The Annals of Statistics
1992, Vol. 20, No. 4, 2161–2167

REWEIGHTED LS ESTIMATORS CONVERGE AT THE SAME RATE AS THE INITIAL ESTIMATOR

BY XUMING HE AND STEPHEN PORTNOY[1]

National University of Singapore and University of Illinois

The problem of combining high efficiency with high breakdown properties for regression estimators has piqued the interest of statisticians for some time. One proposal specifically suggested by Rousseeuw and Leroy is to use the least median of squares estimator, omit observations whose residuals are larger than some constant cut-off value and apply least squares to the remaining observations. Although this proposal does retain high breakdown point, it actually converges no faster than the initial estimator. In fact, the reweighted least squares estimator is asymptotically a constant times the initial estimator if the initial estimator converges at a rate strictly slower than $n^{-1/2}$.

1. Introduction. Consider the linear model

$$Y_i = x_i'\beta + u_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where $x_i \in \mathbb{R}^p$, $\beta \in \mathbb{R}^p$ and $\{u_i\}$ are i.i.d. from c.d.f. F having density f. In matrix form, $Y = X\beta + u$, where X denotes the $n \times p$ matrix with rows x_i'. An intuitively appealing approach to the estimation of β in the presence of outliers is to use a relatively simple high breakdown estimator $\hat{\beta}_0$ and to identify outliers as observations with unusually large residuals $r_i(\hat{\beta}_0)$, where

$$r_i(\beta) = Y_i - x_i'\beta. \tag{1.2}$$

Omitting observations with residuals larger than some constant (called "skipping") would still retain the high breakdown properties of the initial estimator. Rousseeuw and Leroy [(1987), pages 131 and 238] presented in detail a specific proposal using reweighted least squares (LS) based on skipping residuals from the least median of squares (LMS) estimator. Similar proposals appear elsewhere. We, and others apparently, have believed that this would provide a reasonably efficient estimator even if $\hat{\beta}_0$ converged at a rate less than $n^{-1/2}$. Note that the LMS estimator converges at rate $n^{-1/3}$ [Davies (1990)]. However, it turns out that this hope is illusory: The convergence rate of a skipped mean based on trimming large residuals is no better than that of the initial estimator $\hat{\beta}_0$ from which the residuals are computed. The purpose of this note is to provide a formal result showing the asymptotic equivalence of the reweighted least squares estimator and the initial estimator $\hat{\beta}_0$ as presented in Section 2. Before presenting this result, it may be instructive to

Received July 1991, revised March 1992.
[1]Partially supported by NSF Grant DMS-89-22472.
AMS 1980 *subject classifications*. Primary 62G35; secondary 62J05, 62E20.
Key words and phrases. Linear models, reweighted least squares, least median of squares, convergence rates.

consider the simplest case: that of a single location parameter ($p = 1$, $x_i \equiv 1$). Given residuals $r_i(\hat{\beta}_0)$ from an initial estimator, consider a symmetric skipped mean:

$$\hat{\beta} \equiv c \sum_{|r_i(\hat{\beta}_0)| \le a} Y_i, \quad \text{where } c = \frac{1}{\#\left\{i: \left|r_i(\hat{\beta}_0)\right| \le a\right\}}.$$

Without loss of generality, assume $\beta = 0$. If f is symmetric about zero, it is relatively clear that $\hat{\beta}$ is approximately (up to order $n^{-1/2}$)

$$\begin{aligned} \frac{\int_{\hat{\beta}_0-a}^{\hat{\beta}_0+a} uf(u)\,du}{\int_{\hat{\beta}_0-a}^{\hat{\beta}_0+a} f(u)\,du} &\approx \frac{\int_{\hat{\beta}_0-a}^{-a} uf(u)\,du + \int_{a}^{\hat{\beta}_0+a} uf(u)\,du}{2F(a) - 1 + O(\hat{\beta}_0)} \\ &= \frac{2af(a)}{2F(a) - 1}\hat{\beta}_0 + o(\|\hat{\beta}_0\|), \end{aligned} \tag{1.3}$$

assuming that f is continuous at a. Clearly, if $\hat{\beta}_0 - \beta = n^{-\gamma}\tau_n$, where τ_n converges in distribution to a nondegenerate limit, then $(\hat{\beta} - \beta)$ will be of the same order $(n^{-\gamma})$ so long as $f(a)$ is nonzero. The proof of Section 2 differs somewhat from this argument by using asymptotic expansions to show that $\hat{\beta}$ and $\hat{\beta}_0$ have the same rate of convergence for $\gamma < 1/2$ and for a reasonable wide class of weighting functions. Our expansion is reminiscent of (though simpler than) that of Ruppert and Carroll (1980), who consider root-n convergent initial estimators and find an asymptotically convergent expansion for $n^{1/2}(\hat{\beta} - \beta)$. Our result implies that some information is lost if one deletes observations based on residuals from a more slowly converging estimator. Note that one could also consider α trimming, that is, deleting a fixed fraction α of observations with extreme residuals. The expansions of Section 2 are carried out assuming skipping (rather than trimming) primarily because this is what Rousseeuw and Leroy (1987) suggest in order to preserve the high breakdown point of the initial estimator. In the case of α trimming, we can show by similar techniques that the result of Section 2 remains true. We forego the details.

Although our result here is negative, it is important to notice that the problem is in reweighting, and not only in the use of a more slowly convergent initial estimator (like the LMS estimator). For example, one-step M estimators using the LMS estimator as a starting value can be defined so that they converge at rate $n^{-1/2}$ and are efficient. In fact, such one-step estimators were explicitly suggested by Rousseeuw (1984). Further studies were carried out in Jurečková and Portnoy (1987) and Simpson, Ruppert and Carroll (1992). Practical experience suggests that more than one step should be taken, but the first order asymptotic distribution is still the same. Furthermore, if the skipping point (or trimming fraction) tends to infinity (or zero) sufficiently quickly as $n \to \infty$ (instead of remaining constant), then skipped (or trimmed) LS estimators should again converge at rate $n^{-1/2}$. In particular, consider the case of a single location parameter previously considered, and let $\hat{\beta}$ omit

observations with $|r_i(\hat{\beta}_0)| > a_n$, where $a_n \to \infty$. If $\sup\{|x|f(x): |x| > a_n\} = o(n^{-1/2+\gamma})$, then clearly the error in (1.3) converges to zero faster than $n^{-1/2}$, and such asymptotic skipping will work. Similar results should extend to the linear model (1.1). The finite-sample breakdown point of such estimators can remain high, but the sensitivity curves would become unbounded. Finally, an alternative for constructing estimators with high efficiency, bounded influence and high breakdown point is given in He (1991). The idea is to compute one efficient (and bounded influence) estimator and one high breakdown estimator; and to use the former if the two estimators are sufficiently close. This idea appears explicitly in the on-line menu of the new S-Plus package.

REMARKS. (1) In practice, residuals are generally standardized by dividing by an estimate of the standard deviation of the error. If this estimate converges at a rate of $n^{-1/2}$, then the expansion of Section 2 still holds. In fact, in typical cases, the standard deviation estimate based on an initial estimator will converge at a root-n rate, even if the initial estimator converges more slowly. In particular, Rousseeuw and Leroy (1988) and Davies (1990) showed that the scale estimate from the LMS estimate converges at a root-n rate.

(2) Since ν in the theorem is typically less than 1 (and, in fact, may be quite small), the reweighted LS estimator is asymptotically closer to the true parameter value than the initial estimator. However, its rate of convergence is no better, and so the reweighted LS estimator must be inefficient. How an estimator performs in finite sample situations, however, is another question of practical importance. In a Monte Carlo study described in Rousseeuw and Leroy [(1987), pages 208–214] the reweighted least squares estimator substantially improved the finite sample efficiency of the initial least median of squares estimator and actually outperformed a one-step M estimator for a sample size of 50. It is not unlikely that a much larger sample size is needed to see the effect of a slower convergence rate for the initial estimator. The result in this paper should not be interpreted as a rejection of the reweighted least squares method. More extensive comparisons for various proposals are necessary to make any recommendation as to which is the best way to estimate a linear regression model. See Ruppert (1991) for a recent attempt.

2. The asymptotic equivalence result. In this section, we give conditions and some formal expansions providing the desired result. The conditions are formulated more to simplify the proofs than to give the most general theorem possible. The conditions are as follows:

CONDITION F. The density is twice continuously differentiable.

CONDITION W. The weighting function $w(r)$ has at most finitely many jump discontinuities, is continuous and Lipschitz elsewhere, has compact support and satisfies $w(r) \geq 0$ for all r and $w(0) > 0$. Furthermore $\int uw(u)f(u)\,du = 0$.

CONDITION I. $\hat{\beta}_0$ is regression equivariant with $\hat{\beta}_0 - \beta = O_p(n^{-\gamma})$, $\gamma \in (0, 1/2)$.

CONDITION X. $\sum_{i=1}^n \|x_i\|^4 \le bn$ for some constant b, $n^{-1}X'X \to Q$, where Q is a positive definite $p \times p$ matrix and $\max_{1 \le i \le n} \|x_i\| = o(n^{\gamma \wedge 1/4})$.

REMARK. Somewhat more careful analysis would permit Condition F to be weakened to require only one continuous derivative for f. Furthermore, regression equivariance is only introduced to justify taking $\beta = 0$ in the proof without loss of generality. It is actually not needed since the terms introduced by leaving β general cancel exactly. However, computation of these terms would complicate the proof.

Now, the weighted least squares estimator is $\hat{\beta} = (X'WX)^{-1}X'WY$, where W is the $n \times n$ diagonal matrix with diagonal elements $w(r_i(\hat{\beta}_0))$. Following Ruppert and Carroll (1980) we seek to represent $\hat{\beta}$ in terms of $\hat{\beta}_0$. Since $n^{1/2}(\hat{\beta}_0 - \beta) \to \infty$, the Ruppert–Carroll result does not apply directly. However, we can still expand $\hat{\beta}_0$ directly, actually affording somewhat simpler computations. The basic approach uses the chaining argument to obtain the preliminary lemma. Since the proof is quite similar to those in Koenker and Portnoy (1987), we omit some of the details. First, given $\varepsilon > 0$, choose K so that for n large enough,

$$\mathbb{P}\left\{\left\|\hat{\beta}_0 - \beta\right\| \le Kn^{-\gamma}\right\} \ge 1 - \varepsilon. \tag{2.1}$$

LEMMA 2.1. *Define*

$$\Delta \equiv \{\delta : \|\delta\| \le Kn^{-\gamma}\}, \tag{2.2}$$

$$\begin{aligned} T_1(\delta) &\equiv \sum_{i=1}^n w(r_i(\beta + \delta))x_i Y_i, \qquad T_2(\delta) \equiv \sum_{i=1}^n w(r_i(\beta + \delta))x_i x_i', \\ \tilde{T}_k(\delta) &\equiv T_k(\delta) - ET_k(\delta) \quad \text{for } k = 1, 2. \end{aligned} \tag{2.3}$$

Then, under the preceding conditions,

$$\sup\left\{\left\|\tilde{T}_k(\delta)\right\| : \delta \in \Delta\right\} = O_p\left((n \log n)^{1/2}\right) \quad \text{for } k = 1, 2.$$

PROOF. (i) For each fixed $\delta \in \Delta$, use a large deviation result [as in the proposition of Koenker and Portnoy (1987, page 855) or by extending Theorem 4 of Feller [(1966), page 524] to show there is c such that for any $\lambda > c$

$$\mathbb{P}\left\{\left\|\tilde{T}_k(\delta)\right\| \ge \lambda(n \log n)^{1/2}\right\} \le ce^{-(\lambda - c)\log n}.$$

(ii) Now use the chaining argument: Cover Δ with balls S_ν of radius n^{-3}. We first wish to show that for any fixed S_ν,

$$\begin{aligned} &\mathbb{P}\left\{\sup\{\|T_1(\delta_1) - T_1(\delta_2)\| : \{\delta_1, \delta_2\} \subset S_\nu\} \ge \lambda(n \log n)^{1/2}\right\} \\ &\qquad \le ce^{-(\lambda - c)\log n}. \end{aligned} \tag{2.4}$$

To do this, first note that,

$$T_1(\delta_1) - T_1(\delta_2) = \sum_{i=1}^{n} \big(w(u_i - x_i'\delta_1) - w(u_i - x_i'\delta_2)\big)x_i(u_i + x_i'\beta).$$

By Condition X, for $\delta \in \Delta$, $|x_i'\delta|$ is uniformly bounded (it actually tends to zero); and so

$$\begin{aligned} &\|T_1(\delta_1) - T_1(\delta_2)\| \\ &\quad \le \Big| \sum_{|u_i| \le B} \big(w(u_i - x_i'\delta_1) - w(u_i - x_i'\delta_2)\big)x_i(u_i + x_i'\beta) \Big| \end{aligned} \tag{2.5}$$

for some B since w has compact support. Now, if i is such that $(u_i - x_i'\delta_1)$ and $(u_i - x_i'\delta_2)$ are *not* separated by a jump discontinuity, then since w is Lipschitz,

$$\begin{aligned} &\|\big(w(u_i - x_i'\delta_1) - w(u_i - x_i'\delta_2)\big)x_i(u_i + x_i'\beta)\| \\ &\quad \le c|x_i(\delta_1 - \delta_2)|\big(\|x_i\|B + \|x_i\|^2\beta\big) \\ &\quad \le c'\big(\|x_i\|^3 + \|x_i\|^2\big)n^{-3}, \end{aligned}$$

whose sum is uniformly bounded (say, by c_1) for $\{\delta_1, \delta_2\} \subset \Delta$. Otherwise, u_i must lie in an interval J_i of length $|x_i'(\delta_1 - \delta_2)| \le cn^{-2.75}$ on Δ. So

$$\begin{aligned} \mathbb{P}\{\#\{i\colon u_i \in J_i\} \ge \lambda\} &\le \sum_{j=[\lambda]}^{n} \binom{n}{j} (cn^{-2.75})^j \\ &\le \sum_{j=[\lambda]}^{n} \frac{1}{j!}(cn^{-1.75})^j \le e^c n^{-1.75\lambda}. \end{aligned}$$

Therefore, since each term in (2.5) contributes at most $c'\|x_i\| + c''\|x_i\|^2 \le c^* n^{1/2}$, there is c such that

$$\mathbb{P}\{\sup\{\|T_1(\delta_1) - T_1(\delta_2)\|\colon \{\delta_1, \delta_2\} \subset S_\nu\} \ge \lambda c^* n^{1/2} + c_1\} \le ce^{-\lambda \log n}$$

for each ball S_ν. Thus, (2.4) follows for n large enough.

Furthermore, it is direct to show that $E\|T_1(\delta_1) - T_1(\delta_2)\|$ is uniformly bounded on each S_ν. Hence, T_1 can be replaced by $\tilde{T}_1$ in (2.4). An entirely analogous argument holds for $\tilde{T}_2$; and, therefore, using part (i),

$$\mathbb{P}\Big\{\sup\big\{\big\|\tilde{T}_k(\delta)\big\|\colon \delta \in S_\nu\big\} \ge 2\lambda(n \log n)^{1/2}\Big\} \le 2ce^{-(\lambda - c)\log n}.$$

Last, the number of such balls needed to cover Δ is smaller than $c_2 n^{3p}$. Therefore,

$$\mathbb{P}\Big\{\sup\big\{\big\|\tilde{T}_k(\delta)\big\|\colon \delta \in \Delta\big\} \ge 2\lambda(n \log n)^{1/2}\Big\} \le 2cc_2 n^{3p} e^{-(\lambda - c)\log n} \to 0$$

for $\lambda > c + 3p$; and the result follows [using (2.1)]. □

The lemma leads to the following main theorem.

THEOREM. *Under the preceding conditions,*

$$\hat{\beta} - \beta = \nu\big(\hat{\beta}_0 - \beta\big) + o_p(n^{-\gamma}),$$

where $\nu = 1 + \int uw(u)f'(u)\,du / \int w(u)f(u)\,du$.

PROOF. First compute $ET_k(\delta)$ by expanding the density and using the conditions. Without loss of generality, assume $\beta = 0$:

$$\begin{aligned} ET_1(\delta) &= \sum_{i=1}^{n} Ew(Y_i - x_i'\beta - x_i'\delta)x_iY_i \\ &= \sum_{i=1}^{n} \int w(u)\,f(u + x_i'\delta)x_i(u + x_i'\delta)\,du \\ &= \sum_{i=1}^{n} \left\{a_1x_ix_i'\delta + O\big(\|x_i\|(x_i'\delta)^2\big)\right\} \\ &= a_1(X'X)\delta + O(n\|\delta\|^2), \end{aligned}$$

where $a_1 \equiv \int w(u)f(u)\,du + \int uw(u)f'(u)\,du$. Similarly,

$$ET_2(\delta) = a_2(X'X) + O(n\|\delta\|), \qquad a_2 \equiv \int w(u)\,f(u)\,du.$$

Now, except with probability bounded by ε, Lemma 2.1 applies (by 2.1); and

$$\begin{aligned} \tilde{\beta} &= \big(T_2(\hat{\beta}_0)\big)^{-1}T_1(\hat{\beta}_0) \\ &= \Big[a_2(X'X) + O_p(n^{1-\gamma}) + O\big((n\log n)^{1/2}\big)\Big]^{-1} \\ &\quad \times\Big[a_1(X'X)\hat{\beta}_0 + O_p(n^{1-2\gamma}) + O\big((n\log n)^{1/2}\big)\Big] \\ &= \nu\hat{\beta}_0 + O_p(n^{-2\gamma}) + O\big((\log n/n)^{1/2}\big), \end{aligned} \tag{2.6}$$

where the last step follows since $(X'X/n)^{-1}$ exists and has bounded eigenvalues (by Condition X). The result follows. □

REMARK. Note that by (2.6), the result of the theorem does not hold if $\gamma = 1/2$. Here, the more complicated expansion [Ruppert and Carroll (1980)] is required.

REFERENCES

DAVIES, L. (1990). The asymptotics of *S*-estimators in the linear regression model. *Ann. Statist.* **18** 1651–1675.

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

HE, X. (1991). A local breakdown property of robust tests in linear regression. *J. Multivariate Anal.* **38** 294–305.

JUREČKOVÁ, J. and PORTNOY, S. (1987). Asymptotics for one-step *M* estimators in regression with application to combining efficiency and high breakdown point. *Comm. Statist. Theory Methods* **16** 2187–2200.

KOENKER, R. W. and PORTNOY, S. (1987). *L*-Estimation for the linear model. *J. Amer. Statist. Assoc.* **82** 851–857.

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880.

ROUSSEEUW, P. J. and LEROY, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

ROUSSEEUW, P. J. and LEROY, A. M. (1988). A robust scale estimator based on the shortest half. *Statist. Neerlandica* **42** 103–116.

RUPPERT, D. (1991). Computing *S*-estimators for regression and multivariate location/dispersion. *J. Comput. Graphical Statist.* To appear.

RUPPERT, D. and CARROLL, R. J. (1980). Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.* **75** 828–838.

SIMPSON, D., RUPPERT, D. and CARROLL, R. J. (1992). On one-step *GM*-estimates and stability of inferences in linear regressions. *J. Amer. Statist. Assoc.* **87** 439–450.

DEPARTMENT OF MATHEMATICS
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE 0511

DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS 61801