Least Median of Squares Regression

PETER J. ROUSSEEUW*

Classical least squares regression consists of minimizing the sum of the squared residuals. Many authors have produced more robust versions of this estimator by replacing the square by something else, such as the absolute value. In this article a different approach is introduced in which the sum is replaced by the *median* of the squared residuals. The resulting estimator can resist the effect of nearly 50% of contamination in the data. In the special case of simple regression, it corresponds to finding the narrowest strip covering half of the observations. Generalizations are possible to multivariate location, orthogonal regression, and hypothesis testing in linear models.

KEY WORDS: Least squares method; Outliers; Robust regression; Breakdown point.

1. INTRODUCTION

The classical linear model is given by $y_i = x_{i1}\theta_1 + \cdots + x_{ip}\theta_p + e_i (i = 1, \ldots, n)$, where the error e_i is usually assumed to be normally distributed with mean zero and standard deviation σ. The aim of multiple regression is to estimate $\theta = (\theta_1, \ldots, \theta_p)^t$ from the data $(x_{i1}, \ldots, x_{ip}, y_i)$. The most popular estimate $\hat{\theta}$ goes back to Gauss or Legendre (see Stigler 1981 for a recent historical discussion) and corresponds to

$$\underset{\hat{\theta}}{\text{minimize}} \sum_{i=1}^{n} r_i^2, \tag{1.1}$$

where the residuals r_i equal $y_i - x_{i1}\hat{\theta}_1 - \cdots - x_{ip}\hat{\theta}_p$. Legendre called it the *method of least squares* (LS), and it became a cornerstone of statistics. But in spite of its mathematical beauty and computational simplicity, this estimator is now being criticized more and more for its dramatic lack of robustness. Indeed, one single outlier can have an arbitrarily large effect on the estimate. In this connection Hampel (1971) introduced the notion of the *breakdown point* ϵ^*, extending a definition of Hodges (1967): ϵ^* is the smallest percentage of contaminated data that can cause the estimator to take on arbitrarily large aberrant values. In the case of least squares, $\epsilon^* = 0$.

A first step toward a more robust regression estimator came from Edgeworth (1887), improving a proposal of Boscovich. His *least absolute values* or L_1 criterion is

$$\underset{\hat{\theta}}{\text{minimize}} \sum_{i=1}^{n} |r_i|. \tag{1.2}$$

This generalizes the median of a one-dimensional sample and, therefore, has to be made unique (Harter 1977). But whereas the breakdown point of the sample median is 50%, it can be shown that L_1 regression yields the same value $\epsilon^* = 0$ as LS. Although L_1 regression protects against outlying y_i, it cannot cope with grossly aberrant values of $x_i = (x_{i1}, \ldots, x_{ip})$, which have a large influence (called *leverage*) on the fit.

The next step in this direction was the *M estimator* (Huber 1973, p. 800), based on the idea of replacing r_i^2 in (1.1) by $\rho(r_i)$, where ρ is a symmetric function with a unique minimum at zero. Unlike (1.1) or (1.2), however, this is not invariant with respect to a magnification of the error scale. Therefore one often estimates the scale parameter simultaneously:

$$\sum_{i=1}^{n} \psi(r_i/\hat{\sigma})\, x_i = 0 \tag{1.3}$$

$$\sum_{i=1}^{n} \chi(r_i/\hat{\sigma}) = 0, \tag{1.4}$$

where ψ is the derivative of ρ and χ is a symmetric function. (Finding the simultaneous solution of this system of equations is not trivial, and in practice one uses an iteration scheme based on reweighted least squares or Newton–Raphson.) Motivated by minimax asymptotic variance arguments, Huber proposed to use the function

$$\psi(u) = \min(k, \max(u, -k)),$$

where k is some constant, usually around 1.5. As a consequence, such M estimators are statistically more efficient than L_1 at a central model with Gaussian errors. However, again $\epsilon^* = 0$ because of the possibility of leverage points.

Because of this vulnerability to leverage points, *generalized M estimators* (GM estimators) were introduced, with the basic purpose of bounding the influence of outlying x_i, making use of some weight function w. Mallows (1975) proposed to replace (1.3) by

$$\sum_{i=1}^{n} w(x_i)\psi(r_i/\hat{\sigma})\, x_i = 0, \tag{1.5}$$

* Peter J. Rousseeuw is Professor, Department of Mathematics and Informatics, Delft University of Technology, Julianalaan 132, 2628 BL Delft, The Netherlands. This work was supported by the Belgian National Science Foundation. The author is grateful to W. Stahel and F. Hampel for bringing the subject of high breakdown regression to his attention and providing interesting discussions. Helpful comments and suggestions were made by P. Bickel, D. Donoho, D. Hoaglin, P. Huber, R.D. Martin, R.W. Oldford, F. Plastria, A. Samarov, A. Siegel, J. Tukey, R. Welsch, V. Yohai, and two referees. Special thanks go to A. Leroy for assistance with the programming.

December 1984, Volume 79, Number 388
Theory and Methods Section

whereas Schweppe (see Hill 1977) suggested using

$$\sum_{i=1}^{n} w(x_i)\psi(r_i/(w(x_i)\hat{\sigma}))\, x_i = 0. \quad (1.6)$$

Making use of influence functions, good choices of ψ and w were made (Hampel 1978; Krasker 1980; Krasker and Welsch 1982). It turns out, however, that the GM estimators now in use have a breakdown point of at most $1/(p + 1)$, where p is the dimension of x_i (Maronna, Bustos, and Yohai 1979; Donoho and Huber 1983). Various other estimators have been proposed by Theil (1950), Brown and Mood (1951), Sen (1968), Jaeckel (1972), and Andrews (1974); but none of them achieves $\epsilon^* = 30\%$ in the case of simple regression ($p = 2$).

All of this raises the question whether robust regression with a high breakdown point is at all possible. The affirmative answer was given by Siegel (1982), who proposed the *repeated median* with a 50% breakdown point. Indeed, 50% is the best that can be expected (for larger amounts of contamination, it becomes impossible to distinguish between the "good" and the "bad" parts of the sample). Siegel's estimator is defined as follows: For any p observations $(x_{i_1}, y_{i_1}), \ldots, (x_{i_p}, y_{i_p})$, which determine a unique parameter vector, the jth coordinate of this vector is denoted by $\theta_j(i_1, \ldots, i_p)$. The repeated median is then defined coordinatewise as

$$\hat{\theta}_j = \underset{i_1}{\text{med}}(\cdots (\underset{i_{p-1}}{\text{med}}(\underset{i_p}{\text{med}}\, \theta_j(i_1, \ldots, i_p))) \cdots). \quad (1.7)$$

This estimator can be calculated explicitly, but is not equivariant for linear transformations of the x_i. It was applied to a biological problem by Siegel and Benson (1982).

Let us now return to (1.1). A more complete name for the LS method would be *least sum of squares*, but apparently few people have objected to the deletion of the word "sum"—as if the only sensible thing to do with n positive numbers would be to add them. Perhaps as a consequence of this historical name, most people have tried to make this estimator robust by replacing the square by something else, not touching the summation sign. Why not, however, replace the sum by a median, which is very robust? This yields the *least median of squares* (LMS) estimator, given by

$$\underset{\hat{\theta}}{\text{minimize}}\ \underset{i}{\text{med}}\, r_i^2. \quad (1.8)$$

This proposal is essentially based on an idea of Hampel (1975, p. 380). In the next section it is shown that the LMS satisfies $\epsilon^* = 50\%$ but has a very low efficiency. In Section 4 some variants with higher efficiency are given.

2. PROPERTIES OF THE LEAST MEDIAN OF SQUARES METHOD

We shall now investigate the behavior of the LMS technique. The n observations $(x_i, y_i) = (x_{i1}, \ldots, x_{ip}, y_i)$ belong to the linear space of row vectors of dimension $p + 1$, and the unknown parameter θ is a p-dimensional column vector $(\theta_1, \ldots, \theta_p)^t$. The unperturbed linear model states that $y_i = x_i\theta + e_i$, where e_i is distributed according to $N(0, \sigma)$. Throughout this section it is assumed that all observations with $x_i = 0$ have been deleted, because they give no information on θ. (This condition is automatically satisfied if the model has an intercept because then the last coordinate of each x_i equals 1.) Moreover, it is assumed that in the $(p + 1)$-dimensional space of the (x_i, y_i), there is no vertical hyperplane containing more than $[n/2]$ observations. (Here, a vertical hyperplane is a p-dimensional subspace that contains $(0, \ldots, 0)$ and $(0, \ldots, 0, 1)$. The notation $[r]$ stands for the largest integer less than or equal to r.) The proofs of the following results can be found in the Appendix.

Lemma 1. There always exists a solution to (1.8).

In what follows we shall say the observations are in general position when any p of them give a unique determination of θ. For example, in the case in which $p = 2$, this means that any pair of observations (x_{i1}, x_{i2}, y_i) and (x_{j1}, x_{j2}, y_j) determines a unique nonvertical plane through zero, which implies that $(0, 0, 0)$, (x_{i1}, x_{i2}, y_i), and (x_{j1}, x_{j2}, y_j) cannot be collinear. When the observations come from continuous distributions, this event has probability one.

Let us now discuss the breakdown properties of the LMS method. Hampel's (1971) original definition of the breakdown point was asymptotic in nature. In this article, however, I use a version introduced by Donoho and Huber (1983) that is intended for finite samples, like the precursor ideas of Hodges (1967). Take any sample X of n data points (x_i, y_i) and a regression estimator T. Let $\beta(m; T, X)$ be the supremum of $\| T(X') - T(X) \|$ for all corrupted samples X', where any m of the original data points are replaced by arbitrary values. Then the breakdown point of T at X is

$$\epsilon^*(T, X) = \min\{m/n;\ \beta(m; T, X) \text{ is infinite}\}. \quad (2.1)$$

In other words, it is the smallest amount of contamination that can cause the estimator to take on values arbitrarily far from $T(X)$. Note that this definition contains no probability distributions! For least squares, $\epsilon^*(T, X) = 1/n$ because one bad observation can already cause breakdown. For least median of squares, however, this is no longer the case.

Theorem 1. If $p > 1$ and the observations are in general position, then the breakdown point of the LMS method is $([n/2] - p + 2)/n$.

Note that the breakdown point depends only slightly on n. To have only a single value, one often considers the limit for $n \to \infty$ (with p fixed); so it can be said that LS has a breakdown point of 0%, whereas the breakdown point of the LMS technique is as high as 50%, the best that can be expected. The following corollary gives a spe-

cial case that shows the large resistance of the LMS method.

Corollary 1. If $p > 1$ and there exists some θ such that at least $n - [n/2] + p - 1$ of the observations satisfy $y_i = x_i\theta$ exactly and are in general position, then the LMS solution equals θ whatever the other observations are.

Remark 1. The breakdown point in Theorem 1 is slightly smaller than that of the repeated median, although they are both 50% breakdown estimators. I am indebted to A. Siegel (personal communication) for a way to overcome this. Instead of taking the median of the ordered squared residuals, consider the kth order statistic $(r^2)_{k:n}$, where $k = [n/2] + [(p + 1)/2]$, and minimize $(r^2)_{k:n}$. It turns out (analogous to the proof of Theorem 1) that this variant of the LMS has breakdown point $([(n - p)/2] + 1)/n$, which is exactly the same value as for Siegel's repeated median. In the Appendix, it is shown that this is the maximal value for all regression-equivariant estimators. (By regression equivariance, I mean the property

$$T(\{(x_i, y_i + x_i v);\quad i = 1, \ldots, n\}) = T(\{(x_i, y_i);\quad i = 1, \ldots, n\}) + v$$

for any vector v.) For this variant of the LMS, Corollary 1 holds whenever strictly more than $\frac{1}{2}(n + p - 1)$ of the observations are in an exact fit situation, which also corresponds to the repeated median.

It is well known that the LS estimator reduces to the arithmetic mean in the special case of one-dimensional estimation of location, obtained by putting $p = 1$ and $x_i = 1$ for all i. Interestingly, in that special case, the LMS estimator also corresponds to something we know.

Theorem 2. Let $p = 1$ and all $x_i = 1$, so the sample reduces to $(y_i)_{i=1,\ldots,n}$. If

$$m_T^2 : \operatorname*{med}_i r_i^2 = \operatorname*{med}_i (y_i - T)^2$$

equals $\min_\theta \operatorname{med}_i (y_i - \theta)^2$,

then both $T - m_T$ and $T + m_T$ are observations in the sample.

Theorem 2 makes it easy to determine T in the location case because one has to determine only the shortest half of the sample. (This is done by finding the smallest of the values

$$y_{h:n} - y_{1:n}, y_{h+1:n} - y_{2:n}, \ldots, y_{n:n} - y_{n-h+1:n},$$

where $h = [n/2] + 1$ and $y_{1:n} \le y_{2:n} \le \cdots \le y_{n:n}$ are the ordered observations.) By Theorem 2 T simply equals the midpoint of this shortest interval. (In case there are several shortest halfs, which happens with probability zero when the distribution is continuous, one could take the average of their midpoints.) This is reminiscent of the estimator called *shorth* in the Princeton Monte Carlo study (Andrews et al. 1972), where the mean of all of the observations in the shortest half is taken. The shorth converges like $n^{-1/3}$; therefore its influence function is not well defined, but its finite-sample breakdown behavior is extremely good (Andrews et al. 1972, pp. 50, 33, 103). All of these properties are shared by the LMS (details on its asymptotic behavior can be found in the Appendix). The fact that the LMS converges like $n^{-1/3}$ does not trouble me very much, because I consider the LMS mainly as a data analytic tool, for which statistical efficiency is not the most important criterion. In Section 4 I will construct variants with higher efficiency. Although the LMS has no well-defined influence function, it is possible to get some idea of its local robustness properties by constructing stylized sensitivity curves, as was done by Andrews et al. (1972, p. 101) for the shorth. For the LMS this yields Figure 1 for $n = 10$; for larger sample sizes the upward and downward peaks become thinner and higher.

Theorem 2 can also be used in the more general case of regression with a constant, obtained by putting $x_{i,p} = 1$ for all i. From Theorem 2, it follows that for an LMS solution, both hyperplanes $y = x\hat{\theta} - m_T$ and $y = x\hat{\theta} + m_T$ contain at least one observation.

In the special case of simple regression, there are only a single independent variable and a single dependent variable to be fitted to the model $y_i = ax_i + b + e_i$. One therefore has to find the slope and the intercept of a line determined by n points in the plane. By Theorem 2 the LMS solution corresponds to finding the narrowest strip covering half of the observations. (To be exact the thickness of the strip is measured in the vertical direction, and

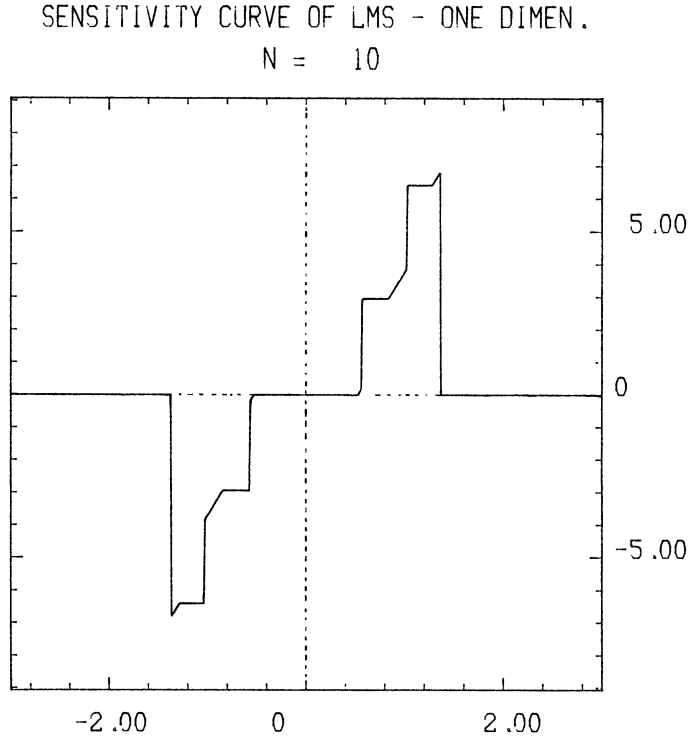

Figure 1. Stylized Sensitivity Curve of the LMS in One Dimension, Given by SC(x) = $n(T_n(y_1, \ldots, y_{n-1}, x) - T_{n-1}(y_1, \ldots, y_{n-1}))$, Where $y_i = \Phi^{-1}(i/n)$, for $n = 10$.

we want at least $h = [n/2] + 1$ points on it.) It is easy to write a computer program to determine the LMS line, because for each value of a, the above algorithm for the location case can be used to calculate

$$m_a^2 = \min_b \operatorname{med}_i ((y_i - ax_i) - b)^2$$

immediately. Therefore, one only has to minimize the continuous function m_a^2 of the single variable a. This technique will be used in Section 3. For data analytic purposes, D. Hoaglin (personal communication) proposed to compute not only the global minimum of the objective function but also the second best local minimum (if it exists), because this might reflect a possible ambiguity in the data.

For problems in higher dimensions, another program has been developed (Leroy and Rousseeuw 1984), making use of brute force minimization of the objective function $\operatorname{med}_i r_i^2$. Table 1 lists some computation times on the CDC computer of the University of Brussels for different values of n and p. (For larger values of n, the time is roughly proportional to n for fixed p.) These times are large, which is not so surprising in view of the relationship of the LMS to the projection pursuit technique (Friedman and Tukey 1974). Indeed, consider the $(p + 1)$-dimensional space of observations (x_i, y_i). We want to find a direction, given by some vector $(\theta, -1)$, such that the projection of the data on the y axis, in the direction orthogonal to $(\theta, -1)$, possesses the smallest dispersion (measured by the median of squares).

In addition to the regression coefficients $\theta_1, \ldots, \theta_p$, the scale parameter σ has to be estimated in a robust way. Once the LMS solution T has been found, with $m_T^2 : \min_\theta \operatorname{med}_i r_i^2$, a natural estimator for σ is

$$S = 1.483\, c(n, p)\, m_T, \tag{2.2}$$

where $1/\Phi^{-1}(.75) \simeq 1.483$ is an asymptotic correction factor for the case of normal errors, because then

$$\operatorname{med}_i r_i^2 \to \sigma^2 \operatorname{med}(\chi_1^2) = \sigma^2(\Phi^{-1}(.75))^2,$$

where Φ denotes the standard normal cumulative. The constant $c(n, p)$ is a finite-sample correction factor larger than 1, which is necessary to make S approximately unbiased when simulating samples with normal errors.

Table 1. Computation Times for the LMS Multiple Regression Program With Intercept for Different n and p*

	p								
n	2	3	4	5	6	7	8	9	10
25	.23	1.28	1.95	2.89	4.13	4.91	5.74	6.76	7.79
50	1.52	2.08	3.16	4.50	5.98	6.89	7.82	8.87	10.02
100	2.34	3.82	5.55	7.65	10.10	11.02	12.11	13.45	16.83
150	3.39	5.51	8.05	10.89	14.00	15.15	16.50	17.98	19.56
200	4.47	7.16	10.28	13.84	17.82	19.33	20.85	22.39	24.42

* In CP seconds on a CDC 750 computer.

Work is in progress to determine $c(n, p)$ empirically. For n tending to infinity, $c(n, p)$ converges to 1.

3. EXAMPLES

In the first example, 30 "good" observations were generated according to the linear relation $y_i = ax_i + b + e_i$, where $a = 1$, $b = 2$, x_i is uniformly distributed on (1, 4), and e_i is normally distributed with mean zero and standard deviation .2. The number p of coefficients to be estimated therefore equals 2. Then a cluster of 20 "bad" observations were added, possessing a spherical bivariate normal distribution with mean (7, 2) and standard deviation $\frac{1}{2}$. This yielded 40% of contamination in the pooled sample, which is high. This amount was actually chosen to demonstrate what happens if one goes above the upper bound $1/(p + 1) \simeq 33.3\%$ on the breakdown point of the GM estimators now in use.

Let us now see which estimator succeeds best in describing the pattern of the majority of the data. The classical least squares method yields $\hat{a} = -.47$ and $\hat{b} = 5.62$: it clearly fails because it tries to suit both the good and the bad data points. Making use of the ROBETH library of subroutines (Marazzi 1980), three robust estimators were applied: Huber's M estimator (1.3)–(1.4) with $\psi(x) = \min(1.5, \max(-1.5, x))$, Mallows's GM estimator (1.5) with Hampel weights, and Schweppe's GM estimator (1.6) with Hampel–Krasker weights (both Mallows's and Schweppe's using the same Huber function ψ). All three methods, however, gave results virtually indistinguishable from the LS solution: the four lines almost coincide in Figure 2. The repeated median estimator (1.7) yields $\hat{a} = .30$ and $\hat{b} = 3.11$. If the cluster of "bad" points is moved further down, the repeated median line follows it a little more and then stops. Therefore this method does not break down. Finally, the LMS (1.8), calculated by means of the algorithm for simple regression described in Section 2, yields $\hat{a} = .97$ and $\hat{b} = 2.09$, which comes close to the original values of a and b. When the cluster of bad points is moved further away, this solution does not change. Moreover, the LMS method does not break down even when only 26 "good" points and 24 outliers are used.

It may seem unfair to consider such large amounts of contamination (although they sometimes occur, e.g., in the case of ancient astronomical observations (Huber 1974) or in certain sloppy medical data sets). The breakdown point of the currently used GM estimators, however, is less than $1/(p + 1)$, which is small in problems with several independent variables; so very common amounts of contamination already necessitate the use of a more robust regression estimator. Moreover, it still has to be investigated empirically whether the upper bound $1/(p + 1)$ on the asymptotic breakdown point can actually be reached in finite sample situations.

Note that looking at the least squares residuals (possibly followed by a rejection of outlying ones) is not sufficient. In fact the least squares fit often masks bad data

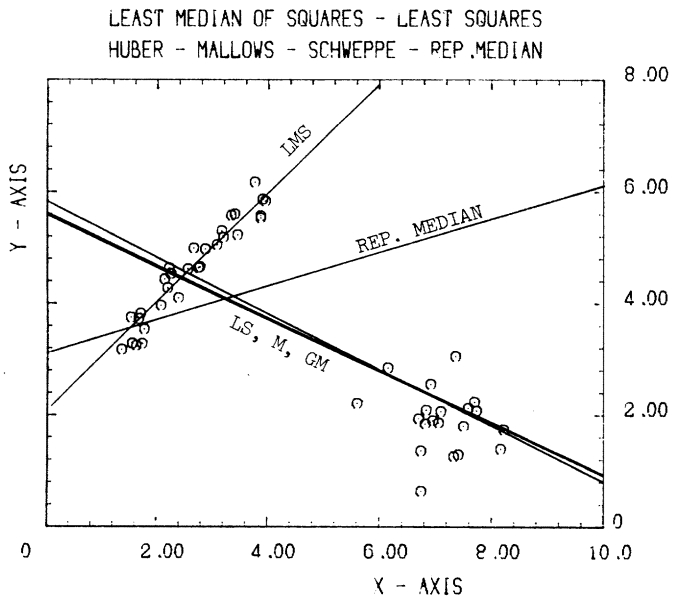

Figure 2. Regression Lines for the Simulated Data of the First Example, Using Six Methods. (LMS = least median of squares; LS = least squares; M = Huber's M estimator; GM = Mallows's and Schweppe's G-M estimator; REP. MEDIAN = repeated median; ⊙ = 30 "good" points generated according to a linear relation $y_i = x_i + 2 + e_i$ and 20 "bad" points in a spherical cluster around (7, 2).

points: in Figure 2, the largest LS residuals correspond to good data! In problems with several variables, a very robust estimator like the LMS can be used for finding the outlying observations, as shall be seen in the next example.

When faced with a practical application, it seems like a good idea to run both an LMS and an LS regression. If they agree closely, the LS result can be trusted. If, on the other hand, there is a significant difference, then we know which observations are responsible by looking at the LMS residuals.

Let us now look at a second example, containing multidimensional real data. It seems that an entirely real example with "messy" data might not be completely convincing, because we would end up with different results for LS and LMS without a conclusive way to decide which analysis is best, possibly causing some debates. Therefore, we start with a real data set that is rather well behaved and contaminate it by replacing a few observations. It would be easy to illustrate the resistance of the LMS by throwing in some very bad outliers, but I would like to put the LMS to a harder test by considering a more delicate situation. To show that the LMS also works in small samples, I selected a data set containing 20 points with six parameters to be estimated. The raw data came from Draper and Smith (1966, p. 227) and were used to determine the influence of anatomical factors on wood specific gravity, with five independent variables and an intercept (Draper and Smith conclude that x_{i2} could be deleted from the model, but this matter is not considered for the present purpose). Table 2 lists a contaminated version of these data, in which a few observations have been replaced by outliers. Applying least squares yields

$$\hat{y}_i = .44069\, x_{i1} - 1.47501\, x_{i2} - .26118\, x_{i3} + .02079\, x_{i4} + .17082\, x_{i5} + .42178.$$

Table 2 lists the LS residuals, divided by the LS scale estimate $\hat{\sigma}_{LS} = .02412$. It is not easy to spot the outliers just by looking at the observations, and the LS result (without a more detailed analysis) is of little help. Indeed, the standardized residuals look very inconspicuous, ex-

Table 2. Modified Data on Wood Specific Gravity With Standardized Residuals From Least Squares and Least Median of Squares

								Residual/Scale	
i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	y_i	LS	LMS
1	.5730	.1059	.4650	.5380	.8410	1.000	.5340	−.7250	−.0827
2	.6510	.1356	.5270	.5450	.8870	1.000	.5350	.0472	.0013
3	.6060	.1273	.4940	.5210	.9200	1.000	.5700	1.2427	.2836
4	.4370	.1591	.4460	.4230	.9920	1.000	.4500	.3547	−7.6137
5	.5470	.1135	.5310	.5190	.9150	1.000	.5480	1.0024	.9020
6	.4440	.1628	.4290	.4110	.9840	1.000	.4310	−.4518	−9.1023
7	.4890	.1231	.5620	.4550	.8240	1.000	.4810	.9067	.3746
8	.4130	.1673	.4180	.4300	.9780	1.000	.4230	−.0349	−8.9077
9	.5360	.1182	.5920	.4640	.8540	1.000	.4750	−.3959	−.3746
10	.6850	.1564	.6310	.5640	.9140	1.000	.4860	−.4150	−.2071
11	.6640	.1588	.5060	.4810	.8670	1.000	.5540	1.9859	.0013
12	.7030	.1335	.5190	.4840	.8120	1.000	.5190	−1.1977	−.9656
13	.6530	.1395	.6250	.5190	.8920	1.000	.4920	−.4854	.0013
14	.5860	.1114	.5050	.5650	.8890	1.000	.5170	−1.2612	−.6709
15	.5340	.1143	.5210	.5700	.8890	1.000	.5020	−.5866	−.1733
16	.5230	.1320	.5050	.6120	.9190	1.000	.5080	.5237	.0013
17	.5800	.1249	.5460	.6080	.9540	1.000	.5200	−.2548	.0013
18	.4480	.1028	.5220	.5340	.9180	1.000	.5060	.2838	−.1090
19	.4170	.1687	.4050	.4150	.9810	1.000	.4010	−1.0837	−10.7265
20	.5280	.1057	.4240	.5660	.9090	1.000	.5680	.5450	.0013

cept for observation 11 (and this is a false trail). Because of this, many people would probably be satisifed with the LS fit (especially when not expecting trouble). By means of the subroutines that are presently at my disposal, I was unable to obtain a GM estimate essentially different from the LS one, although it is possible that a more refined GM program would do the job. The LMS estimate can also be computed, however, as described in Remark 1 of Section 2, taking $c(20, 6) = 1.8$ in (2.2). This yields

$$\hat{y}_i = .26870\, x_{i1} - .23806\, x_{i2} - .53572\, x_{i3} - .29373\, x_{i4} + .45096\, x_{i5} + .43474.$$

Now look at the LMS residuals divided by the LMS scale estimate .0195, which are given in the last column of Table 2. These standardized residuals make it easy to spot the four outliers. This example illustrates the use of the LMS as a data analytic tool: as a next step in the analysis, LS could be computed again without these four observations.

4. RELATED APPROACHES

A disadvantage of the LMS method is its lack of efficiency because of its $n^{-1/3}$ convergence. Of course it is possible to take an extreme point of view, wanting to stay on the safe side, even if it costs a lot. However, it is not so difficult to improve the efficiency of the LMS estimator. One first has to calculate the LMS estimate T and the corresponding scale estimate S given by (2.2). With these starting values, one can compute a *one-step M estimator* (Bickel 1975). If one uses a redescending ψ function, like the one of the hyperbolic tangent estimator (Hampel, Rousseeuw, and Ronchetti 1981) or the biweight (Beaton and Tukey 1974), the large outliers will not enter into the computation. Such a one-step M estimator converges like $n^{-1/2}$ and possesses the same asymptotic efficiency (for normal errors) as a fully iterated M estimator. This was proven by Bickel (1975) when the starting value was $n^{1/2}$ consistent, but in general it even holds when the starting value is better than $n^{1/4}$ consistent (Bickel, personal communication, 1983). In particular, the combined procedure (LMS + one-step M) achieves the asymptotic efficiency $e = (\int \psi' d\Phi)^2/(\int \psi^2 d\Phi)$. For instance, the choice $c = 4.0$ and $k = 4.5$ in table 2 of Hampel, Rousseeuw, and Ronchetti (1981) already yields an efficiency of more than 95%.

Another possibility is to use reweighted least squares. To each observation (x_i, y_i), one assigns a weight w_i that is a nonincreasing function of $|r_i/S|$ and that becomes zero starting from, say, $|r_i/S|$ equal to three or four. Then one replaces all observations (x_i, y_i) by $(w_i^{1/2}x_i, w_i^{1/2}y_i)$, which means that points with large LMS residuals disappear entirely. On these weighted observations, a standard least squares program is used to obtain the final estimate. Actually, applying this estimator to the data in Section 3 amounts to deleting the 20 "bad" points from the first example and deleting the four outliers from the second.

The main idea of the LMS is to minimize the scatter of the residuals. From this point of view, there already exist other estimators with a similar objective, using other measures of scatter. For instance, LS (1.1) minimizes the mean square of the residuals, and L_1 (1.2) minimizes their mean deviation (from zero). Jaeckel's (1972) estimator is also defined by minimizing a dispersion measure of the residuals. For this purpose, Jaeckel uses linear combinations of the ordered residuals. Since these estimators of scale are translation invariant, it is not possible to estimate a constant term in the regression model. Although these scale estimates can be quite robust, their largest breakdown point (with regard to both explosion and implosion of $\hat{\sigma}$) is $\epsilon^* = 25\%$, which is achieved by the interquartile range. Therefore, the breakdown point of a Jaeckel estimator is at most 25%. A way to improve this would be to replace the scale estimator by another one having a larger breakdown point. For this reason Hampel (1975) suggested using the median absolute deviation, which essentially yields the LMS. The LMS, indeed, possesses a 50% breakdown point but unfortunately only converges like $n^{-1/3}$. This slow rate of convergence can be improved by computing a one-step M estimator afterwards, as seen earlier in this section. There is also another way to achieve this, by using a different objective function. The *least trimmed squares* (LTS) estimator (Rousseeuw in press) is given by

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{h} (r^2)_{i:n}, \tag{4.1}$$

where $(r^2)_{1:n} \leq \cdots \leq (r^2)_{n:n}$ are the ordered squared residuals. If $h = [n/2] + 1$ is chosen, the breakdown point of Theorem 1 is obtained, and for $h = [n/2] + [(p + 1)/2]$, the result of Remark 1 holds. In general, h may depend on some trimming proportion α, for instance by means of $h = [n(1 - \alpha)] + 1$. The LTS converges like $n^{-1/2}$ (Rousseeuw in press), with the same asymptotic efficiency at the normal distribution as the M estimator defined by $\psi(x) = x$ for $|x| \leq \Phi^{-1}(1 - (\alpha/2))$ and $\psi(x) = 0$ otherwise, which is called the *Huber skipped mean*. The main disadvantage of the LTS is that its objective function requires sorting of the squared residuals, which takes $O(n \log n)$ operations compared with only $O(n)$ operations for the median, hereby blowing up the already large computation times in Table 1. Another possibility is to use the so-called *S estimators* defined by

$$\underset{\theta}{\text{minimize}}\ S(\theta), \tag{4.2}$$

where $S(\theta)$ is a certain type of M estimator of scale on the residuals $r_1(\theta), \ldots, r_n(\theta)$. (These estimators are now being investigated in collaboration with V. Yohai, following similar suggestions by J. Tukey and R. Martin (personal communications).) It appears that S estimators have essentially the same asymptotic behavior as regression M estimators, but they can also achieve a high breakdown point.

In the case of simple regression, we saw that the LMS

method corresponds to finding the narrowest strip covering half of the points. Taking the word "narrowest" literally amounts to replacing the usual squared residuals by the squared distances from the observations to the fitted line. This way the LMS can be generalized to orthogonal regression. Another generalization can be made to hypothesis testing in linear models, replacing the usual sums of squared residuals by their medians (or, better still, by the objective function of (4.1) or the square of (4.2)). The ratio of two such medians would then be compared to critical values, possibly obtained by simulation.

The LMS technique can also be used to estimate the location of a spherical multivariate distribution. If the sample $x_1, \ldots, x_n$ consists of p-dimensional vectors, then the LMS estimator T is defined by

$$\underset{T}{\text{minimize}} \underset{i}{\text{med}} \| T - x_i \|^2, \tag{4.3}$$

which corresponds to finding the center of the smallest disk (or ball) covering half of the points. The LTS analog would be to minimize the sum of the first h ordered values of $\| T - x_i \|^2$. Both procedures have a breakdown point of 50% and are equivariant with respect to magnification factors, translations, and orthogonal transformations.

Recently a more difficult problem has received attention, namely to construct a high breakdown estimator of multivariate location that is equivariant for affine transformations, which means that $T(Ax_1 + b, \ldots, Ax_n + b) = AT(x_1, \ldots, x_n) + b$ for any vector b and any nonsingular matrix A. The first solution to this problem was obtained independently by Stahel (1981) and Donoho (1982). For each observation x_i, one looks for a one-dimensional projection in which x_i is most outlying and then downweights x_i according to this worst result. The Stahel–Donoho estimator is then defined as the weighted mean of the observations, which is affine equivariant and possesses a breakdown point of 50%. Making use of the same weights, they also compute a covariance estimator. We shall now see that it is also possible to generalize the LMS to meet the joint objectives of affine equivariance and 50% breakdown point: Define T as the center of the minimal volume ellipsoid covering at least h observations (Rousseeuw in press). The corresponding generalization of the LTS, which also yields $\epsilon^* = 50\%$, takes the mean of the h points for which the determinant of the covariance matrix is minimal. Both the minimal volume ellipsoid estimator and the minimal covariance determinant estimator yield robust covariance estimators at the same time (for instance, by computing the covariance of the selected h observations, multiplied with a correction factor to obtain consistency in the case of multivariate normality).

It is possible to use the LMS idea in multivariate location, but one can also extend the Stahel–Donoho estimator to regression. For each observation (x_i, y_i), one could define u_i as the supremum of

$$| r_i(\theta) | / \underset{j}{\text{med}} | r_j(\theta) |$$

over all possible vectors θ. Then one could assign a weight $w(u_i)$ to each point and compute a weighted least squares estimator. Unfortunately this technique involves an optimization for each data point, so it requires about n times the computational cost of the LMS.

Finally, a word of caution. Some people have objected to the notion of the breakdown point, on the ground that it is very crude, and have pointed out the possibility of constructing high breakdown procedures that are not good on other grounds (Oldford 1983). It is indeed true that the breakdown point is only one out of several criteria; so a high breakdown point alone is not a sufficient condition for a good method. I consider a good breakdown point to be more like a necessary condition: If it is not satisfied because the procedure is vulnerable to a certain type of contamination (such as leverage points in regression), one cannot guarantee that such contamination will never occur in practice.

APPENDIX

Proof of Lemma 1

For this proof, we work in the $(p + 1)$-dimensional space E of the observations (x_i, y_i). The space of the x_i is the horizontal hyperplane through the origin, which is denoted by $(y = 0)$ because the y coordinates of all points in this plane are zero (in a space E of dimension $p + 1$, a hyperplane is a p-dimensional subspace). There are two cases.

Case A. This is really a special case for which there exists a $(p - 1)$-dimensional subspace of V of $(y = 0)$ containing (at least) $[n/2] + 1$ of the x_i. The observations (x_i, y_i) corresponding to these x_i now generate a subspace S of E (in the sense of linear algebra), which is at most p dimensional. Because it was assumed that E has no vertical hyperplane containing $[n/2] + 1$ observations, it follows that S does not contain $(0, 1)$; hence the dimension of S is at most $p - 1$. This means that there exists a nonvertical hyperplane H given by some equation $y = x\theta$, which includes S. For this value of θ, clearly $\text{med}_i\, r_i^2 = 0$, which is the minimal value.

Case B. Let us now assume that we are in the general situation in which case A does not hold. The rest of the proof will be devoted to showing that there exists a ball around the origin, which may be very large, to which attention can be restricted for finding a minimum of $\text{med}_i\, r_i^2(\theta)$. Because $\text{med}_i\, r_i^2(\theta)$ is continuous in θ, this is sufficient for the existence of a minimum. Put $\delta = \frac{1}{2} \inf\{\eta > 0$; there exists a $(p - 1)$-dimensional subspace V of $(y = 0)$ such that V^η covers (at least) $[n/2] + 1$ of the $x_i\}$. Here V^η is the set of all x with distance to V not larger than η. Because we are not in case A, $\delta > 0$. Also $M : \max_i | y_i |$. Now attention may be restricted to the ball around the origin with radius $(\sqrt{2} + 1)M/\delta$. Indeed, for any θ with $\| \theta \| > (\sqrt{2} + 1)M/\delta$, it will be shown that

$$\text{med}_i\, r_i^2(\theta) > \text{med}_i\, y_i^2 = \text{med}_i\, r_i^2(0),$$

so smaller objective functions cannot be found outside the ball. A geometrical construction is needed to prove this. Such a θ determines a nonvertical hyperplane H given by $y = x\theta$. By the dimension theorem of linear algebra, the intersection $H \cap (y = 0)$ has dimension $p - 1$. Therefore $(H \cap (y = 0))^\delta$ contains at most $[n/2]$ of the x_i. For each of the remaining observations (x_i, y_i), we construct the vertical two-dimensional plane through (x_i, y_i) and orthogonal to $(H \cap (y = 0))$. (This plane does not pass through zero, so to be called vertical, it has to go through both (x_i, y_i) and $(x_i, y_i + 1)$.) We see that

$$| r_i | = | x_i\theta - y_i | \geq \| x_i\theta | - | y_i \|$$

with $| x_i\theta | > \delta | tg(\alpha) |$, where α is the angle (between $-\pi/2$ and $\pi/2$) formed by H and the horizontal line in P_i. Therefore $| \alpha |$ is the angle between the line orthogonal to H and $(0, 1)$; hence

$$| \alpha | = \arccos\{ | (-\theta, 1)(0, 1)^t | / (\| (-\theta, 1) \| \cdot \| (0, 1) \|)\}$$
$$= \arccos(1/(1 + \| \theta \|^2)^{1/2}),$$

and finally, $| tg(\alpha) | = \| \theta \|$. Because $\| \theta \| > (\sqrt{2} + 1)M/\delta$, it follows that $| x_i\theta | > \delta \| \theta \| > M \geq | y_i |$, so $| r_i | > (\delta \| \theta \| - | y_i |)$. But then $r_i^2 > ((\sqrt{2} + 1)M - | y_i |)^2 > 2M^2$ for at least $n - [n/2]$ observations, hence $\text{med}_i (r_i^2) > M^2 \geq \text{med}_i (y_i^2)$. Such θ outside the ball yield an objective function larger than the one for $\theta = 0$; hence such θ can be disregarded.

Proof of Theorem 1

1. We first show that

$$\epsilon^*(T, X) \geq ([n/2] - p + 2)/n$$

for any sample $X = \{(x_i, y_i); i = 1, \ldots, n\}$ consisting of n points in general position. By Lemma 1 the sample X yields a solution θ of (1.8). We now have to show that the LMS remains bounded when $n - ([n/2] - p + 2) + 1$ points are unchanged. For this purpose construct any corrupted sample $X' = \{(x_i', y_i'); i = 1, \ldots, n\}$ by retaining $n - [n/2] + p - 1$ observations of X—which will be called the "good" observations—and by replacing the others by arbitrary values. It suffices to prove that $\| \theta - \theta' \|$ is bounded, where θ' corresponds to X'. For this purpose some geometry is needed. We work in the $(p + 1)$-dimensional space E of the observations (x_i, y_i). The space of the x_i is the horizontal hyperplane through the origin, denoted by $(y = 0)$ because the y coordinates of all points in this plane are zero. (We call this subspace a hyperplane because its dimension is p, which is one less than the dimension of the total space E.) Put $\rho : \frac{1}{2}\inf\{\eta > 0$; there exists a $(p - 1)$-dimensional subspace V of $(y = 0)$ such that V^η covers at least p of the $x_i\}$. Here V^η is the set of all x with distance to V not larger than η. Because X is in general positon, it holds that $\rho > 0$. Also put $M : \max_i | r_i |$, where r_i are the residuals $y_i - x_i\theta$.

The rest of the proof of part 1 will be devoted to showing that $\| \theta - \theta' \| < 2(\| \theta \| + M/\rho)$, which is sufficient because the right member is a finite constant. Denote by H the nonvertical hyperplane given by the equation $y = x\theta$, and let H' correspond in the same way to θ'. Without loss of generality assume that $\theta' \neq \theta$, hence $H' \neq H$. By the dimension theorem of linear algebra, the intersection $H \cap H'$ has dimension $p - 1$. If $\text{pr}(H \cap H')$ denotes the vertical projection of $H \cap H'$ on $(y = 0)$, it follows that at most $p - 1$ of the good x_i can lie on $(\text{pr}(H \cap H'))^\rho$. A is defined as the set of remaining good observations, containing at least $n - [n/2] + p - 1 - (p - 1) = n - [n/2]$ points. Now consider any (x_a, y_a) belonging to A, and put $r_a = y_a - x_a\theta$ and $r_a' = y_a - x_a\theta'$. Construct the vertical two-dimensional plane P_a through (x_a, y_a) and orthogonal to $\text{pr}(H \cap H')$. It follows, as in the proof of Lemma 1, that

$$| r_a' - r_a | = | x_a\theta' - x_a\theta | > \rho | tg(\alpha') - tg(\alpha) |$$
$$\geq \rho \| tg(\alpha') | - | tg(\alpha) \| = \rho | \| \theta' \| - \| \theta \| |,$$

where α is the angle formed by H and some horizontal line in P_a and α' corresponds in the same way to H'. Since

$$\| \theta' - \theta \| \leq \| \theta \| + \| \theta' \| = 2 \| \theta \| + (\| \theta' \| - \| \theta \|)$$
$$\leq | \| \theta' \| - \| \theta \| | + 2 \| \theta \|,$$

it follows that $| r_a' - r_a | > \rho(\| \theta' - \theta \| - 2 \| \theta \|)$. Now the median of squared residuals of the new sample X' with respect to the old θ, with at least $n - [n/2] + p - 1$ of these residuals being the same as before, is less than or equal to M^2. Because θ' is a solution of (1.8) for X', it also follows that $\text{med}_i (y_i' - x_i'\theta')^2 \leq M^2$. If it is now assumed that $\| \theta' - \theta \| \geq 2(\| \theta \| + M/\rho)$, then for all a in A, it holds that

$$| r_a' - r_a | > \rho(\| \theta' - \theta \| - 2 \| \theta \|) \geq 2M,$$

so $| r_a' | \geq | r_a' - r_a | - | r_a | > 2M - M = M$, and finally, $\text{med}_i (y_i' - x_i'\theta')^2 > M^2$, a contradiction. Therefore, $\| \theta' - \theta \| < 2(\| \theta \| + M/\rho)$ for any X'.

2. Let us now show that the breakdown point can be no larger than the announced value. For this purpose, consider corrupted samples in which only $n - [n/2] + p - 2$ of the good observations are retained. Start by taking $p - 1$ of the good observations, which determine a $(p - 1)$-dimensional subspace L of E. Now construct any nonvertical hyperplane H' through L, which determines some θ' through the equation $y = x\theta'$. If all of the "bad" observations are put on H', then X' has a total of $([n/2] - p + 2) + (p - 1) = [n/2] + 1$ points that satisfy $y_i' = x_i'\theta'$ exactly; so the median squared residual of X' with respect to θ' is zero, hence θ' satisfies (1.8) for X'. By choosing H' steeper and steeper, one can make $\| \theta' - \theta \|$ as large as one wants.

Proof of Corollary 1

There exists some θ such that at least $n - [n/2] + p - 1$ of the observations lie on the hyperplane H given by the equation $y = x\theta$. Then θ is a solution of (1.8), because $\text{med}_i r_i^2(\theta) = 0$. Suppose that there is another solution $\theta' \neq \theta$, corresponding to a hyperplane $H' \neq H$ and yield-

ing residuals $r_i(\theta')$. As in the proof of Theorem 1, $(H \cap H')$ has dimension $p - 1$ and thus contains at most $p - 1$ observations. For all remaining observations in H, it holds that $r_i^2(\theta') > 0$ and there are at least $n - [n/2]$ of them. Therefore $\text{med}_i\, r_i^2(\theta') > 0$, so θ' cannot be a solution.

Remark 1 of Section 2

Let us now show that any regression equivariant estimator T satisfies

$$\epsilon^*(T, X) \leq ([(n - p)/2] + 1)/n$$

at all samples $X = \{(x_i, y_i); i = 1, \ldots, n\}$. Suppose that the breakdown point is strictly larger than this value. This would mean that there exists a finite constant b such that $T(X')$ lies in the ball $B(T(X), b)$ for all samples X' containing at least $m : n - [(n - p)/2] - 1$ points of X. Here $B(T(X), b)$ is defined as the set of all θ for which $\| T(X) - \theta \| \leq b$. Now construct a p-dimensional column vector $v \neq 0$ such that $x_1 v = 0, \ldots, x_{p-1} v = 0$. Inspection shows that $2m - (p - 1) \leq n$. Therefore the first $2m - (p - 1)$ points of X can be replaced by

$$(x_1, y_1), \ldots, (x_{p-1}, y_{p-1}), (x_p, y_p), \ldots, (x_m, y_m),$$
$$(x_p, y_p + x_p \lambda v), \ldots, (x_m, y_m + x_m \lambda v)$$

for any $\lambda > 0$. For this new sample X', the estimate $T(X')$ belongs to $B(T(X), b)$. But looking at X' in another way reveals that $T(X')$ can also be written as $T(X'') + \lambda v$, where $T(X'')$ is in $B(T(X), b)$, hence $T(X')$ also belongs to $B(T(X) + \lambda v, b)$. This is a contradiction, however, because the intersection of $B(T(X), b)$ and $B(T(X) + \lambda v, b)$ is empty for large enough values of λ.

Proof of Theorem 2

Case A. First suppose that n is odd, with $n = 2k - 1$. Then $\text{med}(r_i^2)$ is reached by the kth square. Therefore at least one of the points $T + m_T$ or $T - m_T$ is an observation; without loss of generality, suppose that $T + m_T$ is and $T - m_T$ is not. There is a partition of the r_i^2 into $k - 1$ squares $\geq m_T^2$, 1 square $= m_T^2$, and $k - 1$ squares $\leq m_T^2$. Now take the smallest observation y_j that is larger than $T - m_T$ (choose one if there are several), and define

$$T' = \tfrac{1}{2}((T + m_T) + y_j)$$

and

$$m^2 = (\tfrac{1}{2} \mid (T + m_T) - y_j \mid)^2 < m_T^2.$$

Then there is a partition of the $(r_i')^2 = (y_i - T')^2$ into $k - 1$ squares $\geq m^2$ (corresponding to the same points as before), $k - 2$ squares $\leq m^2$ (corresponding to the same points as before, except y_j), and 2 squares $= m^2$. Finally $\text{med}(r_i')^2 = m^2 < m_T^2$, a contradiction.

Case B. Suppose that n is even, with $n = 2k$. If the ordered squares are denoted by $r_{(1)}^2 \leq \cdots \leq r_{(n)}^2$, then

$$m_T^2 = \tfrac{1}{2}(r_{(k)}^2 + r_{(k+1)}^2).$$

There is a partition of the squared residuals into $k - 1$ squares $\leq r_{(k)}^2$, $r_{(k)}^2$ itself, $r_{(k+1)}^2$ itself, and $k - 1$ squares $\geq r_{(k+1)}^2$. If $T + m_T$ is an observation and $T - m_T$ is not (or conversely), we can repeat the reasoning of case A. Now suppose that neither $T + m_T$ nor $T - m_T$ is an observation, which implies that $r_{(k)}^2 < r_{(k+1)}^2$ because otherwise $r_{(k)}^2 = m_T^2 = r_{(k+1)}^2$. Therefore, at least $r_{(k+1)}^2 > 0$.

Case B.1. Assume that $r_{(k)}^2 = 0$. In that case T coincides with exactly k observations, the nearest other observation (call it y_d) being at a distance $| r_{(k+1)} |$. Putting $T' = \frac{1}{2}(T + y_d)$, however, we find

$$\text{med}(y_i - T')^2 = \tfrac{1}{2}((\tfrac{1}{2}r_{(k+1)})^2 + (\tfrac{1}{2}r_{(k+1)})^2)$$
$$= \tfrac{1}{4}r_{(k+1)}^2 < \tfrac{1}{2}r_{(k+1)}^2 = m_T^2,$$

a contradiction.

Case B.2. Assume that $r_{(k)}^2 > 0$. Denote by y_j some observation corresponding to $r_{(k)}^2$ and by y_d some observation corresponding to $r_{(k+1)}^2$. If the observations leading to $r_{(k)}^2$ and $r_{(k+1)}^2$ are all larger than T or all smaller than T, one can again repeat the reasoning of case A. Therefore, one may assume without loss of generality that $y_j < T < y_d$. Putting $T' = \frac{1}{2}(y_j + y_d)$, we find

$$\text{med}(y_i - T')^2 = \tfrac{1}{2}((y_j - T')^2 + (y_d - T')^2)$$
$$< \tfrac{1}{2}((y_j - T)^2 + (y_d - T)^2)$$

because the function $g(t) = (a - t)^2 + (b - t)^2$ attains its unique minimum at $t = \frac{1}{2}(a + b)$.

Asymptotic Behavior of the LMS Estimator

Suppose that the observations $y_1, \ldots, y_n$ are iid according to $F(y - \theta)$, where F is a symmetric and strongly unimodal distribution with density f. Then the distribution of the LMS estimator T_n converges weakly as follows:

$$\mathcal{L}(n^{1/3}(T_n - \theta)) \to \mathcal{L}(A\tau/f(F^{-1}(.75))).$$

Here $A = (\frac{1}{2}\Lambda^2(F^{-1}(.75)))^{-1/3}$, where $\Lambda = -f'/f$ corresponds to the maximum likelihood scores, and τ is the random time s for which $s^2 + Z(s)$ attains its minimum, where $Z(s)$ is a standard Brownian motion. This result is obtained by repeating parts 1, 2, and 3 of the heuristic reasoning of Andrews et al. (1972, p. 51), putting $\alpha = .25$, which yields the constant A. The remaining part of the calculation is slightly adapted, using the same notation. If $\theta = 0$ and $\hat{t}$ is the minimizing value of t, then the main asymptotic variability of T_n is given by

$$\tfrac{1}{2}(F_n^{-1}((\tfrac{1}{2} + \hat{t}) + .25) + F_n^{-1}((\tfrac{1}{2} + \hat{t}) - .25))$$
$$\simeq \hat{t}(F^{-1})'(\tfrac{1}{2} + .25) = \hat{t}/f(F^{-1}(.75)),$$

where $n^{1/3}\hat{t}$ behaves asymptotically like $A\tau$.

[*Received January 1983. Revised January 1984.*]

REFERENCES

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

ANDREWS, D.F., BICKEL, P.J., HAMPEL, F.R., HUBER, P.J., ROGERS, W.H., and TUKEY, J.W. (1972), *Robust Estimates of Location: Survey and Advances*, Princeton, N.J.: Princeton University Press.

BEATON, A.E., and TUKEY, J.W. (1974), "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," *Technometrics*, 16, 147–158.

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

BROWN, G.W., and MOOD, A.M. (1951), "On Median Tests for Linear Hypotheses," *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*, 159–166.

DONOHO, D.L. (1982), "Breakdown Properties of Multivariate Location Estimators," qualifying paper, Harvard University, Statistics Dept.

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

DRAPER, N.R., and SMITH, H. (1966), *Applied Regression Analysis*, New York: John Wiley.

EDGEWORTH, F.Y. (1887), "On Observations Relating to Several Quantities," *Hermathena*, 6, 279–285.

FRIEDMAN, J.H., and TUKEY, J.W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computers*, C-23, 881–889.

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

——— (1975), "Beyond Location Parameters: Robust Concepts and Methods," *Bulletin of the International Statistical Institute*, 46, 375–382.

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

HAMPEL, F.R., ROUSSEEUW, P.J., and RONCHETTI, E. (1981), "The Change-of-Variance Curve and Optimal Redescending M-estimators," *Journal of the American Statistical Association*, 76, 643–648.

HARTER, H.L. (1977), "Nonuniqueness of Least Absolute Values Regression," *Communications in Statistics*, A6, 829–838.

HILL, R.W. (1977), "Robust Regression When There Are Outliers in the Carriers," unpublished Ph.D. dissertation, Harvard University.

HODGES, J.L., JR. (1967), "Efficiency in Normal Samples and Tolerance of Extreme Values for Some Estimates of Location," *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), 163–168.

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

——— (1974), "Early Cuneiform Evidence for the Planet Venus," paper presented at the Annual Meeting of the American Association for the Advancement of Science, Feb. 25, San Francisco.

——— (1981), *Robust Statistics*, New York: John Wiley.

JAECKEL, L.A. (1972), "Estimating Regression Coefficients by Minimizing the Dispersion of Residuals," *Annals of Mathematical Statistics*, 5, 1449–1458.

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.

LEROY, A., and ROUSSEEUW, P. (1984), "PROGRES: A Program for Robust Regression," Research Report No. 201, University of Brussels, Centrum Voor Statistiek en Operationeel Onderzoek.

MALLOWS, C.L. (1975), "On Some Topics in Robustness," unpublished memorandum, Bell Telephone Laboratories, Murray Hill, N.J.

MARAZZI, A. (1980), "Robust Linear Regression Programs in ROBETH," Research Report No. 23, Fachgruppe für Statistik, ETH Zürich.

MARONNA, R.A., BUSTOS, O., and YOHAI, V. (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, 91–116.

OLDFORD, R.W. (1983), "A Note on High Breakdown Regression Estimators," Technical Report, Massachusetts Institute of Technology, Center for Computational Research in Economics and Management Science.

ROUSSEEUW, P.J. (in press), "Multivariate Estimation With High Breakdown Point," *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics and Probability*, Bad Tatzmannsdorf, Austria, September 4–9, 1983.

SEN, P.K. (1968), "Estimates of the Regression Coefficient Based on Kendall's Tau," *Journal of the American Statistical Association*, 63, 1379–1389.

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

SIEGEL, A.F., and BENSON, R.H. (1982), "A Robust Comparison of Biological Shapes," *Biometrics*, 38, 341–350.

STAHEL, W.A. (1981), "Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen," unpublished Ph.D. dissertation, ETH Zürich, Statistics Dept.

STIGLER, S.M. (1981), "Gauss and the Invention of Least Squares," *Annals of Statistics*, 9, 465–474.

THEIL, H. (1950), "A Rank-Invariant Method of Linear and Polynomial Regression Analysis" (Parts 1–3), *Nederlandsche Akademie van Wetenschappen Proceedings*, Ser. A, 53, 386–392; 521–525; 1397–1412.