19

PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 17, Fasc. 1 (1997), pp. 1–20

M-ESTIMATION FOR LINEAR REGRESSION WITH INFINITE VARIANCE

BY

RICHARD A. DAVIS* (COLORADO STATE UNIVERSITY)
AND
WEI WU (SCHERING – PLOUGH RESEARCH INSTITUTE)

Abstract. The limiting behavior of *M*-estimates for a linear model when the regressors and/or errors have heavy tailed distributions is given. By *heavy tail* we mean that the distribution is in the domain of attraction of a non-normal stable distribution or, equivalently, that the tail probabilities are regularly varying at infinity with exponent $\alpha \in (0, 2)$. These results are applicable to both least squares and least absolute deviation estimators. The limiting distribution of the minimum dispersion estimate is also derived and its performance is compared with that of the *M*-estimate.

1. Introduction. In this paper, the limiting behavior of *M*-estimates for a linear model when the regressors and/or errors have heavy tailed distributions is given. Consider the linear model

$$Y_i = X_i' \boldsymbol{\beta} + Z_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where $X_i = (X_{i1}, \ldots, X_{id})'$, $i = 1, \ldots, n$, are independent and identically distributed (iid) random vectors with joint distribution function F (written more concisely as $\{X_i\}_{i=1}^{n} \overset{\text{iid}}{\sim} F$), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)'$ is the parameter vector, and $\{Z_i\}_{i=1}^{n} \overset{\text{iid}}{\sim} G$ are the errors. The regressors $\{X_i\}_{i=1}^{n}$ and errors $\{Z_i\}_{i=1}^{n}$ are also assumed to be independent. The *M-estimate* $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is defined as any minimizer of the objective function

$$\sum_{i=1}^{n} \varrho(Y_i - \phi_1 X_{i1} - \ldots - \phi_d X_{id}) \tag{1.2}$$

* Research supported by the NSF Grant DMS 9100392.

with respect to $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_d)'$, i.e. $\hat{\boldsymbol{\beta}}$ is a solution to the following equations

$$\sum_{i=1}^{n} \psi(Y_i - \phi_1 X_{i1} - \ldots - \phi_d X_{id}) X_{kd} = 0, \qquad k = 1, \ldots, d,$$

where $\psi(x) = \varrho'(x)$.

The distribution F of the regressors will be *heavy tailed* by assuming that F is in the domain of attraction of a multivariate stable distribution with index $\alpha \in (0, 2)$. This implies, in particular, that the regressors have infinite variance. In some situations, a similar condition will be imposed on the distribution of the errors, only with index $\gamma \in (0, 2]$.

In Section 2 we consider the case of a simple linear regression model ($d = 1$), where the inderlying distribution F of the independent variable X is in the domain of attraction of a stable law. The weak convergence of the M-estimate is established under mild regularity conditions on the score function $\psi(x)$. In particular, if a_n is the $1-n^{-1}$ quantile of the distribution of $|X_1|$, then $a_n(\hat{\beta}-\beta)$ converges in distribution to a random variable which is defined as the minimum of some stochastic process.

In Section 3 we specialize to the class of loss functions given by $\varrho(x) = |x|^\lambda$, $\lambda \geqslant 1$ with $\lambda = 2$ and $\lambda = 1$ corresponding to the least squares (LS) and least absolute deviation (LAD) loss functions, respectively. For this class of loss functions, it is easier to describe the interplay between the loss function and the heaviness of the tails of F and G on the performance of the M-estimator. For example, if the tails of the error distribution are heavier than those of the regressor, then the LAD estimate performs better than the LS estimate.

Section 4 extends the results of Sections 2 and 3 to the multiple linear regression setting. In addition, M-estimates are discussed when a location parameter β_0 is incorporated into the model.

Another popular method of estimation, at least when G has a stable distribution, is to minimize the dispersion of the errors (see for example Blattberg and Sargent [2]). In Section 5, we derive the limit distribution of the dispersion estimator for the simple linear model and compare its performance with that of the M-estimate.

As mentioned earlier, the limit random variable of the normalized M-estimator can be described as the minimizer of some stochastic process. If the distribution of the errors and the stable index of F are known, then it is possible to generate replicates of the minimizer of this stochastic process by repeated replication of the stochastic process. However, in the typical situation both F and α are unknown. To overcome this difficulty, a bootstrap procedure can be implemented to approximate the sampling distribution of the M-estimate. Provided one chooses a bootstrap sample of size m_n with $m_n/n \to 0$, the bootstrap estimate of the sampling distribution of the M-estimate is consistent. This result is similar in spirit to that obtained by Athreya et al. [1] and is discussed in Section 6.

2. Simple linear regression. Let (Y_i, X_i), $i = 1, \ldots, n$, be observations from the simple linear model

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

where $\{Z_i\}_{i=1}^n \overset{\text{iid}}{\sim} G$ and $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} F$. It is further assumed that F belongs to the domain of attraction of a stable law with index $0 < \alpha < 2$ (denoted by $F \in \mathscr{D}(\alpha)$ or $X_i \in \mathscr{D}(\alpha)$), i.e. there exist a slowly varying function $L(x)$ at ∞, constants $0 \leqslant p,\ q \leqslant 1$, $p+q = 1$, and $\alpha \in (0, 2)$, such that

$$1 - F(x) \sim px^{-\alpha} L(x), \qquad F(-x) \sim qx^{-\alpha} L(x) \qquad \text{as } x \to \infty. \tag{2.2}$$

Then the partial sums $\sum_{i=1}^n X_i$, suitably centered and scaled, converge in distribution to a stable distribution. The scaling constants are given by

$$a_n = \inf\{x\colon\ P(|X_1| \geqslant x) \leqslant n^{-1}\}.$$

Now for a given loss function $\varrho(x)$, the M-estimate $\hat{\beta}$ of the regression coefficient β is defined as any minimizer of the objective function

$$g(\phi) := \sum_{i=1}^n \varrho(Y_i - \phi X_i) \tag{2.3}$$

which may also be found as a solution to the equation

$$g'(\phi) = \sum_{i=1}^n \psi(Y_i - \phi X_i) X_i = 0, \tag{2.4}$$

where $\psi(x) = \varrho'(x)$. The traditional derivation of the asymptotic distribution of the M-estimator is to first show that it is consistent and then to expand $g'(\phi)$ in a Taylor series around the true regression parameter β. Replacing Y_i by $\beta X_i + Z_i$, we have

$$\begin{aligned} 0 = g'(\hat{\beta}) &= g'(\beta) + (\hat{\beta} - \beta) g''(\beta) + R_n \\ &= \sum_{i=1}^n \psi(Z_i) X_i - (\hat{\beta} - \beta) \sum_{i=1}^n \psi'(Z_i) X_i^2 + R_n. \end{aligned}$$

If the remainder term R_n is $o_p(n^{1/2})$, then, under suitable moment conditions on $\psi(Z_i)$, $\psi'(Z_i)$, and X_i, the asymptotic normality of

$$n^{1/2}(\hat{\beta} - \beta) \mathrel{\dot{\approx}} \frac{\sum_{i=1}^n \psi(Z_i) X_i / n^{1/2}}{\sum_{i=1}^n \psi'(Z_i) X_i^2 / n}$$

follows directly from the central limit theorem and the strong law of large numbers. In the heavy tailed case, however, a_n plays the role of $n^{1/2}$ and R_n is not $o_p(a_n)$. In fact, if a_n is the correct scaling for $\hat{\beta} - \beta$, then, for any integer k, the k-th term of the Taylor expansion of (2.4) divided by a_n is

$$a_n^{-1}\sum_{i=1}^{n}\psi^{(k)}(Z_i)[(\hat{\beta}-\beta)X_i]^k X_i = [a_n(\hat{\beta}-\beta)]^k a_n^{-(k+1)}\sum_{i=1}^{n}\psi^{(k)}(Z_i)X_i^{k+1}$$

$$\sim [a_n(\hat{\beta}-\beta)]^k O_p(1) = O_p(1).$$

To overcome these obstacles with a Taylor series expansion, we work directly with the objective function – viewing it as a stochastic process indexed by ϕ. This approach is the same as the one employed in Davis et al. [5].

To carry out this program, note that the parameter estimate $\hat{\beta}$ which minimizes (2.3) also minimizes $\sum_{i=1}^{n}[\varrho(Z_i-(\phi-\beta)X_i)-\varrho(Z_i)]$, which can be rewritten as

$$\sum_{i=1}^{n}[\varrho(Z_i-a_n(\phi-\beta)a_n^{-1}X_i)-\varrho(Z_i)].$$

Building the normalization a_n into our pårameterization, we define the sequence of stochastic processes $W_n(u)$ on $\boldsymbol{R}$ by

$$W_n(u)=\sum_{i=1}^{n}[\varrho(Z_i-ua_n^{-1}X_i)-\varrho(Z_i)]. \tag{2.5}$$

Then, for each fixed n, $\hat{u}_n = a_n(\hat{\beta}-\beta)$ minimizes the stochastic process $W_n(u)$ in (2.5). If one could show that the stochastic processes $W_n(u)$ converge in distribution to a limiting process $W(u)$, then one would expect that, under reasonable conditions, $\hat{u}_n$ would also converge in distribution to $\hat{u}$, the minimizer of $W(u)$. This is the content of the following two theorems.

THEOREM 2.1. *Let* $\{(Y_i, X_i)\}_{i=1}^n$ *be observations from model* (2.1), *where* $\{X_i\}_{i=1}^n \overset{iid}{\sim} F$ *with* F *satisfying* (2.2), $\{Z_i\}_{i=1}^n \overset{iid}{\sim} G$, *and the two sequences* $\{X_i\}_{i=1}^n$ *and* $\{Z_i\}_{i=1}^n$ *are independent. Let* $\varrho(\cdot)$ *be a loss function whose score function* $\psi(x)=\varrho'(x)$ *satisfies*:

(a) $\psi(\cdot)$ *is Lipschitz of order* τ_1,

$$|\psi(x)-\psi(y)|\leqslant C|x-y|^{\tau_1},$$

for some constant $\tau_1>\max(\alpha-1, 0)$ *and some positive constant* C;

(b) $E|\psi(Z_1)|^{\tau_2}<\infty$ *for some* $\tau_2>\alpha$;

(c) $E\psi(Z_1)=0$ *if* $\alpha\geqslant 1$.

Then on $C(\boldsymbol{R})$, $W_n(u)$ *converges in distribution to*

$$W(u)=\sum_{k=1}^{\infty}[\varrho(Z_k-u\delta_k\Gamma_k^{-1/\alpha})-\varrho(Z_k)],$$

where $\{Z_k\}$, $\{\delta_k\}$, $\{\Gamma_k\}$ *are sequences of random variables as given in Proposition* A.1 *of the Appendix.* ($C(\boldsymbol{R})$ *denotes the space of continuous functions on* $\boldsymbol{R}$ *and convergence is defined as uniform convergence on compact sets.*)

Proof. The convergence of the finite-dimensional distributions is straightforward using Propositions A.1–A.3 in the Appendix. It thus suffices to show that $W_n(\cdot)$ is tight on $C[-M, M]$ for any $M > 0$ (see Proposition 4.18 in Resnick [13] or Theorem 23, p. 108, in Pollard [11]). For $M > 0$ fixed we see by the mean value theorem that, for all $u, v \in [-M, M]$,

$$|W_n(u) - W_n(v)| = \Big|(u-v)\sum_{i=1}^{n} a_n^{-1} X_i \psi(\xi_i^{(n)})\Big|$$

$$= \Big|(u-v)\sum_{i=1}^{n} a_n^{-1} X_i \psi(Z_i) + (u-v)\sum_{i=1}^{n} a_n^{-1} X_i\big(\psi(\xi_i^{(n)}) - \psi(Z_i)\big)\Big|.$$

Since $|\xi_i^{(n)} - Z_i| \leqslant (|u| \vee |v|)\, a_n^{-1}|X_i| \leqslant M a_n^{-1}|X_i|$, this last term is bounded by

$$|u-v|\,\Big|\sum_{i=1}^{n} a_n^{-1} X_i \psi(Z_i)\Big| + |u-v|\, C M^{\tau_1} \sum_{i=1}^{n} (a_n^{-1}|X_i|)^{1+\tau_1}.$$

From Proposition A3 it follows that

$$\sum_{i=1}^{n} a_n^{-1} X_i \psi(Z_i) = O_p(1) \quad \text{and} \quad a_n^{-(1+\tau_1)} \sum_{i=1}^{n} |X_i|^{1+\tau_1} = O_p(1)$$

so that the above bound is $|u-v|\, O_p(1)$. The tightness of $W_n(\cdot)$ on $C[-M, M]$ is now immediate.

THEOREM 2.2. *If $\varrho(\cdot)$ is convex and satisfies the conditions of Theorem 2.1, and $W(\cdot)$ attains a unique minimum at u^* a.s., then*

$$a_n(\hat{\beta} - \beta) \underset{n\to\infty}{\overset{\mathscr{D}}{\rightarrow}} u^*.$$

For the proof see Lemma 2.2 of Davis et al. [5].

Remark 2.1. If the loss function $\varrho(x)$ is chosen to be strictly convex, then the limiting process $W(u)$ will be strictly convex a.s., and thus $\hat{u}$ is unique a.s.

The distribution of the limiting random variable $\hat{u}$ depends on the choice of the loss function $\varrho(\cdot)$, the error distribution G, the skewness parameter p, and the stability index α of F. Even if all of these parameters are known, tabulating the distribution of $\hat{u}$ via simulation is a formidable task. Nevertheless, resampling methods can be implemented to approximate the sampling distribution of $a_n(\hat{\beta} - \beta)$ (see Section 6).

3. Special case when $\varrho(x) = |x|^\lambda$, $\lambda \geqslant 1$. In this section we specialize to the class of convex loss functions given by $\varrho(x) = |x|^\lambda$, $\lambda \geqslant 1$. We first concentrate on the case when ϱ is strictly convex, i.e. $\lambda > 1$, and then we will return to the LAD ($\lambda = 1$) case at the end of the section.

For $\lambda > 1$, the score function $\psi(x) = \lambda x^{\langle \lambda - 1\rangle} = \lambda |x|^{\lambda-1} \operatorname{sgn} x$ (where $s^{\langle v\rangle} = |s|^v \operatorname{sgn} s$), and hence condition (b) of Theorem 2.1 becomes $E|Z_1|^{(\lambda-1)\tau_2} < \infty$

for some $\tau_2 > \alpha$. This necessarily implies that $E|Z_1|^{(\lambda-1)\alpha} < \infty$. Now, if the distribution of Z_1 is also heavy tailed, say $Z_1 \in \mathscr{D}(\gamma)$, then one should choose λ small enough to lessen the effect of the exceptionally large residuals. In fact, condition (b) of Theorem 2.1 is met, and hence the conclusion of Theorem 2.2 holds, when λ is chosen such that $\lambda - 1 < \gamma/\alpha$. On the other hand, if this inequality is reversed, then the correct scaling for $\hat{\beta} - \beta$ will be of smaller order than a_n. In other words, if possible, one should always choose λ such that $\lambda - 1 < \gamma/\alpha$.

THEOREM 3.1. *Let* $\varrho(x) = |x|^\lambda$, $\lambda > 1$, $\{X_i\}_{i=1}^n \overset{iid}{\sim} F \in \mathscr{D}(\alpha)$, $\{Z_i\}_{i=1}^n \overset{iid}{\sim} G \in \mathscr{D}(\gamma)$, *where* $0 < \alpha < 2$ *and* $0 < \gamma < 2$. *Further, assume that* $E[X_i Z_i^{\langle\lambda-1\rangle}] = 0$ *if* $\gamma/(\lambda-1) \geqslant 1$. *Let* a_n *and* b_n *be the* $1 - n^{-1}$ *quantiles of the distributions of* $|X_1|$ *and* $|Z_1|$, *respectively, i.e.*

$$a_n = \inf\{z\colon\ P(|X_1| > z) \leqslant n^{-1}\} \quad \textit{and} \quad b_n = \inf\{z\colon\ P(|Z_1| > z) \leqslant n^{-1}\}.$$

Then if $\lambda > \gamma/\alpha + 1$, *we have on* $C(\mathbf{R})$

$$W_n(u) \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} \lambda u S\left(\frac{\gamma}{\lambda-1}\right) + |u|^\lambda S\left(\frac{\alpha}{\lambda}\right) \qquad \textit{if } \alpha < \lambda,$$

$$a_n^\lambda n^{-1} W_n(u) \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} |u|^\lambda E|X_1|^\lambda \qquad \textit{if } \alpha > \lambda,$$

where

$$W_n(u) = \sum_{i=1}^n [|c_n Z_i - u a_n^{-1} X_i|^\lambda - |c_n Z_i|^\lambda],$$

$c_n = a_n^{1/(\lambda-1)} b_n^{-1}$, *and* $S(\gamma/(\lambda-1))$ *and* $S(\alpha/\lambda)$ *are independent stable random variables with indices* $\gamma/(\lambda-1)$ *and* α/λ, *respectively. Moreover,*

$$\text{(3.1)} \qquad \begin{aligned} a_n c_n(\hat{\beta}-\beta) &\underset{n\to\infty}{\overset{\mathscr{D}}{\to}} \left[S\left(\frac{r}{\lambda-1}\right) \Big/ S\left(\frac{\alpha}{\lambda}\right)\right]^{\langle 1/(\lambda-1)\rangle} && \textit{if } \alpha < \lambda, \\ a_n c_n(\hat{\beta}-\beta) &\underset{n\to\infty}{\overset{\mathscr{D}}{\to}} 0 && \textit{if } \alpha > \lambda. \end{aligned}$$

Note that $c_n \to 0$, *so that the scaling* $a_n c_n$ *goes to infinity at a slower rate than* a_n.

Proof. By inequalities I.4 and I.5 of the Appendix, we have

$$\Big|W_n(u) + \lambda u \sum_{i=1}^n a_n^{-1} X_i (c_n Z_i)^{\langle\lambda-1\rangle} - |u|^\lambda \sum_{i=1}^n |a_n^{-1} X_i|^\lambda\Big|$$

$$\leqslant \begin{cases} C\sum_{i=1}^n |u a_n^{-1} X_i|^{1+b} |c_n Z_i|^{\lambda-1-b} & \text{if } 1 < 1+b < \lambda < 2, \\ 0 & \text{if } \lambda = 2, \\ C\sum_{i=1}^n [|u a_n^{-1} X_i|^{\lambda-1} |c_n Z_i| + (u a_n^{-1} X_i)^2 |c_n Z_i|^{\lambda-2}] & \text{if } \lambda > 2. \end{cases}$$

Since X_1 and Z_1 are independent, $|X_1|^j \in \mathscr{D}(\alpha/j)$, and $|Z_1|^{\lambda-j} \in \mathscr{D}(\gamma/(\lambda-j))$, it follows that

$$|X_1|^j |Z_1|^{\lambda-j} \in \mathscr{D}\left(\min\left(\frac{\alpha}{j}, \frac{r}{\lambda-j}\right)\right),$$

and hence $c_n = a_n^{1/(\lambda-1)} b_n^{-1} \to 0$. Now, for the case $\lambda > 2$,

$$\begin{aligned}\sum_{i=1}^{n} |a_n^{-1} X_i|^2 |c_n Z_i|^{\lambda-2} &= c_n^{\lambda-2} a_n^{-2} \max(a_n^2, b_n^{\lambda-2}) O_p(1)\\ &= \max\left(c_n^{\lambda-2}, (b_n c_n)^{\lambda-2} a_n^{-2}\right) O_p(1)\\ &= \max(c_n^{\lambda-2}, a_n^{-\lambda/(\lambda-1)}) O_p(1) = o_p(1).\end{aligned}$$

Similarly, for $1 < b+1 < \lambda < 2$, we have

$$\begin{aligned}\sum_{i=1}^{n} |a_n^{-1} X_i|^{1+b} |c_n Z_i|^{\lambda-1-b} &= \sum_{i=1}^{n} a_n^{-(1+b)} c_n^{(\lambda-1-b)} |X_i|^{1+b} |Z_i|^{\lambda-1-b}\\ &= a_n^{-(1+b)} c_n^{(\lambda-1-b)} \max(a_n^{1+b}, b_n^{\lambda-1-b}) O_p(1)\\ &= \max(c_n^{\lambda-1-b}, a_n^{-b\lambda/(\lambda-1)}) O_p(1) = o_p(1).\end{aligned}$$

Therefore,

$$\begin{aligned}W_n(u) &= -\lambda u \sum_{i=1}^{n} a_n^{-1} X_i (c_n Z_i)^{\langle\lambda-1\rangle} + |u|^\lambda \sum_{i=1}^{n} |a_n^{-1} X_i|^\lambda + o_p(1)\\ &= -\lambda u b_n^{-(\lambda-1)} \sum_{i=1}^{n} X_i Z_i^{\langle\lambda-1\rangle} + |u|^\lambda a_n^{-\lambda} \sum_{i=1}^{n} |X_i|^\lambda + o_p(1).\end{aligned}$$

In the case when $\alpha < \lambda$, the convergence of the finite-dimensional distributions of $W_n(u)$ is easily established since $X_i Z_i^{\langle\lambda-1\rangle} \in \mathscr{D}(\gamma/(\lambda-1))$, and $\gamma/(\lambda-1) < \alpha$. In particular, a point process argument, similar to the one used in Davis and Resnick [6], shows that

$$\left(b_n^{-(\lambda-1)} \sum_{i=1}^{n} X_i Z_i^{\langle\lambda-1\rangle}, a_n^{-\lambda} \sum_{i=1}^{n} |X_i|^\lambda\right) \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} \left(S\left(\frac{\gamma}{\lambda-1}\right), S\left(\frac{\alpha}{\lambda}\right)\right),$$

where $S(\gamma/(\lambda-1))$ and $S(\alpha/\lambda)$ are independent stable r.v.'s with indices $\gamma/(\lambda-1)$ and α/λ, respectively. Note that no centering is needed here since $\alpha/\lambda < 1$.

The tightness of $\{W_n\}$ on $C(\boldsymbol{R})$ is immediate since the W_n's have convex sample paths (see Remark 1 of Davis et al. [5]). Thus W_n converges in distribution on $C(\boldsymbol{R})$ to

$$W(u) = -\lambda u S\left(\frac{\gamma}{\lambda-1}\right) + |u|^\lambda S\left(\frac{\alpha}{\lambda}\right),$$

which has a unique minimum at

$$\hat{u} = \left[\frac{S(\gamma/(\lambda-1))}{S(\alpha/\lambda)}\right]^{\langle 1/(\lambda-1)\rangle}.$$

This, combined with the strict convexity of $W_n(u)$, proves (3.1) in the case $\alpha < \gamma$. For $\alpha > \lambda$,

$$\frac{1}{n}\sum_{i=1}^{n} |X_i|^\lambda \underset{n\to\infty}{\overset{\text{a.s.}}{\to}} E|X_1|^\lambda,$$

and hence

$$a_n^\lambda n^{-1} W_n(u) = -a_n^\lambda n^{-1} \lambda u O_p(1) + |u|^\lambda n^{-1} \sum_{i=1}^{n} |X_i|^\lambda + a_n^\lambda n^{-1} o_p(1)$$

$$\underset{n\to\infty}{\overset{\mathscr{D}}{\to}} |u|^\lambda E|X_1|^\lambda.$$

Since the limit process has a unique minimum at $\hat{u} = 0$, the second case of (3.1) follows.

Remark 3.1. The scaling constants for $\hat{\beta}-\beta$ grow at a slower rate for $\lambda \geqslant \gamma/\alpha+1$ than for $\lambda < \gamma/\alpha+1$. Hence one should choose, using prior information about α and γ if possible, a loss function with $\lambda < \gamma/\alpha+1$.

Remark 3.2. For certain choices of λ, such as $\lambda > \max(\alpha, \gamma/\alpha+1, \alpha/(\alpha-\gamma))$, $\hat{\beta}$ is not even consistent. (In this case, $a_n c_n \to 0$.)

By combining the results of Theorems 2.1, 2.2 and 3.1 with $\lambda = 2$, the asymptotic behavior of the LS estimate can be summarized as follows.

Corollary 3.1. *Under the assumptions of Theorem* 3.1, *the least squares estimate* $\hat{\beta}_{\text{LS}}$ *of* β *has the following asymptotic behavior*:

$$a_n(\hat{\beta}_{\text{LS}}-\beta) \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} S(\alpha)/S(\alpha/2) \quad \text{if } 2 \geqslant \gamma > \alpha,$$

$$a_n^2 b_n^{-1}(\hat{\beta}_{\text{LS}}-\beta) \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} S(\gamma)/S(\alpha/2) \quad \text{if } \gamma < \alpha < 2,$$

where

$$S(\alpha) = \sum_{k=1}^{\infty} Z_k \delta_{1k} \Gamma_{1k}^{-1/\alpha}, \quad S(\alpha/2) = \sum_{k=1}^{\infty} \Gamma_{1k}^{-2/\alpha} \quad \text{and} \quad S(\gamma) = \sum_{k=1}^{\infty} Z_k \delta_{2k} \Gamma_{2k}^{-1/\gamma}$$

are stable r.v.'s with indices α, $\alpha/2$ *and* γ, *respectively*, $S(\gamma)$ *is independent of* $S(\alpha/2)$. (*The sequences* $\{\delta_{ik}\}_{k=1}^{\infty}$ *and* $\{\Gamma_{ik}\}_{k=1}^{\infty}$, $i = 1, 2$, *are independent copies of the sequences* $\{\delta_k\}_{k=1}^{\infty}$ *and* $\{\Gamma_k\}_{k=1}^{\infty}$ *described in Proposition* A.1.)

A version of this corollary was established by Daren Cline (personal communication) by using more analytic methods.

Remark 3.3. In the case when $F \in \mathscr{D}(2)$ and $\gamma < 2$, there exists a sequence $\tilde{a}_n = nh(n)$, where $h(\cdot)$ is a slowly varying function such that

$$\sum_{i=1}^{n} X_i^2/\tilde{a}_n \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} 1$$

(see Feller [10]). Thus, with $c_n = \tilde{a}_n^{1/2}/b_n^{-1}$, we have

$$W_n(u) = \sum_{i=1}^{n} [|c_n Z_i - u\tilde{a}_n^{-1/2} X_i|^{\lambda} - |c_n Z_i|^{\lambda}]$$

$$= -2u \sum_{i=1}^{n} b_n^{-1} X_i Z_i + u^2 \sum_{i=1}^{n} \tilde{a}_n^{-1} X_i^2 \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} -2uS(\gamma)+u^2,$$

which implies

$$\tilde{a}_n b_n^{-1} (\hat{\beta}_{\mathrm{LS}} - \beta) \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} S(\gamma).$$

Remark 3.4. In the case $\gamma = \alpha$ and $E|X|^{\alpha} = E|Z|^{\alpha} = \infty$, we have $X_1 Z_1 \in \mathscr{D}(\alpha)$ (Cline [3]). If $\tilde{a}_n$ is the $1-n^{-1}$ quantile of $X_1 Z_1$, then $\tilde{a}_n/a_n \to \infty$, $\tilde{a}_n/b_n \to \infty$ (see Davis and Resnick [6]), and hence

$$W_n(u) = \sum_{i=1}^{n} [|a_n \tilde{a}_n^{-1} Z_i - u a_n^{-1} X_i|^{\alpha} - |a_n \tilde{a}_n Z_i|^{\alpha}]$$

$$= -2u\tilde{a}_n^{-1} \sum_{i=1}^{n} X_i Z_i + u^2 a_n^{-2} \sum_{i=1}^{n} X_i^2 \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} -2uS(\alpha)+u^2 S(\alpha/2),$$

where $S(\alpha)$ and $S(\alpha/2)$ are independent stable r.v.'s with indices α and $\alpha/2$, respectively. Thus

$$a_n^2 \tilde{a}_n^{-1} (\hat{\beta}_{\mathrm{LS}} - \beta) \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} S(\alpha)/S(\alpha/2).$$

We conclude this section with a discussion of the LAD estimate (i.e. $\varrho(x) = |x|$). In this case,

$$W_n(u) = \sum_{i=1}^{n} [|Z_i - u a_n^{-1} X_i| - |Z_i|],$$

and the natural candidate for the limiting process is

$$W(u) = \sum_{k=1}^{\infty} [|Z_k - u\delta_k \Gamma_k^{-1/\alpha}| - |Z_k|],$$

which is well defined when $\alpha < 1$, because $|W(u)| \leqslant \sum_{k=1}^{\infty} |u| \Gamma_k^{-1/\alpha} < \infty$ a.s. If $\alpha \geqslant 1$, additional assumptions on the behavior of Z_1 near zero are required in order to ensure that the defining sum for W converges.

THEOREM 3.2. *Let* $\{X_i\} \overset{iid}{\sim} F \in \mathscr{D}(\alpha)$ *and* $\{Z_i\} \overset{iid}{\sim} G$ *be two independent sequences of random variables with* $0 < \alpha < 2$. *Assume that either*

(a) $\alpha < 1$; *or*

(b) $\alpha > 1$, $E|Z_1|^{\tau} < \infty$ *for some* $\tau < 1-\alpha$ *and* Z_1 *has median* 0; *or*

(c) $\alpha = 1$ *and* $E(\ln|Z_1|) > -\infty$.

Then

$$W_n(\cdot) \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} W(\cdot) \quad \text{in } C(\mathbb{R}).$$

Moreover, if $W(\cdot)$ has a unique minimum a.s., then

$$a_n(\hat{\beta}-\beta) \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} \hat{u},$$

where $\hat{\beta}$ is the LAD estimate of β, and $\hat{u}$ minimizes $W(\cdot)$.

Proof. The proof is nearly identical to the argument given for Theorem 4.1 in Davis et al. [5], and hence is omitted.

4. Multivariate regressors and unknown location. In this section, we describe how the results of the preceding two sections for the simple linear model can be extended to the multivariate linear model,

$$Y_i = X_i'\boldsymbol{\beta}+Z_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)'$, $X_i = (X_{i1}, \ldots, X_{id})'$, $i = 1, \ldots, n$, are in $\boldsymbol{R}^d$. We consider the asymptotic behavior of the M-estimator under the following two types of heavy-tailed conditions.

Assumption 1. F is d-variate regularly varying, i.e., there exists a sequence $a_n \to \infty$ and a Lévy measure μ on $(\boldsymbol{R}^d, B(\boldsymbol{R}^d))$ such that

$$(4.1) \qquad nP(a_n^{-1}X_1 \in \cdot) \underset{n\to\infty}{\overset{v}{\to}} \mu(\cdot)$$

($\overset{v}{\to}$ is vague convergence on $\boldsymbol{R}^d\backslash(0, 0, \ldots, 0)$) or, in polar coordinates,

$$(4.2) \qquad nP(a_n^{-1}\|X_1\| > r, \theta(X_1)\in B) \to r^{-\alpha}S(B),$$

where S is a probability measure on the $(d-1)$-dimensional unit sphere

$$S^{d-1} = \{(t_1, \ldots, t_d)\in \boldsymbol{R}^d, \sum_{i=1}^{d} t_i^2 = 1\}$$

(see Resnick [13]).

Assumption 2. The components of X_1 are independent ($F = F_1 \ldots F_d$) with $F_j\in \mathscr{D}(\alpha_j)$, $j = 1, \ldots, d$.

The M-estimate $\hat{\boldsymbol{\beta}}$ minimizes the objective function

$$\sum_{i=1}^{n} \varrho(Y_i-X_i'\boldsymbol{\phi}) = \sum_{i=1}^{n} \varrho(Z_i-X_i'(\boldsymbol{\phi}-\boldsymbol{\beta}))$$

with respect to $\boldsymbol{\phi}\in \boldsymbol{R}^d$. As before, we build the parameter normalization into the objective function and treat this new object as a stochastic process. Under Assumption 1, the normalization will be the same for each of the coefficient parameters, and the relevant stochastic process is given by

$$W_n^{(1)}(\boldsymbol{u}) = \sum_{i=1}^{n} [\varrho(Z_i-a_n^{-1}X_i'\boldsymbol{u})-\varrho(Z_i)],$$

where $\boldsymbol{u} \in \boldsymbol{R}^d$. Then, for each n, $\hat{\boldsymbol{u}}_n^* = a_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ minimizes $W_n^{(1)}(\boldsymbol{u})$. Under Assumption 2, each component of the parameter vector may require a different normalization, so that the required stochastic process is

$$W_n^{(2)}(u_1, \ldots, u_d) = \sum_{i=1}^{n} [\varrho(Z_i - u_1 a_{n1}^{-1} X_{i1} - \ldots - u_d a_{nd}^{-1} X_{id}) - \varrho(Z_i)],$$

where

$$a_{nj} = \inf\{x\colon P(|X_{1j}| > x) \leqslant 1/n\}, \qquad p_j = \lim_{x \to \infty} \frac{P(X_{1j} > x)}{P(|X_{1j}| > x)}.$$

Then if $\hat{\boldsymbol{u}}_n^* = (\hat{u}_{n1}^*, \ldots, \hat{u}_{nd}^*)'$ minimizes W_n, the j-th component is equal to $a_{nj}(\hat{\beta}_j - \beta_j)$.

The following results are straightforward generalizations of the arguments given in Theorems 2.1, 2.2 and 3.1 for the simple linear model.

COROLLARY 4.1. *Assume that the loss function* $\varrho(x)$ *satisfies conditions* (a)–(c) *of Theorem* 2.1, *where under Assumption* 2, $\alpha := \max(\alpha_1, \ldots, \alpha_d)$. *On* $C(\boldsymbol{R}^d)$,

$$W_n^{(i)}(\boldsymbol{u}) \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} W^{(i)}(\boldsymbol{u}) \quad \textit{for } i = 1, 2,$$

where under Assumption 1,

$$W^{(1)}(\boldsymbol{u}) = \sum_{k=1}^{\infty} [\varrho(Z_k - \Gamma_{k1}^{-1/\alpha} \boldsymbol{\eta}_k' \boldsymbol{u}) - \varrho(Z_k)],$$

and under Assumption 2,

$$W^{(2)}(u_1, \ldots, u_d) = \sum_{k=1}^{\infty} [\varrho(Z_k - u_1 \Gamma_{k1}^{-1/\alpha_1} \delta_{k1} - \ldots - u_d \Gamma_{kd}^{-1/\alpha_d} \delta_{kd}) - \varrho(Z_k)].$$

Here $\{\Gamma_{k1}\}, \ldots, \{\Gamma_{kd}\}, \{\delta_{k1}\}, \ldots, \{\delta_{kd}\}, \{Z_k\}$, *and* $\{\boldsymbol{\eta}_k\}$ *are independent sequences of random variables (vectors) where for each i*, $\{\Gamma_{ki}\} \overset{d}{=} \{\Gamma_k\}$, *and* $\{\delta_{ki}\} \overset{d}{=} \{\delta_k\}$ ($\{\Gamma_k\}$ *and* $\{\delta_k\}$ *are as defined in Proposition* A.1 *with* $p = p_i$), $\{Z_k\} \overset{iid}{\sim} G$, *and* $\{\boldsymbol{\eta}_k\} \overset{iid}{\sim} S$ (S *is the distribution given in* (4.2)). *Furthermore, if the loss function* $\varrho(\cdot)$ *is convex and* $W(\cdot)$ *has a unique minimum* $\hat{\boldsymbol{u}} \in \boldsymbol{R}^d$ *a.s., then*

$$\hat{\boldsymbol{u}}_n \underset{n\to\infty}{\overset{\mathscr{D}}{\to}} \hat{\boldsymbol{u}}.$$

COROLLARY 4.2. *Let* $Z_1 \in \mathscr{D}(\gamma)$, $0 < \gamma < 2$, $\varrho(x) = |x|^\lambda$, $\lambda > \max(\alpha, \gamma/\alpha + 1)$ *and define*

$$W_n(\boldsymbol{u}) = \sum_{i=1}^{n} [\varrho(c_n Z_i - a_n^{-1} \boldsymbol{X}_i' \boldsymbol{u}) - \varrho(c_n Z_i)],$$

where $c_n = a_n^{1/(\lambda-1)} b_n^{-1}$ *and* b_n *is the* $(1-n^{-1})$*-quantile of* $|Z_1|$. *Then, under Assumption* 1,

$$W_n \xrightarrow[n\to\infty]{\mathscr{D}} W,$$

where

$$W(\boldsymbol{u}) = -\lambda \sum_{j=1}^{d} u_j S_j\left(\frac{\gamma}{\lambda-1}\right) + \sum_{j=1}^{d} |u_j|^{\lambda} S_j\left(\frac{\alpha}{\gamma}\right),$$

$\{S_j(\gamma/(\lambda-1))\}_{j=1}^{d}$ *and* $\{S_j(\alpha/\lambda)\}_{j=1}^{d}$ *are stochastically independent, each is a stable vector with intensity measure given by the vague limits of*

$$nP(b_n^{-1} X_i Z_i^{\langle\lambda-1\rangle} \in \cdot) \quad \text{and} \quad nP\left(a_n^{-\lambda}(|X_1|^{\lambda}, \ldots, |X_{1d}|^{\lambda}) \in \cdot\right),$$

respectively. Moreover, since $W(\boldsymbol{u})$ *has a unique minimum at* $\hat{\boldsymbol{u}} = (\hat{u}_1, \ldots, \hat{u}_d)$ *with*

$$\hat{u}_j = \left[S_j\left(\frac{\gamma}{\lambda-1}\right) \Big/ S_j\left(\frac{\alpha}{\lambda}\right)\right]^{\langle 1/(\lambda-1)\rangle}, \quad j = 1, \ldots, d,$$

we have

$$a_n c_n (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow[n\to\infty]{\mathscr{D}} \hat{\boldsymbol{u}}.$$

Unknown location. An intercept term can also be included in the model without much difficulty. For the linear model with intercept β_0,

$$Y_i = \beta_0 + X_i'\boldsymbol{\beta} + Z_i, \qquad i = 1, \ldots, n,$$

the M-estimate $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ of $(\beta_0, \boldsymbol{\beta})$ minimizes

$$\sum_{i=1}^{n} \varrho\,[Y_i - (\phi_0 - X_i'\boldsymbol{\phi})] = \sum_{i=1}^{n} \varrho\,[Z_i - (\phi_0 - \beta_0) - X_i'(\boldsymbol{\phi} - \boldsymbol{\beta})]$$

with respect to ϕ_0 and $\boldsymbol{\phi}$. For brevity we assume that the distribution of X_i satisfies Assumption 1. Let

$$\begin{aligned} W_n^*(u_0, \boldsymbol{u}) &= \sum_{i=1}^{n} [\varrho(Z_i - n^{-1/2} u_0 - a_n^{-1} X_i'\boldsymbol{u}) - \varrho(Z_i)] \\ &= W_n(\boldsymbol{u}) + \sum_{i=1}^{n} [\varrho(Z_i - n^{-1/2} u_0) - \varrho(Z_i)] + R_n(\boldsymbol{u}) \\ &=: W_n(\boldsymbol{u}) + Z_n(u_0) + R_n(\boldsymbol{u}), \end{aligned}$$

where $u_0 = n^{1/2}(\phi_0 - \beta_0)$, and $\boldsymbol{u} = a_n(\boldsymbol{\phi} - \boldsymbol{\beta})$. The minimum of W_n^* occurs at $\hat{u}_0 = n^{1/2}(\hat{\beta}_0 - \beta_0)$, $\hat{\boldsymbol{u}}_n = a_n(\boldsymbol{\phi} - \boldsymbol{\beta})$. If $\varrho(\cdot)$ has a Lipschitz continuous derivative $\psi(\cdot)$, then $R_n = o_p(1)$. To see this note that

$$\begin{aligned} R_n(\boldsymbol{u}) &= \sum_{i=1}^{n} [\varrho(Z_i - n^{-1/2}u_0 - a_n^{-1}X_i'\boldsymbol{u}) - \varrho(Z_i - a_n^{-1}X_i'\boldsymbol{u})] \\ &\quad - \sum_{i=1}^{n} [\varrho(Z_i - n^{-1/2}u_0) - \varrho(Z_i)] \\ &= \sum_{i=1}^{n} [h(Z_i - a_n^{-1}X_i'\boldsymbol{u}) - h(Z_i)], \end{aligned}$$

where

$$h(x) = \varrho(x - n^{1/2}u_0) - \varrho(x),$$

$$|h'(\xi_i^n)| = |\psi(\xi_i^n - n^{-1/2}u_0) - \psi(\xi_i^n)| \leqslant kn^{-1/2}|u_0|,$$

and hence

$$\begin{aligned} |R_n(\boldsymbol{u})| \leqslant kn^{-1/2}|u_0|\,a_n^{-1}\sum_{i=1}^{n}|X_i'\boldsymbol{u}| &= \begin{cases} n^{-1/2}\,O_p(1) & \text{if } \alpha < 1, \\ n^{1/2}a_n^{-1}\,O_p(1) & \text{if } \alpha \geqslant 1, \end{cases} \\ &= o_p(1). \end{aligned}$$

If $\psi(\cdot)$ also has a Lipschitz continuous derivative $\psi'(\cdot)$ and $E|\psi(Z_1)|^2 < \infty$, then on $C(\boldsymbol{R})$

$$\begin{aligned} Z_n(u_0) &= \sum_{i=1}^{n} [\varrho(Z_i - n^{-1/2}u_0) - \varrho(Z_i)] \\ &= -u_0 n^{-1/2}\sum_{i=1}^{n}\psi(Z_i) + \frac{u_0^2}{2n}\sum_{i=1}^{n}\psi'(Z_i + n^{-1/2}\xi_0^n) \\ &= -u_0 n^{-1/2}\sum_{i=1}^{n}\psi(Z_i) + \frac{u_0^2}{2n}\sum_{i=1}^{n}\psi'(Z_i) + o_p(1) \\ &\xrightarrow[n\to\infty]{\mathscr{D}} Z(u_0) = -u_0 Z + \frac{u_0^2}{2}E\psi'(Z), \end{aligned}$$

where Z is a normal r.v. with mean zero and variance $E\psi^2(Z_1)$. The limit process is minimized at

$$Z/E\psi'(Z_1) \sim N(0,\, E\psi^2(Z_1)/[E\psi'(Z_1)]^2).$$

Also, as before,

$$W_n(\cdot) \xrightarrow[n\to\infty]{\mathscr{D}} W(\cdot),$$

and since $W(\cdot)$ and $Z(\cdot)$ are independent, we have

$$\left(n^{1/2}(\hat{\beta}_0 - \beta_0),\, a_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) \xrightarrow[n\to\infty]{\mathscr{D}} \left(\frac{Z}{E\psi'(Z_1)},\, \hat{\boldsymbol{u}}\right),$$

where Z and $\hat{\boldsymbol{u}}$ are independent.

5. Least dispersion estimation. In this section we consider the least dispersion estimate for the simple linear regression model,

$$Y_i = \beta X_i + Z_i, \quad i = 1, \ldots, n, \tag{5.1}$$

where $\{Z_i\}_{i=1}^n \overset{\text{iid}}{\sim} G \in \mathscr{D}(\gamma)$ and $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} F \in \mathscr{D}(\alpha)$ with $0 < \alpha,\ \gamma < 2$. We say that a linear estimate $\hat{\beta} = \sum_{i=1}^n c_i Y_i$, where $c_1, \ldots, c_n$ are functions of $X_1, \ldots, X_n$, is *unbiased* if $\sum_{i=1}^n c_i X_i = 1$ a.s. (If the mean of Y_i exists and $EZ_i = 0$, then such an estimate is unbiased.) The summability constraint on the c_i's implies that

$$\hat{\beta} - \beta = \sum_{i=1}^n c_i Y_i - \beta = \sum_{i=1}^n c_i (\beta X_i + Z_i) - \beta = \sum_{i=1}^n c_i Z_i.$$

Given $X_1, \ldots, X_n$, the dispersion or relative dispersion of $\hat{\beta} - \beta$ (see Cline and Brockwell [4] and Davis and Resnick [6]) is given by

$$\lim_{z \to \infty} \frac{P(|\hat{\beta} - \beta| > z \mid X_1, \ldots, X_n)}{P(|Z_1| > z)} = \lim_{z \to \infty} \frac{P(|\sum_{i=1}^n c_i Z_i| > z \mid X_1, \ldots, X_n)}{P(|Z_1| > z)}$$
$$= \sum_{i=1}^n |c_i|^\gamma,$$

where the last equality follows directly from the proposition on p. 278 of Feller [10]. The least dispersion estimate $\hat{\beta}_{\mathrm{LD}}$ is then defined as the estimate which minimizes the dispersion of $\hat{\beta} - \beta$ among all linearly unbiased estimates. In other words, $\hat{\beta}_{\mathrm{LD}} = \sum_{i=1}^n c_i Y_i$, where the c_i's minimize $\sum_{i=1}^n |c_i|^\gamma$ subject to the constraint $\sum_{i=1}^n c_i X_i = 1$ a.s.

One may interpret the dispersion as the asymptotic scale (raised to the γ power) of $\hat{\beta} - \beta$. In fact, if the Z_i's have a symmetric stable distribution, then the dispersion is equal to the γ power of the scale of $\hat{\beta} - \beta$ conditional on $X_1, \ldots, X_n$. Moreover, for any linear unbiased estimate $\tilde{\beta}$, we have

$$P(|\tilde{\beta} - \beta| > z \mid X_1, \ldots, X_n) \geqslant P(|\hat{\beta} - \beta| > z \mid X_1, \ldots, X_n) \text{ a.s.} \tag{5.2}$$

for all large z (for all z if Z_1 has a symmetric stable distribution).

The least dispersion estimate has an explicit form given by

$$\hat{\beta}_{\mathrm{LD}} = \begin{cases} \sum_{i=1}^n X_i^{\langle 1/(\gamma-1) \rangle} Y_i / \sum_{i=1}^n |X_i|^{\gamma/(\gamma-1)} & \text{if } \gamma > 1, \\ Y_{\tau_n} / X_{\tau_n} & \text{if } \gamma \leqslant 1, \end{cases}$$

where τ_n satisfies: $|X_{\tau_n}| = \max_{j \leqslant n} |X_j|$. Blattberg and Sargent [2] discuss the merits of this estimator relative to LS and LAD. The following theorem gives the asymptotic distribution of $\hat{\beta}_{\mathrm{LD}}$.

THEOREM 5.1. *For the simple linear model given in* (5.1) *where* $EZ_1 = 0$ *if* $\gamma > 1$, *we have*

$$a_n(\hat{\beta}_{\mathrm{LD}}-\beta)\xrightarrow[n\to\infty]{\mathscr{D}}\begin{cases}\dfrac{\sum_{k=1}^{\infty} Z_k\delta_k\Gamma_k^{-1/\alpha(\gamma-1)}}{\sum_{k=1}^{\infty}\Gamma_k^{-\gamma/\alpha(\gamma-1)}} & \text{if } \gamma>1,\\ Z_1\Gamma_1^{1/\alpha} & \text{if } \gamma\leqslant 1,\end{cases}$$

where the sequences $\{a_n\}$, $\{\delta_k\}$, and $\{\Gamma_k\}$ are as defined in Proposition A.1. *Note that in the $\gamma>1$ case, the limit random variable is a ratio of two dependent stable random variables with indices $\alpha(\gamma-1)$ and $\alpha(\gamma-1)/\gamma$, respectively.*

Remark 5.1. While the least dispersion estimate and the M-estimate use the same scaling a_n, the limit distribution for the least dispersion estimate is more tractable, at least if γ is known. The limit distribution in Theorem 5.1 remains valid even if $Z_1\notin\mathscr{D}(\gamma)$. For example, the conclusion of the theorem holds if $Z_1\in\mathscr{D}(\gamma)$ is replaced by $E|Z_1|^\delta<\infty$ for some $\delta>\alpha(\gamma-1)$. This makes the choice of γ less critical. In this case, however, the LD estimate no longer has the interpretation as the linear unbiased estimate which minimizes the asymptotic scale of $\hat{\beta}-\beta$.

Remark 5.2. For $\gamma>1$, the least dispersion estimate $\hat{\beta}_{\mathrm{LD}}$ may be expressed as a weighted least squares estimate. Using the weights,

$$W_i=|X_i|^{-1+1/(\gamma-1)}/c \quad \text{with } \sum_{i=1}^{n} W_i=1,$$

$\hat{\beta}_{\mathrm{LD}}$ minimizes

$$\sum_{i=1}^{n}(Y_i-\beta X_i)^2 W_i.$$

As expected, the weights increase for heavier tails of the noise (γ decreasing).

6. Bootstrapping the M-estimate. Direct application of the results in the preceding sections for making inferences about the parameter vector $\boldsymbol{\beta}$ is difficult without having more detailed information on the distributions of the regressors and noise. For example, in the simple linear model situation of Theorem 2.2, a 95% confidence interval for β is given by

$$\left(\hat{\beta}-\frac{u_{0.025}}{a_n},\ \hat{\beta}+\frac{u_{0.975}}{a_n}\right),$$

where $u_{0.025}$ and $u_{0.975}$ are the 0.025 and 0.975 quantiles of the limit random variable $\hat{u}$. Unless F is completely known, even the normalizing constants a_n are difficult to estimate. The scaling problem may be obviated by using random normalization. Since

$$a_n^{-1}M_n:=a_n^{-1}\max\{|X_1|,\ldots,|X_n|\}\xrightarrow[n\to\infty]{\mathscr{D}}\Gamma_1^{-1/\alpha},$$

it is easy to see that

$$M_n(\hat{\beta}_n-\beta)\xrightarrow[n\to\infty]{\mathscr{D}}\tilde{u}:=\hat{u}\Gamma_1^{1/\alpha}. \tag{6.1}$$

The remaining task is then to compute the quantiles of the distribution of $\tilde{u}$. If one can simulate from the distribution of the noise, and the α and p parameters of F are known, then it is possible to simulate replicates of $\hat{u}$, and hence compute the relevant quantiles. However, in the case when the distribution of the noise and the parameters of F are unknown, bootstrapping methods may be used to approximate the distribution of $M_n(\hat{\beta}_n-\beta)$.

To implement the bootstrap in this context, let (Y_i, X_i) be observations from the simple linear model (2.1) and suppose $\hat{\beta}_n$ is the M-estimate of β under the loss function ϱ which is assumed to satisfy the assumptions of Theorem 2.1. Denote the estimated residuals by

$$\hat{Z}_i = Y_i-\hat{\beta}_n X_i, \quad i = 1, \ldots, n,$$

and let

$$\hat{F}_n(z, x) = \left(n^{-1}\sum_{i=1}^{n} I(\hat{Z}_i \leqslant z)\right)\left(n^{-1}\sum_{i=1}^{n} I(X_i \leqslant x)\right)$$

be the product distribution based on the empirical distributions of $(\hat{Z}_1, \ldots, \hat{Z}_n)$ and $(X_1, \ldots, X_n)$, respectively. Next, a random sample $\{(Z_i^*, X_i^*), i = 1, \ldots, m_n\}$ is generated from the distribution $\hat{F}_n$ from which we get bootstrap replicates of the Y_i's given by

$$Y_i^* = \hat{\beta}_n X_i^*+Z_i^*, \quad i = 1, \ldots, m_n.$$

The bootstrap replicate $\hat{\beta}_{m_n}^*$ of $\hat{\beta}_n$ is then computed as the M-estimate based on the observations $(Y_1^*, X_1^*), \ldots, (Y_{m_n}^*, X_{m_n}^*)$, i.e., $\hat{\beta}_{m_n}^*$ minimizes $\sum_{i=1}^{m_n}\varrho(Y_i^*-\phi X_i^*)$. If M_n^* denotes the maximum of $|X_1^*|, \ldots, |X_{m_n}^*|$, then it is shown in Davis and Wu [9] that, provided $m_n \to \infty$ and $m_n/n \to 0$,

$$P\left(M_n^*(\hat{\beta}_{m_n}^*-\hat{\beta}_n)\in\cdot \mid \mathscr{X}, \mathscr{Y}\right) \underset{n\to\infty}{\overset{\mathscr{P}}{\to}} P(\tilde{u}\in\cdot),$$

where $\overset{\mathscr{P}}{\to}$ is convergence in probability relative to the weak topology on the space of probability measures on $\boldsymbol{R}$ and $\mathscr{X} = \{X_j\}_{j=1}^{\infty}$, $\mathscr{Y} = \{Y_j\}_{j=1}^{\infty}$.

APPENDIX

This section contains the technical complements to Sections 2–5. Much of the requisite background material on point processes, as well as notation and definitions, can be found in Davis and Resnick [6], Resnick [13], and Davis et al. [5].

PROPOSITION A.1. *Let* $\{X_i\}_{i=1}^{n} \overset{iid}{\sim} F$ *and* $\{Z_i\}_{i=1}^{n} \overset{iid}{\sim} G$ *be independent sequences where* $F\in\mathscr{D}(\alpha)$, $0 < \alpha < 2$. *Set*

$$a_n = \inf\{x\colon\ P(|X_1| > x) \leqslant n^{-1}\}.$$

Then

$$N_n = \sum_{i=1}^{n} \varepsilon_{(Z_i, ua_n^{-1} X_i)} \underset{n\to\infty}{\overset{\mathscr{D}}{\rightarrow}} N := \sum_{k=1}^{\infty} \varepsilon_{(Z_k, u\delta_k \Gamma_k^{-1/\alpha})}$$

in the space $M_p\{[-\infty, \infty]\times([-\infty, \infty]\backslash 0)\}$ *of Radon point measures on*

$$[-\infty, \infty]\times([-\infty, \infty]\backslash 0),$$

where $\{Z_k\}$, $\{\delta_k\}$, $\{\Gamma_k\}$ *are independent sequences of random variables,* $\{Z_k\}_{k=1}^{\infty} \overset{iid}{\sim} G$, $\{\delta_k\}$ *are iid with* $P(\delta_k = 1) = p = 1-P(\delta_k = -1)$ *with* p *given in* (2.2), *and* $\Gamma_k = E_1 + \ldots + E_k$, *where* E_i*'s are iid exponential r.v.'s with mean* 1.

The proof is clear from Resnick [13].

Proposition A.2. *Let* $g(x, y) = [\varrho(x+y)-\varrho(x)]\, I(|y| > \delta)$, $\delta > 0$; *then*

$$\begin{aligned} N_n(g) &= \sum_{i=1}^{n} [\varrho(Z_i + ua_n^{-1} X_i) - \varrho(Z_i)]\, I(|ua_n^{-1} X_i| > \delta) \\ &\underset{n\to\infty}{\overset{\mathscr{D}}{\rightarrow}} \sum_{k=1}^{\infty} [\varrho(Z_k + u\delta_k \Gamma_k^{-1/\alpha}) - \varrho(Z_k)]\, I(|u\delta_k \Gamma_k^{-1/\alpha}| > \delta). \end{aligned}$$

The proof is clear from Proposition A.1.

Proposition A.3. *Under the assumptions of Theorem* 2.1, *for any* $\varepsilon > 0$ *we have*

$$\lim_{\delta\to 0} \overline{\lim_{n\to\infty}}\, P(|N_n(g) - N_n(f)| > \varepsilon) = 0,$$

where $f(x, y) = \varrho(x+y) - \varrho(x)$.

Proof. We have

$$\begin{aligned} \sum_{i=1}^{n} [\varrho(Z_i - ua_n^{-1} X_i) - \varrho(Z_i)] &= -\sum_{i=1}^{n} ua_n^{-1} X_i \psi(\xi_i^{(n)}) \\ &= -\sum_{i=1}^{n} ua_n^{-1} X_i \psi(Z_i) + \sum_{i=1}^{n} ua_n^{-1} X_i \big(\psi(Z_i) - \psi(\xi_i^{(n)})\big), \end{aligned}$$

where $|\xi_i^{(n)} - Z_i| \leqslant |u|\, a_n^{-1}\, |X_i|$. It follows easily from pp. 89–91 of Resnick [12] that

$$\sum_{i=1}^{n} \varepsilon_{(a_n^{-1} X_i \psi(Z_i), a_n^{-1} X_i)} \underset{n\to\infty}{\overset{\mathscr{D}}{\rightarrow}} \sum_{k=1}^{\infty} \varepsilon_{(\delta_k \Gamma_k^{-1/\alpha} \psi(Z_k), \delta_k \Gamma_k^{-1/\alpha})}. \tag{A.1}$$

Also the partial sums $\sum_{i=1}^{n} a_n^{-1} X_i \psi(Z_i)$ converge in distribution without centering since the distribution of $X_i \psi(Z_i)$ is in $\mathscr{D}(\alpha)$. For $0 < \alpha < 1$, centering is not required, while for $1 < \alpha < 2$, assumption (c) in Theorem 2.1 implies $E(X_1 \psi(Z_1)) = 0$. For $\alpha = 1$, it is straightforward to check that assumption (b) also implies

$$\frac{P(X_1\psi(Z_1)>t)}{P(|X_1\psi(Z_1)|>t)}\to\frac{1}{2}\quad \text{as } t\to\infty,$$

so that centering is not needed in the $\alpha=1$ case since $X_1\psi(Z_1)$ is in the domain of attraction of a symmetric Cauchy distribution. It follows, by applying standard point process arguments to (A.1), that

$$\Big(\sum_{i=1}^{n} a_n^{-1}X_i\psi(Z_i),\ \sum_{i=1}^{n} a_n^{-1}X_i\psi(Z_i)I(a_n^{-1}|X_i|>\delta)\Big)$$
$$\xrightarrow[n\to\infty]{\mathcal{D}}\Big(\sum_{k=1}^{\infty}\delta_k\Gamma_k^{-1/\alpha}\psi(Z_k),\ \sum_{k=1}^{\infty}\delta_k\Gamma_k^{-1/\alpha}\psi(Z_k)I(\Gamma_k^{-1/\alpha}>\delta)\Big),$$

whence

$$\sum_{i=1}^{n} a_n^{-1}X_i\psi(Z_i)[1-I(a_n^{-1}|X_i|>\delta)]\xrightarrow[n\to\infty]{\mathcal{D}}\sum_{k=1}^{\infty}\delta_k\Gamma_k^{-1/\alpha}\psi(Z_k)[1-I(\Gamma_k^{-1/\alpha}>\delta)].$$

Denoting the above limiting random variable by U_δ, we obtain the following formula for its characteristic function:

$$\text{(A.2)}\quad E\exp\{itU_\delta\}=\exp\Big\{-\iint_{|x|\leqslant\delta}(1-e^{itx\psi(z)})\,G(dz)\,\nu(dx)\Big\}$$
$$=\exp\Big\{-\iint_{|x|\leqslant\delta}\big[1-\cos\big(tx\psi(z)\big)-i\sin\big(tx\psi(z)\big)\big]\,G(dz)\,\nu(dx)\Big\},$$

where $\nu(\cdot)$ is the Lévy measure of a stable random variable. Now

$$\iint_{|x|\leqslant\delta}\big[1-\cos\big(tx\psi(z)\big)\big]\,G(dz)\,\nu(dx)=\int|t\psi(z)|^\alpha\int_{|u|\leqslant\delta|t\psi(z)|}(1-\cos u)\,\nu(du)\,G(dz)$$
$$\leqslant|t|^\alpha\int|\psi(z)|^\alpha G(dz)\int(1-\cos u)\,\nu(du),$$

and since $(1-\cos u)$ is integrable with respect to $\nu(du)$ on $(0,\infty)$ and $\int|\psi(z)|^\alpha G(dz)$ is finite, the double integral converges to 0 as $\delta\to0$. When $0<\alpha<1$, $\sin u$ is integrable with respect to ν, and thus the second term in the double integral converges to 0 as $\delta\to0$. If $\alpha\geqslant1$,

$$\iint_{|x|\leqslant\delta}\sin[tx\psi(z)]\,G(dz)\,\nu(dx)=\iint_{|x|\leqslant\delta}\big[\sin\big(tx\psi(z)\big)-tx\psi(z)\big]\,G(dz)\,\nu(dx),$$

and since $\{\sin[tx\psi(z)]-tx\psi(z)\}I(|x|\leqslant1)$ is integrable with respect to $G(dz)\,\nu(dx)$, the integral in (A.2) converges to 0 as $\delta\to0$. We conclude that, for all $\varepsilon>0$,

$$\lim_{\delta\to0}\overline{\lim_{n\to\infty}}\,P\Big[\Big|\sum_{i=1}^{n}a_n^{-1}X_i\psi(Z_i)I(a_n^{-1}|X_i|\leqslant\delta)\Big|>\varepsilon\Big]=0.$$

The following inequalities were used in the previous sections. We state these without proof.

Inequality I.1. For any fixed $\lambda>2$, there exists a constant C depending on λ only, such that for any $z\in\boldsymbol{R}$ and for any fixed $\lambda>2$ there exists a constant

C depending on λ only, such that for all $z \in \boldsymbol{R}$

$$\big||1+z|^{\lambda}-|z|^{\lambda}\big| \leqslant C \max(|z|^{\lambda-1}, 1).$$

INEQUALITY I.2. For any fixed $\lambda > 2$ there exists a constant C depending on λ only, such that for any $z \in \boldsymbol{R}$

$$\big||1+z|^{\lambda}-1-|z|^{\lambda}-\lambda z\big| \leqslant C \max(|z|^{\lambda-1}, z^2).$$

INEQUALITY I.3. If $1 < \lambda < 2$, then there exists a constant C such that

$$\big||1+z|^{\lambda}-1-|z|^{\lambda}-\lambda z\big| \leqslant C|z|^{b+1},$$

where $b+1 < \lambda < 2$.

INEQUALITY I.4. If $\lambda > 2$, then for any $x, y \in \boldsymbol{R}$ we have

$$\big||x+y|^{\lambda}-|x|^{\lambda}-|y|^{\lambda}-\lambda y x^{\langle\lambda-1\rangle}\big| \leqslant C[|y|^{\lambda-1}|x|+y^2|x|^{\lambda-2}].$$

INEQUALITY I.5. If $\lambda < 2$, then for any $x, y \in \boldsymbol{R}$ we have

$$\big||x+y|^{\lambda}-|x|^{\lambda}-|y|^{\lambda}-\lambda y x^{\langle\lambda-1\rangle}\big| \leqslant C|y|^{b+1}|x|^{\lambda-1-b}.$$

REFERENCES

[1] K. B. Athreya, S. Lahiri and W. Wu, *Inference for heavy tailed distributions*, preprint, 1992.

[2] R. Blattberg and T. Sargent, *Regression with non-Gaussian stable disturbances: some sampling results*, Econometrica 39 (1971), pp. 501–510.

[3] D. B. H. Cline, *Estimation and Linear Prediction for Regression, Autoregression and ARMA with Infinite Variance Data*, PhD Thesis, Department of Statistics, Colorado State University, 1983.

[4] — and P. J. Brockwell, *Linear prediction of ARMA processes with infinite variance*, Stochastic Process. Appl. 19 (1985), pp. 281–296.

[5] R. A. Davis, K. Knight and J. Liu, *M-estimation for autoregressions with infinite variance*, ibidem 40 (1992), pp. 145–180.

[6] R. A. Davis and S. I. Resnick, *Limit theory for moving averages of random variables with regularly varying tail probabilities*, Ann. Probab. 13 (1985), pp. 179–195.

[7] — *Limit theory for the sample correlation function of moving averages*, Ann. Statist. 14 (1986), pp. 533–558.

[8] — *Basic properties and prediction of max-ARMA processes*, Adv. Appl. Probab. 21 (1989), pp. 781–803.

[9] R. A. Davis and W. Wu, *Bootstrapping M-estimates in regression and autoregression with infinite variance*, preprint, 1995.

[10] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd edition, Wiley, New York 1971.

[11] D. Pollard, *Convergence of Stochastic Processes*, Springer, New York 1984.

[12] S. I. Resnick, *Point processes, regular variation and weak convergence*, Adv. Appl. Probab. 18 (1986), pp. 66–138.

[13] – *Extreme Values, Regular Variation, and Point Processes*, Springer, New York 1987.

Colorado State University
Dept. of Statistics
Firt Collins, CO 80523, U.S.A.

Received on 9.3.1995