Statistica Sinica 2(1992), 237-254

M-ESTIMATION OF MULTIVARIATE LINEAR REGRESSION PARAMETERS UNDER A CONVEX DISCREPANCY FUNCTION

Z. D. Bai, C. Radhakrishna Rao and Y. Wu

Temple University, Penn State University and York University

Abstract: There is vast literature on M-estimation of linear regression parameters. Most of the papers deal with special cases by choosing particular discrepancy functions to be minimized or particular estimating equations. A few discuss general results, but prove results under heavy assumptions which seem to exclude important special cases. In this paper, a general theory of M-estimation is developed using a convex discrepancy function under what appear to be a necessary set of assumptions to develop a satisfactory asymptotic theory. Detailed proofs are given for establishing the asymptotic normality of the distribution of M-estimates and the results are applied to several particular cases. Appropriate criteria are developed for tests of hypotheses concerning regression parameters. The problem is discussed in the multivariate situation which includes the univariate case.

Key words and phrases: Gauss-Markoff linear model, least distances estimation, least squares estimation, M-estimation, multivariate linear model.

1. Introduction

Consider a general p-variate regression model

$$Y_i = X_i'\beta + E_i, \qquad i = 1, \dots, n \tag{1.1}$$

where E_i are iid p-vectors, X_i are $m \times p$ given matrices. (When we consider the univariate case of the model (1.1) with $p = 1$, we write lower case letters y_i, x_i and e_i in the place of Y_i, X_i and E_i.) It may be noted that the model (1.1) is more general than the usual multivariate Gauss-Markoff linear model

$$Y_i = Bx_i + E_i, \qquad i = 1, \dots, n \tag{1.2}$$

where B is a $p \times a$ matrix of parameters and x_i is an $a \times 1$ vector of concomitant (or design) variables. Note that (1.2) can be written in the form (1.1) defining $\beta = \text{vec}B$ and $X_i' = I \otimes x_i'$.

For the model with $p = 1$, Huber (1964, 1973) introduced what is called an M-estimate of β defined as any value of β minimizing

$$\sum_{i=1}^{n} \rho(y_i - x_i'\beta) \tag{1.3}$$

for a suitable choice of the function ρ, or any value β satisfying the estimating equation

$$\sum_{i=1}^{n} \psi(y_i - x_i'\beta)x_i = 0 \tag{1.4}$$

for a suitable choice of the ψ function. A natural method of obtaining the estimating equation (1.4) is by taking the derivative of (1.3) with respect to β when ρ is continuously differentiable and equating it to the null vector. However, in general one can use any suitably chosen function ψ and set up the equations (1.4).

There is considerable literature devoted to the asymptotic theory of M-estimation under some assumptions on the ρ and ψ functions. Reference may be made to papers by Huber (1964, 1973, 1981, 1987), Relles (1968), Jurečkova (1971), Jackel (1972), Bickel (1975), Heiler and Willers (1988) and others. The particular case of $\rho(x) = |x|$ has been extensively studied. See, for instance, papers by Bassett and Koenker (1978), Amemiya (1982), Bloomfield and Steiger (1983), Dupacõva (1987), Bai, Chen, Wu and Zhao (1990), Bai, Rao and Yin (1990) and Bai, Chen, Miao and Rao (1990). Most of the papers cited above discuss particular choices of ρ and ψ, or general ρ and ψ under some restrictive conditions which do not cover important special cases.

In this paper we attempt to provide a general theory of M-estimation, defining an M-estimate of β in the model (1.1) as any value of β minimizing

$$\sum_{i=1}^{n} \rho(Y_i - X_i'\beta) \tag{1.5}$$

for any *convex function* ρ, under what we believe to be a minimal set of conditions on X_i and the random error E_i. It is well known that an M-estimate so defined exists although it may not be unique. Our results are true for any choice of such an estimate. We are aware that it is more satisfying to consider an expression of the type

$$\sum_{i=1}^{n} \rho\Big(\Sigma^{-1/2}(Y_i - X_i'\beta)\Big) + \frac{n}{2}\log\Big|\Sigma\Big| \tag{1.6}$$

introducing a scale parameter matrix Σ, for minimization and also to extend the results to non-convex functions. We hope to consider these problems in future communications. In particular, we note that our method can be easily extended to any ρ function which is a difference of two convex functions. This covers most of the ρ functions considered by earlier writers.

In a recent paper, Koenker and Portnoy (1990) consider a discrepancy function of the type: $\rho(x_1, \ldots, x_p) = \rho_*(x_1) + \cdots + \rho_*(x_p)$ where ρ_* is a univariate convex function. In our discussion, we consider a general convex function of p variables.

In Section 2, we state the main theorems in the multivariate case and the basic assumptions under which they are proved. In Section 3, a more general form of the multivariate regression model is introduced and more general theorems are established from which those stated in Section 2 are deduced as special cases.

We use the following notations in the discussion of Sections 2 and 3

$X_1, \ldots, X_n$ denote $m \times p$ matrices,
β denotes an m-vector of regression coefficients,
$Y_1, \ldots, Y_n$ are vector response measurements corresponding to the values $X_1, \ldots, X_n$,
ρ is a convex function of p variables,
ψ is a choice of a derivative of ρ, a $p \times p$ matrix-valued function,
$E_1, \ldots, E_n$ are the components of error.

In Section 4, we give a further discussion of the problem of testing of hypotheses and in Sections 5 and 6, we consider some examples.

2. Statements of Main Theorems and Assumptions

We consider the linear model (1.1)

$$Y_i = X_i'\beta + E_i, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

where X_i' is the transpose of X_i, and the problem of estimating β by minimizing

$$\sum_{i=1}^{n} \rho(Y_i - X_i'\beta). \tag{2.2}$$

Let $\hat{\beta}_n$ be any value which minimizes (2.2) and β_0 be the true value of β. (To simplify the notation we drop the suffix n and represent $\hat{\beta}_n$ simply by $\hat{\beta}$.) We develop the asymptotic theory (as $n \to \infty$) concerning $\hat{\beta}$ and tests of hypotheses on β_0. The following assumptions are made.

(M$_1$). $\rho(Z)$ is a convex function of p variables.

Let $\psi(Z)$ be any choice of the subgradient of $\rho(Z)$ and denote by $\mathcal{D}$ the set of discontinuity points of ψ, which is the same for all choices of ψ.

(M_2). The common distribution function F of E_i satisfies $F(\mathcal{D}) = 0$.

This condition is imposed to provide unique values for certain functionals of ψ which appear in the discussion, and it automatically holds when ρ is differentiable. (For instance if $\rho(x) = |x|^r$, $r > 1$, the condition does not impose any restriction on F.)

(M_3).
$$E[\psi(E_1 + C)] = AC + o(\|C\|) \quad \text{as} \quad \|C\| \to 0 \tag{2.3}$$

where C is a p-vector, $A > 0$ is a $p \times p$ constant matrix and $\|\cdot\|$ denotes the Euclidean norm of matrices or vectors.

It is easy to see that if (M_3) holds for one choice of ψ, then it holds for all choices of ψ with the same constant matrix A. This follows from the assumption (M_2).

(M_4).
$$g(C) = E\|\psi(E_1 + C) - \psi(E_1)\|^2 \tag{2.4}$$

exists for all sufficiently small $\|C\|$, and g is continuous at $C = 0$.

(M_5).
$$E[\psi(E_1)\psi'(E_1)] = B > 0. \tag{2.5}$$

(M_6). $S_n = X_1X_1' + \cdots + X_nX_n'$ is nonsingular for $n \geq n_0$ (some value of n) and

$$d_n^2 = \max_{1\leq i\leq n} \operatorname{tr}[X_i'S_n^{-1}X_i] \to 0 \quad \text{as} \quad n \to \infty. \tag{2.6}$$

The most recent results on M-estimation relevant to our discussion are due to Yohai and Maronna (1979) and Basawa and Koul (1988). Before stating our results, we may point out some differences between the assumptions made by these authors and ours.

Yohai and Maronna (1979) adopt the definition of an M-estimate as a solution of an estimating equation

$$\sum_{i=1}^{n} \psi(y_i - x_i'\beta)x_i = 0 \tag{2.7}$$

and make assumptions on ψ. Our conditions on ψ defined as a chosen subgradient of ρ may look similar to those of Yohai and Maronna, but the most noticeable differences are their (A_1), (A_2) in the place of our (M_1) and (M_2) and their extra condition, the second part of their (C_1), which has no counterpart in ours. Further, their condition (A_2) seems to exclude estimating equations obtained by equating the derivative of

$$\sum_{i=1}^{n} |y_i - x_i'\beta|^p, \quad p > 2$$

to zero, and hence their theory does not cover the L_p-norm estimates for $p > 2$.

In our case, an M-estimate obtained by minimizing (2.2) always exists and has the stated properties. Yohai and Maronna have not given a formal proof of the existence of a solution of (2.7) under their conditions. Of course, continuity of ψ would ensure the existence of a solution, but no specific assumption is made about continuity of ψ in their paper. We do not use these additional conditions, and in this sense our results seem to be more general.

On the other hand, Basawa and Koul adopt our definition (2.2) of an M-estimate; and their main assumptions for the regression problem (their Example 7) appear to be somewhat closer to ours. However, they do not provide details of the proofs but state that some extra smoothness conditions are necessary to prove their results, such as strict convexity of ρ, existence of bounded second derivative of ψ or existence of a bounded density function for the distribution of E_i, and nondecreasing behavior and right continuity of ψ. We do not need these conditions. They have not stated any assumption similar to our (M_4), which we think is necessary.

Now, we state our main results for the multivariate regression model based on the definition (2.2) for M-estimation.

Theorem 2.1. *Under the assumptions* (M_1)-(M_6), *for any fixed* $c > 0$,

$$\sup_{|T_n^{1/2}(\beta-\beta_0)|\leq c}\Big|\sum_{i=1}^{n}[\rho(Y_i - X_i'\beta) - \rho(Y_i - X_i'\beta_0) + (\beta-\beta_0)'X_i\psi(Y_i - X_i'\beta_0)] - \frac{1}{2}(\beta-\beta_0)'K_n(\beta-\beta_0)\Big| \to 0 \quad \text{in probability} \tag{2.8}$$

where β_0 *is the true value for the model* (2.1), *and*

$$T_n = \sum_{i=1}^{n} X_iBX_i', \quad K_n = \sum_{i=1}^{n} X_iAX_i'. \tag{2.9}$$

For notational simplicity, in the sequel, we drop the suffix n from T_n and K_n.

Theorem 2.2. *Under the assumptions* (M_1)-(M_6),

$$\hat{\beta} \to \beta_0 \quad \text{in probability.} \tag{2.10}$$

Theorem 2.3. *Under the assumptions* (M_1)-(M_6), *we have for any* $c > 0$

$$\sup_{|T^{1/2}(\beta-\beta_0)|\leq c}\Big|\sum_{i=1}^{n} T^{-1/2}X_i[\psi(Y_i - X_i'\beta) - \psi(Y_i - X_i'\beta_0)] + T^{-1/2}K(\beta-\beta_0)\Big| \to 0 \quad \text{in probability.} \tag{2.11}$$

Theorem 2.4. *Under the assumptions* (M_1)–(M_6),

$$T^{-1/2}K(\hat{\beta}-\beta_0)\xrightarrow{\mathcal{D}}N(0,I_p). \tag{2.12}$$

Now, consider testing the hypothesis $H_0 : H'\beta = \gamma$ where H is an $m \times q$ matrix of rank q. Let $\tilde{\beta}$ denote the solution of

$$\min_{H'\beta=\gamma}\sum_{i=1}^{n}\rho(Y_i - X_i'\beta) \tag{2.13}$$

and $\hat{\beta}$ be the solution for the unrestricted minimum.

Theorem 2.5. *Under the assumptions* (M_1)–(M_6):

$$\Big|\sum_{i=1}^{n}[\rho(Y_i-X_i'\tilde{\beta})-\rho(Y_i-X_i'\hat{\beta})]-\frac{1}{2}|Q'\sum_{i=1}^{n}X_i\psi(E_i)|^2\Big| \to 0 \text{ in probability}, \tag{2.14}$$

and

$$(H'\hat{\beta}-\gamma)'(H'K^{-1}TK^{-1}H)^{-1}(H'\hat{\beta}-\gamma)\xrightarrow{\mathcal{D}}\chi_q^2, \tag{2.15}$$

where Q is an $m \times q$ matrix such that

$$QQ' = K^{-1}H(H'K^{-1}H)^{-1}H'K^{-1},$$

and χ_q^2 denotes a chi-square random variable with q degrees of freedom.

Theorems 2.1 – 2.5 are established by first proving the results for a more general model

$$Y_{in} = X_{in}'\beta_n + E_{in}, \quad i = 1,\dots,n \tag{2.16}$$

under the assumptions (M_1)–(M_5), with (M_6) modified as

(M_6'). $\sum_{i=1}^{n}X_{in}BX_{in}' = I_m$ and $d_n^2 = \max_{1\le i\le n}|X_{in}|^2 \to 0$ in probability.

The results for the original model (1.1) are obtained by making the transformation

$$Y_{in} = Y_i, \quad E_{in} = E_i, \quad X_{in} = T^{-1/2}X_i, \quad \beta_n = T^{1/2}\beta. \tag{2.17}$$

Since $B > 0$, it is not difficult to prove that (M_6) implies (M_6') after the transformation (2.17).

Note 1. The test statistic

$$\sum_{i=1}^{n}\rho(Y_i - X_i'\tilde{\beta}) - \sum_{i=1}^{n}\rho(Y_i - X_i'\hat{\beta})$$

has the same asymptotic distribution as that of $2^{-1}|Q'\sum_{i=1}^{n}X_i\psi(e_i)|^2$, which, in general, is a mixture of chi-squares.

Note 2. No doubt, it would be of greater interest to consider a non-convex discrepancy function ρ. That would seem to require much stronger conditions. However, we note that if ρ can be written as the difference of two convex functions, each satisfying our conditions (see for instance the papers by Bickel (1975) and Ruppert and Carroll (1980)), then it is easy to see that our Theorem 2.1 is still true while other theorems remain valid by choosing $\hat{\beta}$ as some minimizer of the function in (1.5), instead of the one which provides a global minimum.

3. Proofs of the Main Theorems

As discussed in Section 2, we need only to prove the theorems for the generalized model (2.16) with the Condition (M_6) replaced by (M_6'). In this case, $T = I_p$. To obtain the results for the original model (1.1), we need only to replace $\hat{\beta}$ by $T^{1/2}\hat{\beta}$, where $\hat{\beta}$ is the M-estimate of β in the model (1.1).

In the proofs that follow, we work with the general model (2.16) but drop the suffix n for convenience of notation and write β for β_n, $\hat{\beta}$ for $\hat{\beta}_n$ and without loss of generality assume that the true value of $\beta_n = \beta_{n0} = 0$.

We need the following lemma:

Lemma 1. *Under the assumptions* (M_1)–(M_3), *we have*

$$E[\rho(E_1 + C) - \rho(E_1)] = \frac{1}{2}C'AC + o(\|C\|^2) \quad \text{as} \quad C \to 0. \tag{3.1}$$

Proof. Suppose C is a small p-vector. For any $k \geq i$, it follows from the convexity of ρ

$$\frac{C'}{k}\psi\Big(E_1 + \frac{(i-1)C}{k}\Big) \leq \rho\Big(E_1 + \frac{iC}{k}\Big) - \rho\Big(E_1 + \frac{(i-1)C}{k}\Big) \leq \frac{C'}{k}\psi\Big(E_1 + \frac{iC}{k}\Big).$$

Taking expectation, using (M_3), summing over $i = 1, \ldots, k$, and then letting $k \to \infty$, we get (3.1).

3.1. Proof of Theorem 2.1

Under the additional assumptions made at the beginning of this section, (2.8) reduces to

$$\sup_{\|\beta\|\le c}\Big|\sum_{i=1}^{n}[\rho(E_i - X_i'\beta) - \rho(E_i) + \beta' X_i\psi(E_i)] \qquad (3.1.1)$$
$$- \frac{1}{2}\beta' K\beta\Big| \to 0 \quad \text{in probability as } n \to \infty$$

where K in (3.1.1) stands for $T^{-1/2}KT^{-1/2}$ with K as in (2.9). To prove (3.1.1), one needs only to prove that for any subsequence $\{n'\}$ of all positive integers, there exists a subsequence $\{n''\}$ of the sequence $\{n'\}$ such that

$$\sup_{\|\beta\|\le c}\Big|\sum_{i=1}^{n''}[\rho(E_i - X_i'\beta) - \rho(E_i) + \beta' X_i\psi(E_i)] - \frac{1}{2}\beta' K\beta\Big| \to 0 \quad \text{a.s.} \quad \text{as } n'' \to \infty. \qquad (3.1.2)$$

At first, by Condition (M_6') and the facts that $A > 0$ and $B > 0$, there is a subsequence $\{n^{(3)}\}$ of $\{n'\}$ such that

$$K \to K^0 > 0 \quad \text{as} \quad n^{(3)} \to \infty. \qquad (3.1.3)$$

For each fixed β, we have

$$|\rho(E_i - X_i'\beta) - \rho(E_i) + \beta' X_i\psi(E_i)| \le \|X_i'\beta\|\,\|\psi(E_i - X_i'\beta) - \psi(E_i)\|$$

by the convexity of the function ρ. Thus, by the condition $\max_{1\le i\le n}\|X_i'\beta\| \to 0$, we have (denoting V for variance)

$$V\Big[\sum_{i=1}^{n}(\rho(E_i - X_i'\beta) - \rho(E_i) + \beta' X_i\psi(E_i))\Big]$$
$$\le \sum_{i=1}^{n} E\|\psi(E_i - X_i'\beta) - \psi(E_i)\|^2\|X_i'\beta\|^2 \to 0.$$

Thus,

$$\sum_{i=1}^{n}[\rho(E_i - X_i'\beta) - \rho(E_i) + \beta' X_i\psi(E_i) \qquad (3.1.4)$$
$$- E(\rho(E_i - X_i'\beta) - \rho(E_i))] \to 0 \quad \text{in probability.}$$

But by Lemma 1 and (3.1.3)

$$\sum_{i=1}^{n^{(3)}} E(\rho(E_i - X_i'\beta) - \rho(E_i)) = \frac{1}{2}\beta' K^0\beta + o(1). \qquad (3.1.5)$$

Substituting (3.1.5) into (3.1.4), we get

$$\sum_{i=1}^{n^{(3)}}\Big[\rho(E_i - X_i'\beta) - \rho(E_i) + \beta' X_i\psi(E_i)\Big] \\ -\frac{1}{2}\beta' K^0\beta \to 0 \quad \text{in probability as } n^{(3)} \to \infty. \tag{3.1.6}$$

By a diagonal technique, for any given countable dense set of β in R^m, one can choose a subsequence $\{n''\}$ of $\{n^{(3)}\}$ such that

$$\sum_{i=1}^{n''}\Big[\rho(E_i - X_i'\beta) - \rho(E_i) + \beta' X_i\psi(E_i)] \to \frac{1}{2}\beta' K^0\beta, \quad \text{a.s.} \tag{3.1.7}$$

as $n'' \to \infty$, for each β in the countable set. Since

$$\sum_{i=1}^{n''}\Big[\rho(E_i - X_i'\beta) - \rho(E_i) + \beta' X_i\psi(E_i)] \quad \text{is convex in } \beta$$

and also $\frac{1}{2}\beta' K^0\beta$ is continuous and convex in β, by Theorem 10.8 of Rockafellar (1970, p.90), we obtain

$$\sup_{\|\beta\|\le c}\left|\sum_{i=1}^{n''}[\rho(E_i - X_i'\beta) - \rho(E_i) + \beta' X_i\psi(E_i)] - \frac{1}{2}\beta' K^0\beta\right| \to 0 \quad \text{a.s.}$$

This implies (3.1.2) and the proof of Theorem 2.1 is complete.

3.2. Proof of Theorem 2.2

The conclusion (2.10) of Theorem 2.2 is obviously implied by the following result for the generalized model (2.16)

$$P(\|\hat{\beta}\| \ge c_n) \to 0, \quad \text{as} \quad n \to \infty, \tag{3.2.1}$$

for any sequence $\{c_n\}$ such that $c_n \to \infty$.

By (3.1.1), one can easily choose a sequence of $\{c_n'\}$ such that $c_n' \to \infty$, $c_n' \le c_n$ and

$$\sup_{\|\beta\|\le c_n'}\left|\sum_{i=1}^{n}[\rho(E_i - X_i'\beta) - \rho(E_i) + \beta' X_i\psi(E_i)] - \frac{1}{2}\beta' K\beta\right| \to 0 \quad \text{in probability.} \tag{3.2.2}$$

When $\|\beta\| = c_n'$, we have

$$\frac{1}{2}\beta' K\beta \ge \frac{1}{2}\lambda(AB^{-1})(c_n')^2, \tag{3.2.3}$$

where $\underline{\lambda}(A)$ denotes the smallest eigenvalue of A.

On the other hand,

$$\sum_{i=1}^{n} X_i\psi(E_i) = O_p(1);$$

hence

$$\beta' \sum_{i=1}^{n} X_i\psi(E_i) = O_p(c_n'). \tag{3.2.4}$$

Then, (3.2.2)–(3.2.4) imply that

$$P\Big(\inf_{\|\beta\|=c_n'} \sum_{i=1}^{n}[\rho(E_i - X_i'\beta) - \rho(E_i)] \le 0\Big) \to 0, \quad \text{as} \quad n \to \infty,$$

which, together with the convexity of ρ, implies that

$$P\Big(\inf_{\|\beta\|\ge c_n'} \sum_{i=1}^{n}\rho(E_i - X_i'\beta) \le \sum_{i=1}^{n}\rho(E_i)\Big) \to 0, \quad \text{as} \quad n \to \infty.$$

By the definition of $\hat{\beta}$, we have

$$P(\|\hat{\beta}\| \ge c_n') \to 0, \quad \text{as} \quad n \to \infty.$$

Note that $c_n' \le c_n$, which completes the proof of (3.2.1).

3.3. Proof of Theorem 2.3

(2.11) is equivalent to

$$\sup_{\|\beta\|\le c} \Big\| \sum_{i=1}^{n} X_i[\psi(E_i - X_i'\beta) - \psi(E_i)] + K\beta \Big\| \to 0, \quad \text{in probability}, \tag{3.3.1}$$

for the generalized model (2.16).

Using the technique as in the proof of Theorem 2.1, one needs only to prove (3.3.1) for the subsequence $\{n^{(3)}\}$ with K (noting that K depends on n) replaced by K^0 (independent of $n^{(3)}$). Then (3.3.1) is a consequence of (3.1.2) and Theorem 2.5.7 of Rockafellar (1970, p.248).

3.4. Proof of Theorem 2.4

For the model (2.16), (2.12) is equivalent to

$$K\hat{\beta} \xrightarrow{\mathcal{D}} N(0, I_p). \tag{3.4.1}$$

Let

$$\overline{\beta} = K^{-1}\sum_{i=1}^{n} X_i\psi(E_i). \tag{3.4.2}$$

Noting that

$$K\overline{\beta}\xrightarrow{\mathcal{D}}N(0,I_p), \tag{3.4.3}$$

(3.4.1) follows, if we can prove that

$$\hat{\beta} - \overline{\beta} \to 0 \quad \text{in probability.} \tag{3.4.4}$$

Take $\delta > 0$. By (3.1.1) and the definition of $\overline{\beta}$, we have

$$\sum_{i=1}^{n}[\rho(E_i - X_i'\overline{\beta}) - \rho(E_i)] + \frac{1}{2}\overline{\beta}'K\beta \to 0, \quad \text{in probability,} \tag{3.4.5}$$

and for some sequence $\{c_n\}$ with $c_n \to \infty$,

$$\sup_{\|\beta\|\le c_n+\delta}\left|\sum_{i=1}^{n}[\rho(E_i - X_i'\beta) - \rho(E_i)] + \beta'K\overline{\beta} - \frac{1}{2}\beta'K\beta\right| \to 0 \quad \text{in probability.} \tag{3.4.6}$$

(3.4.5) and (3.4.6) imply that

$$\sup_{\|\beta-\overline{\beta}\|=\delta}\left|\sum_{i=1}^{n}[\rho(E_i-X_i'\beta)-\rho(E_i-X_i'\overline{\beta})]-\frac{1}{2}(\beta-\overline{\beta})'K(\beta-\overline{\beta})\right| \to 0 \quad \text{in probability.} \tag{3.4.7}$$

Note that when $\|\beta - \overline{\beta}\| = \delta$,

$$(\beta - \overline{\beta})'K(\beta - \overline{\beta}) \ge \underline{\lambda}(AB^{-1})\delta^2. \tag{3.4.8}$$

By (3.4.7) and (3.4.8) and the convexity of ρ, we get

$$P(|\hat{\beta} - \overline{\beta}| \ge \delta) \to 0. \tag{3.4.9}$$

This completes the proof of Theorem 2.4.

3.5. Proof of Theorem 2.5

(3.4.4) and (3.4.5) give for the model (2.16)

$$\hat{\beta} = K^{-1}\sum_{i=1}^{n} X_i\psi(E_i) + o_p(1). \tag{3.5.1}$$

By (3.2.1) and (3.2.2) (taking $c_n = c_n'$), we have

$$\sum_{i=1}^{n}[\rho(E_i - X_i'\hat{\beta}) - \rho(E_i)] = -\frac{1}{2}\left\|K^{-1/2}\sum_{i=1}^{n}X_i\psi(E_i)\right\|^2 + o_p(1). \tag{3.5.2}$$

Without loss of generality, we can assume that $H'H = I_q$. Let $G : m\times(m-q)$ be such that

$$G'G = I_{m-k},\ G'H = 0$$

(with two extreme cases where we define $H = 0,\ G = I_m$ if $q = 0$ and $H = I_p,\ G = 0$ if $p = q$). Then

$$H'\beta = 0 \Longleftrightarrow \beta = G\gamma.$$

Let $\hat{\gamma}$ be the value of γ which minimizes $\sum_{i=1}^{n}\rho(E_i - X_i'G\gamma)$. Then the M-estimate $\tilde{\beta}$ of β under the null hypothesis $H'\beta = 0$ satisfies $\tilde{\beta} = G\hat{\gamma}$. Since $\hat{\gamma}$ is the M-estimate of the model

$$Y_i = X_i'G\gamma + \varepsilon_i,$$

similar to (3.5.2), we have

$$\sum_{i=1}^{n}[\rho(E_i - X_i'\tilde{\beta}) - \rho(E_i)] = -\frac{1}{2}\left\|K_*^{-1/2}G'\sum_{i=1}^{n}X_i\psi(E_i)\right\|^2 + o_p(1), \tag{3.5.3}$$

where

$$K_* = G'KG = \sum_{i=1}^{n}G'X_iAX_i'G.$$

Thus,-

$$\sum_{i=1}^{n}[\rho(E_i - X_i'\tilde{\beta}) - \rho(E_i - X_i'\hat{\beta})] = \frac{1}{2}\left\|Q'\sum_{i=1}^{n}X_i\psi(E_i)\right\|^2 + o_p(1), \tag{3.5.4}$$

where Q is an $m \times q$ matrix such that

$$\begin{aligned} QQ' &= K^{-1} - G(G'KG)^{-1}G' \\ &= K^{-1}H(H'K^{-1}H)^{-1}H'K^{-1}. \end{aligned} \tag{3.5.5}$$

By (3.5.1), one gets

$$\hat{\beta}'H(H'K^{-2}H)^{-1}H'\hat{\beta}\xrightarrow{\mathcal{D}}\chi_q^2. \tag{3.5.6}$$

Now, we have already proved Theorem 2.5 for the generalized model (2.16). To get the results for the original model (1.1), one needs only to replace H, Q,

K, X_i (for model (2.16)) by $T_n^{-1/2}H$, $T_n^{1/2}Q_n$, $T_n^{-1/2}K_nT_n^{-1/2}$ and $T^{-1/2}X_i$, respectively.

4. Further Discussion on Test of Hypothesis

Let $\mathcal{N}_m$ denote an m-dimensional normal vector with iid standard normal components. Then by (2.14), the test statistic $\sum_{i=1}^n[\rho(Y_i - X_i'\tilde{\beta}) - \rho(Y_i - X_i'\hat{\beta})]$ is asymptotically distributed as $\mathcal{N}_m'[T^{1/2}K^{-1}H(H'K^{-1}H)^{-1}H'K^{-1}T^{1/2}]\mathcal{N}_m$, which is generally a weighted chi-square distribution, involving the nuisance parameters A and B. Intuitively, the matrix B can be estimated by

$$\hat{B} = \frac{1}{n}\sum_{i=1}^n[\psi(Y_i - X_i'\hat{\beta})][\psi(Y_i - X_i'\hat{\beta})]'.$$

If ψ is continuously differentiable and its derivatives denoted by ξ (a $p\times p$ matrix function), then A can be estimated by

$$\hat{A}_1 = \frac{1}{n}\sum_{i=1}^n \xi(Y_i - X_i'\hat{\beta}).$$

Generally, A can be estimated by

$$\hat{A}_2 = \frac{1}{nh}\sum_{i=1}^n[\psi(Y_i - X_i'\hat{\beta} + he_1)\dots\psi(Y_i - X_i'\hat{\beta} + he_p)](e_1\cdots e_p)^{-1},$$

where $e_1\cdots e_p$ are linearly independent p-vectors and $h = h_n \to 0$, $d_n/h_n \to 0$.

In the present parer, we shall not discuss the consistency of the estimates of matrices A and B. The main purpose of this section is to figure out some alternative approaches to eliminate the nuisance parameters for some special cases.

4.1. A and B are proportional and one is known

Suppose that B is known and $A = aB$, for an unknown constant $a > 0$. Then, by (2.14), we have

$$\sum_{i=1}^n[\rho(Y_i - X_i'\tilde{\beta}) - \rho(Y_i - X_i'\hat{\beta})]$$

$$= a^{-1}\Big[\sum_{i=1}^n X_i\psi(E_i)\Big]'T^{-1}H(H'T^{-1}H)^{-1}H'T^{-1}\Big[\sum_{i=1}^n X_i\psi(E_i)\Big] + o_p(1).$$

On the other hand, by (3.5.1)

$$\hat{\beta} = a^{-1}T^{-1}\sum_{i=1}^{n} X_i\psi(E_i) + o_p(1);$$

hence, (note that we use the original model (1.1) here)

$$\begin{aligned}&(H'\hat{\beta}-\gamma)'(H'T^{-1}H)^{-1}(H'\hat{\beta}-\gamma)\\ &= a^{-2}\Big[\sum_{i=1}^{n} X_i\psi(E_i)\Big]'T^{-1}H(H'T^{-1}H)^{-1}H'T^{-1}\Big[\sum_{i=1}^{n} X_i\psi(E_i)\Big] + o_p(1).\end{aligned}$$

Therefore,

$$\begin{aligned}&\Big\{\sum_{i=1}^{n}[\rho(Y_i - X_i'\tilde{\beta}) - \rho(Y_i - X_i'\hat{\beta})]\Big\}^2/[(H'\hat{\beta}-\gamma)'(H'T^{-1}H)^{-1}(H'\hat{\beta}-\gamma)]\\ &= \Big[\sum_{i=1}^{n} X_i\psi(E_i)\Big]'T^{-1}H(H'T^{-1}H)^{-1}H'T^{-1}\Big[\sum_{i=1}^{n} X_i\psi(E_i)\Big] + o_p(1)\xrightarrow{\mathcal{D}}\chi_q^2.\end{aligned}$$

4.2. Univariate case

For distinguishing from the multivariate case, we rewrite $A = \lambda$ and $B = \sigma^2$ and use lower cases to denote corresponding variables. In such a case, we suggest the following procedure.

Consider an extended linear model

$$y_i = x_i'\beta + Z_i'\gamma + e_i, \quad i = 1,\dots,n \tag{4.2.1}$$

where Z_i are s-vectors satisfying the conditions

$$Z'X = 0, \qquad Z'Z = I_s, \qquad d_n = \max_{1\le i\le n}|Z_i| \to 0 \tag{4.2.2}$$

with $Z = (Z_1 : \dots : Z_n)'$ and $X = (x_1 : \dots : x_n)$. Let (β^*, γ^*) be a solution of

$$\min_{\beta,\gamma}\sum_{i=1}^{n}\rho(y_i - x_i'\beta - Z_i'\gamma).$$

By Theorem 2.5, under the model (1.1),

$$2\lambda\sigma^{-2}\sum_{i=1}^{n}[\rho(y_i - x_i'\hat{\beta}) - \rho(y_i - x_i'\beta^* - Z_i'\gamma^*)]\xrightarrow{\mathcal{D}}\chi_s^2$$

and is asymptotically independent of $2\lambda\sigma^{-2}\sum_{i=1}^{n}[\rho(y_i - x_i'\tilde{\beta}) - \rho(y_i - x_i'\hat{\beta})]$ by (3.5.4) and (4.2.2), whether the hypothesis H is true or not. Then we have:

Theorem 4.1. *For the model* (1.1), *under the assumptions* (U_1)–(U_6)

$$\frac{s\sum_{i=1}^n[\rho(y_i - x_i'\tilde{\beta}) - \rho(y_i - x_i'\hat{\beta})]}{q\sum_{i=1}^n[\rho(y_i - x_i'\hat{\beta}) - \rho(y_i - x_i'\beta^* - Z_i'\gamma^*)]} \xrightarrow{\mathcal{D}} F(q,s)$$

where $F(q,s)$ *denotes the* F *distribution with* q *and* s *degrees of freedom and* (U_1)–(U_6) *correspond to* (M_1)–(M_6) *with* $p = 1$.

5. Some Examples: Univariate Case

We consider some well known special cases to show how our results can be applied to a variety of situations. Some of these cases have been discussed previously by a number of authors. (See, for example, papers on M-estimation by Huber (1973), Relles (1968) and others.)

5.1. Least squares estimation (LSE)

In this case

$$\rho(x) = x^2, \quad \psi(x) = 2x, \quad \lambda = 2, \quad \sigma^2 = \sigma_0^2 = V(e)$$

giving the result

$$S_n^{1/2}(\hat{\beta} - \beta_0)\xrightarrow{\mathcal{D}} N(0, \sigma_0^2 I_m),$$

where $S_n = \sum_{i=1}^n X_i X_i'$.

5.2. Least distances estimation (LDE)

In this case

$$\rho(x) = |x|, \quad \psi(x) = \operatorname{sign} x.$$

If F has density around zero and $F'(0) = f(0) > 0$, then $\lambda = 2f(0)$ and $\sigma^2 = 1$ leading to the result

$$S_n^{1/2}(\hat{\beta} - \beta_0)\xrightarrow{\mathcal{D}} N(0, [2f(0)]^{-2} I_m).$$

5.3. Mixed LS and LD estimation

Many authors considered the case

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } |x| \le c \\ c|x| - \frac{1}{2}c^2, & \text{if } |x| > c \end{cases}$$

where $c > 0$ is a fixed constant. Then, we have

$$\psi(x) = \begin{cases} x, & \text{if } |x| \le c \\ c \text{ sign x.} & \text{if } |x| > c \end{cases}$$

and

$$\lambda = \int_{-c}^{c} dF(x),\ \sigma^2 = c^2 - \int_{-c}^{c} (c^2 - x^2) dF(x) \tag{5.3.1}$$

giving

$$S^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2/\lambda^2)$$

with σ^2 and λ as in (5.3.1). It may be verified that as $c \to 0$, $(\sigma^2/\lambda^2) \to 1/4f^2(0)$, provided $f(0)$ as defined in the Example 4.2 exists.

5.4. L_P-norm

If $\rho(x) = |x|^p$, $p > 1$, then $\psi(x) = p|x|^{p-1}$ sign x and

$$\lambda = \int p(p-1)|x|^{p-2} dF(x),\ \sigma^2 = \int p^2 |x|^{2(p-1)} dF(x)$$

giving

$$S^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2/\lambda^2).$$

5.5. Differentiable ψ

In this case

$$\lambda = \int \psi'(x) dF(x),\ \sigma^2 = \int \psi^2(x) dF(x)$$

giving

$$S_n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2/\lambda^2).$$

6. Some Examples: Multivariate Case

6.1. LDE in the multivariate case

If we choose $\rho(x) = \left(\Sigma x_i^2\right)^{1/2}$, where $x' = (x_1, \ldots, x_p)$, then

$$\psi(x) = \begin{cases} x/|x|, & \text{if } x \ne 0 \\ 0, & \text{if } x = 0; \end{cases}$$

$$\psi'(x) = \begin{cases} \frac{1}{|x|}\left(1 - \frac{xx'}{|x|^2}\right), & \text{if } x \ne 0 \\ 0, & \text{if } x = 0. \end{cases}$$

Hence

$$A = E\Big[\frac{1}{|E_1|}\Big(I - \frac{E_1E_1'}{|E_1|^2}\Big)\Big], \tag{6.1.1}$$

$$B = E\Big(\frac{E_1E_1'}{|E_1|^2}\Big). \tag{6.1.2}$$

Using A amd B as determined above, we can write down the asymptotic distribution of $\hat{\beta}$. The results (6.1.1) and (6.1.2) are reported in Bai, Chen, Miao and Rao (1990).

6.2. Joint distribution of the component medians

In this case $X_i = I_p$ and the component medians are obtained by choosing

$$\rho(x) = \sum |x_i|, \text{ where } x' = (x_1, \dots, x_p).$$

Then

$$\psi(x) = (\operatorname{sign} x_1, \dots, \operatorname{sign} x_p)'$$

and

$$A = \operatorname{diag}(2f_1(0), \dots, 2f_p(0))$$

where $f_i(0)$ is the density of the marginal distribution of y_i the i-th component of the p-vector variable Y at the median value. The matrix $B = (b_{ij})$ where

$$b_{ij} = \begin{cases} 1, & \text{if } i = j \\ 4[P\{y_i \le 0,\ y_j \le 0\} - 1], & \text{if } i \ne j. \end{cases}$$

These results are reported in Babu and Rao (1988).

Acknowledgements

We wish to thank the referees for a number of useful suggestions, which resulted in considerable improvement of our original paper. The research sponsored by the Air Force Office of Scientific Research under Grant AFOSR-89-0279 and Office of Naval Research under Contract N00014-85-K-0292. The United States Government is authorized to reproduce and distribute reprints for governmental purposed notwithstanding any copyright notation hereon.

References

Amemiya, T. (1982). Two stage least absolute deviations estimators. *Econometrika* **50**, 689-711.

Babu, G. Jogesh and Rao, C. Radhakrishna (1988). Joint asymptotic distribution of marginal quantiles and quantile functions in smples from a multivariate population. *J. Multivariate Anal.* **27**, 15-23.

Bai, Z. D., Chen, X. R., Miao, B. Q. and Rao, C. R. (1990). Asymptotic theory of least distances estimate in multivariate linear models. *Statistics* **21**, 503-519.

Bai, Z. D., Chen, X. R., Wu, Y. and Zhao, L. C. (1990). Asymptotic normality of minimum L_1-norm estimates in linear models. *Chinese Sci. Ser.A* **33**, 449-463.

Bai, Z. D., Rao, C. R. and Yin, Y. Q. (1990). Least absolute deviations analysis of variance. *Sankhyā Ser.A*, **52**, 166-177.

Basawa, L. V. and Koul, H. L. (1988). Large-sample statistics based on quadratic dispersion. *Internat. Statist. Rev.* **56**, 199-219.

Basset, G. and Koenker, R. (1978). Asymptotic theory of least absolute error regression. *J. Amer. Statist. Assoc.* **73**, 618-622.

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70**, 428-433.

Bloomfield, P. and Steiger, W. L. (1983). *Least Absolute Deviations.* Birkhauser, Boston Inc.

Dupacova, J. (1987). Asymptotic properties of restricted L_1-estimates of regression. In *Statistical Data Analysis Based on the L_1-Norm and Related Methods* (Edited by Y. Dodge), 263-274, North-Holland.

Heiler, S. and Willers, R. (1988). Asymptotic normality of R-estimates in the linear model. *Statistics* **19**, 173-184.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73-101.

Huber, P. J. (1973). Robust regression. *Ann. Statist.* **1**, 799-821.

Huber, P. J. (1981). *Robust Statistics.* John Wiley, New York.

Huber, P. J. (1987). The place of the L_1-norm in robust estimation. In *Statistical Data Analysis Based on the L_1-Norm and Related Methods* (Edited by Y. Dodge), 23-34, North-Holland.

Jackel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.* **43**, 1449-1458.

Jureckova, J. (1971). Nonparametric estimate of regression coefficients. *Ann. Math. Statist.* **42**, 1328-1338.

Koenker, R. and Portnoy, S. (1990). M-estimation of multivariate regressions. *J. Amer. Statist. Assoc.* **85**, 1060-1068.

Relles, D. (1968). Robust regression by modified least squares. Ph.D. thesis, Yale University.

Rockafellar, R. T. (1970). *Convex Analysis.* Princeton University Press.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in linear models. *J. Amer. Statist. Assoc.* **75**, 828-838.

Yohai, V. J. and Maronna, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *Ann. Statist.* **7**, 258-268.

Department of Statistics, Temple University, Philadephia, PA 19122, U.S.A.
Department of Statistics, Penn State University, University Park, PA 16803, U.S.A.
Department of Mathematics and Statistics, York University, North York, Canada M3J 1P3.

(Received November 1989; accepted August 1991)