The Annals of Statistics
1998, Vol. 26, No. 2, 755–770

LIMITING DISTRIBUTIONS FOR L_1 REGRESSION ESTIMATORS UNDER GENERAL CONDITIONS[1]

BY KEITH KNIGHT

University of Toronto

It is well known that L_1-estimators of regression parameters are asymptotically normal if the distribution function has a positive derivative at 0. In this paper, we derive the asymptotic distributions under more general conditions on the behavior of the distribution function near 0.

1. Introduction. Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \tag{1}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are unknown parameters and $\{\varepsilon_i\}$ are unobservable independent, identically distributed (i.i.d.) random variables each with median 0. For simplicity, we will assume that the x_{ki}'s are nonrandom although the results will typically hold for random x_{ki}'s. We will consider the asymptotic behavior of L_1-estimators of $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$; that is, $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ minimize the objective function

$$g_n(\boldsymbol{\phi}) = \sum_{i=1}^{n} |Y_i - \phi_0 - \phi_1 x_{1i} - \cdots - \phi_p x_{pi}|$$

over all $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_p)$.

The asymptotic behavior of L_1-estimators in regression is well known, at least in the case where the errors have a distribution function $F(t)$ which is differentiable at 0 with the derivative positive. In particular, if we denote this derivative by $\lambda = F'(0)$, we have $(X_n^T X_n)^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a $(p+1)$-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $(4\lambda^2)^{-1} I$ provided that

$$\max_{1 \le i \le n} \mathbf{x}_i^T (X_n^T X_n)^{-1} \mathbf{x}_i \to 0 \quad \text{as } n \to \infty,$$

where $\mathbf{x}_i^T = (1, x_{1i}, \ldots, x_{pi})$ and X_n is the $n \times (p+1)$ matrix whose ith row is $\mathbf{x}_i^T$. [Note that $\mathbf{x}_i^T (X_n^T X_n)^{-1} \mathbf{x}_i$, $i = 1, \ldots, n$, are simply the diagonal elements of the so-called hat matrix $X_n (X_n^T X_n)^{-1} X_n^T$.] If $n^{-1}(X_n^T X_n) \to C$ for some positive definite matrix C, then it will follow that $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to the $(p+1)$-variate normal distribution whose covariance matrix is $(4\lambda^2)^{-1} C^{-1}$. [See, e.g., Bassett and Koenker (1978), Bloomfield and Steiger (1983), Bai, Chen, Wu and Zhao (1990) and Pollard (1991) for various

Received March 1997; revised October 1997.

[1] Supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

AMS 1991 *subject classifications*. Primary 62F12, 60F05; secondary 62E20, 60F17.

Key words and phrases. L_1-estimation, linear regression, asymptotic distribution.

approaches to proving the asymptotic normality.] Second order results are given by Arcones (1996a, b), Babu (1989) and He and Shao (1996).

A natural question to ask is what happens when the distribution function does not have a positive derivative at 0. While these cases may seem pathological, they are, in fact, far from it. Indeed, while assuming the existence of a density seems reasonable, it is an assumption which is difficult to verify. In fact, previous work suggests that the asymptotic behavior of L_1-estimators is very sensitive to this assumption. For i.i.d. observations, Smirnov (1952) identifies four possible types of limiting distributions for sample quantiles and characterizes their domains of attraction. Jurečková (1983) considers the asymptotic behavior of M-estimators of location under nonregular conditions; her results include the L_1-estimator of location (namely, the sample median) as a special case. On another front, Arcones (1994) considers the asymptotic behavior of so-called L_p-median (i.e., minimizers of $\sum_{i=1}^n |Y_i - \theta|^p$) for $0 < p \leq 1/2$ and shows that the convergence rate is slower than $O_p(n^{-1/2})$.

To consider the asymptotic behavior of L_1-estimators, we will start by defining (for some sequence of constants a_n),

$$\Psi_n(t) = \int_0^t \sqrt{n}\big(F(s/a_n) - F(0)\big)\, ds, \tag{2}$$

which for each n is a convex function. If the limit of $\{\Psi_n(t)\}$ exists for each t, define

$$\Psi(t) = \lim_{n\to\infty} \Psi_n(t); \tag{3}$$

$\Psi(t)$ (if it exists) is a convex function taking values in $[0, \infty]$; note that $\Psi(t)$ may equal ∞ although typically $\Psi(t)$ will be finite. [See Examples 3 and 4, in Section 3, for cases where $\Psi(t) = \infty$ for certain t.]

The exact form of Ψ in (3) can be more easily obtained by considering the limit of $\sqrt{n}(F(t/a_n) - F(0))$; if

$$\lim_{n\to\infty} \sqrt{n}(F(t/a_n) - F(0)) = \psi(t), \tag{4}$$

then typically

$$\Psi(t) = \int_0^t \psi(s)\, ds.$$

In the case where $F(x)$ is differentiable at $x = 0$ [with $F'(0) > 0$] then $a_n = \sqrt{n}$ and $\psi(t) = \lambda t$, where $\lambda = F'(0)$, and so $\Psi(t) = \lambda t^2/2$. More generally, condition (4) includes cases where F has one-sided derivatives at 0 [$\psi(t) = \lambda^+ t$ for $t > 0$ and $\psi(t) = \lambda^- t$ for $t < 0$ where $a_n = \sqrt{n}$] or is regularly varying in a neighborhood of 0. These conditions are very similar to those given by Smirnov (1952). These assumptions are somewhat weaker than those used in Jurečková (1983) for the location case. In particular, notice that it is not necessary to assume that F is absolutely continuous (with respect to Lebesgue measure); in fact, F can contain discrete components.

We will show that (under suitable regularity conditions on the design) $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ converges in distribution. To do this, we will first modify the objective function g_n as follows:

$$Z_n(\mathbf{u}) = \frac{a_n}{\sqrt{n}} \sum_{i=1}^{n} \big[|\varepsilon_i - \mathbf{x}_i^T \mathbf{u}/a_n| - |\varepsilon_i| \big]. \tag{5}$$

It is easy to see that the vector $\widehat{\mathbf{u}}_n$ which minimizes Z_n is simply $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$. If one now regards $\{Z_n\}$ as a sequence of random convex functions on R^{p+1} and if the finite-dimensional distributions of $Z_n(\mathbf{u})$ converge in distribution to those of some function $Z(\mathbf{u})$ which has a unique minimum $\mathbf{U}$, then it will follow that

$$\mathbf{U}_n = a_n\big(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\big) \to_d \mathbf{U} = \operatorname{argmin}(Z)$$

as $n \to \infty$ [see Hjørt and Pollard (1993) and Geyer (1996)].

2. Limiting distributions. We will now formally state the regularity conditions needed to find the limiting distribution of the L_1-estimator:

(A1) $\{\varepsilon_i\}$ are i.i.d. random variables with median 0 with distribution function F continuous at 0.

(A2) For some positive definite matrix C,

$$\lim_{n\to\infty} \frac{1}{n} X_n^T X_n = C.$$

(A3) For each $\mathbf{u}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Psi_n(\mathbf{u}^T \mathbf{x}_i) = \tau(\mathbf{u})$$

for some convex function $\tau(\mathbf{u})$ taking values in $[0, \infty]$, where $\{\Psi_n(t)\}$ is defined as in (2) (for some sequence $\{a_n\}$).

At this point, it is worth making a few comments on the regularity conditions. Condition (A2) is standard and implies, for example, that

$$\frac{1}{n} \max_{1\le i\le n} \mathbf{x}_i^T \mathbf{x}_i \to 0.$$

Condition (A3) is similar in spirit to (A2); it is essentially another moment condition for the $\mathbf{x}_i$'s. If $\Psi(t)$ [defined in (3)] is finite for all t, then $\tau(\mathbf{u})$ in (A3) can sometimes be evaluated as

$$\tau(\mathbf{u}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Psi\big(\mathbf{u}^T \mathbf{x}_i\big)$$

(assuming the convergence of Ψ_n to Ψ is sufficiently uniform). If this is the case and (4) holds with $\psi(t) = \lambda t$, then $\Psi(t) = \lambda t^2/2$ and so (A3) is implied by (A2) with

$$\tau(\mathbf{u}) = \frac{\lambda}{2} \mathbf{u}^T C \mathbf{u}.$$

Note that conditions (A2) and (A3) rule out certain designs for which the moment conditions are not appropriate; for example, consider $\mathbf{x}_i^T = (1, i)$. In such cases, it may be possible to reformulate conditions (A2) and (A3) so that a nondegenerate limiting distribution exists.

THEOREM 1. *Assume the model* (1) *for* $Y_1, Y_2, \ldots$ *and define* $Z_n(\cdot)$ *as in* (5). *If assumptions* (A1), (A2) *and* (A3) *hold, then for any* $(\mathbf{u}_1, \ldots, \mathbf{u}_k)$,

$$\big(Z_n(\mathbf{u}_1), \ldots, Z_n(\mathbf{u}_k)\big) \to_d \big(Z(\mathbf{u}_1), \ldots, Z(\mathbf{u}_k)\big),$$

where

$$Z(\mathbf{u}) = -\mathbf{u}^T\mathbf{W} + 2\tau(\mathbf{u})$$

with $\mathbf{W}$ *a* $(p+1)$-*variate normal random vector with mean vector* $\mathbf{0}$ *and covariance matrix* C.

(The minus sign in front of $\mathbf{u}^T\mathbf{W}$ is obviously unnecessary but will be useful in the sequel.)

COROLLARY 2. *Under the hypotheses of Theorem* 1, *if* $Z(\mathbf{u})$ *has unique minimum* (*with probability* 1), *then*

$$a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}(Z).$$

PROOF OF THEOREM 1. We will use the identity

$$|x-y| - |x| = -y\big[I(x>0) - I(x<0)\big] + 2\int_0^y \big[I(x \le s) - I(x \le 0)\big]\,ds,$$

which is valid for $x \neq 0$. [$I(A)$ is the indicator function of the set A.] Now

$$Z_n(\mathbf{u}) = Z_n^{(1)}(\mathbf{u}) + Z_n^{(2)}(\mathbf{u}),$$

where

$$Z_n^{(1)}(\mathbf{u}) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{x}_i^T\mathbf{u}\big[I(\varepsilon_i > 0) - I(\varepsilon_i < 0)\big]$$

and

$$Z_n^{(2)}(\mathbf{u}) = \frac{2a_n}{\sqrt{n}}\sum_{i=1}^n \int_0^{v_{ni}} \big[I(\varepsilon_i \le s) - I(\varepsilon_i \le 0)\big]\,ds$$
$$= \sum_{i=1}^n Z_{ni}^{(2)}(\mathbf{u})$$

(with $v_{ni} = \mathbf{x}_i^T\mathbf{u}/a_n$). By the Lindeberg–Feller central limit theorem, for each $\mathbf{u}$,

$$Z_n^{(1)}(\mathbf{u}) \to_d -\mathbf{u}^T\mathbf{W}$$

[using the fact that $n^{-1}(X_n^T X_n)$ converges to C] and the convergence in distribution holds for any finite collection of $\mathbf{u}$'s. For $Z_n^{(2)}(\mathbf{u})$, we have

$$Z_n^{(2)}(\mathbf{u}) = \sum_{i=1}^n E\big(Z_{ni}^{(2)}(\mathbf{u})\big) + \sum_{i=1}^n \big(Z_{ni}^{(2)}(\mathbf{u}) - E\big(Z_{ni}^{(2)}(\mathbf{u})\big)\big).$$

Letting $v_i = \mathbf{x}_i^T \mathbf{u} = a_n v_{ni}$, it follows that

$$\begin{aligned}
\sum_{i=1}^n E(Z_{ni}^{(2)}(\mathbf{u})) &= \frac{2a_n}{\sqrt{n}} \sum_{i=1}^n \int_0^{v_{ni}} \big(F(s) - F(0)\big)\, ds \\
&= \frac{2}{n} \sum_{i=1}^n \int_0^{v_i} \sqrt{n}\bigg(F\bigg(\frac{s}{a_n}\bigg) - F(0)\bigg)\, ds \\
&= \frac{2}{n} \sum_{i=1}^n \Psi_n\big(\mathbf{u}^T \mathbf{x}_i\big) \\
&\to 2\tau(\mathbf{u}).
\end{aligned}$$

For the remainder term in $Z_n^{(2)}(\mathbf{u})$, we have

$$\begin{aligned}
\operatorname{Var}(Z_n^{(2)}(\mathbf{u})) &= \sum_{i=1}^n E\big[\big(Z_{ni}^{(2)}(\mathbf{u}) - E(Z_{ni}^{(2)}(\mathbf{u}))\big)^2\big] \\
&\le \frac{2}{\sqrt{n}} \max_{1\le i\le n} |\mathbf{x}_i^T \mathbf{u}| \sum_{i=1}^n E\big(Z_{ni}^{(2)}(\mathbf{u})\big) \\
&= \frac{2}{\sqrt{n}} \max_{1\le i\le n} |\mathbf{x}_i^T \mathbf{u}| E\big(Z_n^{(2)}(\mathbf{u})\big).
\end{aligned}$$

Thus if $\tau(\mathbf{u}) < \infty$,

$$Z_n^{(2)}(\mathbf{u}) - E(Z_n^{(2)}(\mathbf{u})) \to_p 0 \quad \text{as } n \to \infty,$$

and so $Z_n^{(2)}(\mathbf{u}) \to_p 2\tau(\mathbf{u})$. If $\tau(\mathbf{u}) = \infty$, then

$$\begin{aligned}
P\big(\big|Z_n^{(2)}(\mathbf{u}) - E\big(Z_n^{(2)}(\mathbf{u})\big)\big| > \varepsilon E\big(Z_n^{(2)}(\mathbf{u})\big)\big) &\le \frac{\operatorname{Var}\big(Z_n^{(2)}(\mathbf{u})\big)}{\varepsilon^2 E\big(Z_n^{(2)}(\mathbf{u})\big)^2} \\
&\le 2\frac{\max_{1\le i\le n}\big|\mathbf{x}_i^T \mathbf{u}\big|/\sqrt{n}}{\varepsilon^2 E(Z_n^{(2)}(\mathbf{u}))} \\
&\to 0,
\end{aligned}$$

which implies that $Z_n^{(2)}(\mathbf{u}) \to_p \infty = \tau(\mathbf{u})$. Thus we have

$$Z_n(\mathbf{u}) \to_d -\mathbf{u}^T \mathbf{W} + 2\tau(\mathbf{u}) = Z(\mathbf{u})$$

and the finite-dimensional convergence holds trivially. □

Proof of Corollary 2. If $Z(\mathbf{u})$ has a unique minimum, then by the convexity of the Z_n's it follows that

$$\operatorname{argmin}(Z_n) = a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}(Z)$$

as $n \to \infty$. [See, e.g., Geyer (1996).] □

In general, argmin(Z) will have a multivariate normal distribution if, and only if, τ is a quadratic function. Moreover, τ will be quadratic if, and only if, the function ψ defined in (4) is linear. When τ is differentiable (with gradient $\nabla\tau$) then

$$a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \mathbf{U},$$

where $\mathbf{U}$ satisfies the equation

$$2\nabla\tau(\mathbf{U}) = \mathbf{W}.$$

Under appropriate regularity conditions, we can evaluate $\nabla\tau$ as

$$\nabla\tau(\mathbf{u}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \psi(\mathbf{u}^T\mathbf{x}_i),$$

where ψ is defined in (4).

Cases where $\tau(\mathbf{u})$ is infinite for $\mathbf{u}$ outside some compact set K can occur in many ways. For example, suppose that F is absolutely continuous with density f, where

$$\lim_{x\to 0-} f(x) = \infty \quad \text{and} \quad \lim_{x\to 0+} f(x) = \lambda > 0.$$

In this case, we have $a_n = \sqrt{n}$ and

$$\Psi(t) = \begin{cases} \infty, & \text{if } t < 0, \\ \lambda t^2/2, & \text{if } t \geq 0. \end{cases}$$

Therefore, $\tau(\mathbf{u})$ will take infinite values. This case is considered further in Example 3 below; see also Example 4.

The results given in this section can be extended in numerous directions. For example, we can obtain similar results for so-called regression quantiles [Koenker and Bassett (1978)] by replacing $|x|$ by the function $\rho_q(x) = |x| - (2q-1)x$ for some $0 < q < 1$. Similarly, we could consider regression M-estimators with discontinuous "ψ" functions similar to Jurečková (1983).

3. Examples. The limiting distributions for $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ are, in general, quite complicated (but not impossible) to determine in closed form.

In Examples 1 and 2, we assume that the ε_i's have a distribution function satisfying

$$F(x) - F(0) = \lambda \operatorname{sgn}(x)|x|^{\alpha} L(|x|) \tag{6}$$

for x in a neighborhood of 0 where $\alpha > 0$ and L is a slowly varying function at 0 [$\mathrm{sgn}(x) = 1$ if x is positive and -1 if x is negative]. In this case, we can take

$$a_n = n^{1/(2\alpha)} L^*(n) \quad \text{and} \quad \psi(t) = \lambda \,\mathrm{sgn}(t)|t|^{\alpha},$$

where L^* is a slowly varying function at infinity. For example, if $F(x)-F(0) = \lambda x \ln(|x|^{-1})$ (for x close to 0), then we can take $a_n = \sqrt{n}\ln(n)/2$ and $\psi(t) = \lambda t$. Moreover, for distributions satisfying (6), we have

$$\Psi(t) = \frac{\lambda}{\alpha+1}|t|^{\alpha+1}.$$

Thus, for larger α, condition (A3) is a more stringent "moment" condition on the $\mathbf{x}_i$'s.

EXAMPLE 1. Suppose that $Y_1, Y_2, \ldots$ are i.i.d. random variables with $Y_i = \mu + \varepsilon_i$, where μ is the median of the distribution of the Y_i's. The sample median $\widehat{\mu}_n$ minimizes

$$g_n(\theta) = \sum_{i=1}^{n} |Y_i - \theta|.$$

If we assume that the distribution function of the ε_i's satisfies (6) for some $\alpha > 0$, then the limit of $Z_n(u)$ as defined in (5) is

$$Z(u) = -uW + \frac{2\lambda}{\alpha+1}|u|^{\alpha+1},$$

where W is normal with mean 0 and variance 1; $Z(u)$ is minimized at U which satisfies

$$2\lambda|U|^{\alpha}\,\mathrm{sgn}(U) = W$$

and so $U = \mathrm{sgn}(W)|W/(2\lambda)|^{1/\alpha}$; the density of U [and hence the limiting density of $a_n(\widehat{\mu}_n - \mu)$] is

$$f(x) = \frac{\lambda\alpha\sqrt{2}}{\sqrt{\pi}}|x|^{\alpha-1}\exp(-2\lambda^2|x|^{2\alpha}).$$

[See also Smirnov (1952) and Jurečková (1983), Corollary 1.] Note that the density has a singularity at 0 if $\alpha < 1$ and that the density is bimodal if $\alpha > 1$.

EXAMPLE 2. Suppose that $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where the ε_i's are i.i.d. random variables whose distribution satisfies (6) for some $\alpha > 0$. Suppose also that half of the x_i's are 1 and the other half are -1. Then

$$\frac{1}{n}\sum_{i=1}^{n}|u_0 + u_1 x_i|^{\alpha+1} \to \frac{1}{2}\big(|u_0 - u_1|^{\alpha+1} + |u_0 + u_1|^{\alpha+1}\big).$$

The limit of $Z_n(u_0, u_1)$ is

$$Z(u_0, u_1) = -(u_0 W_0 + u_1 W_1) + \frac{\lambda}{\alpha+1}\big(|u_0 - u_1|^{\alpha+1} + |u_0 + u_1|^{\alpha+1}\big),$$

where W_0 and W_1 are independent normal random variables each with mean 0 and variance 1. If U_0 and U_1 minimize $Z(u_0, u_1)$, then they satisfy the equations

$$\lambda\big[|U_0+U_1|^\alpha \operatorname{sgn}(U_0+U_1) + |U_0-U_1|^\alpha \operatorname{sgn}(U_0-U_1)\big] = W_0,$$
$$\lambda\big[|U_0+U_1|^\alpha \operatorname{sgn}(U_0+U_1) - |U_0-U_1|^\alpha \operatorname{sgn}(U_0-U_1)\big] = W_1.$$

The joint density of (U_0, U_1) is then

$$f(u_0,u_1) = \frac{2\lambda^2\alpha^2}{\pi}|u_0^2 - u_1^2|^{\alpha-1}\exp\big[-\lambda^2\big(|u_0+u_1|^{2\alpha} + |u_0-u_1|^{2\alpha}\big)\big].$$

Thus $a_n(\widehat{\beta}_{n0}-\beta_0)$ and $a_n(\widehat{\beta}_{n1}-\beta_1)$ are asymptotically independent if, and only if, $\alpha = 1$. The asymptotic marginal densities are, in this example, identical; if $\alpha \le 1/2$, the marginal densities will have a singularity at 0. Figures 1 and 2 give these limiting densities for $\alpha = 1/2$ and $\alpha = 3/2$ (with $\lambda = 1$). (Note that the limiting joint distribution could also be deduced from the limiting distribution in Example 1.) It is also possible to see that, as $\alpha \to \infty$, the asymptotic distribution concentrates around the points $(0,1)$, $(0,-1)$, $(1,0)$ and $(-1,0)$ with equal probability.

EXAMPLE 3. Suppose that $Y_i = \beta_0+\beta_1 x_i+\varepsilon_i$, where the x_i's are uniformly distributed over the interval $[-1,1]$. In this case, we have

$$\frac{1}{n}X_n^T X_n \to C = \begin{pmatrix} 1 & 0 \\ 0 & 1/3 \end{pmatrix}.$$

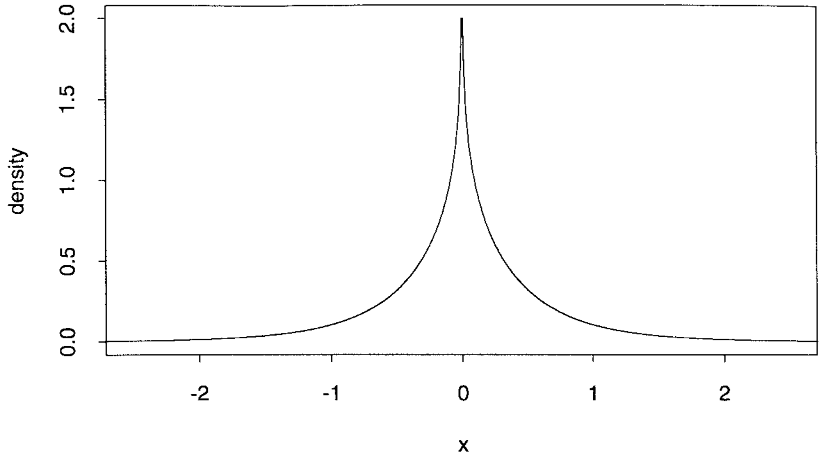

FIG. 1. *Marginal density for $\alpha = 1/2$.*

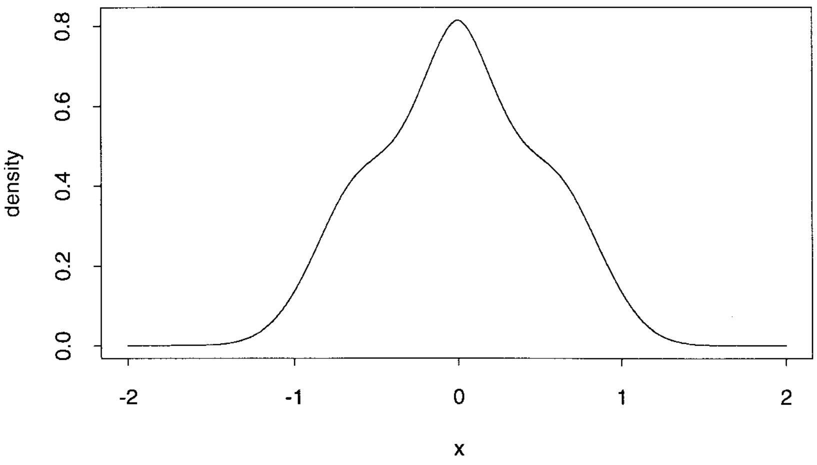

FIG. 2. *Marginal density for* $\alpha = 3/2$.

We will assume that the distribution function F of the ε_i's satisfies

$$\lim_{n\to\infty} \sqrt{n}\left(F(t/\sqrt{n}) - F(0)\right) = \begin{cases} -\infty, & \text{if } t < 0, \\ \lambda t, & \text{if } t \geq 0, \end{cases}$$

and so

$$\Psi(t) = \begin{cases} \infty, & \text{if } t < 0, \\ \lambda t^2/2, & \text{if } t \geq 0. \end{cases}$$

Thus, given that the x_i's are contained in $[-1, 1]$, we have

$$\tau(u_0, u_1) = \begin{cases} \lambda(u_0^2/2 + u_1^2/6), & \text{if } u_0 \geq |u_1|, \\ \infty, & \text{otherwise.} \end{cases}$$

Then letting W_0 and W_1 be independent 0 mean normal random variables with variances 1 and 1/3, respectively, it follows that

$$\sqrt{n}\begin{pmatrix} \widehat{\beta}_{0n} - \beta_0 \\ \widehat{\beta}_{1n} - \beta_1 \end{pmatrix} \to_d \begin{pmatrix} U_0 \\ U_1 \end{pmatrix},$$

where

$$U_0 = \begin{cases} W_0/(2\lambda), & \text{if } W_0 \geq 3|W_1|, \\ 3(W_0 + W_1)/(8\lambda), & \text{if } 0 < W_0 + W_1 < 4W_1, \\ 3(W_0 - W_1)/(8\lambda), & \text{if } 0 < W_0 - W_1 < -4W_1, \\ 0, & \text{if } W_0 \leq -|W_1|, \end{cases}$$

and

$$U_1 = \begin{cases} 3W_1/(2\lambda), & \text{if } W_0 \geq 3|W_1|, \\ 3(W_0 + W_1)/(8\lambda), & \text{if } 0 < W_0 + W_1 < 4W_1, \\ 3(W_1 - W_0)/(8\lambda), & \text{if } 0 < W_0 - W_1 < -4W_1, \\ 0, & \text{if } W_0 \leq -|W_1|. \end{cases}$$

Note that much of the limit distribution is concentrated on the set $B = \{(u_0, u_1)\colon u_0 = |u_1|\}$; a straightforward computation gives

$$\begin{aligned} P\big[(U_0, U_1) \in B\big] &= P\big[W_0 < 3|W_1|\big] \\ &= \frac{1}{\pi\sqrt{3}} \int_0^\infty \int_{-\infty}^x \exp\left(-\frac{1}{6}(3t^2 + x^2)\right) dt\, dx \\ &= \frac{5}{6}. \end{aligned}$$

Likewise, $P(U_0 = U_1 = 0) = P(W_0 \leq -|W_1|) = 1/6$.

EXAMPLE 4. As in Example 3, suppose that $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where the x_i's are uniformly distributed on the interval $[-1, 1]$. Suppose that the density of the ε_i's is

$$f(x) = \frac{1}{2x^2} \exp(-|x|^{-1})$$

so that its distribution function is

$$F(x) = \tfrac{1}{2}\big(1 + \operatorname{sgn}(x) \exp(-|x|^{-1})\big). \tag{7}$$

It is now easy to see that

$$\sqrt{n}\big(F(t/\ln(n)) - F(0)\big) \to \psi(t) = \begin{cases} 0, & \text{for } |t| < 2, \\ \operatorname{sgn}(t)1/2, & \text{for } |t| = 2, \\ \operatorname{sgn}(t)\infty, & \text{for } |t| > 2, \end{cases}$$

so that

$$\Psi_n(t) \to \Psi(t) = \begin{cases} 0, & \text{for } |t| \leq 2, \\ \infty, & \text{for } t > 2. \end{cases}$$

Defining Z_n as in (5) with $a_n = \ln(n)$, it is easy to determine that

$$Z_n(u_0, u_1) \to_d Z(u_0, u_1) = u_0 W_0 + u_1 W_1 + \tau(u_0, u_1),$$

where W_0, W_1 are independent normal random variables with mean 0 (variances 1 and 1/3, respectively) and

$$\tau(u_0, u_1) = \begin{cases} 0, & \text{if } -2 \leq u_0 + u_1 \leq 2 \text{ and } -2 \leq u_0 - u_1 \leq 2, \\ \infty, & \text{otherwise.} \end{cases}$$

[We have

$$\frac{1}{n}\sum_{i=1}^{n}\Psi_n(\mathbf{u}^T\mathbf{x}_i) \to \tau(\mathbf{u})$$

since the $\mathbf{x}_i$'s are contained in a compact set.] Thus determining the limiting distribution of $\ln(n)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ depends on minimizing $u_0W_0 + u_1W_1$ over the region

$$A = \{(u_0, u_1)\colon -2 \le u_0 + u_1 \le 2 \text{ and } -2 \le u_0 - u_1 \le 2\}.$$

Since $P(W_0 = W_1) = P(W_0 = -W_1) = 0$, $Z(u_0, u_1)$ has an almost sure unique minimum; the minimizer will be one of the four corners of the region A with probability 1/4 for each corner. [Likewise, if $Y_i = \mu + \varepsilon_i$, where the ε_i's have the distribution function (7), then it is easy to show that for the sample median $\widehat{\mu}_n$ of $Y_1, \dots, Y_n$ we have

$$\ln(n)(\widehat{\mu}_n - \mu) \to_d U,$$

where $P(U = 2) = P(U = -2) = 1/2$; this is one of the four types of limiting distribution for the sample median given by Smirnov (1952).]

The convergence to the limiting distribution is very slow (as might be expected). Figures 3 and 4 show plots of $\ln(n)(\widehat{\beta}_{n0} - \beta_0)$ versus $\ln(n)(\widehat{\beta}_{n1} - \beta_1)$ for $n = 100$ and $n = 100{,}000$ based on 1000 simulations; note the tendency for the points to concentrate around the corners of A. However, even for the extremely large sample size, there is little evidence that the limiting distribution would be a good approximation to the true distribution. (The simulations were done using S-PLUS with the pseudorandom variates generated using an inverse transformation of uniform pseudorandom variates.)

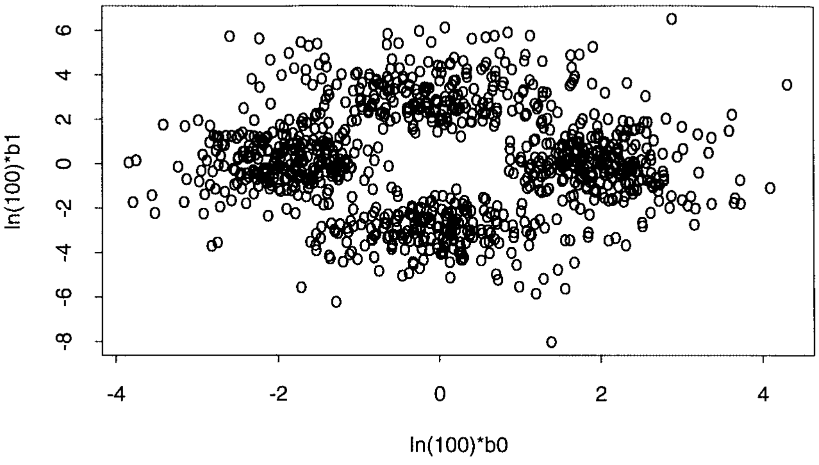

FIG. 3. *Scatterplot of* $\ln(n)(\hat{\beta}_{n1} - \beta_1)$ *versus* $\ln(n)(\hat{\beta}_{n0} - \beta_0)$ *for* $n = 100$.

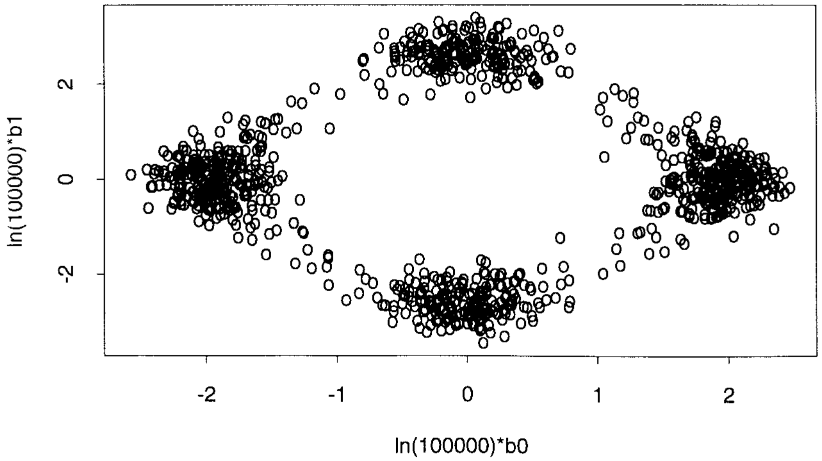

FIG. 4. *Scatterplot of* $\ln(n)(\hat{\beta}_{n1} - \beta_1)$ *versus* $\ln(n)(\hat{\beta}_{n0} - \beta_0)$ *for* $n = 100000$.

More generally, when $Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$ for i.i.d. ε_i's with distribution function (7) then

$$Z_n(\mathbf{u}) \to_d Z(\mathbf{u}) = \mathbf{u}^T \mathbf{W} + \tau(\mathbf{u}),$$

where $\mathbf{W}$ is $(p+1)$-variate normal with mean $\mathbf{0}$ and covariance matrix

$$C = \lim_{n\to\infty} \frac{1}{n} X_n^T X_n$$

and

$$\tau(\mathbf{u}) = \begin{cases} 0, & \text{if } \mathbf{u} \in A, \\ \infty, & \text{otherwise,} \end{cases}$$

with

$$A = \left\{\mathbf{u}\colon \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I\big(|\mathbf{u}^T \mathbf{x}_i| \le 2\big) = 1\right\}.$$

Note that A is always nonempty in that it will always contain $\mathbf{0}$. Moreover, in order to obtain a nondegenerate limit distribution, the $\mathbf{x}_i$'s need to be "essentially" bounded in the sense that a negligible fraction lies outside a bounded set.

4. Other comments.

4.1. *Efficiency.* It is interesting to compare the rate of convergence of L_1-estimators to the best attainable rate of convergence. In this section, we will show that L_1-estimators are not "ratewise robust" when $\alpha \neq 1$ when α is defined as in (6).

When $a_n = o(\sqrt{n})$ then L_1-estimators have an asymptotic relative efficiency of 0 since $\sqrt{n}$-consistent estimators of $\boldsymbol{\beta}$ exist. When $F'(0) = \lambda > 0$, the L_1-estimators are $\sqrt{n}$-consistent; this is typically the optimal rate of convergence in this case. However, the situation is less clear when $a_n/\sqrt{n} \to \infty$. Intuitively, it seems possible to improve on the rate of convergence obtained by L_1-estimators by choosing an estimation method which more fully exploits the concentration of ε_i's near 0.

In order to find the best possible rate of convergence, we will apply the theory developed by Akahira and Takeuchi (1981, 1995). Assume that ε_i has a symmetric density

$$f(x) = |x|^{\alpha-1}L(|x|),$$

where $\alpha < 1$ and L is a slowly varying function at 0. In this case, it follows that

$$F(x) - F(0) = \operatorname{sgn}(x)|x|^{\alpha}L^*(|x|),$$

where $L^*(|x|)/L(|x|) \to \alpha$ as $x \to 0$. Define $\{b_n\}$ so that

$$n\big(F(x/b_n) - F(0)\big) \to \operatorname{sgn}(x)|x|^{\alpha}.$$

(Note that $b_n = a_{n^2}$ and so $b_n/a_n \to \infty$.) Then it follows that

$$\frac{n}{b_n} f\left(\frac{x}{b_n}\right) \to \alpha|x|^{\alpha-1}.$$

Define

$$K_n(\theta) = -\int_{-\infty}^{\infty} [\ln f(x - \theta/b_n) - \ln f(x)]\, f(x)\, dx.$$

It follows that

$$\begin{aligned} K_n(\theta) &\sim \frac{\alpha(1-\alpha)}{n} \int_{-\infty}^{\infty} [\ln(|x-\theta|) - \ln(|x|)]\, |x|^{\alpha-1}\, dx \\ &= \frac{\alpha(1-\alpha)}{n} |\theta|^{\alpha} \int_{-\infty}^{\infty} [\ln(|x-1|) - \ln(|x|)]\, |x|^{\alpha-1}\, dx. \end{aligned}$$

Then provided that

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} |\mathbf{u}^T \mathbf{x}_i|^{\alpha} < \infty,$$

it follows from Theorem 3.5.2 of Akahira and Takeuchi (1995) that there cannot exist a sequence of "regular" estimators $\{\widehat{\boldsymbol{\beta}}_n\}$ of $\boldsymbol{\beta}$ with $c_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = O_p(1)$ and $c_n/b_n \to \infty$.

We will now give a heuristic demonstration of the existence of a b_n-consistent estimator of $\boldsymbol{\beta}$. Consider estimating $\boldsymbol{\beta}$ by minimizing

$$\sum_{i=1}^{n} |Y_i - \boldsymbol{\phi}^T \mathbf{x}_i|^r$$

over all $\boldsymbol{\phi}$, where $r > 0$ and $r + \alpha < 1$. Define

$$Z_n(\mathbf{u}) = b_n^r \sum_{i=1}^n \left[|\varepsilon_i - \mathbf{u}^T\mathbf{x}_i/b_n|^r - |\varepsilon_i|^r\right]$$
$$= \int_{-\infty}^{\infty} \left[|v - \mathbf{u}^T\mathbf{x}|^r - |v|^r\right] \nu_n(dv \times d\mathbf{x}),$$

where ν_n is a random measure with

$$\nu_n(A \times B) = \sum_{i=1}^n I(b_n\varepsilon_i \in A, \mathbf{x}_i \in B).$$

Under appropriate regularity conditions, $\{\nu_n\}$ converges in distribution (with respect to the vague topology) [Kallenberg (1983)] to a Poisson random measure ν with

$$E\big(\nu(A \times B)\big) = \lambda(B) \int_A \alpha x^{\alpha-1}\, dx,$$

where

$$\frac{1}{n}\sum_{i=1}^n I(\mathbf{x}_i \in B) \to \lambda(B).$$

Using standard methods, it can be shown that

$$Z_n(\mathbf{u}) \to_d Z(\mathbf{u}) = \int_{-\infty}^{\infty} \left[|v - \mathbf{u}^T\mathbf{x}|^r - |v|^r\right] \nu(dv \times d\mathbf{x}),$$

where $Z(\mathbf{u})$ is finite if

$$\int |\mathbf{u}^T\mathbf{x}|^{r+\alpha}\lambda(d\mathbf{x}) < \infty.$$

The function Z_n is not convex so the b_n-consistency is not immediate, but this argument can be tightened up (albeit with some difficulty) to yield b_n-consistency of the estimator. Note that this result does not depend on the existence of a density. Some related results are given in Ibragimov and Has'minskii (1981).

4.2. *Second-order results.* The proof of Theorem 1 implies that we can approximate the function Z_n by

$$Z_n^*(\mathbf{u}) = -\mathbf{u}^T\mathbf{W}_n + 2\tau(\mathbf{u}),$$

where

$$\mathbf{W}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{x}_i\big[I(\varepsilon_i > 0) - I(\varepsilon_i < 0)\big].$$

The convexity of Z_n and Z_n^* implies that we can approximate $a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \operatorname{argmin}(Z_n)$ by $\operatorname{argmin}(Z_n^*)$.

The asymptotic behavior of $\operatorname{argmin}(Z_n) - \operatorname{argmin}(Z_n^*)$ follows from the asymptotic behavior of $Z_n - Z_n^*$. In particular, it can be shown (under appropriate regularity conditions including twice-differentiability of τ) that

$$n^{1/4}(Z_n - Z_n^*) \to_d V$$

(where V is a differentiable Gaussian process defined on R^{p+1}), from which it follows that

$$n^{1/4}(\operatorname{argmin}(Z_n) - \operatorname{argmin}(Z_n^*)) \to_d -H^{-1}(\mathbf{U})\mathbf{D}(\mathbf{U}),$$

where H is the Hessian matrix of τ, $\mathbf{D}$ is the gradient of V and $\mathbf{U}$ is the limit of $\operatorname{argmin}(Z_n) = a_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$; it can be shown that $\mathbf{U}$ is independent of $\mathbf{D}$. Details are given in Knight (1997). Under the classical assumption that $F'(0) = \lambda > 0$ such "weak" Bahadur–Kiefer representations [Bahadur (1966), Kiefer (1967)] are given by Arcones (1996a, b). "Strong" Bahadur–Kiefer representations in the classical case are given by He and Shao (1996) as well as Arcones (1996b).

Acknowledgments. The author would like to thank the referees and the co-Editors for some very helpful comments on the original version of this paper as well as Miguel Arcones for sending me preprints of some of his papers.

REFERENCES

Akahira, M. and Takeuchi, K. (1981). *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency.* Springer, New York.

Akahira, M. and Takeuchi, K. (1995). *Non-Regular Statistical Estimation.* Springer, New York.

Arcones, M. A. (1994). Distributional convergence of M-estimators under unusual rates. *Statist. Probab. Lett.* **21** 271–280.

Arcones, M. A. (1996a). The Bahadur–Kiefer representation of L_p regression estimators. *Econometric Theory* **12** 257–283.

Arcones, M. A. (1996b). Second order representations of the least absolute deviation regression estimator. Unpublished manuscript.

Babu, G. J. (1989). Strong representation for LAD estimators in linear models. *Probab. Theory Related Fields* **83** 547–558.

Bahadur, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37** 577–580.

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 Sciences A* **33** 449–463.

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

Bloomfield, P. and Steiger, W. L. (1983). *Least Absolute Deviations: Theory, Applications and Algorithms.* Birkhäuser, Boston.

Geyer, C. J. (1996). On the asymptotics of convex stochastic optimization. Unpublished manuscript.

He, X. and Shao, Q.-M. (1996). General Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24** 2608–2630.

Hjørt, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. Statistical Research Report, Univ. Oslo.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer, New York.

Jurečková, J. (1983). Asymptotic behavior of M-estimators of location in nonregular cases. *Statist. Decisions* **1** 323–340.

Kallenberg, O. (1983). *Random Measures*, 3rd ed. Akademie, Berlin.

KIEFER, J. (1967). On Bahadur's representation of sample quantiles. *Ann. Math. Statist.* **38** 1323–1342.

KNIGHT, K. (1997). Asymptotics for L_1 regression estimators under general conditions. Technical Report 9716, Dept. Statistics, Univ. Toronto.

KOENKER, R. and BASSETT, G. (1978). Regression quantiles. *Econometrica* **46** 33–50.

POLLARD, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7** 186–199.

SMIRNOV, N. V. (1952). Limit distributions for the terms of a variational series. *Amer. Math. Soc. Transl. Ser.* (*1*) **11** 82–143.

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
100 ST. GEORGE STREET
TORONTO, ONTARIO
M5S 3G3 CANADA
E-MAIL: keith@utstat.toronto.edu