Analytical and Bootstrap Approximations to Estimator Distributions in L^1 Regression

D. De Angelis, Peter Hall, and G. A. Young*

Edgeworth and bootstrap approximations to estimator distributions in L^1 regression are described. Analytic approximations based on Edgeworth expansions that mix lattice and nonlattice components and allow for an intercept term in the regression are developed under mild conditions, which do not even require a density for the error distribution. Under stronger assumptions on the error distribution, the Edgeworth expansion assumes a simpler form. Bootstrap approximations are described, and the consistency of the bootstrap in the L^1 regression setting is established. We show how the slow rate $n^{-1/4}$ of convergence in this context of the standard, unsmoothed bootstrap that resamples for the raw residuals may be improved to rate $n^{-2/5}$ by two methods: a smoothed bootstrap approach based on resampling from an appropriate kernel estimator of the error density and a normal approximation that uses a kernel estimator of the error density at a particular point, its median 0. Both of these methods require choice of a smoothing bandwidth, however. Numerical illustrations of the comparative performances of the different estimators in small samples are given, and simple but effective empirical rules for choice of smoothing bandwidth are suggested.

KEY WORDS: Density estimator; Edgeworth expansion; L^1 regression; Normal approximation; Smoothed bootstrap.

1. INTRODUCTION

In the general linear model with independent and identically distributed errors, the L^1 estimator of the regression parameter is widely recognized to enjoy superior robustness properties to the least-squares estimator (see, for example, Bloomfield and Steiger 1983).

Asymptotic theory for the L^1 regression estimator was developed by Bassett and Koenker (1978). In this article we discuss in detail analytic approximations based on Edgeworth expansion of estimator distributions in this setting. We consider also use of the bootstrap in estimation of the sampling distribution of the L^1 estimator. In particular, we establish consistency of the unsmoothed bootstrap and show how a faster convergence rate may be obtained using an appropriate smoothed bootstrap or by a simple normal approximation based on kernel estimation of the error density.

Specifically, we consider the model

$$Y_i = \beta_0^T \mathbf{x}_i + \varepsilon_i, \qquad 1 \le i \le n, \tag{1.1}$$

where $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})^T$ is a vector of unknown parameters, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is a vector of design points, and $\varepsilon_1, \ldots, \varepsilon_n$ are independent, identically distributed random errors from a distribution F with density f and median 0. (Here and throughout the article, β_0 denotes the true value of β.)

Under the L^1 criterion, the estimator $\hat{\beta}$ of β_0 is chosen to minimize

$$L(\beta) = \sum_{i=1}^{n} |Y_i - \beta^T \mathbf{x}_i|.$$

Bassett and Koenker (1978) showed that under appropriate regularity conditions, $\hat{\beta}$ is asymptotically normally distributed with

$$\operatorname{var}\{n^{1/2}(\hat{\beta} - \beta_0)\} = \{2f(0)\}^{-2}\mathbf{V}_0^{-1} + o(1),$$

where $\mathbf{V}_0 = n^{-1}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T$.

Section 2 describes our three main practical approximations to the distribution G of $\hat{\beta}$. These are the ordinary bootstrap approximation $\hat{G}$, a Normal approximation $\tilde{G}$, and a smoothed bootstrap approximation $\check{G}$. The performance of these methods is compared in a Monte Carlo study, which points to the advantages of the Normal approximation and the smoothed bootstrap over the unsmoothed bootstrap. Sections 3 and 4 develop theory that elucidates the numerical results presented in Section 2. In particular, Section 3 describes an Edgeworth expansion of the distribution of $n^{1/2}(\hat{\beta} - \beta_0)$. An important, nonstandard feature of that expansion is that it mixes lattice and nonlattice components, the former arising when we assume an intercept term in the Model (1.1). For related work on Edgeworth expansions for lattice random variables or for one-component lattice, see Yarnold (1972), Babu (1991), and Babu and Singh (1989a,b). Section 4 describes theory for the two bootstrap approximations and relates it to the Normal approximation. We show that the smoothed bootstrap and the Normal approximation both have theoretical convergence rates superior to that of the ordinary bootstrap; the latter is only $n^{-1/4}$. For discussion of Edgeworth expansion methodology with the bootstrap see, for example, Bhattacharya and Qumsiyeh (1989) and Hall (1992). Sections 5 and 6 outline proofs of our main results.

The use of the bootstrap with M estimates in linear models has been considered by Mammen (1989). For a more general discussion of resampling methods in regression analysis, see Wu (1986). Hall (1988) has discussed Edgeworth expansions of the distributions of least-squares estimators in regression and has shown that quite different conclusions should be

* D. De Angelis is Statistician, MRC Biostatistics Unit, Institute of Public Health, Cambridge, CB2 2SR, U.K., and PHLS AIDS Centre at CDSC, London NW9 5EQ, U.K. Peter Hall is Professor, Centre for Mathematics and Its Applications, Australian National University, Canberra ACT 2601, Australia, and CSIRO Division of Mathematics and Statistics, Sydney, Australia. G. A. Young is lecturer, Statistical Laboratory, University of Cambridge, CB2 1SB, U.K. The authors thank two referees and an associate editor for their very helpful comments.

drawn there. The differences arise partly from the fact that variance is easily estimated root-n consistently in the least-squares case; no density estimation is involved. Withers (1987) has described expansions, in n^{-1}, of the cumulants of $\hat{\beta}$. These formulas enable one to provide Edgeworth expansions of the "continuous" part of the distribution of $\hat{\beta}$, but do not express the lattice component.

2. APPROXIMATIONS TO DISTRIBUTION OF $\hat{\beta}$

Assume the model described in Section 1, and let $\hat{\beta}$ have the definition given there. Put $\hat{\varepsilon}_i = Y_i - \hat{\beta}^T \mathbf{x}_i$, $1 \le i \le n$. If we suppose that one component (the first) of $\mathbf{x}_i$ is identically unity, then the median of the $\hat{\varepsilon}_i$'s, as defined by minimizing $\sum |\hat{\varepsilon}_i - \text{med.}|$, equals 0. Let $\varepsilon_1^*, \ldots, \varepsilon_n^*$ denote values drawn randomly, with replacement, from the collection $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$, and put

$$Y_i^* = \hat{\beta}^T \mathbf{x}_i + \varepsilon_i^* \quad \text{and} \quad L^*(\beta) = \sum_{i=1}^n |Y_i^* - \beta^T \mathbf{x}_i|.$$

Choose $\hat{\beta}^*$ to minimize L^*, and let $\mathbf{z} = (z_1, \ldots, z_p)^T \in \mathbb{R}^p$. The bootstrap estimator of

$$G(\mathbf{z}) = P\{n^{1/2}(\hat{\beta}_j - \beta_{0j}) \le z_j,\ 1 \le j \le p\}$$

is given by

$$\hat{G}(\mathbf{z}) = P\{n^{1/2}(\hat{\beta}_j^* - \hat{\beta}_j) \le z_j,\ 1 \le j \le p | \mathcal{X}\},$$

where $\mathcal{X} = \{(\mathbf{x}_i, Y_i),\ 1 \le i \le n\}$ denotes the original data set. We shall show in Section 4 that the error in this approximation is of size $n^{-1/4}$, as $n \to \infty$.

As noted in Section 1, the asymptotic variance matrix of $n^{1/2}(\hat{\beta} - \beta_0)$ equals $\{2f(0)\}^{-2}\mathbf{V}_0^{-1}$, where $\mathbf{V}_0 = n^{-1} \sum \mathbf{x}_i \mathbf{x}_i^T$ and $f(0) = F'(0)$. This suggests a Normal approximation based on a kernel estimator of $f(0)$, as follows. Define

$$\hat{f}(0) = (nh)^{-1} \sum_{i=1}^n K(\hat{\varepsilon}_i / h), \tag{2.1}$$

where K is a symmetric density function. It may be shown that if $h \sim$ const. $n^{-1/5}$, then $n^{2/5}\{\hat{f}(0) - f(0)\}$ is asymptotically Normal $N(b_1, b_2^2)$, where $|b_1| < \infty$ and $0 < b_2 < \infty$. Therefore, if we take $\tilde{G}$ to be the distribution function of the p-variate Normal $(\mathbf{0}, \hat{\mathbf{\Sigma}})$ distribution, where $\hat{\mathbf{\Sigma}} = \{2\hat{f}(0)\}^{-2}\mathbf{V}_0^{-1}$, then

$$\sup_{\mathbf{z}} |\tilde{G}(\mathbf{z}) - G(\mathbf{z})| = O_p(n^{-2/5}). \tag{2.2}$$

Sufficient regularity conditions are those of Theorem 4.1 and that the symmetric density K be bounded and compactly supported. See Silverman (1986, chap. 3) for a discussion of kernel density estimation.

A convergence rate similar to that in (2.2) may be achieved by using the smoothed bootstrap, as follows. Generalizing (2.1), define

$$\hat{f}(x) = (nh_1)^{-1} \sum_{i=1}^n K\{(x - \hat{\varepsilon}_i)/h_1\}, \qquad -\infty < x < \infty.$$

Conditional on $\mathcal{X}$, let $\varepsilon_1^\dagger, \ldots, \varepsilon_n^\dagger$ denote independent variables drawn from the population with density $\hat{f}$. (Alternatively, we may center this distribution at its median first, but it may be shown that this does not affect the convergence rate given in (2.3).) Put

$$Y_i^\dagger = \hat{\beta}^T \mathbf{x}_i + \varepsilon_i^\dagger \quad \text{and} \quad L^\dagger(\beta) = \sum_{i=1}^n |Y_i^\dagger - \beta^T \mathbf{x}_i|.$$

Choose $\hat{\beta}^\dagger$ to minimize $L^\dagger$ and define

$$\check{G}(\mathbf{z}) = P\{n^{1/2}(\hat{\beta}_j^\dagger - \hat{\beta}_j) \le z_j,\ 1 \le j \le p | \mathcal{X}\}.$$

The argument used in Section 6 to prove Theorem 4.1 is readily modified to show that if $h_1 \sim$ const. $n^{-1/5}$, then for each $\eta, \lambda > 0$,

$$\sup_{\|\mathbf{z}\| \le \lambda} |\check{G}(\mathbf{z}) - G(\mathbf{z})| = O_p(n^{-(2/5)+\eta}). \tag{2.3}$$

A longer proof allows the right side to be refined to $O_p(n^{-2/5})$.

The discussion here relies crucially on the fact that the kernel K is nonnegative. Should K take negative values, then of course it is not a density, and so resampling from the distribution whose density is $\hat{f}$ is problematical. This difficulty has been discussed by Hall, DiCiccio, and Romano (1989), who suggested numerical integration as one solution to the problem. Another solution is to slightly modify $\hat{f}$ if K should take negative values, retaining its good convergence properties but removing its negativity.

Next we describe numerical results that compare the three approximations $\hat{G}$, $\tilde{G}$, and $\check{G}$ of G We consider the case $p = 2$ of the model (1.1), where

$$Y_i = \beta_{01} + \beta_{02} x_{i2} + \varepsilon_i, \qquad 1 \le i \le n, \tag{2.4}$$

with $\beta_{01} = 1.0$, $\beta_{02} = 2.0$, and $x_{12}, x_{22}, \ldots, x_{n2}$ fixed design points generated from a uniform distribution on [0, 1]. Four distributions for the independent errors $\varepsilon_1, \ldots, \varepsilon_n$ are considered: standard normal $N(0, 1)$, the normal mixture $.9N(0, 1) + .1N(0, 5)$, the double exponential with density $f(\varepsilon) = \frac{1}{2}e^{-|\varepsilon|}$, and the logistic with density $f(\varepsilon) = e^{-\varepsilon}/(1 + e^{-\varepsilon})$. Three sample sizes were considered: $n = 11$, 21, and 51.

For each of the 12 combinations of error distribution and sample size, the "exact" distribution G of $n^{1/2}(\hat{\beta} - \beta_0)$ was determined from a simulation of 5,000 data sets from the model (2.4), at 2,500 points of a regular grid in $[-17.5, 17.5] \times [-17.5, 17.5]$, which covers the range of values of $n^{1/2}(\hat{\beta} - \beta_0)$ experienced in the simulations.

From a given data set, we may construct three approximations to G—$\hat{G}$, $\tilde{G}$, and $\check{G}$—as described previously. In the study, the bootstrap approximations $\hat{G}$ and $\check{G}$ were estimated at each of the 2,500 grid points by drawing 2,500 resamples from the given data set, and $\tilde{G}$ was evaluated numerically at the same points. The quantities $\sup|\hat{G} - G|$, $\sup|\tilde{G} - G|$, and $\sup|\check{G} - G|$ were then evaluated, and these error measures were averaged over 2,000 simulations from the model. The resulting summaries of the accuracies of the three estimators are given in Table 1.

The smoothed bootstrap estimator $\check{G}$ and the normal approximation $\tilde{G}$ both depend on the choice of a kernel function K and a bandwidth. Throughout the simulation, the Epanechnikov kernel was used. This has the form

$$K(t) = \{3/(4\sqrt{5})\}\{1 - (t^2/5)\}, \quad \text{if } |t| \le \sqrt{5},$$
$$= 0, \quad \text{if otherwise.}$$

Table 1. Average Over Simulation of $\sup|\hat{G} - G|$ (UNS), $\sup|\check{G} - G|$ (SM), and $\sup|\tilde{G} - G|$ (NOR) for the Regression Model

Error distribution		UNS	SM			NOR		
			$c_1 = .5$	$c_1 = 1.0$	$c_1 = 1.5$	$c_2 = .5$	$c_2 = 1.0$	$c_2 = 1.5$
Normal	$n = 11$	.25221	.15123	.12348	.14585	.15920	.12410	.15517
	$n = 21$	.20237	.11476	.09462	.12546	.11915	.09325	.13013
	$n = 51$	.15533	.08226	.07248	.10428	.08200	.06954	.10482
Normal mixture	$n = 11$	.25886	.14731	.18052	.24117	.15245	.18377	.25157
	$n = 21$	.21263	.10529	.15770	.24094	.10554	.15960	.24626
	$n = 51$	.15889	.07562	.13877	.23373	.07431	.13952	.23616
Double exponential	$n = 11$	.24493	.13538	.16328	.22285	.14178	.17013	.23570
	$n = 21$	.19619	.10774	.16723	.24767	.10745	.16989	.25458
	$n = 51$	.15146	.08915	.16521	.24910	.08540	.16436	.25062
Logistic	$n = 11$	.26307	.15130	.13206	.16563	.15974	.13205	.17550
	$n = 21$	.20867	.11014	.10322	.15042	.11353	.10220	.15525
	$n = 51$	.15883	.08179	.07709	.12176	.08127	.07231	.12050

NOTE: Simulation size was 2,000.

Letting $\hat{\sigma}^2$ denote the empirical variance of the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$, we investigate the following empirical rules for choice of the bandwidths h_1, h:

$$h_1 = c_1\hat{\sigma}n^{-1/5},$$
$$h = c_2\hat{\sigma}n^{-1/5}, \tag{2.5}$$

for values of c_1, c_2 to be chosen. In the simulation these values are taken, arbitrarily, as $c_1, c_2 = .5, 1.0$, or 1.5.

Random number generation was done using the NAG subroutine library and L^1 minimization performed using the NAG routine `e02gaf`. The normal approximation $\tilde{G}$ was evaluated using the NAG routine `g01haf` for computation of the bivariate normal distribution function.

The results in Table 1 highlight the slow rate of convergence of the unsmoothed bootstrap (UNS) in this context. Each figure in the table has a standard error of no more than .0015. The smoothed bootstrap (SM) and normal approximation (NOR) methods improve greatly and significantly on the UNS, as the theory predicts. These two improved estimators rarely perform significantly differently from each other. The figures suggest that, at least for the simple case $p = 2$ considered here, the NOR method is to be preferred over the computationally more expensive SM. Success of the simple bandwidth selection rules (2.5) is striking; taking $c_1 = c_2 = .5$ always gives considerable improvement over the UNS, although optimal choice of bandwidth clearly depends on the underlying error distribution.

3. ANALYTICAL APPROXIMATION OF DISTRIBUTION OF $\hat{\beta}$

We assume the model described in Section 1. Our attention is mainly devoted to the case where one component (which we take to be the first) of β_0 is a location or intercept constant, and those components of the design vector $\mathbf{x}_i$ that do not correspond to the location constant vary in a reasonably smooth manner. Other cases are treated in Remarks 3.5 and 3.6 following the first theorem. The second theorem specializes the first result to the case where the error distribution is smooth.

We begin by describing the related Edgeworth expansion in a simpler setting. Then we show how this may be developed into a very general expansion of the distribution of $\hat{\beta}$. Let $z_1, \ldots, z_p$ denote real numbers, and put

$$\delta_i = n^{-1/2}\sum_{j=1}^{p} x_{ij}z_j \quad \text{and}$$
$$\mu_i = 2P(\delta_i < \varepsilon \le 0) \quad \text{if } \delta_i < 0$$
$$= -2P(0 \le \varepsilon < \delta_i) \quad \text{if } \delta_i \ge 0. \tag{3.1}$$

Let J_i, $1 \le i \le n$, denote independent random variables taking the values ± 1 with probabilities $1/2(1 \pm \mu_i)$; note that $E(J_i) = \mu_i$ and put $\Delta_i = J_i - \mu_i$, and let $\mathbf{Z}_i$ be the p vector whose jth component is $x_{ij}\Delta_i$. Write $\{\chi_\nu\}$ for the average of the p-variate cumulant sequence and $\mathbf{V} = n^{-1}\sum(1 - \mu_i^2)\mathbf{x}_i\mathbf{x}_i^T$ for the average variance matrix, of the vectors $(x_{i1}\Delta_i, \ldots, x_{ip}\Delta_i)^T$, $1 \le i \le n$ (Bhattacharya and Rao 1976, p. 71; the X_i there is here replaced by $x_{i1}\Delta_i$); and let $P_k(-\phi_{\mathbf{0},\mathbf{V}} : \{\chi_\nu\})$ denote the usual polynomial multiple of the $N(\mathbf{0}, \mathbf{V})$ density $\phi_{\mathbf{0},\mathbf{V}}$ (Bhattacharya and Rao 1976, pp. 53–54). We generally shall assume that $x_{i1} = 1$ for each i.

Let $\mathbf{S} = \Sigma\mathbf{Z}_i$. If the vectors $\mathbf{Z}_i$ had continuous distributions, then it would typically be the case that the distribution of the vector $\mathbf{S} = (S_1, \ldots, S_p)^T$ would admit an Edgeworth expansion of the form

$$P(n^{-1/2}\mathbf{S} \in B) = \int_B \left\{\phi_{\mathbf{0},\mathbf{V}}(\mathbf{x}) + \sum_{k=1}^{r} n^{-k/2}P_k(-\phi_{\mathbf{0},\mathbf{V}} : \{\chi_\nu\})(\mathbf{x})\right\} d\mathbf{x} + O(n^{-(r+1)/2}), \tag{3.2}$$

uniformly in a large class of Borel sets B. See, for example, Bhattacharya and Rao (1976, p. 194). But here the first component of $\mathbf{S}$ has a lattice distribution, and the other $p - 1$ components have discrete distributions.

Nevertheless, we claim that an analog of (3.2) is true, which allows for the latticeness of the distribution of S_1 and ignores the discreteness of $S_2, \ldots, S_p$. Most important, that analog produces an Edgeworth expansion of the distribution of $n^{1/2}(\hat{\beta} - \beta_0)$ under regularity conditions that do not even demand a density for the error distribution. The reader is

referred to Bhattacharya and Rao (1976, chap. 5) for a discussion of Edgeworth approximations in the lattice case.

Let $r \geq 0$ be fixed.

Theorem 3.1. Assume that $x_{i1} = 1$ for all i, and that $(x_{i2}, \ldots, x_{ip})^T$ for $1 \leq i \leq n$ represent independent values of a $(p-1)$ vector $\mathbf{X} = (X_2, \ldots, X_p)^T$ that satisfies the moment condition

$$\sum_{j=2}^{p} E(|X_j|^{r+3}) < \infty \tag{3.3}$$

and the Cramér-type condition

$$\sup_{0 \leq t_1 \leq \pi, t_1^2 + \cdots + t_p^2 > \eta^2} E\left\{\cos\left(t_1 + \sum_{j=2}^{p} t_j X_j\right)\right\} < 1 \quad \text{for all} \quad \eta > 0 \tag{3.4}$$

and is such that with $\mathbf{X}$ probability 1, the average variance matrix $\mathbf{V}$ converges to a proper, nonsingular limit. Assume too that the distribution function F of the error distribution is Hölder continuous at the origin and satisfies $F(0) = \frac{1}{2}$. Define $v_j = -n^{-1/2} \sum_i \mu_i x_{ij}$. Then for a sequence of design points $\mathbf{x}_1, \mathbf{x}_2, \ldots$ arising with $\mathbf{X}$ probability 1 and for $\zeta > 0$ sufficiently small,

$$\begin{aligned} P\{n^{1/2}(\hat{\beta}_j - \beta_{0j}) \leq z_j,\ 1 \leq j \leq p\} &= 2n^{-1/2} \\ &\times \sum_{k \leq \mu_1 + n^{-1/2} v_1} \int \cdots \int_{u_j \leq v_j, 2 \leq j \leq p} [\phi_{\mathbf{0},\mathbf{V}}\{n^{-1/2}(k - \mu_1), \\ u_2, \ldots, u_p\} &+ \sum_{k=1}^{r} n^{-k/2} P_k(-\phi_{\mathbf{0},\mathbf{V}} : \{\chi_\nu\}) \\ &\times \{n^{-1/2}(k - \mu_1), u_2, \ldots, u_p\}]\, du_2 \ldots du_p \\ &+ O(n^{-(r+1)/2}) \end{aligned} \tag{3.5}$$

uniformly in vectors $\mathbf{z} = (z_1, \ldots, z_p)^T$ satisfying $\|\mathbf{z}\| = (\sum z_i^2)^{1/2} \leq \zeta n^{1/2}$.

The qualification "with $\mathbf{X}$ probability 1" means "for sequences $\mathbf{x}_1, \mathbf{x}_2, \ldots$ arising with probability 1 as independent realizations of $\mathbf{X}$".

Remark 3.1. The moment condition (3.3) is required because the $(k+2)$th sample moment of the collection $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ appears as a coefficent of the polynomial P_k. Because P_{r+1} is a major contributor to the remainder term $O(n^{-(r+1)/2})$ in (3.5), the remainder would be of larger order than $n^{-(r+1)/2}$ if the $(r+3)$ moment were infinite.

Remark 3.2. The Cramér-type condition (3.4) differs from the more usual Cramér condition,

$$\sup_{t_1^2 + \cdots + t_p^2 > \eta^2} \left| E\left\{\exp\left(i \sum_{j=1}^{p} t_j X_j\right)\right\} \right| < 1 \quad \text{for all } \eta > 0 \tag{3.6}$$

(see, for example, Bhattacharya and Rao 1976, p. 207), in that it involves only the cosine function, it is one-sided (i.e., we do not take the absolute value of the expectation in (3.4)), the range of t_1 is bounded, and the coefficient of t_1 is nonrandom. Condition (3.4) holds if $(X_2, \ldots, X_p)$ satisfies the usual $(p-1)$-variate Cramér condition and if $\sum E(X_j^2) < \infty$ (see the Appendix). In particular, it is sufficient for the distribution of $(X_2, \ldots, X_p)$ to have finite variance and a nondegenerate continuous component.

Remark 3.3. The condition that $\|\mathbf{z}\| \leq \zeta n^{1/2}$ is hardly important, because it is readily deduced from (3.5) that

$$\begin{aligned} &\sum_{j=1}^{p} P(|\hat{\beta}_j - \beta_{0j}| > (1/2)\zeta) \\ &\quad = \sum_{j=1}^{p} P(n^{1/2}|\hat{\beta}_j - \beta_{0j}| > (1/2)\zeta n^{1/2}) = O(n^{-(r+1)/2}). \end{aligned}$$

Remark 3.4. The factor $2n^{-1/2}$ outside the summation sign in (3.5) derives from the lattice component $S_1 = \sum \Delta_i$. Note that the span of the lattice of Δ_i equals 2 and that the density of the discrete random variable S_1 is of size $n^{-1/2}$.

Remark 3.5. If there were no intercept in the linear model—for example, if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ were a realization of a sequence of independent p vectors distributed as $(X_1, \ldots, X_p)^T$ and satisfying (3.6)—then the lattice component of the expansion (3.5) would vanish. In this event the right side of (3.5) would be identical to that of (3.2), with $B = \prod (-\infty, v_j]$.

Remark 3.6. There is a version of the theorem for the case of regularly spaced design, in particular where $\mathbf{x}_i = (1, i/n, (i/n)^2, \ldots, (i/n)^{p-1})^T$, $1 \leq i \leq n$. In this circumstance the conditions on the $\mathbf{X}$ distribution, including (3.4), are of course dropped from the theorem.

To conclude, we suppose that the underlying error distribution F has three bounded derivatives in a neighborhood of the origin, with $f \equiv F'$ and $f(0) \neq 0$. Assume also the conditions of Theorem 3.1 for the case $r = 1$. We claim that in this circumstance the expansion (3.5) assumes a simpler form. Indeed, the asymptotic symmetry of the variables Δ_i ensures that third-order cumulants equal $O(n^{-1/2})$, whence $P_1(-\phi_{\mathbf{0},\mathbf{V}} : \{\chi_\nu\}) = O(n^{-1/2})$. This term may be omitted from (3.5) if we seek the latter expansion only up to a remainder of $O(n^{-1})$. It may be shown by Taylor expansion that

$$\begin{aligned} \mathbf{V} &= \mathbf{V}_0 = O(n^{-1}), \\ \mathbf{v} &= (v_1, \ldots, v_p)^T \\ &= 2f(0)\mathbf{V}_0\mathbf{z} + n^{-1/2} f'(0) a(\mathbf{z}) + O(n^{-1}), \end{aligned}$$

where $\mathbf{z} = (z_1, \ldots, z_p)^T$,

$$(\mathbf{V}_0)_{j_1 j_2} = n^{-1} \sum_{i=1}^{n} x_{ij_1} x_{ij_2}, \quad \text{and}$$

$$(a(\mathbf{z}))_j = n^{-1} \sum_{i=1}^{n} (\mathbf{x}_i^T \mathbf{z})^2 x_{ij}.$$

(The formula for $\mathbf{V}$ follows from the fact $\mu_i = O(n^{-1/2})$.) Arguing thus, we may prove the following result.

Theorem 3.2. Assume the conditions imposed in the previous paragraph. Then for a sequence of design points $\mathbf{x}_1, \mathbf{x}_2, \ldots$ arising with $\mathbf{X}$ probability 1, and for $\zeta > 0$ suffi-

ciently small,

$$P\{n^{1/2}(\hat\beta_j - \beta_{0j}) \le z_j,\ 1 \le j \le p\}$$
$$= 2n^{-1/2} \sum_{k \le \mu_1 + n^{-1/2}v_1} \int \cdots \int_{u_j \le 2f(0)(\mathbf{V}_0\mathbf{z})_j + n^{-1/2}f'(0)\{a(\mathbf{z})\}_j}$$
$$\times \phi_{\mathbf{0},\mathbf{V}_0}\{n^{-1/2}(k - \mu_1), u_2, \ldots, u_p\}\, du_2, \ldots, du_p + O(n^{-1}) \quad (3.7)$$

uniformly in vectors $\mathbf{z} = (z_1, \ldots, z_p)^T$ satisfying $\|\mathbf{z}\| \le \zeta n^{1/2}$.

Remark 3.7. Owing to the discrete rounding errors of size $n^{-1/2}$ arising from the series in (3.7), the quantities μ_1 and v_1 appearing there may not be replaced by their Taylor expansions. We do not suggest that (3.7) be used to develop a numerical approximation to the distribution on the left-side, but state it here because of its theoretical interest.

4. THEORY FOR BOOTSTRAP APPROXIMATION TO DISTRIBUTION OF $\hat\beta$

We begin by describing the basic bootstrap approximations, $\hat G$, of the distribution G. (Refer to Sec. 2 for notation.) Our claim is that the error between $\hat G$ and G is of order $n^{-(1/4)+\eta}$ for each $\eta > 0$. In particular, the unsmoothed bootstrap approximant $\hat G$ consistently estimates G, as our next theorem shows.

Theorem 4.1. Assume the conditions of Theorem 3.1, except that we strengthen (3.3) by asking that $P(\|\mathbf{X}\| \le C) = 1$ for some $C > 0$ and insist that the distribution function F of the error distribution have three bounded derivatives in a neighborhood of the origin, with $F'(0) \ne 0$. Then for each $\eta, \lambda > 0$, and for a sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots$ arising with $\mathbf{X}$ probability 1,

$$\sup_{\|\mathbf{z}\| \le \lambda} |\hat G(\mathbf{z}) - G(\mathbf{z})| = O_p(n^{-(1/4)+\eta}).$$

Remark 4.1. A longer proof than that given here may be used to show that the convergence rate of $\hat G$ to G is actually $O_p(n^{-1/4})$, not just $O_p(n^{-(1/4)+\eta})$. We shall make clear in our proof the changes that are needed. The rate $n^{-1/4}$ cannot be improved on, as we briefly explain now. By comparing Edgeworth approximations to the functions G and $\hat G$ (see (3.7) and (6.3)), we may deduce that $\hat G(\mathbf{z}) - G(\mathbf{z}) \equiv o_p(n^{-1/4})$ if and only if $\hat v_j - v_j = o_p(n^{-1/4})$, for $1 \le j \le p$, where $\hat v_j$ (defined in Sec. 6) is an empirical approximation to v_j (defined in Sec. 3). The actual formula for $\hat v_j$ is rather complicated, but this quantity may be shown to have the same first-order asymptotic properties as

$$\tilde v_j = -n^{-1/2} \sum_{l=1}^{n} x_{lj} \Big[I(\delta_l < 0) \sum_{i=1}^{n} I\{\varepsilon_i \in (\delta_l, 0]\} - I(\delta_l \ge 0) \sum_{i=1}^{n} I\{\varepsilon_i \in [0, \delta_l)\} \Big], \quad (4.1)$$

where δ_l is given by (3.1). Now the right-side of (4.1) is a sum of independent random variables with mean v_j and variance asymptotic to a constant multiple of $n^{-1/2}$. In fact, it is asymptotically Normally distributed with this mean and variance and closely resembles a kernel-type density estimator with variable bandwidth $h = \delta_l \simeq n^{-1/2}$. The variance of such a quantity is of course approximately given by $(nh)^{-1} \simeq n^{-1/2}$. Therefore, $\tilde v_j - v_j$ and $\hat v_j - v_j$ are of size $(n^{-1/2})^{1/2} = n^{-1/4}$, as claimed earlier.

Remark 4.2. As noted in Section 1, the asymptotic variance matrix of $n^{1/2}(\hat\beta - \beta)$ equals $\{2f(0)\}^{-2}\mathbf{V}_0^{-1}$, where $\mathbf{V}_0 = n^{-1}\sum \mathbf{x}_i\mathbf{x}_i^T$ and $f(0) = F'(0)$ denotes the error density evaluated at the median. Indeed, it may be proved that under the conditions of Theorem 4.1,

$$\operatorname{var}\{n^{1/2}(\hat\beta - \beta_0)\} = \{2f(0)\}^{-2}\mathbf{V}_0^{-1} + O(n^{-1}). \quad (4.2)$$

Therefore, a Normal approximation to the distribution G of $n^{1/2}(\hat\beta - \beta_0)$ may be used, provided that we estimate $f(0)$. Should our estimate of $f(0)$ be accurate to $O(n^{-c})$, where $0 < c < 1/2$, then in view of (3.7) and (4.2) our Normal approximation to G will also be accurate to $O(n^{-c})$.

Remark 4.3. As indicated in Remark 4.1, the convergence rate $n^{-1/4}$ comes about because the bootstrap is implicitly like using a kernel estimator to estimate $f(0)$ and taking the bandwidth to be $h = n^{-1/2}$. Of course, $h = n^{-1/2}$ is too small a smoothing parameter for such an approach; taking h to be of size $n^{-1/5}$ provides a faster convergence rate of $n^{-2/5}$ when estimating $f(0)$ (see, for example, Silverman 1986, p. 40ff). This leads to the Normal approximation $\tilde G$, suggested in Section 2.

Remark 4.4. Result (2.2) shows that $\tilde G$ improves on the bootstrap estimator $\hat G$, but it has the drawback of requiring the choice of the bandwidth h. This may be done empirically, using methods such as those suggested by Silverman (1986, p. 43ff). Higher orders of approximation may be obtained by using a higher-order kernel K in definition (2.1) (see, for example, Silverman 1986, p. 66ff). But because $f(0)$ generally cannot be estimated root-n consistently without making parametric assumptions, then the convergence rate in (2.2) generally cannot be improved to $O_p(n^{-1/2})$.

Remark 4.5. Our discussion in Remark 4.3 of the bootstrap approximation $\hat G$, which suggests that the bootstrap implicitly estimates the variance of $\hat\beta$ with an error of order $n^{-1/4}$, is in line with the bootstrap's peformance in estimating the variance of a sample quantile. There the convergence rate is also $n^{-1/4}$ (see Babu 1986; Hall and Martin 1988). The issue of smoothing to improve this rate was treated by Hall, DiCiccio, and Romano (1989).

5. PROOF OF THEOREM 3.1

Observe that $L(\beta)$ is a convex, head-down, cup-shaped surface in p-dimensional space with its vertex at $\hat\beta$, and that with probability 1, $\hat\beta \le \beta$ if and only if $\partial L(\beta)/\partial\beta \ge 0$, where both inequalities are interpreted element-wise. It follows, after some algebra, that

$$P\{n^{1/2}(\hat\beta_j - \beta_{0j}) \le z_j,\ 1 \le j \le p\} = P\Big(\sum_l x_{lj}\Delta_l \le v_j,\ 1 \le j \le p\Big),$$

where $\Delta_l = \operatorname{sgn}(\varepsilon_l - n^{-1/2}\sum_k x_{lk}z_k) - \mu_l$. Put $S_j = \sum_l x_{lj}\Delta_l$, $1 \le j \le p$. We smooth $S_2, \ldots, S_p$ by adding to them the

variables $n^{-w}N_2, \ldots, n^{-w}N_p$, where the N_j's are independent standard normal variables independent also of $\varepsilon_1, \ldots, \varepsilon_n$, and $w > 0$ is a large, fixed, positive constant. We seek the joint density (q, say) of $\mathbf{T} = (T_1, \ldots, T_p)^T = (n^{-1/2}S_1, n^{-1/2}S_2 + n^{-w}N_2, \ldots, n^{-1/2}S_p + n^{-w}N_p)$. Put $\mathbf{S} = (S_1, \ldots, S_p)^T$.

For any Borel set $B \subseteq \mathbb{R}^p$,

$$P(\mathbf{T} \in B) = \sum_k \int_{\mathbf{u} \in \mathbb{R}^{p-1}:(n^{-1/2}(k-\mu_1),\mathbf{u}) \in B} q\{n^{-1/2}(k - \mu_1), \mathbf{u}\}\, d\mathbf{u}. \tag{5.1}$$

If we demonstrate that

$$\sum_k \int_{\mathbb{R}^{p-1}} |q\{n^{-1/2}(k - \mu_1), \mathbf{u}\} - q_r\{n^{-1/2}(k - \mu_1), \mathbf{u}\}|\, d\mathbf{u} = O(n^{-(r+1)/2}), \tag{5.2}$$

where q_r denotes the integrand of the integral on the right side of (3.5), then we may deduce from (5.1) that

$$\sup_{B \in \mathcal{B}} |P(\mathbf{T} \in B) - Q_r(B)| = O(n^{-(r+1)/2}), \tag{5.3}$$

where Q_r is the signed measure whose density is q_r and $\mathcal{B}$ is the class of all Borel subsets of $\mathbb{R}^p$. Now the class $\mathcal{C}$ of all semi-infinite p-dimensional rectangles $\prod (-\infty, a_j]$ satisfies

$$\sup_{B \in \mathcal{C}} \int_{(\partial B)^\eta} \phi_{\mathbf{0},\mathbf{I}_p}(\mathbf{z})\, d\mathbf{z} = O(\eta)$$

as $\eta \downarrow 0$, where $(\partial B)^\eta$ denotes the set of all points in $\mathbb{R}^p$ distant η or less from the boundary of ∂B. Therefore, with $\eta = n^{-(w-1)}$ and $w > (r + 3)/2$,

$$\begin{aligned}\sup_{B \in \mathcal{C}} |P(\mathbf{T} \in B) - P(n^{-1/2}\mathbf{S} \in B)| &\le 2nP(|N_1| > n) + 2 \sup_{B \in \mathcal{C}} P\{\mathbf{T} \in (\partial B)^\eta\} \\ &= O(n^{-(w-1)}) = O(n^{-(r+1)/2}).\end{aligned}$$

The theorem follows from this result and (5.3). So the proof may be completed by deriving (5.2), for which the reader is referred to our technical report, De Angelis, Hall, and Young (1991) on which this article is based.

6. PROOF OF THEOREM 4.1

Let δ_l be as defined in Section 3, and put

$$\begin{aligned}\hat{\mu}_l &= 2P(\delta_l < \varepsilon_1^* \le 0 | \mathcal{X}) \quad \text{if } \delta_l < 0 \\ &= -2P(0 \le \varepsilon_1^* < \delta_l | \mathcal{X}) \quad \text{if } \delta_l \ge 0.\end{aligned}$$

Then, conditional on $\mathcal{X}$, the variables $\Delta_l^* = \operatorname{sgn}(\varepsilon_l^* - n^{-1/2}\delta_l) - \hat{\mu}_l$ are independent and identically distributed with 0 mean, taking the values $\pm 1 - \hat{\mu}_l$ with probabilities $\frac{1}{2}(1 \pm \hat{\mu}_l)$. Put $S_j^* = \sum_l x_{lj}\Delta_l^*$, $1 \le j \le p$. A version of Theorem 3.1 may be derived by arguing as in Section 5. The "smoothing" part of that method can be conducted as before, working with $\mathbf{T}^* = (n^{-1/2}S_1^*, n^{-1/2}S_2^* + n^{-w}N_2, \ldots, n^{-1/2}S_p^* + n^{-w}N_p)$ instead of $\mathbf{T}$, where $N_2, \ldots, N_p$ are independent standard normal random variables independent of $\mathcal{X}$ and of $\varepsilon_1^*, \ldots, \varepsilon_n^*$. We need this version of Theorem 3.1 only up to a remainder of order $n^{-1/2}$, and so that is the extent to which we state it here:

$$\hat{G}(\mathbf{z}) = 2n^{-1/2} \sum_{k \le \hat{\mu}_1 + n^{-1/2}\hat{v}_1} \int \cdots \int_{u_j \le \hat{v}_j, 2 \le j \le p} \times \phi_{\mathbf{0},\hat{\mathbf{V}}}\{n^{-1/2}(k - \hat{\mu}_1), u_2, \ldots, u_p\} \times du_2 \ldots du_p + O(n^{-1/2}) \tag{6.1}$$

with probability 1, where $\hat{\mathbf{V}}$ denotes the average p-variate variance matrix of the vectors $(x_{l1}\Delta_l^*, \ldots, x_{lp}\Delta_l^*)^T$, $1 \le l \le p$, conditional on $\mathcal{X}$, and $\hat{v}_j = -n^{-1/2}\sum_l \hat{\mu}_l x_{lj}$. It may be shown by Markov's inequality that for each $\xi, \eta > 0$,

$$P\left\{\max_{1 \le j \le p} \left|\hat{v}_j - 2n^{-1/2}f(0)\sum_{l=1}^n x_{jl}\delta_l\right| > \zeta n^{-(1/4)+\eta}\right\} \to 0. \tag{6.2}$$

Details are given in our technical report. More simply, it may be proved that $\hat{\mathbf{V}} = \mathbf{V}_0 + O_p(n^{-1/2})$, where $\mathbf{V}_0$ was defined just before Theorem 3.2. Therefore, directly from (6.1),

$$\hat{G}(\mathbf{z}) = 2n^{-1/2} \sum_{k \le \hat{\mu}_1 + n^{-1/2}\hat{v}_1} \int \cdots \int_{u_j \le \hat{v}_j, 2 \le j \le p} \times \phi_{\mathbf{0},\mathbf{V}_0}\{n^{-1/2}(k - \hat{\mu}_1), u_2, \ldots, u_p\} \times du_2, \ldots, du_p + O_p(n^{-1/2}). \tag{6.3}$$

Theorem 3.1 follows from combining this formula with (6.2) and (3.7).

APPENDIX: VERIFICATION OF (3.4)

Here we outline a proof that if $E(X_2^2 + \cdots + X_p^2) < \infty$ and $(X_2, \ldots, X_p)$ satisfies the usual Cramér condition,

$$1 - \delta_1(\eta) = \sup_{t_2^2 + \cdots + t_p^2 > \eta^2} |E[\exp\{i(t_2X_2 + \cdots + t_pX_p)\}]| < 1, \quad \text{all } \eta > 0,$$

then the modified condition (3.4) also holds. We treat only the case $p = 2$, because $p \ge 3$ differs only in that the notation is more complex.

Given $\eta > 0$, choose $0 < \rho < 1$ so small that

$$1 - \delta_2 \equiv \sup_{(1-\rho^2)^{1/2}\eta \le t_1 \le \pi} \{\cos t_1 + (1/2)\rho^2\eta^2 E(X_2^2) + \rho\eta\pi(EX_2^2)^{1/2}\} < 1.$$

If $0 \le t_1 \le \pi$, $t_1^2 + t_2^2 > \eta^2$ and $t_2^2 \le \rho^2\eta^2$, then $(1 - \rho^2)^{1/2}\eta \le t_1 \le \pi$, and also

$$\begin{aligned}E\{\cos(t_1 + t_2X_2)\} &= E[\cos t_1 - (\cos t_1)\{1 - \cos(t_2X_2)\} - (\sin t_1)\sin(t_2X_2)] \\ &\le \cos t_1 + (1/2)t_2^2E(X_2^2) + |t_1t_2|(EX_2^2)^{1/2} \\ &\le 1 - \delta_2.\end{aligned}$$

Furthermore, if $t_2^2 > \rho^2\eta^2$, then

$$\begin{aligned}|E\{\cos(t_1 + t_2X_2)\}| &= |\operatorname{Re}\exp(it_1)E\{\exp(it_2X_2)\}| \\ &\le |E\{\exp(it_2X_2)\}| \le 1 - \delta_1(\rho\eta).\end{aligned}$$

Therefore,

$$\sup_{0\le t_1\le\pi, t_1^2+t_2^2>\eta^2} E\{\cos(t_1 + t_2X_2)\} \le \max\{1 - \delta_1(\rho\eta), 1 - \delta_2\} < 1,$$

which verifies (3.4).

[Received December 1991. Revised December 1992.]

REFERENCES

Babu, G. J. (1986), "A Note on Bootstrapping the Variance of Sample Quantile," *The Annals of the Institute of Statistical Mathematics,* 38, 439–443.

——— (1991), "Edgeworth Expansions for Statistics Which are Functions of Lattice and Nonlattice Random Variables," *Statistics and Probability Letters,* 12, 1–7.

Babu, G. J., and Singh, K. (1989a), "A Note on Edgeworth Expansions for the Lattice Case," *Journal of Multivariate Analysis,* 30, 27–33.

——— (1989b), "On Edgeworth Expansions in the Mixture Cases," *The Annals of Statistics,* 17, 443–447.

Bassett, G., and Koenker, R. (1978), "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association,* 73, 618–621.

Bhattacharya, R. N., and Qumsiyeh, M. (1989), "Second Order and L^p Comparisons Between the Bootstrap and Empirical Edgeworth Expansion Methodologies," *The Annals of Statistics,* 17, 160–169.

Bhattacharya, R. N., and Rao, R. R. (1976), *Normal Approximation and Asymptotic Expansions,* New York: John Wiley.

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

De Angelis, D., Hall, P., and Young, G. A. (1991), Analytical and bootstrap approximations to estimator distributions in L^1 regression. Research Report CMA-SR32-91, Centre for Mathematics and its Applications, Australian National Univ.

Hall, P. (1988), "Unusual Properties of Bootstrap Confidence Intervals in Regression Problems," *Probability Theory and Related Fields,* 81, 247–273.

——— (1992), *The Bootstrap and Edgeworth Expansion,* New York: Springer-Verlag.

Hall, P., DiCiccio, T. J., and Romano, J. P. (1989), "On Smoothing and the Bootstrap," *The Annals of Statistics,* 17, 692–704.

Hall, P., and Martin, M. A. (1988), "Exact Convergence Rate of Bootstrap Quantile Variance Estimator, *Probability Theory and Related Fields,* 80, 261–268.

Mammen, E. (1989), "Asymptotics With Increasing Dimension for Robust Regression With Applications to the Bootstrap," *The Annals of Statistics,* 17, 382–400.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis,* London: Chapman and Hall.

Withers, C. S. (1987), "The Bias and Skewness of L_1 Estimates in Regression," *Computational Statistics and Data Analysis,* 5, 301–303.

Wu, C. F. J. (1986), Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis" (With Discussion), *The Annals of Statistics,* 14, 1261–1295.

Yarnold, J. K. (1972), "Asymptotic Approximations for the Probability That a Sum of Lattice Random Vectors Lies in a Convex Set," *The Annals of Mathematical Statistics,* 43, 1566–1580.