On One-Step GM Estimates and Stability of Inferences in Linear Regression

D. G. Simpson; D. Ruppert; R. J. Carroll

Journal of the American Statistical Association, Vol. 87, No. 418. (Jun., 1992), pp. 439-450.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199206%2987%3A418%3C439%3AOOGEAS%3E2.0.CO%3B2-D

Journal of the American Statistical Association is currently published by American Statistical Association.

On One-Step GM Estimates and Stability of Inferences in Linear Regression

D. G. SIMPSON, D. RUPPERT, and R. J. CARROLL*

The folklore on one-step estimation is that it inherits the breakdown point of the preliminary estimator and yet has the same large sample distribution as the fully iterated version as long as the preliminary estimate converges faster than $n^{-1/4}$, where n is the sample size. We investigate the extent to which this folklore is valid for one-step GM estimators and their associated standard errors in linear regression. We find that one-step GM estimates based on Newton–Raphson or Scoring inherit the breakdown point of high breakdown point initial estimates such as least median of squares provided the usual weights that limit the influence of extreme points in the design space are based on location and scatter estimates with high breakdown points. Moreover, these estimators have bounded influence functions, and their standard errors can have high breakdown points. The folklore concerning the large sample theory is correct assuming the regression errors are symmetrically distributed and homoscedastic. If the errors are asymmetric and homoscedastic, Scoring still provides root-n consistent estimates of the slope parameters, but Newton–Raphson fails to improve on the rate of convergence of the preliminary estimates. If the errors are symmetric and heteroscedastic, Newton–Raphson provides root-n consistent estimates, but Scoring fails to improve on the rate of convergence of the preliminary estimate. Our primary concern is with the stability of the inferences associated with the estimates, not merely with the point estimates themselves. To this end we define the notion of standard error breakdown, which occurs if the estimated standard deviations of the parameter estimates can be driven to zero or infinity, and study the large sample validity of the standard error estimates. A real data set from the literature illustrates the issues.

KEY WORDS: Asymmetry; Heteroscedasticity; Least median of squares; Minimum volume ellipsoid; Robust inference; Standard error breakdown.

Consider the linear model $y_i = z_i^t\beta + \varepsilon_i$, for $i = 1, \ldots, n$, where $z_i = (1x_i^t)^t$, x_i is a known $(p - 1)$-dimensional vector of explanatory variables, and y_i is an observed response. Two standard assumptions are: (1) $\varepsilon_1, \ldots, \varepsilon_n$ are identically distributed according to some F, and (2) $F = N(0, \sigma^2)$ for some $\sigma^2 > 0$. The earlier robust regression estimators—for example, M estimators (Andrews 1974; Bickel 1975; Huber 1973), rank estimators (Hettmansperger and McKean 1977; Jaeckel 1972), and trimmed least squares (Ruppert and Carroll 1980)—were designed to maintain efficiency under violations of (2), especially when the error distribution is heavy-tailed. However, it is as important to protect against violations of (1), particularly at outlying x observations, where heteroscedasticity or nonlinearity is likely. The generalized M estimators (GM estimators), such as the proposals of Mallows (1975), Hampel (1978), Krasker (1980), and Krasker and Welsch (1982), and the weighted trimmed least squares estimators of De Jongh, DeWet, and Welsh (1988) were intended to produce stable results when there are possible response outliers at outlying values of x, as can occur when (1) fails. In particular, they have influence functions bounded in both x and y. Unfortunately these bounded-influence estimators have breakdown points of at most $1/(p+1)$, where p is the number of predictor variables (Maronna, Bustos, and Yohai 1979), suggesting that they can be overwhelmed by a cluster of outliers; see, for example, Rousseeuw (1984).

The low breakdown point of the GM estimators has been viewed as a serious deficiency, particularly for multidimensional problems and exploratory data analysis. Several high breakdown point (HBP) estimators have been proposed that achieve breakdown points near $\frac{1}{2}$ for each p, including the least median of squares estimator of Rousseeuw (1984), the S estimators of Rousseeuw and Yohai (1984), and the estimators of Yohai (1987) and Yohai and Zamar (1988), which combine good asymptotic efficiency under the normal linear model with HBP. These estimators do not have bounded influence functions.

The HBP property provides some confidence that one will not be completely fooled by a cluster of poorly fit data. In practice, however, one would like the inferences to be robust to outliers, leverage points, and so on. If a few points can change the estimate by many standard errors or change drastically the standard error, it is small consolation that the change in the estimate is bounded. Routine data are thought to contain 1%–10% gross errors (Hampel, Ronchetti, Rousseeuw, and Stahel 1986). Although this is below the breakdown point of HBP estimators currently available, such a fraction of anomalous data can have a substantial effect if the influence function is unbounded. See, for instance, table 1 of Yohai and Zamar (1988), in which the bias of the Krasker–Welsch bounded-influence estimator is considerably less than that of the HBP unbounded-influence estimators if the level of contamination is 5%. We therefore contend that the

* D. G. Simpson is Associate Professor, Department of Statistics and Institute for Environmental Studies, University of Illinois at Urbana-Champaign, Champaign, IL 61820. D. Ruppert is Professor, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853. R. J. Carroll is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843. Simpson's research was supported by an NSF Mathematical Sciences Postdoctoral Research Fellowship and Air Force Office of Scientific Research Grant AFOSR-87-0041. Ruppert's research was supported by NSF Grant DMS-88-0029 and the U.S. Army Research Office through the Mathematical Sciences Institute at Cornell. Carroll's research was supported by Air Force Office of Scientific Research Grant AFOSR-89-0240. The authors thank the participants of the IMA Conference on Robustness, Diagnostics, Computation, and Graphics for their useful comments; the IMA for providing travel funds and housing; and the referees, who suggested strengthening the technical conditions to substantially simplify the proofs.

Journal of the American Statistical Association
June 1992, Vol. 87, No. 418, Theory and Methods

local stability associated with the bounded-influence property is as important as the global stability suggested by a high breakdown point. Moreover, the stability of the standard errors themselves is important and worthy of investigation.

To construct regression estimators that have bounded influence functions and high breakdown points, we follow a strategy that exists in the folklore: Start with a high breakdown point estimator and perform one iteration of a Newton–Raphson-type algorithm towards solution of the GM estimating equations. Hampel et al. (1986, p. 330) mentioned the possibility of using a one-step GM estimator but gave no details. We find one detail to be crucial for a high breakdown point, namely, the x-dependent weights associated with the GM iteration need to be based on high breakdown point location and scatter estimates rather than on the customary multivariate M estimates. Section 1 provides the specific definitions of our one-step GM estimates. Section 2 provides the breakdown analysis. Clearly one can iterate a fixed finite number of times and retain the breakdown point of the one-step. As a rough measure of the stability of inferences based on the estimates, we consider breakdown of the standard errors as well as the parameter estimates. The influence functions are derived in Section 3.

The large sample theory of the one-step GM estimators requires some care, as one natural initial estimator (least median of squares) converges only like $n^{-1/3}$ rather than the $n^{-1/2}$ rate usually associated with parametric estimation (Davies 1990; Kim and Pollard 1990; Rousseeuw 1984). However, results presented in Section 4 establish that both Newton–Raphson and Scoring versions of the one-step GM estimators converge at the root-n rate provided that the preliminary estimate is better than fourth root-n consistent and that the regression errors are symmetric and homoscedastic. Using a different method of proof, Jureckova and Portnoy (1987) established this kind of result for certain one-step Huber estimators. We find that if the errors are asymmetric and homoscedastic, Scoring still provides root-n consistent estimates of the slope parameters, whereas Newton–Raphson fails to improve on the rate of convergence of the preliminary estimate. On the other hand, if the errors are heteroscedastic and symmetric then Newton–Raphson provides root-n consistent estimates, whereas Scoring fails to improve on the rate of convergence of the preliminary estimate. We study asymptotic validity of the standard errors as well.

A potential objection to bounded-influence estimators is their low efficiency in cases where most of the sample information about β is contained in a few high leverage points. However, Morganthaler (1988) and Stefanski (1991) have shown that no estimator with a breakdown point greater than $1/n$ can have high finite-sample efficiency in the presence of extreme leverage points. In such instances, which involve a kind of extrapolation, it requires considerable faith in the linear model to take seriously the efficiency under the model. Our principal motivation for requiring a bounded-influence function as well as a high breakdown point is stability of inference. Section 5 illustrates some of the issues with a particularly vexing data set.

1. ONE-STEP MALLOWS ESTIMATES

Define residuals, $r_i = y_i - z_i^t\hat{\beta}_0$, where $\hat{\beta}_0$ is a high breakdown preliminary estimate with breakdown value at least m/n. For instance, a modified least median of squares (LMS) estimate has $m = [(n-p)/2] + 1$ (Rousseeuw and Leroy 1987). Let $\hat{\sigma}_0 = \text{med}\{|r_i|\}/\kappa$, where κ is a standardizing constant, and let m_x and C_x be multivariate location and scatter for the $\{x_i\}$ with breakdown point at least m/n. A possible choice for (m_x, C_x), the minimum volume ellipsoid (MVE) estimator, is given by the center and covariance of the smallest ellipsoid containing at least $[(n+p+1)/2]$ points. It has $m = [(n-p+1)/2]$, the best possible for affine equivariant covariance estimators (Rousseeuw and van Zomeren 1990). Cook and Hawkins (1990) discussed certain difficult computational issues associated with MVE.

The estimators we use are one-step estimators taking the form

$$\hat{\beta} = \hat{\beta}_0 + H_0^{-1}g_0, \qquad g_0 = \hat{\sigma}_0 \sum_{i=1}^{n} \psi(r_i/\hat{\sigma}_0)w_iz_i,$$

where there are two viable choices for H_0:

$$\text{Newton–Raphson:} \quad H_0 = \sum_{i=1}^{n} w_iz_iz_i^t\psi^{(1)}(r_i/\hat{\sigma}_0);$$

$$\text{Scoring:} \quad H_0 = n^{-1}\sum_{i=1}^{n} \psi^{(1)}(r_i/\hat{\sigma}_0) \sum_{j=1}^{n} w_jz_jz_j^t.$$

In the regression we employ Mallows weights,

$$w_i = \min\left[1, \left\{\frac{b}{(x_i - m_x)^tC_x^{-1}(x_i - m_x)}\right\}^{\alpha/2}\right]. \qquad (1.1)$$

The case $\alpha = 0$ is the one-step Huber estimate discussed by Bickel (1975) and Jureckova and Portnoy (1987). Jureckova and Portnoy (1987) imposed a nonequivariant bound on the step size to get HBP when $\alpha = 0$. We show that if $\alpha \geq 1$, the Mallows weights automatically bound the step size. The case $\alpha = 1$ is usual for GM estimators, whereas $\alpha = 2$ was used by Giltinan, Carroll, and Ruppert (1986) to force a bounded change of variance function, indicating local stability of the asymptotic variance. Ronchetti and Rousseeuw (1985) gave the form of the change of variance function for GM estimators. An even more extreme case, $\alpha = \infty$, deletes any observation in which the robust Mahalanobis distance from m_x exceeds b. Rousseeuw and van Zomeren (1990) discussed this possibility. We set b equal to the $(1-\gamma)$ quantile of the chi-squared distribution on $p-1$ degrees of freedom, where $\gamma = .1$ or $.05$.

Scoring and Newton–Raphson are asymptotically equivalent if the errors $\{\varepsilon_i\}$ are independent and identically and symmetrically distributed; see Section 4. Another common choice for H_0 is based on iterative weighted least squares, but the resulting one-step estimator has a different asymptotic distribution that depends on that of the initial estimate; we forego the details. For either Newton–Raphson or Scoring, the large sample theory estimate of the covariance matrix of $\hat{\beta}$ is $D = H_0^{-1}M_0H_0^{-1}$, where M_0 has one of two forms:

$$\text{Nonexchangeable:} \quad M_0 = \hat{\sigma}_0^2 \sum_{i=1}^{n} w_i^2 z_i z_i^t \psi^2(r_i/\hat{\sigma}_0);$$

$$\text{Exchangeable:} \quad M_0 = n^{-1}\hat{\sigma}_0^2 \sum_{i=1}^{n} \psi^2(r_i/\hat{\sigma}_0) \sum_{j=1}^{n} w_j^2 z_j z_j^t.$$

If the ε_i are heteroscedastic, then in general D is consistent only if H_0 is "Newton–Raphson" and M_0 is "Nonexchangeable."

2. BREAKDOWN ANALYSIS

The finite sample breakdown point was introduced by Donoho and Huber (1983). Let $X = \{(x_i, y_i): i = 1, \ldots, n\}$ and let T be an estimator of β. Then the breakdown point of T at X is given by

$$\mathrm{BP}(T, X) = \min\{m/n : \sup_{X^*} \|T(X) - T(X^*)\| = \infty\},$$

where the supremum is over all choices of X^* consisting of $(n - m)$ points from X and m arbitrary points. A HBP estimator like Rousseeuw's (1984) LMS estimator has BP $\approx \frac{1}{2}$ for any data set where the z_i's are in general position; that is, any p of them are linearly independent. For the scatter matrix, C_x, breakdown is defined as driving $\lambda_{\max}(C_x) + \{\lambda_{\min}(C_x)\}^{-1}$ to infinity, where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the minimum and maximum eigenvalues of the matrix A. We obtain breakdown points for the one-step GM estimators defined in Section 1 and then consider breakdown of their covariance estimates.

2.1 Breakdown of Estimates

In what follows, we will assume that the first $n - m$ observations are the "good" ones and that the remaining m observations are free to roam. We assume that $n - m \geq n/2 + 1 \geq p$, and, without loss of generality, that the first p observations are such that $(z_1, \ldots, z_p)$ are linearly independent. As usual, $\psi(v)$ is odd and bounded. We make use of the following additional assumptions:

A. Assume that ψ is nondecreasing with the properties

$$\psi(v)/v \geq d_0 > 0 \qquad \text{if } 0 \leq |v| \leq a; \qquad (2.1)$$

$$\psi^{(1)}(v) \geq d_1 > 0 \qquad \text{if } 0 \leq |v| \leq a; \qquad (2.2)$$

and

$$a > \kappa. \qquad (2.3)$$

B. If ψ is redescending, assume (2.1)–(2.3) as well as

$$\sup_{|v| \geq a} |\psi^{(1)}(v)| = d_2, \qquad \text{where } d_1 > d_2. \qquad (2.4)$$

C. Assume that any set of $n - m - n/2$ "good" points has a linearly independent subset of size p.

Theorem 2.1. Either of Assumptions A or B suffice for the breakdown value of the one-step Mallows to be at least m/n under Scoring. For Newton–Raphson, the breakdown value is at least m/n under Assumptions A and C taken together.

Remark 2.1. If ψ is redescending, then $\psi^{(1)}(r_i/\hat{\sigma}_0)$ can go negative. We conjecture that in this case it is possible to manipulate p data points so that the Newton–Raphson version of H_0 equals 0.

2.2 Standard Error Breakdown

Let D be the covariance estimate of $\hat{\beta}$ given in Section 1. Standard-error breakdown occurs if either $\lambda_{\max}(D) \to \infty$ or $\lambda_{\min}(D) \to 0$. The former usually is the only concept considered, as in Hampel et al. (1986), but the latter is important as well. For instance, even if the estimate does not break down, the Wald-type tests for the parameters can break down if D breaks down to 0. He, Simpson, and Portnoy (1990) have discussed breakdown of tests in general. A simple analysis shows that $\lambda_{\max}(D) \leq \lambda_{\max}(M_0)/\lambda^2_{\min}(H_0)$, and we show in the Appendix that $\lambda_{\min}(H_0) > 0$. It is clear that, because $\alpha \geq 1$, $\lambda_{\max}(M_0)$ has a finite upper bound under any arrangement of the "bad" points, and hence the same holds for $\lambda_{\max}(D)$.

Unfortunately, breakdown to 0 may occur unless $\alpha \geq 2$. Because $\lambda_{\min}(D) \to 0$ if $\det(D) \to 0$, this breakdown occurs if either $\det(M_0) \to 0$ or $\det(H_0^{-1}) \to 0$, the latter occurring if $\lambda_{\max}(H_0) \to \infty$. A detailed analysis as presented in Section 2.1 shows that under any arrangement of the "bad" points, $\lambda_{\min}(M_0) > 0$. Thus $\lambda_{\min}(D) \to 0$ if we can show that $\lambda_{\max}(H_0) \to \infty$. This may happen if $\alpha < 2$.

Lemma 2.1. Define $d_i = z_i/\|z_i\|$ and let $\|z_j\| \to \infty$ for $j \geq n - m + 1$ in such a way that for a positive definite matrix S, $\sum_{i=n-m+1}^{n} d_i d_i^t \to S$. Then $\lambda_{\max}(H_0) \to \infty$ if $\alpha < 2$, whereas $\lambda_{\max}(H_0) = O(1)$ if $\alpha \geq 2$.

3. INFLUENCE ANALYSIS

Influence analysis is a method of studying the local stability of estimators in terms of the effect of point-mass perturbations of the data or the underlying distribution. Two approaches to influence analysis of linear regression are in common use: (1) treat $\{(x_i, y_i)\}$ as a random sample and define the influence function on the space of distributions for (x, y) (Hampel et al. 1986) and (2) define the influence function via asymptotic linearity of the estimator (Krasker and Welsch 1982). We show that in either case the influence function of the one-step Mallows estimator is bounded when evaluated at the model. Method (1) requires that the preliminary estimates have influence functions, but they need not be bounded. Method (2) requires only a rate of convergence. Method (2) is perhaps more appropriate for regression, because it yields an influence function even when an iid assumption on the x_i's is inappropriate.

First act as if $\{(x_i, y_i)\}$ is a random sample from a distribution F_0 and consider the effect of perturbation of F_0. We suppose that the preliminary estimates and the location and scatter functionals for x have influence functions, but the influence functions need not be bounded. For instance, the preliminary regression estimate might be a regression S estimate (Rousseeuw and Yohai 1984), and the location scatter estimate might be a multivariate S estimate (Davies 1987; Lopuhaä 1989). The alternative definition of the influence function via asymptotic linearity allows treatment of the minimum volume ellipsoid.

Consider a generic matrix-valued functional $T(F)$ defined on the space of distributions for (x, y). Let F_0 be a fixed distribution representing the target model and let F_λ be a point-mass contamination of F_0: $F_\lambda = (1-\lambda)F_0 + \lambda\Delta_{x,y}$, for $0 \le \lambda \le 1$. Following Hampel et al. (1986), T has an influence function, which we shall denote by $IF(x, y; T)$, if it has a directional derivative at $\lambda = 0$:

$$IF(x, y; T) = \lim_{\lambda \downarrow 0} \{T(F_\lambda) - T(F_0)\}/\lambda.$$

The IF operation preserves matrix dimensions and satisfies the multiplication and chain rules of scalar differentiation.

The one-step Newton–Raphson estimators described in the preceding section correspond to the functional $\hat\beta(F) = \hat\beta_0(F) + \{H(F)\}^{-1}g(F)$, where

$$g(F) = \hat\sigma_0(F)E_F\left[\psi\left(\frac{Y - Z^t\hat\beta_0(F)}{\hat\sigma_0(F)}\right)w(X, m(F^x), C(F^x))Z\right]$$

and

$$H(F) = E_F\left[\psi^{(1)}\left(\frac{Y - Z^t\hat\beta_0(F)}{\hat\sigma_0(F)}\right)w(X, m(F^x), C(F^x))ZZ^t\right].$$

For Scoring, $H(F)$ instead takes the form

$$H(F) = E_F\left[\psi^{(1)}\left(\frac{Y - Z^t\hat\beta_0(F)}{\hat\sigma_0(F)}\right)\right] \times E_{F^x}[w(X, m(F^x), C(F^x))ZZ^t].$$

Here E_F denotes expectation with respect to F, $\hat\beta_0(F)$ and $\hat\sigma_0(F)$ are the functionals corresponding to the preliminary regression and scale estimates, F^x is the marginal distribution for x, $m(F^x)$ and $C(F^x)$ are the location and scatter functionals for x, and the weight function w is of the same form as in (1.1). Assuming F_0 is such that the conditional distribution of $(Y - Z^t\hat\beta_0(F_0))$ given $Z = z$ is independent of z, Newton–Raphson and Scoring reduce to the same functional at F_0. If F_n is the empirical distribution of $\{(x_i, y_i)\}$, then statistics and functionals are related as follows: $\hat\beta = \hat\beta(F_n)$, $\hat\beta_0 = \hat\beta(F_n)$, $g_0 = ng(F_n)$, and $H_0 = nH(F_n)$.

For both Newton–Raphson and Scoring the multiplication rule yields

$$IF(x, y; \hat\beta) = IF(x, y; \hat\beta_0) + \{H(F_0)\}^{-1}IF(x, y; g), \quad (3.1)$$

because $g(F_0) = 0$. In the following we suppress the dependence of IF on (x, y). Fisher consistency and symmetry of the residual distribution for F_0 yield

$$\begin{aligned}IF(g) &= \sigma\psi\left(\frac{y - z^t\beta}{\sigma}\right)w(x, m(F_0^x), C(F_0^x))z \\ &\quad - E_{F_0}\left[\psi^{(1)}\left(\frac{Y - Z^t\beta}{\sigma}\right)w(X, m(F_0^x), C(F_0^x))ZZ^t\right] \\ &\quad \times IF(\hat\beta_0) + \sigma^{-1}g(F_0)IF(\hat\sigma_0) \\ &\quad + \sigma E_{F_0}\left[\psi\left(\frac{Y - Z^t\beta}{\sigma}\right)IF(w(X, m(\cdot), C(\cdot)))Z\right] \\ &\quad + IF(\hat\sigma_0)E_{F_0}\left[\left(\frac{Y - Z^t\beta}{\sigma}\right)\psi^{(1)}\left(\frac{Y - Z^t\beta}{\sigma}\right) \times w(X, m(F_0^x), C(F_0^x))Z\right] \\ &= \sigma\psi\left(\frac{y - z^t\beta}{\sigma}\right)w(x, m(F_0^x), C(F_0^x))z \\ &\quad - \mathrm{H}(F_0)IF(\hat\beta_0).\end{aligned}$$

Inserting the latter expression in (3.1) gives

$$IF(x, y; \hat\beta) = \{H(F_0)\}^{-1}\sigma\psi\left(\frac{y - z^t\beta}{\sigma}\right)w(x, m(F_0^x), C(F_0^x))z.$$

This expression agrees with the influence function of the fully iterated GM estimate with weight function w (Hampel et al. 1986).

Alternatively, observe that by Theorem 4.1 the one-step GM estimator has the following asymptotic representation:

$$\hat\beta = \beta + n^{-1}\sum_{i=1}^{n} Q^{-1}z_iw_i\sigma\psi\left(\frac{y_i - z_i^t\beta}{\sigma}\right) + o_p(n^{-1/2}), \quad (3.2)$$

where Q is as in D1 of Section 4.3. The summand in (3.2) shows the contributions of the observations to the deviation of $\hat\beta$ from β. Following Krasker and Welsch (1982) we call the corresponding function, $Q^{-1}zw(z)\sigma\psi((y - z'\beta)/\sigma)$, the influence function.

Remark 3.1. If an estimator has an influence function, then general results of He and Simpson (in press) imply that the bounded-influence property is necessary rather than sufficient for local stability of the estimator. A stronger result would be to establish that the bias sensitivity is bounded. Martin, Yohai, and Zamar (1989) studied bias properties of certain S estimators and GM estimators.

The Huber estimates, which used bounded ψ but $w(\cdot) = 1$, bound the residuals but not the influence of the position in the design space. These estimators are susceptible to leverage points; that is, to outliers in the design space. On the other hand, if ψ and $\|z\|w(z)$ are both bounded, then the Mallows estimators bound the joint influence of the residuals and the position in the design space.

4. LARGE SAMPLE THEORY

To provide a rigorous large sample theory on which to base precision estimates and other inferences, we derive asymptotic representations for the one-step GM estimators. The preliminary estimates $(\hat{\beta}_0, \hat{\sigma}_0)$ need only be n^{τ}-consistent for some $\tau \in (\frac{1}{4}, \frac{1}{2}]$. For instance, $\hat{\beta}_0$ might be the LMS estimate, which converges at the rate $n^{-1/3}$ (Davies 1990; Kim and Pollard 1990; Rousseeuw 1984), or the least trimmed sum of squares estimate (Rousseeuw 1984), which converges at the rate $n^{-1/2}$.

The rate of convergence of the remainder in the asymptotic representation depends on the rate of convergence of the preliminary estimator. Although any rate better than $n^{-1/4}$ suffices for the one-step estimator to be root-n consistent and asymptotically normal, a better rate of convergence for the preliminary estimator implies a better rate of convergence for the remainder. In the following let

$$Q_n = \sum_{i=1}^{n} E[\psi^{(1)}(\varepsilon_i/\sigma)]w_i z_i z_i^t. \tag{4.1}$$

Theorem 4.1. Assume conditions A1–D2 of Section 4.3. Suppose $\hat{\beta}_0 - \beta = O_p(n^{-\tau})$ and $\hat{\sigma}_0 - \sigma = O_p(n^{-\tau})$ for some $\tau \in (\frac{1}{4}, \frac{1}{2}]$. Then for Newton-Raphson, $n^{-1}(H_0 - Q_n) = O_p(n^{-\tau})$ and

$$n^{-1/2}H_0(\hat{\beta} - \beta)$$
$$= n^{-1/2}\sigma \sum_{i=1}^{n} \psi(\varepsilon_i/\sigma)w_i z_i + O_p(n^{1/2-2\tau}). \tag{4.2}$$

The same is true of Scoring if $n^{-1} \sum_{i=1}^{n} \|z_i\| = O(1)$.

Theorem 4.1 implies that $H_0(\hat{\beta} - \beta)$ is asymptotically normal with mean 0 and covariance A_n, where

$$A_n = \sigma^2 \sum_{i=1}^{n} \operatorname{var}[\psi(\varepsilon_i/\sigma)]w_i^2 z_i z_i^t. \tag{4.3}$$

In practice we estimate A_n by M_0. The following result shows that this works.

Theorem 4.2. Assume conditions A1–D2 and suppose $\hat{\beta}_0 - \beta = O_p(n^{-\tau})$ and $\hat{\sigma}_0 - \sigma = O_p(n^{-\tau})$. If $n^{-1} \sum_{i=1}^{n} w_i^4 \|z_i\|^4 = O(1)$, then nonexchangeable M_0 satisfies

$$n^{-1}(M_0 - A_n) = O_p(n^{-\tau}), \tag{4.4}$$

and hence $M_0^{-1/2}H_0(\hat{\beta} - \beta) = Z_n + O_p(n^{1/2-2\tau})$, where Z_n has mean 0, covariance I, and is asymptotically normal. The same is true for exchangeable M_0 if instead $n^{-1} \sum_{i=1}^{n} \|z_i\| = O(1)$.

4.1 Effect of Asymmetry

Condition D2 of Theorem 4.1 is essentially symmetry of the error distribution. Carroll and Welsh (1988) and Welsh (1989) noted that the Huber and Mallows GM estimates of the slope are consistent even when the errors are asymmetric. This kind of result extends to the one-step versions as well. We show that if the errors are iid, then the asymptotic bias introduced by asymmetry is absorbed in the intercept, and we provide asymptotic expansions for the slope estimates. Asymmetry implies that the Scoring and Newton-Raphson estimators have different limiting behavior. In particular, the Scoring estimate of the slope vector is root-n consistent, whereas the Newton-Raphson estimate fails to improve on the rate of convergence of the preliminary estimate.

Partition $\beta^t = (\eta, \gamma^t)$ into intercept η and slope vector γ, and do the same for $\hat{\beta}_0$ and $\hat{\beta}$. Even if the error distribution is asymmetric, γ is identifiable as the value such that the distribution of $y_i - x_i^t\gamma$ is independent of x_i (Carroll and Welsh 1988). Hence it is reasonable to expect $n^{\tau}(\hat{\gamma}_0 - \gamma) = O_p(1)$ even in the asymmetric case, as long as the errors are homoscedastic. For a fully iterated GM estimate the intercept η may be defined by the condition

$$E\psi\left(\frac{\varepsilon_i}{\sigma}\right) = E\psi\left(\frac{y_i - \eta - x_i^t\gamma}{\sigma}\right) = 0. \tag{4.5}$$

As different choices of ψ give different values of η in the asymmetric case, we can expect only that $n^{\tau}(\hat{\eta}_0 - \eta_0) = O_p(1)$ for some η_0 not necessarily the same as η.

Let $\beta_0 = (\eta_0, \gamma^t)^t$ be the limiting value of the preliminary estimator and define $u_i = y_i - z_i^t\beta_0 = \varepsilon_i + \eta - \eta_0$ for $i = 1, \ldots, n$. Replace ε_i by u_i in the definition of Q_n. In correspondence with the partition of β, we partition the Hessian matrix and Q_n:

$$H_0 = \begin{bmatrix} h_{11} & h_{(1)}^t \\ h_{(1)} & H_{22} \end{bmatrix}, \qquad Q_n = \begin{bmatrix} q_{11} & q_{(1)}^t \\ q_{(1)} & Q_{22} \end{bmatrix}.$$

Here h_{11} and q_{11} are scalars and H_{22} and Q_{22} are $(p - 1) \times (p - 1)$ symmetric matrices. Define $H_{22\cdot 1} = H_{22} - h_{(1)}h_{(1)}^t/h_{11}$ and similarly define $Q_{22\cdot 1}$. To simplify the analysis, we center the x's by their Mallows-weighted means so that

$$\sum_{i=1}^{n} x_i w_i = 0. \tag{4.6}$$

This centering implies that $Q_{22\cdot 1} = Q_{22}$ and, for Scoring, $H_{22\cdot 1} = H_{22}$.

Lemma 4.1. Assume conditions A1–C2. Assume D1, replacing $\{\varepsilon_i\}$ by $\{u_i\}$. Suppose $\hat{\beta}_0 - \beta_0 = O_p(n^{-\tau})$ and $\hat{\sigma}_0 - \sigma = O_p(n^{-\tau})$. Then for Scoring, $n^{-1}(H_{22} - Q_{22}) = O_p(n^{-\tau})$ and

$$n^{-1/2}H_{22}(\hat{\gamma} - \gamma) = n^{-1/2}\sigma \sum_{i=1}^{n} x_i w_i \{\psi(u_i/\sigma)$$
$$- E[\psi(u_1/\sigma)]\} + O_p(n^{1/2-2\tau}).$$

Assume also that $\psi^{(2)}$ has derivative $\psi^{(3)}$ with $\|\psi^{(3)}\|_{\sup}$ and $\|(\cdot)^2\psi^{(3)}(\cdot)\|_{\sup}$ both finite. Then for Newton-Raphson, $n^{-1}(H_{22\cdot 1} - Q_{22}) = O_p(n^{-\tau})$ and

$$\hat{\gamma} = \gamma + Q_{22}^{-1}\sigma \sum_{i=1}^{n} x_i w_i \{\tilde{\psi}(u_i/\sigma) - E[\tilde{\psi}(u_1/\sigma)]\}$$
$$+ \frac{a_0 a_2}{a_1^2}(\hat{\gamma}_0 - \gamma) + O_p(n^{1/2-2\tau}),$$

where $\tilde{\psi}(t) = \psi(t) - a_0 a_1^{-1}\psi^{(1)}(t)$ and $a_k = E[\psi^{(k)}(u_1/\sigma)]$ for $k = 0, 1, 2$.

Remark 4.1. If the preliminary estimate converges more slowly than $n^{-1/2}$, then the expansion for Newton–Raphson implies $n^{\tau}(\hat{\gamma} - \gamma) = (a_0 a_2 / a_1^2) n^{\tau}(\hat{\gamma}_0 - \gamma) + o_p(1)$, and the asymptotic relative efficiency of Newton–Raphson versus $\hat{\gamma}_0$ is $a_1^4 a_0^{-2} a_2^{-2}$. This approaches infinity as the error distribution approaches symmetry.

Remark 4.2. Both the Scoring and Newton–Raphson versions of $\hat{\eta}$ converge in probability to $\eta_0 + \sigma a_0 / a_1$, which is one step of a Newton–Raphson algorithm for solving (4.5). Hence, iteration can drive a_0 to 0. In theory, iterating k_n times to achieve $a_0 = o(n^{\tau - 1/2})$ implies that the Newton–Raphson k_n step has the same asymptotic distribution as does the fully iterated version.

Remark 4.3. In the asymmetric case, asymptotically valid Wald-type inferences on the slope parameters may be obtained by the Scoring method coupled with the following modification of the exchangeable M_0:

$$M_{22} = n^{-1} \hat{\sigma}_0^2 \sum_{i=1}^{n} \{\psi(r_i / \hat{\sigma}_0) - \bar{\psi}\}^2 \sum_{j=1}^{n} w_i^2 x_i x_i^t,$$

where $\bar{\psi} = n^{-1} \sum_{i=1}^{n} \psi(r_i / \hat{\sigma}_0)$. In this case $M_{22}^{-1/2} H_{22}(\hat{\gamma} - \gamma) = Z_{n2} + O_p(n^{1/2 - 2\tau})$, where Z_{n2} has mean 0 and covariance I_{p-1} and is asymptotically normal.

4.2. Effect of Heteroscedasticity

We next consider the large sample behavior of one-step estimators when the errors are symmetrically distributed but heteroscedastic. We show that Newton–Raphson and the nonexchangeable version of M_0 provide valid large sample inferences, whereas Scoring fails to improve on the rate of convergence of the initial estimator.

Lemma 4.2. Suppose the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent with $\varepsilon_i \sim F_i$. Assume A2–D1 of Section 4.3, and assume D2 holds for each F_i. Suppose $n^{\tau}(\hat{\beta}_0 - \beta) = O_p(1)$ and $n^{\tau}(\hat{\sigma}_0 - \sigma) = O_p(1)$. Then both Newton–Raphson and Scoring have expansions of the form

$$n^{-1} H_0(\hat{\beta} - \beta) = n^{-1} \sigma \sum_{i=1}^{n} \psi(\varepsilon_i / \sigma) w_i z_i + T_n + O_p(n^{-2\tau}).$$

For Newton–Raphson $T_n = 0$, whereas for Scoring T_n is asymptotically equivalent to $\Gamma_n(\hat{\beta}_0 - \beta)$ for a symmetric nonstochastic matrix Γ_n.

Because of the heteroscedasticity, the limiting value of $\hat{\sigma}_0$ depends on the estimator. Although σ has an effect on the efficiency of $\hat{\beta}$, the Newton–Raphson covariance estimate $H_0^{-1} M_0 H_0^{-1}$ is asymptotically correct.

Theorem 4.3. Assume the conditions of Lemma 4.2. For the Newton–Raphson version of H_0 and nonexchangeable M_0 we have $M_0^{-1/2} H_0(\hat{\beta} - \beta) = Z_n + O_p(n^{1/2 - 2\tau})$, where Z_n has mean 0 and covariance I and is asymptotically normal.

4.3 Technical Conditions and Remarks

A1. The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent with distribution function F.

A2. The score function ψ is bounded and continuous.

B1. ψ has derivative $\psi^{(1)}$ such that (a) $\|\psi^{(1)}\|_{\sup} < \infty$ and (b) $\|(\cdot)\psi^{(1)}(\cdot)\|_{\sup} < \infty$, where $\|\cdot\|_{\sup}$ is the supremum norm.

B2. $\psi^{(1)}$ has derivative $\psi^{(2)}$ such that (a) $\|\psi^{(2)}\|_{\sup} < \infty$, (b) $\|(\cdot)\psi^{(2)}(\cdot)\|_{\sup} < \infty$, and (c) $\|(\cdot)^2\psi^{(2)}(\cdot)\|_{\sup} < \infty$.

C1. As $n \to \infty$ the design satisfies (a) $n^{-1} \sum_{i=1}^{n} \|z_i\|^4 \times w_i^2 = O(1)$ and (b) $n^{-1} \sum_{i=1}^{n} \|z_i\|^3 w_i = O(1)$.

C2. The design satisfies $\lim_{n\to\infty} \max_{1 \le i \le n} \|z_i\|^2 w_i^2 / \sum \|z_j\|^2 w_j^2 = 0$.

D1. $\lim_{n\to\infty} n^{-1} A_n = A$ and $\lim_{n\to\infty} n^{-1} Q_n = Q$ for some symmetric positive definite matrices A and Q.

D2. $E_F[\psi(\varepsilon v)] = 0$ and $E_F[\varepsilon v \psi^{(1)}(\varepsilon v)] = 0$ for any nonnegative scalar v. For example, ψ is odd and F has a density symmetric about 0.

Remark 4.4. We place heavy conditions on ψ but weak conditions on F. In the context of robust inference it seems appropriate to place conditions on ψ (which is under our control) rather than on F. The differentiability of $\psi^{(1)}$ given in B2 can be weakened by Lipschitz-type conditions, as indicated in Lemma A.1.

Remark 4.5. For appropriately chosen Mallows weights the present design conditions are weaker than the standard conditions for Huber regression. In particular, taking $\alpha = 2$ in (2.1) ensures that $\|z_i\|^2 w_i \le \lambda_{\max}(C_x)$, so it is sufficient that $\lambda_{\max}(C_x) = O(1)$, $n^{-1} \sum \|z_i\| = O(1)$, and $\sum_{i=1}^{n} w_i^2 \|z_i\|^2 \to \infty$. The asymptotics of the preliminary estimator may require additional conditions; for instance, the conditions given by Kim and Pollard (1990) or Davies (1990) for least median of squares.

Remark 4.6. The conditions on ψ exclude piecewise linear score functions such as Hampel's three-part redescender. Simpson, Ruppert, and Carroll (1989) gave an alternative proof for such estimators. Discontinuities in $\psi^{(1)}$ can lead to instability in the large sample variance if there is substantial discreteness in the data (Simpson, Carroll, and Ruppert 1987).

5. LAND USE/WATER QUALITY

Haith (1976) collected data relating land use to water quality. Each case was a river basin in New York State. Basins were selected by two criteria: independence (no basin in the sample being a tributary of another basin in the sample) and completeness of the data. All 20 basins satisfying these criteria were included in the sample. The data, which also were given in Allen and Cady (1982, table 2.1), include five variables, nitrogen concentration and four land use variables given as a percentage of total land usage: N = total nitrogen; AC = active agriculture; FR = forest, brushland, or plantation; RS = residential; and CI = commercial/industrial. Haith (1976) developed linear regression models relating N to subsets of the four other variables. Because the purpose of modeling was to attribute nonpoint source pollution to the various types of land use, the parameter estimates and their standard errors were of primary interest.

The covariates exhibit sizeable linear dependencies, and there are design outliers. *AC* and *FR* have a negative association, except for case #5 (the Hackensack River), which is an outlier in the design space with low *AC* and *FR* values and high *RS* and *CI* values. Much of the variation in *RS* and *CI* is due to five rivers, and the observed *RS* and *CI* values exhibit a strong positive association. Their sample correlation is .86; their sample correlation excluding the rivers with the two highest *RS* values is .93. With such a design it is difficult to disentangle the residential and commercial effects reliably. To alleviate the collinearity, we replace *RS* and *CI* by their sum, $UR := RS + CI =$ percent urban land usage. If the goal were to predict *N*, one might instead use stepwise regression to select a subset of the variables; this was Haith's strategy. However, simpler models do not attain a goal of relating all land uses to water quality. Aggregating *RS* and *CI* is a compromise made necessary by the design that still allows us to relate all land uses to pollution.

Case #5 (the Hackensack River) is such a severe design outlier that data analysts likely would set this point aside rather than including it in a linear least squares analysis. We shall present results both with and without case #5. Although inferences that rely heavily on this point are too unstable to be trusted, it would be of interest to determine whether the Hackensack River conforms roughly to the model suggested by the other rivers or whether it points to some alternative phenomenon in urban rivers. The Mallows weights that we use essentially delete case #5 in the fitting algorithm. Such downweighting of design outliers and response outliers is meant to limit their influence on the fitted model and associated inferences, but it also has the benefit of accentuating the inadequacy of the model for these points, possibly making it easier to discover alternative and more satisfactory models. Thus, although outliers may be downweighted or even deleted during the fitting of the model, this does not imply that they are "discarded" in the analysis of the data. They are in fact emphasized.

For the full data, ordinary least squares (OLS) regression of *N* on the land use variables yields (with standard errors in parentheses):

$$\hat{N} = 1.43(\pm 1.29) + .0085(\pm .016)AC - .0084(\pm .015)FR + .029(\pm .028)UR.$$

Omitting case #5 yields instead

$$\hat{N} = 1.70(\pm .76) + .0021(\pm .0094)AC - .014(\pm .0086)FR + .16(\pm .028)UR.$$

Table 1. Linear Model Parameter Estimates and Standard Errors for New York Rivers Data

	OLS	LMS	M	GM
AC	.0028 (.0043)	.0157	.0175 (.0021)	.0164 (.0030)
FR	.0058 (.0020)	.00019	.0022 (.00096)	.0026 (.0014)
UR	.0437 (.016)	.171	.179 (.0077)	.203 (.046)
OTHER	.0143 (.013)	.0364	.0251 (.0063)	.0239 (.0077)

Table 2. Linear Model Parameter Estimates and Standard Errors Excluding the Hackensack River

	OLS	LMS	M	GM
AC	.0191 (.0026)	.0175	.0177 (.0018)	.0162 (.0029)
FR	.00322 (.0013)	.00114	.0023 (.00089)	.0024 (.0013)
UR	.173 (.025)	.136	.156 (.018)	.195 (.052)
OTHER	.0170 (.0076)	.0335	.0276 (.0054)	.0263 (.0070)

It is clear that case #5 would have considerable effect on the OLS inferences about urban effects were it included.

The parameter estimates and standard errors for the covariates in the above model are somewhat difficult to interpret, because the parameters represent incremental effects over other land uses not measured. We therefore reparameterize the model by replacing the intercept with the constructed variable $OTHER := 100 - AC - FR - UR$. This reparameterization leaves the design space intact but provides directly interpretable parameters. For instance, the *AC* parameter is the nitrogen that can be attributed to each percentage of agricultural use.

Tables 1 and 2 give estimates and standard errors using several methods: OLS; LMS; a three-step Huber estimator (M)—that is, Mallows with $\alpha = 0$, starting from LMS; and a three-step Mallows estimator (GM) with $\alpha = 2$, starting from LMS. The three-step estimates used the scoring method, exchangeable standard errors, and a three-part redescending Hampel ψ function with tuning constants $(a, b, c) = (1.5, 3, 8)$. The normalizing constant in the scale estimate was set equal to $\kappa = .6745$, but standard errors were inflated by the factor $\{W/(W-p)\}^{1/2}$, where W is the number of observations with nonzero weight. The Mallows weights for GM used $b = \chi^2(.95; p - 1)$. MVE estimates of location and scatter for the covariates were computed using a FORTRAN program supplied by B. van Zomeren. LMS was computed via the S-plus (Statistical Sciences, Inc.) function, LMSREG. S functions for the GM steps and diagnostics are available from the authors on request.

On deletion of case #5, the nitrogen concentration attributed to urban use by OLS quadruples and the standard errors become considerably smaller. It is clear that the nitrogen concentration for case #5 is much less than was predicted by linear extrapolation from the remaining data. The LMS and M parameter estimates are not affected drastically by the presence or absence of case #5; however, the M standard error for *UR* is more than doubled by the deletion. The GM parameter estimates and standard errors show little change on deletion of case #5. The M standard error for *UR* seems overly optimistic, even after deleting case #5, given the change in the estimates induced by the deletion and the differences among the estimates. The standard error associated with GM is perhaps more realistic.

Table 3 provides diagnostics for selected rivers based on the full data: diagonals of the OLS projection matrix (h_{ii}); OLS studentized residuals (t_i^{OLS}); standardized residuals for LMS (s_i^{LMS}), M (s_i^{M}), and GM (s_i^{GM}); and the Mallows weights (w_i). The standardized residuals s_i were scaled by $\mathrm{median}\{|\mathrm{residual}|\}/.6745$. McKean, Sheather, and Hett-

Table 3. Diagnostics for Selected Observations From the New York Rivers Data

i	h_{ii}	t_i^{OLS}	s_i^{LMS}	s_i^{M}	s_i^{GM}	w_i
3	.365	.726	0	.206	.302	.286
4	.170	.588	−.712	−1.16	−2.08	.0662
5	.957	−3.29	−44.6	−34.1	−41.2	.000585
6	.053	.839	−1.35	−1.10	−1.76	.0637
7	.063	2.89	0	−.041	−1.15	.0175
19	.315	−2.12	−5.38	−3.52	−3.62	1.00

mansperger (1990) developed a method of studentizing rather than standardizing robust residuals that likely will be helpful in studying outliers.

Case #5 is an OLS leverage point in the full data, and it exhibits a moderately large OLS studentized residual. Clearly this point will have a large effect on the OLS fit (Cook 1977). The OLS residuals not shown were all smaller than 1 in magnitude, perhaps a clue that case #5 has inflated the scale estimate. The extreme discordance of case #5 is obvious from the more robust standardized residuals, and the MVE-based Mallows weight also identifies it as extremely outlying in the design space. The Mallows weights not shown were all equal to 1. The corresponding MVE-based Mahalanobis distances (Rousseeuw and van Zomeren 1990) provide a clear identification of several urban rivers (cases #3–7). The robust residuals also point to case #19 (the Oswegatchie River) as a possible response outlier. It is suggestive that case #19 is the largest river basin and case #5 the smallest (Haith 1976, table 2).

Table 4 presents the same diagnostics after exclusion of case #5. Only case #19 remains as a response outlier. Case #7 emerges as a moderate OLS leverage point. The Mallows weights excluding case #5 are unchanged, because the resampling algorithm (Rousseeuw and van Zomeren 1990) selects the same subsample. Is there a pattern in the residuals? Figure 1 shows plots of residuals versus *UR* for OLS, LMS, M, and GM after excluding case #5. The plot for GM reveals a pattern of negative residuals for the more urban rivers. Coupled with the huge negative residual of the much more urban Hackensack River, there is evidence of nonlinearity for large values of *UR*. The pattern fails to emerge in the other plots, for which the estimators do not have the bounded-influence property. It is clear, however, that additional leverage points could influence the fit in the plots for OLS, LMS, and M.

The nonlinearity revealed by the GM plot suggests that an alternative mechanism might come into play in urban areas. Perhaps urban areas have more efficient waste treatment, which would mitigate the effects of urbanization on water quality. One could attempt to introduce nonlinearity into the model to account for such diminishing effects; however, because nearly all information about the nonlinearity is provided by the four most urban rivers, the effect will be difficult to model reliably.

Table 4. Diagnostics for Selected Observations Excluding the Hackensack River (#5)

i	h_{ii}	t_i^{OLS}	s_i^{LMS}	s_i^{M}	s_i^{GM}	w_i
3	.374	.577	.153	.251	.293	.286
4	.279	−1.09	0	−.952	−2.20	.0662
6	.178	−.650	−.783	−.976	−1.96	.0637
7	.640	.865	2.56	.812	−1.07	.0175
19	.323	−3.021	−7.92	−4.68	−4.51	1.00

The preceding analysis leads us to some tentative conclusions, with caveats about the hazards of interpreting observational data. The significant contribution of agricultural use to nitrogen content persists across estimators, so this appears to be a reliable attribution. Forestland also persists as a minor, marginally significant contributor. Urbanization of rural rivers is associated with relatively large increases in nitrogen content, but there is evidence that further urbanization of substantially urban rivers has less effect. Given the size of the data set and the collinearity, attribution of nitrogen to sources is very difficult; we would not be surprised if others discovered analyses that they prefer to ours.

6. CONCLUSIONS

We have examined the behavior of one-step Mallows type robust regression methods in the linear model using either Scoring or Newton–Raphson. Two major general conclusions have emerged:

1. Under reasonably general conditions, the regression parameter estimates inherit the breakdown properties of the preliminary estimates of the regression parameters and the multivariate location and scale estimates of the design x's.
2. It makes little sense to confine attention to regression parameter estimation and to completely ignore the associated problem of inference. Even when regression parameter estimates have reasonable breakdown properties, their estimated standard errors may change radically with the deletion of a single observation.

We have shown how to construct Mallows regression parameter estimates with the same breakdown properties as their standard error estimates. The Mallows weights depend on a user-chosen parameter α in (1.1). When using a redescending ψ function, the Scoring method with $\alpha \geq 2$ is recommended for inference; $\alpha \geq 1$ suffices for point estimation.

In our analysis of the New York rivers data, we used LMS as the preliminary regression parameter estimate and the MVE scatter matrix estimate for the design. Both have high breakdown points, but they are extremely inefficient estimates and might have undesirable small sample performance; see, for example, Cook and Hawkins (1990). In our example this was not a problem. In other settings, however, one might be more successful in lowering the breakdown requirement from 50% to something less ambitious, such as 20%, to avoid the exact fit property (Rousseeuw and Yohai 1984). Moreover, although any rate of convergence better than $n^{-1/4}$ is sufficient for the one-step GM estimator to be root-n consistent and asymptotically normal, this approximation is more accurate if the preliminary estimator has a better rate of convergence. Hence improved performance

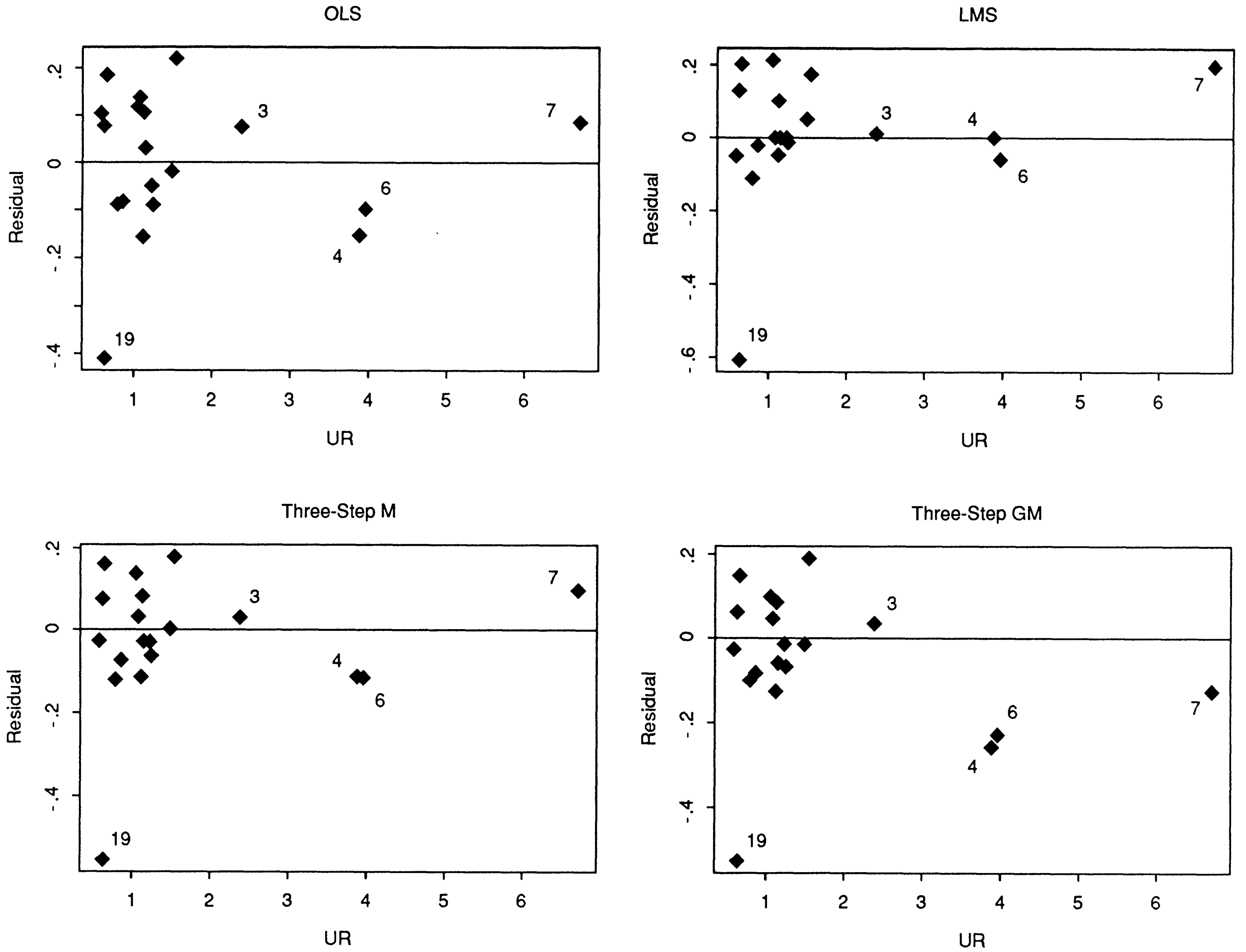

Figure 1. Residuals Versus Percent Urban Usage for Least Squares (OLS), Least Median Squares (LMS), a Three-Step M Estimator, and a Three-Step GM Estimator, Excluding the Hackensack River.

may occur using more efficient preliminary estimates such as S estimates (Rousseeuw and Yohai 1984; Davies 1987). Another way to improve the starting value is to iterate more than once, as in our analysis of the New York rivers data; we have observed empirically that three-step GM estimates starting from LMS or MVE are somewhat more stable than the one-step versions.

The behavior of one-step regression estimators with asymmetric and heteroscedastic errors deserves further study. If the regression errors are iid and symmetrically distributed, then both Scoring and Newton–Raphson have the standard large sample theory of fully iterated GM estimates. If the errors are asymmetric, however, then only Scoring improves on the rate of convergence of the preliminary estimate. On the other hand, if the errors are symmetric and heteroscedastic, then only Newton–Raphson shows this improvement. Fully iterated Mallows estimates work in either case, but they may give up the high breakdown point (Maronna et al. 1979).

The complexities encountered in the analysis of the land use data suggest that several important areas of research need further development, including stability of inference, robust model selection, and robust diagnostics.

APPENDIX: TECHNICAL PROOFS AND LEMMA

Proof of Theorem 2.1. First observe that $\|H_0^{-1}g_0\| \le \|g_0\|/|\lambda_{\min}(H_0)|$. Because $\alpha \ge 1$, we have

$$\begin{aligned}
&\hat{\sigma}^{-2}\|g_0\|^2 \\
&\quad \le \|\psi^2\|_{\sup}\sum_{i=1}^{n}\|z_i\|^2 w_i^2 \le \|\psi^2\|_{\sup} \\
&\quad\quad \times \sum_{i=1}^{n}\{1+\|m_x\|^2+\|x_i-m_x\|^2\}w_i^2 \\
&\quad \le \|\psi^2\|_{\sup}\left\{n(1+\|m_x\|^2)+b\sum_{i=1}^{n}\frac{\|x_i-m_x\|^2}{(x_i-m_x)^t C_x^{-1}(x_i-m_x)}\right\} \\
&\quad \le n\|\psi^2\|_{\sup}\{1+\|m_x\|^2+b\lambda_{\max}(C_x)\}.
\end{aligned}$$

Because ψ is bounded and C_x has breakdown m/n, $\|g_0\|^2$ has breakdown at least m/n. We now must show that no matter what one does with the "bad" points, $\lambda_{\min}(H_0) > 0$.

Scoring. We have that

$$\lambda_{\min}\left(\sum_{i=1}^{n} w_i z_i z_i^t\right) \geq \lambda_{\min}\left(\sum_{i=1}^{p} w_i z_i z_i^t\right)$$
$$\geq (\inf_{1\leq j\leq p} w_j)\lambda_{\min}\left(\sum_{i=1}^{p} z_i z_i^t\right). \quad (A.1)$$

Because by convention the first p of the $\{z_i\}$ are linearly independent, we need only consider the first factor on the right side in (A.1). This term is 0 only if $\sup_{1\leq j\leq p}\{(x_i - m_x)^t C_x^{-1}(x_i - m_x)\} = \infty$, which cannot happen because C_x has breakdown m/n and the first p observations are "good." It thus suffices to show that

$$\sum_{i=1}^{n} \psi^{(1)}(r_i/\hat{\sigma}) > 0. \quad (A.2)$$

By (2.3), there are at least $n/2$ observations with $|r_i|/\hat{\sigma} \leq a$; so that if ψ is nondecreasing, application of (2.2) suffices to prove (A.2). Under Assumption B we find that the left side of (A.2) is at least $n/2(d_1 - d_2)$, and (A.2) then follows from (2.4).

Newton–Raphson. We must show that under arbitrary manipulation of the "bad" points

$$\lambda_{\min}\left\{\sum_{i=1}^{n} \psi^{(1)}(r_i/\hat{\sigma}) w_i z_i z_i^t\right\} > 0. \quad (A.3)$$

When ψ is nondecreasing, $\psi^{(1)}(v) \geq 0$ and $\psi^{(1)}(r_i/\hat{\sigma}) \geq d_1 > 0$ for at least $n - m - n/2$ "good" points. Thus (A.3) follows from Assumption C.

Proof of Lemma 2.1. For the first part of the lemma, it suffices to replace H_0 by $\sum_1^n w_i z_i z_i^t$. As in (A.1), $\lambda_{\max}(\sum_{i=1}^n w_i z_i z_i^t) \geq \lambda_{\max}(\sum_{i=n-m+1}^n w_i z_i z_i^t)$. Now letting $\|z_j\| \to \infty$ for $j \geq n - m + 1$, we have $w_j \sim b^{\alpha/2}(x_j^t C_x^{-1} x_j)^{-\alpha/2} \sim b^{\alpha/2}\|x_j\|^{-\alpha}(g_j C_x^{-1} g_j)^{-\alpha/2} \sim b^{\alpha/2}\|z_j\|^{-\alpha}(g_j C_x^{-1} g_j)^{-\alpha/2}$, where $g_i = x_i/\|x_i\|$, because C_x and m_x have breakdown m/n. But because $g_j^t C_x^{-1} g_j \leq \lambda_{\max}(C_x^{-1}) = \{\lambda_{\min}(C_x)\}^{-1}$, it then follows that in the limit, as $\|z_j\| \to \infty$, $(\inf_{n-m+1\leq j\leq n} w_j) \geq \frac{1}{2}\{b\lambda_{\min}(C_x)\}^{\alpha/2}\|z_j\|^{-\alpha}$, and hence that $\lambda_{\max}(\sum_{i=1}^n w_i z_i z_i^t) \geq \frac{1}{2}\{b\lambda_{\min}(C_x)\}^{\alpha/2}\lambda_{\max}(\sum_{i=n-m+1}^n d_i d_i^t \|z_i\|^{2-\alpha})$. This can be made to diverge to ∞ if $\alpha < 2$. If $\alpha \geq 2$, then $\lambda_{\max}(\sum_{i=1}^n w_i z_i z_i^t) \leq \sum_{i=1}^n \|z_i\|^2 w_i \leq \sum_{i=1}^n (1 + \|m_x\|^2 + \|x_i - m_x\|^2) w_i \leq n\{1 + \|m_x\|^2 + b\lambda_{\max}(C_x)\}$, the last step following because $\alpha \geq 2$.

Proof of Theorem 4.1. We derive a more general result that holds even if the errors are asymmetric, as in Section 4.2. Let η_0 be the limiting value of the preliminary estimate of the intercept. Let $\beta_0 = (\eta_0, \gamma^t)^t$, $u_i = y_i - z_i^t\beta_0$, and $G(\beta, \sigma) = \sigma\sum_{i=1}^n \psi(u_i/\sigma) w_i z_i = \sigma\sum_{i=1}^n \psi((r_i + z_i'(\hat{\beta}_0 - \beta_0))/\sigma) w_i z_i$.

Newton–Raphson. Conditions B1 and B2 and the mean value theorem yield

$$G(\beta, \hat{\sigma}_0) = \hat{\sigma}_0 \sum_{i=1}^{n} \psi(r_i/\hat{\sigma}_0) w_i z_i + \sum_{i=1}^{n} \psi^{(1)}(r_i/\hat{\sigma}_0) w_i z_i z_i'(\hat{\beta}_0 - \beta_0)$$
$$+ \frac{1}{2}\hat{\sigma}_0^{-1} \sum_{i=1}^{n} \psi^{(2)}\left(\frac{r_i + z_i'(\tilde{\beta}_0 - \beta_0)}{\hat{\sigma}_0}\right) w_i z_i (z_i'(\hat{\beta}_0 - \beta_0))^2$$
$$= H_{\mathrm{NR}}(\hat{\beta}_{\mathrm{NR}} - \beta_0) + O\left(\hat{\sigma}_0^{-1}\|\hat{\beta}_0 - \beta_0\|^2 \sum_{i=1}^{n} w_i\|z_i\|^3\right), \quad (A.4)$$

where $\tilde{\beta}_0$ is on the line segment between β_0 and $\hat{\beta}_0$. On the other hand, applying the mean value theorem to $g(s) = G(\beta, s)$ yields, after some simplification,

$$G(\beta, \hat{\sigma}_0) = \hat{\sigma}_0 \sum_{i=1}^{n} \psi(u_i/\sigma) w_i z_i$$
$$- (\hat{\sigma}_0 - \sigma) \sum_{i=1}^{n} \psi^{(1)}(u_i/\sigma)(u_i/\sigma) w_i z_i$$
$$+ \frac{1}{2}(\hat{\sigma}_0 - \sigma)^2 \tilde{\sigma}^{-1} \sum_{i=1}^{n} \psi^{(2)}(u_i/\tilde{\sigma})(u_i/\tilde{\sigma})^2 w_i z_i,$$

where $\tilde{\sigma}$ is between $\hat{\sigma}_0$ and σ. Equating (A.4) and (A.5),

$$H_{\mathrm{NR}}(\hat{\beta}_{\mathrm{NR}} - \beta_0) = \hat{\sigma}_0 \sum \{\psi(u_i/\sigma) - a_0\} w_i z_i$$
$$- (\hat{\sigma}_0 - \sigma) \sum \{\psi^{(1)}(u_i/\sigma)(u_i/\sigma) - b_1\} w_i z_i$$
$$+ \{a_0\hat{\sigma}_0 + b_1(\sigma - \hat{\sigma}_0)\} \sum w_i z_i$$
$$+ O(\hat{\sigma}_0^{-1}\|\hat{\beta}_0 - \beta_0\|^2 \sum w_i\|z_i\|^3)$$
$$+ O((\hat{\sigma}_0 - \sigma)^2\tilde{\sigma}^{-1} \sum w_i\|z_i\|),$$

with $a_0 = E[\psi(u_1/\sigma)]$ and $b_1 = E[\psi^{(1)}(u_1/\sigma)(u_1/\sigma)]$.

The assumption on $\hat{\sigma}_0$, Conditions A1, A2, B1(b), C1, and Chebyshev's inequality imply $n^{-1/2}(\hat{\sigma}_0 - \sigma) \sum \{\psi(u_i/\sigma) - a_0\} w_i z_i = O_p(n^{-\tau})$ and $n^{-1/2}(\hat{\sigma}_0 - \sigma) \sum \{\psi^{(1)}(u_i/\sigma)(u_i/\sigma) - b_1\} w_i z_i = O_p(n^{-\tau})$. Moreover, by C1, $n^{-1/2}\hat{\sigma}_0^{-1}\|\hat{\beta}_0 - \beta_0\|^2 \sum w_i\|z_i\|^3 = O_p(n^{1/2-2\tau})$ and $n^{-1/2}(\hat{\sigma}_0 - \sigma)^2\tilde{\sigma}^{-1} \sum w_i\|z_i\| = O_p(n^{1/2-2\tau})$. Observing also that $\tau \leq \frac{1}{2}$ implies $2\tau - \frac{1}{2} \leq \tau$, we have

$$n^{-1/2}H_{\mathrm{NR}}(\hat{\beta}_{\mathrm{NR}} - \beta_0)$$
$$= n^{-1/2}\sigma \sum_{i=1}^{n} \{\psi(u_i/\sigma) - a_0\} w_i z_i + B_n + O_p(n^{1/2-2\tau}), \quad (A.6)$$

where the bias term is $B_n = n^{-1/2}\{a_0\hat{\sigma}_0 + b_1(\sigma - \hat{\sigma}_0)\} \sum_{i=1}^n w_i z_i$. Condition D2 implies $B_n = 0$, which establishes (4.2) for Newton–Raphson.

Scoring. Observe that

$$H_{\mathrm{S}}(\hat{\beta}_{\mathrm{S}} - \beta_0) = H_{\mathrm{NR}}(\hat{\beta}_{\mathrm{NR}} - \beta_0) + (H_{\mathrm{S}} - H_{\mathrm{NR}})(\hat{\beta}_0 - \beta_0). \quad (A.7)$$

Hence if we show that the components of $(H_{\mathrm{S}} - H_{\mathrm{NR}})$ are of order $O_p(n^{1-\tau})$, it will then follow that expansion (A.6) holds with $\hat{\beta}_{\mathrm{NR}}$ and H_{NR} replaced by $\hat{\beta}_{\mathrm{S}}$ and H_{S}. Setting $g(t) = \psi^{(1)}(t)$ and $c_i = n^{-1}w_i z_{ij} z_{ik}$ in Lemma A.1 shows that

$$n^{-1}H_{\mathrm{NR}} = n^{-1}Q_n + O_p\left(\frac{1}{n^{1+\tau}} \sum_{i=1}^{n} w_i\|z_i\|^2(1 + \|z_i\|)\right.$$
$$\left. + \left\{\frac{1}{n^2} \sum_{i=1}^{n} w_i^2\|z_i\|^4\right\}^{1/2}\right), \quad (A.8)$$

replacing ε_i by u_i in the definition of Q_n. On the other hand, setting $c_i = n^{-1}$ shows that $n^{-1} \sum \{\psi^{(1)}(r_i/\hat{\sigma}_0) - E[\psi^{(1)}(u_i/\sigma)]\} = O_p(n^{-\tau}(1 + n^{-1} \sum \|z_i\|) + n^{-1/2})$, from which it follows that

$$n^{-1}H_{\mathrm{S}} = n^{-1}Q_n$$
$$+ \mathrm{O}_p\left(n^{-\tau}\left(1 + n^{-1} \sum_{i=1}^{n} \|z_i\|\right) n^{-1} \sum_{i=1}^{n} w_i\|z_i\|^2\right). \quad (A.9)$$

Comparing (A.8) and (A.9) shows that $n^{-1}(H_{\mathrm{S}} - H_{\mathrm{NR}}) = O_p(n^{-\tau})$, whence

$$n^{-1/2}H_{\mathrm{S}}(\hat{\beta}_{\mathrm{S}} - \beta_0)$$
$$= n^{-1/2}\sigma \sum_{i=1}^{n} \{\psi(u_i/\sigma) - a_0\} w_i z_i + B_n + O_p(n^{1/2-2\tau}). \quad (A.10)$$

Lemma A.1. Suppose $\{u_i\}$ are independent, $n^\tau(\hat{\beta}_0 - \beta_0) = O_p(1)$, and $n^\tau(\hat{\sigma}_0 - \sigma) = O_p(1)$ with $\sigma > 0$. Let $\{c_i\}$ be a sequence of finite constants. If a measurable function g satisfies the Lipschitz condition

$$|g(s) - g(t)| \le L|s - t|/(1 + |t|), \quad (\text{all } s, t) \quad (\text{A.11})$$

for a finite constant L, then

$$\sum_{i=1}^{n} c_i \left\{ g\left(\frac{u_i + z_i^t(\beta_0 - \hat{\beta}_0)}{\hat{\sigma}_0} \right) - E[g(u_i/\sigma)] \right\}$$

$$= O_p\left(n^{-\tau} \sum_{i=1}^{n} |c_i|(1 + \|z_i\|) + \left\{ \sum_{i=1}^{n} c_i^2 \right\}^{1/2} \right).$$

Proof. Condition (A.11) implies

$$\left| g\left(\frac{u_i + z_i^t(\beta_0 - \hat{\beta}_0)}{\hat{\sigma}_0} \right) - g(u_i/\hat{\sigma}_0) \right| \le L \frac{|z_i^t(\beta_0 - \hat{\beta}_0)|}{\hat{\sigma}_0}$$

$$\le L \frac{\|z_i\| \|\beta_0 - \hat{\beta}_0\|}{\hat{\sigma}_0}$$

and $|g(u_i/\hat{\sigma}_0) - g(u_i/\sigma)| \le L|u_i\|\hat{\sigma}_0^{-1} - \sigma^{-1}|/(1 + |u_i|\sigma^{-1}) \le L\hat{\sigma}_0^{-1}|\sigma - \hat{\sigma}_0|$. Hence

$$\sum_{i=1}^{n} |c_i| \left| g\left(\frac{u_i + z_i^t(\beta_0 - \hat{\beta}_0)}{\hat{\sigma}_0} \right) - g(u_i/\sigma) \right|$$

$$\le L\left(\frac{\|\beta_0 - \hat{\beta}_0\| + |\sigma - \hat{\sigma}_0|}{\hat{\sigma}_0} \right) \sum_{i=1}^{n} |c_i|(1 + \|z_i\|).$$

Condition (A.11) also implies that g is continuous and bounded between $g(0) \pm L$. Hence the sum $\Delta_n = \sum_{i=1}^{n} c_i\{g(u_i/\sigma) - E[g(u_i/\sigma)]\}$ has mean 0 and variance bounded by $\{|g(0)| + L\}^2 \sum c_i^2$. Chebyshev's inequality implies $\Delta_n = O_p(\{\sum c_i^2\}^{1/2})$.

Proof of Theorem 4.2. To prove the result for nonexchangeable M_0, use Lemma A.1 with $g = \psi^2$ and $c_i = n^{-1}w_i^2 z_{ik}z_{il}(k, l \in \{1, \ldots, n\})$. For exchangeable M_0 set $c_i = n^{-1}$ to show $n^{-1} \sum \psi^2(r_i/\hat{\sigma}_0) = E\psi^2(\varepsilon_1/\sigma) + O_p(n^{-\tau}(1 + n^{-1} \sum \|z_i\|))$.

Proof of Lemma 4.1. The expansion for Scoring follows from (A.10), because $\sum_{i=1}^{n} w_i x_i = 0$ and H_S is block diagonal.

For Newton–Raphson, first recall that $n^{-1}(H_{NR} - Q_n) = O_p(n^{-\tau})$ by (A.8), and $q_{(1)} = 0$ due to the centering in (4.6). It follows that $h_{(1)}h_{11}^{-1} = O_p(n^{-\tau})$ and

$$n^{-1}H_{22 \cdot 1} = n^{-1}H_{22} + O_p(n^{-2\tau}) = n^{-1}Q_{22} + O_p(n^{-\tau}). \quad (\text{A.12})$$

Next rearrange (A.6) to obtain

$$H_{22 \cdot 1}(\hat{\gamma} - \gamma) = h_{(1)}h_{11}^{-1}(b_n + S_1) + S_2 + O_p(n^{1-2\tau}), \quad (\text{A.13})$$

where $b_n = \{a_0\hat{\sigma}_0 + b_1(\sigma - \hat{\sigma}_0)\} \sum w_i = \{a_0\hat{\sigma}_0 + O_p(n^{-\tau})\} \sum w_i$, $S_1 = \sigma \sum_{i=1}^{n} \{\psi(u_i/\sigma) - a_0\} w_i = O_p(\{\sum w_i^2\}^{1/2})$, and $S_2 = \sigma \sum_{i=1}^{n} \{\psi(u_i/\sigma) - a_0\} w_i x_i$. In (A.13) the term $h_{(1)}h_{11}^{-1}S_1 = O_p(n^{1/2-\tau}) = o_p(n^{1-2\tau})$, which can be absorbed into the remainder. Further we have $h_{11}^{-1} b_n = a_1^{-1} a_0\hat{\sigma}_0 + O_p(n^{-\tau})$, so it remains to detail the large sample behavior of $\hat{\sigma}_0 h_{(1)}$. An application of the mean value theorem yields

$$\hat{\sigma}_0 h_{(1)} = \hat{\sigma}_0 \sum \psi^{(1)}(r_i/\hat{\sigma}_0) w_i x_i$$

$$= \hat{\sigma}_0 \sum \psi^{(1)}(u_i/\hat{\sigma}_0) w_i x_i + \sum \psi^{(2)}(u_i/\hat{\sigma}_0) w_i x_i z_i^t(\hat{\beta}_0 - \beta_0)$$

$$+ O(\hat{\sigma}_0^{-1} \|\hat{\beta}_0 - \beta_0\|^2 \|\psi^{(3)}\|_{\sup} \sum w_i \|z_i\|^3). \quad (\text{A.14})$$

Further expansion of the first term in (A.14) yields

$$\hat{\sigma}_0 \sum \psi^{(1)}(u_i/\hat{\sigma}_0) w_i x_i$$

$$= \hat{\sigma}_0 \sum \psi^{(1)}(u_i/\sigma) w_i x_i - (\hat{\sigma}_0 - \sigma) \sum \psi^{(2)}(u_i/\sigma)(u_i/\sigma) w_i x_i$$

$$+ O\left(\frac{(\hat{\sigma}_0 - \sigma)^2}{\tilde{\sigma}} \|(\cdot)^2\psi^{(3)}(\cdot)\|_{\sup} \sum w_i \|x_i\| \right),$$

where $\tilde{\sigma}$ is between σ and $\hat{\sigma}_0$. Because of the centering, Chebyshev's inequality and the conditions on ψ and x imply that $\sum \psi^{(1)}(u_i/\hat{\sigma}_0)) w_i x_i = O_p(n^{1/2})$ and $\sum \psi^{(2)}(u_i/\sigma)(u_i/\sigma) w_i x_i = O_p(n^{1/2})$. Hence $\hat{\sigma}_0 \sum \psi^{(1)}(u_i/\hat{\sigma}_0) w_i x_i = \sigma\sum \{\psi^{(1)}(u_i/\sigma) - a_1\} w_i x_i + O_p(n^{1/2-\tau}) + O_p(n^{1-2\tau})$. To handle the second term in (A.14), note that $\sum \psi^{(2)}(u_i/\hat{\sigma}_0) w_i x_i z_i^t = a_2 \sum w_i x_i z_i^t + O_p(n^{-\tau} \sum w_i) + O_p(\{\sum w_i^2\}^{1/2}) = a_1^{-1} a_2[q_{(1)}; Q_{22}] + O_p(n^{1-\tau})$. Thus we have

$$\hat{\sigma}_0 h_{(1)} = \sigma \sum \{\psi^{(1)}(u_i/\sigma) - a_1\} w_i x_i$$

$$+ a_1^{-1} a_2 Q_{22}(\hat{\gamma}_0 - \gamma) + O_p(n^{1-2\tau}). \quad (\text{A.15})$$

Combining (A.12), (A.13), and (A.15) completes the proof.

Proof of Lemma 4.2. The proof of Theorem 4.1 for Newton–Raphson extends immediately to the present case. To handle Scoring, use (A.7) and observe that, by Lemma A.1,

$$n^{-1}(H_{NR} - H_S)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left\{ E\left[\psi^{(1)}\left(\frac{\varepsilon_i}{\sigma} \right) \right] - \frac{1}{n} \sum_{j=1}^{n} E\left[\psi^{(1)}\left(\frac{\varepsilon_j}{\sigma} \right) \right] \right\} w_i z_i z_i^t + O_p(n^{-\tau}).$$

As an example where this matrix fails to vanish asymptotically, let the empirical covariance between $E[\psi^{(1)}(\varepsilon_i/\sigma)]$ and $w_i z_{ip}^2$ converge to unity as $n \rightarrow \infty$.

Proof of Theorem 4.3. This follows from Lemma 4.2 and an application of Lemma A.1 to M_0.

[*Received January 1990. Revised February 1991.*]

REFERENCES

Allen, D. M., and Cady, F. B. (1982), *Analyzing Experimental Data by Regression,* Belmont, CA: Lifetime Learning Publications.

Andrews, D. F. (1974), "A Robust Method for Multiple Linear Regression," *Technometrics,* 16, 523-531.

Bickel, P. J. (1975), "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association,* 70, 428-434.

Carroll, R. J., and Welsh, A. H. (1988), "A Note on Asymmetry and Robustness in Linear Regression," *The American Statistician,* 42, 285-287.

Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics,* 19, 15-18.

Cook, R. D., and Hawkins, D. M. (1990), on "Unmasking Multivariate Outliers and Leverage Points" by P. J. Rousseeuw and B. C. van Zomeren, *Journal of the American Statistical Association,* 85, 640-644.

Davies, P. L. (1987), "Asymptotic Behavior of S-Estimates of Multivariate Location Parameters and Dispersion Matrices," *The Annals of Statistics,* 15, 1269-1244.

Davies, L. (1990), "The Asymptotics of S-Estimators in the Linear Regression Model," *The Annals of Statistics,* 18, 1651-1675.

Donoho, D. L., and Huber, P. J. (1983), "The Notion of Breakdown Point," In *A Festchrift for Erich L. Lehmann,* eds. P. J. Bickel, K. A. Doksum, and J. L. Hodges Jr., Belmont, CA: Wadsworth, pp. 157-184.

De Jongh, P. J., De Wet, T., and Welsh, A. H. (1987), "Mallows-Type Bounded-Influence Trimmed Means," *Journal of the American Statistical Association,* 84, 805-810.

Giltinan, D. M., Carroll, R. J., and Ruppert, D. (1986), "Some New Estimation Methods for Weighted Regression When There Are Possible Outliers," *Technometrics,* 28, 219-230.

Haith, D. A. (1976), "Land Use and Water Quality in New York Rivers," *Journal of the Environmental Engineering Division, Proceedings of the American Society of Civil Engineers,* 102, 1-15.

Hampel, F. R. (1978), "Optimally Bounding the Gross-Error-Sensitivity and the Influence of Position in Factor Space," in *Proceedings of the Statistical Computing Section, American Statistical Association,* pp. 59-64.

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.

He, X., Simpson, D. G., and Portnoy, S. (1990), "Breakdown Robustness of Tests," *Journal of the American Statistical Association,* 85, 446-452.

He, X., and Simpson, D. G. (in press), "Robust Direction Estimation," *The Annals of Statistics.*

Hettmansperger, T. P., and McKean, J. P. (1977), "A Robust Alternative Based on Ranks to Least Squares in Analyzing Linear Models," *Technometrics,* 19, 275-284.

Huber, P. J. (1973), "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *The Annals of Statistics,* 1, 799-821.

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

Jureckova, 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, Theory, and Methods,* 16(8), 2187–2199.

Kim, J., and Pollard, D. (1990), "Cube Root Asymptotics," *The Annals of Statistics,* 18, 191–219.

Krasker, W. S. (1980), "Estimation in Linear Regression Models With Disparate Data Points," *Econometrica,* 48, 1333–1346.

Krasker, W. S., and Welsch, R. E. (1982), "Efficient Bounded-Influence Regression Estimation," *Journal of the American Statistical Association,* 77, 595–604.

Lopuhaä, H. P. (1989), "On the Relation Between S-Estimators and M-Estimators of Multivariate Location and Covariance," *The Annals of Statistics,* 17, 1662–1683.

Mallows, C. L. (1975), "On Some Topics in Robustness," technical memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

Maronna, R. A., Bustos, O. H., and Yohai, V. J. (1979), "Bias- and Efficiency-Robustness of General M-Estimators for Regression With Random Carriers," in *Smoothing Techniques for Curve Estimation,* eds. T. Gasser and M. Rosenblatt, New York: Springer-Verlag, pp. 91–116.

Martin, R. D., Yohai, V. J., and Zamar, R. H. (1989), "Min-Max Bias Robust Regression," *The Annals of Statistics,* 17, 1608–1630.

McKean, J. W., Sheather, S. J., and Hettmansperger, T. P. (1990), "On the Use of Standardized Residuals From a High Breakdown GM-Fit of a Linear Model," in *Proceedings of the Business and Economic Statistics Section, American Statistical Association,* pp. 242–247.

Morgenthaler, S. (1989), "Comment on Yohai and Zamar," *Journal of the American Statistical Association,* 84, 636.

Ronchetti, E., and Rousseeuw, P. J. (1985), "Change-of-Variance Sensitivities in Regression Analysis," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete,* 68, 503–519.

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

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

Rousseeuw, P. J., and van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85, 633–639.

Rousseeuw, P. J., and Yohai, V. (1984), "Robust Regression by Means of S-Estimators," in *Robust and Nonlinear Time Series Analysis,* eds. J. Franke, W. Hardle, and R. D. Martin, New York: Springer-Verlag, pp. 256–272.

Ruppert, D., and Carroll, R. J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association,* 75, 828–838.

Simpson, D. G., Carroll, R. J., and Ruppert, D. (1987), "M-Estimation for Discrete Data: Asymptotic Distribution Theory and Implications," *The Annals of Statistics,* 15, 657–669.

Simpson, D. G., Ruppert, D., and Carroll, R. J. (1989), "One-Step GM-Estimates for Regression With Bounded Influence and High Breakdown Point," Technical Report No. 859, Cornell University, School of Operations Research and Industrial Engineering.

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

Welsh, A. H. (1989), "On M-Processes and M-Estimation," *The Annals of Statistics,* 17, 337–361.

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

Yohai, V. J., and Zamar, R. H. (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.