"FOLDED" ESTIMATES OF REGRESSION[+]

by

C. Zachary Gilstein* and Edward E. Leamer**

Discussion Paper Number 219
October 1981
Department of Economics
University of California, Los Angeles

[+]Support from N.S.F. grant SOC78-09477 is gratefully acknowledged.

*Special Research Associate, Department of Economics, UCLA.

**Professor of Economics, UCLA.

"Folded" Estimates of Regression

by

C. Zachary Gilstein and Edward E. Leamer

0.0 Abstract

Assume that observations y_i $i = 1,\ldots,m$ are drawn independently from a distribution of the form $f(y_i;\underset{\sim}{x}_i,\underset{\sim}{\beta},g) = c_g \exp\{-g(y_i - \underset{\sim}{x}_i'\underset{\sim}{\beta})\}$ where $\underset{\sim}{\beta}$ is a $(k \times 1)$ vector of regression parameters, $\underset{\sim}{x}_i$ is a $(k \times 1)$ vector of known numbers, c_g is a normalizing constant, and $g \in G$ a space of functions. We investigate here the set of alternative maximum likelihood estimates for the regression parameter vector, based on alternative choices for g with $g \in G = \{g|\mathbb{R}^1 \to \mathbb{R}^1$, g has a continuous derivative, $g(0) = 0$, $g(x) = g(-x)$, g is strictly convex$\}$. It is shown that the set of estimates is a finite union of convex polytopes, the set is connected and contractible, and the set may not be convex. Using this information, we develop an efficient algorithm to completely delineate the boundary of the set in two dimensions. Finally, an example is given and a comparison is made between the set of "folded" estimates and classical normal theory confidence ellipsoids.

1.0 Introduction

The assumption of normality is clearly convenient but is usually incredible as well. Nonetheless, the assumption may be made without concern when the inferences drawn from a data set do not depend critically on it. The results in this paper are intended to identify cases when the assumption of normality is critical. Rather than normality, we assume only that the density is symmetric and unimodal with tails that die at least as fast as a double exponential distribution. Given a data set, each density in this family selects a particular maximum likelihood estimate. When this set of maximum likelihood estimates is small, the convenient but incredible assumption of normality is also inconsequential, and does not require scrutiny. When the set of estimates is large, the assumption does require scrutiny; in particular, the equivalent of a prior density over the family of densities is required.

We consider here the familiar problem of estimating a vector of unknown regression parameters, $\underset{\sim}{\beta}$, from a sample of observations on random variables $Y_1,\ldots,Y_m$ drawn independently from a distribution

$$P(Y_i < y) = F(y - \underset{\sim}{x}_i'\underset{\sim}{\beta}) \qquad i = 1,\ldots,m$$

where $\underset{\sim}{x}_i : i = 1,\ldots,m$ are the rows of a known design matrix $X_{(m\times k)}$ and the distribution F is not precisely known. We will assume that F has a density of the form

$$f(y_i;\ \underset{\sim}{x}_i,\underset{\sim}{\beta},g) = c_g \exp\{-g(y_i - \underset{\sim}{x}_i'\underset{\sim}{\beta})\}$$

where c_g is a normalizing constant and $g \in G$ a space of functions. The restrictions we place on G will be made specific below.

A maximum likelihood estimate $\hat{\beta}_g$ is a solution to the problem of minimizing the function

$$h_g(\beta) = \sum_{i=1}^{m} g(y_i - x_i'\beta) ,$$

that is

$$h_g(\hat{\beta}_g) = \min_{\beta} h_g(\beta) .$$

We will be concerned with the set of maximum likelihood estimates

$$B(G) = \{\hat{\beta}_g | h_g(\hat{\beta}_g) = \min_{\beta} h_g(\beta), \ g \in G\}$$

where G is the following space of functions: $G = \{g | \mathbb{R}^1 \to \mathbb{R}^1$, g has a continuous derivative, $g(0) = 0$, $g(x) = g(-x)$, g is strictly convex$\}$. The family of distributions corresponding to $g \in G$ is a set of distributions with symmetric, unimodal densities which have tails that die at least as fast as the tails of the double exponential distribution. Since g has been assumed differentiable, the solution to the minimization problem can be found by solving simultaneously the normal equations:

$$\sum_{i=1}^{m} x_{ij} g'(y_i - x_i'\beta) = 0 \qquad j = 1, \ldots, k.$$

We assume throughout that the design matrix X is of full rank k so that this system of equations always has a unique solution.

Interest in the set B(G) derives on the one hand from the improbability of precise prior knowledge of g, and on the other from the difficulty of adequately dealing in a statistical sense with the infinite dimensional nuisance parameter g. Furthermore, after an experiment or study is completed the data set is fixed. Examination of the set $B(G)$ is tantamount to passing many different models over the data and examining the set of possible estimates that could be supported by the data. In this way we can examine the impact of our assumptions on the estimate. For instance we could compare the set $B(G)$ to a confidence region generated from some assumed distribution, i.e. a normal distribution. If the set $B(G)$ is large compared to the confidence region, then it is clear that the assumption of a particular distribution has a strong influence on the particular estimate we report. In that case it will be important to utilize prior information about the distribution or to use the data to create an adaptive estimate. On the other hand, if the set $B(G)$ is small compared to the confidence region, then we can see that the error arising from the uncertainty of the sampling distribution is overwhelmed by the standard error of the estimate.

Huber (1964, 1973) introduced the concept of M-estimates for robust estimation of a location parameter or more generally of a regression parameter vector. He suggested solving the equations

$$\sum_{i=1}^{m} x_{ij}\psi(y_i - \underset{\sim}{x}_i'\underset{\sim}{\beta}) = 0 \qquad j = 1, \ldots, k$$

simultaneously to find an estimate $\hat{\underset{\sim}{\beta}}$, where ψ is some function of the investigator's choosing. For robustness ψ should not give too much weight to large values of $|x|$. One objection to this method is that the choice of ψ may have a strong effect on the estimates $\hat{\underset{\sim}{\beta}}$. The set $B(G)$ bounds the ambiguity of the possible M-estimates for the class of ψ functions which are required to be continuous, monotone increasing, and satisfy $\psi(x) = -\psi(-x)$.

The title of this paper, "'Folded' Estimates of Regression," refers to the fact that for a special case $B(G)$ is the set of all weighted averages of the folded sample points. If $Y_{(1)}, Y_{(2)}, \ldots, Y_{(m)}$ are the order statistics found from the sample $Y_1, \ldots, Y_m$, then the folded sample is the set of midranges $M_i = (Y_{(i)} + Y_{(n+1-i)})/2 \quad i = 1,2,\ldots, [(n+1)/2]$. Leamer (1981) has shown that if x_i is the scalar one for all observations, that is, if we are estimating the location of the density, then $B(G) = \{\beta | \min_i M_i < \beta < \max_i M_i\}$ unless $\min_i M_i = \max_i M_i$ in which case B(G) reduces to this point. The concept of folding a sample does not generalize intuitively to the regression setting, but our mathematical minimization problem does generalize. This same kind of device has been used by Koenker and Bassett (1978) to produce regression "quantiles." It must be pointed out, however, that these generalizations are not unique because there are many minimization problems which produce as solutions the folded estimates (and also sample quantiles).

The problem of identifying the set B(G) is essentially the same as the problem discussed by Dybvig and Ross (1977) of identifying the set of feasible portfolios for classes of utility functions. Although it is not possible to give an analytic description of the region B(G) as a function of the data points, it is possible to construct an efficient algorithm to determine the boundaries in the two-dimensional case. Section 2 describes a method for testing if a particular $\hat{\underset{\sim}{\beta}} \in B(G)$. The technique leads to a general characterization of the set. Further properties of the set are developed in section 3 and some modifications are considered in section 4. The algorithm is described in section 5 of this paper and several examples are reported in section 6. Finally, in section 4, the algorithm is described and an example is given in which a comparison is made between the set of "folded" regression estimates and classical normal theory confidence ellipsoids.

2.0 The Feasibility of a Regression Estimate

In this section, we will show that a given estimate, $\hat{\underset{\sim}{\beta}}$, is in the set of "folded" estimates if and only if a related linear program has a solution. The linear programming formulation will lead to a general characterization of the set of "folded" estimates as a finite union of convex polytopes. We will illustrate this result with a simple example.

Given the data set and design matrix, we say that an estimate $\underset{\sim}{\beta}$ is feasible if there exists a $g \varepsilon G$ such that for the distribution implied by $g, \underset{\sim}{\beta}$ is the maximum likelihood estimate, that is, if $\underset{\sim}{\beta} \varepsilon B(G)$. The theorem below shows that the feasibility of a given $\underset{\sim}{\beta}$ may be determined by solving a linear program.

Definition 2.1. If $\underset{\sim}{u} = (u_i) \varepsilon R^m$, then

(i) $\underset{\sim}{u}$ is nonnegative, written $\underset{\sim}{u} \geqq 0$, if $u_i \geqq 0$ for all i,

(ii) $\underset{\sim}{u}$ is positive, written $\underset{\sim}{u} > 0$, if $u_i > 0$ for all i,

(iii) $\underset{\sim}{u}$ is semipositive, written $\underset{\sim}{u} \geq 0$, if $\underset{\sim}{u} \geqq 0$ but $\underset{\sim}{u} \neq 0$.

If $\underset{\sim}{u}, \underset{\sim}{v} \varepsilon R^m$, we write $\underset{\sim}{u} \geqq \underset{\sim}{v}$, $\underset{\sim}{u} > \underset{\sim}{v}$, $\underset{\sim}{u} \geq \underset{\sim}{v}$ according as $\underset{\sim}{u} - \underset{\sim}{v}$ is nonnegative, positive, or semipositive, respectively.

Theorem 2.1. Let the residual vector implied by a given $\hat{\underset{\sim}{\beta}}$ be $\hat{\underset{\sim}{e}}$ and assume that $\hat{\underset{\sim}{e}} \neq 0$. Order the subscripts i so that $0 \leq |\hat{e}_1| \leq |\hat{e}_2| \leq \cdots \leq |\hat{e}_m|$. Let $s_i = \text{sgn } e_i$, $S = \text{diag}\{s_1, \ldots, s_m\}$, and

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

Define $\mathscr{I}$ to be the subset of the first m integers for which $|\hat{e}_i| = |\hat{e}_{i-1}|$, $\hat{e}_0 \overset{\text{def}}{=} 0$. Then the given vector $\hat{\underset{\sim}{\beta}} \varepsilon B(G)$, if and only if the following linear program has a solution:

$$X^t SL\underset{\sim}{\gamma} = 0$$

$$\underset{\sim}{\gamma} \geq 0$$

$$\gamma_i = 0 \leftrightarrow i \in \mathscr{I}$$

Proof. The residual vector $\underset{\sim}{\hat{e}}$ with the subscripts ordered as given above can be written

$$\underset{\sim}{\hat{e}} = SL\underset{\sim}{\theta}$$

where $\underset{\sim}{\theta} \geq 0$, $\theta_i = |\hat{e}_i| - |\hat{e}_{i-1}|$ $i = 1,\ldots,m$, and $\theta_i = 0 \leftrightarrow i \in \mathscr{I}$. The cone

$$C = \{\underset{\sim}{e} | \underset{\sim}{e} = SL\underset{\sim}{\gamma},\ \underset{\sim}{\gamma} \geq 0,\ \gamma_i = 0 \leftrightarrow i \in \mathscr{I}\}$$

is the set of all vectors $\underset{\sim}{e}$ which have the same signs and same ordering of absolute values as $\underset{\sim}{\hat{e}}$.

If $\underset{\sim}{\hat{\beta}} \in B(G)$, then $\exists\, g \in G$ such that $X^t g'(\underset{\sim}{\hat{e}}) = 0$, where $g'(\underset{\sim}{\hat{e}})$ represents the vector $(g'(\hat{e}_1), g'(\hat{e}_2), \ldots, g'(\hat{e}_m))^t$. Since g' is monotone strictly increasing and $g'(x) = -g'(-x)$, the vector $g'(\underset{\sim}{\hat{e}}) \in C$. This implies there is a solution to the linear program.

To prove the converse, let $\underset{\sim}{\bar{\gamma}}$ satisfy the linear program and let $\underset{\sim}{e}^* = SL\underset{\sim}{\bar{\gamma}}$; then $\underset{\sim}{e}^* \in C$. Since $\underset{\sim}{\hat{e}}$ and $\underset{\sim}{e}^*$ have the same signs and same ordering of absolute values, $\exists\, g \in G$ such that $g'(\underset{\sim}{\hat{e}}) = \underset{\sim}{e}^*$. Then $X^t SL\underset{\sim}{\bar{\gamma}} = 0$ implies $X^t\underset{\sim}{e}^* = 0$ or $X^t g'(\underset{\sim}{\hat{e}}) = 0$. So $\underset{\sim}{\hat{\beta}}$ is feasible. □

We therefore have the problem: Determine if there exists a semipositive vector $\underset{\sim}{\gamma}$ such that $A^*\underset{\sim}{\gamma} = 0$ where $A^* = X^t SL$ and $\gamma_i = 0 \leftrightarrow i \in \mathscr{I}$. Since possibly for some i's $\gamma_i = 0$, the program may be reduced by removing the ith columns of $X^t SL$ for $i \in \mathscr{I}$. Then the problem becomes: Determine if there exists a positive vector $\underset{\sim}{\gamma}$ such that $A\underset{\sim}{\gamma} = 0$ where A is the reduced A* matrix.

The m-dimensional cone defined by the signs and ordering of the absolute values of the e_i's will be denoted by

$$C(\underset{\sim}{p}, \underset{\sim}{s}) = \{\underset{\sim}{e} \,|\, |e_{p_i}| \le |e_{p_{i+1}}|, \; i = 1, \ldots, m-1; \; e_i s_i \ge 0\}$$

where $\underset{\sim}{p}$ is a permutation of the first m integers and $\underset{\sim}{s}$ is a vector of positive and negative ones and zeros. A given parameter vector $\underset{\sim}{\beta}$ selects a cone in the sense that

$$\underset{\sim}{y} - X\hat{\underset{\sim}{\beta}} \in C(\underset{\sim}{p}(\hat{\underset{\sim}{\beta}}), \underset{\sim}{s}(\hat{\underset{\sim}{\beta}})),$$

where $\underset{\sim}{p}(\hat{\underset{\sim}{\beta}})$ and $\underset{\sim}{s}(\hat{\underset{\sim}{\beta}})$ indicate the ordering and signs of $\underset{\sim}{y} - X\hat{\underset{\sim}{\beta}}$. The preceding theorem asserts that the feasibility of $\hat{\underset{\sim}{\beta}}$ depends only on the cone selected by $\hat{\underset{\sim}{\beta}}$: $C(\underset{\sim}{p}(\hat{\underset{\sim}{\beta}}), \underset{\sim}{s}(\hat{\underset{\sim}{\beta}}))$. Thus if $\hat{\underset{\sim}{\beta}}$ is feasible, then any other value of $\hat{\underset{\sim}{\beta}}$, which selects the same cone, is also feasible.

The space R^m is therefore partitioned by the m^2 hyperplanes defined by

$$\begin{array}{lll} e_i = e_j & i = 1, \ldots, m & \\ & & j \ne i \\ e_i = -e_j & j = 1, \ldots, m & \\ e_i = 0 & i = 1, \ldots, m & \end{array}$$

The set B(G) is the finite union of the convex polytopes defined by the linear mapping of the feasible cones in R^m into the parameter space $\mathbb{R}^k$. In principle, we could check each of these cones individually. However, since there are $2^m m!$ cones, even modestly large m would present numerical difficulties which would tax all but the most extravagent computer budgets.

<u>Example 1</u>   Consider the model $y = \beta_1 + \beta_2 x$ with data values (x_i, y_i) given by (-2, -3), (0, 0), (2, 1), and (3, 4). Figure 1 illustrates the separating lines and the set B(G) for this example. Computations of the feasibility of several points in this set are carried out below in the counterexample showing that the set is not convex.

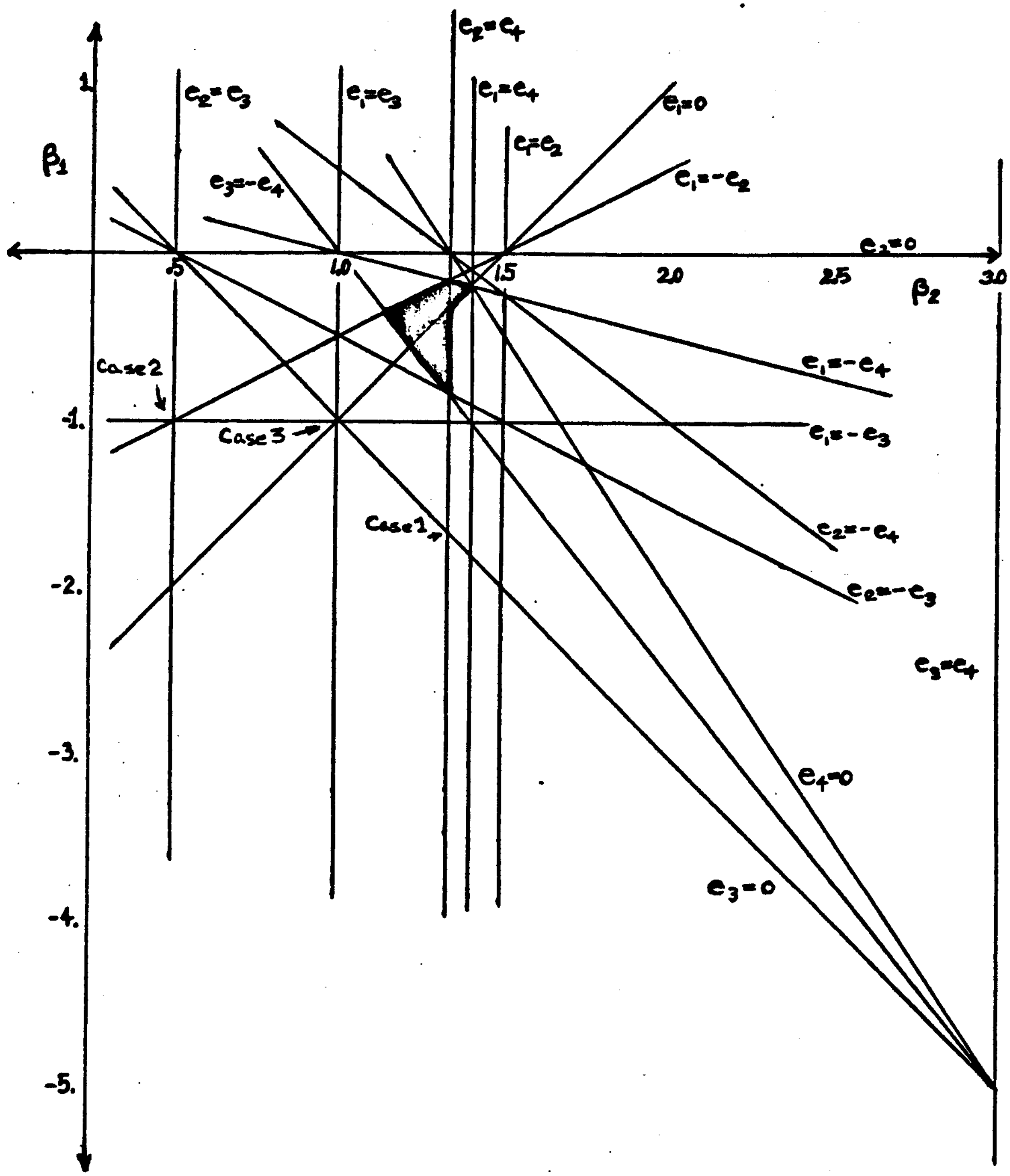


Figure 1. Separating lines and set of maximum likelihood estimates for Example 1. B(G) outlined and shaded.

3.0 Topological Properties of $B(G)$

We now investigate some of the topological properties of the set $B(G)$. These properties will further define the nature of the set and will form the basis for an algorithm to completely delineate the boundary of the set $B(G)$ in the two-dimensional case. We show below that the set is connected and that there are examples of the set which are not convex. Since the set need not be convex we will show that $B(G)$ is contractible, a result which will be important in the development of the algorithm. We then investigate the boundary of the set and give a result on the invariance of the set under location and scale changes of the data vector and reparameterization of the design matrix. The proofs of Theorems 3.1 and 3.3 which are technical in nature are not given below. The reader is referred to Gilstein (1980) for the details of these proofs.

Theorem 3.1. The set $B(G)$ is connected.

The proof of this theorem involves investigating the mapping ϕ from $G \to B(G)$ which maps the function g into the corresponding maximum likelihood estimate. It can be shown that this mapping is continuous. Then, since the space of functions G is connected, $B(G)$, being the continuous image of a connected set, is connected. We next prove by example that the set $B(G)$ may not be convex.

Theorem 3.2. There exist data vectors $\underset{\sim}{y}$ and design matrices X for which the set $B(G)$ is not convex.

<u>Proof</u>. We examine the example above where the model is $\dot{y} = \beta_1 + \beta_2 x$ and the data points (x_i, y_i) are (-2, -3), (0, 0), (2, 1), and (3, 4). We will use the linear program formulation to test the feasibility of three values of $\hat{\beta}$. We will show that $(-.2, 1.38) \in B(G)$, $(-.6, 1.32) \in B(G)$, but that $(-.4, 1.35) \notin B(G)$. For (-.2, 1.38) $|e_1| < |e_4| < |e_2| < |e_3|$,

$$e_1 < 0,\ e_4 > 0,\ e_2 > 0,\ e_3 < 0,$$

$$X'SL = \begin{pmatrix} 0 & 1 & 0 & -1 \\ 3 & 1 & -2 & -2 \end{pmatrix}$$

and $\underset{\sim}{\gamma}' = (2, 4, 1, 4)$ satisfies $X'SL\underset{\sim}{\gamma} = 0$. For (-.6, 1.32) $|e_1| < |e_2| < |e_4| < |e_3|$,

$$e_1 > 0,\ e_2 > 0,\ e_4 > 0,\ e_3 < 0,$$

$$X'SL = \begin{pmatrix} 2 & 1 & 0 & -1 \\ -1 & 1 & 1 & -1 \end{pmatrix}$$

and $\underset{\sim}{\gamma}' = (2, 1, 11, 5)$ satisfies $X'SL\underset{\sim}{\gamma} = 0$. However, for (-.4, 1.35) $|e_1| < |e_4| < |e_2| < |e_3|$,

$$e_1 > 0,\ e_4 > 0,\ e_2 > 0,\ e_3 < 0,$$

$$X'SL\underset{\sim}{\gamma} = \begin{pmatrix} 2 & 1 & 0 & -1 \\ -1 & 1 & -2 & -2 \end{pmatrix}\underset{\sim}{\gamma} = 0$$

has no positive solution since subtracting the second equation from the first implies $3\gamma_1 + 2\gamma_3 + \gamma_4 = 0$. □

Since these sets may not be convex their shape is too ill-defined at this point to attempt to create an algorithm to determine the boundaries of these sets. The following theorem is important because it implies that there is only one continuous border of the set and that the set cannot have holes in it. That is the set is a solid.

<u>Theorem 3.3</u>. The set B(G) is contractible.

For a proof see Gilstein (1980).

We now investigate the boundary of the set. In particular we would like to know if the boundary of the set is included in the set or not. The set B(G) is a finite union of convex polytopes. The question then is whether the exterior faces, edges, and vertices are part of the set. The answer in general is in the negative, however there is a necessary condition for this to be the case and it is possible to design data vectors and design matrices for which this condition is not satisfied. For a given $\hat{\underset{\sim}{\beta}}$ to be an element of the set there must be a positive solution to a linear program of the form $A^*\gamma = 0$ where $A^* = X'SL$. It has been noted above that the constraints that some $\gamma_i = 0$ allows for the deletion of the corresponding columns from A^*. Let the reduced A^* matrix be denoted by A. Note that if $\hat{\underset{\sim}{\beta}}$ is on a face or edge of a polytope then some of the γ_i's are equal to zero and there will be a reduction in the A^* matrix. Recall that the hyperplanes $e_i = 0$ $i = 1, \ldots, m$, $e_i = e_j$, $e_i = -e_j$ $i, j = 1, \ldots, m$ $i \neq j$ partition R^k into a finite collection of convex polytopes. If $\hat{\underset{\sim}{\beta}}$ implies equality in one or more of these equations then $\hat{\underset{\sim}{\beta}}$ is contained in a polytope of degree less than k. This polytope is contained in a number of k dimensional polytopes. The following theorem gives a sufficient condition for the boundary of the set not to be included in the set.

Theorem 3.4. If $\hat{\underset{\sim}{\beta}}$ is contained in a polytope, P, defined by the appropriate set of hyperplanes; if the degree of P is less than k; if the corresponding matrix A_p is of full rank k; and if there exists a positive solution to $A_p\underset{\sim}{\gamma} = 0$; then there exists a positive solution to $A\underset{\sim}{\gamma} = 0$ for the corresponding A matrix for every higher dimensional polytope that contains P.

Definition 3.1. A boundary polytope of the set B(G) is a polytope which intersects a polytope of greater degree contained in B(G) and a polytope of greater degree not contained in B(G).

Corollary 3.4. If the matrix A corresponding to a boundary polytope of B(G) is of full rank k, then the boundary polytope is not contained in B(G).

For proofs of Theorem 3.4 and Corollary 3.4, see Gilstein (1980).

In general, if m is much larger than k, then it is unlikely that the A matrix will be of less than full rank at a boundary polytope. Thus in general the set B(G) will not contain its boundary and is therefore an open set. As noted above, however, it is possible to construct examples where this will not be the case. We give here one example.

Example 2. Consider the model $y = \beta_1 + \beta_2 x$ with data points $(x_i, y_i) = (0,0),(1,2),(2,1),(3,5)$, and consider the set B(G). The point $(\hat{\beta}_1, \hat{\beta}_2) = (0, 1.5)$ implies

$$0 = |\hat{e}_1| < \underset{+}{|\hat{e}_2|} = \underset{+}{|\hat{e}_4|} < \underset{-}{|\hat{e}_3|} .$$

The reduced matrix $X'SL = \begin{pmatrix} 1 & -1 \\ 2 & -2 \end{pmatrix}$ is of rank $1 < k = 2$. There is a positive solution to $X'SL\underset{\sim}{\gamma} = 0$ so this vertex is part of B(G). However the two adjacent polygons $0 < |\underset{-}{e_1}| < |\underset{+}{e_2}| < |\underset{+}{e_4}| < |\underset{-}{e_3}|$ and $0 < |\underset{+}{e_1}| < |\underset{+}{e_4}| < |\underset{+}{e_2}| < |\underset{-}{e_3}|$ are not in B(G) while the two adjacent polygons $0 < |\underset{-}{e_1}| < |\underset{+}{e_4}| < |\underset{+}{e_2}| < |\underset{-}{e_3}|$ and $0 < |\underset{+}{e_1}| < |\underset{+}{e_2}| < |\underset{+}{e_4}| < |\underset{-}{e_3}|$ are in B(G). The set has a butterfly shape (see Figure 2a).

We now give a theorem describing the invariance properties of the set B(G). These properties allow for changes in scale and location of the data vector $\underset{\sim}{y}$ and reparameterization of the design matrix X. These changes may make computations more convenient. To emphasize the dependence on the data vector and design matrix we write $B(G,\underset{\sim}{y},X)$ for B(G).

<u>Theorem 3.5</u>. If $\hat{\underset{\sim}{\beta}} \in B(G,\underset{\sim}{y},X)$ then the following are elements of the solution of the specified transformed problems:

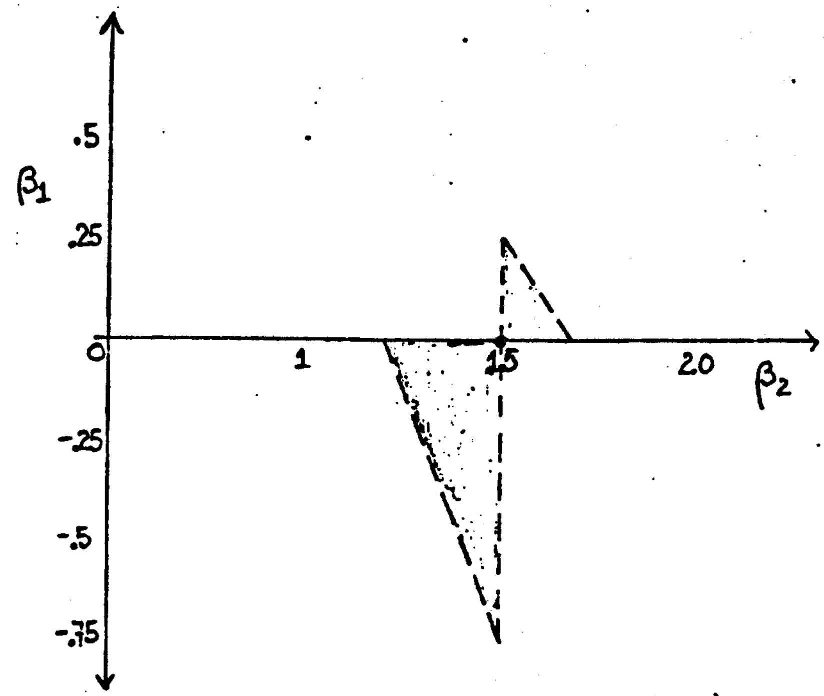
(i) $\lambda\hat{\underset{\sim}{\beta}} + \underset{\sim}{b} \in B(G,\lambda\underset{\sim}{y} + X\underset{\sim}{b},X)$ $\qquad \underset{\sim}{b} \in R^k, \quad \lambda \in (0,\infty)$

(ii) $A^{-1}\hat{\underset{\sim}{\beta}} \in B(G,\underset{\sim}{y},XA)$ $\qquad A_{k\times k}$ nonsingular

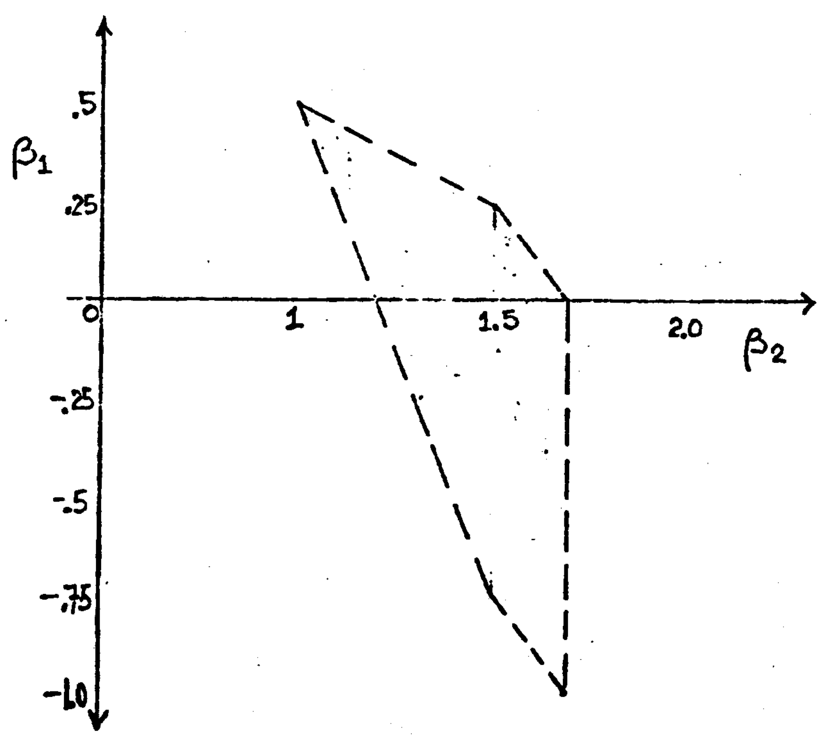
<u>Proof</u>: Left to reader.

<u>Remark</u>. Theorem 3.5 implies that under the specified transformations the new set $\tilde{B}(G)$ may be obtained by making the appropriate transformation on the vertices of B(G).

Similar results to those given in this Section and the previous Section can be derived under slightly weaker conditions on the space of functions, G. If we relax the assumption of strict convexity to convexity, for instance, a linear program can again be used to test the feasibility of an estimate of $\underset{\sim}{\beta}$.



<u>Figure 2a</u>. B(G) for Example 2.



<u>Figure 2b</u>. B($\tilde{G}$) for Example 3.

Let $B(\tilde{G})$ denote the set of regression estimates under the convexity assumption. Clearly $B(\tilde{G})$ is at least as large as $B(G)$. It may in fact be much larger. $B(\tilde{G})$ is illustrated in Figure 2b for the data of Example 2. One important implication of the convexity assumption is that the resultant set $B(\tilde{G})$ is open, that is, it never contains its boundary polytopes. For further discussion of modifications of G and proofs of these properties see Gilstein (1980).

4.0 Example

In this section, we will discuss an example involving 14 observations on two regressors and a response variable. To delineate the boundaries of the set of "folded" estimates in two dimensions, we have developed an efficient edge-tracing algorithm. Before discussing the example, we will briefly describe the algorithm.

Recall that the set $B(G)$ is a finite union of convex polytopes and that the set is connected and contractible. In two dimensions, these properties imply that $B(G)$ is a union of polygons and that $B(G)$ has a single piecewise linear boundary.

The algorithm separates into two sections. The first objective is to find the edge of the set by finding a line segment which separates a non-feasible polygon from a feasible one. After the edge of the set has been found, the objective is to follow the edge around the set until we return to the original segment. To find an edge segment, we start with a point inside the set, say the least-squares point, and move in one direction, say the direction of increasing β_1, checking the feasibility of the polygons we enter by using the linear programming test given in Theorem 2.1. When a nonfeasible polygon is found, then the line segment that was crossed to enter this polygon has been found to be an edge segment. To follow the edge of the set, we test the feasibility of polygons adjacent to the one containing the previously determined edge segment. By so testing, we can determine the adjacent line segment which separates a feasible polygon from a nonfeasible polygon. The process can be continued until a return is made to the original segment. The details of the algorithm are described in Gilstein (1980).

The algorithm described above is designed for determining the "folded" set of estimates for two regression parameters. In higher dimensions, the problem is substantially more complicated (especially because the set is not convex) and we have not developed an algorithm to completely delineate the boundary of the set for the case of more than two parameters. In multiple regression settings, the above algorithm can be used to determine the set of estimates for two of the parameters after fixing the value of the remaining parameters. In the example discussed below, we have set the intercept value to the least-squares estimate and determined the set of estimates for the two regression parameters.

The example we will examine is a gasoline demand function estimated with data from Maddala (1977, p. 129). The dependent variable g is the number of gallons of gasoline consumed per person, per year. The explanatory variables are p, the retail price of gasoline divided by the consumer price index (1953 = 100), and I, per-capita disposable income in 1958 dollars. The least-squares estimate using the fourteen annual observations (1947-1960) is:

$$g = \underset{(77.5)}{799.1} - \underset{(.706)}{2.563}\, p + \underset{(.0015)}{.0616}\, I \quad ,$$

where standard errors are reported in parentheses. This equation suggests that a ten percent relative increase in the price of gasoline that raised the index from 100 to 110 would reduce per capita consumption by 26 gallons with a standard error of 7 gallons.

The sensitivity of this conclusion to the assumption of normality is revealed by the set of "folded" estimates shown in Figure 3. The lines which cross in the middle of the figure are the major axes of the 50% confidence ellipsoid based on the normal assumption. The least absolute-

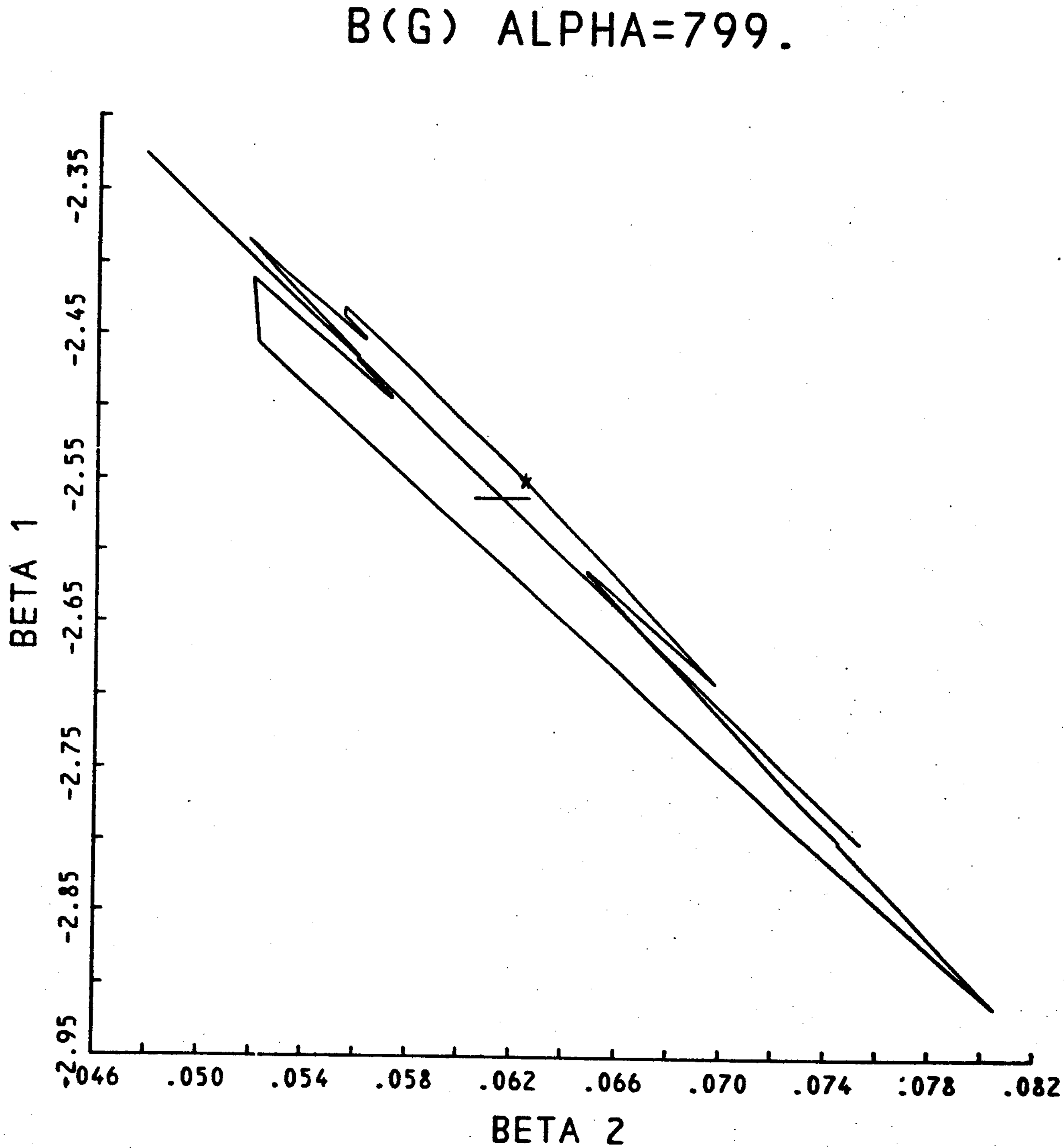


Figure 3

The Set of Maximum Likelihood Estimates: Gasoline Demand Example error point (*), which is necessarily in the closure of B(G), is also shown.

In this example, the set of "folded" estimates is very similar to the 50% ellipsoid. The shape of these two regions can be expected to be similar because both require the residual vector e to be small. In this case the traditional 95% ellipsoid would completely capture the folded set and it may be concluded that sampling uncertainty is more important than "specification uncertainty". In other examples, it has been noticed that small samples yield 95% confidence ellipsoids that are generally larger than the sets of folded estimates, while large samples yield the reverse. The tendency for the sampling uncertainty to converge to zero more rapidly than the specification uncertainty has been studied in detail for the location parameter problem in Gilstein (1981). There it is shown that while the confidence ellipsoid may shrink at a rate $O(1/\sqrt{n})$ the set B(G) cannot shrink at a rate faster than roughly $O(1/\log n)$ for any distribution determined for $g \in G$.

5.0 Summary and Concluding Remarks

We have proposed in this paper a method for studying the robustness of least-squares estimates to choice of error distribution. We compute a set of alternative estimates based on alternative assumptions about the error distribution. When this set of estimates is small, the choice of a particular distribution is inconsequential. When the set is too large to be useful, either estimation from the given data set is suspended, or a narrower family of distributions must be identified.

In the one-dimensional location case the set of estimates is the interval between the smallest and largest folded sample point. Algorithms for encompassing the set in higher dimensions are difficult because the set is not necessarily convex. We have presented an edge-tracing algorithm for the two-dimensional case, and we have shown generally that the feasibility of any particular estimate can be established by a linear programming algorithm. A grid search is therefore possible for the higher dimensional problems, but will necessarily suffer from the non-convexity of the set. The alternative, which we have illustrated here, is to fix all but two of the parameters at their least-squares values and to generate the "folded" set for the remaining couple. Though this is neither a slice nor a projection of the set B(G), it does usefully indicate cases when the assumption of normality is consequential.

It is fair to object that the set of distributions we have considered here is either too narrow or too wide. Assymetric distributions have been excluded, as have distributions with fatter tails than normals (e.g. Cauchy). Distributions with thin tails (e.g. Uniform) have been included. We think it is interesting both to enlarge and to shrink the set of distributions and to study the resulting sets of estimates. The intent would be to identify maximal sets of distributions which admit usefully small sets of estimates. For results based on other sets of distributions see Leamer (1981), Gilstein and Leamer (1981) and Gilstein (1980).

The results which we have presented are bounds for point estimates. We would like also to be able to provide bounds for interval estimates, or more ideally for posterior probabilities. Because the mapping from distributional assumptions into posterior probabilities is quite complex, we are not altogether hopeful that this important problem can be solved.

References

[1] Dybvig, Phillip H. and Stephen A. Ross (1977), "Portfolio Efficient Sets," Econometrica, forthcoming.

[2] Gale, David (1960), The Theory of Linear Economic Models, McGraw-Hill, New York.

[3] Gilstein, C. Zachary(1981), "On the Joint Asymptotic Distribution of Extreme Midranges", mimeo.

[4] Gilstein, C. Zachary (1980), Sets of Maximum Likelihood Estimates for Regression Parameter Vectors, unpublished Ph.D. thesis, UCLA.

[5] Gilstein, C. Zachary, and Edward E. Leamer (1981), "Sets of Weighted Regressions," mimeo.

[6] Huber, Peter J. (1964), "Robust Estimation of a Location Parameter," Annals of Math. Stat., 35, 73-101.

[7] Huber, Peter J. (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," Annals of Statistics, 1 799-821.

[8] Koenker, Roger and Bassett, Gilbert, Jr. (1978), "Regression Quantiles," Econometrica, 46, 33-50.

[9] Leamer, E. E. (1981), "Sets of Estimates of Location," Econometrica, 49, 193-204.

[10] Maddala, G. S. (1977) Econometrics, McGraw-Hill, New York.