SOME RESULTS IN NONLEAST SQUARES REGRESSION ANALYSIS

A Dissertation

Presented to

the Faculty of the

Department of Industrial and Systems Engineering

University of Houston

In Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

---

by

Albert G. Gray

August 1972

SOME RESULTS IN NONLEAST SQUARES REGRESSION ANALYSIS

An Abstract of a Dissertation

Presented to

the Faculty of the

Department of Industrial and Systems Engineering

University of Houston

In Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

---

by

Albert G. Gray

August 1972

ABSTRACT

This thesis considers the regression analysis problem in which the estimators of the parameters are selected according to some criterion other than least squares. Two basic areas are considered. First, some basic properties are derived for the estimators that minimize the sum of the absolute values of the residuals raised to the λ power. Both the homoscedastic and heteroscedastic cases are considered. Second, procedures for estimating weights for two types of heteroscedastic models are presented.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Fitting a curve to a set of data points is one of the basic problems in data analysis. This problem is usually approached by determining a general form of the curve such as $\tilde{Y} = X\beta$ where X is a vector of variables and β is a vector of parameters. However, due to the stochastic nature of the available data, the response $\tilde{Y}$ cannot be observed. The actual observed quantity, Y, is the true response, $\tilde{Y}$, plus some random error, e. Thus, if n observations are made, the model becomes $Y = X\beta + e$ where Y is an $(n\times 1)$ vector of observed responses, X is an $(n\times p)$ matrix of known values of the independent mathematical variables, β is a $(p\times 1)$ vector of unknown parameters, and e is an $(n\times 1)$ vector of random error with $E(e) = 0$.

The regression problem in statistics is to select the β such that some objective function is extremized. The choice of the objective function depends upon the application and the distribution of the random error and no universal agreement has been reached.

The usual assumption is that the random error has a dispersion matrix equal to $I\sigma^2$ where I is the $(n\times n)$ identity. Models with this form of variance are said to possess homoscedasticity. If, in addition, the random errors are normally distributed, then the maximum likelihood estimators of β

and the least squares estimators, that is, estimators that minimize the sum of the squares of the residuals, are identical. Furthermore, the Gauss-Markoff theorem shows that these estimators are unbiased and have minimum variance among all linear unbiased estimators [1]. In this context linear means that the estimators are linear combinations of the Y vector. In addition, the mathematics for testing various hypotheses is very tractable. For these reasons the least squares estimators are usually accepted by the practitioner for this model.

If the variance does not meet the homoscedastic assumption, then the dispersion matrix of ee^T can be modeled by $V\sigma^2$ where V is an $(n \times n)$ symmetric positive definite matrix. This type of dispersion matrix is called heteroscedastic. Often the least squares estimators are used for this model. There are two basic reasons for this usage of the least squares estimators. First, and from a practical point of view perhaps the most important, the least squares procedure is much wider known and understood than any other technique. Second, due to the central limit theorem, the errors are often assumed approximately normal.

However, it is known [2] that least squares does not produce the linear unbiased minimum variance estimators for the heteroscedastic model, even if the data is normally distributed. There is even criticism of the use of least

squares for certain cases of the homoscedastic model. For example, if the errors are double exponential, then the estimators that minimize the sum of the absolute value of the residuals are identical to the maximum likelihood estimators [2]. The criterion of linear unbiased minimum variance estimators has been criticized since it may exclude some desirable estimators that are either biased or non-linear [2, 3, 4, 5]. The criterion of the mean square error, which is the variance plus the square of the bias, has been proposed by some of the critics as an alternate to the linear unbiased minimum variance estimators.

There have been two general methods proposed to find estimators that may be in some sense superior to the least squares estimators. The first of these are estimates that

$$\min \sum_j |Y_j - \hat{Y}_j|^{\lambda} \qquad (1)$$

where Y_j is the j^{th} observation, $\hat{Y}_j$ is the j^{th} predicted response, and λ a known constant. The case most often considered has $\lambda = 1$. The estimators so produced are called the minimum sum of absolute deviations estimators and have been proposed by Eddington in a footnote to one of Fisher's papers [6], Herrey [7] and others.

The second general method for the heteroscedastic model consists of estimating the dispersion matrix and then performing a least squares analysis where each data point has been weighted based upon the estimated dispersion matrix.

This model has been proposed by many authors including Draper and Smith [2] and Daniels [8].

The remaining chapters of this thesis will be devoted to estimators of the types described above. For the homoscedastic case, Chapter II will consider the expectation and variance of the estimators determined by the norm in (1) for $\lambda = 1$ and $\lambda \to \infty$. Chapter III will develop some basic properties of the estimators that satisfy the above norm for a general λ. Chapter IV will consider the problem of estimating weights for weighted least squares. Finally, Chapter V will propose some related research projects.

CHAPTER II

ESTIMATORS OBTAINED FROM LINEAR PROGRAMMING

For the model $Y = X\beta + e$ with X known, estimators of β that satisfy the criterion of minimizing the sum of absolute deviations or of minimizing the absolute value of the maximum deviations, Chebyshev's criterion, have been proposed as alternates for fitting a curve to data [9, 10, 11].

It is also known that linear programming may be used to find the estimates for these criteria. However, the expectation, variance and covariance of these estimators are not known. Under very general conditions, this chapter will present formulas for the dispersion matrix and show that these estimators are unbiased. The variance of the predicted mean of future observations will also be derived. Finally, the variance of the estimators and the predicted mean of future observations will be shown to be at least as large as the corresponding variances obtained by the least squares criterion.

Model

This chapter will consider the model

$$Y = X\beta + e \qquad (1)$$

where Y is an $(n\times 1)$ matrix of observed responses, X is an $(n\times p)$ matrix of known constants, β is a $(p\times 1)$ matrix of

unknown parameters and e is an (n×1) matrix of random error. Furthermore, it is assumed that $E(e) = 0$ and the dispersion matrix of ee^T is $D(ee^T) = I\sigma^2$ where $\sigma^2 > 0$ and is unknown. Let X_i be the i^{th} row of X, then an individual observation of the response, Y_i, may be expressed

$$Y_i = X_i\beta + e_i, \qquad i=1,2,\ldots,n, \qquad (2)$$

where e_i is the i^{th} element of e.

The symbol b will denote the (p×1) matrix of the estimators of β. The i^{th} element of b will be written b_i. Let $\hat{Y}$ be the (n×1) matrix of the predicted response and $\hat{Y}_i$ the i^{th} element of this matrix. Thus

$$\hat{Y}_i = X_i b, \qquad i=1,2,\ldots,n. \qquad (3)$$

The assumption that $b_i \neq 0$, $i=1,2,\ldots,p$, will be made for the remainder of this chapter. As stated above, b_i is an estimate of β_i. The probability that the continuous random variable $b_i = 0$ is zero, even if $\beta_i = 0$. Thus the assumption is not a critical restriction to the following results. Actually, it is sufficient to make the weaker assumption that b_i, $i=1,2,\ldots,p$, be basic variables in the optimal linear programming solution that will be described later.

Minimum Sum of Absolute Deviations

The minimum sum of absolute deviations criterion, MSAD, selects the estimators of β such that

$$\sum_{i=1}^{n} |Y_i - \hat{Y}_i| = \sum_{i=1}^{n} |Y_i - X_i b| \qquad (4)$$

is minimum. According to Crocker [9], this criterion of fitting a curve to data was first proposed by Boscovich in 1757. The numerical difficulties that arise in the minimization of the expression in (4) have limited the development of the procedure.

However, in the late 1950's and early 1960's, Wagner [10], Karst [12] and Fisher [13] formulated the problem as a linear program. Let d be an (n×1) matrix of deviations, and denote the i^{th} element by d_i. The n equations analogous to (2) may be written

$$Y_i = X_i b + d_i, \qquad i=1,2,\ldots,n. \qquad (5)$$

Since d_i is not restricted as to sign and the simplex algorithm for solving linear programs requires the variables to be positive, the well-known device [14] of writing $d_i = d_i^+ - d_i^-$ where $d_i^+ \geq 0$ and $d_i^- \geq 0$ will be adopted.

With this notation, equation (5) became

$$Y_i = X_i b + d_i^+ - d_i^-, \quad i=1,2,\ldots,n. \qquad (6)$$

The objective is to

$$\min\Sigma|Y_i - \hat{Y}_i| = \min\Sigma|X_i b + d_i^+ - d_i^- - X_i b|$$
$$= \min\Sigma|d_i^+ - d_i^-| = \min\Sigma(d_i^+ + d_i^-). \qquad (7)$$

The last equality follows since in the simplex method d_i^+ and d_i^- cannot both be simultaneously greater than zero. Thus the numerical estimates for the MSAD criterion can be found by solving the linear programming problem with the

objective function given by (7) and constraints given by (6). It should be noted that b_i is not restricted as to sign either and needs to be written as $b_i = b_i^+ - b_i^-$ for the simplex method. However, the more notationally convenient b_i will be used except where it is necessary to consider the sign of the estimates.

It is known [11] that for the nondegenerate case a linear program with n constraints will have n basic variables, i.e., nonzero variables, and the remaining nonbasic variables will be zero. Under the assumption that the estimates are not zero, then $b_i > 0$, $i=1,2,\ldots,p$, and thus are basic variables. There is a total of n basic variables, so n-p of the deviations must be basic. Also, there are n deviations, consequently p of the deviations must be zero or nonbasic variables. The specific set of the p nonbasic deviation variables is readily available from the solution to the linear programming problem. For ease of notation, relabel the constraints in equations (6) so that the p nonbasic deviations are in the first p rows. The initial tableau for $b_i > 0$, $i=1,2,\ldots,p$, can be symbolized as

b	d^+		d^-		Y
X_p	I_p	$0_{p,n-p}$	$-I_p$	$0_{p,n-p}$	Y_p
X_r	$0_{n-p,p}$	I_{n-p}	$0_{n-p,p}$	$-I_{n-p}$	Y_r

where X_p is the (p×p) submatrix of X that corresponds to the p constraints that have zero deviations, X_r is the [(n-p)×p] submatrix of X corresponding to the remaining deviations, Y_p is the (p×1) submatrix of Y corresponding to the p nonbasic deviations, Y_r is the [(n-p)×1] submatrix of the remaining elements of Y, I_s is an (s×s) identity and O_{st} is an (s×t) null matrix.

In this notation b is given by the solution of the system

$$X_p b = Y_p. \tag{8}$$

However, it remains to be shown that the solution to (8) is identical to the MSAD estimators. It will now be proven that X_p is nonsingular and the MSAD estimators are uniquely given by $b = X_p^{-1} Y_p$.

The final tableau may be partitioned as

b	d^+		d^-		y
I_p	A	$O_{p,n-p}$	-A	$O_{p,n-p}$	b_m
$O_{n-p,p}$	Z	D	-Z	-D	d_n

where A is a (p×p) matrix, Z is an [(n-p)×p] matrix, b_m is a (p×1) vector containing the MSAD estimates of β and d_n is the [(n-p)×p] vector of numerical values of the nonzero deviations, D is an [(n-p)×(n-p)] matrix with 1 or -1 on the principal diagonal and zeros elsewhere. The specific

signs of the diagonal elements of D depend on the signs of the deviations. It is not necessary to describe this matrix in greater detail for this discussion.

Define B^{-1} to be the matrix that occupies the same columns as the original basic feasible solution, i.e., the columns corresponding to d^+ in the initial tableau. The initial tableau can be transformed to the final tableau by premultiplying by the matrix

$$B^{-1} = \begin{pmatrix} A & 0_{p,n-p} \\ Z & D \end{pmatrix}.$$

The inverse of B^{-1}, B, can be found in the columns of the initial tableau that correspond to the identity matrix in the final tableau. Thus B will have the form

$$B = \begin{pmatrix} X_p & 0 \\ X_r & D \end{pmatrix}$$

$$I_n = BB^{-1} = \begin{pmatrix} X_p & 0 \\ X_r & D \end{pmatrix} \begin{pmatrix} A & 0 \\ Z & D \end{pmatrix} = \begin{pmatrix} X_p A & 0 \\ X_r A+DZ & I_{n-p} \end{pmatrix}$$

Thus

$$X_p A = I_p$$

and

$$X_p^{-1} = A.$$

Multiply B^{-1} times the last column of the initial tableau to obtain

$$\begin{pmatrix} A & 0 \\ Z & D \end{pmatrix} \begin{pmatrix} Y_p \\ Y_r \end{pmatrix} = \begin{pmatrix} AY_p \\ ZY_p+DY_r \end{pmatrix} = \begin{pmatrix} b_m \\ d_n \end{pmatrix}.$$

The MSAD estimators are then

$$b_m = AY_p = X_p^{-1} Y_p.$$

Thus the MSAD estimators are equivalent to the solution of the system in (8). For computational purposes, the matrix X_p and Y_p can be found from the linear programming solution.

As an example, consider the problem of fitting the model

$$Y = b_1 + b_2X$$

to the data

Y	X
1	1
2	4
3	3 .

The initial simplex tableau is

b_1	b_2	d_1^+	d_2^+	d_3^+	d_1^-	d_2^-	d_3^-	Y
1	1	1	0	0	-1	0	0	1
1	4	0	1	0	0	-1	0	2
1	3	0	0	1	0	0	-1	3

and the final tableau is

b_1	b_2	d_1^+	d_2^+	d_3^+	d_1^-	d_2^-	d_3^-	Y
1	0	4/3	-1/3	0	-4/3	1/3	0	2/3
0	1	-1/3	1/3	0	1/3	-1/3	0	1/3
0	0	-1/3	-2/3	1	1/3	2/3	-1	4/3

Thus

$$X_p = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}, \qquad Y_p = \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

$$B^{-1} = \left(\begin{array}{cc|c} 4/3 & -1/3 & 0 \\ -1/3 & 1/3 & 0 \\ \hline -1/3 & -2/3 & 1 \end{array}\right) \qquad X_p^{-1} = A = \begin{pmatrix} 4/3 & -1/3 \\ -1/3 & 1/3 \end{pmatrix}$$

and

$$b = X_p^{-1} Y_p = \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix}$$

as indicated by b_m in the final tableau.

If some $b_j < 0$, a similar result can be derived. Let $X_{p,i}$ denote the i^{th} column of X_p and denote the matrix that corresponds to X_p for the case by X_p^*.

$$X_p^* = (X_{p,1}, X_{p,2}, \ldots, X_{p,j-1}, -X_{p,j}, X_{p,j+1}, \ldots, X_{p,n})$$

The minus appears on $X_{p,j}$ since b_j^- and not b_j^+ is basic. The matrix corresponding to A in B^{-1}, A^*, will differ from A only in that the i^{th} row of A^* has been multiplied by -1 in order to place the correct (minus) sign on b_j^-. Let α_i be the i^{th} row of A. Then A^* can be written

$$A^* = \begin{pmatrix} \alpha_1 \\ \cdots \\ \alpha_{j-1} \\ -\alpha_j \\ \alpha_{j+1} \\ \cdots \\ \alpha_p \end{pmatrix}$$

Let $(A^*X^*_p)s,t$ denote the s,t element of $A^*X^*_p$. Then

$$(A^*X^*_p)s,t = \begin{cases} 0 & s\neq j\ ,\ t\neq j\ ,\ s\neq t \\ 1 & s\neq j\ ,\ t\neq j\ ,\ s=t\ . \end{cases}$$

If $s=j\ ,\ t\neq j$

$$(A^*X^*_p)s,t = (-\alpha_j\ X_{p,t}) = -(\alpha_j\ X_{p,t}) = 0\ .$$

If $s\neq j\ ,\ t=j$

$$(A^*X^*_p)s,t = \alpha_s(-X_{p,j}) = -(\alpha_s\ X_{p,j}) = 0\ .$$

If $s=j\ ,\ t=j$

$$(A^*X^*_p)s,t = -\alpha_j(-X_{p,j}) = \alpha_j\ X_{p,j} = 1\ .$$

Thus the appropriate matrix for the equivalent system of equations is nonsingular and a unique solution exists. The above proof of nonsingularity can be easily extended to the general case of any subset of the b_i's being negative. It will be assumed for the remainder of this chapter that the correct signs have been applied to X_p and A in the system in (8).

Before continuing with development of the properties of the MSAD estimators, it is convenient to describe Chebyshev's criterion for fitting a curve to data.

Chebyshev's Criterion

It has been shown [10] that the criterion of minimizing the absolute value of the maximum deviation may be formulated as a linear programming problem. However, a modification of the formulation given in [10] is more convenient for the present purpose. Let d^* be the absolute value of the

maximum deviation. Any deviation d_i can be written $d_i = d^* - c_i$ where $0 \leq c_i \leq 2d^*$. Substitute into equations analogous to (2), to obtain $Y_i = X_i b + d^* - c_i$. Define $r_i \geq 0$ such that $-2d^* + c_i + r_i = 0$ for i=1,2,...,n. The linear program,

$$\begin{aligned} &\text{minimize } d^* \\ &\text{subject to} \quad Y_i = X_i b + d^* - c_i \\ &\qquad\qquad -2d^* + c_i + r_i = 0 \\ &\qquad\qquad d^* > 0,\ c_i \geq 0,\ r_i \geq 0 \\ &\qquad\qquad \text{for } i=1,2,\ldots,n, \end{aligned} \tag{9}$$

will minimize the absolute value of the maximum deviation. The program in (9) has 2n equations and 2n+1 variables. Thus, there are 2n basic variables. If the regression curve is not an exact fit, d^* will be nonzero and thus a basic variable. As in the MSAD case, assume that each b_i, i=1,2,...,p, is a basic (nonzero) variable. Thus, 2n-p-1 of the 2n variables c_i and r_i, i=1,2,...,n, will be basic variables. The p+1 nonbasic variables, that is, variables that are zero, will be the remaining c_i or r_i variables.

If c_i is a nonbasic variable, i.e., $c_i = 0$, then from (9) $r_i > 0$; thus r_i is a basic variable. Furthermore, $c_i = 0$ implies that $d_i = d^*$. If r_i is a nonbasic variable, i.e., $r_i = 0$, then $c_i = 2d^*$ and $d_i = -d^*$. Let $S_i = \pm 1$ such that $d_i = S_i d^*$. Relabel the equation in (9) such that the p+1

equations of the form $Y_i = X_i b + S_i d^*$ are first. Relabel the columns of the linear programming tableau such that the p columns corresponding to b_i are in the first p positions and the column corresponding to d^* is in the $(p+1)^{st}$ position. Let X_{p+1} be a $[(p+1)\times(p+1)]$ matrix such that the first p elements of the i^{th} row of X_{p+1} are the row of X that correspond to $d_i = S_i d^*$ and the $(p+1)^{st}$ element of X_{p+1} is S_i for $i=1,2,\ldots,p+1$. Let Y_{p+1} be a $[(p+1)\times 1]$ matrix consisting of the first (p+1) elements of Y. The linear program in (9) is equivalent to the system of equations

$$Y_{p+1} = X_{p+1} b \tag{10}$$

with solution

$$b = (X_{p+1})^{-1} Y_{p+1} \tag{11}$$

where b is a (p+1) matrix whose first p elements are the p estimates of β and the $(p+1)^{st}$ element is an estimate of d^*. All the information required to form (10) and find its solution (11) is given by the solution to (9). The uniqueness proof is essentially identical to the MSAD case.

The Chebyshev's criterion is equivalent to

$$\min \lim_{\lambda\to\infty} \Sigma |d_i|^{\lambda}.$$

If necessary, rescale the data so $d_i > 1$ for some i. Then the sum will be dominated by $(d^*)^{\lambda}$ as λ increases without bound. Thus minimizing the above limit is equivalent to minimizing the absolute value of the largest deviation. This observation will be developed further in the next chapter.

Properties of the Estimators

The properties of MSAD and Chebyshev's estimators may be derived together. Adopt the notation that β is the vector of parameters for the MSAD case and β is the vector of parameters plus the quantity d^* in the $(p+1)^{st}$ position for Chebyshev's criterion, i.e., let $d^* = b_{p+1}$. Let $t = p$ for MSAD and $t = p+1$ for Chebyshev's criterion.

For either criterion, write the system of equations (8) or (10) as

$$Y_t = X_t b$$

with solution

$$b = AY_t.$$

An element of Y_t, $Y_{t,i}$, $i=1,2,\ldots,t$, has expectation $E(Y_{t,i}) = E(X_{t,i}\beta + \varepsilon_i) = X_{t,i}\beta$ where $X_{t,i}$ is the i^{th} row of X_t. It is also apparent that

$$E(Y_t) = E\begin{pmatrix} Y_{t,1} \\ Y_{t,2} \\ \cdots \\ Y_{t,t} \end{pmatrix} = \begin{pmatrix} X_{t,1}\beta \\ X_{t,2}\beta \\ \cdots \\ X_{t,t}\beta \end{pmatrix} = X_t\beta.$$

The expectation of the estimators obtained by either MSAD or Chebyshev's criterion may now be derived.

$$E(b) = E(AY_t) = AE(Y_t) = AX_t\beta = \beta ,$$

since $A^{-1} = X_t$. Thus the estimators are unbiased.

The dispersion matrix of any linear combination of the Y's, say LY, is $LL^T\sigma^2$ [15]. Thus

$$D(b) = AA^T\sigma^2.$$

Now consider the properties of the mean of future observations at a p dimensional point, S, as predicted by the model, that is by

$$\hat{Y} = Sb.$$

Then

$$E(\hat{Y}) = SE(b) = S\beta.$$

So $\hat{Y}$ is unbiased. The variance of $\hat{Y}$ is given by

$$V(\hat{Y}) = \sum_{i=1}^{p} \sum_{j=1}^{p} S_i S_j \text{ Cov } (b_j, b_i) = S^T AA^T S\sigma^2. \quad (12)$$

Comparison with Least Squares

The estimators of β using either MSAD or Chebyshev's criterion are linear combinations of Y_t, thus they are a linear combination of Y. Furthermore, the estimators are unbiased. Thus by the Gauss-Markoff theorem the variance of a b_i based on the MSAD or Chebyshev criteria is at least as large as the variance of an estimator for β_i obtained from the least squares criterion.

To compare the variance of the mean of future observations for the MSAD case, partition $X = \begin{pmatrix} X_p \\ X_r \end{pmatrix}$. Then

$$X^TX = X_p^T X_p + X_r^T X_r.$$

The rank of X_p is p so $X_p^T X_p$ is positive definite and $X_r^T X_r$ is positive semidefinite. Under these conditions, it can be shown [16] that

$$[(X_p)^T X_p]^{-1} - (X^T X)^{-1} = AA^T - (X^T X)^{-1}$$

is positive semidefinite. Thus

$$S^T AA^T S\sigma^2 \geq S^T (X^T X)^{-1} S\sigma^2.$$

The quantity on the left was shown in (12) to be the variance of the mean of future observations at the point, and it is well known that the quantity on the right is the corresponding variance obtained from least squares.

To compare the variance of the mean of future observations for Chebyshev's criterion, partition

$$X_t = \begin{pmatrix} X_p & C \\ R & Q \end{pmatrix} \tag{13}$$

where X_p is the $(p\times p)$ matrix of coefficients of $b_1, b_2, \ldots, b_p$ in the first p equations, R is a $(1\times p)$ matrix of coefficients of $b_1, b_2, \ldots, b_p$ for the $(p+1)^{st}$ equation, C is the $(p\times 1)$ matrix of coefficients of d^* for the first p equations, and Q is the sign of the coefficient of d^* in the $(p+1)^{st}$ equation. Thus

$$X = \begin{pmatrix} X_p \\ R \\ X_r \end{pmatrix}$$

where X_r has the same definition as for the MSAD case. Hence,

$$X^T X = X_p{}^T X_p + R^T R + X_r{}^T X_r. \tag{14}$$

Let V_d be the $[(p+1)\times(p+1)]$ dispersion matrix for $b_1, b_2, \ldots, b_p$ and d^*. From (13)

$$(V_d)^{-1} = \frac{1}{\sigma^2}\,(X_t{}^T X_t)^{-1} = \frac{1}{\sigma^2}\begin{pmatrix} X_p{}^T X_p + R^T R \\ C^T X_p + S^T R \end{pmatrix}\begin{pmatrix} X_p{}^T C + R^T Q \\ C^T C + Q^T Q \end{pmatrix}.$$

The $(p\times p)$ dispersion matrix V for $b_1, b_2, \ldots, b_p$ in Chebyshev's criterion is obtained by deleting the last row and column from V_d. It can be shown [16] that

$$V^{-1} = X_p{}^T X_p + R^T R - (X_p{}^T C + R^T Q)(C^T C + Q^T Q)^{-1}(C^T X_p + Q^T R).$$

Since $(C^T C + Q^T Q)$ is a scalar composed of the sum of p+1 squares of 1 or -1, and $(C^T X_p + Q^T R)^T = X_p{}^T C + R^T Q$. Since the weighted sum of positive semidefinite matrices is positive semidefinite for non-negative weights [16],

$$X_r{}^T X_r + \frac{1}{p+1}\,(C^T X_p + Q^T R)^T\,(C^T X_p + Q^T R)$$

is positive semidefinite. As before, $V - (X^T X)^{-1}$ is positive semidefinite and thus the variance of the mean of future observations for Chebyshev's criterion will be at least as large as the variance for least squares.

CHAPTER III

REGRESSION ACCORDING TO A GENERALIZED NORM

The three criteria discussed in Chapter II can be written as

$$\min \Sigma |d_i|^{\lambda} \qquad (1)$$

where d_i is the i^{th} deviation and $\lambda = 1$ for MSAD, $\lambda = 2$ for LS and λ increases without bound for Chebyshev's criterion. This observation suggests a study of the criterion in (1) for an arbitrary λ. This chapter will show that the variance of the unbiased estimators for the homoscedastic model is at least as large as the least squares variance. For the heteroscedastic model it will be shown that for certain cases unbiased linear estimators exist that have smaller variance than least squares.

The first part of this chapter will assume that $\lambda > 1$. This assumption will be relaxed toward the end of the chapter. Also, the notation of Chapter II will be adopted where applicable.

Separable Convex Programming Formulation

If $\lambda > 1$, the criterion in (1) becomes convex and separable. For the model $Y = X\beta = e$, write as before $Y_i = X_i b + d_i$. Let $d_i = a_i - u_i$ where $a_i \geq 0$ and $u_i \geq 0$. The variable a_i corresponds to d_i^+ and u_i to d_i^- in Chapter II. The change in notation is used since a_i and u_i

must be raised to a power. The criterion in (1) is expressed in this notation as

$$\min\Sigma|d_i|^{\lambda} = \min\Sigma|a_i - u_i|^{\lambda} = \min(\Sigma a_i^{\lambda} + \Sigma u_i^{\lambda}) \qquad (2)$$

since a_i and u_i cannot be simultaneously nonzero.

Let $f(z) = z^{\lambda}$ for $\lambda > 1$ then $f''(z) = \lambda(\lambda-1)z^{\lambda-2}$, if $z > 0$ then $f''(z) > 0$ and $f(z)$ is convex. Thus a_i^{λ} and u_i^{λ} are convex functions. Since the sum of convex functions is convex [17], the expression in (2) is convex. The system is obviously separable. The estimates, b, which satisfy the general criterion are obtained by the solution of the separable convex programming problem

$$\min(\Sigma a_i^{\lambda} + \Sigma u_i^{\lambda}) \qquad (3)$$

subject to

$$Y_i = X_i b + a_i - u_i, \qquad i=1,2,\ldots,n$$
$$a_i \geq 0,\ u_i \geq 0 .$$

An algorithm for solving separable convex programs is given in [17].

Let M be an upper bound on the absolute value of the deviations, that is, $d^* < M$. Define $c > 0$ and q, a positive integer, such that $cq = M$.

First consider only the portion of the formulation relating to a_i. Let $f_i(a_i)$ be a piecewise linear function whose linear segments connect the points $[(k-1)c, a_i^{\lambda}]$ and $[kc, a_i^{\lambda}]$ for $k=1,2,\ldots,q$. Denote the slope of the k^{th} line segment by S_{ik}, $k=1,2,..,q$. Since $f_i(a_i)$ is convex,

$0 < S_{i1} < S_{i2} < \ldots < S_{iq}$. By choosing c large $f_i(a_i)$ can be made arbitrarily close to a_i^{λ}.

Define the auxiliary variables

$$a_{ik} = \begin{cases} 0 & a_i < (k-1)c \\ a_i - (k-1)c & (k-1)c < a_i \leq kc \\ c & kc \leq a_i \end{cases}$$

for $k=1,2,\ldots,q$

$i=1,2,\ldots,n.$

These variables are bounded by

$$0 \leq a_{ik} \leq c \qquad \text{for } k=1,2,\ldots,q$$

$i=1,2,\ldots,n.$

Let k^* be such that

$$(k^*-1)c \leq a_i \leq k^*c.$$

Then

$$c < 2c < \ldots < (k^*-1)c \leq a_i \leq k^*c < (k^*-1)c < \ldots < qc$$

and

$$a_{ik} = \begin{cases} c & k=1,2,\ldots,k^*-1 \\ a_i - (k^*-1)c & k=k^* \\ 0 & k=k^*+1,k^*+2,\ldots,q \end{cases}$$

$$\sum_{k=1}^{q} a_{ik} = (k^*-1)c + a_i - (k^*-1)c + (q-k^*)0 = a_i. \quad (5)$$

With this notation

$$f_i(a_i) = \Sigma S_{ik} a_{ik} \approx a_i^{\lambda}. \quad (6)$$

In an identical manner the equations

$$u_i = \Sigma u_{ik}, \quad (7)$$

and

$$u_i^{\lambda} \approx \Sigma S_{ik} u_{ik} \tag{8}$$

can be developed.

The separable convex program in (3) can be written

$$\min\left(\sum_{i=1}^{n} \sum_{k=1}^{q} S_{ik} a_{ik} + \sum_{i=1}^{n} \sum_{k=1}^{q} S_{ik} u_{ik} \right)$$

subject to

$$Y_i = Xb + \sum_{k=1}^{q} a_{ik} - \sum_{k=1}^{q} i_{ik} \tag{9}$$

$$0 \leq a_{ik} \leq c$$

$$0 \leq u_{ik} \leq c,$$

$$\text{for } i=1,2,\ldots,n$$

$$k=1,2,\ldots,q$$

where it is understood that $a_{ik} = 0$ if $a_{i(k-1)} < c$ and $u_{ik} = 0$ if $u_{i(k-1)} < c$. Hiller and Lieberman [17] show that the constraints on the slopes S_{ik} will cause the simplex method to automatically satisfy the last two restrictions. Thus the solution to the linear program in (9) gives the estimates for the criterion in (1).

Introduce the slack variables r_{ik} and t_{ik}, $i=1,2,\ldots,n$, $k=1,2,\ldots,q$, such that the upper bounds on a_{ik} and u_{ik} hold as equalities. The linear program can be written in simplex form as

$$\min\left(\sum_{i=1}^{n} \sum_{k=1}^{q} S_{ik} a_{ik} + \sum_{i=1}^{n} \sum_{k=1}^{q} S_{ik} u_{ik} \right)$$

subject to

$$Y_i = \sum_{j=1}^{p} X_{ji} b_j + \sum_{k=1}^{q} a_{ik} - \sum_{k=1}^{q} u_{ik}$$

$$c = a_{ik} + r_{ik}$$

$$c = u_{ik} + t_{ik} \qquad (10)$$

$$\text{for } a_{ik} \geq 0,\ u_{ik} \geq 0,\ r_{ik} \geq 0,\ t_{ik} \geq 0$$

$$i=1,2,\ldots,n,\ k=1,2,\ldots,q.$$

The initial simplex tableau can be expressed as

b	a_{1i} ... a_{ni}	u_{1i} ... u_{ni}	r_{1i} ... r_{ni}	t_{1i} ... t_{ni}	Y
X	A_1 ... A_n	$-A_1$ ... $-A_n$	0 ... 0	0 ... 0	Y
0	I ... 0	0 ... 0	I ... 0	0 ... 0	c_ν
...					...
0	0 ... I	0 ... 0	0 ... I	0 ... 0	c_ν
0	0 ... 0	I ... 0	0 ... 0	I ... 0	c_ν
...					...
0	0 ... 0	0 ... I	0 ... 0	0 ... I	c_ν

where X is the (n×p) matrix of known coefficients, Y is the (n×1) matrix of observations, A_i is an (n×q) matrix that contains zeros everywhere except for the i^{th} row which is all ones, c_ν is a (q×1) vector with each element equal to c, I is a (q×q) identity and 0 denotes a null matrix.

The simplex formulation has n constraints on Y, nq constraints relating a_{ik} and r_{ik}, and nq constraints relating u_{ik} and t_{ik}. The feasible region is enclosed by the n+2nq equations. There are n+4nq variables, n of the

Y's and nq of the a's, u's, r's and t's. The simplex method is thus equivalent to solving a system of n+2nq equations in n+2nq basic variables. The remaining 2nq variables will be identically equal to zero.

As before, assume that none of the estimates are identically zero, that is, the variables $b_1, b_2, \ldots, b_p$ are basic. The remaining n+2nq-p basic variables must be some of the a_{ik}, u_{ik}, r_{ik} and t_{ik}, i=1,2,...,n, k=1,2,...,q. The particular set of the basic variables is available in the final tableau of the simplex method.

Let the [(n+2nq)×(n+2nq)] matrix of coefficients of the equivalent system be denoted by H. The columns of H are the columns of the tableau in (10) that correspond to the n+2nq basic variables. Let H^{-1} be the inverse of H and denote the (i,j) element of H^{-1} by h_{ij}. If necessary, relable the rows of the tableau such that the i^{th} row of H^{-1} defines b_i. With this notation

$$b_i = \sum_{j=1}^{n} h_{ij}Y_j + c \sum_{j=n+1}^{n+2nq} h_{ij}. \qquad (11)$$

The first n elements of the first p rows of H^{-1} are orthogonal to the first p elements of the first n rows of H. This observation may be proved by multiplying the i^{th} row of H^{-1} by the column in (10) that corresponds to b_i

$$\sum_{j=1}^{n} h_{ij}X_{ij} + \sum_{j=n+1}^{n+2nq} h_{ij}(0) = 1.$$

$$\sum_{j=1}^{n} h_{ij}X_{ji} = 1$$

The first equality holds since the column in (10) is a submatrix of H, the inverse of H^{-1}. Multiply the i^{th} row of H^{-1} by the column that corresponds to b_j, $j \neq i$, in (10)

$$\sum_{j=1}^{n} h_{ij}X_{ji} + \sum_{j=1}^{n} h_{ij}(0) = 0$$

$$\sum_{j=1}^{n} h_{ij}X_{ji} = 0.$$

The expected value of b_i may now be found.

$$E(b_i) = E\left(\sum_{j=1}^{n} h_{ij}Y_j + c\sum_{j=n+1}^{n+nq} h_{ij}\right)$$

$$= \sum_{i=1}^{n} h_{ij}E(Y_i) + c\sum_{j=n+1}^{n+nq} h_{ij}$$

$$= \sum_{j=1}^{n} h_{ij}\left(\sum_{k=1}^{p} X_{jk}\beta_k\right) + c\sum_{j=n+1}^{n+nq} h_{ij}$$

$$= h_{i1}(X_{11}\beta_1+\ldots+X_{1p}\beta_p) + h_{i2}(X_{21}\beta_1+\ldots+X_{2p}\beta_p)$$

$$+\ldots+h_{in}(X_{n1}\beta_1+\ldots+X_{np})\beta_p + c\sum_{j=n+1}^{n+nq} h_{ij}$$

$$= (h_{i1}X_{11} + h_{i2}X_{21}+\ldots+h_{in}X_{p1})\beta_1+\ldots+$$

$$(h_{i1}X_{1p} + h_{i_2}X_{2p}+\ldots+h_{in}X_{np})\beta_p + c\sum_{j=n+1}^{n+nq} h_{ij}$$

$$E(b_i) = \beta_i + c\sum_{j=n+1}^{n+nq} h_{ij} . \qquad (12)$$

The last equation results from the orthogonality of X and the first p rows of H^{-1}.

The estimators that satisfy the criterion in (1) are linear combinations of the Y vector plus a constant which is identical to the bias, (11) and (12). Thus

$$b_i = H_i Y + K \qquad i=1,2,\ldots,p \tag{13}$$

$$E(b_i) = \beta_i + K \qquad i=1,2,\ldots,p \tag{14}$$

where $K = c \sum_{j=n+1}^{n+2nq} h_{ij}$ and H_i is the i^{th} row of H. The quantity K is a constant for a given set of data and may be found from the linear programming solution. It may or may not be zero.

Unbiased Estimators

Since the bias, K, is known (13), introduce

$$b_i^* = b_i - K \qquad i=1,2,\ldots,p\,. \tag{15}$$

Then

$$b_i^* = H_i Y \tag{16}$$

$$E(b_i^*) = \beta_i.$$

If $V(b_i)$ is the variance of b_i, then the variance of b_i^* is

$$V(b_i^*) = V(b_i - K) = V(b_i). \tag{18}$$

Thus b_i^*, i=1,2,...,p, are unbiased linear estimators with the same variance as b_i.

For the homoscedastic model, the Gauss-Markoff theorem applies. Thus the variance of b_i^* will be at least as large as the variance of the least squares estimators. However, the assumptions of the Gauss-Markoff theorem are not met by

the heteroscedastic model. Therefore, this theorem cannot be used to compare the variances to the variances of the least squares estimators.

Contaminated Normal Model

Although the variance of the λ-norm estimators for the general heteroscedastic model has not been compared to the variance of the least squares estimators, a comparison can be made for an interesting special case.

Tukey [18] has proposed the contaminated normal distribution as a model for the type of data often encountered. A sample from a contaminated normal with contamination coefficient G may be generated by considering a base distribution which is normal with mean μ and variance σ^2 and a contamination distribution which is normal with mean μ and variance $\phi^2\sigma^2$. The observation is selected from the base distribution with probability 1-G and from the contamination distribution with probability G. Thus, data from a contaminated normal is heteroscedastic since the variance is σ^2 or $\phi^2\sigma^2$. Tukey suggests that $\phi = 3$ and $0 < G < .1$. He shows that although it takes literally thousands of data points to detect the difference between a normal and a contaminated normal, the properties of the two distributions differ greatly.

For the special model $Y = b_1 + b_2X + e$ where the errors e are uncorrelated and selected from a contaminated

normal, Forsythe [19] has shown by simulation that for $G > 0$ values of λ between 1 and 2 produce estimators that have smaller mean square error than the least squares estimators. The reduction seems significant since Forsythe states that "with 10% contamination least squares is approximately only 50% efficient and so it might be said that the usual technique throws away half of the data."

Since the mean square error is the variance plus the square of the bias [20] and the least square estimate of β_i is unbiased

$$V(b_i)_\lambda + K^2 < V(b_i)_{LS}$$

$$V(b_i)_\lambda < V(b_i)_{LS}$$

$$V(b_i^*)_\lambda < V(b_i)_{LS} \qquad i=1,2,\ldots,p$$

where K is the bias of b_i. Thus for this type data which is practically impossible to detect from a true normal [18], the estimators b_i^*, $i=1,2,\ldots,p$, based on the λ-norm are unbiased, linear combinations of the Y vector and have variance less than the least squares estimators.

The Case of $\lambda \leq 1$

The previous sections in this chapter have considered the λ-norm with $\lambda > 1$. This section will first extend the results to the case where $\lambda \in (0,1)$ and then show that the case of $\lambda < 0$ is uninteresting.

If $\lambda \in (0,1)$, the separable convex programming formulation does not apply since the objective function is concave.

However, for $\lambda \in (0,1)$ an algorithm that differs from the simplex in only the way incoming and outgoing variables are selected has been developed [21]. Since the above discussion only uses the fact that the simplex solution is equivalent to solving a linear system of equations, the proofs will apply for $\lambda \in (0,1)$ as well.

If $\lambda \leq 0$, the criterion in (1) reduces to models that have little applicability. If $\lambda = 0$, then any b such that $d_i \neq 0$, for $i=1,2,\ldots,n$, will produce n as the minimum value of (1).

If $\lambda < 0$, then write

$$\Sigma|d_i|^{\lambda} = \Sigma(1./|d_i|^{-\lambda})$$

and let $$t_i = 1./|d_i|^{-\lambda}.$$

Since $\lambda < 0$, then $-\lambda > 0$, a t_i could be made as small as desired by choosing b such that d_i was very large. It is obvious that a choice of b exists that makes all d_i, $i=1,2,\ldots,n$, simultaneously as large as desired. Thus each t_i can be made arbitrarily close to zero and the sum can be made to approach zero. Since the expression in (1) is bounded below by zero, the minimum will occur when the deviations are arbitrarily large. The general regression problem attempts to find b vectors such that the deviations are in some sense small. The criterion in (1) with $\lambda < 0$ thus produces uninteresting results.

The minimization of $\Sigma|d_i|^{\lambda}$ produces large residuals, maximizing the sum will yield residuals small in a given

sense. However, the maximizing criterion will generate a class of uninteresting b vectors. Again write

$$\max \Sigma |d_i|^{\lambda} = \max \Sigma (1/|d_i|^{-\lambda})$$

where $-\lambda > 0$. If $d_i = 0$, $i=1,2,\ldots,n$, then the sum would contain an infinite term and would itself be infinite. Thus only b vectors that had at least one $d_i = 0$ would maximize the expression.

Future Study

To actually find the estimates of β for the contaminated normal that are linear unbiased and have variance smaller than least squares, it is necessary to solve the linear programming problem in (10). Unfortunately, the program in (10) may be very large and difficult to solve. The most efficient method of solving it is not known. Perhaps one of the existing separable convex programming algorithms can solve the problem efficiently. If not, perhaps the number of variables and iterations could be reduced by solving the problem like Forsythe [19] and using the results to initialize the linear programming procedure.

CHAPTER IV

WEIGHTED LEAST SQUARES ANALYSIS

This chapter will consider the problem of estimating weights that will transform the heteroscedastic regression model to the homoscedastic model. Specifically, the model to be considered is

$$Y = X\beta + e \tag{1}$$

where Y is an (n×1) vector of observations, X is an (n×p) matrix of known constants, β is a (p×1) vector of unknown parameters, and e is an (n×1) vector of random errors such that $E(e) = 0$ and the dispersion matrix is $D(ee^T) = V\sigma^2$, where V is an unknown symmetric positive definite matrix and $\sigma^2 > 0$.

Two methods of estimating weights that will transform the model in (1) to a model in which $D(ee^T) = I\sigma^2$ will be developed. In addition, a criterion for selecting between sets of weights will be presented. Finally, examples of the procedures will be given.

Introduction to Weighted Least Squares

It is well known that least squares will not produce minimum variance estimators of β if the errors are not distributed as $I\sigma^2$, where I is the identity matrix [2, 8]. Draper and Smith [2] show that there exists a symmetric

matrix W such that

$$W^T W = V$$

and that if the model in (1) is weighted by W^{-1} then the errors of the weighted model will have a dispersion matrix of $I\sigma^2$. The weighted model thus fulfills the hypothesis of the Gauss-Markoff theorem [1] and the weighted least squares estimators will have minimum variance unbiased estimators.

The matrix V and the variance σ^2 are rarely known and must be estimated. Historically, the residuals obtained from the model with V = I are used to estimate V. A plot of the residuals against the estimated response will appear as a horizontal band if the condition of constant variance is met [2]. However, if the range of the residuals varies with the predicted response, then heteroscedasticity is indicated, that is, nonconstant variance. Thus a guess at V is used to produce a weighted model and the plot of the residuals is again examined for any signs of heteroscedasticity. If it appears necessary, another estimate of V can be made and the process continued until the analyst is satisfied with the residual plots.

This procedure for estimating V appears deficient in two areas. First, there is no numerical procedure to aid in determining the estimates of V. Secondly, it relies entirely on a graphical method for choosing between estimators.

This chapter will be addressed to these difficulties. The selection problem will be discussed in the next section followed by considerations of the estimation problem.

Selection Criterion

A criterion for choosing between estimates of V can be developed by estimating $V\sigma^2$, not V itself. Since V must be positive definite and σ^2 is positive, $V\sigma^2$ is positive definite [2]. So there exists a nonsingular matrix W such that

$$W^TW = WW^T = V\sigma^2.$$

Let $f = W^{-1}e$. Then the expectation of f may be determined

$$E(f) = E(W^{-1}e) = W^{-1}E(e) = 0.$$

Since $E(e) = 0$, the dispersion matrix of ff^T is

$$\begin{aligned} D(ff^T) &= E(ff^T) \\ &= W^{-1}D(ee^T)W^{-T} \\ &= W^{-1}V\sigma^2W^{-T} \\ &= W^{-1}WW^TW^{-T} \\ &= I \cdot 1. \end{aligned}$$

Thus weights based on a knowledge of $V\sigma^2$ not only transform to a model with constant variance but also a model with $\sigma^2 = 1$.

Denote the usual unbiased estimator of σ^2 by s^2. Hence, for the transformed model, $E(s^2) = 1$. Estimators based on complete knowledge of $V\sigma^2$ produce estimates of s^2 that have expectation of 1.

The heuristic selection criterion to be used in this chapter compares various estimates of $V\sigma^2$ by examining residual plots and by measuring the nearness of s^2 to 1. Such a criterion might not produce the optimal transformation but at least the selected weights will have some properties of the optimal weights. Numerical studies, to be discussed later, show that fairly small changes in the weights can greatly influence s^2. Thus the criterion appears to be reasonably sensitive.

It might seem that it is more difficult to estimate weights that will transform to a model with constant variance equal to 1 than it is to find weights that produce models that have only constant variance. However, as will be shown in the next section, it is actually easier to estimate $V\sigma^2$ than V itself.

Estimating $V\sigma^2$

Williams [22] has shown that one cannot expect to produce reliable estimates of weights unless there are at least ten data points for each weight estimated. This much replicated data is rarely available. In order to estimate weights with less data, some additional assumptions are necessary. For the remainder of this chapter, the matrix $V\sigma^2$ will be assumed to be diagonal and will be denoted by D. The j^{th} element on the main diagonal of D will be denoted

by σ^2_j. In addition, it will be assumed that σ_j is some function of X. Two specific functions will be considered. First, σ_j will be assumed to be a linear function of X. Next, σ_j will be assumed to be proportional to some power of the absolute value of the mean response at the j^{th} point in the factor space.

When $\sigma_j = \Sigma X_{ji}\gamma_i$, Rutemiller and Bowers have developed a method of obtaining maximum likelihood estimators of β and γ by using the method of scoring. Before considering Rutemiller and Bowers' work in detail, it is worthwhile to pursue Rao's method of scoring.

The Method of Scoring

Maximum likelihood estimators are traditionally determined by solving the system of equations that result from setting to zero the partial derivatives of the likelihood function with respect to the various parameters. However, sometimes the resulting system cannot be analytically solved. In this case a numerical procedure can be used.

A general method for solving such systems would be to assume a trial solution and determine the difference between zero and the partials evaluated at the trial point. Based upon these differences, a new trial solution could be determined. This process can be repeated until the corrections become negligible. The method of scoring [15]

is a mechanization of such a general procedure. Thus the method of scoring can be considered as a procedure of obtaining numerical solutions to the maximum likelihood equations that are so complex the traditional analytical solutions cannot be found.

Specifically, consider the random sample $x_1, x_2, \ldots, x_n$ with the joint distribution function $L(x_1, \ldots, x_{n_j}, \theta_1, \ldots, \theta_K)$, where $\theta_1, \theta_2, \ldots, \theta_k$ are parameters to be estimated. Define A as a $(k \times 1)$ vector such that the i^{th} element of A is $\frac{\partial \ln L}{\partial \theta_i}$, $i=1,2,\ldots,k$. Denote by B the $(k \times k)$ matrix with the ij^{th} element equal to $-E\left(\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right)$ where the expectation is over the joint distribution. The matrix B is called the Fisher information matrix. Let α_1 be a $(k \times 1)$ vector of initial estimates of the parameters $\theta_1, \theta_2, \ldots, \theta_k$. An iterative procedure for finding an improved estimate of the parameters, α_{i+1}, is given by

$$\alpha_{i+1} = \alpha_i + [B(\alpha_i)]^{-1} A(\alpha_i), \qquad i=1,2,\ldots$$

where $A(\alpha_i)$ is the A matrix evaluated at α_i and $B(\alpha_i)$ is the Fisher information matrix evaluated at α_i. The process stops when $|\alpha_{i+1} - \alpha_i| < \varepsilon$ for some predetermined vector ε. Although it seems the $(k \times k)$ matrix $B(\alpha_i)$ would have to be inverted for each iteration, after some number of iterations, $B(\alpha_i)$ changes very slowly with changes in the parameters α_i. Thus $B(\alpha_i)$ need not be inverted each time.

Wald [23] has shown that at convergence for a large sample size the matrix B^{-1} is the dispersion matrix of the parameters.

Gamma Weighted Least Squares

Rutemiller and Bowers [24] have applied the method of scoring to the problem of estimating β and γ for the model

$$Y_j = X_j\beta + e_j \qquad j=1,2,\ldots,n$$

where X_j is the j^{th} row of X and the e_j, $j=1,2,\ldots,n$ are uncorrelated normal random variables with zero mean and the standard deviation σ_j is a linear function of the X variables, that is,

$$\sigma_j = \sum_i X_{ji}\gamma_i. \tag{2}$$

The log of the likelihood function for the normal distributed data becomes

$$\ln L = -n\ln\sqrt{2\pi} - \sum_j \ln\left(\sum_i X_{ji}\beta_i\right) - (1/2)\sum_j (Y_j - \sum_i X_{ji}\beta_i)^2\left(\sum_i X_{ji}\gamma_i\right)^{-2}.$$

For the method of scoring, let $k = 2p$ where the first k elements of A correspond to β and the last k correspond to γ. Now

$$A_\ell = \frac{\partial \ln L}{\partial \beta_\ell} = \sum_j \left(\sum_i X_{ji}\gamma_i\right)^{-2}\left(Y_j - \sum_i X_{ji}\beta_i\right)^2 X_{\ell j}$$

$$\text{for } \ell=1,2,\ldots,k$$

$$A_{\ell+p} = \frac{\partial \ln L}{\partial \gamma_\ell} = -\sum_j X_{\ell j}\left(\sum_i X_{ji}\gamma_i\right)^{-1} + \sum_j (Y_j - \sum_i X_{ji}\beta_i)^{-2} X_{\ell j}\left(\sum_i X_{ji}\gamma_i\right)^{-3}$$

$$\text{for } \ell=1,2,\ldots,k.$$

$$B_{\ell,m} = -E\left(\frac{\partial^2 \ln L}{\partial\beta_\ell \partial\beta_m}\right) = \sum_j X_{\ell j} X_{mj} \left(\sum_i X_{ji}\gamma_i\right)^{-2}$$

for $\ell=1,2,\ldots,p$

$m=1,2,\ldots,p$

$$B_{\ell,m+p} = -E\left(\frac{\partial^2 \ln L}{\partial\beta_\ell \partial\gamma_m}\right) = 0$$

for $\ell=1,2,\ldots,p$

$m=1,2,\ldots,p$

$$B_{\ell+p,m+p} = -E\left(\frac{\partial^2 \ln L}{\partial\gamma_\ell \partial\gamma_m}\right) = 2 \sum X_{\ell j} X_{mj} \left(\sum X_{ji}\gamma_i\right)^{-2}$$

for $\ell=1,2,\ldots,p$

$m=1,2,\ldots,p$

$$B_{\ell+p,m} = -E\left(\frac{\partial^2 \ln L}{\partial\beta_\ell \partial\gamma_m}\right) = 0$$

for $\ell=1,2,\ldots,p$

$m=1,2,\ldots,p$

Note that B may be partitioned as

$$B = \begin{pmatrix} B_{\ell,m} & 0 \\ 0 & 2B_{\ell m} \end{pmatrix}$$

or

$$B^{-1} = \begin{pmatrix} B_{\ell,m}^{-1} & 0 \\ 0 & 1/2\ B_{\ell,m} \end{pmatrix}.$$

Thus, only the $(p\times p)$ matrix $B_{\ell,m}$ need be inverted. This derivation is a special case of Rutemiller and Bowers' work since they did not require Y and σ_j to be a linear function of the same variables. Apparently they did not notice the above partitioning since one of their examples meets the requirement but does not have this property.

The estimators published in [24] are correct but there are some errors in the dispersion matrix which are possibly due to the inversion of the (2p×2p) matrix and not the (p×p) matrix.

After the estimates of γ have been obtained, the matrix $D = V\sigma^2$ can be estimated. By the assumptions of this case the j^{th} diagonal element of D will be equal to σ_j^2. If W is an (n×n) diagonal matrix with the j^{th} diagonal element equal to the estimate of σ_j, then $W^TW = WW^T = D$. The inverse of W, W^{-1}, will be a diagonal matrix with the j^{th} diagonal element equal to $1/\hat{\sigma}_j = 1/(\Sigma X_{ji}\hat{\gamma}_i)$. This matrix can then be used as a weight matrix in a weighted least squares analysis. For future reference this weighted model will be called the gamma weighted least squares model (GWLS).

An Attempt to Extend the Method of Scoring

Much of the historical work on variance stabilization transformations [20] is based on the assumption that the variance is proportional to some power of the mean. Let

$$\mu_j = \sum_i X_{ji}\beta_i, \qquad j=1,2,\ldots,n. \tag{3}$$

Such a model could be formulated by requiring the standard deviation of an observation to be

$$\sigma_j = |c|\,|\mu_j|^q, \qquad j=1,2,\ldots,n. \tag{4}$$

It is necessary to include the absolute value signs since σ_j must be positive.

It first appears that the method of scoring could be applied to this model. However, certain difficulties arise.

Let A be a (p+1) vector of parameters with the first p corresponding to β and the (p+1)st to q. If normality is assumed,

$$A_k = \frac{\partial \ln L}{\partial \beta_k} = -q \sum_j |\mu_j|^{-1} S(\mu_j) X_{jk}$$

$$+ |c|^{-2} \sum_j X_{jk} |\mu_j|^{-2q-1} [q(Y_j-\mu_j)^2 S(\mu_j) + (Y_j-\mu_j)|\mu_j|],$$

$$k=1,2,\ldots,p$$

$$A_{p+1} = \frac{\partial \ln L}{\partial q} = -\sum_j \ln|\mu_j| + |c|^{-2} \sum_j (Y_j-\mu_j)^2 |\mu_j|^{-2q} \ln|\mu_j| \qquad (5)$$

where
$$S(z) = \begin{cases} 1 & z > 0 \\ 0 & z = 0 \\ -1 & z < 0 . \end{cases}$$

The model has only p+1 parameters since the $\frac{\partial \ln L}{\partial c} = 0$ can be solved for c in terms of $\beta_1,\ldots\beta_p$, and q, yielding

$$c = \frac{\sum (Y_j-\mu_j)^2 |\mu_j|^{-2q}}{n} . \qquad (6)$$

Thus c acts like a variance where the residuals have been weighted by $|\mu_j|^{-2q}$.

The B matrix may be defined as

$$B_{k,\ell} = -E\left(\frac{\partial^2 \ln L}{\partial \beta_k \partial \beta_\ell}\right) = |c|^{-2} \sum_j X_{jk} X_{j\ell} [2q^2 |c|^2 |\mu_j|^{-2} + |\mu_j|^{-2q}]$$

$$k=1,2,\ldots,p$$

$$\ell=1,2,\ldots,p$$

$$B_{k,p+1} = B_{p+1,k} = -E\left(\frac{\partial^2 \ln L}{\partial \beta_k \partial q}\right) = 2q \sum_j X_{jk} S(\mu_j) |\mu_j|^{-1} |n|\mu_j| \qquad k=1,2,\ldots,p$$

$$B_{p+1,p+1} = -E\left(\frac{\partial^2 \ln L}{\partial q^2}\right) = 2 \sum_j (\ln|\mu_j|)^2.$$

Upon actually implementing this model, it was found that the method of scoring very seldom converged. The convergence seems to depend very heavily on the initial estimate of q.

While investigating methods for obtaining initial estimates of q, a method of estimating the weights of the original model was discovered. This method which is described below does not have as many numerical problems as the method of scoring and produces usable results. For these reasons, the method of scoring as applied to the model in (4) will not be pursued further.

CQ Weighted Least Squares

If the standard deviation and the means are related as in (4), an iterative procedure for estimating c and q is as follows. First, perform a least squares analysis to obtain the predicted responses, $\hat{Y}_j$, and the residuals $r_j = Y_j - \hat{Y}_j$ for $j=1,2,\ldots,n$. Partition the range of the $\hat{Y}_j$ into k mutually exclusive classes such that $n_i \geq 2$ where n_i is the number of observations in the i^{th} class. Let

$z_i = \Sigma \hat{Y}_i / n_i$ and $s_i^2 = \Sigma\, r_i^2/(n_i-1)$ where the sums are over all members of the i^{th} class. Regress the k points according to the model $\ln s_i = \ln c + q \ln z_i$. The regression will yield values of q and ln c.

With estimates of β, c and q known, a weight matrix W^{-1} may be constructed with the j^{th} diagonal element of W^{-1} equal to $\hat{c}^{-1}|\hat{\mu}_j|^{-\hat{q}}$, $j=1,2,\ldots,n$. A weighted least squares analysis can be performed to produce new residuals for the next iteration. The method must be iterative since both s_i and z_i depend upon the current estimates of β.

When the weight matrix becomes an identity, the process should be terminated. Typically this happens when q becomes 0 and c = 1, since $\sigma = 1$ for the optimal weights.

This model will be referred to as the CQ weighted least squares model (CQWLS).

Examples

A program was developed to implement the methods and criterion cited above. Four sets of data were analyzed by least squares, gamma weighted least squares and CQ weighted least squares. Two of the data sets were data that had been previously published [25, 26]. One of the remaining data sets was data simulated according to the CQ model and the other was data simulated according to the gamma model.

CQ Simulated Data

The first data set will be called the CQ simulated data since it was simulated with the standard deviation given in (4). The data consist of 93 points simulated from a line with $\beta_1 = 1$, $\beta_2 = 1$, $c = 0.1$ and $q = 1.5$. The factor space is given in Appendix 1 along with the simulated response and the actual standard deviation. The independent variables consist of 63 points identical to the data in [25] with 30 additional points augmented in order to increase the density of the data in the region of high variability. The random numbers were generated by the IBM-provided subroutine GAUSS [27] with a starting seed of 139. Since the first few numbers from such a subroutine are often poor samples from the population, the first 10 numbers were discarded. It should be noted that this simulation produced results with a large range of standard deviation (1.11 to 26.25). As a result, the simulated response contained a large degree of variation.

The application of the method of scoring proposed by Rutemiller and Bowers did not converge, so the gamma weighted least squares procedure could not be used. Apparently, this case departs too far from the assumptions of linearity (2) made by Rutemiller and Bowers.

The convergence of the CQ weighted least squares method is shown by

	b_1	b_2	% error of b_1	% error of b_2	s
Actual	1.00	1.00			
LS	6.08	0.67	508	33	
CQ Iteration 1	1.98	0.88	98	12	4.48
CQ Iteration 2	1.53	0.91	53	9	2.42
CQ Iteration 3	1.24	0.94	24	6	1.56
CQ Iteration 4	1.02	0.96	2	4	1.07

The first iteration produced values of c and q of 0.097 and 1.52, respectively. These values are very close to the actual simulated values of 0.1 and 1.5. The values of c and q for the final CQWLS results were 0.88 and 0.09, respectively, which are close to the theoretical values for the heteroscedastic model of 1.0 and 0. The successive weights drove s in the weighted factor spaces from 4.48 to 1.07. The estimate of 1.07 is close to the theoretical value of 1. Therefore, it can be concluded that the weights were very near optimal. The residual plot for the LS case is shown in Figure 1 and for the CQWLS model in Figure 2. Figure 1 exemplifies the increasing range of the residuals which is typical of heteroscedastic models. On the other hand, the residuals from the CQWLS model in Figure 2 show no sign of nonconstant variance. For comparison purposes it should be noted that the scales of Figures 1 and 2 are different. However, it is the overall shape and not the scale that is of interest.

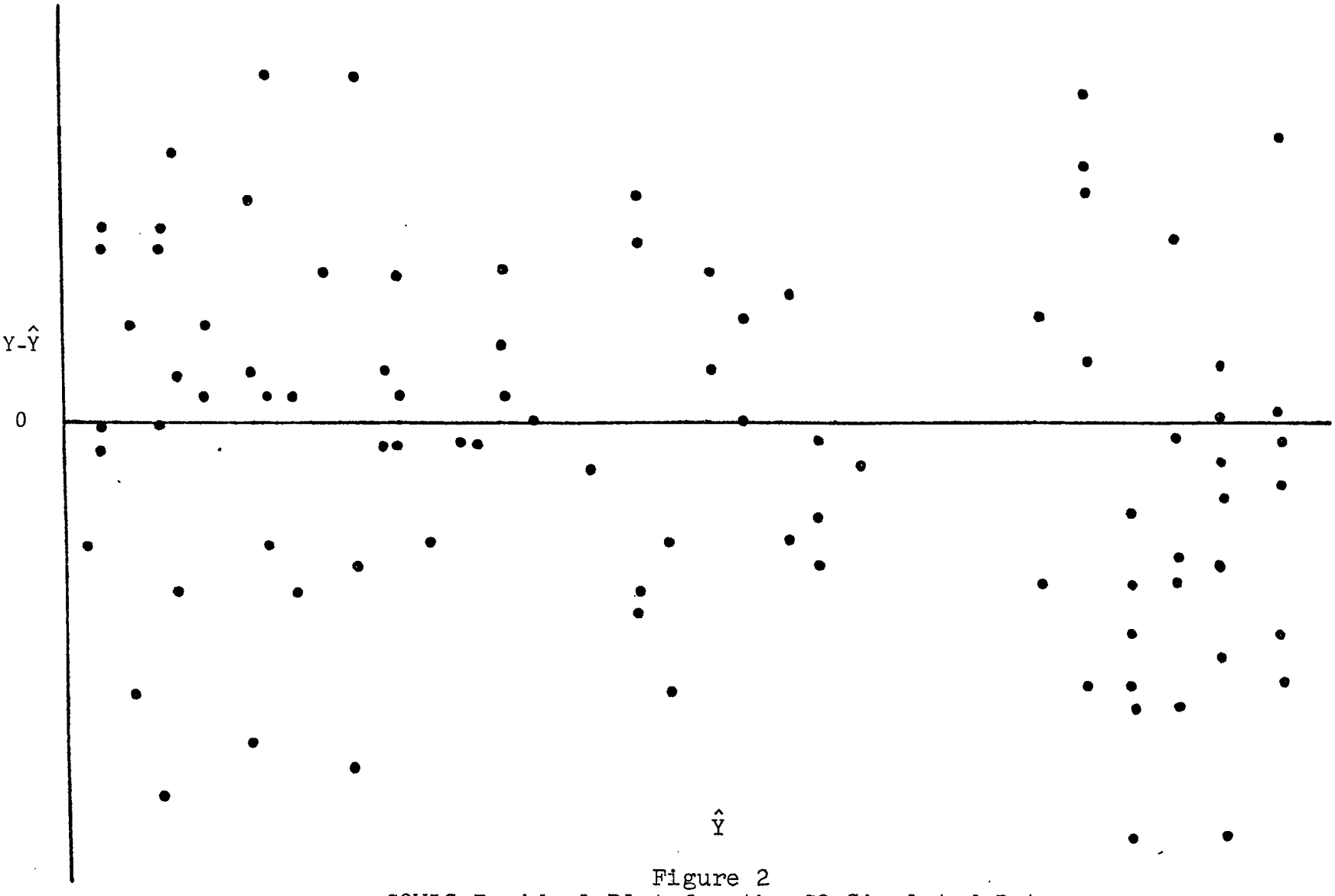

Figure 2
CQWLS Residual Plot for the CQ Simulated Data

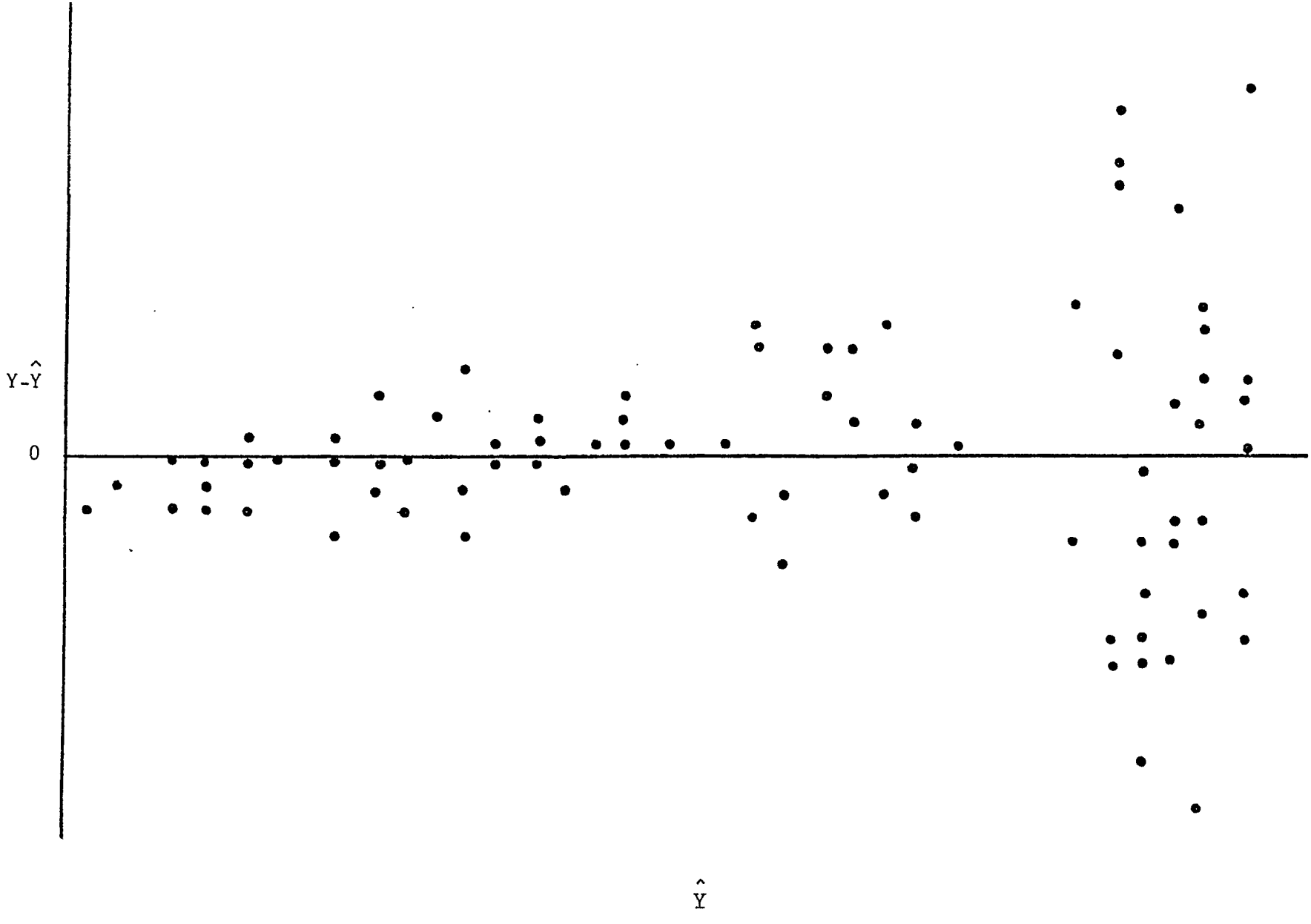

Figure 1
LS Residual Plot for the CQ Simulated Data

The final results for both the LS and the CQWLS cases are

	Actual	LS	CQWLS
b_1	1	6.08	1.02
b_2	1	0.67	0.96
% of error in b_1		508.00	2.00
% of error in b_2		37.00	4.00

The estimated dispersion matrix for the LS case is

$$\begin{pmatrix} 20.36 & -0.65 \\ -0.65 & 0.026 \end{pmatrix}$$

with an estimate of σ^2 of 353.82. The estimated dispersion matrix for the LS case is

$$\begin{pmatrix} 13.21 & -1.06 \\ -1.06 & 0.125 \end{pmatrix}.$$

The estimated variance and covariance of the estimators of β must be interpreted very carefully. For the model $Y = X\beta + e$ it is well known that $(X^TX)^{-1}\sigma^2$ is the dispersion matrix of the estimators of β. Thus $(X^TX)^{-1}s^2$ is an estimation of this matrix. This estimated dispersion matrix is with respect to the original factor space.

If the model is weighted by W^{-1}, where W^{-1} is the weight that will transform $D(ee^T)$ to I, then the weighted model becomes $W^{-1}Y = W^{-1}X\beta + W^{-1}e$.

An estimate of the dispersion matrix of the weighted estimators of β is $(X^TW^{-T}W^{-1}X)^{-1}$. The estimators for each

model have the same expected value, namely β, but their variances and covariances are not preserved under the transformation. Thus the matrix $(X^{T}W^{-T}W^{-1}X)^{-1}$ is an estimate of the dispersion matrix with respect to the weighted factor space, not the original space. Therefore, the estimate of the dispersion matrix from the LS analysis cannot be compared directly with the estimate of the dispersion matrix from the CQWLS analysis, since the independent variables are in different spaces.

A comparison can be made by considering the variance of the mean of future predictions at a known point. If S is a known point, then the predicted mean is $\hat{Y} = bS$. It is well known that if V(b) is the estimated dispersion matrix of b, then the estimated variance of $\hat{Y}$, $V(\hat{Y})$, is

$$V(\hat{Y}) = S^{T} V(b)S.$$

If the estimate of V(b) is with respect to a weighted space, and S is weighted by a scalar weight w, then

$$\hat{Y}_{W} = b(wS)$$

$$V(\hat{Y}_{W}) = (wS)^{T}V(b)(wS)$$

and

$$V(\hat{Y}_{W}) = w^{2}V(\hat{Y})$$

$$V(\hat{Y}) = w^{-2}V(\hat{Y}_{W}).$$

Thus if w is the appropriate weight to transform S into the weighted space, the $V(\hat{Y})$ in the original space may be computed by the above formula. The value of w for one of the points in the factor space may be obtained directly from the w^{-1} used to transform the data. Weights for other

points may be found by (4) where c, q and μ_i are replaced by their estimated value.

A meaningful way to study the relationships between the variance and the mean is to construct limits which give 95% confidence that 95% of the distribution lies within these limits. Such limits are called 95% tolerance limits. [28].

Tolerance limits were constructed and are shown in Figure 3 along with the data points.

Since the limits in Figure 3 are limits on the mean of future observations, not all the data points need to fall within the band. However, the expected value should be in them. For simulated data the expected value is known. The LS limits do not include the expected or actual value for $28 < X \leq 40$. In addition, they do not appreciably increase in width as the variance increases. However, the CQWLS bands always include the expected value and their width increases in the regions of large variance. In addition, when the variance is very small, CQWLS produces limits that are smaller than the LS limits. Therefore, one must conclude that the CQWLS method not only produces better point estimators but also produces better interval estimators for the CQ simulated data.

Gamma Simulated Data

The other simulated data set was constructed to conform to the gamma model (2). The same factor space that was used

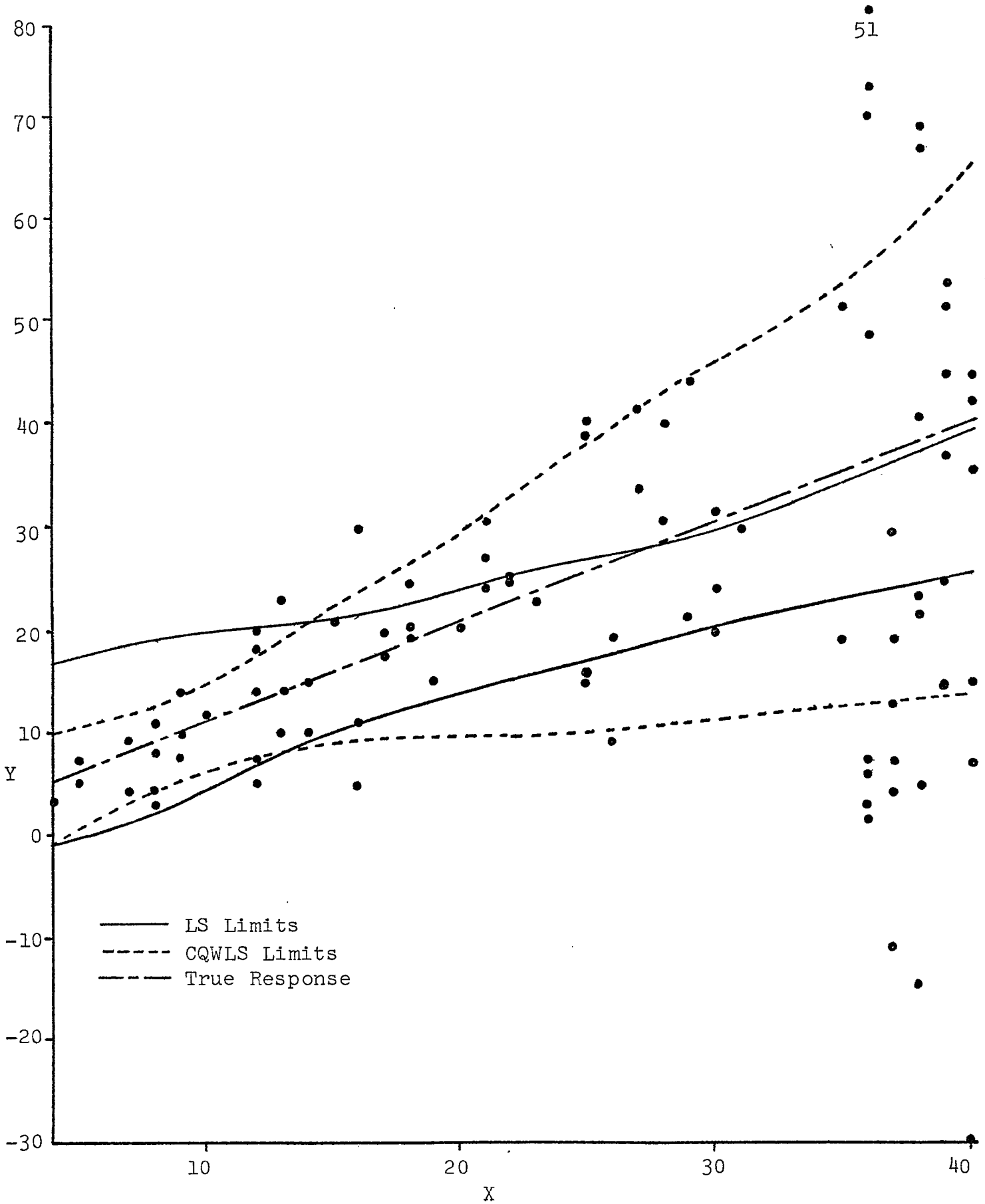

Figure 3
Tolerance Limits for the CQ Simulated Data

in the CQ simulation was used with $\gamma_1 = 0.1$, and $\gamma_2 = 0.1$. The simulated responses and actual standard deviations are summarized in Appendix 1. Note that this example is not as variable as the CQ simulated data.

Starting with the LS estimates the Rutemiller and Bowers procedure converged in 15 iterations to produce the following results:

$$b = \begin{pmatrix} 1.01 \\ 0.98 \end{pmatrix}$$

$$\gamma = \begin{pmatrix} 0.045 \\ 0.105 \end{pmatrix}$$

$$D(b) = \begin{pmatrix} 0.0593 & -0.0034 \\ -0.0034 & 0.00032 \end{pmatrix}$$

$$D(\gamma) = \begin{pmatrix} 0.029 & -0.0017 \\ -0.0017 & 0.00016 \end{pmatrix}$$

These values were then used to determine weights for the gamma weighted least squares analysis. The results are

	Actual	LS	GWLS	CQWLS
b_1	1	1.81	1.01	0.96
b_2	1	0.94	0.99	0.99
% error in b_1		82.0	1.0	4.0
% error in b_2		6.0	1.0	1.0
Weighted s^2			1.41	1.25

The estimated dispersion matrix for the LS estimators in the original factor space is

$$\begin{pmatrix} 0.55 & -0.018 \\ -0.018 & 0.0007 \end{pmatrix}$$

with an estimated variance of 9.69. The estimated dispersion matrix for the GWLS and CQWLS estimator in the respected weighted factor spaces are

$$\begin{pmatrix} 0.88 & -0.004 \\ -0.004 & 0.0004 \end{pmatrix}$$

and

$$\begin{pmatrix} 0.077 & -0.004 \\ -0.004 & 0.0004 \end{pmatrix}$$

The LS analysis produced results much better than for the CQ simulated data, perhaps due to the smaller variability of this data. The GWLS and CQWLS methods produced practically the same results, neither being able to drive s^2 as close to 1 as CQWLS did in the last example. However, comparing the estimates of β to the known values indicates that the values of s^2 produced (1.4 and 1.2) are close enough to 1 to warrant use of the weights. Although not reproduced here, the residual plot for LS is very much like Figure 1 and the plots for GWLS and CQWLS are very much like Figure 2.

The variance of $\hat{Y}$ at various points was calculated as before and 95% tolerance limits were set. These limits are summarized in Figure 4. The scale of Figure 4 is such that the GWLS and CQWLS limits cannot be discerned.

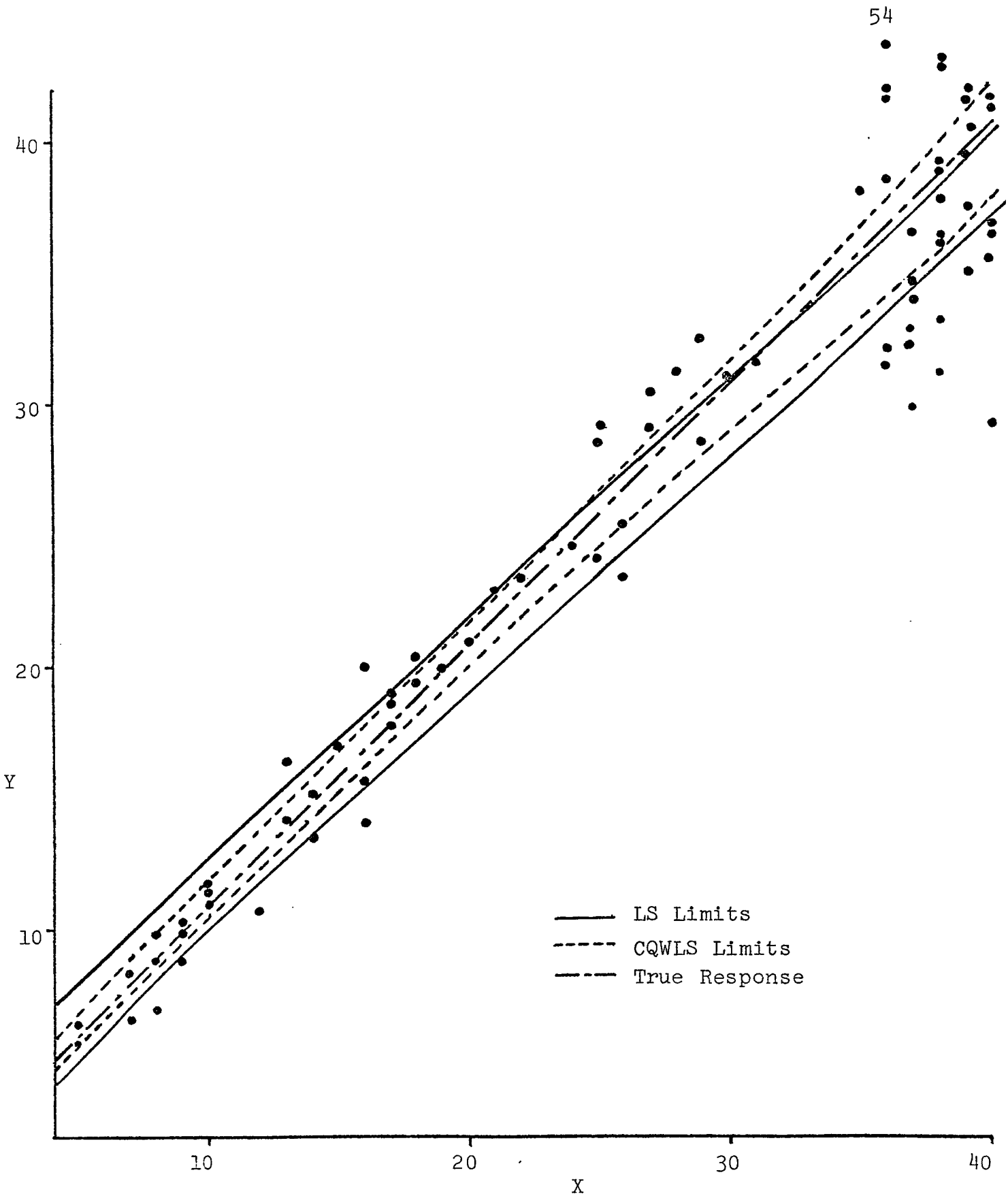

Figure 4
Tolerance Limits for the Gamma Simulated Data

At the upper level of X, the LS limits again did not include the actual value as they should 95% of the time. The GWLS and CQWLS limits are tighter than LS in the lower range of X and increase slightly as the variability increases.

It can be concluded that for this example GWLS and CQWLS produced practically the same results although the data was simulated according to the gamma model. The results produced were somewhat superior to the LS results but not as markedly improved as for the last example.

Radio Sales Data

The next set of data is actual data that relates the sales of radios in the continental states and the District of Columbia to the effective buying income. The data consists of 49 points for the model $Y = b_1 + b_2X$ and is recorded in [26]. The results of the three analyses are

	LS	GWLS	CQWLS
b_1	-11.79	-1.84	-1.79
b_2	0.032	0.029	0.029
Weighted s^2		3.92	3.20

The estimated dispersion matrix in the original factor space for LS is

$$\begin{pmatrix} 24.25 & -0.002 \\ -0.002 & 0.00000041 \end{pmatrix}$$

with an estimated variance of 686.64. The estimated

dispersion matrices in the appropriate weighted factor space for GWLS and CQWLS are

$$\begin{pmatrix} 12.89 & -0.004 \\ -0.004 & 0.0000035 \end{pmatrix}$$

and

$$\begin{pmatrix} 6.35 & -0.007 \\ -0.007 & 0.000014 \end{pmatrix}$$

respectively.

The slope b_2 was practically the same for each method, but the intercept, b_1, from LS was different from the intercept from the other two methods. Again both GWLS and CQWLS gave about the same results. Neither set of weights was able to drive s^2 to 1 but produced results of 3.9 and 3.2. However, these transformations were begun with $s^2 = 686.67$, for the data prior to transformation by W^{-1}.

The residual plot for LS residuals, Figure 5, displays the spread of variance that is characteristic of a heteroscedastic model. The residual plots from GWLS and CQWLS are quite similar and only the GWLS plot is shown in Figure 6. The most striking property exhibited in Figure 6 is not the constant band of residuals which indicates a constant variance, but the one observation that seems to be an outlier. The supposition that this point is indeed a sample from another distribution is strengthened by noting that this residual was generated by the District of Columbia data point and not from one of the data points from one of the states.

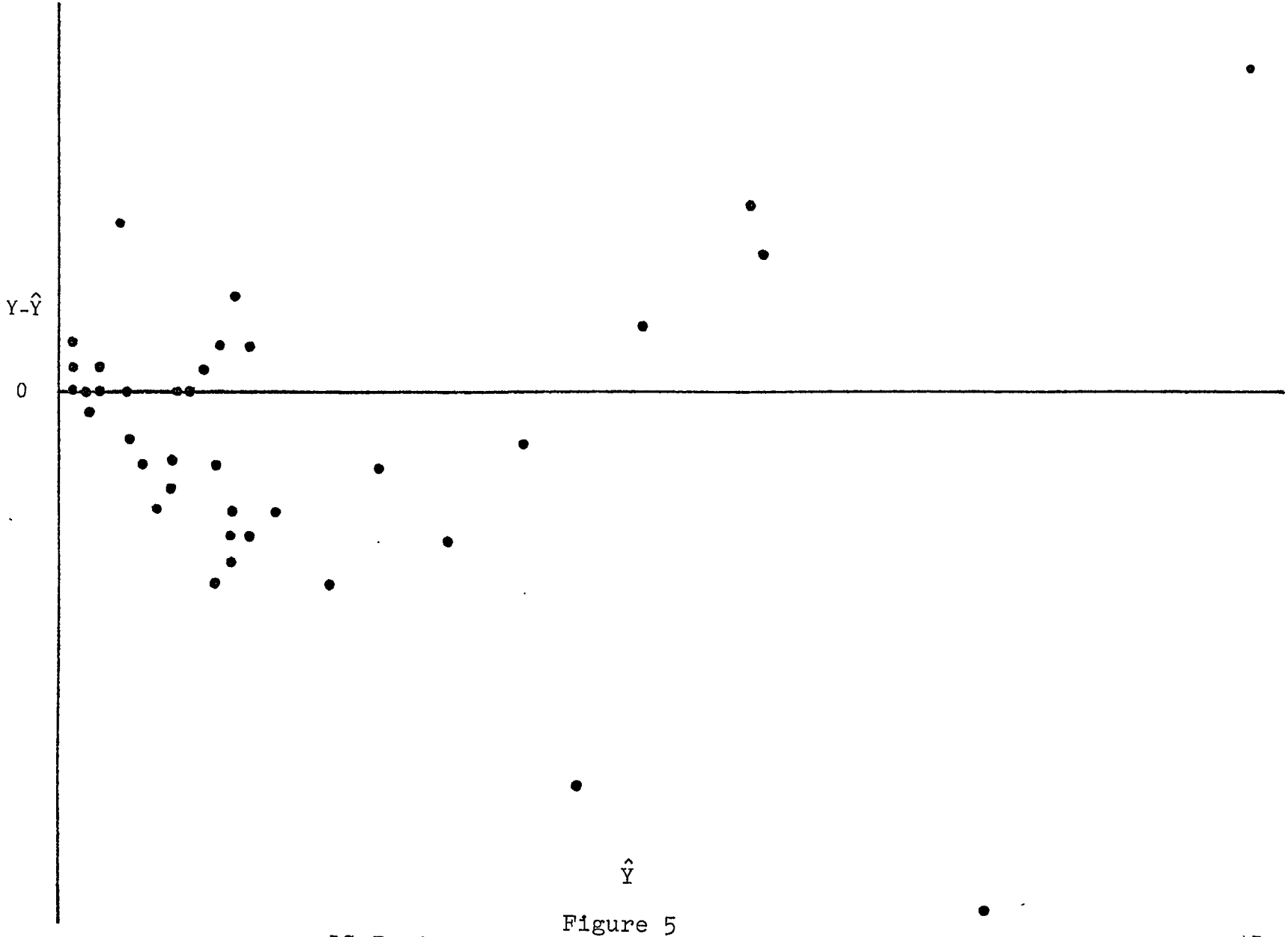

Figure 5
LS Residual Plot for the Radio Sales Data

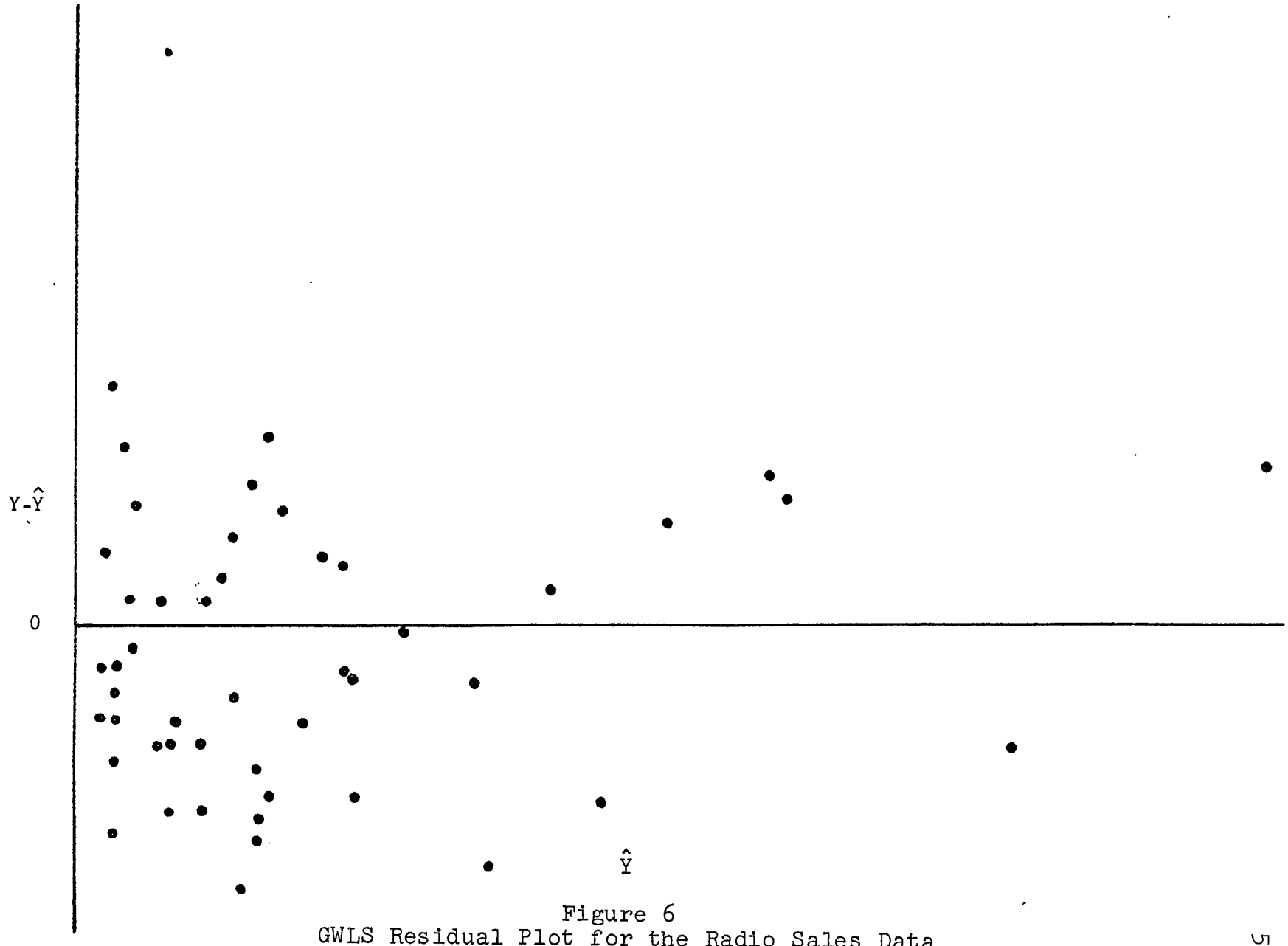

Figure 6
GWLS Residual Plot for the Radio Sales Data

Tolerance limits were calculated and the LS and CQWLS limits are shown in Figure 7. The CQWLS and GWLS limits differ somewhat. The GWLS limits are wider in the lower ranges of X and are tighter in the upper ranges of X. The GWLS limits are not plotted, but a comparison with CQWLS limits for some specific points is given in Appendix 2. The CQWLS and GWLS limits increase in width in the upper levels of X where an increase in variability is indicated by the residual plots, Figures 5 and 6. The LS limits do not have this property. Since this data is not simulated, the true relation is not known. Thus it is difficult to choose between the GWLS and CQWLS limits.

Stopping Distance Data

The last example is data relating the stopping distance of an automobile and its speed [25] for the model $Y = b_1 + b_2X + b_3X^2$. The final point estimate results are

	LS	GWLS	CQWLS
b_1	1.838	1.593	1.545
b_2	0.369	0.395	0.401
b_3	0.066	0.066	0.065
Weighted s^2		2.81	2.63

The LS estimated dispersion matrix in the original factor space is

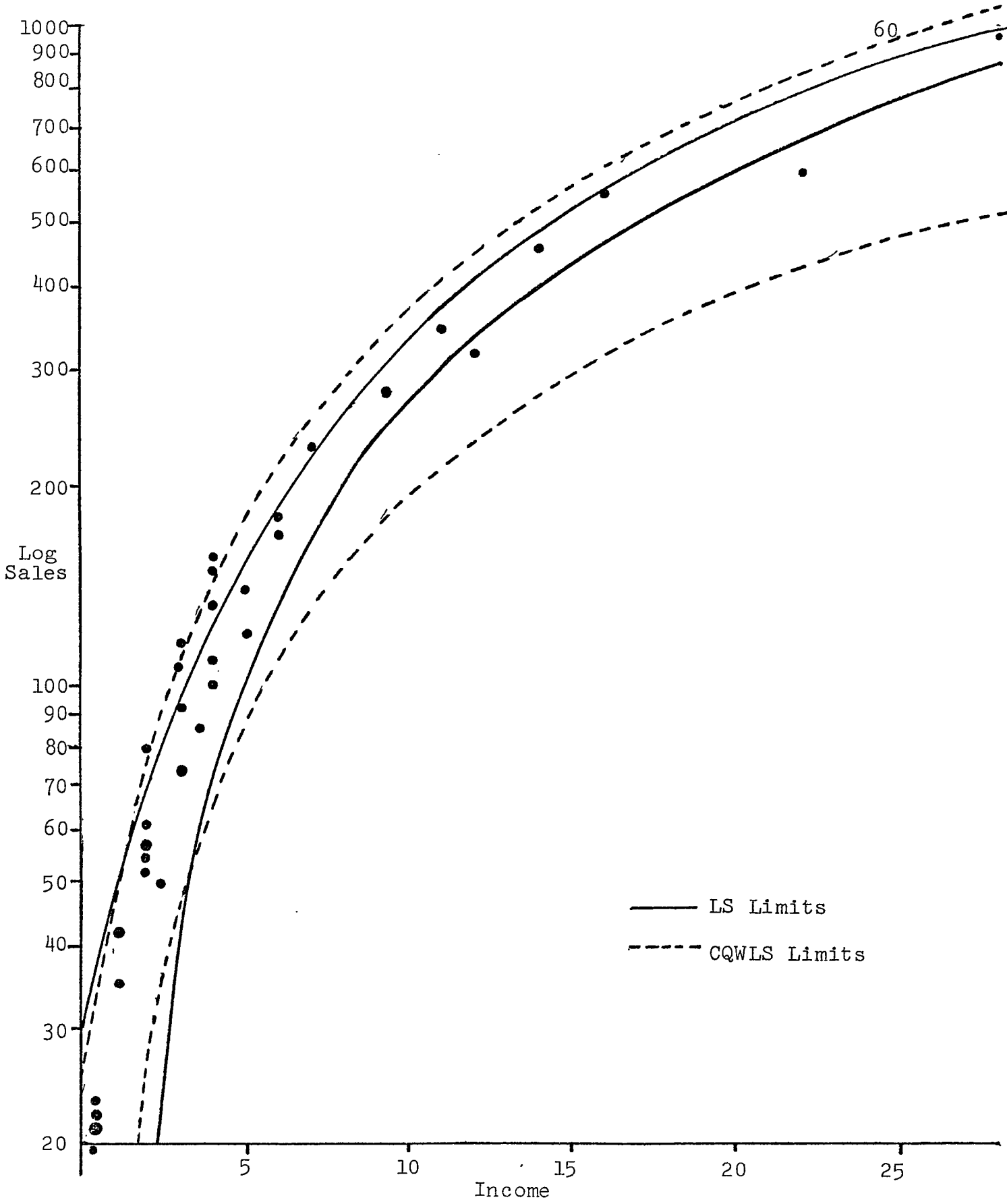

Figure 7
Tolerance Limits for the Radio Sales Data

$$\begin{pmatrix} 24.786 & -2.506 & 0.532 \\ -2.506 & 0.292 & -0.007 \\ 0.532 & -0.007 & 0.0001 \end{pmatrix}$$

with an estimate of σ^2 of 94.68. The GWLS estimated dispersion matrix in the GWLS weighted factor space is

$$\begin{pmatrix} 9.764 & -1.595 & 0.045 \\ -1.595 & 0.298 & -0.009 \\ 0.045 & -0.009 & 0.0003 \end{pmatrix}$$

and the matrix for CQWLS in the CQWLS weighted factor space is

$$\begin{pmatrix} 13.499 & -2.064 & 0.057 \\ -2.064 & 0.360 & -0.01 \\ 0.057 & -0.01 & 0.0003 \end{pmatrix}$$

The interval estimates are shown in Figure 8. Again the GWLS and CQWLS methods produced essentially the same results. The estimates of β differed from the LS estimates only slightly in the first component.

The residual plot for LS, Figure 9, indicates a non-constant variance, while the residual plot for CQWLS, Figure 10, indicates constant variance. Thus the increase in width of the weighted limits is more indicative of the true state than the narrower LS limits. Based primarily on the tolerance limits, either one of the weighted estimators is preferred over the LS estimators.

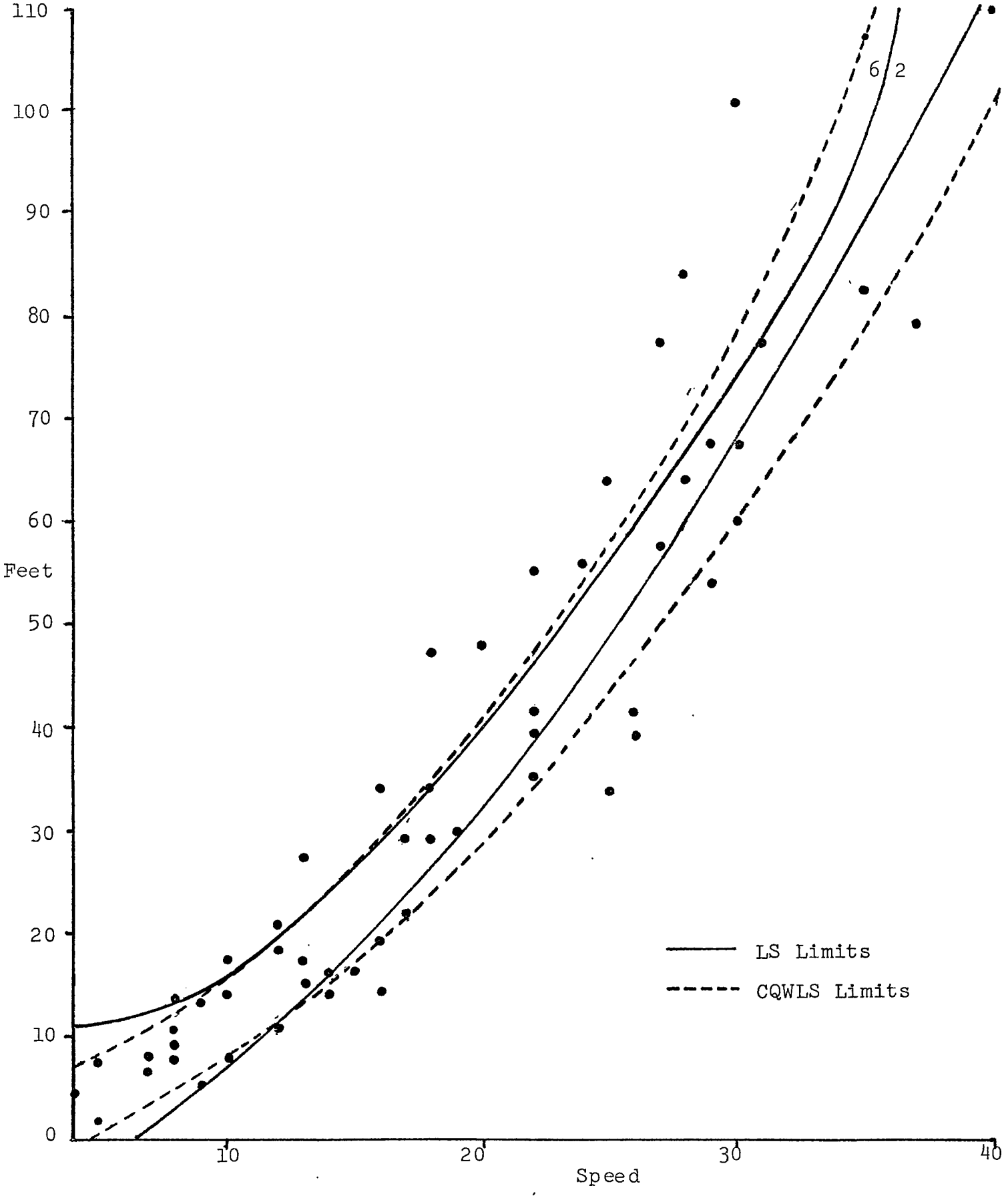

Figure 8
Tolerance Limits for the Stopping Distance Data

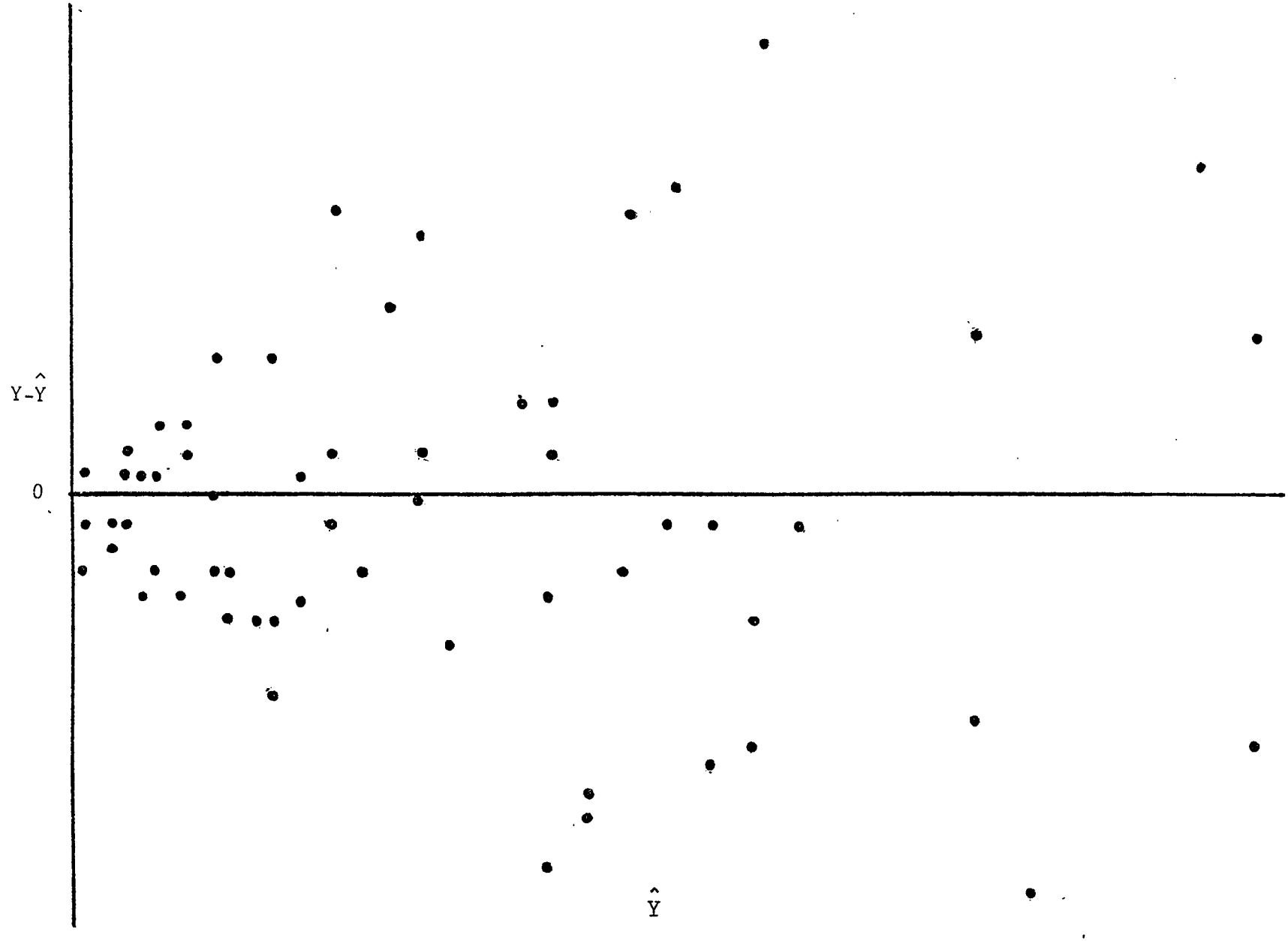

Figure 9
LS Residual Plot for the Stopping Distance Data

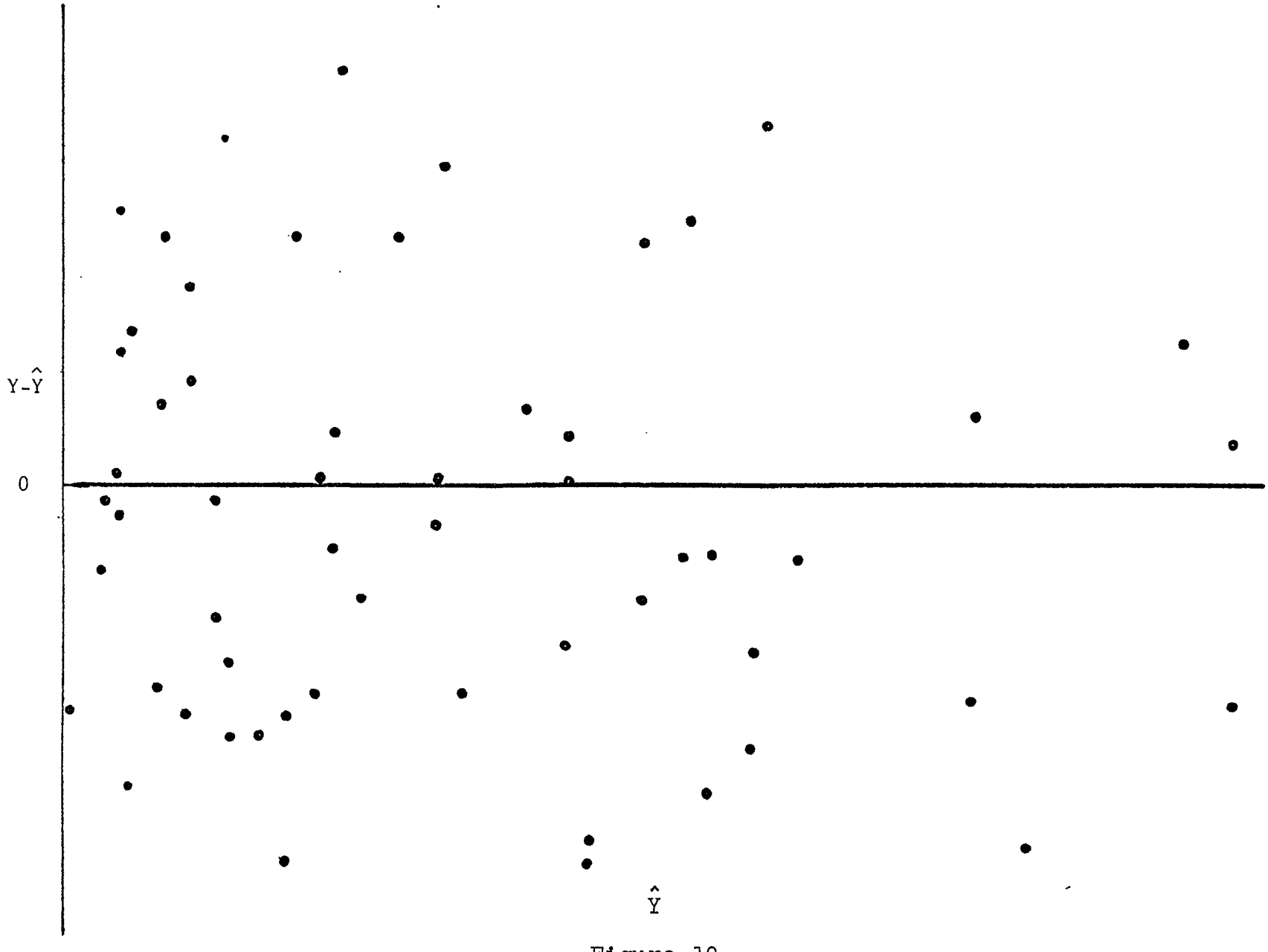

Figure 10
CQWLS Residual Plot for the Stopping Distance Data

Evaluation of Methods

The four examples discussed in the last section along with others not reproduced can be used to evaluate the methods. For the heteroscedastic model, the weighting methods produced better overall results than LS. The weighted point estimates were markedly better than the LS estimates for the CQWLS simulated data. However, in other cases, such as the stopping distance data, the weighted and LS point estimates were essentially identical. In these cases the weighted interval estimates were more indicative of the underlying process, that is, they were wider than the LS limits in regions of high variability.

It may be possible to subjectively classify data according to the number of data points relative to the variability of the data. If there is very little data available, then LS seems to produce as good or better results than a weighted method because there is not enough information to accurately estimate the weights. If an extremely large amount of data is available, LS might not use the data as efficiently as GWLS or CQWLS, but LS apparently uses the large amount of data to produce good estimates.

The GWLS simulated data has 93 data points and is not extraordinarily variable. Thus this data is an example of the large data class. Only when moderate amounts of data

relative to variability are available will weighted methods produce better estimates. An example of such a case is the CQWLS simulated data. It has the same number of data points as the GWLS simulated data, but in this case 93 points can be considered as a moderate amount of data due to the large variability that the data possesses in the upper region.

Although both GWLS and CQWLS are of value, there are some reasons to prefer CQWLS over GWLS even though they often produce essentially the same results. First, CQWLS does not require the assumption of normality nor the assumption that the standard deviation is a linear function of the independent variables. The GWLS procedure seems to be sensitive to the violation of these assumptions, especially the second. This assertion is exemplified by the failure of the GWLS procedure to converge with the CQWLS simulated data. Of course, there are assumptions on the nature of the standard deviation for the CQWLS model as well (4). In general, CQWLS assumes that the standard deviation is a non-linear function of the independent variables. However, for $q = 1$, this model reduces to a restrictive case of the GWLS model. The restriction is

$$\gamma_i = c\beta_i, \qquad i=1,2,\ldots,p.$$

Thus CQWLS reduces to some of the linear cases, while no special case of the GWLS model is non-linear.

Secondly, CQWLS is based entirely on the well-established concepts of least squares while GWLS is based on the method

of scoring. Thus it seems that it would be easier to explain CQWLS than GWLS to a potential user. This advantage of CQWLS is certainly less technical than the one discussed previously. However, it may be more important to the practitioner who is trying to persuade his managerial staff. In addition, CQWLS is relatively easy to add to an existing LS computer program while GWLS requires the programming of some rather difficult and not well-known equations.

The above advantages of CQWLS are inherent in the procedure, while the next two observations may only have occurred because of the specific examples chosen. In these examples, at any rate, the tolerance limits produced by CQWLS were slightly wider in the very variable regions. Also, the estimates of s^2 from CQWLS were always closer to the optimal value of one than the s^2 produced by GWLS.

From these examples one may conclude that if a weighted procedure is indicated, then CQWLS should be tried first. If it does not stabilize the variance, then the GWLS method may be tried.

Of course, weights should not be applied to the homoscedastic data. This data can be found by an examination of the residual plots or by one of the quantitative methods available in the literature [29, 30, 31].

CHAPTER V

AREAS OF FUTURE STUDY

The research described in the previous chapters has revealed some areas that would provide interesting research projects. This chapter will present these proposed studies and outline a procedure for conducting the research.

General Objectives

An interesting model studied is the heteroscedastic model $Y = X\beta + e$ where e is normal with zero mean and a diagonal dispersion matrix D with σ^2_j, $j=1,2,\ldots,n$, on the main diagonal where

$$\sigma_j = |c||\Sigma X_{ji}\beta_i|^q.$$

This model was described in detail in Chapter IV. This model was selected for future study for three reasons. First, it is the historical model used in establishing variance stabilizing transformations [20] such as the square root, log and reciprocal transformations. Observed data is often better analyzed by one of these methods [8]. Thus, actual processes seem to exist that can be reasonably approximated by the proposed model. Second, it is a generalization of a special case of Rutemiller and Bowers' model. Third, the model reduces to the $\sigma^2 I$ model when $q = 0$.

It is proposed that this model be used to predict future responses at points in a range, R. Then it is

desirable to determine estimates of β, c and q that will make good predictions. This study will be limited to linear estimators of β, that is, estimators of the form $b = AY$, where A is a $(p \times n)$ matrix.

Let S, a $(1 \times p)$ vector, be a point in the factor space. Then a prediction of the mean at S, $\hat{Y}_S$, is $\hat{Y}_S = Sb$ where b is the estimate of β. However, the prediction of the mean is not the only quantity of interest. The variance of the prediction of the mean at S, σ_S^2 is also important.

The best estimators of β, c and q might be defined as the estimators that produce the minimum varianced unbiased estimators of Y_S and σ_S^2. However, such estimators may not exist. Therefore, it may be better to define the best estimators at point S as the estimators that minimize the sum of the mean square error of Y_S and σ_S. That is,

$$(Y_S-\hat{Y}_S)^2 + (\sigma_S-\hat{\sigma}_S)^2. \qquad (2)$$

The standard deviation and not the variance is used so that both terms will have the same physical units.

If it is desired to predict within a given range R, then errors in the prediction for the entire range should be considered. Therefore, define the best estimators over R as the estimators that minimize

$$\int_{s \epsilon R} [(Y_S-\hat{Y}_S)^2 + (\sigma_S-\hat{\sigma}_S)^2]ds.$$

This definition appears difficult to apply. A related discrete measure will therefore be used. Let $S_1, S_2, \ldots, S_r$ be r prechosen points in R. Then the best estimators are those which minimize

$$\frac{\sum_{i=1}^{r} [(Y_{si} - \hat{Y}_{si})^2 + (\sigma_{si} - \hat{\sigma}_{si})^2]}{r} . \qquad (2)$$

Evaluating σ_s^2

The expression in (2) requires a knowledge of σ_s^2. For the class of linear estimators, that is, estimators of the form

$$b = AY \qquad (3)$$

where A is a $(p \times n)$ known matrix, the formula for the covariance of b_i and b_j may be derived. Let A_i be the i^{th} row of A, then

$$b_i = A_i Y.$$

Let A_{ij} be the $(i,j)^{th}$ element of A. The covariance of b_i and b_j as given by

$$\begin{aligned} \mathrm{Cov}(b_i, b_j) &= E(b_i b_j) - E(b_i)E(b_j) \\ &= E(A_i Y A_j Y) - E(b_i)E(b_j) \\ &= \sum_k \sum_\ell A_{ik} A_{j\ell} E(Y_k Y_\ell) - E(b_i)E(b_j). \end{aligned}$$

Since Y_k and Y_ℓ are uncorrelated,

$$\mathrm{Cov}(b_i,b_j) = \sum_k \sum_{\ell \neq k} A_{ik}A_{jk}\, E(Y_k)E(Y_\ell)$$

$$+ \sum_k A_{ik}A_{jk}[\sigma_k^2 + E(Y_k)^2] - E(b_i)E(b_j)$$

$$= \sum_k \sum_\ell A_{ik}A_{jk}\, E(Y_k)E(Y_\ell) - E(b_i)E(b_j)$$

$$+ \sum_k A_{ik}A_{jk}\sigma_k^2.$$

Thus

$$\mathrm{Cov}(b_i,b_j) = \sum_k A_{ik}A_{jk}\sigma_k^2. \tag{4}$$

For the model under study this becomes

$$\mathrm{Cov}(b_i,b_j) = c^2 \sum_k A_{ik}A_{jk}\,|\Sigma X_{k\ell}\beta_\ell|^{2q}. \tag{5}$$

The dispersion matrix for b, $D(\beta,c,q)$, depends on β, c and q. Then the variance at S is

$$\sigma_S^2 = SD(\beta,c,q)S^T \tag{6}$$

and

$$\hat{\sigma}_S^2 = SD(b,\hat{c},\hat{q})S^T. \tag{7}$$

Simulation Study

The complex nature of σ_S^2 indicates that isolating the best estimators by analytical methods would be very difficult, if not impossible. The best estimators may even be obtained by different methods for various types of data. However, it may be possible to develop an insight into the behavior of some estimators by a simulation study. This study will analyze the behavior of some possible estimators on

simulated data. Hopefully the results can be used to completely eliminate some of the estimators and develop criteria for selecting among the others. Some of the problems that will be encountered will be discussed in this section.

Estimators of β

Estimators of β that will be considered will have the general form of

$$b = AY.$$

Specifically, the study would consider LS, weighted LS, MSAD, weighted MSAD and Chebyshev's estimators. The earlier chapters of this thesis showed how the A matrix may be determined for all estimators except weighted MSAD.

The weighted MSAD estimates may be obtained by merely changing the objective function of the linear programming problem to a weighted sum of deviations. Thus the A matrix can be determined in the same manner as for MSAD. However, it is not obvious what form the optimal weights should have. Two types of weights appear reasonable. First, the weights could have the same form as the LS weights, that is, weighted inversely to the standard deviation. Second, since the optimal weights for LS are the reciprocal of the standard deviations, it seems that the reciprocal of the mean deviation, δ, would be good weights for MSAD. Herrey [7]

states that for the normal distribution the mean deviation is given by

$$\delta = \sqrt{2/\pi}\ \sigma.$$

Thus the weights $1/\delta_i$ could be used. As little as is known about the relationship between LS, MSAD, either type of weighted MSAD and Chebyshev's method, a preliminary study in which c and q are assumed known is proposed. Perhaps this study could at least determine the proper form of the weights for MSAD and exhibit data that could be analyzed better by a form of MSAD than by LS or weighted LS.

Estimates of c and q

There are two possibilities for estimating c and q. The first was described in Chapter IV in connection with the CQWLS method. This method produced reasonable estimates as shown by the examples.

In addition, it may be possible to determine maximum likelihood type estimators of c and q as an alternate to the first approach. First eliminate c from equations (5) and (6) of Chapter IV to yield

$$\frac{\sum_j (Y_j-\mu_j)|\mu_i|^{-2q}\ln|\mu_j|}{\sum_j \ln|\mu_j|} - \frac{[\sum(Y_j-\mu_j)^2|\mu_j|^{-2q}]^2}{2n} = 0 \quad (8)$$

where $\mu_j = \sum_i X_{ji}\beta_i$. If β were known, equation (8) could be solved numerically for q. The value of c could then be

found by

$$c = \frac{\Sigma(Y_j - \mu_j)^2 |\mu_j|^{-2q}}{n}$$

which is equation (6) of Chapter IV. The valuesof c and q so obtained would be the maximum likelihood estimates. However, if β is estimated, the values of c and q are the maximum likelihood estimates with respect to the assumed value of β. This type of estimate of c and q will be called the relative maximum likelihood estimates and can be found if (8) can be solved numerically.

When weights are estimated, an iterative procedure for estimating c and q may be used. Let b^i denote the i^{th} estimation of β, $\hat{c}^i$ and $\hat{q}^i$ be the i^{th} relative maximum likelihood estimations of c and q. Use the LS or MSAD estimates for b^1. Find $\hat{c}^i$ and $\hat{q}^i$. Use these values to estimate weights and use the weighted estimate of β for b^{i+1}. The iterations are repeated for i=1,2,..., and terminated when the weight matrix approaches the identity.

Since the relative maximum likelihood estimators have not been studied, a comparison with the CQWLS estimators has not been made. Therefore, another preliminary study needs to be made to determine the best method of estimating c and q.

Selection of Estimators

Perhaps selecting the best estimators for a given set of data is the most difficult problem that will be encountered in the study. Hopefully the simulation results can reduce the list of possible estimators to only a few candidates. The choice between the remaining estimators will have to be based on observable quantities. Some of the quantities that appear meaningful are the number of data points, n, the number of elements in β, p, the distribution of the data in the factor space and some measure of the behavior of the variances over the factor space. These variables, possibly along with some others, could be used as independent variables in a regression analysis to predict the sum of the mean square errors of Y_s and σ_s. Possibly a separate prediction equation would have to be developed for each type of estimator. Then the estimator that had the lowest predicted mean square error could be selected.

Two of the variables need to be quantified in order to be used in the regression analysis. These variables are the distribution of the data in the factor space and the variability of the data.

It appears that more data is needed in regions of high variance than in regions of low variance. Thus, the distribution of the data points relative to the predicted mean seems important. One approach to quantify this property is concerned

with the range of the predicted responses. Let Y_{max} be the largest predicted mean in the entire region of interest and let Y_{min} be the minimum of such predictions. Define Y_λ as the point such that λ percent of the range of the prediction of Y is less than Y_λ, that is,

$$Y_\lambda = \frac{\lambda}{100} (Y_{max}-Y_{min}) + Y_{min}.$$

The distribution of the data could be measured by the ratio of the number of data points that yield predictions between Y_{80} and Y_{100} to the number of data points that predict between Y_0 and Y_{20}.

In a similar way several measures of the behavior of the variance can be presented. Define the coefficient of variation from $Y_{\lambda 1}$ to $Y_{\lambda 2}$ as

$$\psi(\lambda_1,\lambda_2) = \frac{S(\lambda_1,\lambda_2)}{\mu(\lambda_1,\lambda_2)}$$

where $S(\lambda_1,\lambda_2)$ is the estimated standard deviation of the prediction between Y_{λ_1} and Y_{λ_2} and $\mu(\lambda_1,\lambda_2)$ is the mean of the predictions between Y_{λ_1} and Y_{λ_2}. Some measures of the heteroscedasticity are $\psi(80,100)/\psi(0,20)$, $\psi(80,100)-\psi(0,20)$, $\psi(\lambda_{max},\lambda_{max}+10)/\psi(\lambda_{min},\lambda_{min}+10)$ where λ_{max} is the value from the set $\{0,10,20,30,40,50,60,70,80,90\}$ that maximizes $\psi(\lambda,\lambda+10)$ and λ_{min} is the value from the same set that minimizes $\psi(\lambda,\lambda+10)$.

In order for the regression curves that will be used for selection of estimators to be generally applicable,

the simulated data used in the study needs to vary over fairly large ranges. Numerical experience gained from the research already conducted indicates that n should vary from 10 to about 150, p from 2 to 5, c needs to be small, say .1 or .05, q ranges from 0 to 3 and the distribution of the data points should vary from data with high concentration in the low ranges of $\hat{Y}$ to evenly distributed data with high concentration in the upper ranges of $\hat{Y}$. Later it may be advisable to extend the range of p. It is felt that these ranges would make the selection curves valid over a reasonable range of applicability.

BIBLIOGRAPHY

1. Graybill, F. A., An Introduction to Linear Statistical Models, McGraw-Hill, New York, 1961.

2. Draper, N. R. and Smith, H., Applied Regression Analysis, John Wiley and Sons, New York, 1966.

3. Allen, David M., "Mean Square Error of Prediction as a Criterion for Selecting Variables," Technometrics, pp. 469-481 (August, 1971).

4. Toro-Vizcarrondo, Carlos and Wallace, T. D., "A Test of the Mean Square Error Criterion for Restrictions in Linear Regression," Journal of the American Statistical Association, Vol. 63, pp. 558-572 (1968).

5. Koltz, J. H., Milton, R. C. and Zacks, S., "Mean Square Efficiency of Estimators of Variance Components," Journal of the American Statistical Association, Vol. 69, p. 1383 (1969).

6. Fisher, R. A., "A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error and by the Mean Square Errors," Monthly Notes of the Royal Astronomical Society, Vol. 80, pp. 758-770.

7. Herrey, E. M. J., "Confidence Intervals Based on the Mean Absolute Deviations of a Normal Sample," Journal of the American Statistical Association, Vol. 60, pp. 118-132 (1965).

8. Daniel, C. and Wood, F. S., Fitting Equations to Data, Wiley-Interscience, New York, 1971.

9. Crocker, D. C., "Linear Programming Techniques in Regression Analysis: The Hidden Danger," AIIE Transactions, Vol. 1, pp. 112-125 (1969).

10. Wagner, H. M., "Linear Programming Techniques for Regression Analysis," Journal of the American Statistical Association, Vol. 54, pp. 206-212 (1959).

11. Dantzig, G. B., Linear Programming and Extensions, Princeton University Press, 1963.

12. Karst, O. J., "Linear Curve Fitting Using Least Deviations," Journal of the American Statistical Association, Vol. 53, pp. 118-132 (1958).

13. Fisher, W. D., "A Note on Curve Fitting with Minimum Deviations by Linear Programming," Journal of the American Statistical Association, Vol. 56, pp. 339-362 (1961).

14. Lelewellyn, R. W., Linear Programming, Holt, Rinehart and Winston, New York, 1964.

15. Rao, C. R., Linear Statistical Inference and Its Applications, John Wiley and Sons, New York, 1965.

16. Graybill, F. A., Introduction to Matrices with Applications in Statistics, Wadsworth, Belmont, California, 1969.

17. Hillier, F. S. and Lieberman, G. J., Introduction to Operations Research, Holden-Day, Inc., San Francisco, 1967.

18. Tukey, J. W., "A Survey of Sampling from Contaminated Distributions," Technical Report No. 33, Department of Mathematics, Princeton University, Princeton, New Jersey, 1959.

19. Forsythe, A. B., "Robust Estimation of Straight Line Regression Coefficients by Minimizing p^{th} Power Deviations," Technometrics, Vol. 14, pp. 159-168 (1972).

20. Brownlee, K. A., Statistical Theory and Methodology in Science and Engineering, John Wiley and Sons, New York, 1965.

21. Alloin, G. A., "A Simplex Method for a Class of Nonconvex Separable Problems," Management Science, Vol. 17, pp. 66-77 (1970).

22. Williams, J. S., "The Variance of Weighted Regression Estimators," Journal of the American Statistical Association, Vol. 62, pp. 1290-1301 (1967).

23. Wald, A., "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large," Transactions, American Mathematical Society, Vol. 54, p. 426 (1943).

24. Rutemiller, H. C. and Bowers, D. A., "Estimation in a Heteroscedastic Regression Model," Journal of the American Statistical Association, Vol. 63, pp. 552-557 (1968).

25. Ezekiel, M. and Fox, K. A., Methods of Correlation and Regression Analysis, John Wiley and Sons, New York, 1959.

26. O'Hara, J. B. and McClelland, R. C., Effective Use of Statistics in Accounting and Business, Holt, Rinehart and Winston, New York, 1954.

27. System/360 Scientific Subroutine Package, Version III, Programmer's Manual, International Business Machines, H20-0205-3, 1968.

28. Hald, A., Statistical Theory with Engineering Applications, John Wiley and Sons, 1952.

29. Glejser, H., "A New Test for Heteroskedasticity," Journal of the American Statistical Association, Vol. 64, pp. 316-323 (1969).

30. Goldfeld, S. M. and Quandt, R. E., "Some Tests for Homoskedasticity," Journal of the American Statistical Association, Vol. 60, pp. 539-547 (1965).

31. Park, R. E., "Estimation with Heteroscedastic Error Terms," Econometrica, Vol. 34, p. 888 (1966).

APPENDIX 1

CQ AND GAMMA DATA

X	CQ σ_j	CQ Y	γ σ_j	γ Y
4	1.11	3.34	.5	4.26
5	1.46	6.96	.6	6.39
5	1.46	5.29	.6	5.71
5	1.46	5.04	.6	5.60
5	1.46	7.16	.6	6.47
7	2.26	8.95	.8	8.33
7	2.26	3.91	.8	6.55
8	2.70	3.02	.9	7.00
8	2.70	11.60	.9	9.86
8	2.70	11.12	.9	9.70
8	2.70	8.54	.9	8.84
9	3.16	14.16	1.0	11.31
9	3.16	10.49	1.0	10.15
9	3.16	6.74	1.0	8.97
10	3.64	11.11	1.1	11.03
10	3.64	11.49	1.1	11.14
10	3.64	12.55	1.1	11.46
12	4.68	18.35	1.3	14.48
12	4.68	14.45	1.3	13.40
12	4.68	5.38	1.3	10.88
13	5.23	23.25	1.4	16.47
13	5.23	10.87	1.4	13.16
13	5.23	14.69	1.4	14.18
14	5.80	9.80	1.5	13.66
14	5.80	15.62	1.5	15.16
15	6.40	21.08	1.6	17.27
16	7.00	30.29	1.7	20.22
16	7.00	12.37	1.7	15.87
16	7.00	5.21	1.7	14.14
17	7.63	20.24	1.8	18.52
17	7.63	17.54	1.8	17.89
18	8.28	19.16	1.9	19.03
18	8.28	20.67	1.9	19.38
18	8.28	25.75	1.9	20.54
19	8.94	15.87	2.0	19.07
20	9.62	21.31	2.1	21.06
21	10.31	27.20	2.2	23.11
21	10.31	24.18	2.2	22.46
21	10.31	31.31	2.2	23.98
22	11.03	25.18	2.3	23.45
24	12.50	23.81	2.5	42.76
25	13.25	42.87	2.6	29.30
25	13.25	15.15	2.6	23.87
25	13.25	16.38	2.6	24.11
25	13.25	39.55	2.6	28.65

APPENDIX 1 (continued)

X	CQ σ_j	CQ Y	γ σ_j	γ Y
26	14.03	9.13	2.7	23.56
26	14.03	19.47	2.7	25.56
27	14.81	42.21	2.8	30.68
27	14.81	34.10	2.8	29.15
28	15.61	32.72	2.9	29.69
28	15.61	41.04	2.9	31.23
29	16.43	44.55	3.0	32.65
29	16.43	22.87	3.0	28.69
30	17.26	20.81	3.1	29.17
30	17.26	32.17	3.1	41.21
30	17.26	24.53	3.1	29.83
31	18.10	30.50	3.2	31.73
35	21.60	52.88	3.6	38.81
35	21.60	19.84	3.6	33.33
36	22.50	7.43	3.7	32.14
39	25.29	10.45	4.0	35.32
40	26.25	36.56	4.1	40.30
40	26.25	-33.51	4.1	29.36
*36	22.50	71.11	3.7	42.60
36	22.50	73.38	3.7	42.98
36	22.50	81.41	3.7	44.30
36	22.50	4.81	3.7	31.70
36	22.50	48.17	3.7	38.83
36	22.50	3.56	3.7	31.50
37	23.42	4.29	3.8	32.53
37	23.42	19.02	3.8	34.92
37	23.42	13.42	3.8	34.01
37	23.42	7.69	3.8	33.08
37	23.42	-11.56	3.8	29.95
37	23.42	30.27	3.8	36.74
38	24.35	42.33	3.9	39.53
38	24.35	22.18	3.9	36.30
38	24.35	67.48	3.9	43.56
38	42.35	4.85	3.9	33.53
38	24.35	69.47	3.9	43.87
38	24.35	23.90	3.9	36.58
39	25.29	54.75	4.0	42.33
39	25.29	52.58	4.0	41.99
39	25.29	25.51	4.0	37.71
39	25.29	37.76	4.0	39.64
39	25.29	45.75	4.0	40.91
39	25.29	-14.87	4.0	31.32
40	26.25	7.84	4.1	35.82
40	26.25	89.16	4.1	48.52
40	26.25	44.73	4.1	41.58
40	26.25	14.69	4.1	36.89
40	26.25	45.95	4.1	41.77
40	26.25	15.19	4.1	36.97

* Points below the line were added to the factor space in [10].

APPENDIX 2

GWLS AND CQWLS TOLERANCE LIMITS

POINT	X	GWLS LOWER	GWLS UPPER	CQWLS LOWER	CQWLS UPPER
43	388	2	17	5	13
18	5933	148	192	119	221
12	11367	281	374	226	430
9	16968	418	562	336	645
47	22505	553	748	445	858
8	28220	693	940	557	1078