The Annals of Statistics
1974, Vol. 2, No. 3, 562-567

ROBUST ESTIMATION IN THE LINEAR MODEL

BY VICTOR J. YOHAI

University of La Plata

The purpose of this paper is to present robust estimates for the regression parameters in the general linear model. We start with a family of M-estimators, and using the observations, we estimate the asymptotic efficiency of each member in the family. Then we choose the estimate in the family with greatest estimated asymptotical efficiency. We prove that this procedure has the same asymptotical efficiency as the member of the family with the greatest asymptotical efficiency for the unknown distribution of the error.

1. Introduction. Let $\{Y_j^{(n)}, 1 \leqq j \leqq n\}$, $n \geqq 1$ be independent random variables such that

$$P(Y_j^{(n)} \leqq y) = F(y - \theta' X_j^{(n)}) ,$$

where $\theta' = (\theta_1, \theta_2, \cdots, \theta_s)$ are the regression parameters, F is a symmetric distribution function, $(X_j^{(n)})' = (X_{j1}^{(n)}, \cdots, X_{js}^{(n)})$ are some known regression constants and "$'$" denotes transposition.

Classically the vector parameter θ is estimated by the method of least squares, i.e., by the θ that minimizes

$$\sum_{j=1}^n (Y_j^{(n)} - \theta' X_j^{(n)})^2 .$$

Although this estimate is optimal when F is normal, it is not robust when F has longer tails.

Huber in [3] proposes using as robust estimates for regression the M-estimates that he introduced in [2] for estimating a location parameter. Given a non-negative and even real function ρ, the corresponding M-estimate is defined as the value of θ that minimizes

$$\sum_{j=1}^n \rho(Y_j^{(n)} - \theta' X_j^{(n)}) . \tag{1}$$

If ρ is convex and has derivative Ψ, minimizing (1) is equivalent to the following system:

$$\sum_{j=1}^n \Psi(Y_j^{(n)} - \theta' X_j^{(n)}) X_{ji}^{(n)} = 0 \qquad i = 1, 2, \cdots, s . \tag{2}$$

An important special case is when Ψ is given by Ψ_k, where

$$\begin{aligned} \Psi_k(t) &= -k \quad &&\text{if} \quad t < -k \\ &= t &&\text{if} \quad |t| \leqq k \\ &= k &&\text{if} \quad t > k \end{aligned} \tag{3}$$

and $0 < k \leqq \infty$. The case $k = \infty$ corresponding to the least squares estimator.

Received September 1972; revised April 1973.
AMS 1970 *subject classifications*. Primary 62; Secondary 55.
Key words and phrases. Robust estimation, linear model, M-estimates,

It seems reasonable to require that an estimate $\hat{\theta}^{(n)}$ of θ should be shift and scale equivariant, i.e. if we denote by $Y^{(n)} = (Y_1^{(n)}, \cdots, Y_n^{(n)})$ and by $X^{(n)}$ the matrix whose j, i entry is $X_{ji}^{(n)}$, then for any s-dimensional vector Δ, we have

$$\hat{\theta}^{(n)}(Y^{(n)} + X^{(n)}\Delta) = \hat{\theta}^{(n)}(Y^{(n)}) + \Delta \,, \tag{4}$$

and if λ is a scalar, then

$$\hat{\theta}^{(n)}(\lambda Y^{(n)}) = \lambda\hat{\theta}^{(n)}(Y^{(n)}) \,. \tag{5}$$

The estimate based on Ψ_k, with fixed k is shift equivariant but is not scale equivariant. In order to get a scale and shift equivariant estimate we should estimate θ by solving

$$\textstyle\sum_{j=1}^n \Psi_{\hat{k}}(Y_j^{(n)} - \theta' X_j^{(n)})X_{ji}^{(n)} = 0 \qquad i = 1, 2, \cdots, s \,, \tag{6}$$

where $\hat{k}$ is a statistic based on the first n observations (we omit the index n from $\hat{k}$) satisfying

$$\hat{k}(\lambda Y^{(n)}) = |\lambda|\hat{k}(Y^{(n)}) \tag{7}$$

and

$$\hat{k}(Y^{(n)} + X^{(n)}\Delta) = \hat{k}(Y^{(n)}) \,. \tag{8}$$

Consider now the following assumptions.

A1. There exist numbers a_n, with $a_n \to \infty$ as $n \to \infty$ such that

$$\lim_{n\to\infty} X^{(n)\prime}X^{(n)}/a_n = \Sigma \,, \tag{9}$$

where Σ is a positive definite matrix.

A2. Put $b_n = \sup\{(X_{ji}^{(n)})^2, 1 \leqq j \leqq n, 1 \leqq i \leqq s\}$. Then

$$\lim_{n\to\infty} b_n/a_n = 0 \,. \tag{10}$$

A3. $F(x)$ is symmetric.

A4. $\hat{k} \to k$ in probability, where k is a constant.

A5. $F(x)$ is continuous at k and $F(k) - F(-k) > 0$.

The following theorem has been shown first, under slightly different conditions by Relles in [5] and then by Yohai in [6].

THEOREM 1. *Assume* A1—A5. *Let* $\hat{\theta}_{\hat{k}}^{(n)}$ *be the solution of* (6), *then* $a_n^{\frac{1}{2}}(\hat{\theta}_{\hat{k}}^{(n)} - \theta)$ *converges in distribution to the multivariate normal distribution with mean* 0 *and covariance matrix* $A(k, F)\Sigma^{-1}$, *where*

$$A(k, F) = \left[\textstyle\int_{-k}^{k} u^2\, dF(u) + k^2\left(1 - \int_{-k}^{k} dF(u)\right)\right] \Big/ \left(\int_{-k}^{k} dF(u)\right)^2 \,. \tag{11}$$

Then the efficiency of the estimate $\hat{\theta}_{\hat{k}}^{(n)}$ *is inversely proportional to* $A(k, F)$, *independently of the matrices* $X^{(n)}$.

Following an idea used by Jaeckel in [4], we are going to estimate the value of k that gives the smallest value of $A(k, F)$ for k belonging to a given interval, and use this value as the $\hat{k}$ of (6).

Let σ be a scale parameter, and $0 < k_1 < k_2 < \infty$, such that $F(k_1\sigma) - F(-k_1\sigma) > 0$. Suppose also that there exists a unique number k^* such that

(12) $$A(k^*, F) = \min_{k \in [k_1\sigma, k_2\sigma]} A(k, F) .$$

In the next section we construct an estimator $\hat{k}^*$ of k^* that satisfies (7) and (8), and such that $\hat{k}^*$ converges to k^* in probability. Then $\hat{\theta}^{(n)}_{\hat{k}^*}$ will be scale and shift equivariant. Moreover, using Theorem 1 we will have that $a_n^{\frac{1}{2}}(\hat{\theta}^{(n)}_{\hat{k}^*} - \theta)$ converges in distribution to the multivariate normal distribution with mean 0 and covariance matrix $A(k^*, F)\Sigma^{-1}$.

2. Estimation of k^*. Assume now that we have an initial estimate $\tilde{\theta}^{(n)}$ of θ such that

A6. $\tilde{\theta}^{(n)}$ is shift and scale equivariant.

A7. $a_n^{\frac{1}{2}}(\tilde{\theta}^{(n)} - \theta)$ is bounded in probability, i.e. given $\varepsilon > 0$, there exists K such that

(13) $$P(a_n^{\frac{1}{2}}|\tilde{\theta}^{(n)} - \theta| \geqq K) \leqq \varepsilon \qquad \text{for all } n .$$

Put

(14) $$U_j^{(n)} = Y_j^{(n)} - \theta' X_j^{(n)} ; \qquad 1 \leqq j \leqq n .$$

Then $U_j^{(n)}$, $1 \leqq j \leqq n$ are independent identically distributed random variables with distribution function equal to F.

We can estimate the values taken by these variables by

(15) $$\hat{U}_j^{(n)} = Y_j^{(n)} - \tilde{\theta}^{(n)\prime} X_j^{(n)} = U_j^{(n)} - (\tilde{\theta}^{(n)} - \theta)' X_j^{(n)} ;$$

and $A(k, F)$ by

(16) $$\hat{A}(k, F) = \left(\sum_{j=1}^n \Psi_k^2(\hat{U}_j^{(n)})/n\right)/\left(\sum_{j=1}^n I_{[-k,k]}(\hat{U}_j^{(n)})/n\right)^2 ,$$

where $I_{[-k,k]}$ is the indicator function of $[-k, k]$.

Let $\hat{\sigma}$ be an estimate of σ and assume

A8. $\hat{\sigma}$ satisfies

(17) $$\hat{\sigma}(Y^{(n)} + X^{(n)}\Delta) = \hat{\sigma}(Y^{(n)})$$

and if λ is a scalar

(18) $$\hat{\sigma}(\lambda Y^{(n)}) = |\lambda|\hat{\sigma}(Y^{(n)}) .$$

A9. $\hat{\sigma} \to \sigma$ in probability.

Define now $\hat{k}^*$ by

(19) $$\hat{k}^* = \min\{k \in [k_1\hat{\sigma}, k_2\hat{\sigma}] : \hat{A}(k, F) = \text{minimum}\} .$$

We omit the index n from $\hat{\sigma}$ and $\hat{k}^*$.

It is easy to see that A6 and A8 imply that if $\hat{k}^*$ satisfies (7) and (8), then $\hat{\theta}^{(n)}_{\hat{k}^*}$ is shift and scale equivariant. In the next theorem we prove that $\hat{k}^*$ converges in probability to k^*.

THEOREM 2. *Assume* A1, A2, A3, A7, A9, $F(x)$ *continuous and* $F(k_1\sigma) - F(-k_1\sigma) > 0$. *Suppose also that* k^* *is unique. Then* $\hat{k}^*$ *converges to* k^* *in probability.*

In order to prove Theorem 2 we need to prove Lemmas 1 and 2.

LEMMA 1. *Let* $U_1, \cdots, U_n, \cdots$ *be a sequence of* i.i.d. *random variables. Let* (f_k), $k \in C$, *where* C *is a compact, a family of Borel measurable real functions such that*

(i) $|f_k| \leqq f$ *where* $E(f(U_1)) < \infty$.
(ii) $\lim_{t\to k} f_t(U_1) = f_k(U_1)$ a.s. *for all* k *in* C.
(iii) $|E(f_k(U_1))| \leqq A$ *for all* k *in* C.

Then

$$\limsup_{n\to\infty} \sup_{k\in C} |\textstyle\sum_{j=1}^n f_k(U_j)/n| \leqq A \quad \text{a.s.}$$

PROOF. Given $\varepsilon > 0$, for any $k \in C$ we can find by dominated convergence a neighborhood of k, C_k such that

$$|E(\sup\{f_t(U_1): t \in C_k\})| \leqq A + \varepsilon \tag{20}$$

and

$$|E(\sup\{-f_t(U_1): t \in C_k\})| \leqq A + \varepsilon\,. \tag{21}$$

Since C is compact, there exists $k_1, k_2, \cdots, k_h$ such that $U_{k=1}^h C_k = C$.
Then

$$\sup_{k\in C} |\textstyle\sum_{j=1}^n f_k(U_j)/n|$$
$$\leqq \sup\{\sup_{1\leqq l\leqq h} \sup_{k\in C_{k_l}} (\textstyle\sum_{j=1}^n f_k(U_j)/n), \sup_{1\leqq l\leqq h} \sup_{k\in C_{k_l}} (\sum_{j=1}^n -f_k(U_j)/n)\}\,.$$

Then it will be enough to show that

$$\limsup_{n\to\infty} \sup_{1\leqq l\leqq h} \sup_{k\in C_{k_l}} \textstyle\sum_{j=1}^n f_k(U_j)/n \leqq A + \varepsilon \quad \text{a.s.} \tag{22}$$

and

$$\limsup_{n\to\infty} \sup_{1\leqq l\leqq h} \sup_{k\in C_{k_l}} \textstyle\sum_{j=1}^n -f_k(U_j)/n \leqq A + \varepsilon \quad \text{a.s.} \tag{23}$$

Since the proofs of (22) and (23) are identical we prove only (22).
We have

$$\limsup_{n\to\infty} \sup_{1\leqq l\leqq h} \sup_{k\in C_{k_l}} \textstyle\sum_{j=1}^n f_k(U_j)/n = \sup_{1\leqq l\leqq h} \limsup_{n\to\infty} \sup_{k\in C_{k_l}} \sum_{j=1}^n f_k(U_j)/n\,. \tag{24}$$

Moreover, using (20) and the strong law of large numbers we have

$$\limsup_{n\to\infty} \sup_{k\in C_{k_l}} \textstyle\sum_{j=1}^n f_k(U_j)/n \leqq \limsup_{n\to\infty} \sum_{j=1}^n \sup_{k\in C_{k_l}} f_k(U_j)/n \leqq A + \varepsilon \quad \text{a.s.} \tag{25}$$

From (24) and (25), (22) follows.

LEMMA 2. *Assume* A1, A2, A3, A7 *and* $F(k_1) - F(-k_1) > 0$. *Then*

$$p\lim_{n\to\infty} \sup_{k\in[k_1,k_2]} |\hat{A}(k, F) - A(k, F)| = 0\,. \tag{26}$$

(p lim denotes limit in probability.)

PROOF. Since $\int_{-k_1}^{k_1} dF(u) > 0$, it will be enough to show that

$$p\lim_{n\to\infty} \sup_{k\in[k_1,k_2]} [\textstyle\sum_{j=1}^n (\Psi_k^2(\hat{U}_j^{(n)}) - \int_{-\infty}^{\infty} \Psi_k^2(u)\, dF(u))/n] = 0 \tag{27}$$

and

$$p\lim_{n\to\infty} \sup_{k\in[k_1,k_2]} [\textstyle\sum_{j=1}^n (I_{[-k,k]}(\hat{U}_j^{(n)}) - \int_{-k}^{k} dF(u))/n] = 0\,. \tag{28}$$

Since the proofs of (27) and (28) are similar we only prove (27). In order to prove (27) it will be enough to show that

$$p\lim_{n\to\infty} \sup_{k\in[k_1,k_2]} [\textstyle\sum_{j=1}^n (\Psi_k^2(U_j^{(n)}) - \int_{-\infty}^{\infty} \Psi_k^2(u)\, dF(u))/n] = 0 \tag{29}$$

and

$$p\lim_{n\to\infty} \sup_{k\in[k_1,k_2]} [\textstyle\sum_{j=1}^n (\Psi_k^2(\hat{U}_j^{(n)}) - \Psi_k^2(U_j^{(n)}))/n] = 0\,. \tag{30}$$

Since the functions $f_k = \Psi_k^2 - \int_{-\infty}^{\infty} \Psi_k^2(u)\, dF(u)$ satisfy the assumptions of Lemma 1 with $A = 0$, and the $U_j^{(n)}$, $1 \leqq j \leqq n$ are i.i.d., (29) follows.

From A1, A2 and A7 it follows that

$$p\lim_{n\to\infty} \sup_{1\leqq j\leqq n} (\tilde{\theta}^{(n)} - \theta)' X_j^{(n)} = 0\,. \tag{31}$$

Then, according to (15) and (31), in order to prove (30) it suffices to show that

$$\lim_{\delta\to 0} p\lim\sup_{n\to\infty} \sup_{k\in[k_1,k_2]} \sup_{|\delta'|\leqq\delta} |\textstyle\sum_{j=1}^n (\Psi_k^2(U_j^{(n)} + \delta') - \Psi_k^2(U_j^{(n)}))/n| = 0\,. \tag{32}$$

But

$$\begin{aligned}\sup_{k\in[k_1,k_2]} \sup_{|\delta'|\leqq\delta} |\textstyle\sum_{j=1}^n (\Psi_k^2(U_j^{(n)} + \delta') - \Psi_k^2(U_j^{(n)}))/n| \\ \leqq \textstyle\sum_{j=1}^n \sup_{k\in[k_1,k_2]} \sup_{|\delta'|\leqq\delta} |(\Psi_k^2(U_j^{(n)} + \delta') - \Psi_k^2(U_j^{(n)}))/n|\,.\end{aligned} \tag{33}$$

It is easy to show that for δ_0 sufficiently small we have

$$\begin{aligned}\sup_{k\in[k_1,k_2]} \sup_{|\delta'|\leqq\delta_0} |\Psi_k^2(U_j^{(n)} + \delta') - \Psi_k^2(U_j^{(n)})| \\ \leqq \sup_{k\in[k_1,k_2]} (2k\delta_0 - \delta_0^2) = 2k_2\delta_0 - \delta_0^2 \leqq \varepsilon\,.\end{aligned} \tag{34}$$

Then from (33) and (34) we obtain

$$p\lim\sup_{n\to\infty} \sup_{k\in[k_1,k_2]} \sup_{0\leqq\delta'\leqq\delta_0} |\textstyle\sum_{j=1}^n (\Psi_k^2(U_j^{(n)} + \delta') - \Psi_k^2(U_j^{(n)}))/n| \leqq \varepsilon\,;$$

and then (30) follows.

PROOF OF THEOREM 2. Let $\varepsilon > 0$; since $A(k, F)$ is continuous and k^* is the only minimum of $A(k, F)$ in the interval $[k_1\sigma, k_2\sigma]$, we have

$$\begin{aligned}\lambda(\sigma, \varepsilon) = \min\{A(k, F): k \in [k_1\sigma, k_2\sigma], |k - k^*| \geqq \varepsilon\} \\ - \min\{A(k, F): k \in [k_1\sigma, k_2\sigma]\} > 0\,.\end{aligned} \tag{35}$$

Using once more the continuity of $A(k, F)$ we can find $\delta > 0$ and $\mu > 0$ such that

$$\lambda(\sigma', \varepsilon) > \mu > 0 \qquad \text{if} \quad |\sigma' - \sigma| \leqq \delta\,. \tag{36}$$

But, according to (19), (35) and (36), we have

$$P(|\hat{k}^* - k^*| \geqq \varepsilon)$$
$$\leqq P(\sup\{|\hat{A}(k, F) - A(k, F)| : k \in [k_1(\sigma - \delta), k_2(\sigma + \delta)]\} > \mu/2)$$
$$+ P(|\hat{\sigma} - \sigma| > \delta) .$$

Then by Lemma 2 and A9 we have

$$P(|\hat{k}^* - k^*| > \varepsilon) \to 0 \qquad \text{as } n \to \infty . \ \square$$

From Theorems 1 and 2 we obtain immediately the following theorem.

THEOREM 3. *Assume the same conditions as in Theorem* 2. *Then* $a_n^{\frac{1}{2}}(\hat{\theta}_{\hat{k}^*}^{(n)} - \theta)$ *converges in distribution to the multivariate normal with mean* 0 *and covariance matrix* $A(k^*, F)\Sigma^{-1}$.

REMARK 1. The case in which k^* is not unique may be treated as the similar case in Lemma 3 of [4].

REMARK 2. If $F(x)$ has second moment we can take as $\tilde{\theta}^{(n)}$ the least squares estimator.

REMARK 3. In [1] Bickel proposes taking as scale parameter $\sigma = F^{-1}(\frac{3}{4})/\Phi^{-1}(\frac{3}{4})$, where Φ is the standard normal cumulative distribution.

In this case we can use as estimator of σ

$$\hat{\sigma} = (\hat{U}_{(n-[n/4]+1)}^{(n)} - \hat{U}_{([n/4])}^{(n)})/2\Phi^{-1}(\tfrac{3}{4}),$$

where $\hat{U}_{(i)}^{(n)}$ $1 \leqq i \leqq n$ are the order statistics of $\hat{U}_i^{(n)}$ $1 \leqq i \leqq n$.

REMARK 4. Bickel introduced in [1] the one step $(\Psi_{\hat{k}})$ estimates that have the same asymptotic efficiency as the M-estimator based on $\Psi_{\hat{k}}$. Then the one step $\Psi_{\hat{k}^*}$ estimates will have the same asymptotic efficiency as the corresponding M-estimators.

REFERENCES

[1] BICKEL, P. J. (1973). A note on approximate (M) estimates in the linear model. To appear in *J. Amer. Statist. Assoc.*

[2] HUBER, P. J. (1964). Robust estimates of a location parameter. *Ann. Math. Statist.* **35** 73-101.

[3] HUBER, P. J. (1968). Robust estimation. Mathematical Centre Tracts **27** *Math. Centrum Amsterdam* 3-25.

[4] JAECKEL, L. A. (1971). Some flexible estimates of location. *Ann. Math. Statist.* **42** 1540-1552.

[5] RELLES, D. (1968). Robust regression by modified least squares. Ph. D. thesis, Yale Univ.

[6] YOHAI, V. J. (1972). Robust (M) estimates for the general linear model. *Notas de Matemática* n° 23, Departamento de Matemática, Universidad Nacional de La Plata.

CARLOS PELLEGRINI 743, (8P 37)
BUENOS AIRES, ARGENTINA