Fama (1973)). For example, the boundedness argument implies that current actuarial tables are "incorrect" or "not useful" since they predict with probability greater than zero that some person will live until age 1000. The probability of this event is roughly 1 in $10^{10^{36}}$ (see Feller, 1968, p. 7) and this statistical description does not conflict with any observations yet.

Mixtures. This argument says that the peakedness and fatter-tails of empirical frequency functions can be explained by a mixture of normal distributions instead of some symmetric stable distribution. This conclusion is valid; but unless the factors that lead to the mixture can be pointed out, it is also vacuous since any distribution can be approximated by combining normal (or any other) distributions. A mixture of normal distributions is then equivalent to some symmetric stable distribution, and it does not matter how we view F(u). (For some interesting attempts at unmixing, see Clark (1973).)

## Empirical Evidence

Most of the direct empirical evidence on F(u) has centered on estimating the characteristic exponent $\alpha$, given F(u) a member of the symmetric stable class. These tests have

been mostly applied to price changes in speculative markets; see Mandlebrot (1963), Fama (1965), although the first evidence of nonnormality goes back to Pareto's investigation of income distribution. The evidence from these studies suggest that F(u) is a nonnormal stable distribution with $\alpha$ in the range (1.5 to 1.7). This evidence is also consistent with the common result in which empirical frequency functions have too many small and too many large deviations relative to the normal distribution.[11]

For our purposes the indirect evidence is more valuable since it is based on a comparison of the least squares and L.A.E. estimators. Oveson (1968) used the L.A.E. estimator and estimated the equations in Houtthaker and Taylor's (1965) study of consumer demand, and then compared the forecasting accuracy of the resulting equations with the least squares equations. He found L.A.E. had smaller mean absolute and smaller mean squared forecasting error than least squares. Similar results have been obtained by Meyer and Glauber (1964) and Arrow and Hoffenberg (1959). These results provide some evidence that F(u) is nonnormal since, other things equal, least squares should outperform L.A.E. if data are generated by a normal distribution. It is possible that the specification of the models or assumptions A1 to A5 were incorrect in these studies, but this only implies that L.A.E. is less sensitive to these assumptions than least squares.

We therefore consider F(u) in F* and in particular consider the possibility that F(u) is more-peaked and fatter-tailed than the normal distribution.

## 6. Intuitive Reasons for Studying L.A.E.

In view of our P(F) we now present at an intuitive level the reasons for studying the L.A.E. estimator.

In Figure 2 we have plotted some observations for the model;

$$Y = \beta_1 + \beta_2 X \quad .$$

We have also indicated the L.A.E., least squares, and true relationships. Here least squares yields a poor estimate since in minimizing squared deviations the estimate is affected greatly by large values of u. On the other hand, the L.A.E. estimate is not so strongly influenced by large values of u so that in this example the L.A.E. gives a better estimate than least squares.

We know that when F(u) is the normal distribution, least squares is the minimum variance unbiased estimator. Situations like that depicted in Figure 2 can occur, but they occur with such a small probability that their contribution to variance is small. However, when F(u) is allowed to vary in F* so that it becomes more-peaked and fatter-tailed, the probability of observations as in Figure 2 increases. One might then expect the least squares variance to increase relative to the variance of the L.A.E. estimator, and after some point for the variance of the L.A.E. estimator to be less than the variance of least squares. The variance of the L.A.E. estimator might also be expected to be less sensitive than the least squares variance as F(u) varies in F*.

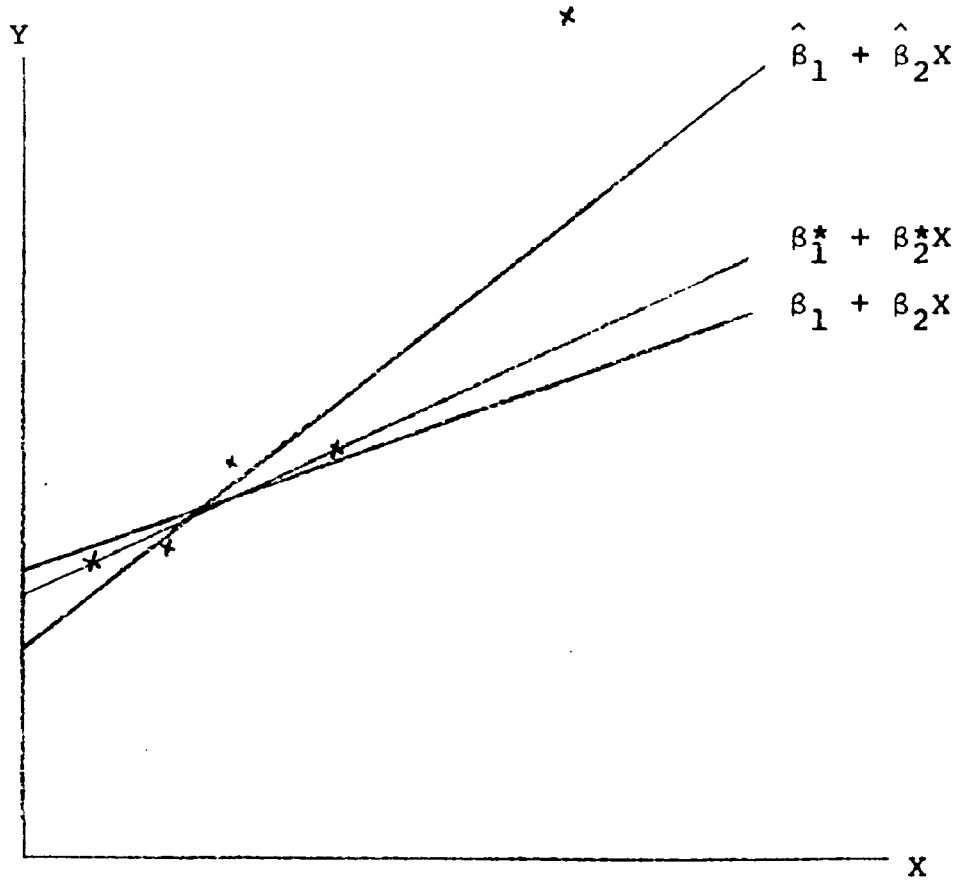



Fig. 2.--Comparison of L.A.E. and Least Squares Estimates

## 7. Summary

In this chapter we have set out the assumptions which are used throughout this study. Our only departure from the classical assumptions which lead to the choice of the least squares estimator is that we assume F(u) is some unknown member of F*. This provides a framework with which we can compare the L.A.E. and least squares estimators. We believe the assumption that F(u) is in F* with F(u) possibly more-peaked and fatter-tailed than the normal may provide an adequate description of the data-generating processes in economic models. This leads to a consideration of alternatives to least squares and the L.A.E. estimator, in particular.

A necessary first step in analyzing sampling properties is a characterization of an estimator in terms of the data. For least squares this is a simple exercise since the least squares minimization problem can be readily solved, and the solution expressed in terms of Y and X. The corresponding situation for L.A.E. is complicated by the fact that there is no known closed form expression for the L.A.E. estimator in terms of Y and X. Therefore, a large part of this study is devoted to obtaining an explicit characterization of the L.A.E. estimator. This is the subject of Chapters II, III, and IV.

In Chapter II the L.A.E. estimator is defined and some properties are derived. In Chapter III an equivalent definition is presented in which the L.A.E. estimator is the solution to a certain linear programming problem. This

result, due originally to Charnes, Cooper and Ferguson (1955), brings the computational costs of L.A.E. in line with least squares and, more importantly, allows us to use linear programming methods in analyzing the L.A.E. estimator. In Chapter IV we use these methods and obtain an explicit characterization of the L.A.E. estimator in terms of Y and X. In the remaining chapters we use this characterization to derive some sampling properties.

In Chapter V we deal with the case in which the least squares estimator is the sample mean and the L.A.E. estimator is the sample median (see Property 7, Chapter III). This simple case serves as a prototype for the general comparison of the L.A.E. and least squares estimators.

Sampling properties in the general case are investigated in Chapter VI. Some properties are derived and we point to some, as yet unsolved, mathematical difficulties which are needed for a complete sampling theory for L.A.E.

Finally, a summary of findings along with some suggestions for future research are included in Chapter VII.

## Footnotes

[1]We note that the Gauss-Markov Theorem is sometimes presented as a justification for the choice of the least squares estimator and this justification does not depend on the normality assumption. However, this justification is relevant only if one restricts attention to linear estimators, and apart from mathematical or computational considerations there appears to be no reason for introducing this restriction (see the discussion in Section 2 below).

[2]Least squares is sometimes viewed as the (unconditionally) "best" estimator. In this chapter we assume that this means; first, there is a strong belief that data-generating processes are adequately described by the normal distribution, and second, there is a preference ordering under which least squares is the best estimator. This assumption motivates our discussion of alternatives to the normal distribution in Section 5.

[3]In some cases the symbol u is used to represent a random variable and in others, for example F(u), it represents a real number in the range of a random variable. It is hoped that this ambiguity is resolved by the context in which u occurs.

[4]We note that in assuming F(u) in F* our definition of the linear regression model differs slightly from the standard one in which $E(Y) = X\beta$.

[5]For a Bayesian analysis of the multivariate linear regression model, see Lindley and Smith (1972).

[6]See Theil (1971), p. 384 for the definition of regular.

[7]For the mean squared error objective function;

$$\min E(\beta-E(\tilde{\beta}))^2 = E(\tilde{\beta})^2 + \text{Var } \tilde{\beta}$$

Stein (1956) showed that for $K > 2$ least squares was not the best estimator even when F(u) is the normal distribution. Heuristically, the reason being that it is possible to trade some bias for a larger decrease in variance. To my knowledge the lexicographic preference ordering over unbiasedness and variance is the only ordering over these characteristics which leads to the choice of least squares.

[8]Box and Tiao (1964) have studied this problem in testing the equality of means by considering Bayes estimates for a parametric family of probability distributions. They essentially use a P(F) to represent prior beliefs on F(u), and then update this prior in light of sample information. To use such a scheme in a multivariate regression problem requires an examination of the frequency of estimated residuals, but these residuals depend on the estimation method used in the first place. The process to be used in updating the prior on F(u) is therefore not obvious. An alternative is to use robust estimators where "robust" is defined in terms of P(F), certain characteristics such as variance, and a new objective function; see, for example, Huber (1972). This is the approach we adopt.

[9]For a discussion of other sufficient conditions, see Dhrymes (1970, p. 103).

[10]However, if it is also assumed that all of the random variables which determine u are normal, then F(u) is normal.

[11]Also, the fact that empirical economic relationships are frequently reported with high $R^2$ and significant coefficients, but yield prediction errors which are inconsistent with these results, might be explained as departures from normality. That is, we might isolate three sources of prediction error: first, sampling error of estimates; second, neglect of future disturbance term, given F(u) normal; and third, departures of F(u) from the normal distribution. When F(u) is fatter-tailed than the normal, these sources of error imply larger prediction errors than is the case when F(u) is normal.

# CHAPTER II

# THE L.A.E. ESTIMATOR: DEFINITION AND PROPERTIES

## 1. Introduction

The L.A.E. estimator was initially proposed in the late eighteenth century prior to the discovery of least squares, and some properties were derived by Gauss and Laplace (see Plackett (1972)). L.A.E. estimation was rejected in favor of least squares because of a special property of L.A.E. to be described in the next chapter, and more importantly, because the computational costs of L.A.E. were far greater than least squares. This situation did not change until 1955 when Charnes, Cooper and Ferguson (1955) realized the L.A.E. estimate could be obtained by the linear programming techniques which were then coming into widespread use. This finding brought the computational costs of L.A.E. in line with least squares, but it did not lead immediately to the discovery of sampling properties. Since 1955 a variety of conjectures concerning the sampling properties have been obtained via extensive Monte Carlo investigations (see especially Taylor (1970), and Rosenberg and Carlson (1971)), and recently some exact sampling results have been obtained.

In this chapter the L.A.E. estimator is defined as the solution to a certain minimization problem. Some properties

which follow directly from this definition are derived so that no prior knowledge of linear programming is required. We also indicate the nature of the computational and mathematical difficulties associated with the L.A.E. estimator. The equivalence between the L.A.E. estimator as defined in this chapter and the solution to a linear programming problem is described in Chapter III.

## 2. The L.A.E. Minimization Problem

Definition

The L.A.E. minimization problem is defined by:

$$(1) \qquad \min_{b \in R^K} \sum_{t=1}^{T} \left| y_t - \sum_{k=1}^{K} x_{tk} b_k \right|$$

where the notation is the same as in Chapter I, and $|\cdot|$ is the absolute value function. Letting,

$$v_t = y_t - \sum_{k=1}^{K} x_{tk} b_k \qquad t = 1 \ldots T$$

we can rewrite (1) as:

$$\min S(b_1, \ldots, b_k, v_1, \ldots, v_T) = \sum_{t=1}^{T} |v_t| \quad .$$

Finally, in matrix notation we obtain;

Problem 1.

$$(2) \qquad \min S(b,v) = i'|v| \qquad \text{such that}$$

$$(3) \qquad Y = Xb + v$$

where $i' = (1, \ldots, 1)$ and $|v| = (|v_1|, \ldots, |v_T|)'$

The set of feasible solutions to Problem 1 is denoted by $C(Y,X)$ where;

$$C(Y,X) = \{(b,v) : Y = Xb+v\} \quad .$$

A (optimal-feasible) solution to Problem 1 is a $(\beta^*, u^*)$ in C such that;

$$S(\beta^*,u^*) \leq S(b,v) \quad \text{for all } (b,v)\varepsilon C \quad .$$

By assumption A4, X is considered fixed and so we write $(\beta^*(Y), u^*(Y))$ to denote the dependence of the solution on Y. When it is necessary to account for the dependence on X we write $(\beta^*(Y:X), u^*(Y:X))$.

A $(\beta^*,u^*)$ is the unique solution to Problem 1 if $(\beta^*, u^*)$ is in $C(Y,X)$ and;

$$S(\beta^*,u^*) < S(b,v) \quad \text{for all } (b,v)\varepsilon C;\ (b,v) \neq (\beta^*,u^*) \quad .$$

The solution to Problem 1 is not necessarily unique (see Property 1 below), and therefore $(\beta^*,u^*)$ is not necessarily an estimator. For $(\beta^*,u^*)$ to be an estimator we have to place certain restrictions on X so that there is a unique solution to Problem 1. These restrictions are derived in Chapter IV.

For the multivariate regression model we show that non-unique solutions to Problem 1 are unlikely so that the restrictions on X do not seriously limit the applicability of the L.A.E. estimation method. However, this situation is in contrast to the least squares estimator which is defined

for all X. In the remaining sections of this chapter ($\beta^*$, $u^*$) are referred to as the L.A.E. estimators for the parameters ($\beta$,u), although ($\beta^*$,$u^*$) may be more properly viewed as a (not necessarily unique) solution to Problem 1.

Geometry

In studying the properties of ($\beta^*$,$u^*$) we use the algebraic properties of the solution to Problem 1. However, the geometry of Problem 1 provides an alternative view of the L.A.E. estimator which can be compared with the geometry of the least squares estimator (see Malinvaud (1965), Chapter 5). Here we reformulate Problem 1 in terms of convex sets in the metric space $(R^T,d^*)$ where $d^*$ is the distance function;

$$d^*(Y,\bar{Y}) = i'|Y-\bar{Y}| \qquad Y,\bar{Y}\epsilon R^T \quad .$$

Some of the properties of the L.A.E. estimator which are derived in this chapter are illustrated for the case T = 2 and K = 1. The reader can then extend this to higher dimensional spaces, and also apply the geometry to results derived in later chapters.

The linear subspace spanned by the columns of X is given by:

$$L(X) = \{\bar{Y} \epsilon R^T : \bar{Y} = Xb, \quad b\epsilon R^K\} \quad .$$

Given a point (observation) $Y\epsilon R^T$, and the distance function $d^*$, we can rewrite Problem 1 as:

$$\min d^*(Y,\bar{Y}) \qquad \text{such that}$$

$$\bar{Y} \in L(X) \quad .$$

To define Problem 1 in terms of the intersection of convex sets, we define the convex set $D^*$ (centered at Y) with radius r as; (see Figure 1)

$$D^*_r = \{\tilde{Y} \in R^T : d^*(Y, \tilde{Y}) \leq r\} \quad .$$

Problem 1 can then be written;

$$\min_r (D^*_r \cap L) \qquad \text{such that}$$
$$D^*_r \cap L \neq \phi$$

where $\phi$ is the empty set.

Least squares also minimize distance, but the distance function is given by:

$$\hat{d}(Y,\bar{Y}) = (Y-\bar{Y})'(Y-\bar{Y}) \quad Y, \bar{Y} \in R^T \quad .$$

We can then define the set $\hat{D}_r$ as: (see Figure 1)

$$\hat{D}_r = \{\tilde{Y} \in R^T : d(Y,\tilde{Y}) \leq r\} \quad .$$

Therefore the difference between the L.A.E. and least squares estimators can be traced to the different characterizations of the sets $D^*$ and $\hat{D}$.

## 3. Properties of $\beta^*, u^*$

Here we derive some properties of $(\beta^*, u^*)$ which follow directly from the feasibility and optimality requirements

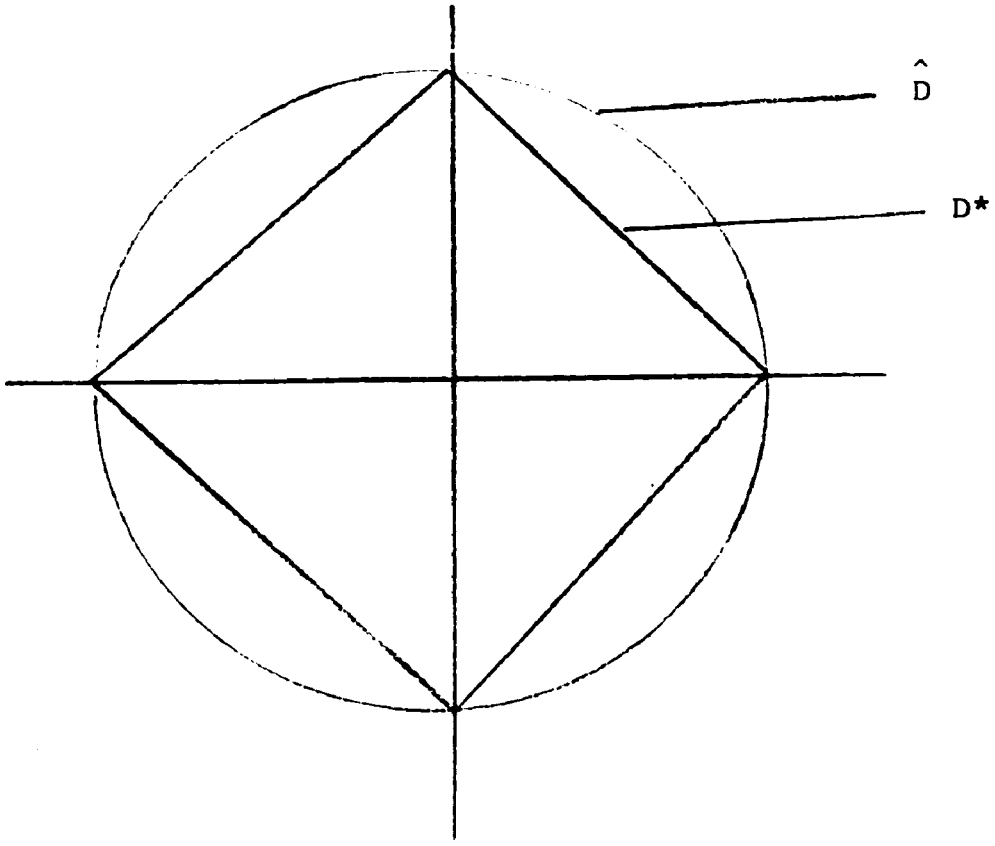



Fig. 3.--The Sets $\hat{D}$ and D*

which define the solution to Problem 1.[1]

Property 1. $(\beta^*,u^*)$ is not necessarily unique.

Proof: The simplest case in which $(\beta^*,u^*)$ is not unique is when;

$$y_1 = \beta + u_1$$

$$y_2 = \beta + u_2 \quad .$$

The solution to Problem 1 for this model is $\beta^* = (\gamma y_1 + (1-\gamma)y_2)$ and $u^* = (Y - X\beta^*)$ for any $\gamma$; $0 \leq \gamma \leq 1$. The value of S evaluated at $\beta^*,u^*$ is;

$$\begin{aligned} S(\beta^*,u^*) &= |y_1 - \gamma y_1 - (1-\gamma)y_2| + |y_2 - \gamma y_1 - (1-\gamma)y_2| \\ &= |(1-\gamma)(y_1-y_2)| + |\gamma(y_2-y_1| \\ &= |y_1-y_2| \end{aligned}$$

To show that this is the solution, suppose $y_1 > y_2$ and take $b' > y_1$. Then $(b', Y-Xb')$ is a feasible solution and;

$$S(b', v') = |y_1-b'| + |y_2-b'| > |y_2 - b'| > |y_1-y_2|$$

so that

$$S(b', v') > S(\beta^*,u^*) \quad .$$

Now let $b'' < y_2$. Then $(b'', Y-Xb'')$ is a feasible solution and;

$$S(b'',v'') = |y_1-b''| + |y_2-b''| > |y_1-b''| > |y_1-y_2|$$

hence,

$$S(b'',v'' > S(\beta^*,u^*) \quad .$$

Similar relations hold when $y_1 < y_2$. Therefore $(\beta^*,u^*)$ is the solution set for Problem 1.

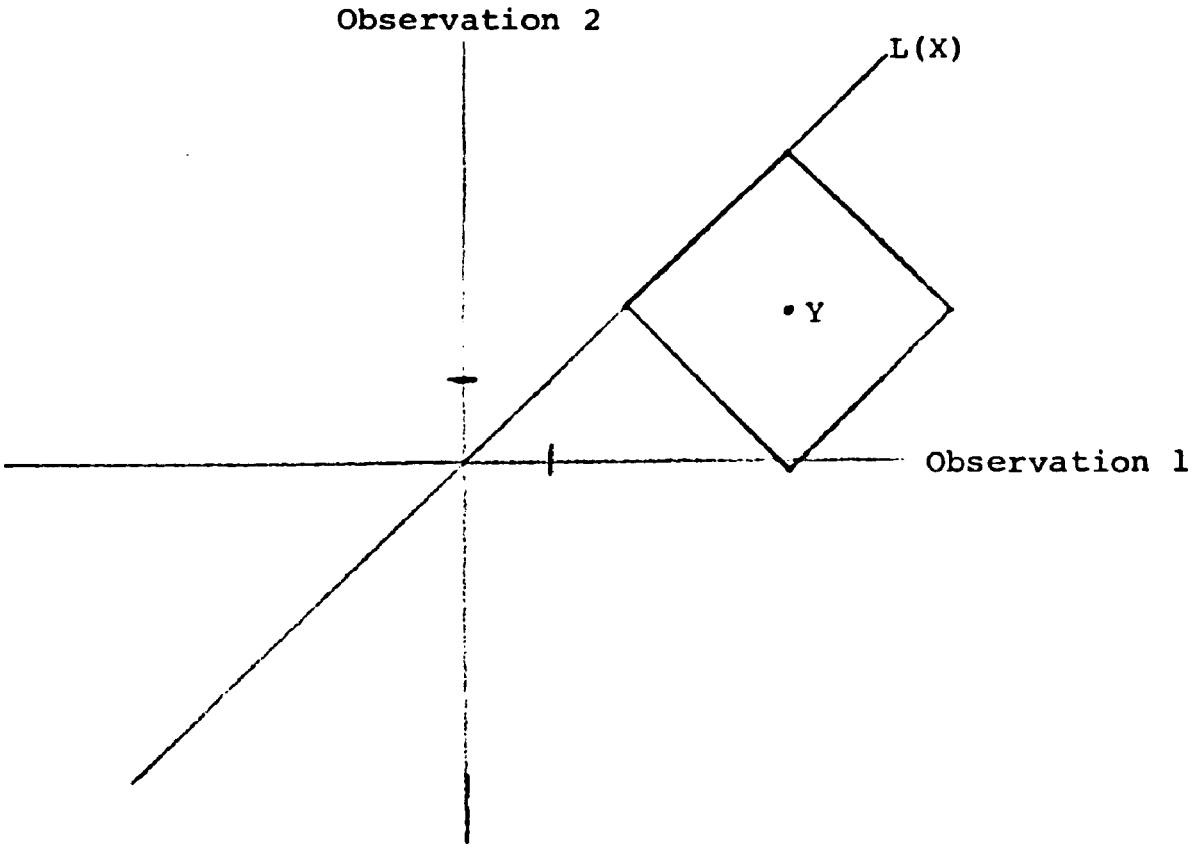



Fig. 4.--Illustration of Property 1

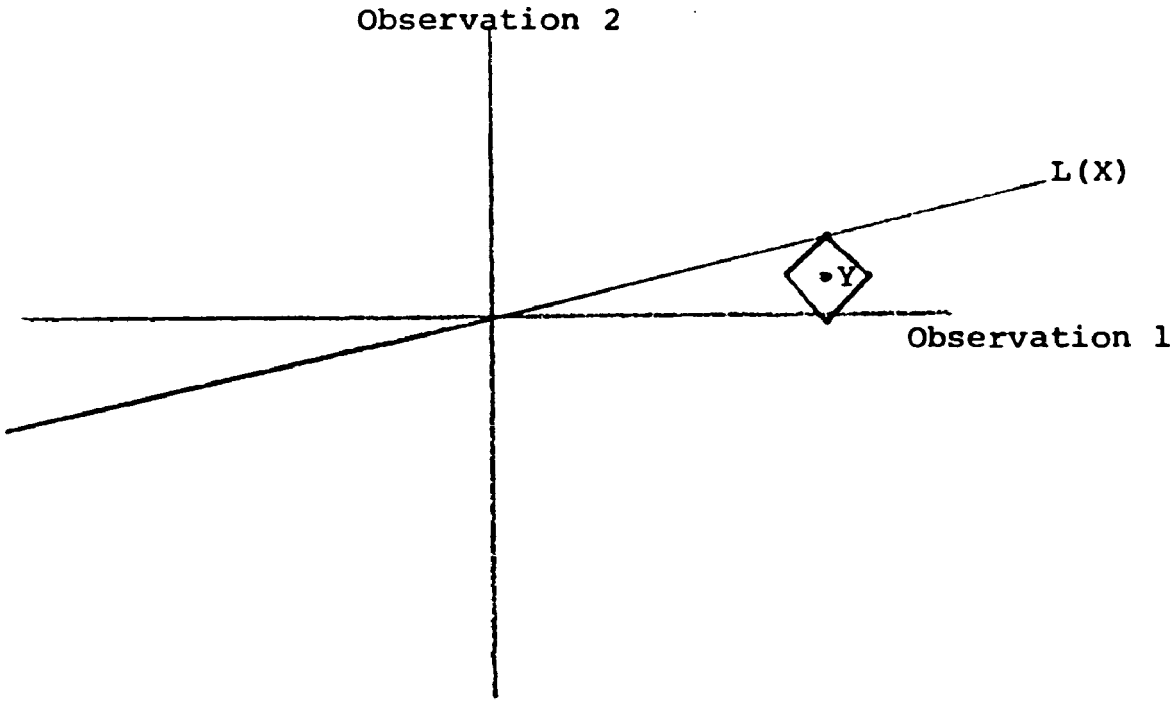



Fig. 5.--An Example of a Unique Solution

In Figure 4 the above case is illustrated, and in Figure 5 we illustrate a case in which the solution to Problem 1 is unique. In $R^2$ the restrictions on X which are needed for the solution to Problem 1 to be unique imply that L cannot be parallel to an edge of D* (see Chapter IV), and this is easy to see in Figures 4 and 5.

Property 2. $\lambda(\beta^*(Y),u^*(Y)) = (\beta^*(\lambda Y),u^*(\lambda Y))$, where $\lambda$ is real.

Proof: From the definition of C it follows that;

$$C(\lambda Y,X) = \{\lambda(b,v) : \lambda Y = X\lambda b + \lambda v,\ (b,v)\epsilon C(Y,X)\}$$
$$= \lambda C(Y,X)$$

and in particular;

$$\lambda(\beta^*(Y),u^*(Y))\epsilon C(\lambda Y,X) \quad .$$

This feasible solution is also optimal since,

$$S(\lambda\beta^*(Y),\lambda u^*(Y)) = i'|\lambda u^*| \leq i'|\lambda v| = S(\lambda b,\lambda v)$$

for $(\lambda b,\lambda v)\epsilon C(\lambda Y,X)$. This shows that for the data $(\lambda Y,X)$ a solution to Problem 1 is $\lambda(\beta^*(Y),u^*(Y))$ where $(\beta^*(Y),u^*(Y))$ is a solution given $(Y,X)$. To show the converse let $(\beta^*(\lambda Y), u^*(\lambda Y))$ be a solution given the data $(\lambda Y,X)$. Since the first part of the proof holds for all $\lambda$, take $\lambda^{-1}$. Then for $(Y,X)$ the solution is $\lambda^{-1}(\beta^*(\lambda Y),u^*(\lambda Y))$. Therefore,

$$\lambda(\beta^*(Y),u^*(Y)) = (\beta^*(\lambda Y),u^*(\lambda Y))$$

and this proves Property 2.

Property 2 holds for all linear estimators, and in particular, least squares. Property 2 is illustrated in Figure 6.

Property 3. $(\beta^*(u), u^*(u)) = (\beta^*(Y) - \beta, u^*(Y))$.

Proof. First, express the set of feasible solutions in terms of (u,X) using assumption A1. We get;

$$C(u,X) = \{(b-\beta,v) : (b,v) \in C(Y,X)\}$$

and, $(\beta^*(Y) - \beta, u^*(Y)) \in C(u,X)$. Also,

$$S(\beta^*(Y) - \beta, u^*(Y)) = i'|u^*(Y)| \leq i'|v| = S(b-\beta,v)$$

for $(b-\beta,v) \in C(u,X)$. Therefore, $(\beta^*(Y) - \beta, u^*(Y))$ is a solution given (u,X). To show the converse, suppose $(\beta^*(u), u^*(u))$ is a solution given (u,X). Then $(\beta^*(u) + \beta, u^*(u))$ is a feasible solution given (Y,X) and,

$$S(\beta^*(u) + \beta, u^*(u)) = i'|u^*(u)| \leq i'|v| = S(b,v)$$

for $(b,v) \in C(Y,X)$. Hence, $\beta^*(Y) - \beta = \beta^*(u)$.

This property does not hold for every linear estimator, but it does hold for least squares.[2] The property is illustrated in Figure 7.

Property 4. $A^{-1}\beta^*(Y:X) = \beta^*(Y:XA)$, where A is a K x K matrix and rank A = K.

Proof: If (b,v) is in C(Y,X) then $Y = Xb+v$ and $Y = XA(A^{-1}b)+v$ so that;

$$C(Y,XA) = \{(A^{-1}b,v) : (b,v) \in C(Y,X)\}$$

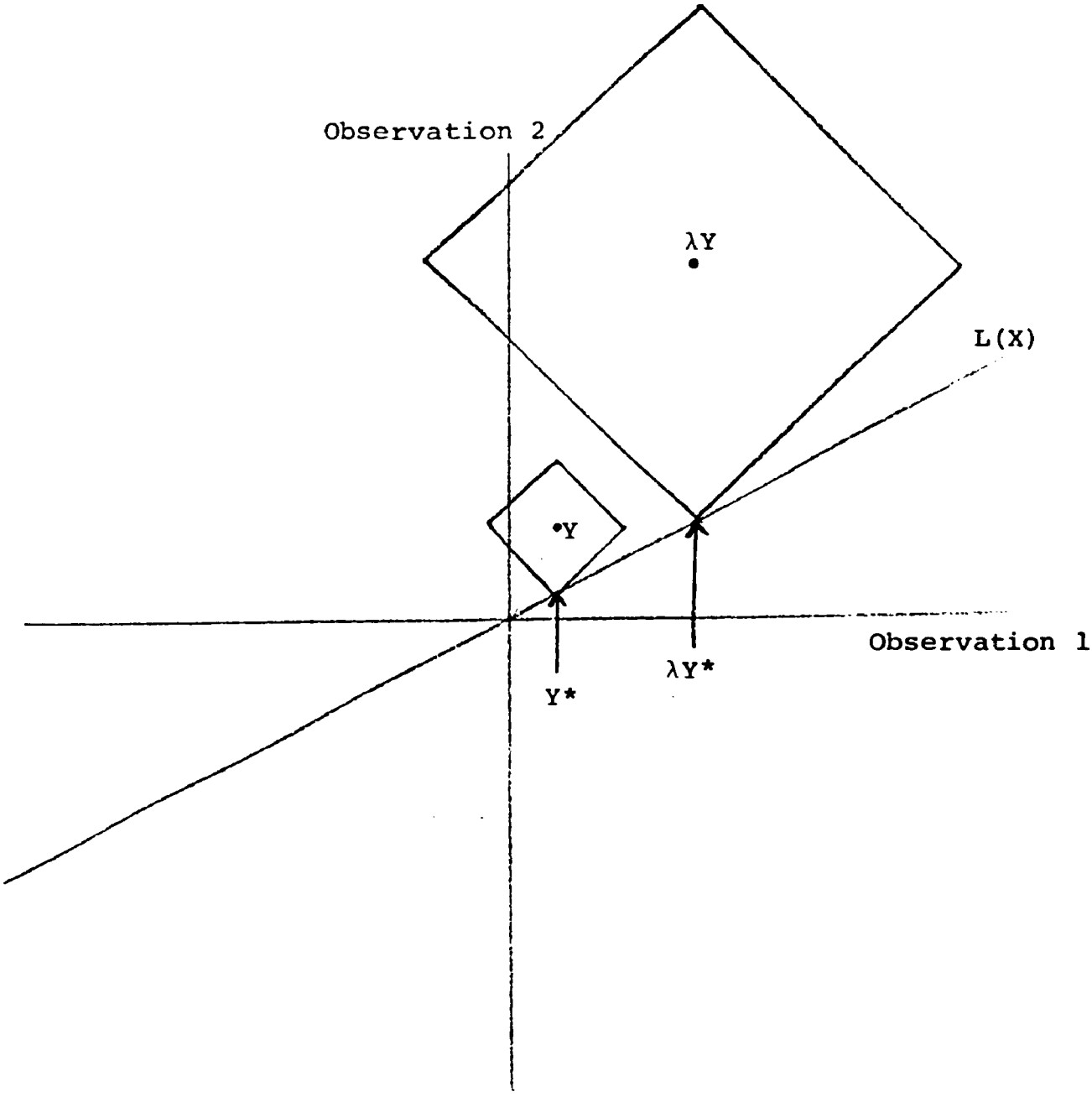



Fig. 6.--Illustration of Property 2

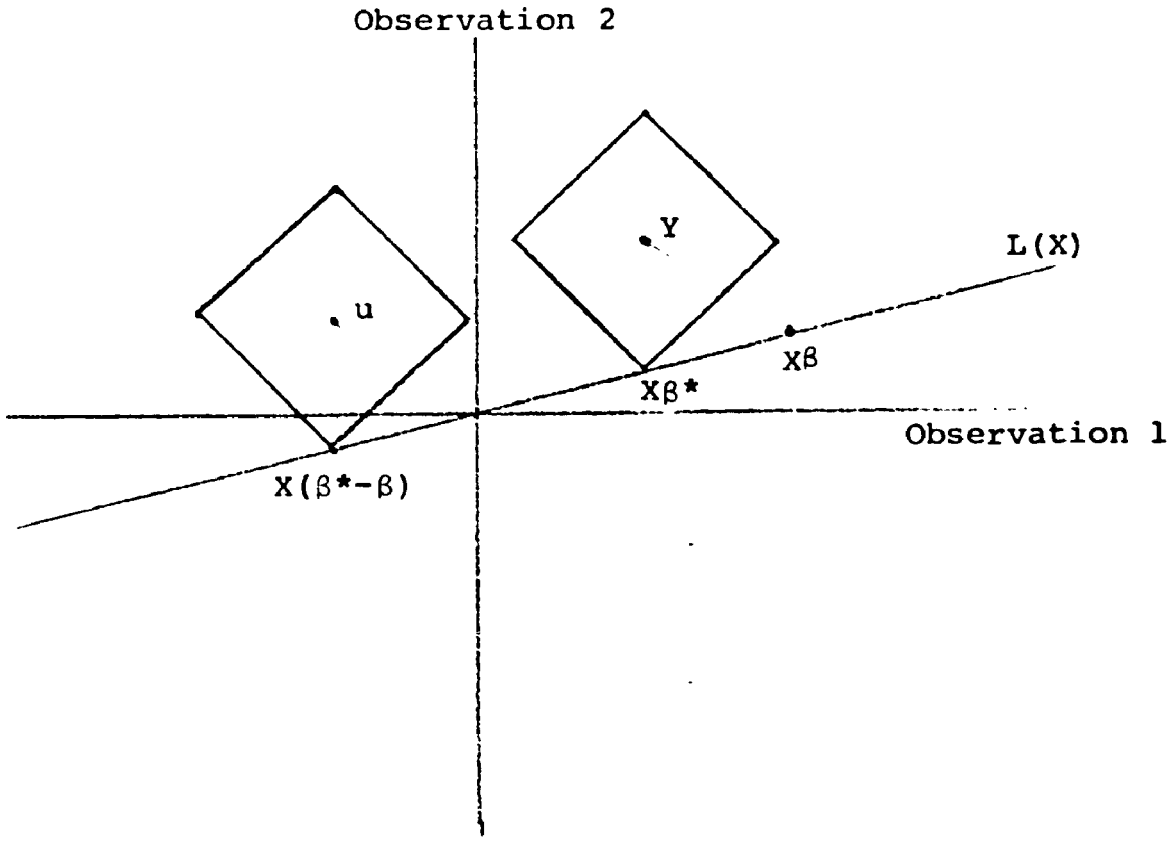



Fig. 7.--Illustration of Property 3

In particular, $(A^{-1}\beta^*(Y:X), u^*(Y:X))$ is in $C(Y, XA)$. This feasible solution is also optimal since;

$$S(A^{-1}\beta^*(Y:X), u^*(Y:X)) = i'|u^*(Y)| \leq i'|v| = S(A^{-1}b,v)$$

where $(A^{-1}b,v)$ is in $C(Y,XA)$. This shows $(A^{-1}\beta^*(Y:X), u^*(Y))$ is a solution given the data $(Y,XA)$. To show the converse suppose $(\beta^*(Y:XA),u^*(Y:XA))$is a solution given the data $(Y,XA)$. Then$(A\beta^*(Y:XA), u^*(Y:XA))$ is a feasible solution given the data $(Y,X)$ and by the first part of the proof, it is also an optimal solution. Hence,

$$\beta^*(Y:XA) = A^{-1}\beta^*(Y:X) \quad .$$

This property also holds for the least squares estimator,[3] and it implies that if the explanatory variables are multiplied by some factor (that is, A diagonal) then the L.A.E. estimates are multiplied by the inverse of this factor. Also, if A is chosen so that $AA' = (X'X)^{-1}$ then $(XA)'(XA) = I$, and the transformed explanatory variables are orthogonal. We can therefore use A to transform X and thereby restrict attention to orthogonal explanatory variables. However, this device is not as useful for the L.A.E. estimator as it is for least squares since we show later that the estimators $(\beta_1^*,\ldots,\beta_K^*)$ are correlated even when X is orthogonal.

Properties 2, 3 and 4 are all shared by the least squares estimator. To the extent that these properties are considered desirable for least squares they must also be considered desirable for the L.A.E. estimator. We now show one of the ways in which the L.A.E. and least squares estimators differ.

It is well-known that the least squares estimator is a linear function of Y for any X. That is, $\hat{\beta}(Y^1+Y^2) = \hat{\beta}(Y^1) + \hat{\beta}(Y^2)$. The corresponding situation for the L.A.E. estimator is more complicated as is evidenced by the next property.

Property 5. $\beta^*(Y:X)$ is a nonlinear function of Y for some X, and is linear for some other X.

Proof: We first present two examples. In the first $\beta^*$ is nonlinear and, in the second it is linear. We then take the general case and show the difficulties which arise in attempting to show nonlinearity.

For the first example take $X = (1,1,1)'$. In the next chapter we show that for this case $\beta^*$ is the median observation on Y. The nonlinearity of the median is easily demonstrated by the following example.

$$\begin{aligned} Y^1 &= (4,5,8) \quad ; \quad \beta^*(Y^1) = 5 \\ Y^2 &= (6,8,7) \quad ; \quad \beta^*(Y^2) = 7 \\ Y^1 + Y^2 &= (10,13,15) \quad ; \quad \beta^*(Y^1 + Y^2) = 13 \neq 12 \quad . \end{aligned}$$

For the second example take $X = (3,1,1)'$. The value of S at $\beta^* = y_1/3$ is;

$$|y_1 - 3\beta^*| + |y_2 - \beta^*| + |y_3 - \beta^*| = |u_2^*| + |u_3^*|$$

To show that this is the L.A.E. estimator, consider the value of S at $\beta^* + \Delta$ where $\Delta \neq 0$. We get

$$|y_1 - 3\beta^* - 3\Delta| + |y_2 - \beta^* - \Delta| + |y_3 - \beta^* - \Delta| =$$

$$|- 3\Delta| + |u_2^* - \Delta| + |u_3^* - \Delta| \geq |u_2^*| + |u_3^*| + \Delta$$

$$> |u_2^*| + |u_3^*| \quad .$$

Therefore, $\beta^* = (\frac{1}{3},0,0) \cdot (y_1,y_2,y_3)' = y_1/3$ and $\beta^*$ is a linear function of Y.

In the general case we have by definition;

$$C(Y^1 + Y^2, X) = \{(b,v) : Y^1 + Y^2 = Xb+v\} \quad .$$

But

$$Y^1 = Xb^1+v^1 \quad \text{for } (b^1,v^1) \in C(Y^1,X)$$
$$Y^2 = Xb^2+v^2 \quad \text{for } (b^2,v^2) \in C(Y^2,X)$$

and

$$Xb+v = X(b^1+b^2) + (v^1+v^2)$$

so that

$$C(Y^1+Y^2,X) = C(Y^1,X) + C(Y^2,X) \quad .$$

Hence, $(\beta^*(Y^1) + \beta^*(Y^2), u^*(Y^1) + u^*(Y^2))$ is in $C(Y^1+Y^2,X)$. To see whether this feasible solution is optimal we calculate the value of S which is;

$$i'|u^*(Y^1) + u^*(Y^2)| \leq i'|u^*(Y^1)| + i'|u^*(Y^2)|$$

$$\leq i'|v^1| + i'|v^2| \quad .$$

But

$$i'|v^1| + i'|v^2| \geq i'|v^1 + v^2| = S(b,v) \quad .$$

Therefore $\beta^*$ is not necessarily a linear function. But the inequalities are not strict and so the possibility that $\beta^*$

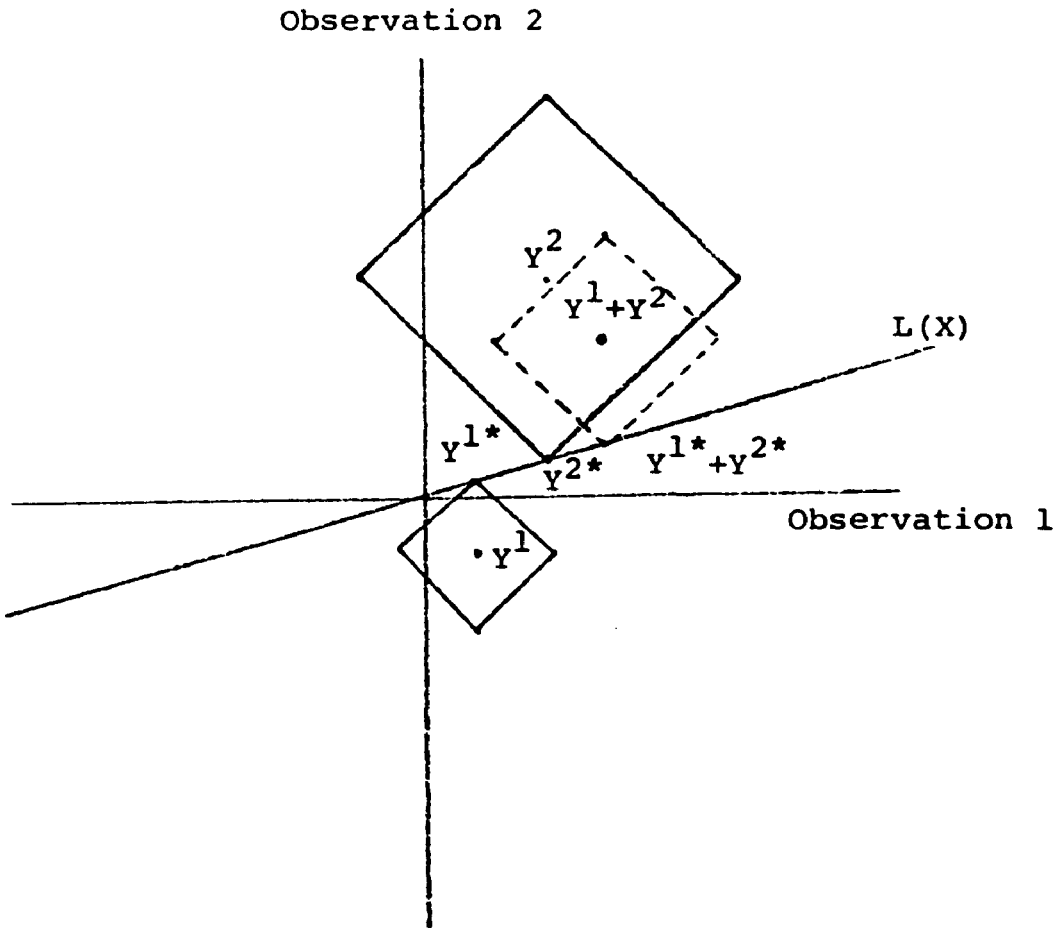



Fig. 8.--Illustration of Property 5

is linear cannot be ruled out. The fact that $\beta^*$ can be linear is demonstrated by the above example.

For our purposes it is especially important to isolate the cases in which $\beta^*(Y)$ is a linear function. If $\beta^*(Y)$ is linear then according to the Gauss-Markov Theorem, it has at least as large a variance as the least squares estimator for any $F(u)$ in $F^*$.[4] In Chapter IV we derive the conditions under which $\beta^*(Y)$ is linear. We show that in almost all cases $\beta^*(Y)$ is nonlinear so that the Gauss-Markov Theorem cannot be used to exclude the L.A.E. estimator.

In Figure 8 we illustrate a case in $\beta^*(Y)$ is linear. By trying other examples the reader can easily discover that in this case the linearity of $\beta^*(Y)$ does not depend on L, $y_1$, or $y_2$ so that for T = 2 and K= 1, $\beta^*$ is always a linear estimator. This result is generalized in Chapter IV.

## 4. Computational and Mathematical Problems

The computational and mathematical difficulties associated with the analysis of the L.A.E. estimator arise because the absolute value function is not continuously differentiable and therefore the solution to Problem 1 cannot be obtained by the standard Lagrangian saddle point method. To calculate the L.A.E. estimate one essentially has to search by some trial and error method over the infinite number of elements in C(Y,X) for the $(\beta^*,u^*)$ which minimizes S. Until recently the computational costs associated with this

search ruled out the L.A.E. estimator as a practical alternative to least squares.

Two related events changed this situation. In 1955, Charnes, Cooper and Ferguson showed that the L.A.E. estimator was the solution to a certain linear programming problem. It is well-known that any linear program can be solved in finite time; the reason being that it is only necessary to examine a finite number of points in the set of feasible solutions. This along with the development of high-speed and low-cost computers which could be used to rapidly search for feasible and optimal solutions has brought the computational costs of the L.A.E. estimator close to the costs of the least squares estimator.[5]

For analytical purposes the nondifferentiability of the absolute value function implies there is no obvious way of obtaining a closed form expression for $\beta^*$ in terms of Y and X. Without such an expression it is unclear how F(u) is transformed under the function $\beta^*$. This is in sharp contrast to the linear least squares estimator whose properties can be obtained by using the theorems on linear transformations of random variables.

Closed form expressions for $\beta^*$, however, are not necessary in order to obtain the sampling properties of L.A.E. Properties 1 to 4 indicate some results on L.A.E. which can be obtained without such an expression, and some are readily transformed into sampling results. For example, Properties 2 and 3 along with F(u) symmetric implies $(\beta^*-\beta)$ is

symmetrically distributed about zero, and $\beta^*$ is therefore unbiased (if the expectation exists). Also, the formulation of the L.A.E. estimator as the solution to a linear programming problem allows us to apply the methods associated with these programs in analyzing the sampling properties of the L.A.E. estimator. The development of these linear programming methods as they apply to the L.A.E. estimator is pursued in the remaining parts of this study. In Chapter IV we obtain conditions which characterize $\beta^*$ explicitly in terms of Y and X and this provides the first stage in the development of a sampling theory for the L.A.E. estimator.

## Footnotes

[1]Properties 2, 3 and 4 are equivalent to Theorems 1 and 3 in Rosenberg and Carlson (1972). It is likely that these properties along with properties derived in Chapter III were known by Gauss and Laplace in the early nineteenth century. See the discussion in Chapter III, Section 3.

[2]Any linear estimator can be written $\bar{\beta} = AY$, where A is a K x T matrix. In terms of the model we have, $\bar{\beta} = AX\beta + Au$ and $\bar{\beta} - \beta = (AX-I)\beta + Au$. Therefore, Property 3 holds for all linear estimators such that $(AX-I) = 0$. It is well-known that this is satisfied by the least squares estimator since, $(X'X)^{-1}X'X-I = 0$.

[3]The least squares estimator given the data Y, XA is;

$$\begin{aligned}\hat{\beta}(Y:XA) &= ((XA)'(XA))^{-1}(XA)'Y \\ &= A^{-1}(X'X)^{-1}A'^{-1}A'X'Y \\ &= A^{-1}(X'X)^{-1}X'Y \\ &= A^{-1}\hat{\beta}(Y:X) \quad .\end{aligned}$$

Hence Property 4 is shared by the least squares estimator.

[4]The Gauss-Markov Theorem holds for F(u) in $\hat{F}$. However, the only F(u) which are in F* and not in F are those with infinite variance, and here the variance of any linear estimator is infinite. This extension of the Gauss-Markov Theorem implies that if $\beta^*$ is linear then $Var(\beta^*) \geq Var(\beta)$ for F(u) in F*.

[5]Rosenberg and Carlson (1972, p. 3) report that for a seven-variable regression with eighty observations computation time for L.A.E. is 2.866 seconds as compared to .348 seconds for least squares. It is probable that this difference can be reduced by taking advantage of the particular form of the L.A.E. linear programming problem. For example, see Property 7 in Chapter III and the accompanying footnote.

# CHAPTER III

## THE L.A.E. LINEAR PROGRAMMING PROBLEM

In this chapter we demonstrate the result due originally to Charnes, Cooper, and Ferguson (1955) which shows that the solution to Problem 1 can be obtained by solving a linear programming problem. This result greatly alleviates the computational difficulties associated with L.A.E., and brings the computational cost of L.A.E. in line with least squares. More importantly, the equivalence between Problem 1 and a linear programming problem allows us to use the analysis associated with linear programs in our study of L.A.E.

This chapter is divided into five sections. In Section 1 the L.A.E. linear programming problem is defined, and we demonstrate its equivalence with Problem 1. Section 2 includes a statement of the dual L.A.E. linear programming problem, and two linear programming theorems. In Section 3 some properties of L.A.E. are derived by using the L.A.E. linear programming problem. Section 4 presents a summary of the Monte Carlo findings and conjectures which have been obtained by Taylor (1970), and Rosenberg and Carlson (1972). The conclusions are included in Section 5.

## 1. A Linear Programming Problem

Definition

A linear programming problem consists of a linear objective function which is minimized (or maximized) subject to linear equality and/or inequality constraints. Problem 1 does not satisfy these requirements since the objective function, $i'|v|$, is not a linear function of v. It is possible, however, to create a linear programming problem whose solution is equivalent to the solution to Problem 1. (The precise meaning of equivalent is made clear in Theorem 1 below.) The importance of this result is that the computational and mathematical analysis associated with linear programming can be used in studying the L.A.E. estimator. Here we indicate the reasons for the equivalence between these two apparently different definitions of the L.A.E. estimator.

The L.A.E. linear programming problem is defined by:

Problem 2

$$\min \bar{S}(b,r^+,r^-) = i'r^+ + i'r^- \quad \text{such that} \tag{1}$$

$$Y = Xb + r^+ - r^- \tag{2}$$

$$b \in R^K;\ r_t^+ \geq 0;\ r_t^- \geq 0 \quad t = 1,\ldots,T \quad . \tag{3}$$

Rewriting (1) and (2) we obtain:

$$\min \bar{S}(b,r^+,r^-) = (0',i',i') \cdot (b,r^+,r^-)' \tag{4}$$

$$Y = [X\ I\ -\ I] \cdot (b,r^+,r^-)' \quad . \tag{5}$$

This is a linear programming problem in (K+2T) variables, and T equality constraints.[1]

To show equivalence between the solutions to Problem 1 and Problem 2 we first prove the following lemma.

Lemma 1. If $(\beta^*, r^{+*}, r^{-*})$ is a solution to Problem 2, then $\min \{r_t^{+*}, r_t^{-*}\} = 0$ for $t = 1,\ldots,T$.

Proof. Suppose $(\beta^*, r^{+*}, r^{-*})$ is a solution to Problem 2, and for some t, say $t = 1$; $r_1^{+*} > 0$ and $r_1^{-*} > 0$. Let $\min \{r_1^{+*}, r_1^{-*}\} = a$, and define,

$$(r_1^{+*} - a,\ r_1^{-*} - a) \quad . \tag{6}$$

(6) still satisfies the first constraint,

$$y_1 = \sum_{k=1}^{K} x_{1k}\beta^*_k + r_1^{+*} - r_1^{-*}$$

$$= \sum_{k=1}^{K} x_{1k}\beta^*_k + (r_1^{+*} - a) - r_1^{-*} - a) \quad .$$

The objective function with (6) is:

$$(r_1^{+*} - a) + (r_1^{-*} - a) + \sum_{t=2}^{T} r_t^{+*} + \sum_{t=2}^{T} r_t^{-*} = \bar{S}(\beta^*, r^{+*}, r^{-*}) - 2a < \bar{S} \ .$$

So $(\beta^*, r^{+*}, r^{-*})$ is a solution only if $a = 0$. The same method can now be applied to any particular t, or any subset of $t = 1,\ldots,T$.

Lemma 1 shows that the constraint:

$$\min \{r_r^{+}, r_t^{-}\} = 0 \quad t = 1,\ldots,T \tag{7}$$

can be added to Problem 2 without altering the solution. The equivalence between Problems 1 and 2 is now easily demonstrated.

Theorem 1. $(\beta^*, r^{+*}, r^{-*})$ is a solution to Problem 2 if and only if $(\beta^*, u^*) = (\beta^*, r^{+*} - r^{-*})$ is a solution to Problem 1.

Proof. Let $(b, r^+, r^-)$ be a feasible solution to Problem 2; that is, it satisfies (2), (3) and (7). Then,

$$i'r^+ + i'r^- = \Sigma|r^+ - r^-| \quad .$$

Now let $(\beta^*, r^{+*}, r^{-*})$ be a (feasible and optimal) solution to Problem 2. Clearly $(\beta^*, r^{+*} - r^{-*})$ is a feasible solution for Problem 1, and

$$\bar{S}(\beta^*, r^{+*}, r^{-*}) = \Sigma|r^{+*} - r^{-*}| \leq \Sigma \mid r^+ - r^-|$$

Hence $(\beta^*, r^{+*} - r^*)$ is also an optimal solution to Problem 1.

To show the converse let $(b, v)$ be a feasible solution to Problem 1. Let

$$(8) \qquad \begin{aligned} r_t^+ &= v_t \quad r_t^- = 0 \quad \text{if} \quad v_t \geq 0 \\ r_t^+ &= 0 \quad r_t^- = -v_t \quad \text{if} \quad v_t < 0 \end{aligned}$$

so that $(b, r^+, r^-)$ satisfies the constraints (2), (3) and (7) in Problem 2. Suppose $(\beta^*, u^*)$ is a solution to Problem 1. Using the construction in (8) we obtain a feasible solution to Problem 2, and,

$$\begin{aligned} S(\beta^*, u^*) = i'|u^*| = i'|r^{+*} - r^{-*}| = i'r^{+*} + i'r^{-*} &\leq i'|v| \\ &= i'r^+ + i'r^- \end{aligned}$$

and therefore this feasible solution is also an optimal solution to Problem 2.

Comments

Theorem 1 shows how the L.A.E. estimate can be obtained from the solution to Problem 2. This, along with high-speed low-cost computers and the simplex algorithm, solves the computational problems associated with L.A.E. At least as an approximation method, L.A.E. becomes a practical alternative to least squares.

As an estimation method, the equivalence between Problems 1 and 2 does not immediately make L.A.E. a practical alternative to the least squares estimator. We have to characterize the L.A.E. estimator in terms of the data, (Y,X), and then see how F(u) is transformed by the function $\beta^*$. The equivalence between L.A.E. and the solution to a linear programming problem, however, allows us to use the analysis associated with linear programs to characterize $\beta^*$. This analysis is used repeatedly in the rest of this study, and it leads eventually to the sampling properties of the L.A.E. estimator.

A practical consequence of Theorem 1 is that linear inequality and/or equality constraints can be imposed on the L.A.E. estimates without significantly changing computational costs, or the analysis of sampling properties. To introduce the constraints we merely add them on to Problem 2 so that the constraints to Problem 2 become

$$\begin{bmatrix} Y \\ Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} X & I & -I \\ A_1 & A_2 & A_3 \\ A_4 & A_5 & A_6 \end{bmatrix} \begin{matrix} b \\ r^+ \\ r^- \end{matrix}$$

where $Z_i$ and $A_i$ express the constraints on the L.A.E. estimates, and vector inequalities are read componentwise. Alternatively one may want to weight the observations differently, and this is accomplished by changing the objective function to:

$$\min \bar{S} = (0, C_1, C_2)(b, r^+, r^-)'$$

where $(C_1, C_2)$ are the weights. In either of these cases we have a linear programming whose solution can be obtained in the same fashion as the solution to Problem 2.

The ease with which constraints can be introduced into Problem 2 permits the inclusion of prior knowledge (which is believed with probability one) directly into the estimation method. For example, in estimating a quadratic production function the decreasing marginal productivity assumption can be introduced directly into Problem 2 in the form of linear inequality constraints (see Ashar and Wallace (1962)). Another production function example arises when one wants to estimate the most-efficient unit isoquant, and not just the "average" isoquant for an industry. This problem is easily handled by imposing constraints on the residual $u_t^*$.[2] The constraints might also be used to handle the common problem of multicollinearity by introducing new information in the form of linear constraints.

The ease with which linearly constrained estimates are obtained for L.A.E. is in contrast to least squares where inequality constraints on $\hat{\beta}$ or $\hat{u}$ require the solution of a quadratic programming problem subject to linear inequality

constraints. A discussion of the computational and mathematical difficulties associated with the constrained least squares estimator can be found in Malinvaud (1965, Chapter 9).

In this study we concentrate on the unconstrained L.A.E. estimator. However, since we study the L.A.E. estimator as the solution to a stochastic linear programming problem, the properties of the linearly constrained estimator can be obtained by merely reanalyzing Problem 2 with constraints. In the constrained and unconstrained case the method of analysis is the same. Therefore, for linearly constrained estimation problems, the L.A.E. estimator solves the computational, and some of the mathematical difficulties, which are associated with the constrained least squares estimator.

## 2. The Dual L.A.E. Problem

In some cases the properties of the solution to Problem 2 are more easily obtained by using the well-known duality correspondences associated with linear programs. In this section the dual L.A.E. problem and two linear programming theorems are stated, and these are used to further characterize the L.A.E. estimate in the next section. For a discussion of duality in linear programming, see Spivey and Thrall (1970), and for the general case, see Rockafeller (1970).[3]

The dual L.A.E. problem is obtained by applying the dual transformation to Problem 2; we get,

$$\max Y'd \qquad \text{such that}$$

$$Xd = 0$$

$$\begin{bmatrix} I \\ -I \end{bmatrix} d \leq 1$$

$$d \in R^T \quad .$$

Simplifying the inequality constraints yields:

$$\max Y'd \qquad \text{such that}$$

$$Y'd = 0$$

$$-1 \leq d_t \leq 1 \quad t = 1,\ldots,T \quad .$$

The dual L.A.E. problem is a linear maximization problem in T bounded variables, and K linear homogeneous constraints.

Geometrically, the constraint X'd = 0 implies that d is contained in the subspace orthogonal to the subspace spanned by the columns of X, L(X). This orthogonal subspace is denoted $L(X)^{\perp}$ (see Figure 9). Let d* be a solution to the dual L.A.E. problem, and let Y* = Xβ*. It then follows that for the random variables d* and Y*,

$$d^* \cdot Y^* = 0 \quad \text{for all } Y \text{ in } R^T$$

and, in particular,

$$E(d^* \cdot Y^*) = 0 \quad .$$

This is analogous to the least squares result in which $\hat{Y}$, $\hat{u}$

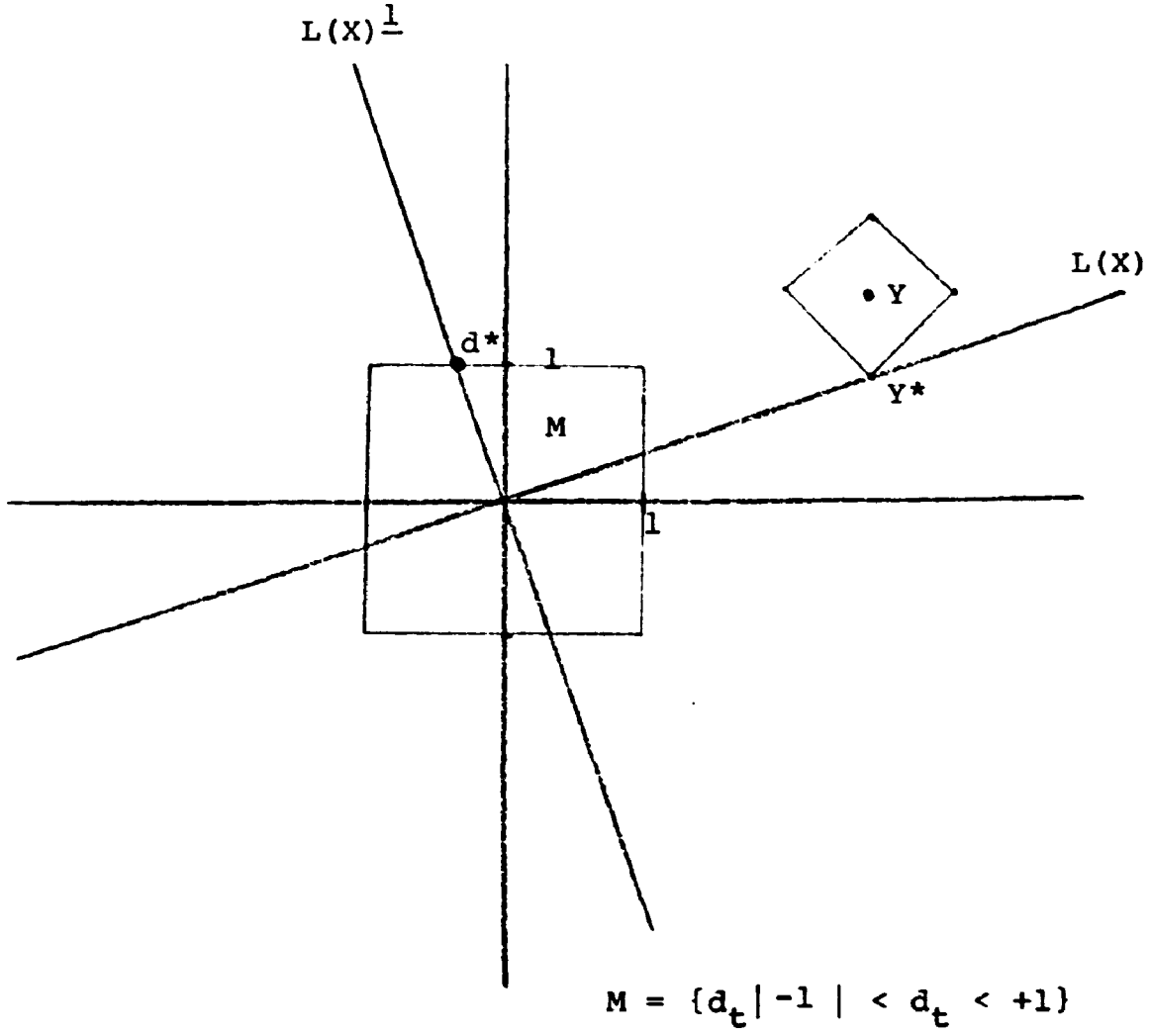



Fig. 9.--Geometry of the Dual L.A.E. Problem

are uncorrelated, and it raises the question of whether the dual variables, instead of the estimated L.A.E. residuals, might be used along with $\beta^*$ to obtain a statistic for testing hypotheses when the dispersion of F(u) is unknown. The answer to this question awaits further study.

The dual L.A.E. problem can be interpreted by first expressing it in terms of u. For the objective function we have;

$$(X\beta+u)'d = u'd + \beta X'd \quad .$$

Therefore, the dual L.A.E. problem in terms of the unobservable error term can be written;

$$\max u'd \qquad \text{such that}$$

$$d \varepsilon L(X)^{\perp}$$

$$-1 \leq d_t \leq 1 \qquad t = 1,\ldots,T \quad .$$

If the constraint $d \varepsilon L(X)^{\perp}$ is, for the moment, ignored then $u'd = \Sigma u_t d_t$ is maximized subject to, $-1 \leq d_t \leq 1$ by setting,

$$d_t^* = +1 \qquad \text{if } u_t > 0$$

$$d_t^* = -1 \qquad \text{if } u_t < 0 \quad .$$

Here the dual variables indicate the sign of $u_t$. In the dual L.A.E. problem the $d_t^*$ continue to be indicators, but subject to the additional, $d^* \varepsilon L(X)^{\perp}$ constraint. Heuristically, if $d_t^* = 1$ the sample information indicates $u_t > 0$, and we should then let $u_t^* > 0$; and similarly for $d_t^* = -1$. That this interpretation is valid is shown by;

Theorem 2. (complementary - slackness). If $(\beta^*,u^*)$ is the unique solution to Problem 1 (or Problem 2 via the transformation in Theorem 1) and $d^*$ is the solution to the dual L.A.E. problem, then;

$$u_t^* > 0 \quad \text{iff} \quad d_t^* = 1$$
$$u_t^* < 0 \quad \text{iff} \quad d_t^* = -1$$
$$u_t^* = 0 \quad \text{iff} \quad |d_t^*| < 1 \quad .$$

Proof: See Appendix I.

Finally, we state the following well-known linear programming theorem using the notation in Problem2.

Theorem 3. If the (T X (K+2T)) matrix [X,I,-I] has rank T, then there is a solution to Problem 2 with at least (K+T) of the variables in $(\beta^*,r^{+*},r^{-*})$ equal to zero with the nonzero variables being linearly independent.[4]

Proof: See Spivey and Thrall (1970), Chapter 3.

This theorem is completely general, and a further discussion can be found in the next chapter. The theorem implies the well-known result that for a linear technology with T resources and (K+2T) activities the economy will use at most T activities. Also, note that the lemma to Theorem 1 is a special case of this theorem since we showed the linear dependent variables, $r_t^{+*}$, $r_t^{-*}$, could not both be positive.

## 3. Properties of the L.A.E. Linear Programming Problem

Inthis section we obtain further properties of the L.A.E. estimator which follow from the L.A.E. linear programming problem. Unlike Properties 2, 3 and 4, none of the properties in this section are shared by the least squares estimator.

Property 5. If $(\beta^*,u^*)$ is a solution to Problem 1, then at least K of the $u_t^* = 0$.

Proof: This follows directly from the equivalence between Problems 1 and 2 along with Theorem 3. According to Theorem 3, at least K+T of the variables in $(\beta^*, r^{+*}, r^{-*})$ are zero and at most T are nonzero in a solution to Problem 2. Of these T nonzero variables, K will correspond to $\beta^*$ since the corresponding "cost" in the objective function is zero. (Here we use assumption A3, rank X=K, so that the K columns of X are linearly independent.) Therefore, at most (T-K) of the variables in $(r^{+*}, r^{-*})$ are nonzero. The equivalence between Problems 1 and 2 then implies at most (T-K) of the $u_t^* \neq 0$ and at least K of $u_t^* = 0$.[5]

According to Property 5, the L.A.E. estimator may be viewed in the following manner. Let $h_i$ be a set of K integers taken from the set, $\{1,2,\ldots,T\}$. Let H be the set of all $h_i$. There are $\binom{T}{K} = \frac{T!}{K!(T-K)!}$ ways of choosing K integers in this way, so,

$$H = \{h_1,h_2,\ldots,h_z; \quad z = \binom{T}{K}\} \quad .$$

We can now identify with each $h_i$ the subset of $R^T$ such that for the K observations corresponding to $h_i$, $u_t^* = 0$. That is, define the random variable H* whose range is H, and Pr[H* = $h_i$] is the probability that the K observations corresponding to $h_i$ have $u_t^* = 0$. We can now view the L.A.E. estimator in two stages. At the first stage we find $h_i \varepsilon H$, and at the second stage we solve for $\beta^*$ by setting $u_t^* = 0$ for the K observations corresponding to $h_i$. The importance of this view arises because we can derive the sampling properties of $\beta^*$ by solving for the necessary and sufficient conditions such that $H = h_i$. This is done in the following chapter.

Property 5 shows that in a sense the L.A.E. estimator finds some set of K observations, and then discards the remaining (T-K) observations in order to solve for $\beta^*$. This characteristic is perhaps the most distinguishing feature of the L.A.E. estimator, and it has led some to exclude L.A.E. from further consideration. For example, Wonnacott and Wonnacott (1970) after observing that the L.A.E. estimated regression hyperplane always contains K observations, write;

> But [L.A.E.] is not a good commonsense solution to the problem (of curve-fitting) since it pays no attention whatever to [excluded observations]. (p. 5.)

It therefore seems that some have rejected L.A.E. because it "ignores" some of the observations in the sample.

The reason why Property 5 cannot be used to exclude L.A.E. from consideration is that, except when L.A.E. is a linear estimator (see Chapter IV), the L.A.E. estimator does not ignore any observations. The function (estimator) $\beta^*(Y)$

is a function of all the observations on Y, even though for any given Y, the estimate $\beta^*$ is determined by only K of the observations. The crucial point is that the set of K observations which determines the estimate $\beta^*$ is itself a random variable determined by all of the observations on Y.

Another property of the L.A.E. estimator which follows from the linear programming problem is that we can interpret K of the constraints in (5) as "binding" or "tight" ($u_t^* = 0$) and (T-K) constraints as being slack. For linear programs it is well-known (see Spivey and Thrall (1970), Chapter 3) that any increase in a resource which is not already being fully utilized (slack) will lead to no change in technique. We now show how this applies to the particular L.A.E. linear program.

Suppose we have the L.A.E. estimates, $(\beta^*(Y), u^*(Y))$, given the data (Y,X). We construct a new observation Y' by setting;

$$Y' = X\beta^*(Y) + Z\, u^*(Y)$$

where Z is a TxT diagonal matrix all of whose diagonal elements are greater than or equal to zero. In terms of the individual observations we have;

$$\begin{aligned} y_t' &> \Sigma x_{tk}\beta^*_k \quad \text{if} \quad u_t^* > 0 \\ y_t' &= \Sigma x_{tk}\beta^*_k \quad \text{if} \quad u_t^* = 0 \\ y_t' &< \Sigma x_{tk}\beta^*_k \quad \text{if} \quad u_t^* < 0 \quad . \end{aligned} \tag{9}$$

This is illustrated in Figure 10. For these new observations we obtain;

Property 6. $\beta^*(Y) = \beta^*(Y')$ where $Y' = X\beta^*(Y) + Z\, u^*(Y)$.

Proof: See Appendix I.

This property shows that the L.A.E. estimate does not change in response to changes in $y_t$ so long as (9) is satisfied. This property is not shared by least squares. Also note that this property is consistent with the interpretation given in Section 2 where the dual variables indicate the sign of $u_t$.[6]

Finally, it is well-known that for the model;

$$y_t = \beta + u_t \qquad t = 1,\ldots,T \tag{10}$$

the least squares estimator is the mean observation on $y_t$. It also is well-known that the L.A.E. estimator is the median observation on $y_t$. We show how this follows directly from Theorem 2 and also the way in which it extends to the multivariate case.[7]

Property 7'. Given (10) and T odd the L.A.E. estimator is the median observ ation on $y_t$.

Proof: Let P denote the number and the set of $u_t^* > 0$; and N the number and set of $u_t^* < 0$. Since K = 1, $u_t^* = 0$ for one of the observations, and for notational ease suppose $u_1^* = 0$. From the constraints to the dual problem we get;

$$i'd^* = \sum_{t=1}^{T} d_t^* = d_1^* + \sum_{t\in P} d_t^* + \sum_{t\in N} d_t^* = 0 \quad .$$

From Theorem 2 (complementary-slackness) we obtain,

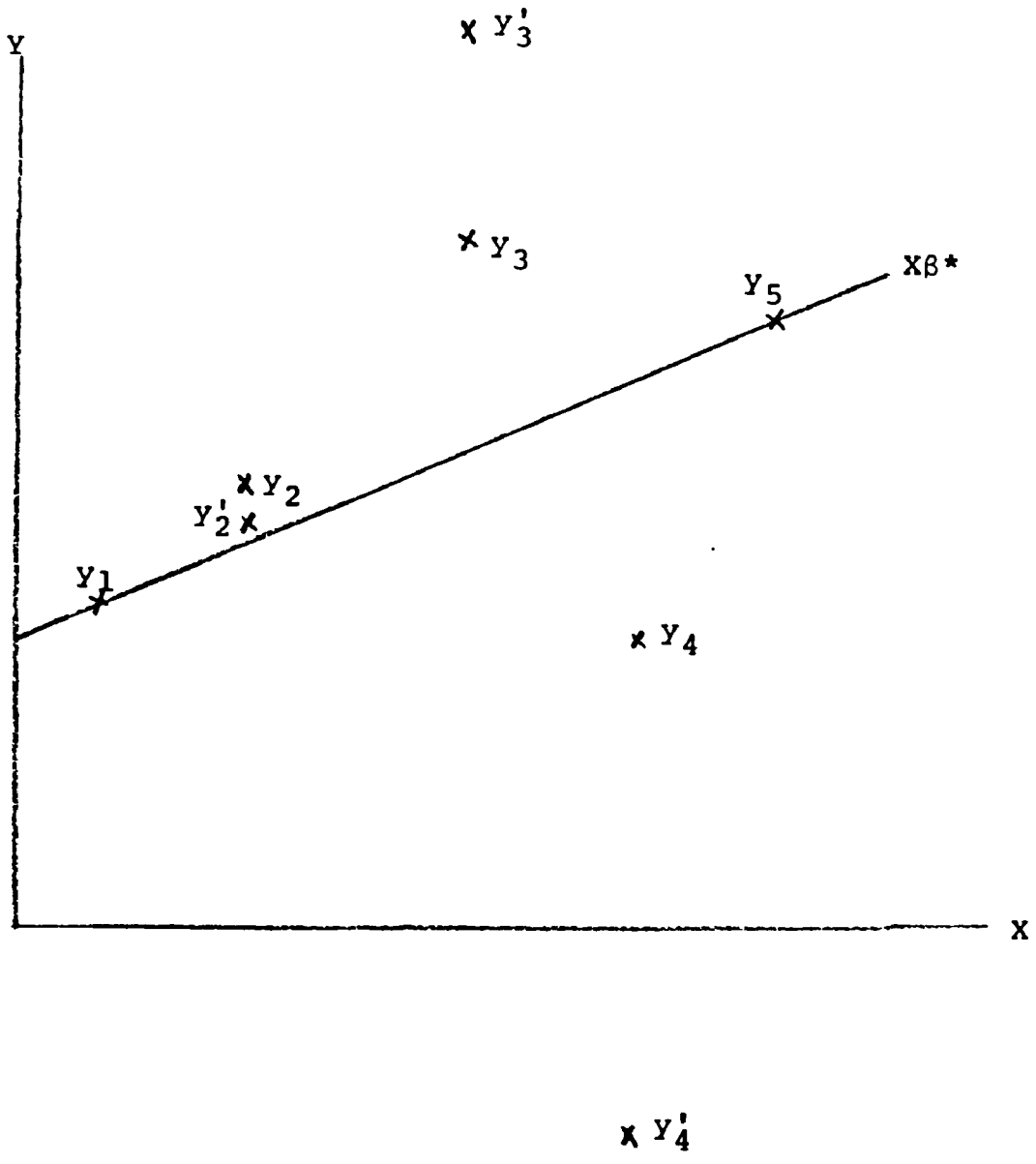



Fig. 10.--Illustration of Property 6

$$d_1^* = |N - P| \quad ,$$

$$-1 < d_1^* < 1 \quad ,$$

and this implies, N = P. Therefore, $\beta^*$ is chosen so that the number of positive and negative $u_t^*$ are equal and such that one of the $u_t^* = 0$. But this implies that $\beta^*$ is the median $y_t$.

The reason for assuming T odd is that when T is even, the solution to Problem 1 (or Problem 2) is not unique, and therefore the L.A.E. estimator (or equivalently, the median $y_t$) is not well defined. The case T = 2 has already been illustrated in the demonstration of Property 1.

In the multivariate case the linear regression model with a constant term is written;

$$Y = (i,Z)\ \beta + u \tag{11}$$

where Z is a (K-1)xT matrix of nonconstant explanatory variables (and $\beta$ is partitioned, $(\beta_1 | \beta_2,\ldots,\beta_k)$). We then obtain the following generalization of Property 7'.

Property 7. Given the model (11), $\beta^*$ is chosen so that $K > |P-N|$ where P and N are the number of positive and negative estimated errors, respectively.

Proof: Since $u_t^* = 0$ for K of the observations, reorder the observations so that $u_t^* = 0$; $t = 1,\ldots,K$ (this again, is for notational ease). The dual constraints are:

$$i'd^* = 0 \tag{12}$$

$$Z'd^* = 0 \quad . \tag{13}$$

Rewriting (12) we get:

$$\sum_{t=1}^{T} d_t^* = \sum_{t=1}^{K} d_t^* + \sum_{t\epsilon P} d_t^* + \sum_{t\epsilon N} d_t^* = 0 \quad .$$

Using Theorem 2 we obtain:

$$\left| \sum_{t=1}^{K} d_t^* \right| = |P-N|$$

$$-1 < d_t^* < 1 \qquad t = 1,\ldots,K \quad .$$

Hence,

$$K > |\Sigma d_t^*| = |P-N| \quad .$$

It is easily seen that for K=1, Property 7 is equivalent to Property 7'.

Combining Properties 5 and 7 we see that when X includes a constant $\beta^*$ is chosen so that K of the $u_t^* = 0$ with almost, in the sense of Property 7, the same number of positive and negative $u_t^*$.

## 4. Monte Carlo Results

In this section we summarize the Monte Carlo results obtained by Taylor (1970), and Rosenberg and Carlson (1972) (for some other results see Blattberg and Sargant (1970), and Smith and Hall (1972)). These particular experiments are reviewed since the properties derived in latter chapters explain some of the Monte Carlo findings, and our approach to L.A.E. has been motivated by these findings.

Taylor (1970)

In his experiments Taylor investigated the properties of the random variable H*. The model was:

$$y_t = 2x_{1t} + 1x_{2t} + u_t \quad t = 1,\ldots,5$$

so that H contained $\binom{5}{2} = 10$ elements. The problem was to discover how Y was mapped to $h_i \varepsilon H$ via the function H*. Taylor investigated this problem under alternative assumptions on F(u). A description of the various Monte Carlo experiments is included in Table 1. The motivation for these experiments was the belief that the sampling distribution of $\beta^*$ might be found by first examining the sampling distribution of H*.

Taylor's main finding was that certain pairs of observations never determined the L.A.E. hyperplane. This result was completely unexpected. In our notation there were some $h_i$ such that, $Pr[H^* = h_i] = 0$. Taylor conjectured that this was not a sampling result, but instead, there was something in the solution to Problem 1 which excluded certain pairs of observations from ever determining the L.A.E. estimates.

The attempt to discover whether Taylor's conjecture was true provided the initial motivation for our study of the L.A.E. estimator. (The prior belief was that the conjecture was almost certainly false.) In Chapter IV we derive the necessary and sufficient conditions such that $H^* = h_i$, and these conditions show that Taylor's conjecture is valid.

More importantly, the conditions enable us to show explicitly how $\beta^*$ depends on Y and X, and, therefore, allows a representation of the sampling distribution of $\beta^*$.

In summary, Taylor's findings were:

1. For some $h_i$; $Pr[H^* = h_i] = 0$.
2. $Pr[H^*]$ is independent of $F(u)$, and is determined solely by X.
3. $E(\beta^*) = \beta$ if $F(u)$ is symmetric.
4. When $F(u)$ is the bilateral Pareto distribution then the distribution of $\beta^*$ is symmetric, but is not Pareto.

We will see that only conjecture 2 is incorrect.

Rosenberg and Carlson (1972)

In seeking an easily computed approximation for the L.A.E. sampling distribution, Rosenberg and Carlson conjectured: (1) the variance-covariance matrix of $\beta^*$ is proportional to $(X'X)^{-1}$; (2) the scaling factor for the variance-covariance matrix is $\lambda(F(u),T)$, where $\lambda/T$ is the variance of the median for samples of size T, and the distribution $F(u)$; and (3) the distribution of $\beta^*$ is approximately normal. The distribution of $\beta^*$ was therefore conjectured to be:

$$C_{L.A.E.} : N(\beta, \lambda(F(u),T)(X'X)^{-1})$$

where $N(e,v)$ is a normal distribution with expectation e, and variance v. Except for the scale factor this is the asymptotic distribution of the least squares estimator (given the well-known conditions on X and $F(u)$; see Dhrymes (1970),

Chapter 4). The least squares approximation is,

$$C_{L.S.} : N(\beta, \sigma^2(F(u),T)(X'X)^{-1})$$

where $\sigma^2/T$ is the variance of the mean for some size T and distribution F(u) (hence, $\sigma^2$ is the variance of F(u)).

As Rosenberg and Carlson point out:

> This approximation has the great virtue of computational simplicity, since tables of (F(u),T) can be computed.... Moreover, since the approximate distribution differs from the normal approximation to the least squares sampling theory only in the scale factor, all of the familiar hypothesis tests and inference procedures of least squares can be applied as an approximation to L.A.E., with only this scale factor changed.

The validity of the conjecture would also imply that L.A.E. has smaller variance than least squares for those F(u) such that the median has smaller variance than the mean. Given the preferences discussed in Chapter I, one could therefore easily translate beliefs concerning F(u) into a choice between L.A.E. and least squares.

To test their approximation, Rosenberg and Carlson tried a number of different experiments, computed L.A.E. and least squares estimates, and then compared the critical regions, using $C_{L.A.E.}$ and $C_{L.S.}$, with the known values of the parameters. The Kolmogorov statistic was used to test the distribution as a whole. The experiments were conducted for alternative T, K, X, and F(u). The various experiments are summarized in Table 2. Note that for each experiment X was fixed, but that the nonconstant explanatory variables were

TABLE 1

TAYLOR'S MONTE CARLO EXPERIMENTS

| | | | |
|---|---|---|---|
| Explanatory variables: | X = | 6 | 1 |
| | | 2 | 7 |
| | | 3 | 4 |
| | | 7 | 5 |
| | | 2 | 3 |

F(u): Uniform in (-1,1)
Triangular in (-1,1)
Laplace
Two-tailed Pareto

Number of replications per experiment: 100

TABLE 2

ROSENBERG AND CARLSON MONTE CARLO EXPERIMENTS

Explanatory variables (fixed):

(a) K = 1
1. constant
2. normal
3. widely-dispersed with high kurtosis

(b) K = 3
1. constant; normal; normal

F(u):

(a) normal (kurtosis = 3)
(b) contaminated normal (kurtosis = 11.8)
(c) contaminated normal (kurtosis = 13.4)

Sample Size:

(a) T = 31
(b) T = 59

Estimator:

(a) L.A.E.
(b) Least Squares

Number of replications per experiment: 1000

alternatively drawn from a normal and a widely dispersed, high kurtosis distribution.

For $C_{L.S.}$ the sampling results were consistent with the conjectured variance, although when F(u) had high kurtosis the null hypothesis that $\hat{\beta}$ was normally distributed was rejected 19 out of 32 times at the 99 percent level. In contrast, the null hypothesis that $\beta^*$ was normally distributed was never rejected at the 99 percent level. However, when the kurtosis of F(u) was high, the variance in $C_{L.A.E.}$ underestimated the variance of $\beta^*$, and this underestimate increased as the kurtosis of X (not F(u)) increased and decreased as T increased. In all other cases $C_{L.A.E.}$ and $C_{L.S.}$ are roughly consistent with the sampling results. Rosenberg and Carlson concluded that the $C_{L.A.E.}$ is an acceptable approximation, but that when the dispersion and kurtosis of X is high the variance estimate for $\beta^*$ must be revised upward.

The Monte Carlo results obtained by Rosenberg and Carlson, and Taylor raise two questions. First, why does $Pr[H^* = h_i] = 0$ for some $h_i$; and second, why does the dispersion or kurtosis of X affect the variance estimate in $C_{L.A.E.}$? We will see that the answer to either of these apparently different questions provides the answer to the other. This is accomplished in the next chapter where we derive the conditions such that $H^* = h_i$.

## 5. Conclusion

In this chapter the L.A.E. estimator was defined as the solution to a linear programming problem, and this was shown to be equivalent with the definition of the L.A.E. estimator given in the previous chapter. Further properties of L.A.E. were derived using the linear programming problem. We have also described some important findings obtained via Monte Carlo methods. The importance of the L.A.E. linear programming problem is illustrated in the next chapter where we derive conditions which yield an explicit characterization of the L.A.E. estimator.

## Footnotes

[1]Problem 2 can be put in standard form where all the variables are subject to nonnegatively constraints by defining;

$$b = b^+ - b^-$$

$$b^+ \geq 0 \; ; \; b^- \geq 0 \quad .$$

(See Ashar and Wallace (1964). The standard form is the starting point for many of the algorithms used in solving linear programs. For our purposes we do not need the standard form since the "cost" of b is zero in (1).

[2]See Aigner and Chu (1968).

[3]The dual L.A.E. problem was first analyzed by Wagner (1959) and later by Appa and Smith (1971).

[4]Note that rank [X,I,-I] = T even when rank X < K.

[5]In this study we assume f(u) is continuous so that the probability of (K+1) or more of the $y_t$ lying in the same hyperplane is zero. With probability one, K of the $u_t^* = 0$, and we ignore cases in which more than K of the $u_t^* = 0$.

[6]Properties 5 and 6 follow directly from the L.A.E. linear programming problem, and one might therefore think that until recently they were unknown. That this was not the case is evidenced by the following quote from Gauss, taken from Plackett (1972):

> Laplace made use of another principle for the solution of linear equations, the number of which is greater than the number of unknown quantities, which had been previously proposed by Boscovich, namely that the differences themselves, but all of them taken positively, should make up as small a sum as possible. It can be easily shown, that a system of values of unknown quantities, derived from this principle alone, must necessarily exactly satisfy as many equations out of the number proposed, as there are unknown quantities, so that the remaining equations come under consideration only so far as they help to determine the choice: if, therefore, the equation V = M, for example, is of the number of those which are not satisfied, the system of values found according to this principle would in no respect be changed,

> even if any other value N had been observed instead of M, provided that, denoting the computed value by n, the differences M - n, N - n, were affected by the same signs.

The remarkable fact about this quote is that it is a complete statement of Properties 5 and 6 made in the absence of any knowledge of linear programming.

[7]To my knowledge, Property 7 and 7' were first demonstrated by Appa and Smith (1971).

# CHAPTER IV

## SOLUTION TO THE L.A.E. LINEAR PROGRAMMING PROBLEM

In this chapter we solve the L.A.E. linear programming problem for the conditions such that any set of K observations determines the L.A.E. estimate. These conditions show explicitly the dependence of the L.A.E. estimator on Y and X, and they are used in Chapter VI to derive the sampling properties of the L.A.E. estimator.

The chapter is divided into seven sections. Section 1 is an introduction to linear programming methods and it is presented to illustrate the methods used in Section 2 where we solve the L.A.E. linear programming problem. In Section 3 we obtain the restrictions on X which are required for the solution of the L.A.E. linear programming problem to be an estimator. In Section 4 the L.A.E. estimator is characterized as a linear function over certain subsets of $R^T$. This characterization is used in Chapter VI to analyze sampling properties. In Section 5 we prove a mathematical theorem relating to the characterization of the L.A.E. estimator. In Section 6 we isolate the special cases in which the L.A.E. estimator is a linear function. The conclusions are presented in Section 7.

## 1. Linear Programming

In this section we outline a method, the simplex algorithm, for solving linear programming problems. With a slight modification, the simplex algorithm yields the conditions which we use as a starting point in developing a sampling theory for the L.A.E. estimator. The outline illustrates the approach we use to solve the L.A.E. linear programming problem, and it is not intended as a complete review of linear programming. The outline follows the presentation in Spivey and Thrall (1970).

In this section we analyze the following linear programming problem;[1]

Problem 1

$$\min S(Z) = \gamma + CZ \tag{1}$$

$$Y = AZ \tag{2}$$

$$Z \geq 0 \tag{3}$$

$\gamma$ is a constant; C is a given (1 x M) vector; A is a given (T x M) matrix with $M \geq T$, rank A = T; Y is a given (T x 1) vector; and Z is a (M x 1) variable vector. The solution to this problem is a $Z = Z^*$ satisfying (2) and (3) (feasible) such that S is a minimum (optimal).[2] We are interested in characterizing $Z^*$, and the conditions such that $Z^*$ is a unique solution.

The linear equality constraints in (2) are a system of T equations in M variables. Since rank A = T the system can

be solved for T of the $z_m$ in terms of the remaining (M - T) $z_m$. Partitioning (2) gives:

$$Y = (A_1, A_2) \cdot (Z_1, Z_2) \tag{4}$$

where $A_1$ is (T x T), rank $A_1$ = T, $A_2$ is (T x (M-T)), $Z_1$ = $(z_1,\ldots,z_T)$, and $Z_2$ = $(z_{T+1},\ldots,z_M)$. Premultiplying (4) by $A_1^{-1}$ yields:

$$A_1^{-1}Y = (I, A_1^{-1}A_2) \cdot (Z_1, Z_2) \tag{5}$$

so that $(Z_1, Z_2) = (A_1^{-1}Y, 0)$ satisfies (2). This is called the basic solution corresponding to $Z_1$, or equivalently, the basic solution corresponding to the T columns in $A_1$. (The T columns form a basis.) The nonbasic variables (columns) are the $z_m$ in $Z_2$ (the (M-T) columns in $A_2$). Any basic solution can be found by choosing T linearly independent columns from A, and then solving for the corresponding basic variables by setting the nonbasic variables equal to zero.

Note that the set of basic solutions is finite since there are at most $\binom{M}{T}$ ways of choosing T linearly independent columns from A.

The importance of the set of basic solutions is given by;

Theorem 1. There is a basic solution which is a feasible and optimal solution to Problem 1. This is called a feasible-optimal basic solution.[3]

Corollary 1. If the solution to Problem 1 is unique then it corresponds to a basic solution.

Corollary 2. The set of feasible-optimal solutions to Problem 1 is a convex combination of the feasible-optimal basic solutions.

The proofs of the theorem and corollaries can be found in Spivey and Thrall (1970, Chap. 3).

The theorem and corollaries show that a solution to Problem 1 can be found by listing the (finite) set of basic solutions in terms of increasing S, and then finding the first element in the list which is feasible. Alternatively, a solution can be found by searching the (finite) set of feasible basic solutions for the element such that S is a minimum. In fact it is never necessary to examine every basic solution since necessary and sufficient conditions exist which indicate when a basic solution is feasible and optimal. These conditions along with some efficient search method provides an algorithm for solving any linear program. The simplex, dual-simplex, and revised simplex algorithms are merely different ways of proceding through the list of basic solutions. For our purposes we are interested in the conditions such that a basic solution is feasible and optimal.

In order to find the conditions such that $(A_1^{-1}Y, 0)$ is feasible and optimal the usual approach is to express (1) solely in terms of $Z_2$. Using (5) we obtain;

Problem 2

$$S = \gamma + C_1A_1^{-1}Y + (C_2-C_1A_1^{-1}A_2)Z_2$$

$$A_1^{-1}Y = Z_1 + A_1^{-1}A_2Z_2$$

$$(Z_1,\ Z_2) \geq 0 \quad .$$

In the terminology of linear programming the transformation from Problem 1 to Problem 2 represents a simplex iteration. The transformation is important because feasibility and optimality can be easily checked in Problem 2.

The above transformation can be more easily represented by matrix multiplication. Problem 1 can be written;

$$\begin{bmatrix} S - \gamma \\ Y \end{bmatrix} = \begin{bmatrix} C_1 & C_2 \\ A_1 & A_2 \end{bmatrix} (Z_1,\ Z_2)$$

$$Z \geq 0$$

Premultiplying both sides by $Q^{-1}$ where

$$Q = \begin{bmatrix} 1 & C_1 \\ 0 & A_1 \end{bmatrix}$$

and using partitioned inverse rules;

$$Q^{-1} = \begin{bmatrix} 1 & -C_1A_1^{-1} \\ 0 & A_1^{-1} \end{bmatrix}$$

gives,

$$\begin{bmatrix} S - \gamma - C_1A_1^{-1}Y \\ A_1^{-1}Y \end{bmatrix} = \begin{bmatrix} 0 & C_2 - C_1A_1^{-1}A_2 \\ I & A_1^{-1}A_2 \end{bmatrix} (Z_1, Z_2)$$

$$Z \geq 0 \quad .$$

Feasibility follows directly from the nonnegativity constraint on Z. We get:

Feasibility Condition. $(A_1^{-1}Y,0)$ is a feasible solution if $A_1^{-1}Y \geq 0$.

The optimality condition is obtained by asking whether the objective function in Problem 2 can be decreased by making some element in $Z_2$ positive. This is clearly not possible if $(C_2 - C_1A_1^{-1}A_2) \geq 0$. Hence:

Optimality Condition. $(A_1^{-1}Y, 0)$ is an optimal basic solution if $(C_2 - C_1A_1^{-1}A_2) \geq 0$.

Combining these two conditions yields:

Theorem 2. If the feasibility and optimality conditions are satisfied then $(A_1^{-1}Y,0)$ is a solution to Problem 1.

For the complete set of conditions for unique optimality the reader is referred to Spivey and Thrall (1970). Here we state a sufficient condition which follows directly from Problem 2.

Unique Optimality Condition. If $(C_2 - C_1A_1^{-1}A_2) > 0$, then $(A_1^{-1}Y,0)$ is a unique optimal basic solution.

This condition is slightly stronger than the optimality condition since it requires every element in $(C_2 - C_1A_1^{-1}A_2)$ to be greater than zero. This condition is only sufficient for uniqueness since it is possible for some element in $(C_2 - C_1A_1^{-1}A_2)$ to be zero with a unique solution. This occurs when some of the elements in $Z_1$ and $Z_2$ are linearly dependent.

## 2. The Solution to the Linear Programming Problem

In this section the above linear programming method is used to solve the L.A.E. linear programming for the conditions such that any set of K observations determine the L.A.E. estimate. We need a substantial amount of notation, and unfortunately this tends to obscure some of the results. We have therefore included in Appendix II a parallel presentation using the linear regression model which Taylor (1970) used in his Monte Carlo experiments.

### Notation

$\bar{T} = (1,\ldots,T)$.

$h_i$ is a subset of $\bar{T}$ containing K elements (integers).

$\bar{h}_i$ is the complement of $h_i$; that is, the (T - K) integers not in $h_i$.

H is the set of all $h_i$. This set contains $\binom{T}{K}$ elements; $H = \{h_1,\ldots,h_Z;\ Z = \binom{T}{K}\}$.

Given $Y = (y_1,\ldots,y_T)$, $Y(h_i)$ is the K dimensional vector which contains $y_t$ for $t \in h_i$.

$X(h_i)$ is the (K x K) matrix containing the K rows of X corresponding to $h_i$.

The function (random variable) H* maps Y in $R^T$ to $h_i$ in H where:

$$H^*(Y) = h_i \quad \text{if} \quad \beta^* = X(h_i)^{-1}Y(h_i)$$

where $\beta^*$ is the L.A.E. estimate. The function $H^*$ identifies the K observations which determine the L.A.E. estimate.

$\bar{I}$ is the set of all $(T - K \times T - K)$ diagonal matrices with +1 or -1 along the diagonal. There are $2^{(T-K)}$ elements in $\bar{I}$, and the typical element is denoted $I^*$.

The Solution

To solve for the conditions such that $H^* = h_i$ we reorder the observations so that;

$$Y = \begin{bmatrix} Y(h_i) \\ Y(\bar{h}_i) \end{bmatrix} \qquad X = \begin{bmatrix} X(h_i) \\ X(\bar{h}_i) \end{bmatrix} .$$

The L.A.E. linear programming problem (see Chapter III) can now be written in the following partitioned form:

$$(6) \qquad \begin{bmatrix} S \\ Y(h_i) \\ Y(\bar{h}_i) \end{bmatrix} = \begin{bmatrix} 0 & i_K & i_{T-K} & i_K & i_{T-K} \\ X(h_i) & I_K & 0 & -I_K & 0 \\ X(\bar{h}_i) & 0 & I_{T-K} & 0 & I_{T-K} \end{bmatrix} \phi$$

$$\phi = (b,\ r^+(h_i),\ r^+(\bar{h}_i),\ r^-(h_i),\ r^-(\bar{h}_i))$$

$$b \varepsilon R^K,\ r^+ \geq 0,\ r^- \geq 0$$

where I is an identity matrix, $i = (1,\ldots,1)$, and subscripts denote the dimension of the vector or matrix. We are interested in the conditions such that $\beta^* = X(h_i)^{-1}Y(h_i)$. (We assume rank $X(h_i) = K$.)[4] The first K estimated residuals therefore are zero, and in (6) this implies we are interested in the conditions such that $(r^{+*}(h_i),\ r^{-*}(h_i)) = 0$.

Note that (6) differs from Problem 1 in the preceding section in that b is unconstrained. This causes no difficulties since the "cost" of b in the objective function is zero.

In order to find the conditions such that $H^*(Y) = h_i$ we have to choose T linearly independent columns (a basis) from the constraint matrix in (6). However, we cannot choose any of these columns from:

$$\begin{bmatrix} i_K & & i_K \\ I_K & , & -I_K \\ 0 & & 0 \end{bmatrix}$$

since we want the corresponding variables $(r^+(h_i), r^-(h_i))$ to be nonbasic, and equal to zero.

Of the remaining (K + 2T) columns in the constraint matrix every optimal-feasible basic solution will contain the K columns of X since the "cost" of b is zero (by assumption, the K columns of X are linearly independent). The remaining (T - K) columns for a basic solution must be linearly independent and come from;

(7)
$$\begin{bmatrix} i & & i \\ 0 & , & 0 \\ I_{T-K} & & -I_{T-K} \end{bmatrix}$$

There are $2^{(T-K)}$ ways of choosing (T-K) linearly independent columns from (7), and we account for all these possibilities by using:

$$\begin{bmatrix} i \\ 0 \\ I^* \end{bmatrix}$$

where $I^*$ is in $\bar{I}$. The T linearly independent columns which form a basis are then given by;

$$\begin{bmatrix} X(h_i) & 0 \\ X(\bar{h}_i) & I^* \end{bmatrix} \quad . \tag{8}$$

One can view (8) as a variable basis since a different basic solution for the linear program is determined by each $I^*$ in $\bar{I}$.

In terms of the L.A.E. problem the reason we make the basis a variable is that we can have $u_t^* = 0;\ t \epsilon h_i$, in $2^{T-K}$ different ways. That is, $u_t^*$ for $t \epsilon \bar{h}_i$ can be positive or negative for all $t \epsilon \bar{h}_i$. By defining $I^*$ as some element in $\bar{I}$ we allow for all of the different ways in which $u_t^* = 0;\ t \epsilon h_i$.

Following the approach discussed in the previous section, (6) can now be solved by premultiplying by $Q^{-1}$ where;

$$Q = \begin{bmatrix} 1 & 0 & i \\ 0 & X(h_i) & 0 \\ 0 & X(\bar{h}_i) & I^* \end{bmatrix} \quad .$$

Using partitioned inverse rules (see Goldberger (1964), p. 27); and noting that for any $I^*$, $I^{*-1} = I^*$ yields;

$$Q^{-1} = \begin{bmatrix} 1 & iI^*X(\bar{h}_i)X(h_i)^{-1} & -iI^* \\ 0 & X(h_i)^{-1} & 0 \\ 0 & -I^*X(\bar{h}_i)X(h_i)^{-1} & I^* \end{bmatrix} \quad .$$

Premultiplying both sides of (6) gives;

$$(9)\qquad \begin{bmatrix} D_1 \\ X(h_i)^{-1}Y(h_i) \\ D_2 \end{bmatrix} = \begin{bmatrix} 0 & i+iD_3 & i-iI^* & i-iD_3 & i+iI^* \\ I & X(h_i)^{-1} & 0 & -X(h_i)^{-1} & 0 \\ 0 & -D_3 & I^* & D_3 & -I^* \end{bmatrix} \phi$$

$$D_1 = S-iI^*X(\bar{h}_i)X(h_i)^{-1}Y(h_i)-iI^*Y(\bar{h}_i)$$

$$D_2 = I^*X(\bar{h}_i)X(h_i)^{-1}Y(h_i)+I^*Y(\bar{h}_i)$$

$$D_3 = I^*X(\bar{h}_i)X(h_i)^{-1} \quad .$$

This represents a simplex iteration on (6). All we have done is transform (6) into an equivalent problem where feasibility and optimality can be easily checked. From (9) we want to know whether $\beta^* = X(h_i)^{-1}Y(h_i)$ is an optimal and feasible solution.

From the previous section it follows that this solution is optimal if every element in the first row of the right hand side of (9) is nonnegative. This condition is satisfied by $(i - iI^*)$ and $(i + iI^*)$ since for any $I^*$ every element is 0 or 2.[5] The crucial condition for optimality is given by:

$$i + iI^*X(\bar{h}_i)\ X(h_i)^{-1} \geq 0$$

$$i - iI^*X(\bar{h}_i)\ X(h_i)^{-1} \geq 0 \quad .$$

Combining these conditions gives;

Condition 1. (Optimality). A necessary condition for $\beta^* = X(h_i)^{-1}Y(h_i)$ is;

$$-i \leq iI^*X(\bar{h}_i)X(h_i)^{-1} \leq i$$

for some I* in $\bar{I}$.

The unique optimality condition is given by;

Condition 1' (Unique Optimality). A necessary condition for $\beta^* = X(h_i)^{-1}Y(h_i)$ to be the unique solution to the L.A.E. linear programming problem is;

$$i < iI^*X(\bar{h}_i)X(h_i)^{-1} < i$$

for some I* in $\bar{I}$.

Note that these conditions do not depend on Y.

The feasibility condition follows directly from the $(r^+, r^-) \geq 0$ constraint. We get;

Condition 2 (Feasibility). A necessary condition for $\beta^* = X(h_i)^{-1}Y(h_i)$ is;

$$-I^*X(\bar{h}_i)X(h_i)^{-1}Y(h_i) + I^*Y(\bar{h}_i) \geq 0$$

from some I* in $\bar{I}$.[6]

By combining these conditions we obtain the following necessary and sufficient conditions for the solution to the L.A.E. linear programming problem.

Theorem 3. If there is an I* in $\bar{I}$ which satisfies both Conditions 1 and 2 then $\beta^* = X(h_i)^{-1}Y(h_i)$ is a solution to the L.A.E. linear programming problem, and conversely.

Theorem 4. If there is an I* in $\bar{I}$ which satisfies Condition 2 and;

$$-i \leq iI^*X(\bar{h}_i)X(h_i)^{-1} \leq i$$

then $\beta^* = X(h_i)^{-1}Y(h_i)$ is a nonunique solution, and conversely.

Theorem 5. If there is an I* in $\bar{I}$ such that both Conditions 1' and 2 are satisfied then $\beta^* = X(h_i)^{-1}Y(h_i)$ is the unique solution to the L.A.E. linear programming problem, and conversely.

We now examine some of the consequences of this theorem.

## 3. Conditions for a Unique Solution

Using the above conditions we can now derive the restrictions on X which insure that the solution to the L.A.E. minimization problem is an estimator.

Theorem 6. The solution to the L.A.E. linear programming problem is unique for all Y in $R^T$ iff for every $h_i$;

$$-i < iI^*X(h_i)X(h_i)^{-1} < i \tag{11}$$

for every I* in $\bar{I}$ which satisfies Condition 1.

Proof: If for every $h_i$, (11) is satisfied by every I* in $\bar{I}$ which satisfies Condition 1 then, according to Theorem 5, $\beta^*$ will be unique. The converse follows from the fact that if (11) is not satisfied for some $h_i$, then there must exist an I*, such that

$$-i \leq iI_1^* X(\bar{h}_i)X(h_i)^{-1} \leq i \quad .$$

Then for all Y such that;

$$-I_1^* X(\bar{h}_i)X(h_i)^{-1}Y(h_i) + I_1^*Y(\bar{h}_i) \geq 0 \quad .$$

Condition 2 is satisfied. It follows from Theorem 4 that for these Y, $\beta^*$ is not unique. Hence (11) must be satisfied for

all $h_i$ to insure uniqueness.

Theorem 6 implies that if we want a well-defined estimator we have to exclude those X which do not satisfy (11). In the multivariate case this does not appear to be a serious restriction. If it turns out that X does not satisfy (11) then to get uniqueness one only has to slightly change the values of some explanatory variables. This might be accomplished by measuring or recording the values of the (nonconstant) explanatory variables to one more (or less) decimal place.[7] This option for the multivariate regression problem is always available, and, therefore, $\beta^*(Y)$ can be made a well-defined estimator.

In the rest of this study we assume that X satisfies (11) so that $\beta^*$ is an estimator. We have not examined all of the consequences of (11) except for the following two examples. The first example illustrates a well-known case in which X does not satisfy (11).

Example 1. In Chapter III, Property 7', we showed that for $X = (1,\ldots,1)$ the L.A.E. estimator was the median observation on Y. As is well-known the median is not well-defined when T is even. We show here how this follows from Theorem 6.

Let $X = (1,\ldots,1)$ and $T = 2n$ where n is a positive integer. Let the diagonal elements of $I^*$ be $e_t$ where $t = 1,\ldots,2n-1$ and $e_t = \pm 1$. Then for any $h_i$;

$$iI^*X(\bar{h}_i)X(h_i)^{-1} = \sum_{t=1}^{2n-1} e_t \quad .$$

When n of the $e_t = +1$ and $(n-1)$ of the $e_t = -1$ we have

$\Sigma e_t = +1$. When n of the $e_t = -1$ and (n-1) of the $r_t = +1$ we have, $\Sigma e_t = -1$. Here it is impossible to satisfy (11) (for any $h_i$) so that $\beta^*$ is not unique. It is also easy to show that for $T = 2n+1$, (11) is satisfied for every $h_i$.

Example 2. Let $I^* = I$ and suppose;

$$iIX(\bar{h}_i)X(h_i)^{-1} = -i \tag{12}$$

so that (11) is not satisfied and nonunique solutions are possible. We can rewrite (12) as;

$$iX(\bar{h}_i) = -iX(h_i)$$

or

$$iX(\bar{h}_i) + iX(h_i) = 0$$

$$i[X(\bar{h}_i), X(h_i)] = 0$$

$$iX = 0$$

and this implies;

$$\sum_{t=1}^{T} x_{tk} = 0 \qquad k = 1,\ldots,K \quad .$$

Therefore, the L.A.E. estimator is not well defined if for each explanatory variable the sum over the T observations is zero. From Condition 2 we know that the Y in $R^T$ which lead to nonunique solutions are given by;

$$-X(\bar{h}_i)X(h_i)^{-1}Y(h_i) + Y(\bar{h}_i) \geq 0 \quad . \tag{13}$$

An example is presented in Appendix III.

Note that when the L.A.E. estimated hyperplane is forced through the point of means we have, $\sum_{t=1}^{T} x_{tk} = 0$. In this case, however, the domain of Y is no longer $R^T$ since Y has to satisfy;

$$\Sigma y_t = 0 \quad . \tag{14}$$

(Here $x_{tk}$ and $y_t$ represent observations expressed as derivations from their respective means.) With this restriction on

Y, Theorem 6 is only a sufficient condition for uniqueness and forcing the L.A.E. hyperplane through the point of means does not necessarily lead to nonunique solutions. To demonstrate this we show that with probability zero there are no Y which satisfy both (13) and (14)

Suppose there is a Y which satisfies (13). It then follows that;

$$i[-X(\bar{h}_i)X(h_i)^{-1}Y(h_i) + Y(\bar{h}_i)] \geq 0$$

and using (12) we obtain;

$$i[Y(h_i), Y(\bar{h}_i)] \geq 0$$

or

$$\Sigma y_t \geq 0 \quad .$$

But this must be an equality if (14) holds and therefore,

$$-X(\bar{h}_i)X(h_i)^{-1}Y(h_i) + Y(\bar{h}_i) = 0 \quad . \tag{15}$$

Since Y has a continuous density function the probability of (15) is zero. In this case, then the fact $\sum_{t=1}^{T} x_{tk} = 0$ does not imply that nonunique solutions will occur. This is illustrated in Appendix III.

## 4. A Characterization of $\beta^*$

In this section we use Conditions 1' and 2 to show that $\beta^*(Y)$ can be written as a linear function defined on certain sets which partition $R^T$. This follows from the fact that Condition 1' is independent of Y and therefore the space $R^T$ can be partitioned independently of the realizations of the random variable Y. This particular representation of $\beta^*(Y)$ is known as a linear bifunction; see Rockafellar (1970, Chapter 31).

In order to partition the sample space $R^T$ we (implicitly) solve for all $I^*$ in $\bar{I}$ which satisfy;

$$-i < iI^*X(\bar{h}_i)X(H_i)^{-1} < 1 \quad . \tag{14}$$

First notice that if $I^*$ satisfies (14) then so does $-I^*$. We can therefore restrict attention to those $I^*$ in $\bar{I}$ such that the first diagonal element is +1. If such an $I^*$ satisfies (14) then so does $-I^*$.

For any $h_i$ we define $\bar{I}(h_i)$ to be the set of all $I^*$ in $\bar{I}$ with first element +1, and which satisfies (14). Let $W_i$ be the number of elements in this set. When $\bar{I}(h_i)$ is empty we let $W_i = 0$. The typical element in $\bar{I}(h_i)$ is denoted by $I^*(h_i, w_i)$ so that;

$$-i < iI^*(h_i, w_i)X(\bar{h}_i)X(h_i)^{-1} < i \quad \text{for } w_i \neq 0$$

and

$$\bar{I}(h_i) = \{I^*(h_i,w_i) : w_i = 0,1,\ldots,W_i;\ W_i \neq 0\}$$

$$= \phi(\text{empty}) \text{ if } W_i = 0 \quad .$$

The set of all $I^*$ in $\bar{I}$ which satisfy (14) for a given $h_i$ is then;

$$\bar{I}(h_i)\bigcup - \bar{I}(h_i) \quad .$$

The reason for not excluding $W_i = 0$ in the above is that, depending on X, there may be some $h_i$ for which (14) is not satisfied by any $I^*$. This means that the K observations corresponding to $h_i$ can never satisfy Condition 1' and therefore never yield the L.A.E. estimate. This gives the answer to Taylor's Monte Carlo result in which he found certain sets of observations never determined the L.A.E. estimate (see Appendix II). Taylor's conjecture that this was not a sampling result, but a property of the L.A.E. estimator is correct.

Using $I^*(h_i,w_i)$, we define the sets (empty if $W_i = 0$);

$$P(h_i,w_i) = \{Y \epsilon R^T;\ - I^*(h_i,w_i)X(\bar{h}_i)X(h_i)^{-1}Y(h_i) +$$

$$I^*(h_i,w_i)Y(\bar{h}_i) \geq 0\}$$

and

$$P(h_i) = \{P(h_i,w_i) : w_i = 1,\ldots,W_i\}$$

and finally;

(15) $$P(h_i)\bigcup - P(h_i) \quad .$$

Notice that for a $I^*(h_i,w_i)$ which satisfies Condition 1',

$P(h_i,w_i)$ is the set of all Y which satisfy Condition 2. Similarly, (15) is the set of all Y in $R^T$ such that the K observations corresponding to $h_i$ determine the L.A.E. estimate. Therefore;

$$P(h_i)\bigcup -P(h_i) = \{Y\varepsilon R^T : H^*(Y) = h_i\} \quad .$$

All we have done here is partition[8] the sample space into those sets such that $\beta^* = X(h_i)^{-1}Y(h_i)$. We can therefore characterize the L.A.E. estimator as a linear function of Y, given Y in $P(h_i)\bigcup -P(h_i)$. That is;

$$\beta^* = X(h_i)^{-1}Y(h_i) \quad \text{if} \quad Y\varepsilon P(h_i) \text{ or } -P(h_i) \quad . \tag{16}$$

The characterization of the L.A.E. estimation error, $(\beta^*-\beta)$ is identical to (16) except that u is in the place of Y. This follows from Property 3, Chapter II, but it can be shown directly here. Suppose Y is in $P(h_i)$ and substitute

$$Y(h_i) = X(h_i)\beta + u(h_i)$$
$$Y(\bar{h}_i) = X(\bar{h}_i)\beta + u(\bar{h}_i)$$

into (16). We get;

$$\beta^* = X(h_i)^{-1}X(h_i)\beta + X(h_i)^{-1}u(h_i)$$

or

$$\beta^*-\beta = X(h_i)^{-1}u(h_i) \quad .$$

Also,

$$-I^*X(\bar{h}_i)X(h_i)^{-1}X(h_i)\beta - I^*X(\bar{h}_i)X(h_i)^{-1}u(h_i)\beta + I^*X(\bar{h}_i) +$$

$$I^*u(h_i) = -I^*X(\bar{h}_i)X(h_i)^{-1}u(h_i) + I^*u(h_i) \geq 0 \quad .$$

Hence;

$$\beta^*-\beta = X(h_i)^{-1}u(h_i) \quad \text{if} \quad u\varepsilon P(h_i) \text{ or } -P(h_i) \quad .$$

Again, this is just a restatement of Property 3, Chapter II.

We now examine the way in which the $P(h_i)$ partition $R^T$.

Properties of $P(h_i)$

Property 1. If Y is in $P(h_i,w_i)$ then $\lambda Y$ is in $P(h_i,w_i)$ for $\lambda \geq 0$, and $\lambda Y$ is in $-P(h_i,w_i)$ if $\lambda \leq 0$.

Proof: By assumption there is a $I^*(h_i,w_i)$ such that;

$$(17) \quad -I^*(h_i,w_i)X(\bar{h}_i)X(h_i)^{-1}Y(h_i) + I^*(h_i,w_i)Y(\bar{h}_i) \geq 0 \quad .$$

Multiplying both sides by $\lambda \geq 0$ leaves (17) satisfied so that $\lambda Y$ is in $P(h_i,w_i)$. Multiplying both sides by $\lambda < 0$ reverses the inequality, but since $I^*(h_i,w_i)$ satisfies Condition 1' then so does $-I^*(h_i,w_i)$ and this gives the required result.

Notice that this property implies Property 2, Chapter II.

Property 2. If $Y^1,Y^2$ are in $P(h_i,w_i)$ then $(Y^1+Y^2)$ is in $P(h_i,w_i)$.

Proof: By assumption:

$$(18) \quad -I^*(h_i,w_i)X(\bar{h}_i)X(h_i)^{-1}Y^1(h_i) + I^*(h_i,w_i)Y^1(\bar{h}_i) \geq 0$$

and

$$(19) \qquad -I^*(h_i,w_i)X(\bar{h}_i)X(h_i)^{-1}Y^2(h_i) + I^*(h_i,w_i)Y^2(\bar{h}_i) \geq 0 \quad .$$

Adding (18) and (19) gives the required result.

Properties 1 and 2 imply that $P(h_i,w_i)$ is closed under addition and positive scaler multiplication; that is, $P(h_i,w_i)$ is a convex cone.

Heuristically the L.A.E. estimator can now be viewed in the following manner. For any given X we use Condition 1' to partition $R^T$ into the convex cones $P(h_i,w_i)$ and $-P(h_i,w_i)$. We associate with each $P(h_i,w_i)$ the linear function $X(h_i)^{-1}$. Then for any Y in $R^T$ we find the set $P(h_i,w_i)$ in which it is contained and apply the linear transformation $X(h_i)^{-1}$. This yields the L.A.E. estimator. This is the characterization we use in analyzing sampling properties.

## 5. A Mathematical Result

In order to derive all of the sampling properties of the L.A.E. estimator the only mathematical problem which remains is an explicit characterization of all those I* in $\bar{I}$ which satisfy (11). This problem is described and discussed in Chapter VI. Here we present a result concerning the (T-KxK) matrix, $X(\bar{h}_i)X(h_i)^{-1}$, which appears in (11). The result is used below in analyzing those cases in which $\beta^*$ is a linear function.

Theorem 7. Every element in the matrix $X(\bar{h}_i)X(h_i)^{-1}$ is of the form;

$$\pm \frac{|X(h_j)|}{|X(h_i)|} \qquad i \neq j$$

where $|X(h_j)|$ is the determinant of the (K x K) matrix $X(h_j)$.

Proof: For notational ease let $h_i = h_1 = (1,2,\ldots,K)$. Then,

$$X(h_1) = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & & \vdots \\ \vdots & & \ddots & \\ x_{K1} & \cdots\cdots & & x_{KK} \end{bmatrix}$$

and $X(h_1)^{-1} = \frac{1}{|X(h_1)|}$ adj $X(h_1)$ where the adjoint matrix is given by;

$$\text{adj } X(h_1) = \begin{bmatrix} |C_{11}| & |C_{21}| & \cdots & |C_{K1}| \\ |C_{12}| & |C_{22}| & & \vdots \\ \vdots & & \ddots & \\ |C_{1K}| & \cdots\cdots & & |C_{KK}| \end{bmatrix}$$

where $|C_{ij}|$ is the determinant of the $ij^{th}$ cofactor.

The matrix $X(\bar{h}_i)$ is;

$$X(\bar{h}_1) = \begin{bmatrix} x_{K+1,1} & x_{K+1,2} & \cdots & x_{K+1,T} \\ x_{K+2,1} & x_{K+2,2} & & \vdots \\ \vdots & & & \\ x_{T,1} & \cdots\cdots & & x_{T,K} \end{bmatrix} .$$

Multiplying $X(\bar{h}_i)$ adj $X(h_1)$ then gives;

$$(20) \qquad \begin{bmatrix} \sum_{i=1}^{K} x_{K+1,i}|C_{1i}| & \cdots & \sum_{i=1}^{K} x_{K+1,i}|C_{K,i}| \\ \vdots & & \vdots \\ \sum_{i=1}^{K} x_{T,i}|C_{1i}| & \cdots & \sum_{i=1}^{K} x_{T,i}|C_{N,i}| \end{bmatrix}$$

Now look at $X(h_2)$ where $h_2 = (2,3,\ldots,K, K+1)$, or

$$X(h_2) = \begin{bmatrix} x_{21} & x_{22} \cdots x_{2K} \\ x_{31} & \\ \vdots & \vdots \\ x_{K+1,1} & x_{K+1,K} \end{bmatrix}$$

To find $|X(h_2)|$ expand by the last row. Except for the sign, this gives;

$$\sum_{i=1}^{K} x_{K+1,i}|C_{1i}|$$

where the cofactor matrix is the same as in (20) (except possibly for a sign change). Hence this is the first element in (20). This process can now be repeated for any $h_i$. Therefore, every element in $X(\bar{h}_i)X(h_i)^{-1}$ is of the form;

$$\pm \frac{|X(h_j)|}{|X(h_i)|} \quad .$$

An illustration of this result can be found in Appendix II.

This theorem shows that in "solving" (11) for I* the relevant data are the $|X(h_i)|$.

## 6. $\beta^*(Y)$ As A Linear Estimator

In this section we examine the special case in which $\beta^*(Y)$ is a linear function. When $\beta^*(Y)$ is linear we know from the Gauss-Markov Theorem that for any $F(u)$ in $F^*$, $\mathrm{Var}(\hat{\beta}) \leq \mathrm{Var}(\beta^*)$ so that it is important for these cases to be isolated. The conditions such that $\beta^*$ is linear provides the initial insight into the way in which X affects the sampling properties of $\beta^*$, and also begins to demonstrate the relationship between Taylor's and Rosenberg and Carlson's Monte Carlo results.

From the representation of $\beta^*$ in (16) we easily obtain;

Theorem 8. If for some $h_i$,

$$P(h_i) \cup -P(h_i) = R^T \tag{21}$$

then,

$$\beta^* = (X(h_i)^{-1}, 0) \cdot (Y(h_i), Y(\bar{h}_i)) = X(h_i)^{-1} Y(h_i)$$

for all Y; and conversely. That is, $\beta^*$ is a linear function of Y.

Proof: If (21) holds then from (16) it follows that;

$$\beta^* = X(h_i)^{-1} Y(h_i) \quad \text{if} \quad Y \epsilon R^T$$

and this proves the first part of the theorem. Since, $\beta^* = X(h_i)^{-1} Y(h_i)$ for $Y \epsilon P(h_i)$ or $-P(h_i)$, the only way for $\beta^*$ to be linear for Y in $R^T$ is for (21) to hold. This proves the converse.

Note that if (21) holds for some $h_i$, say $h_1$, then the sets $P(h_i)$, $i \neq 1$ must be empty. The K observations corresponding to $h_1$ always determine the L.A.E. estimator. This is the limiting case of Taylor's Monte Carlo result where he found that some of the $P(h_i)$ were empty.

Also, when $\beta^*$ is linear the approximate sampling distribution conjectured by Rosenberg and Carlson (1972) will lead to poor results. They conjectured that $Var(\beta^*)$ is scaled by the variance of the median, and therefore if this variance is less than the variance of the mean one would conclude that $Var(\beta^*) \leq Var(\hat{\beta})$. By the Gauss-Markov Theorem this conjecture is incorrect when $\beta^*$ is linear. However, when $\beta^*$ is the median none of the $P(h_i)$ are empty (see Chapter V). One might therefore conjecture that the comparison of the median and mean becomes less relevant as a way of comparing the multivariate L.A.E. and least squares estimators as the number of $P(h_i)$ which are empty increases. If widely-dispersed, high kurtosis explanatory variables imply that "a large number" of $P(h_i)$ are empty then the conjecture would provide an explanation of Rosenberg and Carlson's Monte Carlo results. This conjecture and explanation are discussed in the following chapters.

We now derive some consequences of Theorem 8.

Corollary 1. If K = T-1 then $\beta^*$ is linear.

Proof: $I^*$ is (T-KxT-K) = 1 x 1, a scalar, and $\bar{I}$ contains the two elements, {+1, -1}. Find an $h_i$, say $h_1$, such that;

$$-i < X(\bar{h}_1)X(h_1)^{-1} < i \tag{22}$$

so that $I^* = +1$ satisfies Condition 1'. Then,

$$P(h_1) = \{Y \epsilon R^T : -X(\bar{h}_1)X(h_1)^{-1}Y(h_1) + Y(\bar{h}_i) \geq 0\}$$

$$-P(h_1) = \{Y \epsilon R^T : X(\bar{h}_1)X(h_1)^{-1}Y(h_1) - Y(\bar{h}_i) \geq 0\}$$

and note that, $-X(\bar{h}_i)X(h_1)^{-1}Y(h_1) + Y(\bar{h}_1)$ is a scaler. Therefore, $P(h_1) \bigcup -P(h_1) = R^T$ and $\beta^*$ is linear.

For the case in which K = T-1 we can use the result of the previous section to indicate precisely which K observations determine the L.A.E. estimator.

When K = T-1 there are $\binom{T}{T-1}$ elements in H. That is,

$$H = \{h_1, \ldots, h_T\} \quad .$$

Let $|X(h_i)|$ be the absolute value of the determinant of $X(h_i)$. Arrange these determinants so that;

$$|X(h_1)| > |X(h_2)| > \ldots > |X(h_T)| > 0 \quad . \tag{23}$$

We then show;

Corollary 2. If K = T-1 and the $|X(h_i)|$ are arranged as in (23) then $\beta^* = X(h_1)^{-1}Y(h_1)$ for all Y in $R^T$.

Proof: From Corollary 1 we know that $\beta^*$ is linear. Therefore we have to show;

$$-i < X(\bar{h}_1)X(h_1)^{-1} < i \quad .$$

Using Theorem 7 we can write the (1 x T-1) vector $X(\bar{h}_1)X(h_1)^{-1}$ (except for sign) as

$$\left(\frac{|X(h_2)|}{|X(h_1)|}, \frac{|X(h_3)|}{|X(h_1)|}, \ldots, \frac{|X(h_T)|}{|X(h_1)|}\right) \quad .$$

From the construction in (23) it follows that

$$-1 < \frac{|X(h_i)|}{|X(h_1)|} < 1$$

for $i = 2,\ldots,T$. Therefore $\beta^* = X(h_1)^{-1}Y(h_1)$.

In practice, Corollary 2 is of no use since for $K = T-1$ no one would ever be led to choose the L.A.E. estimator. (The L.A.E. estimator actually ignores the one observation corresponding to $\bar{h}_i$.) However, since the L.A.E. estimator is determined by the $T-1 = K$ observations with the largest determinant it minimizes the generalized variance for all linear estimators which are functions of (T-1) observations. In other words, $\beta^*$ at least chooses the best set of (K-1) observations.

We now present a sufficient condition such that $\beta^*$ is nonlinear.

Corollary 3. For the model;

$$Y = (i,Z)\beta + u$$

and $K \leq (\frac{T-1}{2})$, $\beta^*$ is a nonlinear estimator.

Proof: To prove the corollary we show that there is an $I^*$ in $\bar{I}$ which cannot satisfy Condition 1' for any $h_i$. From Property 7, Chapter III we know:

$$|P - N| < X$$

where P and N are the number of positive and negative $u_t^*$, respectively. Let $I^* = I$ so that this corresponds to having

(T-K) of the $u_t^* > 0$ and none of the $u_t^* < 0$ (see footnote 6 this chapter). Hence:

$$|P - N| = T - K < K \quad .$$

But this is impossible if $K \leq (\frac{T-1}{2})$. Therefore, $I^* = I$ cannot satisfy Condition 1' for any $h_i$, and (T-K) of the $u_t^* > 0$ is impossible. Therefore, $\beta^* \neq X(h_i)^{-1}Y(h_i)$ for every Y in $R^T$ and $\beta^*$ is nonlinear.

This corollary implies that for almost all regression models in economics $\beta^*$ is nonlinear, and the Gauss-Markov Theorem therefore cannot be used to exclude $\beta^*$. For example, if the model contains a constant explanatory variable and, say, 10 nonconstant explanatory variables (for economic time series there are few, if any, models with 10 or more explanatory variables) then for $T \geq 23$, $\beta^*$ is nonlinear.

## 7. Conclusion

In this chapter we solved the L.A.E. linear programming problem for the conditions such that any set of K observations determine the L.A.E. estimate. The conditions show explicitly how $\beta^*$ depends on Y and X and they are used in Chapter VI to analyze the sampling properties of $\beta^*$. We have also derived the restrictions on X which are needed for $\beta^*$ to be an estimator, and also the conditions such that $\beta^*$ is linear.

In the next chapter we begin to analyze the sampling properties of $\beta^*$ by focusing on the simple case in which $\beta^*$ is the median.

## Footnotes

[1]Vector inequalities are defined by:

$$X \geq Y \quad \text{iff} \quad x_t \geq y_t$$

$$X > Y \quad \text{iff} \quad x_t > y_t$$

$$X \geqq Y \quad \text{iff} \quad X \geq Y \text{ and } x_t = y_t \text{ for at least one } t.$$

[2]Problem 1 is actually quite general since the constraints for any linear programming problem can be put in the form of (2), (3).

[3]Here we have implicitly assumed that there exists a feasible and optimal solution.

[4]If rank X = K and rank $X(h_i)$ < K then the K observations corresponding to $h_i$ never determine the L.A.E. estimate.

[5]The fact that some of these elements are 0 causes no problems for unique optimality since the corresponding variables $r_t^+$, $r_t^-$ are linearly dependent.

[6]Note that if a diagonal element of I* is +1 (-1) then the corresponding $u_t^*$ is positive (negative).

[7]The only case in which this cannot be done is when X only contains a constant vector (or dummy variables).

[8]For any Y there is some $\beta^*$ so that the union of all $P(h_i,w_i)$ is $R^T$. Also, since the L.A.E. estimator is assumed to be well defined there is a unique $P(h_i,w_i)$ which contains Y. Therefore, the sets $P(h_i,w_i)$ partition $R^T$.

# CHAPTER V

## THE L.A.E. ESTIMATOR: THE MEDIAN

In this chapter we examine the sampling properties of the L.A.E. estimator for the special case in which $X = (1,\ldots,1)'$ so that the L.A.E. estimator is the median. The median serves as a prototype for our study of the L.A.E. estimator. Insights are obtained in the same way that insights into least squares are obtained by studying the mean.

The primary reason for studying the median is to illustrate how sampling properties can be derived from the characterization of $\beta^*$ presented in Chapter IV. Another reason for studying the median is that for the preferences described in Chapter I the choice between the mean and median is a relatively simple problem to analyze. Whether this analysis can be extended to the general case is discussed in the next chapter.

The chapter is divided into six sections. In Section 1 we show how the median estimator follows from our characterization of $\beta^*$. This characterization is used in Section 2 to derive the sampling density of $(\beta^*-\beta)$. In Section 3 we state some of the properties of the median sampling density. In Section 4 we examine the F in F* such that the variance of the median is less than the variance of the mean. In Section

5 we analyze the median and mean when F(u) is a Student distribution. Conclusions are included in Section 6.

## 1. The Median

Definition

In this chapter we deal exclusively with the simple model:

$$(1) \qquad y_t = \beta + u_t \quad t = 1,\ldots,T : T = 2n+1$$

where n is a positive integer. In Chapter III we demonstrated the well-known fact that for (1) the L.A.E. estimator is the median $y_t$. We now briefly show how this follows from the characterization of $\beta^*$ presented in Chapter IV.

In the previous chapter we showed;

$$\beta^* = X(h_i)^{-1}Y(h_i) \quad \text{if} \quad y \in P(h_i) \text{ or } -P(h_i)$$

where,

$$(2) \qquad P(h_i) = \{P(h_i,w_i) : w_i = 0,1,\ldots,W_i\}$$

and

$$(3) \qquad \{P(h_i,w_i)\} = \{Y \in R^T : -I^*(h_i,w_i)X(\overline{h_i})X(h_i)^{-1}Y(h_i) + I^*(h_i,w_i)Y(\overline{h_i}) \geq 0\} \quad .$$

and where $I^*(h_i,w_i)$ satisfies Condition 1'. Using (1) we can write Condition 1' as;

$$-1 < iI^*(1,\ldots,1)' \cdot 1 < +1$$

where (1,...,1) contains (2n) ones. Let the diagonal elements of $I^*$ be $e_t$ where $e_1 = +1$ and $e_t = \pm 1$, $t = 1,\ldots,(2n)$ so that Condition 1' is;

$$-1 < \sum_{t=1}^{2n} e_t < +1 \quad .$$

Since (2n) is even this is satisfied if and only if n of the $e_t = +1$ and n of the $e_t = -1$. Therefore, $I^*(h_i,w_i)$ is a (T-1 x T-1) diagonal matrix with first diagonal element equal to +1, (n-1) of the remaining diagonal elements equal to +1, and n of the remaining elements equal to -1. This can occur in $\frac{(T-1)!}{2n!n!}$ ways so that

$$W_i = \frac{(T-1)!}{2n!n!} \quad . \tag{4}$$

Also note that this holds for each $h_i$, i=1,...,T.

We note that while the calculation of $I^*(h_i,w_i)$ is trivial for the model (1) it is much more complicated in the general case. For an arbitrary X the sets $\bar{I}(h_i)$ will in general not be identical for every $h_i$, and they cannot be easily characterized in terms of the number of +1 and -1 in $I^*(h_i,w_i)$. These issues are discussed in the next chapter.

Using $I^*(h_i,w_i)$ we can write (3) as;

$$P(h_i,w_i) = \{Y \epsilon R^T : -I^*(h_i,w_i)Y(h_i) + I^*(h_i,w_i)Y(\overline{h_i}) \geq 0\} \quad . \tag{5}$$

For example, if $h_i = h_1$ and;

$$I^*(h_1,w_1) = \begin{vmatrix} +1 & & & \\ & \ddots & & \\ & & +1 & \\ & & & -1 \\ & & & & \ddots \\ & & & & & -1 \end{vmatrix}$$

then $(P(h_1,w_1)$ is the set of all Y such that;

$$y_2 \geq y_1;\ldots;y_{n+1} \geq y_1;\ y_{n+2} \leq y_1;\ldots;y_T \leq y_1 \quad .$$

This is easily recognized as a sufficient condition for the first observation to be the median. Another sufficient condition is given by Y in $-P(h_1,w_1)$. Therefore

$$\beta^* = Y(h_1) = y_1 \text{ if } Y\varepsilon P(h_1,w_1) \text{ or } -P(h_1,w_1) \quad .$$

In a similar fashion we obtain the necessary and sufficient conditions such that $y_1$ is the median;

$$\beta^* = y_1 \text{ if } Y\varepsilon P(h_1,w_i) \text{ or } -P(h_1,w_i)$$

for $w_i = 1,\ldots,W_i$.[1]

It is then easy to see that

(6) $$\beta^* = Y(h_i) \text{ if } Y\varepsilon P(h_i,w_i) \text{ or } -P(h_i,w_i)$$

for $h_i = 1,\ldots,T$; $w_i = 1,\ldots,W_i$; is the definition of the median. Also the median estimator error, $(\beta^*-\beta)$ is;

(7) $$\beta^*-\beta = u(h_i) \text{ if } Y\varepsilon P(h_i,w_i) \text{ or } -P(h_i,w_i) \quad .$$

## 2. Sampling Density of $(\beta^*-\beta)$

In this section the sampling density of $(\beta^*-\beta)$ is derived using the above characterization of the median. This

illustrates the method used to derive the sampling density in the general case.

To obtain the probability density of $(\beta^*-\beta)$ we calculate the probability of the event;

$$(8) \qquad (\beta^*-\beta) \; \varepsilon \; \theta \pm \frac{1}{2} d \; \theta \quad .$$

From (7) it follows that the events which are equivalent to (8) are;

$$(9) \qquad (\beta^*-\beta) = u(h_i) \; \varepsilon \; \theta \pm \frac{1}{2} d \; \theta \quad \text{and}$$

$$u \varepsilon P(h_i,w_i) \text{ or } -P(h_i,w_i); \quad h_i = 1,\ldots,T; \; w_i = 1,\ldots,W_i \quad .$$

Since the sets $P(h_i,w_i)$ partition $R^T$ we have;

$$\Pr[u \varepsilon P(h_i,w_i) \cap P(h_j,w_j)] = 0 \quad i \neq j$$

and therefore (9) represents the $T.W_i$ mutually exclusive events which lead to (8). Therefore, the probability of (8) can be written;

$$(10) \qquad \sum_{h_i=1}^{T} \sum_{w_l=1}^{W_i} \Pr[u(h_i) \varepsilon \; \theta \pm \frac{1}{2} d \; \theta \text{ and } u \varepsilon (P(h_i,w_i) \text{ or } -P(h_i,w_i))] \quad .$$

To obtain the probability of one of the terms in this sum we write the joint probability as the product of marginal and conditional probabilities;

$$(11) \qquad \Pr[u(h_i) \varepsilon \; \theta \pm \frac{1}{2} d \; \theta] \Pr[u \varepsilon (P(h_i,w_i) \text{ or } -P(h_i,w_i)) \mid u(h_i) \varepsilon \; \theta \pm \frac{1}{2} d \; \theta] \quad .$$

For the first factor in (11) we get;

$$\Pr[u(h_i)\varepsilon\ \theta \pm \tfrac{1}{2}\, d\ \theta] = f_{h_i}(\theta)d\ \theta = f(\theta)d\theta \quad .$$

where $f_{h_i}$ is the density of $u(h_i)$. We have used assumption A5 ($u(h_i) = u_t$; identically distributed) so that $f_{h_i}$ does not depend on $h_i$.

For the conditional probability in (11) we have;

$$(12) \qquad \Pr[u\varepsilon P(h_i,w_i) \,|\, u\varepsilon\ \theta \pm \tfrac{1}{2}\, d\ \theta] +$$

$$\Pr[u\varepsilon -P(h_i)\ u\varepsilon\ \theta \pm \tfrac{1}{2}\, d\ \theta] \quad .$$

Substituting (5) for $P(h_i,w_i)$ in the first term yields;

$$\Pr[-I^*(h_i,w_i)u(h_i) + I^*(h_i,w_i)u(\overline{h_i}) \geq 0 \,|\, u(h_i)\varepsilon\ \theta \pm \tfrac{1}{2} d\ \theta]$$

or,

$$(13) \qquad \Pr[-I^*(h_i,w_i)\theta + I^*(h_i,w_i)u(\overline{h_i}) \geq 0] =$$

$$\Pr[I^*(h_i,w_i)u(\overline{h_i}) \geq I^*(h_i,w_i)\theta] \quad .$$

To simplify (13) we use assumption A5 ($u(h_i)$ independent and identically distributed). Notice that when the diagonal element in $I^*$ is +1 the corresponding term in (13) is;

$$\Pr[u_t > \theta] = (1-F(\theta))$$

and when the element is -1 we have

$$\Pr[-u_t > -\theta] = F(\theta) \quad .$$

Since n of the elements in $I^*(h_i,w_i)$ are +1 and n elements

are -1 we can write (13) as;

$$F(\theta)^n(1-F(\theta))^n \quad .$$

A similar relation holds for the second term in (12).

Combining these results allows us to write (10) as;

$$\sum_{h_i=1}^{T} \sum_{w_i=1}^{W_i} 2f(\theta)d\,\theta\, F(\theta)^n(1-F(\theta))^n \quad .$$

The terms in the sum do not depend on $h_i$ or $w_i$ so we have

$$T\cdot W_i \cdot 2f(\theta)d\,\theta\, F(\theta)^n(1-F(\theta))^n \quad .$$

Substituting (4) for $W_i$ gives;

$$(14) \qquad g(\theta)d\,\theta = \frac{T!}{n!n!} F(\theta)^n(1-F(\theta))^n f(\theta)d\theta \quad .$$

This is the well-known expression for the median sampling density. The expression can be obtained in a more straightforward fashion than we have presented (see Kendall and Stuart (1958), p. 329). Again, we have derived the median sampling density from our characterization of $\beta$ since it illustrates the methods used in the next chapter.

## 3. Properties of $g(\theta)$

In this section we present some well-known properties of the median, for F in F*.

### Finite Properties

1. Symmetry: $g(\theta) = g(-\theta)$. This is a consequence of the linear homogeneity of $(\beta^*-\beta)$, and the symmetry of $f(u)$.

2. From symmetry it follows that if $E(\beta^*-\beta)$ exists then $E(\beta^*-\beta) = 0$, and $\beta^*$ is unbiased. We show in the next section that the existence of $E(u)$ is not necessary for the existence of $E(\beta^*-\beta)$.

3. $g(0) = \frac{T!f(0)}{n!n!4^n} < \infty$. This follows from the assumption $f(0) < \infty$.

4. $g(\theta)$ is continuous. This follows from the continuity of f and F, and the fact that a product of continuous functions is continuous.

5. Let $G(\theta)$ be the probability distribution of $(\beta^*-\beta)$. Properties 1, 3, and 4 then imply G is in F*.

Property 5 shows that all of the properties of F* are preserved by the transformation, $(\beta^*-\beta)$. This is surprising in view of the fact that the median is a relatively complicated nonlinear function.

## Asymptotic Properties

1. If f is continuously differentiable then

$$f(0)2\sqrt{T}(\beta^*-\beta) \tag{15}$$

is asymptotically normal with expectation zero and variance one.

Proof: We outline the proof in Cramer (1946, p. 368). The density of (15) is;

$$\left[\frac{T!}{n!n!2\sqrt{T}]^n 2^n}\right] \left[\frac{f(s\ )}{f(0)}\right] [2F(s\theta)2(1-F(s\theta)]^n \tag{16}$$

where $s = (f(0)2\sqrt{T})^{-1}$. For any fixed $\theta$ take limits as $T \to \infty$. The first factor goes to $(2\pi)^{-1/2}$ and the second factor goes to one. For the third factor expand $2F(s\theta)$ about zero. Neglecting terms which go to zero gives;

$$2F(s\theta) = 1 + \frac{\theta}{\sqrt{T}} \quad .$$

Substituting this into the third factor in (16) and taking limits gives $e^{-x^2/2}$.

2. From the above it follows that $\beta^*$ is a consistent estimator of $\beta$.

We note that the differentiability assumption is not necessary for asymptotic normality. For example, Chu and Hotelling (1955) show that if F(u) is a Laplace distribution (which is not continuously differentiable) then the sampling density of the median is asymptotically normal.

Also note that the consistency of the median, as opposed to the mean, does not require any assumptions on the moments of u.

## 4. Variance of $\beta^*$

In this section we characterize a subset of F* such that $\text{Var}(\beta^*) \leq \text{Var}(\hat{\beta})$, and a subset such that $\text{Var}(\beta^*) \geq \text{Var}(\hat{\beta})$. For the preference ordering discussed in Chapter I, we can then translate an assumption on F(u) into a preference for $\beta^*$ or $\hat{\beta}$.

## F(u) in $F_1^*$

Let $F_1^*$ be the set of all probability distributions in $F^*$ such that var(u) = $\infty$. Included in $F_1^*$ are all nonnormal symmetric stable distributions. To obtain a sufficient condition for var($\beta^*$) < $\infty$, given F(u) in $F_1^*$ we use the following result due to Rosenberg and Carlson (1972).

Lemma. If $E|u|^J < \infty$ then $\lim_{u\to\infty} u^J(1-F(u)) = A < \infty$ for $u \varepsilon (0,+\infty)$.

Proof. See Rosenberg and Carlson (1972, p. 18) and for a partial converse see Cramer (1946, p. 71).

We now use this lemma to obtain a sufficient condition for var($\beta^*$) < $\infty$.[2]

Theorem 1. Given F(u) in $F^*$ and $E|u|^J < \infty$ implies $E|\beta^*-\beta|^{nJ} < \infty$.

Proof:

$$E|\beta^*-\beta|^{nJ} = \frac{2T!}{n!n!} \int_0^{+\infty} \theta^{nJ} f(\theta) F(\theta)^n (1-F(\theta))^n d\theta \quad .$$

We now use the facts;

$$F(\theta)^n < 1 \quad \text{for} \quad \theta \varepsilon (0,+\infty)$$

and

$$\theta^{nJ}(1-F(\theta))^n = A^N < \infty \quad \text{for } \theta \varepsilon (0,+\infty) \quad .$$

This last step follows from the assumption $E|u|^J < \infty$ and the above lemma. Substitution then yields;

$$E|\beta^*-\beta|^{nJ} < \frac{2T!A^N}{n!n!} \int_0^{\infty} f(\theta)\, d\,\theta < \infty \quad .$$

Theorem 1 implies that if $\text{Var}(u) = \infty$ and the $E(u)$ exists than $\text{Var}(\beta^*) < \infty$ for $T > 5$. For $F(u)$ in $F_1^*$ we therefore obtain:

Corollary. For $F(u)$ in $F_1^*$ (that is, $J < 2$);

$$\text{Var}(\beta^*) = \text{Var}(\hat{\beta}) = \infty \quad \text{if} \quad T < \frac{4}{J} + 1$$
$$\text{Var}(\beta^*) < \text{Var}(\hat{\beta}) = \infty \quad \text{if} \quad T > \frac{4}{J} + 1 \quad .$$

Hence for the preference ordering described in Chapter I and $F(u)$ in $F_1^*$ the median is preferred to the mean (or indifferent if $T < \frac{4}{J} + 1$).

We recognize that, for F in $F_1^*$ no one who uses the estimator $\hat{\beta}$ will continue to use the variance as a criteria for deciding among alternative estimators. Instead, preference might be determined by using $E|\hat{\beta}-\beta|^J$ as a measure of dispersion where $E|u|^J < \infty$. For example, if $E|u| < \infty$ then one might use a mean absolute deviation criteria. The examination of $\hat{\beta}$ and $\beta^*$ for these other measures of dispersion, while straightforward, is beyond the scope of this study.

The fact that $\text{Var}(\hat{\beta}) = \infty$ for F in $F_1^*$, however, is important if there is uncertainty concerning F in $F^*$, and the possibility that F is in $F_1^*$ is not excluded. Then, given the preferences discussed in Chapter I, it is not at all obvious that there are any objective functions which lead to $\hat{\beta}$ being preferred to $\beta^*$. The preference for the mean in this framework is analogous in game theory to choosing an act whose payoff ($\text{Var}(\hat{\beta})$) is an infinite loss for certain states ($F \epsilon F_1^*$). In contrast, the median because it has finite

variance for moderate T provides some insurance against these states. We return to this point below when discussing the Student distribution.

F(u) in $F_2^*$

We now examine $\hat{\beta}$ and $\beta^*$ for F in $F_2^*$ where $F_2^*$ is the set of all finite variance distributions in F*. The variance of the median, unlike the variance of the mean, cannot be easily expressed in terms of characteristics of F(u), and it is therefore not possible to completely partition $F_2^*$ into the subsets such that $\text{Var}(\beta^*) \leq \text{Var}(\hat{\beta})$ and $\text{Var}(\beta^*) \geq \text{Var}(\hat{\beta})$. The integrand in $\text{Var}(\beta^*)$ involves $F(\cdot)$ so that except when $F(\cdot)$ can be expressed without integrals, approximation methods and bounds must be used in evaluating $\text{Var}(\beta^*)$. Before reviewing these approximation methods, we obtain an exact result which shows that the set of F in $F_2^*$ such that $\text{Var}(\beta^*) < \text{Var}(\hat{\beta})$ is not empty.

The Laplace Distribution. This distribution, sometimes called the bilateral exponential (Feller (1966), p. 48) has density and distribution functions,

$$f(u) = \frac{1}{2}\alpha e^{-\alpha|u|} \qquad \alpha > 0$$

$$F(u) = 1 - \frac{1}{2}e^{-\alpha u} \qquad u \geq 0$$

$$= \frac{1}{2}e^{+\alpha u} \qquad u < 0 \quad .$$

The variance of this distribution is $2\alpha^{-1}$. An exact expression

for Var(β*) can be obtained since F(u) does not involve (explicit) integration. Note that the median is the maximum likelihood estimator for the Laplace distribution.

Given T = 2n+1, Var(β*) can now be calculated directly as,

$$\text{Var}(\beta^*) = \frac{T!}{n!n!} \, 2\int_0^\infty \theta^2 \left(\tfrac{1}{2}\alpha e^{-\alpha\theta}\right)\left(\tfrac{1}{2}e^{-\alpha\theta}\right)^n \left(1-\tfrac{1}{2}e^{-\alpha\theta}\right)^n d\theta \quad .$$

However, since,

$$\left(1-\tfrac{1}{2}e^{-\alpha\theta}\right)^n = \sum_{s=0}^{n} \binom{n}{s} \left(-\tfrac{1}{2}\right)^s e^{-\alpha\theta s}$$

and therefore,

$$\text{Var}(\beta^*) = \frac{T!\,2\alpha\left(\frac{1}{2}\right)^{n+1}}{n!n!} \sum \binom{n}{s} \left(-\tfrac{1}{2}\right)^s \int_0^\infty \theta^2 e^{-\theta(\alpha n+\alpha+\alpha s)} d\theta \quad .$$

From the properties of the gamma distribution (see Feller (1966), p. 48), it follows that:

$$\int_0^\infty 2e^{-\theta(\alpha n+\alpha+\alpha s)} d\theta = \frac{2}{\alpha^3 (n+s+1)^3} \quad .$$

Substituting this into the above yields:

$$\text{Var}(\beta^*) = \frac{T!\,2^{(1-n)}}{n!n!\alpha^2} \sum_{s=0}^{n} \binom{n}{s} \left(-\tfrac{1}{2}\right)^s (n+s+1)^3 \quad .$$

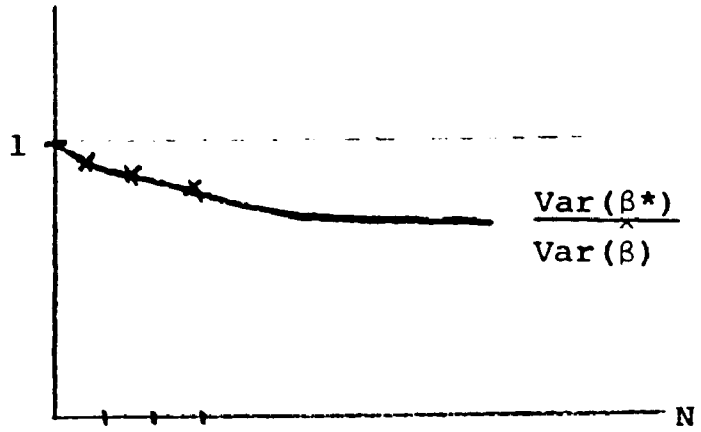
For the mean we have:

$$\text{Var}(\hat{\beta}) = \frac{2}{\alpha^2 T} \quad .$$

Finally,

$$\frac{\text{Var}(\beta^*)}{\text{Var}(\hat{\beta})} = \frac{TT!}{n!n!2^n} \sum_{s=0}^{n} \binom{n}{s} \left(-\tfrac{1}{2}\right)^s (n+s+1)^{-3} \quad .$$

In Figure 11 this relation is plotted and we have also graphed the standardized normal and Laplace densities so that



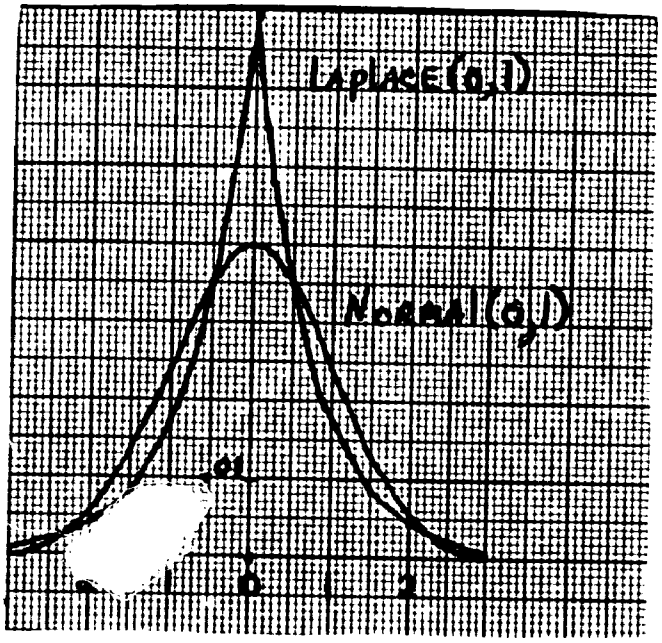




Fig. 11.--The Laplace and Normal Densities

one can see a type of density relative to the normal which leads to $\text{Var}(\beta^*) < \text{Var}(\hat{\beta})$.

A Lower Bound. We now state a result due originally to Chu (1955) which gives a lower bound on $\text{Var}(\beta^*)$.

Theorem 2 (Chu). For $F(u) \in F^*$;

$$\text{var}(\beta^*) > f(o)^{-2}(4(T+2))^{-1} \quad .$$

Proof. For $u > 0$, $F(u) - \frac{1}{2}$ equals,

$$\int_0^u f(s)ds < \int_0^u f(0)ds = f(0)u$$

since (by assumption) $f(0) > f(s)$; $s \neq 0$. A similar relation holds for $u < 0$ so that:

$$u^2 > f(0)^{-2}(F(u)-\tfrac{1}{2})^2 \quad .$$

Substituting this into $\text{var}(\beta^*)$ we obtain:

$$\text{Var}\,\beta^* > f(0)^{-2} \int_{-\infty}^{+\infty} \frac{T!}{n!n!} (F(\theta)-\tfrac{1}{2})^2 F(\theta)^n (1-F(\theta))^n dF(\theta)$$

$$= f(0)^{-2} \int_0^1 \frac{T!}{n!n!} (F-\tfrac{1}{2})^2 F^n (1-F)^n dF \quad .$$

The integral is the variance of a beta distribution with parameters (n+1, n+1) and this variance is (see Cramer (1946), p. 252):

$$(4(2n+3)^{-1} \quad .$$

Hence:

$$\text{Var}\hat{\beta} > f(0)^{-2}(4(2n+3))^{-1}$$

or in terms of T:

$$\text{Var}\ \hat{\beta} > f(0)^{-2}(4(T+2))^{-1} \quad .$$

Notice that this lower bound is almost identical with the variance of the asymptotic normal distribution. For finite T then, the variance of the median is greater than the asymptotic variance so that in using the asymptotic distribution for hypothesis testing one must make sure the finite sample distribution is "close to" the asymptotic distribution. Otherwise, the finite sample variance will be understated.

Since $\text{Var}(\hat{\beta}) = \sigma^2/T$ where $\sigma^2$ is the variance of u, we obtain;

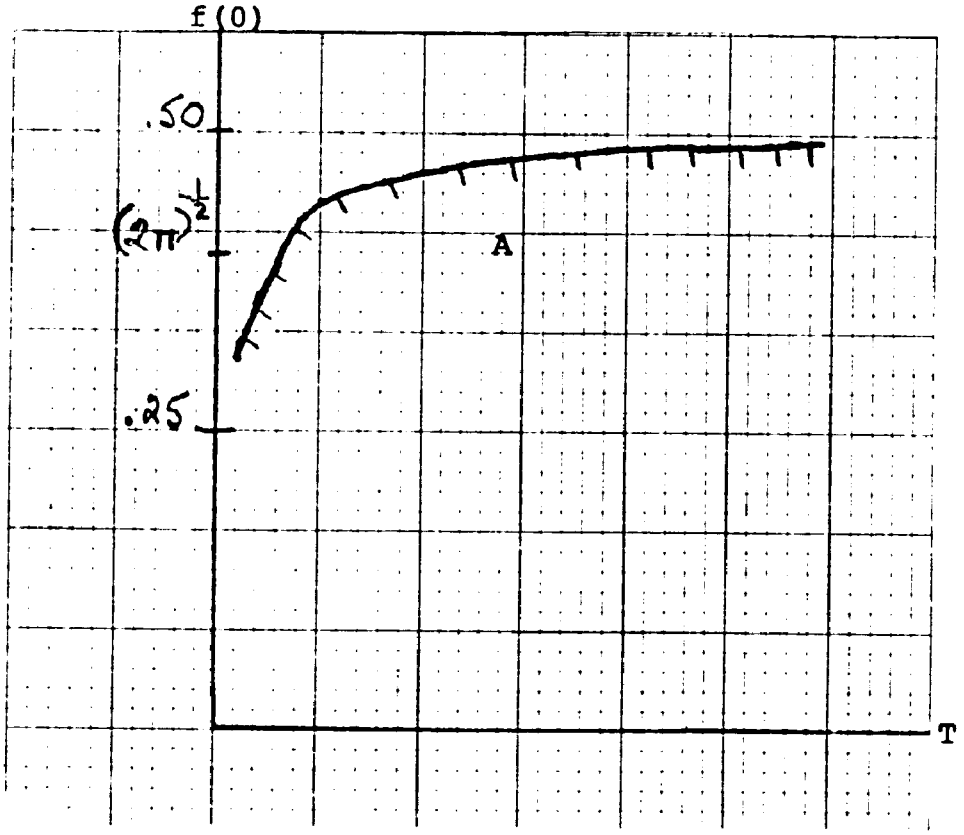
Corollary. If $f(0)^{-2}(4(T+2))^{-1} \geq \sigma^2/T$ then $\text{Var}(\beta^*) \geq \text{Var}(\hat{\beta})$.

In Figure 12 this relation is plotted for $\sigma^2 = 1$ and various T. For the standardized F in $F_2^*$ we see that $\text{Var}(\beta^*) \leq \text{Var}(\hat{\beta})$.

$$f(0) \leq \sqrt{3/20} = .387 \quad ; \quad \text{for } T = 3$$
$$f(0) \leq \sqrt{1/4} = .50 \quad ; \quad \text{for } T = \infty \quad .$$

This lower bound identifies a subset of $F_2^*$ such that $\text{Var}(\hat{\beta}) < \text{Var}(\beta^*)$. However, it does not completely characterize all of the F(u) such that $\text{Var}(\hat{\beta}) < \text{Var}(\beta^*)$. For example, if F(u) is normal then for all T, $\text{Var}(\hat{\beta}) < \text{Var}(\beta^*)$, and yet $f(0) = (2\pi)^{-1/2} = .399$ so that for T = 3 the lower bound does not exclude the possibility that $\text{Var}(\beta^*) < \text{Var}(\hat{\beta})$.

An Approximation. Here we assume f'(u) and f"(u) are continuous. To obtain an approximation for $\text{Var}(\beta^*)$ we expand $u(F)^2$




$$A = \{(f(0),T) : f(0)^{-2}(4(T+2))^{-1} > T^{-1}\}$$

Fig. 12.--Densities Such That, $\mathrm{Var}\hat{\beta} < \mathrm{Var}\beta^*$

around $F = \frac{1}{2}$ where $u(F)$ is the inverse of $F(u)$. This is known as Pearson's expansion (see Kendall and Stuart (1957), p. 328), and the approximation method has been used by David and Johnson (1954).

Expanding $u(F)^2$ around $F = \frac{1}{2}$ yields;

$$u^2 = f(0)^{-2}(F-\tfrac{1}{2})^2 - \tfrac{1}{3}f(0)^{-5} f''(0)(F-\tfrac{1}{2})^4 + R \quad .$$

The remainder term is complicated, and it is dropped (see Kendall and Stuart (1958)). Substituting the approximation for $u^2$ yields;

$$\mathrm{Var}(\beta^*) = f(0)^{-2} \frac{T!}{n!n!} \int_0^1 (F-\tfrac{1}{2})^2 F^n (1-F)^n dF - \tfrac{1}{3} f''(0) f(0)^{-5} \frac{T!}{n!n!} \int_0^1 (F-\tfrac{1}{2})^4 F^n (1-F)^n dF.$$

The integrands in this expression are the second and fourth moments about the mean of a beta distribution with parameters (n+1, n+1). Substituting these moments into the above eventually yields:

$$(17) \qquad \mathrm{var}\beta^* = \frac{1}{f(0)^2 4(T+2)} - \frac{f''(0)}{f(0)^5 16(T+2)(T+4)} \quad .$$

Kendall and Stuart (1957, p. 329) report that for T = 11 and F(u) = normal, the approximation gives a variance of .137 agreeing to three decimal places with the known variance of .137337. The accuracy of the approximation for other $F \varepsilon F_2^*$ is not known.

From the approximation we see that $\mathrm{Var}(\beta^*)$ decreases as T and f(0) increase, and $\mathrm{Var}(\beta^*)$ increases as $|f''(0)|$ increases.

In Figure 13 the approximation is plotted and compared with $\mathrm{Var}(\hat{\beta})$.

## 5. The Student Density

To illustrate the above approximation we take the Student density and compare $\mathrm{Var}(\beta^*)$ and $\mathrm{Var}(\hat{\beta})$. The Student density is:

$$f(u) = v^{-1/2} B(\tfrac{1}{2},\tfrac{1}{2}v)^{-1} (1+\frac{u^2}{v})^{-1/2(V+1)}$$

where

$$v \geq 1$$

$$B(\tfrac{1}{2},\tfrac{1}{2}v) = \frac{\Gamma(\frac{1}{2})\,\Gamma(\frac{v}{2})}{\Gamma(\frac{1}{2}+\frac{v}{2})}$$

where $\Gamma(\cdot)$ is the gamma function. For $v = 1$ this is the Cauchy density, and $v = \infty$ is the normal density. Also,

$$\mathrm{Var}(u) = \frac{v}{v-2} \qquad v > 2$$

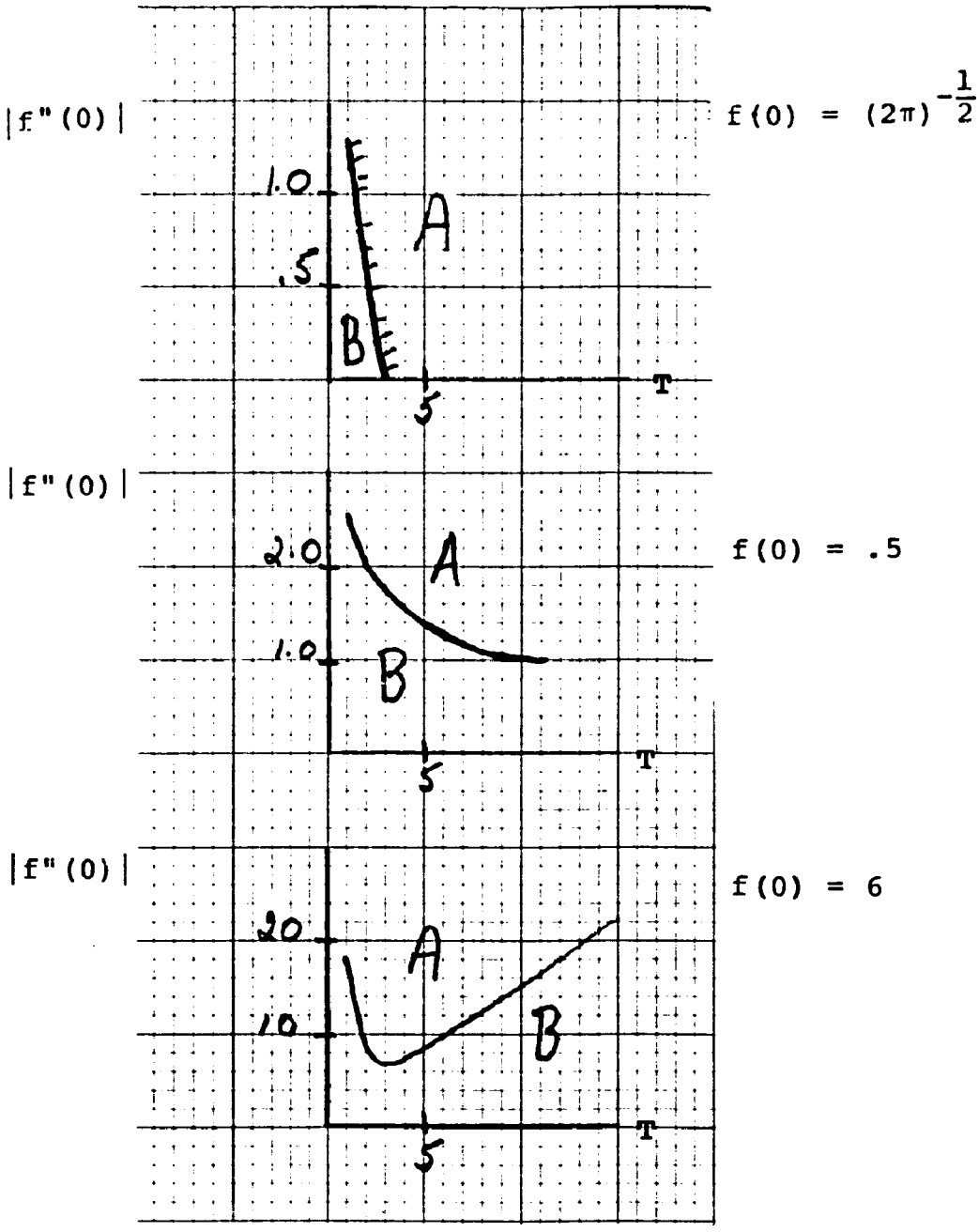
$$= \infty \qquad v \leq 2 \quad .$$

Substituting $f(0)^{-2}$, $f(0)^{-5}$, and $f''(0)$ into (17) yields:

$$\mathrm{Var}\beta^* = \frac{vB(\frac{1}{2},\frac{1}{2}v)^2}{4(T+2)} + \frac{(v^2+v)B(\frac{1}{2},\frac{1}{2}v)^4}{16(T+4)(T+2)} \quad .$$

In Figures 14, 15 and 16 this approximation is plotted and compared with $\mathrm{Var}(\hat{\beta})$.

The figures illustrate how beliefs concerning $F \epsilon F^*$ might be translated into a choice of $\beta^*$ or $\hat{\beta}$. We might characterize our beliefs by assigning probabilities to the parameter v of the Student distribution. For example, if probabilities




Var(u) = 1 and the approximation in (17) imply;

$$A = \{f(0), f''(0), T : Var(\hat{\beta}) < Var(\beta^*)\}$$

$$B = \{f(0), f''(0), T : Var(\beta^*) < Var(\hat{\beta})\}$$

Fig. 13.--An Approximate Partitioning of $F_2^*$

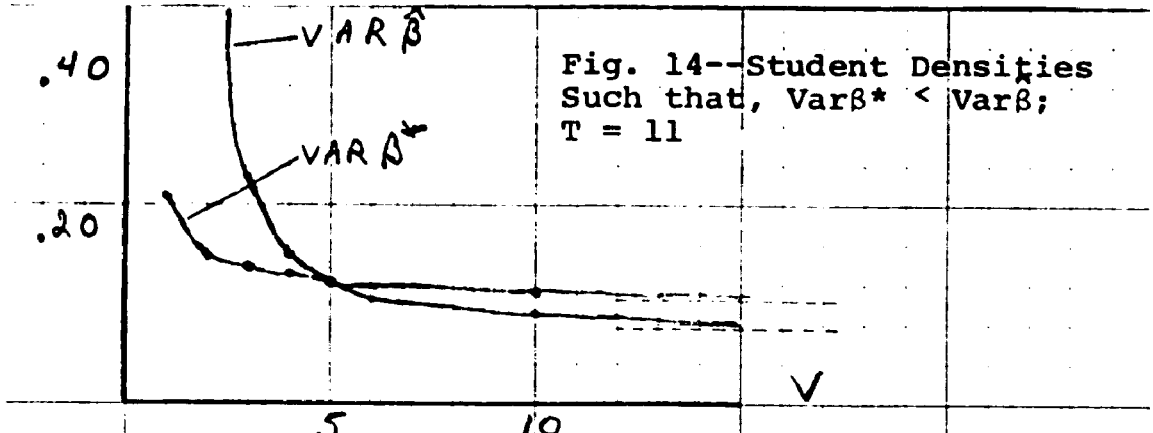



Fig. 14--Student Densities Such that, Var$\beta^*$ < Var$\beta$; T = 11

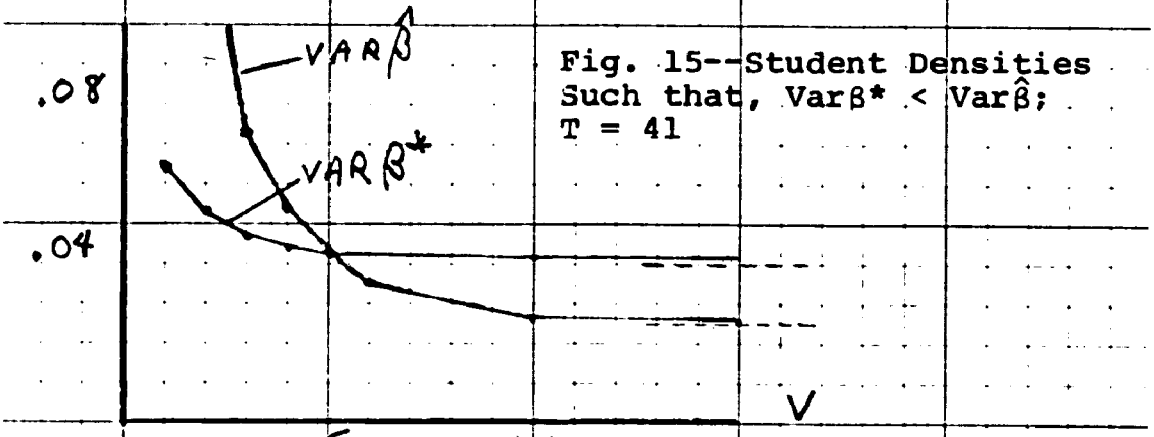



Fig. 15--Student Densities Such that, Var$\beta^*$ < Var$\hat{\beta}$; T = 41

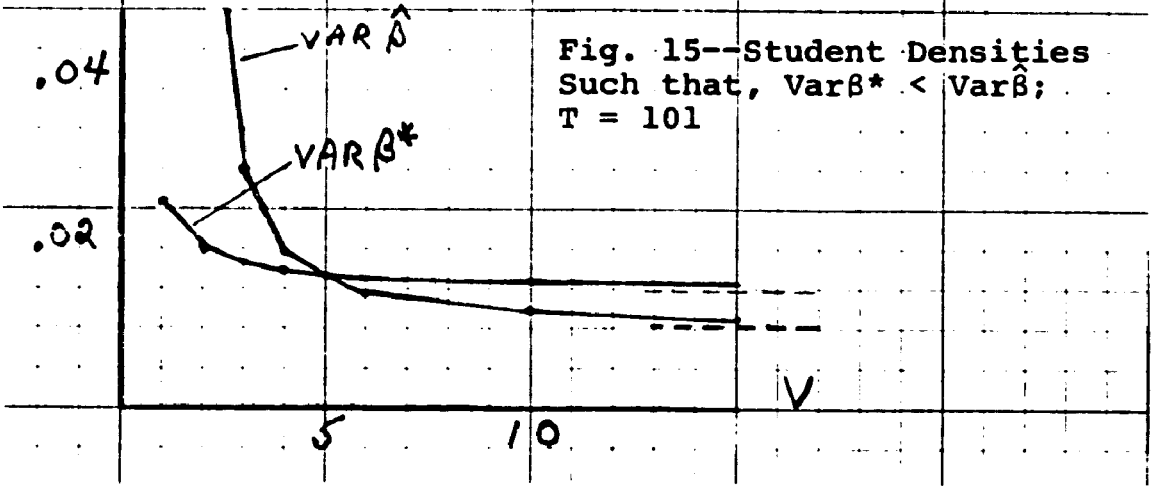



Fig. 15--Student Densities Such that, Var$\beta^*$ < Var$\hat{\beta}$; T = 101

are assigned so that $Pr[1 \leq v < 5] = 1$ then the preference for minimum variance implies $\beta^*$ is preferred to $\hat{\beta}$. Similarly, if $Pr[5 < v \leq \infty] = 1$ then $\hat{\beta}$ is preferred to $\beta^*$. In these cases one estimator dominates the other. (See Figure 17 for a plot of the Student densities.)

If probabilities are assigned to v in $(1,\infty)$ so that the probability that v is in any interval is nonzero then neither estimator dominates. However, it is difficult to think of any standard objective function which leads to $\hat{\beta}$ being preferred to $\beta^*$. A "minimize the maximum variance for v in $(1,\infty)$" rule never leads to $\hat{\beta}$ since for, $1 \leq v \leq 2$; $var(\hat{\beta}) = \infty$. Also, a "minimize the expected variance" rule does not lead to $\hat{\beta}$ since:

$$Pr(1 \leq v \leq 2) > 0$$

$$var(\hat{\beta};\ 1 \leq v \leq 2) \cdot Pr(1 \leq v \leq 2) = \infty \quad .$$

It appears that one must believe that v is almost certainly greater than 5, or alternatively, F is almost certainly normal for $\hat{\beta}$ to be preferred to $\beta^*$. The appeal of $\beta^*$ is that it is better than $\hat{\beta}$ for some v; it is not too much worse than $\hat{\beta}$ for some other v (for $v = \infty$, $T = \infty$, $Var(\beta^*)/Var(\hat{\beta}) = 1.57$); and most importantly, it is not as sensitive as $\hat{\beta}$ to v in $(1,\infty)$. For economic models, these properties seem desirable in view of our ignorance concerning $F \epsilon F^*$.

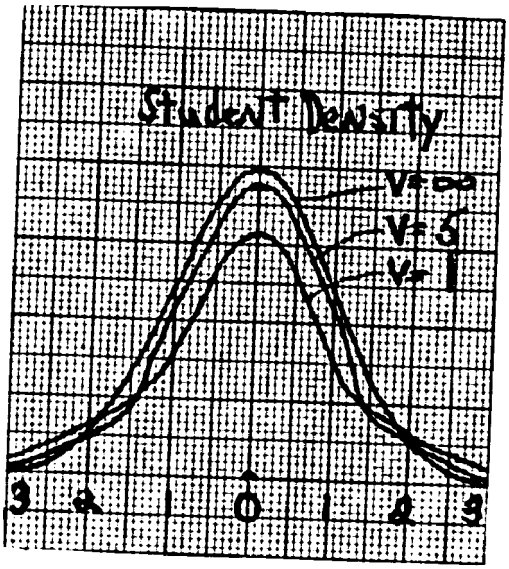



Fig. 17.--Student Densities

## 6. Conclusion

In this chapter we have shown how the sampling distribution of the median follows from the characterization of the median presented in Chapter IV. In the next chapter we show how this characterization extends to the multivariate L.A.E. estimator.

We have also indicated some of the properties of F such that $\mathrm{Var}(\beta^*) < \mathrm{Var}(\hat{\beta})$. Hopefully these properties aid in making the transition from beliefs concerning F to a choice of $\beta^*$ or $\hat{\beta}$. We might also expect, subject to certain qualifications (for example when $\beta^*$ is linear), that these properties may extend to the multivariate case.

We now examine the properties of the L.A.E. estimator for arbitrary X by using the methods outlined in this chapter.

## Footnotes

[1]From (4) it follows that $W_i \neq 0$ and so we have excluded zero in $w_i$.

[2]To my knowledge Theorem 1 was first proved by Rosenberg and Carlson (1972).

# CHAPTER VI

# THE SAMPLING DISTRIBUTION OF THE L.A.E. ESTIMATOR

In this chapter we derive the sampling density of the L.A.E. estimator. The derivation is a straightforward generalization of the methods used in obtaining the sampling density of the median.

The chapter is divided into five sections. Section 1 includes the derivation of the L.A.E. sampling density. In Section 2 we present some comments on the mathematical and computational problems involved in analyzing $\beta^*$ for arbitrary X. Section 3 includes some finite sampling properties of the L.A.E. estimator. In Section 4 we illustrate some requirements which are needed for $\beta^*$ to be a consistent estimator of $\beta$. Whether these requirements are sufficient for consistency is not known. A summary is presented in Section 5.

## 1. The L.A.E. Sampling Density

In this section we derive the probability density of the L.A.E. estimation error using the characterization of $(\beta^*-\beta)$ presented in Chapter IV. We begin by introducing the following notation.

Notation

We have previously used $f(\cdot)$ to represent the probability density of $u_t$. In the following we also use $f(\cdot)$ to denote the density of the K dimensional random variable $u(h_i)$. Here we use assumption A5 ($u_t$; t = 1,...,T; independent and identically distributed) so that $f(\cdot)$ does not depend on $h_i$. In a similar fashion we use $F(\cdot)$ to represent both the probability distribution function of $u_t$, and the (T-K) dimensional random variable $u(\bar{h}_i)$. Any ambiguity in this notation is resolved by specifying the argument in $f(\cdot)$ or $F(\cdot)$.

The probability element of f is defined by (see Cramer, 1946, p. 167);

$$(1) \qquad f(\phi)d\phi = \Pr[u\varepsilon\phi \pm \tfrac{1}{2}d\phi] \quad .$$

When u is a vector random variable, say $u(h_i)$, (1) is the probability that $u(h_i)$ falls in an infinitesimal region in $R^K$.

In the following we need an expression for the probability element of the K-dimensional random variable, $X(h_i)^{-1}u(h_i)$, and this is given by;

$$(3) \qquad |X(h_i)|f(X(h_i)\phi)d\phi = \Pr[X(h_i)^{-1}u(h_i)\varepsilon\phi \pm \tfrac{1}{2}d\phi] \quad .$$

We assume F in F* and in particular, $f(u) = f(-u)$. It follows that, $1 - F(u) = F(-u)$, and this along with assumption A5 gives the following identity;

$$(4) \qquad \Pr[I^*u(h_i) \geq I^*X(\bar{h}_i)\phi] = F(-I^*X(\bar{h}_i)\phi)$$

where $I^*$ is in $\bar{I}$ and $X(\bar{h}_i)$ is a $(T - K \times K)$ matrix.

The Density of $(\beta^*-\beta)$

Let $g(\phi)$ be the probability density of $(\beta^*-\beta)$. The probability element is then given by;

$$g(\phi)d\phi = Pr[(\beta^*-\beta)\varepsilon\phi \pm \tfrac{1}{2}d\theta] \quad . \tag{5}$$

From our characterization of $(\beta^*-\beta)$ in Chapter IV, the event;

$$(\beta^*-\beta)\varepsilon\phi \pm \tfrac{1}{2}d\phi \tag{6}$$

is equivalent to the events;

$$X(h_i)^{-1}u(h_i)\varepsilon\phi \pm \tfrac{1}{2}d\phi \quad \text{and} \quad u\varepsilon P(h_i) \quad \text{or} \quad -P(h_i) \tag{7}$$

for all $h_i \varepsilon H$. (7) can be viewed as merely a list of all the events (realizations of u in $R^T$) which lead to (6).

From the definition of the (mutually exclusive) sets $P(h_i)$ we can write the event $u\varepsilon(P(h_i)$ or $-P(h_i))$ as;

$$u\varepsilon(P(h_i,w_i) \text{ or } -P(h_i,w_i) \tag{8}$$

for all $w_i = 0,1,\ldots,W_i$. The index $w_i$ includes zero to allow for the possibility that $P(h_i)$ is empty; that is,

$$P(h_i,0) = \text{(empty)} \quad . \tag{9}$$

The sets $P(h_i,w_i)$ and $-P(h_i,w_i)$ for all $(h_i,w_i)$ partition $R^T$ and so the events in (8) are mutually exclusive. Therefore we can write;

$$(10) \qquad g(\phi)d\phi = \sum_{h_i}^{H} \sum_{w_i}^{W_i} Pr[X(h_i)^{-1}u(h_i)\varepsilon\phi \pm \tfrac{1}{2}d\phi \quad \text{and}$$

$$u\varepsilon P(h_i,w_i) \text{ or } u\varepsilon -P(h_i,w_i)] \quad .$$

Writing this joint probability as the product of marginal and conditional probabilities gives;

$$(11) \qquad g(\phi)d\phi = \sum_{h_i} \sum_{w_i} Pr(A)\{Pr[u\varepsilon P(h_i,w_i)|A] + Pr[u\varepsilon -P(h_i,w_i)|A]$$

where A is the event, $X(h_i)^{-1}u(h_i)\varepsilon\phi \pm \frac{1}{2}d\phi$. This is the general expression for the probability density of the L.A.E. estimation error. We now introduce our assumptions on u so that (11) can be expressed in terms of f, F, and X.

From (1) we have;

$$(12) \qquad Pr[A] = |X(h_i)|f(X(h_i)\phi)d\phi \quad .$$

For the first conditional probability in (11) we get;[1]

$$Pr[u\varepsilon P(h_i,w_i)|A] =$$

$$Pr[-I^*(h_i,w_i)X(\bar{h}_i)\phi + I^*(h_i,w_i)u(\bar{h}_i) \geq 0] =$$

$$Pr[I^*(h_i,w_i)u(\bar{h}_i) \geq I^*(h_i,w_i)X(\bar{h}_i)\phi]$$

and using (4) we have;

$$(13) \qquad F(-I^*(h_i,w_i)X(\bar{h}_i)\phi) \quad .$$

In a similar fashion we obtain;

$$(14) \qquad \Pr[u\varepsilon - P(h_i,w_i)\,|A] = F(I^*(h_i,w_i)X(\bar{h}_i)\phi) \quad .$$

Substituting (12), (13), and (14) into (11) then yields;

Theorem 1. Given assumptions A1 to A5 and F in F*, the probability density of the L.A.E. estimation error $(\beta^*-\beta)$ is;

$$(15) \qquad g(\phi) = \sum_{h_i} \sum_{w_i} |X(h_i)| f(X(h_i)\phi)\,[F(I^*(h_i,w_i)X(\bar{h}_i)\phi + F(-I^*(h_i,w_i)X(\bar{h}_i)\phi]$$

where $I^*(h_i,w_i)$ satisfies;

$$(16) \qquad -i < iI^*(h_i,w_i)X(\bar{h}_i)X(h_i)^{-1} < i \quad .$$

In order to illustrate Theorem 1 we present two simple examples.

Example 1. When $X = (1,\ldots,1)$ and T is odd we know that for any $h_i$, every $I^*(h_i,w_i)$ contains an equal number of +1 and -1 so that;

$$F(I^*(h_i,w_i)i\phi) = F(\phi)^n(1-F(\phi))^n = [F(\phi)F(-\phi)]^n \quad .$$

Also, $|X(h_i)| = 1$, and $f(X(h_i)\phi) = f(\phi)$. The indices $h_i,w_i$ are given by;

$$h_i = 1,\ldots,T$$

$$w_i = 0,\ldots,\frac{(T-1)!}{2n!n!} \quad .$$

Substituting into (15) then yields the median sampling density.

Example 2. Suppose $\beta^*$ is a linear function of Y so that;

$$\beta^* = X(h_1)^{-1}Y(h_1) \quad \text{all } Y \text{ in } R^T \quad .$$

From the discussion in Chapter IV, Section 6, we know that only $P(h_1)$ is nonempty and $P(h_1) \cup -P(h_1) = R^T$. We can therefore write (15) as;

$$(17) \qquad g(\phi) = |X(h_1)| f(X(h_1)\phi) \sum_{w_i} [(F(I^*(h_1,w_i)X(\bar{h}_1)\phi) + F(-I^*(h_1,w_i)X(\bar{h}_1)\phi)] \quad .$$

Using (13) we can write the summation in (17) as;

$$\Pr[u \varepsilon (P(h_1) \quad \text{or} \quad -P(h_1)) | A] =$$

$$\Pr[u \varepsilon R^T | A] = 1 \quad .$$

Hence,

$$g(\phi) = |X(h_1)| f(X(h_1)\phi)$$

and this is just the density of the random variable, $X(h_1)^{-1}u(h_1)$.

## 2. Some Mathematical Problems

Theorem 1 shows how one can, at least in principle, calculate the sampling density of the L.A.E. estimator. We say "in principle" because even with computers this calculation will probably be too costly. The primary difficulty is in the calculation of the $I^*(h_i,w_i)$ in (16). To find all the $I^*(h_i,w_i)$ requires an examination of every $I^*$ in $\bar{I}$ for every $h_i$. Since there are $\binom{T}{K}$ of the $h_i$ and $2^{T-K}$ elements in $\bar{I}$ it takes,

$$\frac{2^{T-K}T!}{K!(T-K)!}$$

steps to find all of the $I^*(h_i,w_i)$. For 5 explanatory variables and only 20 observations this requires more than 750 million steps. For practical purposes the search for $I^*(h_i,w_i)$ is an impossible task, unless there is some other method besides trying every possibility which can be employed.

The calculation of $I^*(h_i,w_i)$ also illustrates the one remaining mathematical problem which must be solved in order to obtain a sampling theory for the L.A.E. estimator. What is needed is an explicit relation for $\bar{I}(h_i)$ in terms of $X(h_i)$. In the special case of the median such an expression was available; namely,

(18) $\bar{I}(h_i) = \{I^* \epsilon \bar{I} : \text{the number of +1 and -1 are equal}\}$ .

This essentially allowed us to substitute (16) into (15) and thereby obtain an expression for $g(\phi)$ which did not depend on $\bar{I}(h_i)$. For arbitrary X an explicit relation like (18) is unknown and this severely restricts the general results which might be otherwise obtained from Theorem 1.

In analyzing the sampling properties of the L.A.E. estimator we therefore have to separate the consequences of Theorem 1 into two categories. The first includes those results which only depend on the general definition of I*; namely as a (T - K x T - K) diagonal matrix with +1 or -1

along the diagonal. Here we do not have to consider the particular I* which satisfy (16). The second category includes those results which depend on $I^*(h_i,w_i)$ and in this category are the results relating to the variance-covariance matrix of $\beta^*$. Here the only way to procede at present is by considering special cases of X. Again, this problem arises because we do not know how to characterize the general relationship between $I^*(h_i,w_i)$ and $X(h_i)$.

Two results in this study may however provide some insight into the solution to this problem. The first is Property 7, Chapter III, which showed that for models with a constant term;

$$|P - N| < K \tag{19}$$

where P and N are the number of positive and negative $u^*_t$, respectively. For I*, P is the number of +1 and N is the number of -1. This property might be used in the following manner.

Suppose we have the model

$$y_t = \beta_1 + \beta_2 x_t + u_t \qquad t = 1,\ldots,T$$

where T is an even number. Using (19) we get;

$$|P - N| < 2$$

or

$$|P - N| = 0$$

since P + N is an even number. Therefore, for this model we know that $I^*(h_i,w_i)$ has to contain an equal number of +1 and

-1. We therefore do not have to search over every $I^*$ in $\bar{I}$ for the solution to (16). (This case is not equivalent to the median since some $\bar{I}(h_i)$ may be empty.)

The second result which gives some insight into the determination of $I^*(h_i,w_i)$ is Theorem 8 in Chapter V, Section 5. We showed that the elements in $X(\bar{h}_i)X(h_i)^{-1}$ were;

$$\pm \frac{|X(h_j)|}{|X(h_i)|} \qquad i \neq j \quad .$$

These determinants therefore are the data which are used in checking whether $I^*$ satisfies (16). In Appendix II it can be seen that the determinants $|X(h_i)|$ continually recur in calculating $I^*$, and it may be that there is some method which can be used to find $I^*(h_i,w_i)$ without trying every $I^*$ in $\bar{I}$.

## 3. Finite Properties of $g(\phi)$

In this section we present some properties of $g(\phi)$ for $F$ in $F^*$. The results only depend on the general definition of $I^*$, and therefore they hold for any $X$.

1. Symmetry. $g(\phi) = g(-\phi)$. This follows from the symmetry of $f(\cdot)$, and the result is obtained by substituting $-\phi$ into (15). This result does not require an expression for $g(\phi)$ as it can be obtained from the linear homogeneity of $\beta^*(Y)$ (Property 2, Chapter II).

2. Unbiasedness. $E(\beta^*-\beta) = 0$ if the expectation exists. This follows directly from Property 1. The conditions for the

existence of the expectation depend on the particular X, and therefore we cannot obtain general results which are analogous to the case of the median.

3. Continuity. $g(\phi)$ is continuous. This follows from the assumed continuity of f and F along with the fact that the product of continuous functions is continuous and the sum of (a finite number of) continuous functions is continuous.

4. $g(0) = \sum_{h_i} |X(h_i)| f(0)^K \cdot W_i (\frac{1}{2})^{T-K-1} < \infty$. This follows from the assumptions; $F(0) = 1/2$ and $f(0) < \infty$.

5. Let $G(\phi)$ be the probability distribution of $(\beta^*-\beta)$. The above properties then imply $G(\phi)$ is in $F^*$.

## 4. Consistency of $\beta^*$

In this section we present some examples which illustrate the nature of the restrictions which have to be imposed for $\beta^*$ to be a consistent estimator of $\beta$. Whether these restrictions are sufficient for consistency is unknown. The primary difficulty in deriving asymptotic properties can be traced to the problem of characterizing the $I^*$ in $\bar{I}$ which satisfy (16).

### A Restriction on f(u)

Consider an f(u) with the following property;

(20) $$f(u) = 0 \quad \text{for} \quad u \in (-C, +C) \quad C > 0 \quad .$$

Note that probability distributions with this property are not contained in $F^*$. (20) implies;

$$\Pr[-ci < Y - X\beta < ci] = 0$$

where $ci = (c,\ldots,c)'$. This means that with probability zero none of the observations on $y_t$ are within some distance (determined by c) of the "true" relationship, $X\beta$. This is illustrated in Figure 18. The L.A.E. estimator is determined by the K observations corresponding to $h_i$, and therefore for any T.

$$\Pr[-ci < Y - X\beta^* < ci] = 0 \quad . \tag{21}$$

Consistency requires that for any $\delta, \varepsilon > 0$ there exist some T such that for $t > T$;

$$\Pr[|\beta_t^* - \beta| < \varepsilon] > 1 - \delta \quad . \tag{22}$$

It can now be seen that by choosing an $\varepsilon$ which satisfies (21) that (22) can never be satisfied, and therefore $\beta^*$ is not a consistent estimator of $\beta$.

This example shows we have to assume $f(u) > 0$ for u in some neighborhood of zero in order for $\beta^*$ to be consistent. This also holds for the special case of the median where the variance of the asymptotic normal distribution is given by, $(4f(0)^2T)^{-1}$; this being undefined when $f(0) = 0$. However, when $f(0) > 0$ the median is consistent, independently of any assumptions on the moments of u.

We might conjecture that a similar relation holds for the multivariate L.A.E. estimator (subject to certain other restrictions on X discussed below). This is in contrast to

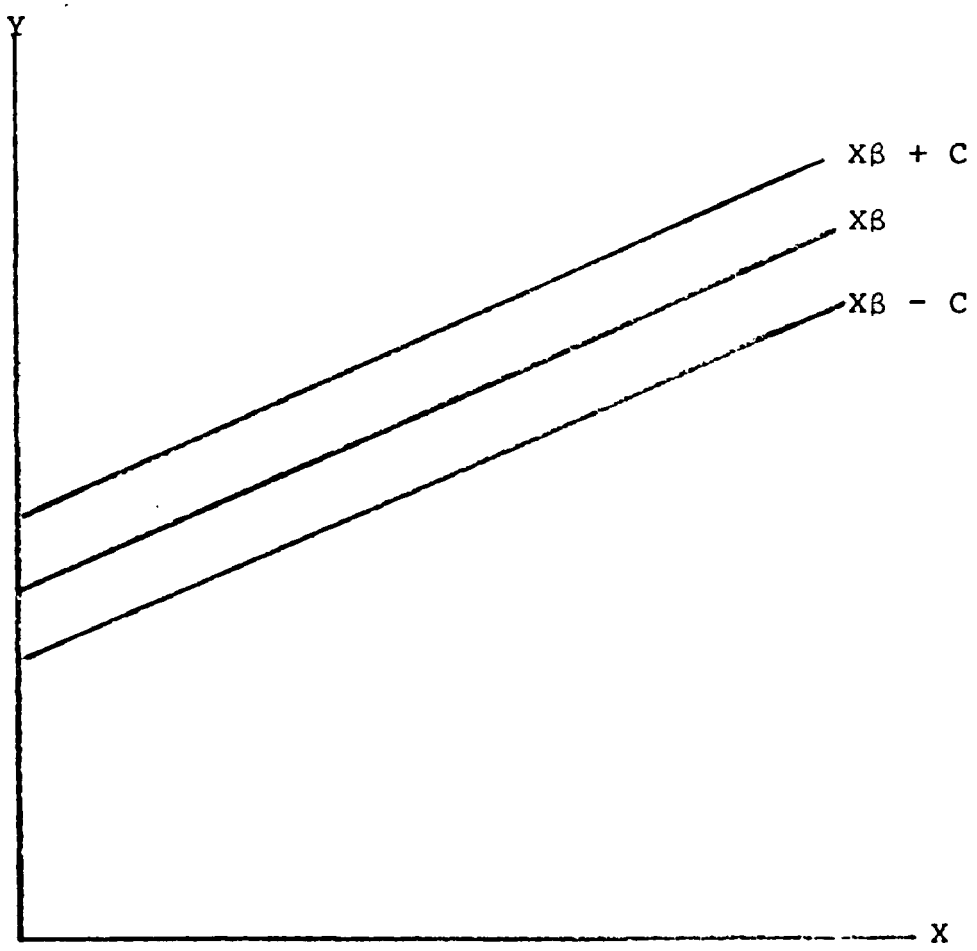



Fig. 18.--An Example Where β* is Inconsistent

the least squares estimator which is consistent when f(u) = 0, but which requires assumptions on the convergence of the moments of u; see Theil (1971, Chapter 8).

Restrictions on X

To show that certain restrictions must be put on X to insure consistency we illustrate with an example taken from Theil (1970, p. 364).

Consider the model;

$$y_t = \beta x_t + u_t \qquad t = 1,2,\ldots \tag{23}$$

where $x_1 = 3$ and $x_t = (1/2)^{t-1}$ for $t = 2,3,\ldots$. The variance of the least squares estimator for any T is given by;

$$\sigma^2\Big(\sum_{t=1}^{T} x_t^2\Big)^{-1} = \sigma^2\Big(9 + \sum_{t=2}^{T} \big(\tfrac{1}{4}\big)^{t-1}\Big)^{-1}$$

$$= \sigma^2\Big(9 + \frac{1-\big(\frac{1}{4}\big)^{T-1}}{12}\Big)^{-1}$$

$$= \sigma^2\Big(\frac{12}{109-\big(\frac{1}{4}\big)^{T-1}}\Big) \quad .$$

Therefore, the least squares estimator has a variance which is never less than $\sigma^2(12/109)$ for any T and is not consistent.

In this example we have;

$$\lim_{T\to\infty} T^{-1}(X'X) = \lim_{T\to\infty}\Big(\frac{9+\sum_{t=2}^{T}\big(\frac{1}{4}\big)^{t-1}}{T}\Big) = 0 \quad .$$

Theil uses the example to illustrate a case in which the sufficient condition;

$$\lim_{T\to\infty} T^{-1}(X'X) = A \qquad 0 < A < \infty \tag{24}$$

for least squares consistency is violated. We now show that some type of condition like (24) also has to be imposed for the L.A.E. estimator to be consistent.

We show that $\beta^*$ is not a consistent estimator given the model (23) by showing;

$$\beta^* = y_1/3 \quad \text{for all } T \quad .$$

Take $X(h_i) = x_1 = 3$ so that for Condition 1' we have;

$$-1 < iI^*X(\bar{h}_i)\tfrac{1}{3} < 1 \quad . \tag{25}$$

Denote the diagonal elements of $I^*$ by $e_t$, where $e_t = \pm 1$. Rewriting (25) then gives;

$$-1 < \tfrac{1}{3} \sum_{t=2}^{T} e_t (\tfrac{1}{2})^{t-1} < 1 \quad . \tag{26}$$

The expression $\sum_{t=2}^{T} e_t (\frac{1}{2})^{t-1}$ is maximized by setting $e_t = +1$ for all t, but then

$$-1 < \tfrac{1}{3} \sum_{t=2}^{T} (\tfrac{1}{2})^{T-1} < 1 \tag{27}$$

and (27) holds for all T. Therefore, any combination of $e_t = +1$ and $e_t = -1$ satisfies (26), and every $I^*$ in $\bar{I}$ satisfies Condition 1'. We can then always find an $I^*$ which satisfies

$$-I^*X(\bar{h}_i)(\tfrac{1}{3})(y_1) + I^*Y(\bar{h}_i) \geq 0 \quad .$$

Hence,

$$\beta^* = y_1/3 \quad \text{for all } T.$$

Theil's intuitive reason for the inconsistency of the least squares estimator for the model (23) is that the values of $x_t$ "converge to zero so rapidly that from a certain point onward, the contribution of all later observations to precision is negligible." A similar interpretation holds for the L.A.E. estimator. The values of $x_t$ converge to zero so rapidly that after the first observation, the contribution of all later observations to precision is zero.

Note that for (23) the L.A.E. estimator is a linear function for all T, and this leads to the inconsistency of $\beta^*$. In proving consistency we have to impose (at least) some restrictions on X so that $\beta^*$ will be nonlinear. In Chapter IV, Section 6, we presented a corollary which stated that for models containing a constant $\beta^*$ is nonlinear for all $T > 2K + 1$. In these cases then the nonlinearity restriction is satisfied. Whether this alone is sufficient for consistency is not known at present.

## 5. Summary

In this chapter we have derived the sampling density of $(\beta^*-\beta)$. However, few general results were obtained from this density because of the complex relationship between $\beta^*$ and X. We could only get results which do not depend on X, or by considering particular X show when certain properties failed to hold. In the next chapter we offer some suggestions which might be used in obtaining further consequences from Theorem 1.

## Footnotes

[1]Here we have used the identity;

$$\Pr[u_1 > u_2 | u_2 \varepsilon \phi \pm \tfrac{1}{2} d\phi] = \Pr[u_1 > \phi] \quad .$$

# CHAPTER VII

## SUMMARY AND SUGGESTIONS FOR FUTURE RESEARCH

In this chapter we compare our findings to previous studies of the L.A.E. estimator. We also examine the ways in which L.A.E. might be used as an alternative to the least squares estimator given the decision problem posed in Chapter I.

The chapter is divided into four sections. In Section 1 the properties of the L.A.E. estimator are summarized. Section 2 includes an explanation of the Monte Carlo results obtained by Taylor (1970) and Rosenberg and Carlson (1972). In Section 3 we indicate some directions which might be taken in obtaining further properties of the L.A.E. estimator. Finally, in Section 4 we offer some conjectures on the ways in which the L.A.E. estimator might be used when there is uncertainty concerning F(u).

### 1. Summary of L.A.E. Properties

Chapters II and III include properties which follow directly from the definition of the L.A.E. estimator. Most have been derived previously (some date back at least to Gauss; see footnote 7, Chapter III), although there is no single statement of all of the properties in the literature.

For Property 1 we used the well-known and obvious result in which the median is not unique for a sample of size two to show that the solution to the L.A.E. minimization problem is not necessarily unique. Previous studies have noted this possibility, but no one has analyzed the conditions under which unique solutions occur. Finding these conditions would seem to be a trivial first step in analyzing L.A.E.; otherwise we do not even know whether we are dealing with well-defined estimators.

Property 2 establishes that $\beta^*(Y)$ is a linear homogeneous function of Y; Property 3 shows that the L.A.E estimation error $(\beta^*-\beta)$ is independent of $\beta$; and Property 4 shows that a linear transformation of the explanatory variables (that is, XA) results in the same linear transformation of the L.A.E. estimator. These properties are all shared by the least squares estimator. Rosenberg and Carlson (1972) have most recently derived these properties by focusing on the objective function, $\Sigma|V|$, in a neighborhood of the minimum value. We derived the properties by using the feasibility and optimality requirements which define the solution to the L.A.E. minimization problem. This parallels our approach in Chapter IV where we apply the feasibility and optimality conditions to the L.A.E. linear programming problem.

In Property 5 we showed that $\beta^*(Y)$ could not be classified as a linear or nonlinear estimator independently of X. This property gives the first hint at the way in which X might affect the sampling properties of $\beta^*$, and also raises the

possibility of rejecting the L.A.E. estimator via the Gauss-Markov Theorem. In all of the previous studies with which we are aware it has been either explicitly or implicitly stated that L.A.E. is a nonlinear estimator. This conclusion has probably resulted from the known nonlinearity of the median along with the belief that a nonlinear estimator for some X must be nonlinear for all X. Property 5 shows that this conclusion is invalid.

Properties 6 and 7 in Chapter III follow directly from the result due originally to Charnes, Cooper and Ferguson (1955) in which the L.A.E. estimator is redefined as the solution to a linear programming problem. These properties, however, are not recent and were known by Gauss in the early nineteenth century.

Our derivation of Property 8 follows the original derivation by Appa and Smith (1971). The property shows one way in which the properties of the median extend to the multivariate L.A.E. estimator (for models containing a constant). The precise relationship between the L.A.E. estimator and its special case, the median, however, is left unanswered. That is, exactly what properties of the median extend to the multivariate L.A.E. estimator? An answer to this question would allow us to apply the intuition and known results of the median to the L.A.E. estimator.

The results in Chapters II and III can be viewed as falling into two categories according to whether the properties do or do not depend on X. The latter category includes

Properties 2, 3, 4, 6 and 7 and most of the previous studies of L.A.E. have focused on these properties. Since the properties do not depend on X, they can be derived without explicitly characterizing $\beta^*$ in terms of Y and X, and it is possible that these exhaust all of the nontrivial properties in this category.

Properties 1, 5 and 8 fall into the first category since they depend on X. Whether L.A.E. is well-defined or linear or shares a median property depends on the particular X under consideration. The Monte Carlo findings of Taylor (1970) and Rosenberg and Carlson (1972) also fall into this category since their unexplained results depend on X. An understanding of these properties and results requires some representation of $\beta^*$ in terms of Y and X, and this was obtained in Chapter IV.

In Chapter IV we derived an explicit characterization of $\beta^*$ in terms of Y and X. This was accomplished by essentially solving the L.A.E. linear programming problem using a variant of the simplex method. To apply the simplex method we had to make the basic solution a variable, and this was accomplished by introducing $I^*$. The feasibility and optimality conditions for the linear programming problem provided the two conditions which we used to obtain further properties.[1]

In Theorem 6 we applied the uniqueness conditions for linear programs to obtain the uniqueness conditions for $\beta^*$. The hypothesis of the theorem only depends on X so that when L.A.E. is well-defined for some X it is well defined for any

F(u) (or any domain on Y). However, in the accompanying example 2 and in Appendix III, we showed that restrictions on the domain of Y (for example, $\Sigma y_t = 0$) could lead to uniqueness even when the hypothesis of Theorem 6 is violated.

Using the uniqueness conditions we were also able to represent the L.A.E. estimator as a linear function defined over certain sets (convex cones) which partition the sample space $R^T$. The partitioning of $R^T$ only depended upon X and this provided the conditions under which $\beta^*$ is linear. That is, when $R^T$ is partitioned into the single set $R^T$ then $\beta^*$ is a linear estimator. Some of the consequences of this were examined in the corrollaries to Theorem 7. We obtained Corollary 2 which showed that for models containing a constant and $T > 2K + 1$, $\beta^*$ is nonlinear so that the Gauss-Markov Theorem cannot be used to reject the L.A.E. estimator.

The representation of the L.A.E. estimator as a linear function defined over certain sets also applies to the median, and this provided an explanation of the relationship between the median and the general L.A.E. estimator.

In Chapter V we presented a variety of results on the median. Most of these results are well-known. We attempted to provide some intuition concerning those F(u) such that L.A.E. is better than (where "better" is in terms of the preferences in Chapter 1) least squares by comparing the median and mean.

In Chapter VI we derived the sampling density of $\beta^*$, and presented some cases in which $\beta^*$ is not a consistent estimator.

We also showed that $\beta^*$ is unbiased for F(u) in F*, although this result follows directly from Properties 2 and 3 and does not require the results in Chapter VI.

## 2. Explanation of Monte Carlo Results

In his Monte Carlo experiments, Taylor found that certain sets of K observations never determined the L.A.E. estimate. He conjectured that this was not a sampling result, but instead would hold for any Y. In Appendix II we applied Condition 1 to the explanatory variables used by Taylor to show that his conjecture was correct.[2] The reason certain observations never determine the L.A.E. estimate can be found in the representation of $\beta^*$ in Chapter IV. That is, $\beta^* \neq X(h_i)^{-1}Y(h_i)$ when the set, $P(h_i) \cup - P(h_i)$, is empty. In Chapter IV, we also showed that in the limiting case where $P(h_i) \cup - P(h_i)$ is empty for all but one $h_i$, the L.A.E. estimator is linear.

As we indicated in Chapter III, Rosenberg and Carlson conjectured;

$$\text{(1)} \qquad \text{Var}(\beta^*) = \lambda(F,T)(X'X)^{-1}$$

where $\lambda/T$ is the variance of the median given F. When $X = (1,\ldots,1)$ the L.A.E. estimator is the median and the conjecture is correct. The question is whether (1) extends in some approximate way to arbitrary X. Rosenberg and Carlson found that (1) worked well except when the values of the explanatory variables were widely dispersed. We now present an example of extremely dispersed $x_t$ which shows how (1) can fail to be an adequate approximation.

Suppose we have the model;

$$y_t = x_t\beta + u_t \quad t = 1,\ldots,T \quad .$$

Order the observations so that;

$$|x_1| \geq |x_2| \geq \ldots \geq |x_T| \quad .$$

and assume that;

$$|x_1| > \sum_{t=2}^{T} |x_t| \quad . \tag{2}$$

We take this as a model of extremely dispersed $x_t$. We now show, $\beta^* = y_1/x_1$ for all Y.

Let $X(h_i) = x_1$ and let the diagonal elements of $I^*$ be denoted by $e_t$. For Condition 1' we have;

$$-1 < iI^*X(\bar{h}_i)X(h_i)^{-1} < 1 \tag{3}$$

or,

$$-1 < \frac{\sum_{t=2}^{T} e_t x_t}{x_1} < 1$$

and finally,

$$\left|\sum_{t=2}^{T} e_t x_t\right| < |x_1| \quad . \tag{4}$$

Now using (2) we get;

$$|\Sigma e_t x_t| \leq \Sigma|e_t x_t| = \sum_{t=2}^{T} |x_t| < |x_1| \quad .$$

Therefore (3) is satisfied by any $e_t = \pm 1$ and every $I^*$ in $\bar{I}$ satisfies Condition 1'. Hence $\beta^* = y_1/x_1$ for all Y.

Here we have a case in which extremely dispersed $x_t$ lead to $\beta^*$ being a linear function, so that

$$\text{Var}(\beta^*) = \sigma^2 (x_1^2)^{-1} \quad .$$

The use of (1) in this case will therefore lead to poor results. For example, if F(u) is a distribution such that;

$$\text{Var (median)} < \text{Var (mean)}$$

then the use of (1) would lead one to incorrectly conclude;

$$\text{Var}(\beta^*) < \text{Var}(\hat{\beta}) \quad .$$

## 3. The Relationship Between $\beta^*$ and X

The above example along with the results in Chapters IV and VI illustrate what is perhaps our most important finding; namely the crucial role that X plays in determining the sampling properties of the L.A.E. estimator. This is in contrast to most of the Monte Carlo work which has focused on the sampling properties of $\beta^*$ for a restricted number of X. For example, in a recent Monte Carlo study, Smith and Hall (1972) take the Laplace distribution and ask whether the maximum likelihood estimator (L.A.E.) has smaller variance than the best linear unbiased estimator (least squares). (They find that L.A.E. has smaller sample variance.) From our investigation of the L.A.E. estimator we know that the question posed by Smith and Hall does not have an answer independently of X. That is, if X = (1,1,1) then from Chapter V we know;

$$\frac{\text{Var}(\beta^*)}{\text{Var}(\hat{\beta})} = \frac{23}{24} \quad .$$

But from the above example we know that if X = (3,1,1) then;

$$\frac{\text{Var}(\beta^*)}{\text{Var}(\hat{\beta})} = \frac{11}{9} \quad .$$

This suggests that in future Monte Carlo investigations one should take some F(u) and examine the properties of $\beta^*$ under a wide range of alternative X.

The problems which arise in analyzing the relationship between $\beta^*$ and X are in some respects analogous to the problems faced by Durbin and Watson (1950). They obtained a statistic, d, to test for first-order autocorrelation of the $u_t$. The statistic d however depended on X in a complex fashion, and so instead of calculating the sampling density for every X they calculated the sampling density of $d_U$ and $d_L$ so that for any X;

$$d_L < d < d_U \quad .$$

It is not known whether a similar method can be used for the L.A.E. estimator. The problem is that we do not know the X which might lead to upper and lower bounds on the distribution of $\beta^*$. If we did have this information, however, we could use Theorem 1 in Chapter VI to calculate the sampling density of $\beta^*$ for the bounding cases, and thereby avoid the relationship between $\beta^*$ and X.

In a sense we have adopted the methods used by Durbin and Watson by focusing on two extreme cases; namely, $\beta^*$ as a linear estimator and $\beta^*$ as the median. These two cases lead to widely different sampling properties, but whether they are the bounding cases is unknown.

## 4. The Use of the L.A.E. Estimator

In Chapter I we attempted to motivate the study of alternatives to least squares by introducing a decision problem in which there is uncertainty concerning F(u). In particular, we considered the possibility that for economic models F(u) is more peaked and fatter tailed than the normal distribution, and on an intuitive basis this led to the study of the L.A.E. estimator. We also adopted a lexicographic preference ordering over unbiasedness and minimum variance so that the L.A.E. and least squares estimators could be compared. Within this framework the choice of an estimator depends on beliefs concerning F(u) (as well as X; see discussion below), and a solution to the decision problem would allow one to translate beliefs on F(u) into a choice of the L.A.E. vis-a-vis the least squares estimator.

In an attempt to solve this decision problem, we have essentially analyzed the validity of the following statements:

S1. The solution to the decision problem only depends on the beliefs concerning F(u).

S2. Therefore, $\text{Var}(\beta^*) < \text{Var}(\hat{\beta})$ and $\beta^*$ is preferred to $\hat{\beta}$ if Var(median) < Var(mean).

If statement S1 was valid we could ignore X and use the comparison of the median and mean as a way of deciding between the L.A.E. and least squares estimators. In previous studies statements S1 and S2 have been implicitly analyzed as evidenced by the attempts to extend various properties of the median to the multivariate L.A.E. estimator.

Based on the above examples in which $\beta^*$ is linear, we know that statements S1 and S2 are in general false. This seems to be a surprising result since it is not at all obvious from the formulation of the decision problem that the particular X under consideration should play any role whatsoever in determining the choice of L.A.E. vis-a-vis least squares.

The reader should note that our only counterexamples to statements S1 and S2 have involved cases in which $\beta^*$ is a linear estimator (such a case was also used to illustrate a necessary condition for the consistency of $\beta^*$). This suggests that statements S1 and S2 might be modified in the following manner:

S1'. When $\beta^*$ is a nonlinear estimator, the solution to the decision problem only depends on the beliefs concerning F(u).

S2'. Therefore (when $\beta^*$ is nonlinear), $\text{Var}(\beta^*) < \text{Var}(\hat{\beta})$ and $\beta^*$ is preferred to $\hat{\beta}$ if Var(median) < Var(mean).

Here we restrict the domain of $\beta^*$ so that it is nonlinear. From Theorem 8 in Chapter IV we know that this is equivalent to a restriction in the domain of X, and furthermore for

models containing a constant and $T > 2K + 1$ the restriction is vacuous since $\beta^*$ is nonlinear. Whether statements S1' and S2' are valid, however, is unknown. The validity essentially involves the question of whether $Var(\beta^*)$ is a discontinuous function of $X^3$.

As we have already indicated, Smith and Hall found that the L.A.E. estimator had smaller (Monte Carlo) sampling variance than the least squares estimator for an F(u) (the Laplace distribution) such that Var(Median) < Var(mean). This provides support for the validity of statements S1' and S2', and suggests that at least as a rough first approximation the comparison of the median and mean is relevant for the choice between the L.A.E. and least squares estimators. In Chapter V we showed that in terms of our assumed preference ordering the median was significantly better than the mean for a range of F(u) in F* and not too much worse than the mean for some other F(u) in F*. In view of the uncertainty concerning F(u) in economic models, this provides some support for the use of the L.A.E. estimator,[4] and suggests further research into the properties of the L.A.E. estimator as an alternative to the least squares estimator.

## Footnotes

[1]By analyzing $\Sigma|V_t|$ in a neighborhood of the minimum value, Rosenberg and Carlson (1972) obtained in their Lemma 4 an equivalent statement of Conditions 1 and 2. They do not, however, use their Lemma to analyze the conditions under which $\beta^*$ is linear or well-defined. They also fail to realize that Conditions 1 and 2 provided an explanation of Taylor's Monte Carlo results.

[2]In a recent letter to Professor Taylor, G. Appa shows that the conjecture is correct since certain bases cannot satisfy the constraints in the dual L.A.E. problem.

[3]That is, if $\beta^*(Y:X_1)$ is a linear L.A.E. estimator, $\beta^*(Y:X_2)$ is nonlinear and $X_1$ is "close to" $X_2$ (here we view the (TxK) matrix X as a vector in $R^{TxK}$) then is $Var(\beta^*(Y:X_1))$ "close to" $Var(\beta^*(Y:X_2))$?

[4]The reader will note that we have been almost solely concerned with estimation problems. The related areas of inference and data reporting generate other problems as well as potential applications for the L.A.E. estimator, but these areas are beyond the scope of the present study. However, we briefly point to one potential application which might be justified by the low computational costs of L.A.E.

The least squares estimates provide all of the relevant sample information contained in the sample when F(u) is the normal distribution (that is, the least squares estimates are sufficient), but when F(u) is not normal the least squares estimates hide sample information. Strictly, the only way to report data when there is uncertainty concerning F(u) is to actually report all of the sample data and this is usually infeasible (without some method of organizing the data it is also probably useless). A potential solution to this problem is to report both the least squares and L.A.E. estimates. This might be viewed as a way of conveying sample information conditional on the normal distribution and conditional on distributions which are more peaked and fatter tailed than the normal. The low computational costs associated with the L.A.E. estimator make this a feasible method of conveying sample information.

## APPENDIX I

In this appendix we show how Theorem 2 and Property 6 in Chapter III follow from the solution to the L.A.E. linear programming problem.

The solution to the dual L.A.E. problem can be read directly from the first row of (11) in Chapter IV. This duality result is discussed in Spivey and Thrall (1970, Chapter 5). That is, if,

$$\beta^* = X(h_i)^{-1}Y(h_i)$$

then

(1) $$d^*(h_i) = iI^*X(\bar{h}_i)X(h_i)^{-1}$$

(2) $$d^*(\bar{h}_i) = I^*i \quad .$$

Since the diagonal elements in $I^*$ indicate the sign of $u_t^*$ we have for (2);

$$d_t^* = 1 \quad \text{iff} \quad u_t^* > 0$$

$$d_t^* = -1 \text{ iff} \quad u_t^* < 0 \quad .$$

From the optimality condition we know that;

$$-i < iI^*X(\bar{h}_i)X(h_i)^{-1} < i$$

so that for the $d(h_i)^*$ in (1) we get;

$$-1 < d_t^* < 1 \quad \text{iff} \quad u_t^* = 0 \quad .$$

Here we see the general result in which the optimality condition for the (primal) L.A.E. linear programming problem is the feasibility condition for the dual problem.

Property 6. $\beta^*(Y) = \beta^*(Y')$ where $Y' = X\beta^*(Y) + Zu^*(Y)$ and $Z \geq 0$.

Proof: We first write $Y'$ in terms of $Y$ using the characterization of $\beta^*(Y)$;

$$\beta^*(Y) = X(h_i)^{-1}Y(h_i) \quad \text{if} \tag{3}$$

$$-I^*X(\bar{h}_i)X(h_i)^{-1}Y(h_i) + I^*Y(\bar{h}_i) \geq 0 \tag{4}$$

where $I^*$ satisfies Condition 1'. We get;

$$Y' = X\beta^*(Y) + Z(Y - X\beta^*(Y))$$

$$= ZY + X\beta^* - ZX\beta^* \quad .$$

Using the same $h_i$ as in (3) we get;

$$\begin{bmatrix} Y'(h_i) \\ Y'(\bar{h}_i) \end{bmatrix} = \begin{bmatrix} Z(h_i)Y(h_i) \\ Z(\bar{h}_i)Y(\bar{h}_i) \end{bmatrix} + \begin{bmatrix} Y(h_i) \\ X(\bar{h}_i)X(h_i)^{-1}Y(h_i) \end{bmatrix} - \begin{bmatrix} Z(h_i)Y(h_i) \\ Z(\bar{h}_i)X(\bar{h}_i)X(h_i)^{-1}Y(h_i) \end{bmatrix}$$

$$= \begin{bmatrix} Y(h_i) + 0 \\ Z(\bar{h}_i)Y(\bar{h}_i) + X(\bar{h}_i)X(h_i)^{-1}Y(h_i) \end{bmatrix} \begin{matrix} \\ \\ 0 \\ -Z(\bar{h}_i)X(\bar{h}_i)X(h_i)^{-1}Y(h_i) \end{matrix}$$

Now using the same I* as in (4) we obtain for Condition 2;

$$-I^*X(\bar{h}_i)X(h_i)^{-1}Y(h_i) + I^*Z(\bar{h}_i)Y(\bar{h}_i) + I^*X(\bar{h}_i)X(h_i)^{-1}Y(h_i)$$

$$- I^*Z(\bar{h}_i)X(\bar{h}_i)X(h_i)^{-1}Y(h_i) = Z(\bar{h}_i)[-I^*X(\bar{h}_i)X(h_i)^{-1}Y(h_i)$$

$$+ I^*(Y(h_i))] \geq 0 \quad .$$

The last step follows from (4) and the fact that all the elements in the diagonal matrix Z are greater than or equal to zero. Hence;

$$\beta^*(Y') = X(h_i)^{-1}Y(h_i) = \beta^*(Y) \quad .$$

## APPENDIX II

The model studied by Taylor (1970) was;

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + u_t \quad t = 1,\ldots,5$$

where,

| Observation | $X_1$ | $X_2$ |
|---|---|---|
| 1 | 6 | 1 |
| 2 | 2 | 7 |
| 3 | 3 | 4 |
| 4 | 7 | 5 |
| 5 | 2 | 3 |

There are $\binom{5}{2} = 10$ elements in H. Let

$h_1 = (1,2)$ $\quad$ $h_6 = (2,4)$

$h_2 = (1,3)$ $\quad$ $h_7 = (2,5)$

$h_3 = (1,4)$ $\quad$ $h_8 = (3,4)$

$h_4 = (1,5)$ $\quad$ $h_9 = (3,5)$

$h_5 = (2,3)$ $\quad$ $h_{10} = (4,5)$

I* is a (3 x 3) diagonal matrix with +1 or -1 along the diagonal. There are eight elements in $\bar{I}$ and four elements such that the first diagonal element is +1. That is,

$$iI^* = +1, +1, +1$$
$$+1, +1, -1$$
$$+1, -1, -1$$
$$+1, -1, +1$$

If one of these I* satisfy Condition 1' then so does -I*. We now solve for the I* satisfying Condition 1'.

$h_1 = (1,2)$

$$X(\bar{h}_1)X(h_i)^{-1} = \frac{1}{40}\begin{bmatrix} 13 & 21 \\ 39 & 23 \\ 8 & 16 \end{bmatrix}$$

| iI* | $iI^*X(\bar{h}_1)X(h_1)^{-1}$ | satisfy $-i < iI^*X(\bar{h}_1)X(h_1) < i$ |
|---|---|---|
| +++ | 60/40 , 60/40 | no |
| ++- | 44/40 , 28/40 | no |
| +-- | -34/40 , -18/40 | yes |
| +-+ | -18/40 , 14/40 | yes |

In a similar fashion we obtain;

$h_2$

$$X(\bar{h}_2)X(h_2)^{-1} = \frac{1}{21}\begin{bmatrix} -13 & 40 \\ 13 & 23 \\ -1 & 16 \end{bmatrix}$$

| iI* | satisfy Condition 1' |
|---|---|
| +++ | no |
| ++- | no |
| +-- | no |
| +-+ | no |

$h_3$

$$X(\bar{h}_3)X(h_3)^{-1} = \frac{1}{23}\begin{bmatrix} -39 & 40 \\ -13 & 21 \\ -11 & 16 \end{bmatrix}$$

| iI* | satisfy Condition 1' |
|---|---|
| +++ | no |
| ++- | no |
| +-- | yes |
| +-+ | no |

$h_4$

$$X(\bar{h}_4)X(h_4)^{-1} = \frac{1}{16}\begin{bmatrix} -8 & 40 \\ 1 & 21 \\ 11 & 23 \end{bmatrix}$$

| iI* | satisfy Condition 1' |
|---|---|
| +++ | no |
| ++- | no |
| +-- | no |
| +-+ | no |

$h_5$

$$X(\bar{h}_5)X(h_5)^{-1} = \frac{1}{13}\begin{bmatrix} -21 & 40 \\ -13 & 39 \\ 1 & 8 \end{bmatrix}$$

| iI* | satisfy Condition 1' |
|---|---|
| +++ | no |
| ++- | no |
| +-- | yes |
| +-+ | yes |

$h_6$

$$X(\bar{h}_6)X(h_6)^{-1} = \frac{1}{39}\begin{bmatrix} -23 & 40 \\ 13 & 13 \\ 11 & 8 \end{bmatrix}$$

| iI* | satisfy Condition 1' |
|---|---|
| +++ | no |
| ++- | no |
| +-- | yes |
| +-+ | yes |

$h_7$

$$X(\bar{h}_7)X(h_7)^{-1} = \frac{1}{8}\begin{bmatrix} -16 & 40 \\ -1 & 13 \\ -11 & 39 \end{bmatrix}$$

| iI* | satisfy Condition 1' |
|---|---|
| +++ | no |
| ++- | no |
| +-- | no |
| +-+ | no |

$h_8$

$$X(\bar{h}_8)X(h_8)^{-1} = \frac{1}{13}\begin{bmatrix} -23 & 21 \\ 39 & -13 \\ 11 & -1 \end{bmatrix}$$

| iI* | satisfy Condition 1' |
|---|---|
| +++ | no |
| ++- | yes |
| +-- | no |
| +-+ | no |

$h_9$

$$X(\bar{h}_9)X(h_9)^{-1} = \begin{bmatrix} 16 & -21 \\ -8 & 13 \\ 11 & -13 \end{bmatrix}$$

| iI* | satisfy Condition 1' |
|---|---|
| +++ | no |
| ++- | no |
| +-- | no |
| +-+ | no |

$h_{10}$

$$X(\bar{h}_{10})X(h_{10})^{-1} = \frac{1}{11}\begin{bmatrix} 16 & -23 \\ -8 & 39 \\ 1 & 13 \end{bmatrix}$$

| iI* | satisfy Condition 1' |
|---|---|
| +++ | no |
| ++- | yes |
| +-- | no |
| +-+ | no |

Since for $h_2$, $h_4$, $h_7$, and $h_9$ there are no I* which satisfy Condition 1' we have;

$$P(h_2) = \phi$$
$$P(h_4) = \phi$$
$$P(h_7) = \phi$$
$$P(h_9) = \phi$$

where $\phi$ is the empty set. These sets of observations never determine the L.A.E. estimator.

Also notice that the elements in $X(\bar{h}_i)X(h_i)^{-1}$ are of the form;

$$\pm \frac{|X(h_j)|}{|X(h_i)|} \qquad i \neq j$$

## APPENDIX III

In this appendix we illustrate the result in Chapter IV, Section 3, example 2. We first examine a case in which $\sum_{t=1}^{T} x_{tk} = 0$ and $\sum_{t=1}^{T} y_t \neq 0$ and show that the solution to the L.A.E. linear programming problem is not unique. We then show that uniqueness results when $\Sigma y_t = 0$.

Suppose we are given the following observations;

| Observation | $y$ | $x_1$ | $x_2$ |
|---|---|---|---|
| 1 | 0 | 2 | 1 |
| 2 | -2 | -4 | -6 |
| 3 | 2 | -1 | +1 |
| 4 | 1 | 3 | 4 |

For $h_1 = (1,2)$ we have for Condition 1;

$$i \begin{bmatrix} +1 & 0 \\ 0 & +1 \end{bmatrix} \begin{bmatrix} -2 & 5 \\ 10 & 3 \end{bmatrix} - \frac{1}{8} = (-1,-1) \quad .$$

And for Condition 2;

$$- \begin{bmatrix} +1 & 0 \\ 0 & +1 \end{bmatrix} \begin{bmatrix} -2 & 5 \\ 10 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ -2 \end{bmatrix} \frac{-1}{8} + \begin{bmatrix} +1 & 0 \\ 0 & +1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{matrix} 3/4 \\ 1/4 \end{matrix} \geq \begin{matrix} 0 \\ 0 \end{matrix} \quad .$$

Therefore,

$$\beta^* = \begin{bmatrix} 2 & 1 \\ -4 & -6 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ -2 \end{bmatrix} = \begin{matrix} -1/4 \\ +1/2 \end{matrix}$$

is a solution to the L.A.E. linear programming problem.

Now for $h_2 = (2,3)$ we have;

$$i \begin{bmatrix} +1 & 0 \\ 0 & +1 \end{bmatrix} \begin{bmatrix} 3 & 8 \\ 7 & 2 \end{bmatrix} \frac{-1}{10} = (-1,-1)$$

and

$$\begin{bmatrix} +1 & 0 \\ 0 & +1 \end{bmatrix} \begin{bmatrix} 3 & 8 \\ 7 & 2 \end{bmatrix} \begin{bmatrix} -2 \\ 2 \end{bmatrix} \frac{-1}{10} + \begin{bmatrix} +1 & 0 \\ 0 & +1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{matrix} 1 \\ 0 \end{matrix} \geq \begin{matrix} 0 \\ 0 \end{matrix} .$$

Therefore,

$$\beta^* = \begin{bmatrix} -4 & -6 \\ -1 & +1 \end{bmatrix}^{-1} \begin{bmatrix} -2 \\ 2 \end{bmatrix} = \begin{matrix} 1 \\ -2/5 \end{matrix}$$

is another solution to the L.A.E. linear programming problem. Hence, $\beta^*$ is not unique and is not a well-defined estimator.

Now suppose $y_4 = 0$ with all other observations unchanged. Then $\Sigma y_t = 0$.

For $h_1 = (1,2)$ Condition 1 is still satisfied, but for Condition 2;

$$\begin{bmatrix} +1 & 0 \\ 0 & +1 \end{bmatrix} \begin{bmatrix} -2 & 5 \\ 10 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ -2 \end{bmatrix} \frac{-1}{8} + \begin{bmatrix} +1 & 0 \\ 0 & +1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 3/4 \\ -6/8 \end{bmatrix} .$$

Hence Condition 2 is not satisfied. From the example in Chapter IV, we know that this will never be satisfied when $\Sigma y_t = 0$.

For $h_2 = (2,3)$, Condition 1 is still satisfied with $I^* = I$ but for Condition 2;

$$- \begin{bmatrix} +1 & 0 \\ 0 & +1 \end{bmatrix} \begin{bmatrix} 3 & 8 \\ 7 & 2 \end{bmatrix} \begin{bmatrix} -2 \\ 2 \end{bmatrix} \frac{-1}{10} + \begin{bmatrix} +1 & 0 \\ 0 & +1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} +1 \\ -1 \end{bmatrix} .$$

Hence Condition 2 is not satisfied.

In a similar fashion the reader can verify that it is not possible to satisfy both Conditions 1 and 2 for any $h_i$ when $I^* = I$.

The reader can also verify that with $h_2 = (2,3)$ and;

$$I^* = \begin{bmatrix} +1 & 0 \\ 0 & -1 \end{bmatrix}$$

both Conditions 1 and 2 are satisfied and the solution to the L.A.E. minimization problem is uniquely given by;

$$\beta^* = \begin{bmatrix} -4 & -6 \\ -1 & +1 \end{bmatrix}^{-1} \begin{bmatrix} -2 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ -2/5 \end{bmatrix} \quad .$$

## REFERENCES

Aigner, D., and Chu, S. "On Estimating the Industry Production Function," Amer. Econ. Rev., Sept. 1968, 826-9.

Appa, G., and Smith, C. "On $L_1$ and Chebyshev Estimation," London School of Economics (mimeographed).

Arrow, K., and Hoffenberg, M. A Time Series Analysis of Interindustry Demands, Amsterdam, 1959.

Ashar, V., and Wallace, T. "A Sampling Study of Minimum Absolute Deviation Estimates," Operations Research, Sept. 1963, 747-58.

Blattberg, R., and Sargent, T. "Regression with Non-Gaussian Stable Distributions. Some Sampling Results," Econometrica, May 1971, 501-10.

Box, G., and Tiao, G. "A Further Look at Robustness via Bayes' Theorem," Biometrica, 1962, 419-32.

Charnes, A., Cooper, W., and Ferguson, R. "Optimal Estimation of Executive Compensation by Linear Programming," Management Science, Jan. 1955, 138-51.

Chu, J. "On the Distribution of the Sample Median," Ann. Math. Stat., 1955, 112-16.

Chu, J., and Hotelling, H. "The Moments of the Sample Median," Ann. Math. Stat., 1955, 593-606.

Clark, P. "A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices," Econometrica (forthcoming), 1973.

Cramer, H. Mathematical Methods of Statistics, New Jersey, 1946.

David, F., and Johnson, N. "Statistical Treatment of Censored Data," Biometrica, 1954, 225-36.

Dhrymes, P. Econometrics, New York, 1970.

Durbin, J., and Watson, G. "Testing for Serial Correlation in Least Squares Regression," Biometrica, 1950, 409-28.

Fama, E. "Mandelbrot and the Stable Paretian Hypothesis" in P. Cootner, ed., The Random Character of Stock Market Prices, M.I.T., 1965.

Fama, E., and Roll, R. "Some Properties of Symmetric Stable Distributions," Journal of Amer. Stat. Assoc., Sept. 1968, 817-36.

_____. "Parameter Estimates for Symmetric Stable Distributions," Journal of Amer. Stat. Assoc., June 1971, 331-38.

Feller, W., An Introduction to Probability Theory and Its Applications, New York, 1966, Vol. I.

_____. An Introduction to Probability Theory and Its Applications, New York, 1968, Vol. II.

Gnedenko, B., and Kolmogorov, A. Limit Distributions for Sums of Random Variables, Cambridge, Mass. (1954).

Gorman, W. "Preference, Revealed Preference, and Indifference," in J. Chipman et al., Preference, Utility and Demand, New York, 1971.

Granger, C., and Orr, D. "Infinite Variance and Research Strategy in Time Series Analysis," Journal of Amer. Stat. Assoc., June 1972, 275-85.

Hogan, W. "Norm Minimizing Coefficients are Unbiased," unpublished manuscript, U.S. Air Force Academy, 1973.

Houthakker, H., and Taylor, L. D. Consumer Demand in the U.S., Harvard, 1965.

Huber, P. "Robust Statistics: A Review," Ann. Math. Stat., 1972, 1041-67.

Johnston, J. Econometric Methods, New York, 1963.

Kendall, M., and Stuart, A. The Advanced Theory of Statistics, New York, 1958, Vols. I and II.

Lindley, D., and Smith, A. "Bayes Estimates for the Linear Model," Journal of the Royal Stat. Soc., Ser. B, 1972, 1-18.

Malinvaud, E. Statistical Method of Econometrics, Chicago, 1966.

Mandelbrot, B. "New Methods in Statistical Economics," Journal of Political Economy, 1963, 1-27.

Meyer, J., and Glauber, R. Investment Decisions, Economic Forecasting, and Public Policy, Mass., 1964.

Oveson, R. "An Analysis of Forecasting Accuracy Using MSAE," unpublished Ph.D. dissertation, Harvard, 1966.

Plackett, R. L. "The Discovery of the Method of Least Squares," Biometrica, 1972, 239-51.

Relles, D. "Robust Regression by Modified Least Squares," unpublished Ph.D. dissertation, Yale, 1968.

Rockafeller, R. Convex Analysis, New Jersey, 1970.

Rosenberg, B., and Carlson, D. "The Sampling Distribution of Least Absolute Residuals Regression Estimates," Univ. of Calif., Berkley, 1971 (mimeographed).

Smith, V., and Hall, T. "A Comparison of Maximum Likelihood Versus BLUE Estimator," Rev. of Econ. and Stat., 1972, 186-90.

Spivey, W., and Thrall, R. Linear Optimization, Preliminary Edition, Univ. of Michigan, 1970.

Taylor, L. D. "On Estimation by Minimizing the Sum of Absolute Errors," unpublished manuscript, Univ. of Michigan, 1970.

Theil, H. Principles of Econometrics, New York, 1971.

Tsiang, S. "Risk, Return and Portfolio Analysis," Journal of Political Economy, May 1973, 748-52.

Wagner, H. "Linear Programming Techniques for Regression Analysis," Journal of Amer. Stat. Asso., 1959, 206-12.