Taylor & Francis
Taylor & Francis Group

Rank-Based Estimates in the Linear Model with High Breakdown Point

Author(s): Ola Hossjer

Source: *Journal of the American Statistical Association*, Vol. 89, No. 425 (Mar., 1994), pp. 149-158
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/2291211
Accessed: 03-02-2016 22:26 UTC

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Taylor & Francis, Ltd. and American Statistical Association are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association.*

http://www.jstor.org

This content downloaded from 130.237.165.40 on Wed, 03 Feb 2016 22:26:14 UTC
All use subject to JSTOR Terms and Conditions

Rank-Based Estimates in the Linear Model With High Breakdown Point

Ola HÖSSJER*

An estimator in the linear model is defined by minimizing an objective function, the derivative of which is a signed rank statistic. The scores are generated from a function $h^+: (0, 1) \to [0, \infty)$, which is not necessarily nondecreasing, as is usually assumed. It is shown that this estimator can be chosen with a maximal breakdown point of .5. Moreover, strong consistency and asymptotic normality (with convergence rate $n^{-1/2}$) of the proposed estimator are proved under various regularity conditions. Because the objective function generally is not convex in the regression parameters, the usual proofs of asymptotic normality do not carry over. Instead the proof is based on an asymptotic linearity result, similar to that obtained by Huber for M estimates, and some moment estimates for signed rank statistics. Numerical examples illustrate the behavior of the estimator.

KEY WORDS: Asymptotic normality; Exact fit property; Robust estimation; Signed rank statistic; Trimmed residuals.

1. INTRODUCTION

Consider the linear model

$$y_i = \theta_0 \mathbf{x}_i' + e_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{x}_i$ and θ_0 are vectors in $\mathbb{R}^p$ and the scalars e_i represent the error terms. We are interested in the case when there is uncertainty in the carrier variables $\mathbf{x}_i$. In particular, we are interested in estimates of θ_0 with high breakdown point. A finite sample version of the breakdown point was introduced by Donoho and Huber (1983). Let $\mathbf{T}_n$ be an estimate of θ_0 and let $\mathbf{Z}_n$ represent any sample $\mathbf{z}_1 = (\mathbf{x}_1, y_1), \ldots, \mathbf{z}_n = (\mathbf{x}_n, y_n)$. Given m, the maximal bias $b_n(m, \mathbf{T}_n, \mathbf{Z}_n)$ is defined as the supremum of $\|\mathbf{T}_n(\mathbf{Z}_n^*) - \mathbf{T}_n(\mathbf{Z}_n)\|$ over all samples $\mathbf{Z}_n^*$ with at most m vectors $\mathbf{z}_i$ differing from those of $\mathbf{Z}_n$. The breakdown point of $\mathbf{T}_n$ (with replacement) at $\mathbf{Z}_n$ is then defined as

$$\varepsilon_n^*(\mathbf{T}_n, \mathbf{Z}_n) = \min\left\{\frac{m}{n}\,;\, b_n(m, \mathbf{T}_n, \mathbf{Z}_n) \text{ is infinite}\right\}. \tag{2}$$

Consequently, ε_n^* is the smallest fraction of outliers that can carry the estimate $\mathbf{T}_n$ over all bounds. With $\mathbf{T} = \{\mathbf{T}_n\}_{n \geq p}$ representing the whole sequence of estimates, we define

$$\varepsilon^*(\mathbf{T}) = \varliminf_{n\to\infty} \inf_{\mathbf{Z}_n \in \mathcal{Z}_n} \varepsilon_n^*(\mathbf{T}_n, \mathbf{Z}_n) \tag{3}$$

as the asymptotic breakdown point of $\mathbf{T}$. Here $\mathcal{Z}_n$ denotes the set of all samples in general position (see Sec. 2). (Other definitions of breakdown point have been given in Hodges 1967 and Hampel 1971.) Many of the well-known estimators, such as least squares (LS), M estimators with a nondecreasing ψ function (Huber 1973), and R estimators based on signed rank statistics with nondecreasing scores (Hettmansperger and McKean 1983), have $\varepsilon^* = 0$ and thus bad protection against outliers in the x direction. The first example of an estimator with the maximal $\varepsilon^* = .5$ was given by Siegel (1982): the repeated median (RM). Rousseeuw (1984, 1985) studied the least median of squares (LMS) and least trimmed squares (LTS), which are defined by minimizing the median or the trimmed mean of the squared residuals. Other examples are S estimators (Rousseeuw and Yohai 1984), MM estimators (Yohai 1987), and τ estimators (Yohai and Zamar 1989). The latter two estimators can be chosen with maximal breakdown point and at the same time an efficiency arbitrarily close to 1 when the errors are independent and indentically distributed (iid) with a normal distribution. Jurečkovà and Portnoy (1987) showed that a one-step M estimator based on a preliminary robust estimate inherits the breakdown point of the latter estimate at the same time as the efficiency can be chosen arbitrarily close to 1. Until recently, there has been little research on finding rank-based estimators resistant to leverage points (i.e., outliers in the x direction). Sievers (1983) considered an estimator based on minimizing a weighted Gini's mean difference of the residuals, and the weights can be chosen so that the estimator has a bounded influence function. Tableman (1990) defined a one-step rank-based estimator with bounded influence function.

In this article we consider estimates based on signed rank statistics, but the scores are not necessarily nondecreasing as is usually assumed. More precisely, define the estimate $\hat{\theta}_n$ as any solution of

$$\hat{\theta}_n = \arg\min_{\theta} D_n(\mathbf{Y}_n - \theta \mathbf{X}_n), \tag{4}$$

where $\mathbf{Y}_n = (y_1, \ldots, y_n)$ and $\mathbf{X}_n = (\mathbf{x}_1', \ldots, \mathbf{x}_n')$ is a $p \times n$ matrix. The objective function D_n is defined as

$$D_n(\mathbf{Y}_n - \theta \mathbf{X}_n) = \frac{1}{n} \sum_{i=1}^{n} a_n(R_{ni}^+)|r_i|, \tag{5}$$

where $r_i = r_i(\theta) = y_i - \theta \mathbf{x}_i'$ and $R_{ni}^+ = R_{ni}^+(\theta)$ is the rank of $|r_i|$ among $|r_1|, \ldots, |r_n|$. The numbers $a_n(i)$ are scores, usually chosen according to $a_n(i) = h^+(i/(n+1))$, with $h^+: (0, 1) \to [0, \infty)$ a given score generating function. We are interested in functions h^+ such that

$$\sup\{u; h^+(u) > 0\} = \alpha, \tag{6}$$

* Ola Hössjer is Research Associate, Department of Mathematical Statistics, Lund Institute of Technology, S-22100 Lund, Sweden. This work was supported by the Swedish Board for Technical Development under Contract 712-89-1073. The author thanks Svante Janson for valuable remarks, especially in connection with Lemma B.2; Uwe Rösler for enlightening discussions, and the associate editor and a referee for suggestions that essentially improved the appearance of the article.

© 1994 American Statistical Association
Journal of the American Statistical Association
March 1994, Vol. 89, No. 425, Theory and Methods

This content downloaded from 130.237.165.40 on Wed, 03 Feb 2016 22:26:14 UTC
All use subject to JSTOR Terms and Conditions

with $0 < \alpha \leq 1$. With such a choice of h^+, a proportion $1 - \alpha$ of the residuals with largest absolute values do not contribute to D_n. In Section 2 we show that $\varepsilon^* = \min(\alpha, 1 - \alpha)$ for $\{\hat{\theta}_n\}_{n \geq p}$ when (6) is satisfied and h^+ is sufficiently regular (e.g., right continuous). In particular, we obtain the maximal breakdown point when $\alpha = .5$.

The estimator of Jaeckel (1972), where the dispersion measure D_n to be minimized is based on the ranks of the residuals (not on their absolute values), is often used to estimate a regression parameter without an intercept. But by varying the scores of this estimator, we can not obtain a breakdown point higher than .25, because a larger fraction of outliers will carry D_n over all bounds if the corresponding residuals are all either large and positive or large and negative.

The asymptotic normality of $\hat{\theta}_n$ is well known when h^+ is nonnegative and nondecreasing. First, one introduces the derivative

$$\mathbf{S}_n(\theta) = (S_{n1}(\theta), \ldots, S_{np}(\theta)) = -\frac{\partial D_n(\mathbf{Y}_n - \theta\mathbf{X}_n)}{\partial\theta}$$
$$= \frac{1}{n}\sum_{i=1}^{n} a_n(R_{ni}^+)\operatorname{sgn}(r_i)\mathbf{x}_i, \quad (7)$$

which is a piecewise constant function of θ, making jumps when some residual equals 0 or when a tie occurs (i.e., the modulus of two residuals is the same). Next, one proves asymptotic normality for $\mathbf{S}_n(\theta_0)$ (cf. Håjek and Šidåk 1967) and uniform asymptotic linearity (in probability) of $\mathbf{S}_n(\theta)$ as a function of θ for local neighborhoods of θ_0 with diameters of size $O(n^{-1/2})$ (van Eeden 1972). This implies that $D_n(\mathbf{Y}_n - \theta\mathbf{X}_n)$ may be approximated by a quadratic function of θ locally around θ_0. Asymptotically, the argument of the minimum of this quadratic function is normally distributed and equivalent to the argument of a local minimum of $D_n(\mathbf{Y}_n - \theta\mathbf{X}_n)$. Finally, the convexity of $D_n(\mathbf{Y}_n - \theta\mathbf{X}_n)$ as a function of θ (McKean and Schrader 1980, thm. 2.1) implies that the local minimum of D_n is actually a global one. But for h^+ to satisfy (6), h^+ cannot be nondecreasing and $D_n(\mathbf{Y}_n - \theta\mathbf{X}_n)$ need not be convex; however, the preceding argument implies that we may find a sequence $\tilde{\theta}_n$ of estimates that locally minimize D_n and are asymptotically normal. This requires only that $h^+ = h_1^+ - h_2^+$, with both h_1^+ and h_2^+ nondecreasing and square integrable (cf. van Eeden 1972). To establish asymptotic normality (with the same asymptotic covariance matrix) for the global minimizer $\hat{\theta}_n$ of D_n, other methods are needed. We first prove that $\hat{\theta}_n$ is a consistent estimator of θ_0 (Sec. 3), and we then extend the asymptotic linearity of $\mathbf{S}_n(\theta)$ to neighborhoods of θ_0 of size $O(1)$ (Lem. 4.2). These results, together with the asymptotic normality of $\mathbf{S}_n(\theta_0)$, imply asymptotic normality of $\hat{\theta}_n$ (Thm. 4.1) under stronger regularity conditions on h^+.

The R estimator $\hat{\theta}_n$ in (4) and the LTS estimator can actually be put into a general framework by minimizing an objective function

$$\sum_{i=1}^{n} \rho(|r(\theta)|_{(i)})a_n(i), \quad (8)$$

where $|r(\theta)|_{(1)}, \ldots, |r(\theta)|_{(n)}$ are the order statistics for the absolute values of the residuals, $\rho : [0, \infty) \to [0, \infty)$ is a nondecreasing function, and the scores $a_n(i)$ generated from a function h^+ as before (6). The R estimator then corresponds to $\rho(x) = x$, and the LTS estimator corresponds to $\rho(x) = x^2$ and $h^+(u) = I(u \leq \alpha)$, given some trimming point α. This suggests that consistency and asymptotic normality may be proved for the more general model (8), with methods similar to those used in this article. Yohai and Maronna (1976) proved that local minimizers of the LTS objective function are asymptotically normal in the location model.

In Section 5 the efficiency of the proposed estimators is discussed; in Section 6 numerical examples are given. Finally, many of the proofs are collected in the Appendixes.

We close the section with some remarks on notation. The l_p norm, $1 \leq p \leq \infty$ of vectors in $\mathbb{R}^n$ is denoted $|\cdot|_p$, with $p = 2$ as a default value (omitting the subscript in this case). The L_p norm with respect to the Lebesgue measure of functions defined on a subset of $\mathbb{R}$ is denoted $\|\cdot\|_p$. C will refer to constants that may vary from line to line. Unless otherwise stated, these constants do not depend on other quantities (such as h^+, the underlying distribution of e_i and $\mathbf{x}_i$). In cases when C depends on such quantities we write $C = C(h^+)$, and so on. The integer part of the real number x is denoted $[x]$, and the smallest integer greater or equal than x is denoted $\lceil x \rceil$.

2. BREAKDOWN POINTS

In this section let n be fixed. Also assume that the scores $a_n(i)$ are nonnegative and define

$$k = \max\{i; a_n(i) > 0\}. \quad (9)$$

The following lemma gives the link between D_n in (5) and $|r|_{(k)}$.

Lemma 2.1. Assume that k is given by (9). Then there exist positive constants α and β such that $\alpha|r|_{(k)} \leq D_n \leq \beta|r|_{(k)}$. In particular, if $k = [n/2] + 1$, then we have $\alpha\text{median}\{|r_i|\} \leq D_n \leq 2\beta\text{median}\{|r_i|\}$.

Proof. Take $\alpha = a_n(k)/n$ and $\beta = 1/n \sum_{i=1}^{k} a_n(i)$.

For the rest of this section we assume that the regression data $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are positioned so that no more than $k - 1$ (where $k - 1 \geq p - 1$) of them lie on any vertical proper linear subspace of $\mathbb{R}^{p+1}$; that is, a subspace containing (0, 1). We then have the following (cf. Rousseeuw 1984, lem. 1):

Lemma 2.2. There always exists a solution to (4).

Proof. Let $M = \max\{y_i\}$, so that $D_n(\mathbf{Y}_n) \leq \beta M$ according to Lemma 2.1. Because no k data points are contained in a single vertical subspace of $\mathbb{R}^{p+1}$, it follows that

$$\inf_{|\theta|=1} |\theta\mathbf{x}'|_{(k)} = m > 0, \quad (10)$$

with $\{|\theta\mathbf{x}'|_{(i)}\}$ denoting the ordered $|\theta\mathbf{x}_i'|$. Combining (10), Lemma 2.1, and the fact that $|r_i(\theta)| \geq |\theta\mathbf{x}_i'| - M$ yields $D_n(\mathbf{Y}_n - \theta\mathbf{X}_n) \geq \alpha(|\theta|m - M) \geq 2\beta M$ whenever $|\theta| \geq (2(\beta/\alpha) + 1)M/m$. Because $D_n(\mathbf{Y}_n - \theta\mathbf{X}_n)$ is continuous

This content downloaded from 130.237.165.40 on Wed, 03 Feb 2016 22:26:14 UTC
All use subject to JSTOR Terms and Conditions

in $\boldsymbol{\theta}$, there exists a minimum inside a ball of radius $(2(\beta/\alpha) + 1)M/m$.

We say that the observations $\mathbf{z}_i$ are in general position whenever any p of them give a unique determination of $\boldsymbol{\theta}$. For such a sample we have the following result extending theorem 1 of Rousseeuw (1984):

Theorem 2.1. Let $\hat{\boldsymbol{\theta}}_n$ be an R estimator as defined in (4) and let $k \geq p$ be given by (9). Then for any sample $\mathbf{Z}_n$ in general position, we have $\varepsilon_n^*(\hat{\boldsymbol{\theta}}_n, \mathbf{Z}_n) = \min(n - k + 1, k - p + 1)/n$. In particular, if $k = [n/2] + 1$, we get $\varepsilon_n^* = ([n/2] - p + 2)/n$ when $p > 1$ and $[(n + 1)/2]/n$ when $p = 1$, the breakdown point of the LMS estimator.

Proof. See Appendix A.

Remark 2.1. It is clear that Lemma 2.2 and Theorem 2.1 still hold if Lemma 2.1 is replaced by the more general requirement $g_1(|r|_{(k)}) \leq D_n \leq g_2(|r|_{(k)})$, where g_1 and g_2 are strictly increasing functions with $g_1(0) = g_2(0) = 0$ and $g_1(\infty) = g_2(\infty) = \infty$. The L_2 distance $D_n = \sum r_i^2/n$ has $k = n$, $g_1(s) = s^2/n$, $g_2(s) = s^2$, and hence $\varepsilon_n^* = 1/n$. For the trimmed L_2 distance, $D_n = \sum_1^{n'} |r|_{(i)}^2/n$, k equals the trimming point n' $(p \leq n' \leq n)$, $g_1(s) = s^2/n$, and $g_2(s) = n's^2/n$. The L_1 norm, corresponding to $a_n(i) \equiv 1$ in (5), yields $k = n$, $g_1(s) = s/n$, $g_2(s) = s$, and $\varepsilon_n^* = 1/n$.

Remark 2.2. The breakdown point in Theorem 2.1 is maximized for $k = [(n+p)/2]$ or $k = \lceil (n+p)/2 \rceil$, and the corresponding breakdown point is $[(n - p)/2] + 1$, which is the maximal value among all regression equivariant estimators (see Rousseeuw 1984, rem. 1).

Remark 2.3. As a corollary of Theorem 2.1 we have the following *exact fit property:* If all observations are in general position and at least $n + 1 - n*\varepsilon_n^* = \max(k, n + p - k)$ of them satisfy $\mathbf{y} - \boldsymbol{\theta}\mathbf{x}'$ exactly for some $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}$ independently of the other observations. This may be shown directly, or it follows from Rousseeuw and Leroy (1987, p. 123).

Remark 2.4. If the scores are generated from a right-continuous (say) function h^+ satisfying (6), it follows that $k/n \to \alpha$ as $n \to \infty$, and hence by Theorem 2.1, $\varepsilon^*(\hat{\boldsymbol{\theta}}) = \min(\alpha, 1 - \alpha)$, where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_{p+1}, \ldots)$.

3. CONSISTENCY

The results of Section 2 were data-oriented in nature and assumed nothing about the distribution of the vectors $\mathbf{z}_i = (\mathbf{x}_i, y_i)$. Assume now that $\mathbf{z}_i$ are iid random vectors such that $\mathbf{x}_i$ and $e_i = y_i - \boldsymbol{\theta}_0\mathbf{x}_i'$ are independent with distributions $G\{d\mathbf{x}\}$ and $F\{dy\}$, and denote the distribution of $\mathbf{z}_i$ by $K\{d\mathbf{z}\}$. We also assume the following.

Assumption 1. The score-generating function h^+ is nonnegative and bounded with at most a finite number of discontinuities. Furthermore, (6) holds with $0 < \alpha \leq 1$.

In Assumptions 2 and 3, let $r = 1$ if either $\alpha < 1$ or $\alpha = 1$ and let $h^+(u) \leq C(1 - u)^\delta$ for some constant C and $\delta > 0$. Otherwise, choose r arbitrarily so that $r > 1$.

Assumption 2. $P_G(\boldsymbol{\theta}\mathbf{x}' = 0) < \alpha$ for all $\boldsymbol{\theta} \in \mathbb{R}^p$, $\boldsymbol{\theta} \neq 0$, and $E_G|\mathbf{x}|^r < \infty$.

Assumption 3. F has a density $f(e)$ that is even and strictly decreasing for positive values of e and $E_F|e|^r < \infty$.

To prove consistency for $\hat{\boldsymbol{\theta}}_n$, we start by proving that $\hat{\boldsymbol{\theta}}_n$ is bounded almost surely.

Lemma 3.1. Suppose that Assumptions 1–3 are satisfied. Then there exists a constant $M < \infty$ such that the estimate $\hat{\boldsymbol{\theta}}_n$ defined in (4) satisfies $\overline{\lim}_{n\to\infty} |\hat{\boldsymbol{\theta}}_n| \leq M$ a.s.

Proof. See Appendix A.

We now come to the main theorem of this section.

Theorem 3.1. Given that Assumptions 1–3 hold, $\hat{\boldsymbol{\theta}}_n$ is a strongly consistent estimate of $\boldsymbol{\theta}_0$.

Proof. See Appendix A. □

4. ASYMPTOTIC NORMALITY

To prove asymptotic normality for the estimate $\hat{\boldsymbol{\theta}}_n$ defined by (4), we impose some additional regularity conditions that, together with the assumptions made in Section 3, will be used throughout this section. It will sometimes be convenient to work with

$$h(u) = h^+(2u - 1), \qquad \frac{1}{2} \leq u < 1,$$
$$= -h^+(1 - 2u), \qquad 0 < u < \frac{1}{2} \tag{11}$$

instead of h^+.

Assumption 4. The function h defined in (11) is absolutely continuous on (0, 1) with $\|h'\|_\infty < \infty$. Moreover, h' has at most a finite number of discontinuities, outside which h'' exists, is continuous, and is bounded.

Assumption 5. The pdf of the error distribution f is absolutely continuous with finite Fisher information $I(f) = \int_{-\infty}^{\infty} f'(x)^2/f(x)\,dx$, and its derivative f' is bounded.

Assumption 6. $E_G|\mathbf{x}|^4 < \infty$.

Next we introduce some additional quantities. Put $\boldsymbol{\Sigma} = E_G(\boldsymbol{x}'\mathbf{x})$, where $\boldsymbol{\Sigma}$ is nonsingular because of Assumption 2, and let

$$A(h) = \int_0^1 h(u)^2\,du \tag{12}$$

and

$$B(h, F) = \int_0^1 h(u)h_F(u)\,du$$
$$= -\int_{-\infty}^{\infty} h(F(y))f'(y)\,dy, \tag{13}$$

with $h_F(u) = -f'(F^{-1}(u))/f(F^{-1}(u))$. Note that $B(h, F) > 0$, because of Assumptions 1 and 3. We will assume that $\boldsymbol{\theta}_0 = 0$ throughout this section (without loss of generality). Define the symmetric distribution function

This content downloaded from 130.237.165.40 on Wed, 03 Feb 2016 22:26:14 UTC
All use subject to JSTOR Terms and Conditions

$$H_\theta(t) = \frac{1}{2}(P_K(r_i(\theta) \le t) + P_K(-r_i(\theta) \le t))$$
$$= \frac{1}{2}(E_G F(t + \mathbf{x}\cdot\theta) + E_G F(t - \mathbf{x}\cdot\theta)) \quad (14)$$

(where the last identity holds, because f is symmetric) and put

$$F_{\theta,j}(t) = \frac{1}{2} E_G(x_j F(t + \mathbf{x}\cdot\theta)), \qquad j = 1, \ldots, p, \quad (15)$$

where $\mathbf{x} = (x_1, \ldots, x_p)$. Finally, $\lambda(\theta) = (\lambda_1(\theta), \ldots, \lambda_p(\theta))$ is defined by

$$\lambda_j(\theta) = 2\int_{-\infty}^{\infty} h(H_\theta(t))\, dF_{\theta,j}(t). \quad (16)$$

The vector $\lambda(\theta)$ is related to the derivative $\mathbf{S}_n(\theta)$ of D_n [cf. (7)], because it follows from Corollary B.1 that $\mathbf{S}_n(\theta) \xrightarrow{p} \lambda(\theta)$ as $n \to \infty$. We start with the following preliminary lemma.

Lemma 4.1. Under Assumptions 1–6, $\lambda(0) = 0$ and $\lambda'(0) = -B(h, F)\mathbf{\Sigma}$.

Proof. See Appendix B.

The following lemma proves asymptotic linearity in probability of $\mathbf{S}_n(\theta)$ as a function of θ uniformly over small enough neighborhoods of 0. The lemma is crucial for proving asymptotic normality of $\hat{\theta}_n$, and the argument is similar to that of lemma 3 of Huber (1967).

Lemma 4.2. Put

$$Z_n(\tau, \theta) = \frac{|\mathbf{S}_n(\tau) - \mathbf{S}_n(\theta) - \lambda(\tau) + \lambda(\theta)|}{n^{-1/2} + |\lambda(\tau)|}.$$

Then for small enough $\delta > 0$, $\sup_{|\tau|\le\delta} Z_n(\tau, 0) \xrightarrow{p} 0$ as $n \to \infty$.

Proof. See Appendix B.

We are now ready to prove asymptotic normality for the estimate $\hat{\theta}_n$.

Theorem 4.1. The estimate $\hat{\theta}_n$ defined by (4) is asymptotically normal in the sense that $n^{1/2}\hat{\theta}_n \xrightarrow{d} N(0, A\mathbf{\Sigma}^{-1}/B^2)$, with $A = A(h)$ and $B = B(h, F)$ given in (12) and (13), and $\mathbf{\Sigma}$ as defined after Assumption 6.

Proof. See Appendix B.

5. EFFICIENCY

As we see from Theorem 4.1, the R estimate $\hat{\theta}_n$ has the optimal rate of convergence $n^{-1/2}$, and the asymptotic efficiency relative to the Cramer–Rao lower bound is [cf. (12), (13) and Assumption 5]

$$e(h, F) = \frac{B(h, F)^2}{A(h)I(f)}. \quad (17)$$

It is well known that $h = h_F$ yields $e = 1$, and thus an asymptotically optimal estimate. To see how much efficiency is lost by imposing (6), we maximize formally the expression (17) subject to this constraint. The maximal value of $e(h, F)$ with F fixed (and f symmetric), is attained by the function $h_{F,\alpha}(u) = h_F(u)I((1-\alpha)/2 \le u \le (1+\alpha)/2)$, or equivalently [cf. (11)], $h^+_{F,\alpha}(u) = h_F((u+1)/2)I(0 \le u \le \alpha)$ (see, for example, Hampel, Ronchetti, Rousseeuw, and Stahel 1986, sec. 2.6c). Because $h_{F,\alpha}$ has two discontinuities, Assumption 4 is violated. But the supremum of $e(h, F)$ over all functions h satisfying Assumption 4 and (6), after the transformation (11), equals $e(h_{F,\alpha}, F)$, so this number certainly has a significance. Moreover, even when $h = h_{F,\alpha}$, a sequence of *local* minima of $D_n(\mathbf{Y}_n - \theta\mathbf{X}_n)$ is asymptotically normal, with the same asymptotic covariance matrix as in Theorem 4.1 (see Sec. 1).

Table 1 shows values of $e(h_{F,\alpha}, F)$ for normal, Laplace, and Cauchy distributions. As we can see, the loss in efficiency for a given breakdown point is smaller for the Cauchy distribution, which has heavier tails. For normally distributed errors, the efficiencies in Table 1 are the same as for the LTS estimator, and in this case

$$h^+_{F,\alpha}(u) = \Phi^{-1}\left(\frac{u+1}{2}\right)I(0 \le u \le \alpha), \quad (18)$$

where Φ is the standard normal distribution function.

What are the advantages of our estimator compared to LMS, LTS, and S? In terms of asymptotic efficiency, LMS is inferior because of the $n^{1/3}$-rate of convergence, whereas the S estimator is preferable in this aspect. Our estimator and LTS have intermediate performance for normal errors, with a rather low efficiency, as can be seen from Table 1. But the efficiency may be improved by computing a one-step M estimator based on a high breakdown initial estimator.

What about finite sample efficiencies? Stefanski (1991) and Morgenthaler (1991) have shown that high breakdown estimators may have arbitrarily low efficiency for certain configurations of design vectors. The reason is that local linear trends with different slope than the global linear trend of the data may be detected by high breakdown estimators. For higher values of ε^*, the probability for this to happen is larger. This problem with the finite-sample efficiency is an unavoidable price one has to pay for the high breakdown point. It thus may be advisable to choose a breakdown point of .20–.30 for small sample sizes. Figure 1 exhibits an artificial data set similar to the one of Stefanski (1991, fig. 1). The local trend consists of five out of nine points, and the R estimator is changed dramatically when the trimming point k [cf. (9)] is changed from five to six.

The asymptotic behavior of $\hat{\theta}_n$ indicates that the ith residual $r_i(\theta)$ has variance

Table 1. Values of the maximal efficiency $e(h_{F,\alpha}, F)$ for different trimming proportions α, breakdown points ε^* and distributions F

α	ε^*	Normal	Laplace	Cauchy
0.5	0.5	0.07	0.50	0.50
0.6	0.4	0.13	0.60	0.69
0.7	0.3	0.22	0.70	0.85
0.8	0.2	0.35	0.80	0.95
0.9	0.1	0.56	0.90	0.99
1.0	0.0	1.00	1.00	1.00

This content downloaded from 130.237.165.40 on Wed, 03 Feb 2016 22:26:14 UTC
All use subject to JSTOR Terms and Conditions

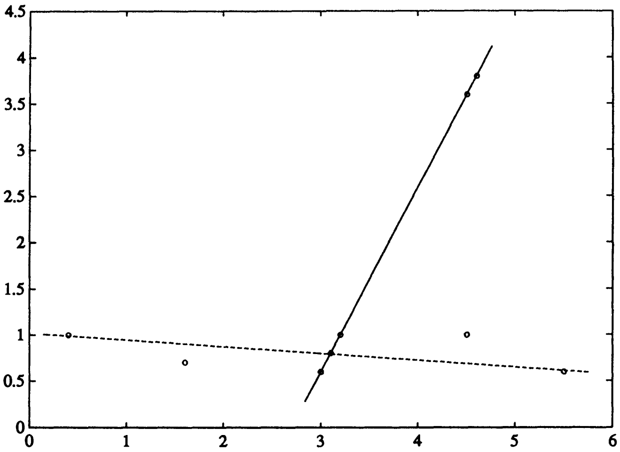

Figure 1. Data Configuration With Different Local and Global Trends. The artificial data set consists of nine points. The two regression lines are computed by means of an R estimator with trimmed normal scores and truncation points k = 5 (solid line) and k = 6 (dashed line).

$$\mathrm{var}(r_i(\hat{\theta}_n)) \approx \sigma^2(1 - Kh_{ii}), \qquad (19)$$

up to approximations of order n^{-1}. Here $\sigma^2 = \mathrm{var}_F(e)$, h_{ii} is the ith diagonal element of the hat matrix $\mathbf{X}_n'(\mathbf{X}_n\mathbf{X}_n')^{-1}\mathbf{X}_n$, and

$$K = \frac{2B(h, F)E_F(eh(F(e))) - A(h)}{\sigma^2 B(h, F)^2}.$$

(See McKean, Sheather, and Hettmansperger 1990 for a derivation of the corresponding result for linear rank statistics.) Observe that (19) holds exactly for least squares, with $K = 1$. For a data set with one extreme leverage point, the corresponding h_{ii} is large, and the variance of the residual is small. For least squares this can be explained by the fact that the LS estimator tries to fit the leverage point. In the same way, the variance of the R estimator residual $r_i(\hat{\theta}_n)$ should be small for a leverage point, because of (19). On the other hand, we know that the R-estimator $\hat{\theta}_n$ will not be influenced by a bad leverage point when ε^* is large enough, and the corresponding residual thus will be large. This seems contradictive. However, formula (19) is conditional on $\mathbf{x}_1, \ldots, \mathbf{x}_n$. When the design vectors are given and $\mathbf{x}_i$ is outlying, the probability is small that $(\mathbf{x}_i, y_i)$ is a bad leverage point, whereas the probability is large that $(\mathbf{x}_i, y_i)$ is a good leverage point. Because a good leverage point $(\mathbf{x}_i, y_i)$ will have large influence on $\hat{\theta}_n$, resulting in a small residual r_i, the overall variance of $r_i(\hat{\theta}_n)$ becomes small.

6. NUMERICAL EXAMPLES

In the numerical examples we use a grid search algorithm for simple linear regression and the PROGRESS algorithm (cf. Rousseeuw and Leroy 1987, chap. 5), for higher dimensions. The latter algorithm computes an approximation of the true R estimate, by modifying (4), so that the minimization is performed over a finite set of $\boldsymbol{\theta}$ values. When $\{\mathbf{z}_i\}$ are in general position, these regression parameters correspond to all (or a random subsample of) $\le \binom{n}{p}$ possible hyperplanes determined by exact fits from p data points. Because the computation of $D_n(\mathbf{Y}_n - \boldsymbol{\theta}\mathbf{X}_n)$ requires ordering of $\{|r_i(\boldsymbol{\theta})|\}$, the computation time for this estimator is of the same order as for (the PROGRESS version of) the LTS estimator. An improvement of the PROGRESS algorithm has been considered by Ruppert (1992), where at each step the objective function is evaluated at a regression parameter that is a convex combination of the current best estimate and the last exact fit. In this way the search for the regression estimate is concentrated at the region around the current best estimate.

We use the trimmed normal scores (18) for all the R estimators in the examples; that is,

$$a_n(i) = \Phi^{-1}\left(\frac{i + n + 1}{2(n + 1)}\right), \qquad i = 1, \ldots, k,$$
$$= 0, \qquad i = k + 1, \ldots, n,$$

where k is the trimming point. Our simulations indicate that these R estimators have very similar performance to that of the LTS estimator with the same trimming point. Indeed, the asymptotic efficiency is the same for normally distributed errors. But there are some data configurations for which the R estimator performs better. An example is given in Figure 2, with $n = 13$ and $k = 7$. The six points in the lower right corner are recognized as outliers by both estimators. The LTS estimator will act as a LS estimator on the remaining seven points, whereas the high breakdown point R estimator will act as a traditional normal scores R estimator. The LS estimator is more sensitive to vertical outliers than is the normal scores R estimator. This explains why the LTS line is more influenced by the vertical outlier in Figure 2.

Figure 3 displays the stars data (Rousseeuw and Leroy 1987, p. 27), with three different R-estimation fits: $\alpha = .5$, .7, and .9. We see clearly that the former two regression lines are not influenced by the outliers in the upper left corner.

Next we give a multiple linear regression example with several leverage points, the Hawkins–Bradu–Kass data,

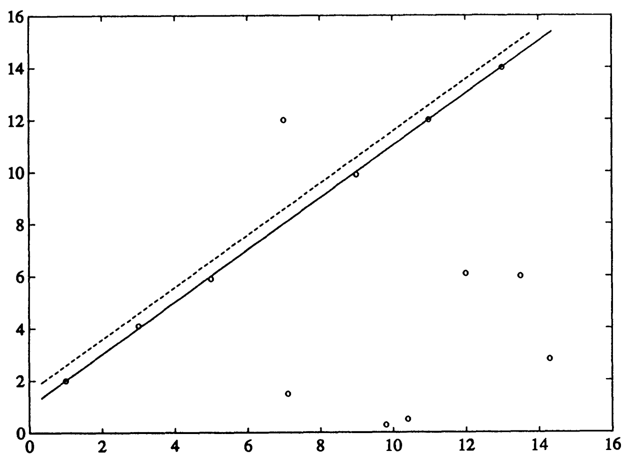

Figure 2. Comparison of the R Estimator and LTS Estimator. The artificial data set consists of 13 points. The two regression lines are computed by means of an R estimator with trimmed normal scores, k = 7 (solid line), and an LTS estimator with k = 7 (dashed line).

This content downloaded from 130.237.165.40 on Wed, 03 Feb 2016 22:26:14 UTC
All use subject to JSTOR Terms and Conditions

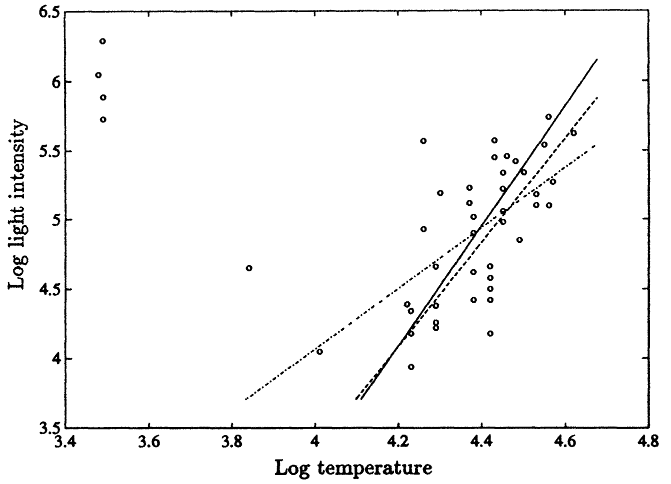

Figure 3. The Stars Data. The data consist of three different R estimates, based on trimmed normal scores. The trimming proportions α are .5 (solid line), .7 (dashed line), and .9 (dashed-dotted line).

which has three explanatory variables and an intercept. The first ten observations are known to be bad leverage points, and the next four are good leverage points (cf. Rousseeuw and Leroy 1987, p. 93). Figure 4 shows a residual plot for the R estimator when $\alpha = .5$. We have standardized the residuals by the median of the absolute residuals, $\hat{s} = 1.483(1 + 5/(n - p))\text{median}_{1 \le i \le n} |r_i(\hat{\theta})|$, where the multiplicative factor 1.483 makes $\hat{s}$ a consistent estimate of the standard deviation for normal errors, with $1 + 5/(n - p)$ a finite sample correction factor. Alternatively, we could have used $D_n(\mathbf{Y}_n - \hat{\theta}\mathbf{X}_n)$, the minimal value of the objective function (properly standardized) as a residual scale estimate. But then each α requires a separate multiplicative constant. We see from Figure 4 that the R estimator manages to identify all the bad leverage points, but does not flag the good leverage points as outliers.

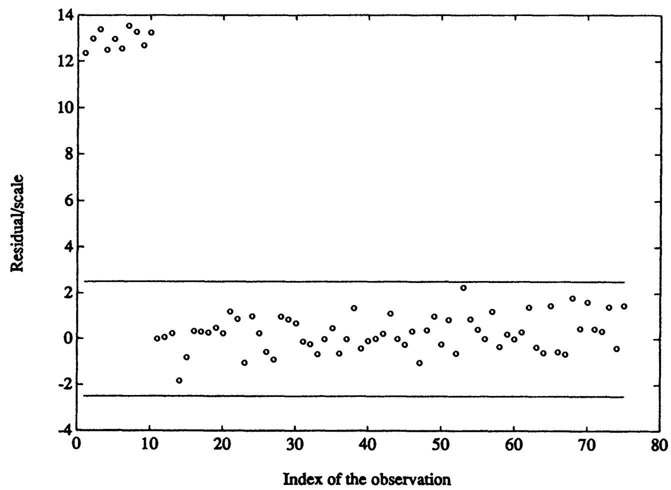

Figure 4. Multiple Linear Regression with p = 4. This figure is a residual plot for the Hawkins–Bradu–Kass data, using an R estimator with trimmed normal scores and trimming proportion α = .5. The number of random p subsets in the PROGRESS algorithm is 10,000.

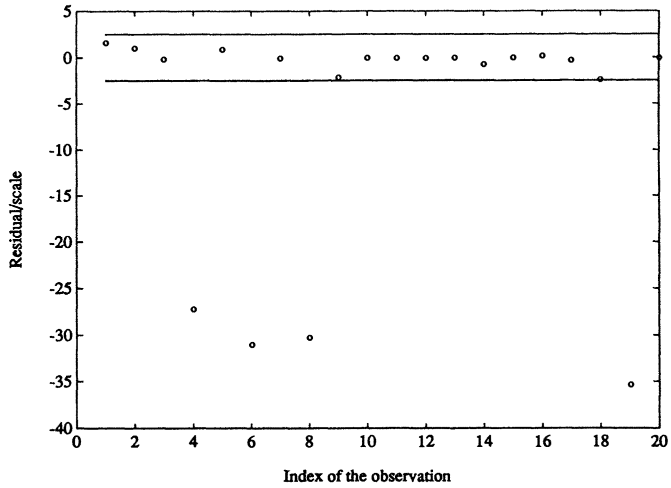

Figure 5. Multiple Linear Regression with p = 6. This figure is a residual plot for the modified wood specific gravity data, using an R estimator with trimmed normal scores and trimming proportion α = .5. The number of random p subsets in the PROGRESS algorithm is 10,000.

Finally, we consider the modified wood specific gravity data, described by Rousseeuw and Leroy (1987, p. 243). The model contains five explanatory variables and an intercept and describes the influence of anatomical factors on wood specific gravity. The observations with indicies 4, 6, 8, and 19 have been replaced by outliers. Looking at the residual plot in Figure 5, we see that the R estimator with a 50% breakdown point manages to identify these four points.

APPENDIX A: PROOFS FROM SECTIONS 2 AND 3

Proof of Theorem 2.1. In theorem 1 of Rousseeuw (1984), ε_n^* is determined for the LMS estimator. It is clear from the proof of this theorem and Lemma 2.1 that ε_n^* must have the same value for $\hat{\theta}_n$ when $k = [n/2] + 1$. The first part in the proof of Rousseeuw's theorem can be extended directly to prove $\varepsilon_n^*(\hat{\theta}_n, \mathbf{Z}_n) \ge \min(n - k + 1, k - p + 1)/n$, and the second part can be extended to prove $\varepsilon_n^*(\hat{\theta}_n, \mathbf{Z}_n) \le (k - p + 1)/n$. It remains to show

$$\varepsilon_n^*(\hat{\theta}_n, \mathbf{Z}_n) \le (n - k + 1)/n. \tag{A.1}$$

For this purpose assume that a subsample $\mathbf{W}_n \subset \mathbf{Z}_n$ with $n - k + 1$ data points is replaced by $\mathbf{W}_n^*$. Denote the new sample $\mathbf{Z}_n^*$, consisting of the elements $\mathbf{z}_1^* = (\mathbf{x}_1^*, y_1^*), \ldots, \mathbf{z}_n^* = (\mathbf{x}_n^*, y_n^*)$. Choose $\mathbf{W}_n^*$ so that all its elements belong to $H_{\theta^*} = \{(\mathbf{x}\ y);\ y - \theta^* \cdot \mathbf{x} = 0\}$, where θ^* will be chosen later. Let $r_i^*(\theta)$ denote the residuals of the new sample and let $\hat{\theta}_n$ and $\hat{\theta}_n^*$ be any two vectors minimizing D_n for the old and new samples. We intend to show that given any $b > 0$, it is possible to choose $\mathbf{W}_n^*$ and θ^* so that $|\hat{\theta}_n^* - \hat{\theta}_n| > b$. To this end, pick θ^* so that $|\theta^* - \hat{\theta}_n| > b$ (but otherwise arbitrarily) and let $\mathbf{x}_i^* = M(\theta^* - \hat{\theta}_n)$ for all data points in $\mathbf{W}_n^*$. We will show that if $M > 0$ is large enough, then $D_n(\mathbf{Y}_n^* - \theta^*\mathbf{X}_n^*) < D_n(\mathbf{Y}_n^* - \tilde{\theta}\mathbf{X}_n^*)$ for any $\tilde{\theta} \in B(\hat{\theta}_n, b)$ — the closed ball in $\mathbb{R}^p$ of radius b (with respect to the L_2 norm) around $\hat{\theta}_n$. This will imply that $\hat{\theta}_n^*$ lies outside this ball. Here $\mathbf{Y}_n^* = (y_1^*, \ldots, y_n^*)$ and $\mathbf{X}_n^* = (\mathbf{x}_1^{*\prime}, \ldots, \mathbf{x}_n^{*\prime})$.

First, notice that

$$\begin{aligned} |r^*(\theta^*)|_{(k)} &\le \max_{\mathbf{z}_i^* \in \mathbf{Z}_n^* \setminus \mathbf{W}_n^*} |y_i^* - \theta^* \cdot \mathbf{x}_i^*| \\ &\le \max_{\mathbf{z}_i \in \mathbf{Z}_n} |y_i| + |\theta^*| \max_{\mathbf{z}_i \in \mathbf{Z}_n} |\mathbf{x}_i|. \end{aligned} \tag{A.2}$$

This content downloaded from 130.237.165.40 on Wed, 03 Feb 2016 22:26:14 UTC
All use subject to JSTOR Terms and Conditions

Because $|\mathbf{W}_n^*| = n - k + 1$, it follows that

$$|r^*(\tilde{\theta})|_{(k)} \ge \min_{\mathbf{z}_i^* \in \mathbf{W}_n^*} |y_i^* - \tilde{\theta}\cdot \mathbf{x}_i^*|$$

$$= \min_{\mathbf{z}_i^* \in \mathbf{W}_n^*} |(\tilde{\theta} - \theta^*)\cdot \mathbf{x}_i^*|$$

$$= M|(\tilde{\theta} - \theta^*)\cdot(\theta^* - \hat{\theta}_n)|$$

$$\ge M(|\theta^* - \hat{\theta}_n|^2 - b|\theta^* - \hat{\theta}_n|) \quad \text{(A.3)}$$

for any $\tilde{\theta} \in B(\hat{\theta}_n, b)$. Notice that the upper bound in (A.2) is independent of M, whereas the lower bound in (A.3) can be made arbitrarily large by increasing M. Together with Lemma 2.1, this implies that $\inf_{\tilde{\theta}\in B(\hat{\theta}_n,b)} D_n(\mathbf{Y}_n^* - \tilde{\theta}\mathbf{X}_n^*) > D_n(\mathbf{Y}_n^* - \theta^*\mathbf{X}_n^*)$ for large enough M. Hence $\hat{\theta}_n^*$ must lie outside $B(\hat{\theta}_n, b)$. Because b was arbitrary, (A.1) is proved.

Remark A.1. The proof of (A.1) was based on positioning $n - k + 1$ data points ($\mathbf{W}_n^*$) in a way that has probability 0. But because D_n is a continuous function of the residuals, we may easily construct a neighborhood of W_n^* with positive probability such that $|\hat{\theta}_n^* - \hat{\theta}_n| > b$ still holds.

Proof of Lemma 3.1. Put $\theta_0 = 0$ without loss of generality (because $\hat{\theta}_n$ is regression equivariant). We introduce a random variable $\tilde{M}(\mathbf{Z}_n)$ such that

$$|\hat{\theta}_n| \le \tilde{M}(\mathbf{Z}_n) \quad \text{(A.4)}$$

and show that $\overline{\lim}_{n\to\infty} \tilde{M}(\mathbf{Z}_n) \le M$ a.s. Let $\varepsilon_n = \varepsilon_n(\mathbf{X}_n) = \min_{|\theta|=1} D_n(\theta\mathbf{X}_n)$. Because D_n is continuous as a function of θ, ε_n is a well-defined random variable. It follows from Lemma A.1 that

$$\frac{1}{n}|\mathbf{Y}_n|_1 \le \frac{\varepsilon_n}{2\|h^+\|_\infty} \Rightarrow |D_n(-\theta\mathbf{X}_n) - D_n(\mathbf{Y}_n - \theta\mathbf{X}_n)| \le \frac{1}{2}\varepsilon_n.$$

Because D_n is scale equivariant, we thus have

$$D_n(\mathbf{Y}_n - \theta\mathbf{X}_n) = |\theta|\, D_n\!\left(\frac{\mathbf{Y}_n}{|\theta|} - \frac{\theta}{|\theta|}\mathbf{X}_n\right) \ge |\theta|\varepsilon_n/2 > D_n(\mathbf{Y}_n)$$

whenever

$$|\theta| > \max\!\left(\frac{2D_n(\mathbf{Y}_n)}{\varepsilon_n}, \frac{2\|h^+\|_\infty |\mathbf{Y}_n|_1}{n\varepsilon_n}\right) \triangleq \tilde{M}(\mathbf{Z}_n). \quad \text{(A.5)}$$

In particular, with this choice of $\tilde{M}(\mathbf{Z}_n)$, (A.4) holds. By the strong law of large numbers, $|\mathbf{Y}_n|_1/n \overset{\text{a.s.}}{\to} E_F(|e|) < \infty$. Because $D_n(\cdot)$ is a L statistic, we may apply results of Wellner (1977) to obtain the limiting behavior of $D_n(\cdot)$. It is easy to see from Wellner's theorem 3-4 and from example 1, which covers most of our regularity conditions, that Assumptions 1–3 are sufficient to guarantee that

$$D_n(\mathbf{Y}_n - \theta\mathbf{X}_n) \overset{\text{a.s.}}{\to} \mu(\theta) = E_U(\tilde{F}_\theta^{-1}(u)h^+(u)) < \infty \quad \text{(A.6)}$$

for all θ, with $\tilde{F}_\theta$ the distribution of $|y_i - \theta\mathbf{x}_i'|$ and U denoting a uniform distribution on (0, 1). In view of (A.5), it thus remains to show that

$$\underline{\lim}_{n\to\infty} \varepsilon_n \ge m \quad \text{a.s.} \quad \text{(A.7)}$$

for some $m > 0$. As in (A.6) we have $D_n(\theta\mathbf{X}_n) \overset{\text{a.s.}}{\to} \tilde{m}(\theta) = E_U(\tilde{G}_\theta^{-1}(u)h^+(u)) < \infty$, with $\tilde{G}_\theta$ the distribution of $|\theta\mathbf{x}_i'|$. It follows from Assumptions 1 and 2 that $\tilde{m}(\theta) > 0$ for all $\theta \ne 0$. It is clear that $\tilde{m}(\theta)$ is continuous in θ, and hence $\underline{m} = \inf_{|\theta|=1} \tilde{m}(\theta) > 0$. Let $\eta = \underline{m}/(2\|h^+\|_\infty E_G|\mathbf{x}|_1)$ and pick $\theta_1, \ldots, \theta_N$, $N = N(\eta)$ from the unit sphere so that $\sup_{|\theta|=1} \min_{1\le i\le N(\eta)} |\theta - \theta_i|_\infty \le \eta$. Hence for any θ, $|\theta| = 1$ we may choose θ_j, $j = j(\theta)$ from these vectors so that $|\theta - \theta_j|_\infty \le \eta$. It then follows from Lemma A.1 that

$$D_n(\theta\mathbf{X}_n) \ge D_n(\theta_j\mathbf{X}_n) - \|h^+\|_\infty \frac{|(\theta - \theta_j)\mathbf{X}_n|_1}{n}$$

$$\ge \min_{1\le i\le N} D_n(\theta_i\mathbf{X}_n) - \frac{\|h^+\|_\infty\eta}{n}\sum_{i=1}^n |\mathbf{x}_i|_1 \overset{\text{a.s.}}{\to} \min_{1\le i\le N} \tilde{m}(\theta_i)$$

$$- \|h^+\|_\infty \eta E_G|\mathbf{x}|_1$$

$$\ge \underline{m} - \|h^+\|_\infty \eta E_G|\mathbf{x}|_1 = \underline{m}/2. \quad \text{(A.8)}$$

Because the lower bound of $D_n(\theta\mathbf{X}_n)$ in (A.8) is independent of θ, it follows that (A.7) holds with $m = \underline{m}/2$.

Proof of Theorem 3.1. Put $\theta_0 = 0$ without loss of generality. In view of Lemma 3.1, it suffices to show

$$P(m < |\hat{\theta}_n| < M \text{ i.o.}) = 0 \quad \text{(A.9)}$$

for any $0 < m < M$. Recall that $D_n(\mathbf{Y}_n - \theta\mathbf{X}_n) \overset{\text{a.s.}}{\to} \mu(\theta)$, with $\mu(\theta)$ defined in (A.6). We claim that $\mu(\theta) > \mu(0)$ for any $\theta \ne 0$. This is so because for any $t > 0$ and $\theta \ne 0$,

$$\tilde{F}_\theta(t) = P_K(|e - \theta\mathbf{x}'| \le t) = E_G P_F(|e - \theta\mathbf{x}'| \le t|\mathbf{x})$$

$$< E_G P_F(|e| \le t) = \tilde{F}_0(t)$$

according to Assumptions 2 and 3, and hence $\tilde{F}_\theta^{-1}(u) > \tilde{F}_0^{-1}(u)$ for all $0 < u < 1$. It is also obvious that $\mu(\theta)$ is continuous in θ and hence $\underline{\mu} = \min_{m\le|\theta|\le M} \mu(\theta) > \mu(0)$. A construction similar to the one in Lemma 3.1 for proving (A.7) gives

$$\underline{\lim}_{n\to\infty} \min_{m\le|\theta|\le M} D_n(\mathbf{Y}_n - \theta\mathbf{X}_n) \ge \frac{\mu(0) + \underline{\mu}}{2} \quad \text{a.s.,}$$

which implies (A.9).

The following lemma is needed in the proofs of Lemma 3.1 and Theorem 3.1.

Lemma A.1. Let $\mathbf{u}$ and $\mathbf{v}$ be vectors in $\mathbb{R}^n$. Then $|D_n(\mathbf{u}) - D_n(\mathbf{v})| \le \|h^+\|_\infty |\mathbf{u} - \mathbf{v}|_1/n$.

Proof. Put $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n)$ and let $\{|u|_{(i)}\}$ and $\{|v|_{(i)}\}$ denote the order statistics corresponding to the absolute values of the components of each vector. Then

$$|D_n(\mathbf{u}) - D_n(\mathbf{v})| = \left|\frac{1}{n}\sum_{i=1}^n a_n(i)|u|_{(i)} - \frac{1}{n}\sum_{i=1}^n a_n(i)|v|_{(i)}\right|$$

$$\le \frac{\|h^+\|_\infty}{n}\sum_{i=1}^n \big||u|_{(i)} - |v|_{(i)}\big|$$

$$\overset{*}{\le} \frac{\|h^+\|_\infty}{n}\sum_{i=1}^n \big||u_i| - |v_i|\big|$$

$$\le \frac{\|h^+\|_\infty}{n}|\mathbf{u} - \mathbf{v}|_1,$$

where the inequality marked "*" follows from Cambanis, Simons, and Stout (1976, thm. 2).

APPENDIX B: PROOFS FROM SECTION 4

Proof of Lemma 4.1. First, the symmetry of f and the skew-symmetry of h around $u = \frac{1}{2}$ gives $\lambda(0) = 0$. Next, the regularity conditions admit us to differentiate under the integral sign with respect to θ_k in (16). This yields (with $f_{\theta,j}(t) = dF_{\theta,j}(t)/dt$)

$$\left[\frac{\partial\lambda_j(\theta)}{\partial\theta_k}\right]_{\theta=0} = 2\int h'(F(t))\left[\frac{\partial H_\theta(t)}{\partial\theta_k}\right]_{\theta=0} f_{0,j}(t)\,dt$$

$$+ 2\int h(F(t))\left[\frac{\partial f_{\theta,j}(t)}{\partial\theta_k}\right]_{\theta=0} dt.$$

This content downloaded from 130.237.165.40 on Wed, 03 Feb 2016 22:26:14 UTC
All use subject to JSTOR Terms and Conditions

Differentiating under the integral sign again in (14) and (15) gives $[\partial H_\theta(t)/\partial\theta_k]_{\theta=0} = 0$ and $[\partial f_{\theta,j}(t)/\partial\theta_k]_{\theta=0} = \frac{1}{2}f'(t)E_G(x_jx_k)$. Inserting this into the preceding expression gives

$$\left[\frac{\partial\lambda_j(\theta)}{\partial\theta_k}\right]_{\theta=0} = \int h(F(t))f'(t)\,dtE_G(x_jx_k) = -B(h,F)E_G(x_jx_k),$$

which proves the lemma.

Proof of Lemma 4.2. We may replace the Euclidean norm $|\cdot|$ by the max norm $|\cdot|_\infty$ for vectors in $\mathbb{R}^p$. It obviously suffices to prove

$$\sup_{|\tau|_\infty\le\delta} Z_{nj}(\tau,0) \xrightarrow{p} 0, \qquad j=1,\ldots,p \tag{B.1}$$

as $n\to\infty$, where

$$Z_{nj}(\tau,\theta) = \frac{|S_{nj}(\tau)-S_{nj}(\theta)-\lambda_j(\tau)+\lambda_j(\theta)|}{n^{-1/2}+|\lambda(\tau)|_\infty}.$$

First, we need some preliminary estimates on $\lambda(\theta)$. It follows from Lemma 4.1 and the facts that Σ is nonsingular and $B(h,F)>0$ that

$$a|\tau|_\infty \le |\lambda(\tau)|_\infty \le a'|\tau|_\infty, \quad \text{whenever} \quad |\tau|_\infty < \delta, \tag{B.2}$$

for some constants $0<a<a'$, if δ is small enough. In the rest of the proof we assume, without loss of generality, that $\delta=1$. Next, observe that

$$|\lambda_j(\tau)-\lambda_j(\theta)| \le |\lambda_j(\tau)-S_{nj}(\tau)| + |\lambda_j(\theta)-S_{nj}(\theta)| + Q_{nj}(\theta,|\tau-\theta|_\infty), \tag{B.3}$$

where $Q_{nj}(\theta,d)$ is a quantity defined in the proof of Lemma B.1 satisfying

$$Q_{nj}(\theta,d) \ge \sup_{\{\tau;|\tau-\theta|_\infty\le d\}} |S_{nj}(\tau)-S_{nj}(\theta)|, \qquad j=1,\ldots,p. \tag{B.4}$$

Taking expectations on both sides of (B.3) and letting $n\to\infty$, it follows from Lemma B.1 and Corollary B.1 that

$$|\lambda_j(\tau)-\lambda_j(\theta)| \le b|\tau-\theta|_\infty, \tag{B.5}$$

with $b>0$ the same constant as in Lemma B.1. Following the construction of Huber (1967, lem. 3), we now subdivide the unit cube around the origin into a number of smaller cubes. Let $C_k = \{\theta; |\theta|_\infty \le (1-q)^k\}$, $k=0,1,\ldots$, where $q=1/M$ and M is a positive integer to be chosen. Subdivide C_0 as a disjoint (except boundaries) union of C_{k_0} and $C_{(1)},\ldots,C_{(N)}$, where each $C_{(i)}$ is a cube lying in $C_{k-1}\backslash C_k$ for some $k\le k_0$ with edges of length $q(1-q)^{k-1}$. There are at most $(2M)^p$ cubes $C_{(i)}$ in each $C_{k-1}\backslash C_k$. Given $\varepsilon>0$, we have

$$P(\sup_{\tau\in C_0} Z_{nj}(\tau,0)\ge 2\varepsilon) \le P(\sup_{\tau\in C_{k_0}} Z_{nj}(\tau,0)\ge 2\varepsilon) + \sum_{i=1}^N P(\sup_{\tau\in C_{(i)}} Z_{nj}(\tau,0)\ge 2\varepsilon). \tag{B.6}$$

The integer M is chosen sufficiently large that

$$\frac{q}{1-q} \le \frac{\varepsilon a}{3b}, \tag{B.7}$$

and $k_0 = k_0(n)$ is selected so that

$$(1-q)^{k_0} \le n^{-\gamma} < (1-q)^{k_0-1}, \tag{B.8}$$

where $\frac{1}{2}<\gamma<1$ is an arbitrary fixed number. Hence $k_0(n) = O(\log n)$ and $N = O(\log n)$.

We first estimate each term of the sum in (B.6). Suppose that $C_{(i)} \in C_{k-1}\backslash C_k$ has center ξ and side length $2d$. Then estimate $Z_{nj}(\tau,0)$ according to

$$Z_{nj}(\tau,0) \le \frac{|S_{nj}(\tau)-S_{nj}(\xi)-\lambda_j(\tau)+\lambda_j(\xi)|}{n^{-1/2}+|\lambda(\tau)|_\infty} + \frac{|S_{nj}(\xi)-S_{nj}(0)-\lambda_j(\xi)|}{n^{-1/2}+|\lambda(\tau)|_\infty}. \tag{B.9}$$

Because for each $\tau\in C_{(i)}$ we have $|\lambda_j(\tau)-\lambda_j(\xi)| \le bd \le bq(1-q)^{k-1}$ and $|\lambda(\tau)|_\infty \ge a(1-q)^k$ according to (B.2) and (B.5), it follows from (B.9) that

$$\sup_{\tau\in C_{(i)}} Z_{nj}(\tau,0) \le \frac{Q_{nj}(\xi,d)+bq(1-q)^{k-1}}{a(1-q)^k} + \frac{|S_{nj}(\xi)-S_{nj}(0)-\lambda_j(\xi)|}{n^{-1/2}+a(1-q)^k} \triangleq Z_{nj,1}+Z_{nj2}. \tag{B.10}$$

It then follows from Lemma B.1, (B.7), and (B.8) that

$$\begin{aligned} P(Z_{nj,1}\ge\varepsilon) &\le P(Q_{nj}(\xi,d)-EQ_{nj}(\xi,d) \\ &\ge \varepsilon a(1-q)^k - bq(1-q)^{k-1} - EQ_{nj}(\xi,d)) \\ &\le P(Q_{nj}(\xi,d)-EQ_{nj}(\xi,d) \\ &\ge bq(1-q)^{k-1}) \le \frac{cq(1-q)^{k-1}n^{-1}}{(bq(1-q)^{k-1})^2} \\ &= \frac{cn^{-1}}{b^2q(1-q)^{k-1}} = O(n^{\gamma-1}), \end{aligned} \tag{B.11}$$

with c the same constant as in Lemma B.1. To estimate $Z_{nj,2}$, we see from Lemma B.2 that $E(S_{nj}(\xi)-S_{nj}(0)-\lambda_j(\xi))^2 \le C(n^{-5/4}+|\xi|_\infty^2 n^{-1})$, where $C=C(h,F,G)$. Formula (B.10) thus gives

$$\begin{aligned} EZ_{nj,2}^2 &\le C\left(\frac{n^{-5/4}}{n^{-1}} + \frac{|\xi|_\infty^2 n^{-1}}{a^2(1-q)^{2k}}\right) \\ &\le Cn^{-1/4} + \frac{Cn^{-1}}{a^2(1-q)^2} = O(n^{-1/4}). \end{aligned}$$

It follows then from Chebyshev's inequality that

$$P(Z_{nj,2}\ge\varepsilon) = O(n^{-1/4}). \tag{B.12}$$

Putting things together, we get from (B.10)–(B.12) that

$$P(\sup_{\tau\in C_{(i)}} Z_{nj}(\tau,0)\ge 2\varepsilon) = O(n^{\max(-1/4,\gamma-1)}) \tag{B.13}$$

uniformly in i. It remains to estimate the first term in (B.6). Let now $2d = 2(1-q)^{k_0}$ be the side length of C_{k_0}. We then have $\sup_{\tau\in C_{k_0}} Z_{nj}(\tau,0) \le n^{1/2}(Q_{nj}(0,d)+a'd)$, with a' given in (B.2). Hence

$$P(\sup_{\tau\in C_{k_0}} Z_{nj}(\tau,0)\ge 2\varepsilon) \le P(Q_{nj}(0,d)-EQ_{nj}(0,d) \ge 2\varepsilon n^{-1/2} - EQ_{nj}(0,d)-a'd).$$

Because $EQ_{nj}(0,d)+a'd \le (b+a')d = (b+a')(1-q)^{k_0} = O(n^{-\gamma})$ according to Lemma B.1, it follows that $2\varepsilon n^{-1/2} - EQ_{nj}(0,d)-a'd \ge \varepsilon n^{-1/2}$ for all n exceeding some integer n_0. Thus $n\ge n_0$ yields

$$P(\sup_{\tau\in C_{k_0}} Z_{nj}(\tau,0)\ge 2\varepsilon) \le P(Q_{nj}(0,d)-EQ_{nj}(0,d)\ge\varepsilon n^{-1/2}) \le \frac{cdn^{-1}}{\varepsilon^2 n^{-1}} \le \frac{c(1-q)^{k_0}}{\varepsilon^2} = O(n^{-\gamma}). \tag{B.14}$$

Summarizing, (B.6), (B.13), (B.14), and the fact that $N = O(\log n)$ proves (B.1) and thus concludes the proof of the lemma.

Proof of Theorem 4.1. The following proof is reminiscent to that of theorem 3 of Huber (1967). By Lemma 4.1 and Theorem 3.1, it suffices to show $n^{1/2}\lambda(\hat\theta_n) \xrightarrow{d} N(0,A\Sigma)$ as $n\to\infty$. But because

This content downloaded from 130.237.165.40 on Wed, 03 Feb 2016 22:26:14 UTC
All use subject to JSTOR Terms and Conditions

$$n^{1/2}\mathbf{S}_n(0) \xrightarrow{d} N(0, A\Sigma) \quad (B.15)$$

(cf. Hàjek and Šidàk 1967, p. 166), we have only to show

$$n^{1/2}(\mathbf{S}_n(0) + \lambda(\hat{\theta}_n)) \xrightarrow{p} 0 \quad (B.16)$$

as $n \to \infty$. The fact that $\hat{\theta}_n$ is consistent implies that with probability tending to 1 as $n \to \infty$,

$$\frac{|\mathbf{S}_n(0) + \lambda(\hat{\theta}_n)|}{n^{-1/2} + |\lambda(\hat{\theta}_n)|} \le \sup_{|\theta| \le \delta} Z_n(\theta, 0) + n^{1/2} \overline{\lim_{\theta \to \hat{\theta}_n}} |\mathbf{S}_n(\hat{\theta}_n)|, \quad (B.17)$$

where $Z_n(\theta, 0)$ is defined in Lemma 4.2. The "limsup" in (B.17) takes into account the fact that $\mathbf{S}_n(\hat{\theta}_n)$ may not be well defined when a tie occurs or when some residual equals 0. For a fixed θ, $\mathbf{S}_n(\theta)$ is well defined with probability 1, because the error distribution is continuous, but at $\hat{\theta}_n$, $\mathbf{S}_n$ is typically *not* well defined. As $\hat{\theta}_n$ minimizes $D_n(\mathbf{Y}_n - \theta\mathbf{X}_n)$, either $\mathbf{S}_n(\hat{\theta}_n) = 0$ if it is well defined, or otherwise

$$n^{1/2} \overline{\lim_{\theta \to \hat{\theta}_n}} |S_{nj}(\hat{\theta}_n)| \le N \cdot 2\|h^+\|_\infty \max_{1 \le i \le n} \frac{|x_{ij}|}{n^{1/2}}, \qquad j = 1, \ldots p. \quad (B.18)$$

Here the second factor on the right side of (B.18) is an upper bound for the size of jumps that $S_{nj}(\theta)$ makes at each tie or change of signs. Assumption 6 implies that this factor tends to 0 in probability as $n \to \infty$ (actually, a finite second moment of $|\mathbf{x}|$ is enough). The first factor N denotes the maximal number of such occurrances at any point. Hence

$$N \le N_0 + N_1 = \sup_\theta |\{i; r_i(\theta) = 0\}| + \sup_\theta |\{i; \ \exists j \ne i \ \text{ such that } \ |r_i(\theta)| = |r_j(\theta)|\}|.$$

Clearly N depends on $\mathbf{Z}_n$, but because the error distribution is continuous, it follows that $P(N_0 > p) = 0$ and $P(N_1 > 2p) = 0$ independently of n. This may be seen by considering all possible subsets of $\mathbf{Z}_n$ of size $p + 1$ and $2p + 1$ and then conditioning on the corresponding subsets of $\mathbf{X}_n$. Altogether this entails

$$n^{1/2} \overline{\lim_{\theta \to \hat{\theta}_n}} |\mathbf{S}_n(\hat{\theta}_n)| \xrightarrow{p} 0, \quad \text{as} \quad n \to \infty. \quad (B.19)$$

Next, it follows from Hàjek and Šidàk (1967, p. 166) that besides (B.15),

$$nE|\mathbf{S}_n(0)|^2 \to A \operatorname{trace}(\Sigma) \quad (B.20)$$

also holds as $n \to \infty$. Let now $\varepsilon > 0$ be given and choose $L > 0$ so that $L^2 = 3A \operatorname{trace}(\Sigma)/\varepsilon$. It then follows from Lemma 4.2, (B.17), (B.19), and (B.20) that for $n \ge n_0(\varepsilon)$, both of the inequalities

$$n^{1/2}|\mathbf{S}_n(0)| \le L \quad (B.21)$$

and

$$|\mathbf{S}_n(0) + \lambda(\hat{\theta}_n)| \le \varepsilon(n^{-1/2} + |\lambda(\hat{\theta}_n)|) \quad (B.22)$$

are satisfied with probability at least $1 - \varepsilon/2$. But (B.21) and (B.22) imply that

$$n^{1/2}|\lambda(\hat{\theta}_n)| \le \frac{\varepsilon + n^{1/2}|\mathbf{S}_n(0)|}{1 - \varepsilon} \le \frac{L + \varepsilon}{1 - \varepsilon}.$$

Inserting this inequality into (B.22) yields that

$$n^{1/2}|\mathbf{S}_n(0) + \lambda(\hat{\theta}_n)| \le \frac{\varepsilon(L + 1)}{1 - \varepsilon} \quad (B.23)$$

holds with probability at least $1 - \varepsilon$. Because the right side of (B.23) can be made arbitrarily small, (B.16) is proved and hence the theorem.

The remaining results of the Appendix are needed for proving uniform asymptotic linearity of $\mathbf{S}_n(\theta)$ in Lemma 4.2. Lemma B.1 gives an upper bound for the first two moments of the quantity $Q_{nj}(\theta, d)$ used in (B.4), by writing it as a U statistic. A result related to Lemma B.1 for a certain weighted version of simple linear rank statistics with Wilcoxon scores ($h^+(u) = u$) was proved by Sievers (1983, thm 5.1). Lemma B.2 on the other hand corresponds to pointwise asymptotic (i.e., θ is fixed) linearity of $\mathbf{S}_n(\theta)$. The proof is based on approximating $\mathbf{S}_n(\theta)$ by a (Chernoff–Savage-type) integral involving empirical distributions of the data. (See Denker and Rösler 1985 for a similar approach.)

Lemma B.1. There exists an upper bound $Q_{nj}(\theta, d)$ of $\sup_{\{\tau; |\tau - \theta|_\infty \le d\}} |S_{nj}(\tau) - S_{nj}(\theta)|$ satisfying $EQ_{nj}(\theta, d) \le bd$ and $\operatorname{var} Q_{nj}(\theta, d) \le cdn^{-1}, j = 1, \ldots, p$, under Assumptions 1–6, where $b = C(h^+ F, G)$ and $c = C(h^+, F, G)$ are constants not depending on d, n, and θ.

Proof. See lemma A.2 of Hössjer (1991).

Lemma B.2. Suppose that Assumptions 1–6 are satisfied and that $\theta_0 = 0$. Then $E(S_{nj}(\theta) - S_{nj}(0) - \lambda_j(\theta))^2 \le C(n^{-5/4} + |\theta|^2 n^{-1})$, $j = 1, \ldots, p$, where $\lambda_j(\theta)$ is defined in (16) and $C = C(h, F, G)$ is a constant independent of θ and n.

Proof. See lemmas A.2 and A.3 of Hössjer (1991).

Corollary B.1. Suppose that Assumptions 1–6 hold and that $\theta_0 = 0$. Then $E(S_{nj}(\theta) - \lambda_j(\theta))^2 \to 0$ as $n \to \infty, j = 1, \ldots, p$.

Proof. Because

$$ES_{nj}(0)^2 \le Cn^{-1}, \quad (B.24)$$

with $C = C(h, F, G)$ [cf. (B.20)], the corollary holds for $\theta = 0$. For $\theta \ne 0$, observe that $E(S_{nj}(\theta) - \lambda_j(\theta))^2 \le 2ES_{nj}(0)^2 + 2E(S_{nj}(\theta) - S_{nj}(0) - \lambda_j(\theta))^2$. Hence the result follows from (B.24) and Lemma B.2.

[Received March 1991. Revised November 1992.]

REFERENCES

Cambanis, S., Simons, G., and Stout, W. (1976), "Inequalities for $Ek(X, Y)$ When the Marginals are Fixed," *Zeitschrift fr Wahrscheinlichkeitstheorie und verwandte Gebiete,* 36, 285–294.

David, H. A. (1970), *Order Statistics,* New York: John Wiley.

Denker, M., and Rösler, U. (1985), "A Moment Estimate for Rank Statistics," *Journal of Statistical Planning and Inference,* 12, 269–284.

Donoho, D. L., and Huber, P. J. (1983), "The Notion of Breakdown Point," in *A Festschrift for Erich Lehmann,* eds. P. Bickel, K. Doksum, and J. L. Hodges, Jr., Belmont, CA: Wadsworth.

Hàjek, J., and Šidàk, Z. (1967), *Theory of Rank Tests,* New York: Academic Press.

Hampel, F. R. (1971), "A General Qualitative Definition of Robustness," *The Annals of Mathematical Statistics,* 42, 1887–1896.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics, The Approach Based on Influence Functions,* New York: John Wiley.

Hettmansperger, T. P., and Mckean, J. W. (1983), "A Geometric Interpretation of Inferences Based on Ranks in the Linear Model," *Journal of the American Statistical Association,* 78, 885–893.

Hodges, J. L., Jr. (1967), "Efficiency in Normal Samples and Tolerance of Extreme Values for Some Estimates of Location," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* pp. 163–168.

Hössjer, O. (1991), "Robust Linear Regression by Means of M and R Statistics," Ph.D. dissertation, Uppsala University, Dept. of Mathematics.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates Under Non-Standard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* pp. 221–233.

——— (1973), "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *The Annals of Statistics,* 1, 799–821.

This content downloaded from 130.237.165.40 on Wed, 03 Feb 2016 22:26:14 UTC
All use subject to JSTOR Terms and Conditions

Jaeckel, L. A. (1972), "Estimating Regression Coefficients by Minimizing the Dispersion of the Residual," *The Annals of Mathematical Statistics,* 43, 1449–1458.

Jurečkovà, J., and Portnoy, S. (1987), "Asymptotics for One-Step M Estimators in Regression with Application to Combining Efficiency and High Breakdown Point," *Communications in Statistics, Part A—Theory and Methods,* 16, 2187–2199.

Mckean, J. W., and Schrader, R. M. (1980), "The Geometry of Robust Procedures in Linear Models," *Journal of the Royal Statistical Society,* Ser. B, 42, 366–371.

Mckean, J. W., Sheather, S. J., and Hettmansperger, T. (1990), "Regression Diagnostics for Rank-Based Methods," *Journal of the American Statistical Association,* 85, 1018–1028.

Morgenthaler, S. (1991), "A Note on Efficient Regression Estimators With Positive Breakdown Point," *Statistics and Probability Letters,* 11, 469–472.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871–880.

——— (1985), "Multivariate Estimation With High Breakdown Point," in *Mathematical Statistics and Applications* (Vol. B), eds. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Dordrecht, The Netherlands: Reidel, pp. 283–297.

Rousseeuw, P. J., and Leroy, A. (1987), *Robust Regression and Outlier Detection,* New York: John Wiley.

Rousseeuw, P. J., and Yohai, V. Y. (1984), "Robust Regression by Means of S Estimators," in *Robust and Nonlinear Time Series Analysis* (Lecture Notes in Statistics No. 26), eds. J. Franke, W. Härdle, and R. D. Martin, New York: Springer Verlag, pp. 256–272.

Ruppert, D. (1992), "Computing *S* Estimators for Regression and Multivariate Location / Dispersion," *Journal of Computational and Graphical Statistics,* 1, 253–270.

Siegel, A. F. (1982), "Robust Regression Using Repeated Medians," *Biometrika,* 69, 242–244.

Sievers, G. L. (1983), "A Weighted Dispersion Function for Estimation in Linear Models," *Communications in Statistics, Part A—Theory and Methods,* 12, 1161–1179.

Stefanski, L. A. (1991), "A Note on High-Breakdown Estimators," *Statistics and Probability Letters,* 11, 353–358.

Tableman, M. (1990), "Bounded-Influence Rank Regression: A One-Step Estimator Based on Wilcoxon Scores," *Journal of the American Statistical Association,* 85, 508–513.

Van Eeden, C. (1972), "An Analogue, for Signed Rank Statistics, of Jurečková's Asymptotic Linearity Theorem for Rank Statistics," *The Annals of Mathematical Statistics,* 43, 791–802.

Wellner, J. A. (1977), "A Glivenko–Cantelli Theorem and Strong Laws of Large Numbers for Functions of Order Statistics," *The Annals of Statistics,* 5, 473–480, corr, 6, 1394.

Yohai, V. (1987), "High Breakdown-Point and High Efficiency Robust Estimates for Regression," *The Annals of Statistics,* 15, 642–656.

Yohai, V., and Maronna, R. (1976), "Location Estimators Based on Linear Combinations of Modified Order Statistics," *Communications in Statistics, Part A—Theory and Methods,* 5, 481–486.

Yohai, V., and Zamar, R. (1988), "High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale," *Journal of the American Statistical Association,* 83, 406–413.

This content downloaded from 130.237.165.40 on Wed, 03 Feb 2016 22:26:14 UTC
All use subject to JSTOR Terms and Conditions