Econometric Theory, 7, 1991, 186–199. Printed in the United States of America.

ASYMPTOTICS FOR LEAST ABSOLUTE DEVIATION REGRESSION ESTIMATORS

DAVID POLLARD
Yale University

The LAD estimator of the vector parameter in a linear regression is defined by minimizing the sum of the absolute values of the residuals. This paper provides a direct proof of asymptotic normality for the LAD estimator. The main theorem assumes deterministic carriers. The extension to random carriers includes the case of autoregressions whose error terms have finite second moments. For a first-order autoregression with Cauchy errors the LAD estimator is shown to converge at a $1/n$ rate.

1. THE PROBLEM

Suppose random variables $y_1, y_2, \ldots$ are generated by a linear regression, $y_i = x_i'\beta_0 + u_i$, for observed $\{x_i\}$, unknown β_0 in $\mathbb{R}^d$, and unknown errors $\{u_i\}$. The least absolute deviations (LAD) estimator $\hat{\beta}_n$ is chosen to minimize the random criterion function

$$\sum_{i \le n} |y_i - x_i'\beta|.$$

In view of its ancient lineage, it is surprising that the asymptotic theory of LAD estimation has only recently been developed.

Bassett and Koenker [5] established a central limit theorem for $\sqrt{n}(\hat{\beta}_n - \beta_0)$, assuming the $\{u_i\}$ to be independent and identically distributed (i.i.d.) random variables and $\{x_i\}$ to be a deterministic sequence for which

$$\frac{1}{n}\sum_{i \le n} x_i x_i' \to Q,$$

with Q a positive definite matrix. They checked pointwise convergence of the density functions. Bloomfield and Steiger ([7], pp. 44–49), using a smoothing technique similar to that of Amemiya [1], extended the central limit theorem to cover stationary, ergodic, martingale differences $\{x_i, y_i\}$. The smoothing allowed them to locate a minimum by equating partial derivatives of zero. Ruppert and Carroll [20] proved central limit theorems for various estima-

This research was partially supported by NSF grants no. DMS-8503347 and DMS-8806900. I am grateful to Gib Bassett, Peter Bloomfield, Jana Jurečková, Keith Knight, Peter Phillips, and two referees for valuable advice and correspondence.

 0266-4666/91 $5.00 + .00

tors related to LAD, relying on a stochastic equicontinuity result of Bickel [6] to develop uniform approximations to a subgradient vector, then applying an argument due to Jurečková [10]. Van de Geer [22] applied empirical process methods to the case of i.i.d. $\{(x_i, y_i)\}$ in order to establish the uniform bounds needed to deduce asymptotic normality directly from the minimizing property of $\hat{\beta}_n$. Sanz [21] also applied empirical process theory to establish an unusual rate of convergence for the LAD estimator generated by a particular long-tailed error distribution. More recently, Knight [11,12] and Davis, Knight, and Liu [8] have applied methods similar to the ones to be presented in this paper. (Their work has developed independently from mine.)

Most of these proofs were built around some sort of stochastic equicontinuity argument. That is, they required uniform smallness for the changes in some sequence of stochastic processes $\{X_n\}$ due to small perturbations of the parameters; they required something like

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}\{ \sup_{|s|<\delta} |X_n(s)| > \epsilon \} = 0 \qquad \text{for each } \epsilon > 0.$$

Moreover, they usually involved some sort of preliminary consistency argument. In the Jurečková paper [10] monotonicity properties of the criterion function greatly simplified this task; the behavior of a process outside a compact region was controlled by the behavior on the boundary of that region.

As I have argued many times (for example, in Pollard [14–18] and in Pakes and Pollard [13]), stochastic equicontinuity often does capture precisely the key technical difficulty in an asymptotic proof. Unfortunately it also tends to make the arguments less accessible to many potential users. By contrast, in this paper I introduce a simpler technique, which depends crucially on the convexity property of the criterion function, in order to derive the necessary uniformity of approximation.

CONVEXITY LEMMA. *Let $\{\lambda_n(\theta) : \theta \in \Theta\}$ be a sequence of random convex functions defined on a convex, open subset Θ of $\mathbb{R}^d$. Suppose $\lambda(\cdot)$ is a real-valued function on Θ for which $\lambda_n(\theta) \to \lambda(\theta)$ in probability, for each θ in Θ. Then for each compact subset K of Θ,*

$$\sup_{\theta \in K} |\lambda_n(\theta) - \lambda(\theta)| \to 0 \qquad \text{in probability.}$$

The function $\lambda(\cdot)$ is necessarily convex on Θ. ■

Andersen and Gill [4] have already proved this result, reducing its proof by means of a subsequencing argument to an application of its well-known nonstochastic analogue (Rockafellar [19], Theorem 10.8). But, for completeness' sake, a simple direct proof will be given in Section 6. The lemma will allow us to derive the limit distribution directly, without preliminary consistency arguments, using a technique analogous to the method of proof for

Lemma 5.2 of Jurečková [10]. The lemma could also be adapted to simplify the asymptotic theory for other estimators defined by minimization of a convex criterion function, but I will not pursue that in the present paper. (See Section 14 of Pollard [18] for one possibility. Heiler and Willers [9] have also applied an analogue of the CONVEXITY LEMMA to establish central limit theorems for R-estimators.)

Throughout the paper I use linear functional notation, writing $\mathbb{P}Z$ instead of the more traditional $\mathbb{E}Z$ for the expectation of a random variable Z. Also I identify sets with their indicator functions. For example, $\mathbb{P}u_i\{n\delta \le u_i \le n\}$ will replace what might be written $\mathbb{E}u_i\chi\{n\delta \le u_i \le n\}$, with $\chi\{\cdots\}$ denoting an indicator function. The symbol $\rightsquigarrow$ will denote convergence in distribution.

Section 2 contains three limit theorems for LAD estimators, corresponding to increasingly complex behavior of the $\{x_i\}$. For the first theorem, whose proof occupies Section 3, they are assumed deterministic. I believe the analysis there comes as close to being elementary as any central limit theorem for an LAD estimator can. The second theorem, whose proof occupies Section 4, allows random $\{x_i\}$; its proof involves only small martingale-type modifications to the first proof. It covers the case of simple autoregressions with errors $\{u_i\}$ having finite second moments. For longer-tailed error distributions a much more delicate argument is needed. As an illustration the third theorem, whose proof occupies Section 5, deals with a first-order autoregression with Cauchy distributed errors. This special case has already received some attention in the literature (An and Chen [3], Bloomfield and Steiger [7]). It is notable for the rapid $O_p(1/n)$ rate of convergence of its LAD estimator.

Remark on History

I first presented the convexity method in lectures at the University of Iowa in July 1988. In the published notes for these lectures (Pollard [18]) I replaced the analysis of the LAD example by a more complicated analysis, which depends on stochastic equicontinuity arguments, for the more complicated problem of LAD fitting to a censored regression. After reading the August 1988 version of the present paper, Peter Bloomfield asked about possible extensions to autoregressions. Peter Phillips then helped me understand the unusual asymptotics for the case with Cauchy errors; I worked out the details for a Yale graduate course in the spring of 1989. A referee pointed out the relevance of the papers by Ruppert and Carroll [20] and Jurečková [10]. While preparing the revised manuscript I received a preprint by Knight [11], which presented similar convexity arguments to mine for the case of deterministic $\{x_i\}$, together with extensions to other problems involving convex criterion functions. Knight [12] and Davis, Knight, and Liu [8] have devel-

oped the arguments further. Clearly the convexity argument is an idea whose time has come.

2. THREE LIMIT THEOREMS

Throughout the section we make the following assumption about the $\{u_i\}$.

ERROR ASSUMPTION. *The regression errors $\{u_i\}$ are independent. They are identically distributed, with median 0 and a continuous, positive density $f(\cdot)$ in a neighborhood of 0.*

The assumption ensures that the function

$$M(t) = \mathbb{P}(|u_i - t| - |u_i|)$$

has a unique minimum at zero, and

$$M(t) = t^2 f(0) + o(t^2) \qquad \text{near zero.} \tag{1}$$

Also, if we define $U(t)$ as $\mathbb{P}\{|u_i| \le t\}$ then $U(t) \to 0$ as $t \to 0$. As will be pointed out later, the assumption could be weakened slightly, but that might distract attention from the more interesting problems connected with the behavior of the $\{x_i\}$ sequence.

THEOREM 1. *Suppose the $\{u_i\}$ satisfy the* ERROR ASSUMPTION *and that $\{x_i\}$ is a deterministic sequence for which the matrix $\sum_{i\le n} x_i x_i'$ eventually has a positive definite square root V_n. If $\max_{i\le n} |V_n^{-1} x_i| \to 0$ as $n \to \infty$, then $2f(0)V_n(\hat{\beta}_n - \beta_0) \rightsquigarrow N(0, I_d)$.* ∎

The conditions on the $\{x_i\}$ are satisfied if there exists a positive definite matrix V for which

$$\frac{1}{n}\sum_{i\le n} x_i x_i' \to V^2.$$

For if such a V exists, then $V_n/\sqrt{n} \to V$ and $|x_n|^2/n \to 0$. Consequently,

$$\max_{i\le n} |V_n^{-1} x_i| = O\Big(\max_{i\le n} |x_i|/\sqrt{n}\Big) \to 0.$$

Similar behavior can be expected of random $\{x_i\}$ under mild conditions. For example, if $\{x_i\}$ were a stationary, ergodic sequence with $\mathbb{P}x_i x_i'$ finite and nonsingular, then almost all realizations would have these limiting properties.

THEOREM 2. *Suppose the $\{u_i\}$ satisfy the* ERROR ASSUMPTION. *Let $\{\mathcal{T}_i\}$ be an increasing sequence of σ-fields and V_n be a sequence of (possibly random) positive definite matrices such that:*

(i) u_i *is independent of* $\mathcal{F}_{i-1}$ *for every* i;
(ii) $V_n^{-1}x_i$ *is* $\mathcal{F}_{i-1}$ *measurable for every* i;
(iii) $\max_{i\le n}|V_n^{-1}x_i| \to 0$ *in probability*;
(iv) $\sum_{i\le n} V_n^{-1}x_i x_i' V_n^{-1} \to I_d$ *in probability.*

Then $2f(0)V_n(\hat{\beta}_n - \beta_0) \rightsquigarrow N(0, I_d)$. ■

The conditions imposed on the $\{x_i\}$ anticipate two slightly different applications.

Example 1

Suppose that $\{x_i\}$ are random vectors independent of the $\{u_i\}$. Let each $\mathcal{F}_i$ be the same, equal to the σ-field generated by all the $\{x_i\}$. Suppose $\max_{i\le n}|x_i| = o_p(\sqrt{n})$ and also that there exists a positive definite, deterministic matrix V for which

$$\frac{1}{n}\sum_{i\le n} x_i x_i' \to V^2 \qquad \text{in probability,}$$

or even just that $[(1/n)\sum_{i\le n} x_i x_i']^{-1}$ is of order $O_p(1)$. Then the theorem applies with either V_n equal to $V\sqrt{n}$ or the positive definite square root of the matrix $\sum_{i\le n} x_i x_i'$. ■

Example 2

Suppose $x_i = y_{i-1}$. That is, the $\{y_i\}$ are a first-order autoregressive sequence. (Extension to higher-order autoregressions is not hard.) Suppose $|\beta_0| < 1$ and the u_i have zero means and finite variance $\sigma^2 > 0$. Let $\mathcal{F}_i$ be the σ-field generated by the random variables $y_0, u_1, \ldots, u_i$. Let us check the conditions of Theorem 2.

It is a standard time series result (see Example VIII.4 of Pollard [15]) that

$$\frac{1}{n}\sum_{i\le n} y_{i-1}^2 \to \frac{\sigma^2}{1-\beta_0^2} \qquad \text{in probability.}$$

Write τ^2 for the limit. Take V_n equal to $\tau\sqrt{n}$.

For the condition regarding the maximum, first use finiteness of second moments to control the errors:

$$\begin{aligned}\mathbb{P}\{\max_{i\le n}|u_i| > \epsilon\sqrt{n}\} &\le \frac{1}{n\epsilon^2}\sum_{i\le n}\mathbb{P}u_i^2\{|u_i| > \epsilon\sqrt{n}\}\\ &= \frac{1}{\epsilon^2}\mathbb{P}u_1^2\{|u_1| > \epsilon\sqrt{n}\}\\ &\to 0 \qquad \text{by dominated convergence.}\end{aligned}$$

Then observe that

$$\max_{i\le n}|y_i| \le O_p(|y_0| + \max_{i\le n}|u_i|) = o_p(\sqrt{n}).$$

The standardized LAD estimator $\sqrt{n}(\hat{\beta}_n - \beta_0)$ has an asymptotic $N(0,(2\tau f(0))^{-2})$ distribution. ■

The case of an autoregression whose errors have infinite second moment is much more interesting. An and Chen [3] analyzed the prototypical case of Cauchy errors. They showed that for each $\delta > 0$, the LAD estimator $\hat{\beta}_n$ lies within $O_p(n^{-1+\delta})$ of the true β_0. A slightly better result is possible.

THEOREM 3. *Suppose* $y_i = \beta_0 y_{i-1} + u_i$, *where* $|\beta_0| < 1$ *and the* $\{u_i\}$ *are independent Cauchy errors. Then* $\hat{\beta}_n = \beta_0 + O_p(n^{-1})$. ■

It is worthwhile to determine the limiting distribution for $n(\hat{\beta}_n - \beta_0)$. Knight [12] has applied the same method to the case $\beta_0 = 1$ for various long-tailed error distributions, establishing existence of a limit distribution expressible as a stochastic integral. More interestingly, for the same case he has also established existence of a limiting normal distribution under random norming. In a recent preprint, Davis, Knight, and Liu [8] have solved the general problem.

3. PROOF OF THEOREM 1

Write $z_{i,n}$ for $V_n^{-1}x_i$. By definition, $\sum_{i\le n} z_{i,n}z_{i,n}' = I_d$ and $\max_{i\le n}|z_{i,n}| \to 0$. Also

$$\sum_{i\le n}|z_{i,n}|^2 = \text{trace}\sum_{i\le n} z_{i,n}z_{i,n}' = d.$$

For θ in $\mathbb{R}^d$, define

$$G_n(\theta) = \sum_{i\le n}(|u_i - z_{i,n}'\theta| - |u_i|).$$

This is a convex function of θ that is minimized by

$$\hat{\theta}_n = V_n(\hat{\beta}_n - \beta_0).$$

The idea behind the proof is to approximate G_n by a quadratic function whose minimizing value has an asymptotic normal distribution, and then to show that $\hat{\theta}_n$ lies close enough to that minimizing value to share its asymptotic behaviour.

Two terms contribute to the approximation. One is a deterministic quadratic function obtained via a Taylor expansion of the expected value $\Gamma_n(\theta) = \mathbb{P}G_n(\theta)$ using (1):

$$\Gamma_n(\theta) = \sum_{i\le n} M(z_{i,n}'\theta) = f(0)|\theta|^2 + o(1).$$

The other term is random and linear in θ. It comes from a sort of Taylor expansion of $G_n(\theta)$ around $\theta = 0$. The usual style of argument—a pointwise expansion of each summand to quadratic terms, followed by appeals to standard limit theory for the sums of coefficients—fails, because $|u_i - t|$ is not everywhere differentiable. Amemiya ([2], Section 4.6) has explained the difficulty. (Peter Phillips has developed an interesting heuristic method using generalized functions, which suggests another way around this difficulty.) However we do benefit from a linear approximation to $|u_i - t|$ obtained by treating the difference of indicator functions,

$$D_i = \{u_i < 0\} - \{u_i \geq 0\},$$

as if it were a first derivative at $t = 0$. Notice that $\mathbb{P}D_i = 0$ because u_i has a zero median. Define

$$R_{i,n}(\theta) = |u_i - z_{i,n}'\theta| - |u_i| - D_i z_{i,n}'\theta$$

and

$$W_n = \sum_{i \leq n} D_i z_{i,n}.$$

Then

$$G_n(\theta) = \Gamma_n(\theta) + W_n'\theta + \sum_{i \leq n} (R_{i,n}(\theta) - \mathbb{P}R_{i,n}(\theta)).$$

The properties of the $\{z_{i,n}\}$ and the multivariate central limit theorem ensure that W_n has an asymptotic $N(0, I_d)$ distribution. It will turn out that $\hat{\theta}_n$ lies close to $-\frac{1}{2} W_n / f(0)$.

For fixed θ, the sum of centered terms $\xi_{i,n} = R_{i,n}(\theta) - \mathbb{P}R_{i,n}(\theta)$ will contribute only a $o_p(1)$ to $G_n(\theta)$. It is easy to show this by means of a second moment bound based on the inequality

$$|R_{i,n}(\theta)| \leq 2|z_{i,n}\theta|\{|u_i| \leq |z_{i,n}\theta|\}.$$

Because of cancellation of cross-product terms, we get

$$\begin{aligned}
\mathbb{P}\left|\sum_{i \leq n} \xi_{i,n}\right|^2 &\leq \sum_{i \leq n} \mathbb{P}R_{i,n}(\theta)^2 \\
&\leq 4 \sum_{i \leq n} |z_{i,n}'\theta|^2 U(|z_{i,n}'\theta|) \qquad \text{where} \quad U(t) = \mathbb{P}\{|u_i| \leq t\} \\
&\leq 4|\theta|^2 U\Big(|\theta| \max_{i \leq n} |z_{i,n}|\Big) \sum_{i \leq n} |z_{i,n}|^2 \\
&\to 0.
\end{aligned}$$

Thus, for each fixed θ,

$$G_n(\theta) = f(0)|\theta|^2 + o(1) + W_n'\theta + o_p(1).$$

The CONVEXITY LEMMA from Section 1, applied to $\lambda_n(\theta) = G_n(\theta) - W_n'\theta$, strengthens the pointwise result to uniform convergence on compact subsets of $\mathbb{R}^d$. With $\eta_n = -\frac{1}{2} W_n/f(0)$, we may write the resulting convergence assertion in the suggestive form

$$G_n(\theta) = f(0)|\theta - \eta_n|^2 - f(0)|\eta_n|^2 + r_n(\theta),$$

where, for each compact set K in $\mathbb{R}^d$,

$$\sup_{\theta \in K} |r_n(\theta)| \to 0 \qquad \text{in probability.}$$

The argument will be complete if we can show for each $\delta > 0$ that

$$\mathbb{P}\{|\hat{\theta}_n - \eta_n| > \delta\} \to 0.$$

This convergence will be a consequence of the convexity of G_n and the behaviour of r_n on the closed ball $B(n)$ with center η_n and radius δ. (The argument is similar to the proof of Jurečková's [10], Lemma 5.2.) Because η_n converges in distribution, it is stochastically bounded. The compact set K can be chosen to contain $B(n)$ with probability arbitrarily close to one, thereby implying that

$$\Delta_n = \sup_{\theta \in B(n)} |r_n(\theta)| \to 0 \qquad \text{in probability.}$$

Now consider the behavior of G_n outside $B(n)$. Suppose $\theta = \eta_n + \beta v$, with $\beta > \delta$ and v a unit vector. Define θ^* as the boundary point of $B(n)$ that lies on the line segment from η_n to θ, that is, $\theta^* = \eta_n + \delta v$. Convexity of G_n and the definition of Δ_n imply

$$\begin{aligned} \frac{\delta}{\beta} G_n(\theta) + \left(1 - \frac{\delta}{\beta}\right) G_n(\eta_n) &\geq G_n(\theta^*) \\ &\geq f(0)\delta^2 - f(0)|\eta_n|^2 - \Delta_n \\ &\geq f(0)\delta^2 + G_n(\eta_n) - 2\Delta_n. \end{aligned}$$

The last expression does not depend on θ. It follows that

$$\inf_{|\theta - \eta_n| > \delta} G_n(\theta) \geq G_n(\eta_n) + \frac{\beta}{\delta} [f(0)\delta^2 - 2\Delta_n].$$

When $2\Delta_n < f(0)\delta^2$, which happens with probability tending to one, the minimum of G_n cannot occur at any θ with $|\theta - \eta_n| > \delta$; with probability tending to one, $|\hat{\theta}_n - \eta_n| \leq \delta$, as required. ■

The reader will observe that independence of the $\{u_i\}$ was required only to ensure asymptotic normality of W_n. Identical distribution of the $\{u_i\}$ gave the limiting quadratic form for the sum of expectations,

$$\sum_{i<n} \mathbb{P}(|u_i - z_{i,n}'\theta| - |u_i| - D_i z_{i,n}'\theta).$$

The proof would also go through under any assumption on the $\{u_i\}$ that took care of these two points.

4. PROOF OF THEOREM 2

Most of the argument will follow the lines established in the previous section. We need concentrate only on the differences caused by randomness of the $\{x_i\}$.

As before define $z_{i,n} = V_n^{-1}x_i$. This time $\sum_{i\le n} z_{i,n}z_{i,n}' \to I_d$ and $\max_{i\le n}|z_{i,n}| \to 0$ in probability, and

$$\sum_{i\le n}|z_{i,n}|^2 = \text{trace}\sum_{i\le n} z_{i,n}z_{i,n}' = O_p(1).$$

As before $W_n = \sum_{i\le n} D_i z_{i,n}$ has an asymptotic $N(0, I_d)$ distribution, but this time by virtue of a martingale central limit theorem (see Theorem VIII.1 of Pollard [15], for example).

Write $\mathbb{P}_i(\cdot)$ for the conditional expectation operator $\mathbb{P}(\cdot\,|\,\mathcal{F}_i)$. The quadratic part of the approximation to $G_n(\theta)$ is now also random. It comes from the process

$$\begin{aligned}\Gamma_n(\theta) &= \sum_{i\le n}\mathbb{P}_{i-1}(|u_i - z_{i,n}'\theta| - |u_i|)\\ &= \sum_{i\le n} M(z_{i,n}'\theta)\\ &= f(0)|\theta|^2 + o_p(1).\end{aligned}$$

Now $\xi_{i,n}$ stands for the variable $R_{i,n}(\theta) - \mathbb{P}_{i-1}R_{i,n}(\theta)$, which has been centered at zero conditional expectation. Much as before,

$$\sum_{i\le n}\mathbb{P}_{i-1}\xi_{i,n}^2 \le 4\sum_{i\le n}|z_{i,n}|^2|\theta|^2 U(|z_{i,n}'\theta|) \to 0 \qquad \text{in probability.}$$

Denote by $S(n)$ the sum of conditional variances that appears on the left-hand side. A simple martingale argument will show that the sum of martingale differences $\sum_{i\le n}\xi_{i,n}$ converges to 0 in probability. For some sequence of real numbers ϵ_n converging to zero, there exist stopping times τ_n for which $S(\tau_n) \le \epsilon_n$ and $\mathbb{P}\{\tau_n \ne n\} \to 0$. The second property of the stopping times ensures that

$$\mathbb{P}\left\{\sum_{i\le n}\xi_{i,n} \ne \sum_{i\le n}\{i\le\tau_n\}\xi_{i,n}\right\} \to 0.$$

The zero conditional expectations account for the vanishing of cross-product terms, leaving

$$\mathbb{P}\left(\sum_{i\le n}\{i\le\tau_n\}\xi_{i,n}\right)^2 = \mathbb{P}\left(\sum_{i\le n}\{i\le\tau_n\}\mathbb{P}_{i-1}\xi_{i,n}^2\right) \le \epsilon_n \to 0.$$

Once again we have, for each fixed θ,

$$G_n(\theta) = f(0)|\theta|^2 + o(1) + W_n'\theta + o_p(1).$$

The rest of the proof now proceeds as for Theorem 1. ■

The ERROR ASSUMPTION on the $\{u_i\}$ could be weakened. The proof of the theorem would also work for stationary, ergodic $\{(x_i, u_i)\}$, provided W_n had a limiting distribution. That would follow from an assumption that the $\{(x_{i+1}, \{u_i < 0\} - \{u_i \geq 0\})\}$ are martingale differences, as was presumably intended in the last paragraph on page 49 of Bloomfield and Steiger [7].

5. PROOF OF THEOREM 3

The proof will be based on ideas from An and Chen [3], as modified by Bloomfield and Steiger ([7], Chapter 3).

To avoid unimportant details let us assume that $y_0 = 0$. Of course that makes $\{y_i\}$ nonstationary. Nevertheless, we will still be able to appeal to the ergodic theorem in the course of the proof.

Anticipating the $1/n$ rate of convergence, let us define

$$G_n(\theta) = \sum_{i \leq n} \left(\left| u_i - \frac{\theta}{n} y_{i-1} \right| - |u_i| \right).$$

It will suffice if we show for each $\epsilon > 0$ that, for T large enough,

$$\mathbb{P}\{G_n(\pm T) > 0\} > 1 - \epsilon \qquad \text{eventually.} \tag{2}$$

That will force $\hat{\theta}_n = n(\hat{\beta}_n - \beta_0)$ into the interval $[-T, T]$ with probability greater than $1 - \epsilon$ eventually, because $G_n(\hat{\theta}_n) \leq G_n(0) = 0$ and G_n is convex.

With the $1/n$ standardization the remainder function becomes

$$R_{i,n}(\theta) = \left| u_i - \frac{\theta}{n} y_{i-1} \right| - |u_i| - \frac{\theta}{n} D_i y_{i-1}.$$

This time it will be the remainder that makes G_n large.

The proof has two parts. First, as a special case of the argument on pages 342–343 of An and Chen [3], or from the result cited on page 97 of Bloomfield and Steiger [7], we have

$$W_n = \frac{1}{n} \sum_{i \leq n} D_i y_{i-1} = O_p(1). \tag{3}$$

For the other part we show, for each $\epsilon > 0$ and each constant C, that

$$\mathbb{P}\left\{ \sum_{i \leq n} R_{i,n}(\pm T) > C|T| \right\} > 1 - \epsilon \qquad \text{eventually, for } |T| \text{ large enough.} \tag{4}$$

Assertion (2) follows easily from (3) and (4).

In order to apply the ergodic theorem we will need to augment the sequence $\{u_i\}$ by new independent observations $u_0, u_{-1} u_{-2}, \ldots$ from the error distribution. Define for each integer n,

$$A_n = \sum_{i \le n} |\beta_0^{n-i+1} u_i|.$$

The sum now runs over all integers i, both positive and negative, less than n. As in the proof of Lemma 3.1 of Bloomfield and Steiger [7], the series for A_n converges almost surely by virtue of the Borel–Cantelli lemma, because

$$\sum_{i \le n} \mathbb{P}\{|u_i| > \lambda^{|i|}\} \le \sum_{i \le n} \lambda^{-|i|} < \infty \qquad \text{if} \quad \lambda > 1.$$

The $\{A_n\}$ sequence is both stationary and ergodic. In particular, there exists a constant K for which

$$\mathbb{P}\{A_n \le K\} = \mathbb{P}\{A_1 \le K\} > 0.$$

Choose and hold fixed such a K.

Now we can establish (4). For simplicity suppose T is positive, and consider only the behavior at $+T$. If $t > 0$,

$$\begin{aligned} |u_i - t| - |u_i| - tD_i &= (2t - 2u_i)\{0 \le u_i \le t\} \\ &\ge t\{0 \le u_i \le t/2\}. \end{aligned}$$

Let δ be a small positive constant. Apply the inequality with t equal to Ty_{i-1}/n on the set where

$$0 \le u_i \le K,$$

$$A_{i-2} \le K,$$

$$n\delta \le u_{i-1} \le n.$$

Together with the inequality $y_{i-1} \ge u_{i-1} - A_{i-2}$, these constraints imply

$$\frac{T}{n} y_{i-1} \ge \frac{T}{n} (u_{i-1} - K) \ge \frac{T}{2n} u_{i-1} \ge 2u_i$$

whenever $n \ge 2K/\delta$ and $T \ge 4K/\delta$. Consequently, for these n and T,

$$R_{i,n}(T) \ge \frac{T}{2n} u_{i-1}\{0 \le u_i \le K, A_{i-2} \le K, n\delta \le u_{i-1} \le n\}. \tag{5}$$

Write $X_{i,n}$ for the coefficient of T on the right-hand side.

The constraints have been chosen so that $X_{i,n}$ has a large conditional expectation given $\mathcal{T}_{i-2}$ (note the choice of σ-field) if δ is small enough:

$$\mathbb{P}_{i-2} X_{i,n} = \frac{1}{2n} \{A_{i-2} \le K\} \mathbb{P}\{0 \le u_i \le K\} \mathbb{P} u_{i-1}\{n\delta \le u_{i-1} \le n\}.$$

The factor $\mathbb{P}\{0 \le u_i \le K\}$ is a fixed positive constant C_1. The other expectation is large when δ is small:

$$\mathbb{P}u_{i-1}\{n\delta \le u_{i-1} \le n\} = \frac{2}{\pi}\int_{n\delta}^{n} \frac{x}{1+x^2}\,dx \ge \frac{2}{\pi}\log\left(\frac{n}{n\delta}\right) = C_\delta.$$

Thus

$$\sum_{i\le n} \mathbb{P}_{i-2}X_{i,n} \ge \frac{1}{2}C_1C_\delta\frac{1}{n}\sum_{i\le n}\{A_{i-2} \le K\}$$

$$\to \frac{1}{2}C_1C_\delta\mathbb{P}\{A_1 \le K\} \qquad \text{by the ergodic theorem.} \tag{6}$$

For the second moments the choice of δ does not matter:

$$\mathbb{P}X_{i,n}^2 \le \frac{1}{4n^2}\mathbb{P}u_{i-1}^2\{n\delta \le u_{i-1} \le n\} < \frac{1}{n}.$$

This bound lets us keep $\sum_{i\le n} X_{i,n}$ relatively close to the sum of conditional expectations, by means of a second moment calculation in which most of the cross-product terms vanish. Write $Z_{i,n}$ for $X_{i,n} - \mathbb{P}_{i-2}X_{i,n}$. Then

$$\mathbb{P}\left|\sum_{i\le n} Z_{i,n}\right|^2 \le \sum_{i\le n}\mathbb{P}Z_{i,n}^2 + 2\sum_{i\le n}\mathbb{P}Z_{i,n}Z_{i+1,n}$$

$$\le \sum_{i\le n}\left(\mathbb{P}X_{i,n}^2 + 2\sqrt{\mathbb{P}X_{i,n}^2\mathbb{P}X_{i+1,n}^2}\right)$$

$$\le 3. \tag{7}$$

From (5), and (6) for a small enough δ, and (7) we deduce (4) by an application of Tchebychev's inequality. ■

6. PROOF OF THE CONVEXITY LEMMA

The inequalities that establish convexity of $\lambda(\cdot)$ are obtained by a passage to the limit from the corresponding inequalities for the $\lambda_n(\cdot)$.

For the uniformity of the convergence it is enough to consider the case where K is a cube with edges parallel to the coordinate directions $e_1, \ldots, e_d$. Every compact subset of Θ can be covered by finitely many such cubes.

Fix $\epsilon > 0$. Since convexity implies continuity, there is a $\delta > 0$ such that λ varies by less than ϵ over each cube of side 2δ that intersects K. For convenience we may assume that the edge length of K is an integer multiple of δ. Partition K into a union of cubes with side δ, then expand K to a larger cube K^δ by adding an extra layer of these δ-cubes around each face. We may as-

sume that δ is small enough to ensure that K^δ lies within Θ. Write $\mathcal{V}$ for the finite set of all vertices of all the δ-cubes that make up K^δ. The convergence in probability is uniform over $\mathcal{V}$:

$$M_n = \max_{t \in \mathcal{V}} |\lambda_n(t) - \lambda(t)| \to 0 \qquad \text{in probability.}$$

Each θ in K lies within a δ-cube with vertices $\{\theta_i\}$ in $\mathcal{V}$; it can be written as a convex combination $\sum_i \alpha_i \theta_i$ of those vertices. Convexity of λ_n gives

$$\begin{aligned}\lambda_n(\theta) &\le \sum_i \alpha_i \lambda_n(\theta_i) \\ &\le M_n + \max_i |\lambda(\theta_i) - \lambda(\theta)| + \lambda(\theta).\end{aligned}$$

The contribution from the maximum over the $\{\theta_i\}$ is less than ϵ, by construction. Thus

$$\mathbb{P}\{\sup_K \lambda_n(\theta) - \lambda(\theta) > 2\epsilon\} \to 0.$$

The companion lower bound is slightly harder to establish. Each θ in K lies within a δ-cube with a vertex θ_0 in $K \cap \mathcal{V}$:

$$\theta = \theta_0 + \sum_i \delta_i e_i \qquad \text{with} \quad |\delta_i| < \delta.$$

Without loss of generality suppose $0 \le \delta_i < \delta$ for each i. Define θ_i to be the vertex $\theta_0 - \delta e_i$ in $\mathcal{V}$. Then θ_0 can be written as a convex combination of θ and the θ_i:

$$\theta_0 = \frac{\delta}{\delta + \sum_j \delta_j}\theta + \sum_i \frac{\delta_i}{\delta + \sum_j \delta_j}\theta_i.$$

Denote these convex weights by β and $\{\beta_i\}$. Notice that

$$\beta \ge \frac{\delta}{\delta + d\delta} = \frac{1}{1+d}.$$

From convexity of λ_n,

$$\begin{aligned}\beta\lambda_n(\theta) &\ge \lambda_n(\theta_0) - \sum_i \beta_i \lambda_n(\theta_i) \\ &\ge \lambda(\theta_0) - \sum_i \beta_i \lambda(\theta_i) - 2M_n \\ &\ge \lambda(\theta) - \epsilon - \sum_i \beta_i [\lambda(\theta) + \epsilon] - 2M_n \\ &\ge \beta\lambda(\theta) - 2\epsilon - 2M_n.\end{aligned}$$

Thus

$$\mathbb{P}\{\inf_K \lambda_n(\theta) - \lambda(\theta) < -3(d+1)\epsilon\} \to 0.$$

The asserted uniform convergence follows. ∎

NOTE ADDED IN PROOF

Dr. Z.D. Bai has brought to my attention his paper, "Asymptotic normality of minimum L_1 norm estimates in linear models" (with Chen, Wu, and Zhao), where similar central limit theorems are proved by direct calculations with directional derivatives. Bai's article appeared in *Chinese Sciences A* 33 (1990): 449–463.

REFERENCES

1. Amemiya, T. Two stage least absolute deviations estimators. *Econometrica* 50 (1982): 689–711.
2. Amemiya, T. *Advanced Econometrics.* Cambridge, MA: Harvard University Press, 1985.
3. An, H.-Z. & Z.-G. Chen. On convergence of LAD estimates in autoregression with infinite variance. *Journal of Multivariate Analysis* 12 (1982): 335–345.
4. Andersen, P.K. & R. Gill. Cox's regression model for counting processes: a large sample study. *Annals of Statistics* 10 (1982): 1100–1120.
5. Bassett, G. & R. Koenker. Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association* 73 (1978): 618–622.
6. Bickel, P.J. One-step Huber estimates in the linear model. *J. American Statistical Association* 70 (1975): 428–433.
7. Bloomfield, P. & W.L. Steiger. *Least Absolute Deviations: Theory, Applications, and Algorithms.* Boston: Birkhauser, 1983.
8. Davis, R.A., K. Knight & J. Liu. M-estimation for autoregressions with infinite variance. Preprint (1990).
9. Heiler, S. & R. Willers. Asymptotic normality of R-estimates in the linear model. *Statistics* 19 (1988): 173–184.
10. Jurečková, J. Asymptotic relations of M-estimates and R-estimates in linear regression model. *Annals of Statistics* 5 (1977): 464–472.
11. Knight, K. A proof of asymptotic normality of LAD and L-estimates in linear regression. Preprint, University of Toronto (1989a).
12. Knight, K. Limit theory for autoregressive-parameter estimates in an infinite-variance random walk. *Canadian Journal of Statistics* 17 (1989b): 261–278.
13. Pakes, A. & D. Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica* 57 (1989): 1027–1058.
14. Pollard, D. A central limit theorem for k-means clustering. *Annals of Probability* 10 (1982): 919–926.
15. Pollard, D. *Convergence of Stochastic Processes.* New York: Springer-Verlag, 1984.
16. Pollard, D. New ways to prove central limit theorems. *Econometric Theory* 1 (1985): 295–314.
17. Pollard, D. Asymptotics via empirical processes. *Statistical Science* 4 (1989): 341–366.
18. Pollard, D. Empirical processes: theory and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 2.* Hayward, CA: Institute of Mathematical Statistics, 1990.
19. Rockafellar, R.T. *Convex Analysis.* Princeton, NJ: Princeton University Press, 1970.
20. Ruppert, D. & R.J. Carroll. Trimmed least squares estimation in the linear model. *J. American Statistical Association* 75 (1980): 828–838.
21. Sanz, G. n^r-consistency of certain optimal estimators, $0 < r < \frac{1}{2}$. Preprint, Universidad de Zaragoza, 1988.
22. Van de Geer, S. Asymptotic normality of minimum L_1-norm estimators in linear regression. Preprint, University of Bristol, 1988.