ROBUST REGRESSION BY MEANS OF S-ESTIMATORS

P. Rousseeuw and V. Yohai

Vrije Universiteit Brussel, CSOO (M 205), Pleinlaan 2,
B-1050 Brussels, Belgium

Departamento de Matemáticas, Facultad de Ciencias Exactas y Naturales, Ciudad Universitaria, Pabellon 1,
1428 Buenos Aires, Argentina

1. Introduction

There are at least two reasons why robust regression techniques are useful tools in robust time series analysis. First of all, one often wants to estimate autoregressive parameters in a robust way, and secondly, one sometimes has to fit a linear or nonlinear trend to a time series. In this paper we shall develop a class of methods for robust regression, and briefly comment on their use in time series. These new estimators are introduced because of their invulnerability to large fractions of contaminated data. We propose to call them "S-estimators" because they are based on estimators of scale.

The general linear model is given by $y_i = x_i^t\theta + e_i$ for $i=1,..,n$ where x_i and θ are p-dimensional column vectors, e_i is the error term and y_i is the dependent variable. Our aim is to estimate θ from the data $(x_1,y_1),..,(x_n,y_n)$. The classical least squares estimator of Gauss and Legendre corresponds to

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

where the residuals r_i equal $y_i - x_i^t\hat{\theta}$. This estimator is not robust at all, because the occurence of even a single (bad) outlier can spoil the result completely.

In connection with this effect, the breakdown point was invented by Hampel (1971), who gave it a rigourous asymptotic

J. Franke et al. (eds.), *Robust and Nonlinear Time Series Analysis*

definition. Recently, Donoho and Huber (1983) introduced a simplified version which works on finite samples, like the precursor ideas of Hodges (1967). In this paper we shall use the finite-sample version of the breakdown point (based on replacements). Take any sample X of n data points (x_i,y_i) and any estimator T of the parameter vector θ . Let $\beta(m,T,X)$ be the supremum of $\|T(X')-T(X)\|$ for all corrupted samples X' where any m of the original points of X are replaced by arbitrary values. Then the breakdown point of T at X is defined as

$$\varepsilon_n^*(T,X) = \min\{\frac{m}{n} \; ; \; \beta(m,T,X) \text{ is infinite}\} \quad . \tag{1.2}$$

In words, ε_n^* is the smallest fraction of contaminated data that can cause the estimator to take on values arbitrarily far from T(X). For least squares, one bad observation can already cause breakdown, so $\varepsilon_n^*(T,X) = \frac{1}{n}$ which tends to 0% when the sample size n becomes large.

The purpose of this paper is to find more robust estimators. A first step came from Edgeworth (1887) who proposed to replace the square in (1.1) by an absolute value, which does not grow so fast. His least absolute values or L_1 criterion is

$$\underset{\hat{\theta}}{\text{minimize}} \; \Sigma_{i=1}^n \; |r_i| \quad . \tag{1.3}$$

It turns out that this protects against outlying y_i , but still cannot cope with outlying x_i (called "leverage points"), which have a large effect on the fit. Therefore, we still have $\varepsilon_n^*(T,X) = \frac{1}{n} \to 0\%$.

The next step in this direction is Huber's (1973,1981) M-estimation method, 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. However, unlike (1.1) or (1.3) this does not yield estimators which are invariant with respect to a magnification of the error scale. Therefore, Huber estimates the scale parameter simultaneously (which corresponds to "Proposal 2" of Huber (1964) for the location problem):

$$\Sigma_{i=1}^n \; \psi(r_i/\hat{\sigma}) \; x_i = 0 \tag{1.4}$$

$$\Sigma_{i=1}^n \; \chi(r_i/\hat{\sigma}) = 0 \tag{1.5}$$

where ψ is the derivative of ρ . (Finding the simultaneous solution of the system of equations (1.4) and (1.5) is not trivial, and in practice one uses some iteration scheme, which makes it hard to use non-monotone ψ-functions.) 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 $\varepsilon_n^*(T,X) = \frac{1}{n}$ because of the possibility of leverage points.

Because of this vulnerability to leverage points, generalized M-estimators ("GM-estimators" for short) were considered. Their basic idea is to bound the influence of outlying x_i , making use of some weight function w. Mallows (1975) proposed to replace (1.4) by

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

whereas Schweppe (see Hill 1977) suggested to use

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

Making use of influence functions, good choices of ψ and w were made (Hampel 1978, Krasker 1980, Krasker and Welsch 1982, Ronchetti and Rousseeuw 1982). However, there are some problems finding suitable defining constants, making the application of these estimators to real data a nontrivial task. Moreover, it turns out that all GM-estimators 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). In fact, some numerical experiments by the first author indicate that most commonly used GM-estimators do not reach $1/(p+1)$ at all. Several other estimators have been proposed by Theil (1950), Brown and Mood (1951), Sen (1968), Jaeckel (1972) and Andrews (1974), but they all have a breakdown point less than 30% in the case of simple regression (p=2).

All this raises the question whether robust regression with 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 we can expect (for larger amounts of contamination, it becomes impossible to distinguish the "good" and the "bad" parts of the data). However, the repeated median is not affine equivariant, by which we mean that it depends on the choice of the coordinate axes of the x_i . The least median of squares (LMS) technique

$$\underset{\hat{\theta}}{\text{minimize}}\ \text{median}(r_i^2) \tag{1.6}$$

(Rousseeuw 1982, based on an idea of Hampel 1975) also possesses a 50% breakdown point and is affine equivariant, but has the disadvantage of converging like $n^{-1/3}$. Finally, the least trimmed squares (LTS) estimator (Rousseeuw 1983) is given by

$$\underset{\hat{\theta}}{\text{minimize}}\ \Sigma_{i=1}^{h}\ (r^2)_{i:n} \tag{1.7}$$

where h is the largest integer $\leq \frac{n}{2}+1$ and $(r^2)_{1:n} \leq \ldots \leq (r^2)_{n:n}$ are the ordered squared residuals. The LTS also has a 50% breakdown point, is affine equivariant, and even converges like $n^{-1/2}$ (Rousseeuw 1983). On the other hand, the computation of its objective function (for fixed p) takes O(n log n) steps (because of the ordering), compared to only O(n) steps for the LMS. The rest of the paper is devoted to the construction of a 50% breakdown, affine equivariant estimator with rate $n^{-1/2}$, with an O(n) objective function and a higher asymptotic efficiency than the LTS.

Our aim is to find simple high-breakdown regression estimators which share the flexibility and nice asymptotic properties of M-estimators. Looking at (1.4) and (1.5), we see that the function χ needed for the estimate $\hat{\sigma}$ is independent of the choice of the function ψ needed for $\hat{\theta}$. On the other hand, a completely different approach would be to take the scale of the residuals as the central notion, and to derive the estimate $\hat{\theta}$ from it.

We will consider one-dimensional estimators of scale defined by a function ρ satisfying

(R1) ρ is symmetric, continuously differentiable and $\rho(0)=0$;

(R2) there exists $c > 0$ such that ρ is strictly increasing on $[0,c]$ and constant on $[c,\infty)$.

For any sample $\{r_1,\ldots,r_n\}$ of real numbers, we define the scale estimate $s(r_1,\ldots,r_n)$ as the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\rho(r_i/s) = K \qquad (1.8)$$

where K is taken to be $E_\Phi[\rho]$, where Φ is the standard normal. (If there happens to be more than one solution to (1.8), we put $s(r_1,\ldots,r_n)$ equal to the supremum of the set of solutions; if there is no solution to (1.8), then we put $s(r_1,\ldots,r_n) = 0$.)

DEFINITION Let $(x_1,y_1),\ldots,(x_n,y_n)$ be a sample of regression data with p-dimensional x_i . For each vector θ , we obtain residuals $r_i(\theta)=y_i-x_i^t\theta$ of which we can calculate the dispersion $s(r_1(\theta),\ldots,r_n(\theta))$ by (1.8), where ρ satisfies (R1) and (R2). Then the S-estimator $\hat\theta$ is defined by

$$\underset{\theta}{\text{minimize}}\; s(r_1(\theta),\ldots,r_n(\theta)) \qquad (1.9)$$

and the final scale estimator is

$$\hat\sigma = s(r_1(\hat\theta),\ldots,r_n(\hat\theta)) \quad . \qquad (1.10)$$

We have decided to call $\hat\theta$ an S-estimator because it is derived from a scale statistic in an implicit way, like the derivation of R-estimators from rank statistics. S-estimators are clearly affine equivariant, and we shall see later on that they possess a high breakdown point and are asymptotically normal.

An example of a ρ-function for (1.8) is

$$\rho(x) = \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} \qquad \text{for } |x| \le c$$
$$= \frac{c^2}{6} \qquad \text{for } |x| \ge c \quad , \qquad (1.11)$$

the derivative of which is Tukey's biweight function:

$$\psi(x) = x(1-(\tfrac{x}{c})^2)^2 \qquad \text{for } |x| \leq c$$
$$= 0 \qquad \text{for } |x| \geq c \quad . \tag{1.12}$$

Another possibility is to take a ρ corresponding to the function $\tilde{\psi}$ proposed by Hampel, Rousseeuw and Ronchetti (1981). In general, $\psi(x) = \rho'(x)$ will always be zero for $|x| \geq c$ because of condition (R2); such ψ-functions are usually called "redescending".

2. The breakdown point of S-estimators

Let us start by considering a function ρ for which (R1) and (R2) hold, and also

$$\text{(R3)} \quad \frac{E_\Phi[\rho]}{\rho(c)} = \frac{1}{2} \quad .$$

This is not impossible: in the case of (1.11) it is achieved by taking $c \simeq 1.547$. Let us now look at the scale estimator $s(r_1,\ldots,r_n)$ which is defined by (1.8).

LEMMA 1. For each ρ satisfying (R1) to (R3) and for each n there exist positive constants α and β such that the estimator s given by (1.8) satisfies $\alpha\text{median}(|r_1|,\ldots,|r_n|) \leq s(r_1,\ldots,r_n) \leq \beta\text{median}(|r_1|,\ldots,|r_n|)$.

Proof. Verification shows that these inequalities hold with $\alpha = \frac{1}{c}$ and $\beta = 1/(\rho^{-1}[\rho(c)/(n+1)])$ for n odd, respectively $\beta = 2/(\rho^{-1}[2\rho(c)/(n+2)])$ for n even. When either $\text{median}(|r_1|,\ldots,|r_n|)$ or $s(r_1,\ldots,r_n)$ is zero, the other is zero too.

From now on, we work with regression data $\{(x_1,y_1),\ldots,(x_n,y_n)\}$. We assume that all observations with $x_i = 0$ have been deleted, because they give no information on θ . Moreover, no more than half of the points (x_i,y_i) should lie on a vertical subspace (here, a vertical subspace is one containing (0,1).)

LEMMA 2. For any ρ satisfying (R1) to (R3), there always exists a solution to (1.9).

Proof. Making use of Lemma 1, this follows from the proof of Theorem 1 of (Rousseeuw 1982) where the result was essentially given for minimization of median$(|r_1|,\ldots,|r_n|)$.

We shall say the observations are in general position when any p of them give a unique determination of θ . In the case of simple regression (p=2), this means that no two points may coincide or determine a vertical line.

THEOREM 1. An S-estimator constructed from a function ρ satisfying (R1) to (R3) has breakdown point

$$\varepsilon_n^* = ([\tfrac{n}{2}]-p+2)/n$$

at any sample $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ in general position.

Proof. Follows from Theorem 2 of (Rousseeuw 1982) using Lemma 1.

The breakdown point only depends slightly on n, and for $n \to \infty$ we obtain $\varepsilon^* = 50\%$, the best we can expect. A nice illustration of this high resistance is given by the following result.

COROLLARY. If there exists some θ such that at least $[\frac{n}{2}]+p$ of the points satisfy $y_i=x_i^t\theta$ exactly and are in general position, then $\hat{\theta} = \theta$ whatever the other observations are.

For example, in the case of simple regression this implies that when 12 out of 20 points lie exactly on a non-vertical straight line, then this line will be found.

If condition (R3) is replaced by $E_\Phi[\rho]/\rho(c) = \lambda$ where $0 < \lambda \leq \frac{1}{2}$, then the corresponding S-estimator has a breakdown point tending to $\varepsilon^* = \lambda$ for $n \to \infty$. In the case of (1.11), values of c larger than 1.547 yield better asymptotic efficiencies at a Gaussian central model, but smaller breakdown points.

3. Examples

In order to demonstrate the usefulness of S-estimators, let us look at an example with a large fraction of outliers. The Belgian Statistical Survey (edited by the Ministry of Economy) gives, among other things, the total number of international phone calls made. These data are listed in Table 1 (in tens of millions), and seem to show an upward trend over the years. However, this time series contains heavy contamination from 1964 till 1969 (and probably the 1963 and 1970 values are also affected). In Table 1, we marked the spurious values with (*). Actually, it turns out that the discrepancy was due to the fact that in this period another recording system was used, which only gave the total number of minutes of these calls!

Table 1.

x	y	x	y
50	0.44	62	1.61
51	0.47	63	2.12(*)
52	0.47	64	11.9(*)
53	0.59	65	12.4(*)
54	0.66	66	14.2(*)
55	0.73	67	15.9(*)
56	0.81	68	18.2(*)
57	0.88	69	21.2(*)
58	1.06	70	4.30(*)
59	1.20	71	2.40
60	1.35	72	2.70
61	1.49	73	2.90

If someone would not look carefully at these data and just apply the least squares method in a routine way, he would obtain $y = 0.5041x - 26.01$ which corresponds to the dotted line in Figure 1. This dotted line has been attracted very much by the 1964-1969 values, and does not fit the good or the bad data points. In fact, some of the good values (such as the 1972 one) yield even larger LS residuals than some of the bad values! Now let us apply the S-estimator corresponding to the function (1.11) with $c = 1.547$. This yields $y = 0.1121x - 5.42$ (plotted as a solid line in Figure 1), which neglects the outliers and gives a good approximation to the other points. (This is not to say that a linear fit is necessarily

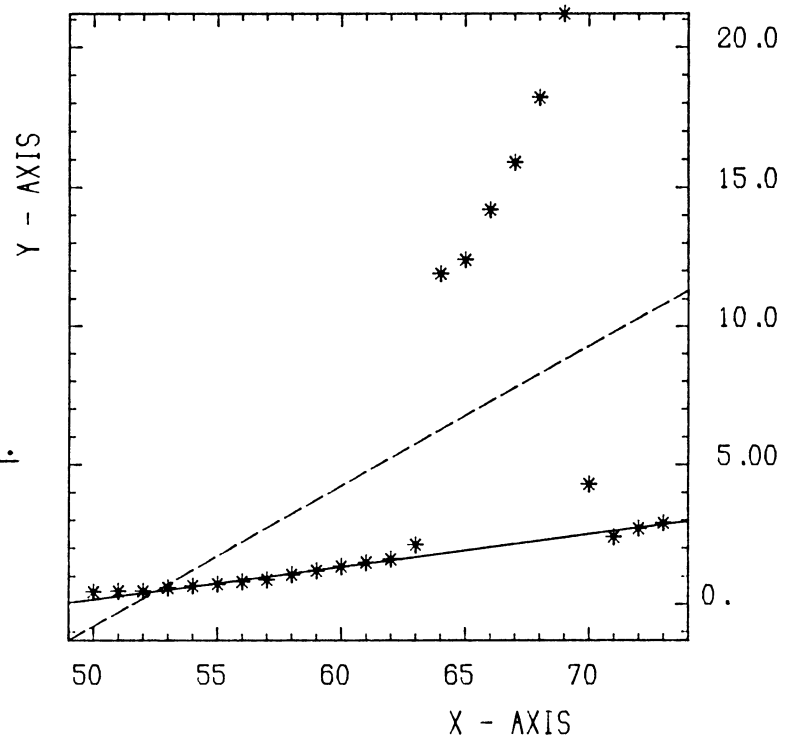

Figure 1.

the best model, because collecting more data might reveal a more complicated kind of relationship.)

Let us now look at an example from astronomy. The data in Table 2 form the Hertzsprung-Russell diagram of the star cluster CYG OB1, which contains 47 stars in the direction of Cygnus. Here, x is the logarithm of the effective temperature at the surface of the star, and y is the logarithm of its light intensity. These numbers were given to us by Claude Doom (personal communication) who extracted the raw data from Humphreys (1978) and performed the calibration according to Vansina and De Greve (1982).

The Hertzsprung-Russel diagram itself is shown in Figure 2. It is a scatterplot of these points, where the log temperature x is plotted from right to left. Our eye sees two groups of points: the majority which seems to follow a steep band, and the four stars in the upper right corner. These parts of the diagram are well-known in astronomy: the 43 stars are said to lie on the main sequence, whereas the four

Table 2.

x	y	x	y
4.37	5.32	4.56	5.74
4.26	4.93	4.56	5.74
4.30	5.19	4.46	5.46
3.84	4.65	4.57	5.27
4.26	5.57	4.37	5.12
3.49	5.73	4.43	5.45
4.48	5.42	4.01	4.05
4.29	4.26	4.42	4.58
4.23	3.94	4.42	4.18
4.23	4.18	3.49	5.89
4.29	4.38	4.29	4.22
4.42	4.42	4.49	4.85
4.38	5.02	4.42	4.66
4.29	4.66	4.38	4.90
4.22	4.39	3.48	6.05
4.38	4.42	4.56	5.10
4.45	5.22	3.49	6.29
4.23	4.34	4.62	5.62
4.53	5.10	4.45	5.22
4.53	5.18	4.43	5.57
4.38	4.62	4.45	5.06
4.50	5.34	4.45	5.34
4.55	5.54	4.45	4.98
4.42	4.50		

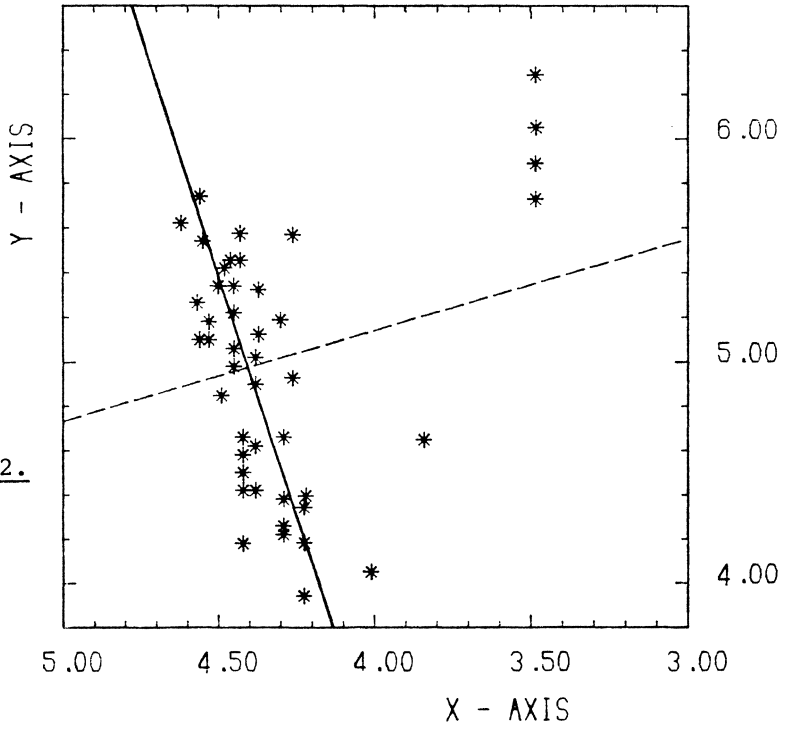

Figure 2.

remaining stars are called giants. Application of our S-estimator to this data yields the solid line y = 3.289x - 9.59 , which fits the main sequence nicely. On the other hand, the least squares solution y = -0.408x + 6.77 corresponds to the dotted line in Figure 2, which has been pulled away by the four giant stars (which it does not fit well either).

4. Asymptotic behaviour of S-estimators

In section 2 we saw that S-estimators do not break down easily when the data are contaminated. Of course, we also want our estimators to behave well when the data are not contaminated, that is, when they satisfy the classical assumptions. Therefore, let us look at the asymptotic behaviour of S-estimators at the central Gaussian model, where (x_i, y_i) are i.i.d. random variables satisfying

$$y_i = x_i^t \theta_0 + e_i \quad , \tag{4.1}$$

x_i follows some distribution H, and e_i is independent of x_i and distributed like $\Phi(e/\sigma_0)$ for some $\sigma_0 > 0$ (here, Φ is the standard normal cdf).

THEOREM 2 (CONSISTENCY). Let ρ be a function satisfying (R1) and (R2), with the derivative $\rho' = \psi$. Assume that
(i) $\psi(u)/u$ is nonincreasing for $u > 0$;
(ii) $E_H[\|x\|] < \infty$, and H has a density.
Let (x_i, y_i) be i.i.d. according to the model (4.1), and let $\hat{\theta}_n$ be a solution of (1.9) for the first n points, and $\hat{\sigma}_n = s(r_1(\hat{\theta}_n), .., r_n(\hat{\theta}_n))$. Then

$$\hat{\theta}_n \to \theta_0 \quad \text{a.s.}$$

$$\hat{\sigma}_n \to \sigma_0 \quad \text{a.s.}$$

Proof. This follows from Theorems 2.2 and 3.1 of Maronna and Yohai (1981) because S-estimators satisfy the same first-order necessary conditions as M-estimators. Indeed, let θ be any p-dimensional parameter vector. By definition, we know

that $s(\theta)=s(r_1(\theta),\ldots,r_n(\theta))$ is larger than or equal to $\hat{\sigma}_n = s(\hat{\theta}_n)$. Keeping in mind that $s(\theta)$ satisfies $\frac{1}{n}\Sigma_{i=1}^{n}\rho(r_i(\theta)/s(\theta))=K$ and that $\rho(u)$ is nondecreasing in $|u|$, it follows that always $\frac{1}{n}\Sigma_{i=1}^{n}\rho(r_i(\theta)/\hat{\sigma}_n) \leq K$. At $\theta = \hat{\theta}_n$, this becomes an equality. Therefore, $\hat{\theta}_n$ minimizes $\frac{1}{n}\Sigma_{i=1}^{n}\rho(r_i(\theta)/\hat{\sigma}_n)$. (This fact cannot be used for determining $\hat{\theta}_n$ in practice, because $\hat{\sigma}_n$ is fixed but unknown.) Differentiating with respect to θ , we find $\frac{1}{n}\Sigma_{i=1}^{n}\psi(r_i(\theta)/\hat{\sigma}_n)x_i = 0$. If we denote ρ-K by χ we conclude that $(\hat{\theta}_n,\hat{\sigma}_n)$ is a solution of the system of equations (1.4)-(1.5). Unfortunately, these equations cannot be used directly because there are infinitely many solutions (ψ is redescending) and the familiar iteration procedures easily end in the wrong place if there are leverage points. (This means we still have to minimize (1.9) with brute force in order to actually compute the S-estimate in a practical situation.) But anyway, the fact that $(\hat{\theta}_n,\hat{\sigma}_n)$ satisfies (1.4)-(1.5) is sufficient to apply the results of Maronna and Yohai (1981) to S-estimators. This ends the proof.

THEOREM 3 (ASYMPTOTIC NORMALITY) Let $\theta_0 = 0$ and $\sigma_0 = 1$ for simplicity. If the conditions of Theorem 2 hold and also

(iii) ψ is differentiable in all but a finite number of points, $|\psi'|$ is bounded and $\int\psi' d\Phi > 0$;

(iv) $E_H[xx^t]$ is nonsingular and $E_H[\|x\|^3] < \infty$, then

$$n^{1/2}(\hat{\theta}_n-\theta_0) \xrightarrow{\mathcal{L}} N(0,E_H[xx^t]^{-1}\int\psi^2 d\Phi/(\int\psi' d\Phi)^2)$$

$$n^{1/2}(\hat{\sigma}_n-\sigma_0) \xrightarrow{\mathcal{L}} N(0,\int(\rho(y)-K)^2 d\Phi(y)/(\int y\psi(y) d\Phi(y))^2) \ .$$

Proof. This follows from Theorem 4.1 of Maronna and Yohai (1981).

As a consequence of this theorem, we can compute the asymptotic efficiency e of an S-estimator at the Gaussian model as $e = (\int\psi' d\Phi)^2/\int\psi^2 d\Phi$. Table 3 gives the asymptotic efficiency of the S-estimator corresponding to (1.11), for different values of the breakdown point ε^* .

We note that taking c=2.560 yields a value of e which is larger than that of L_1 (for which e is about 64%), and gains

Table 3.

ε^*	e	c	K
50%	28.7%	1.547	.1995
45%	37.0%	1.756	.2312
40%	46.2%	1.988	.2634
35%	56.0%	2.251	.2957
30%	66.1%	2.560	.3278
25%	75.9%	2.937	.3593
20%	84.7%	3.420	.3899
15%	91.7%	4.096	.4194
10%	96.6%	5.182	.4475

us a breakdown point of 30%. In practice, we do not recommend the estimators in the table with a breakdown point smaller than 25%. In fact, it seems like a good idea to apply the c=1.547 estimator because of its 50% breakdown point, and to make up for its low efficiency by computing a one-step M-estimator (Bickel 1975) from this first solution, with a redescending and more efficient ψ . Such a two-stage procedure inherits the 50% breakdown point from the first stage, and the high asymptotic efficiency from the second.

5. Outlook

The computation of S-estimators is not at all easy. In fact, there is a relation to the projection pursuit technique (Friedman and Tukey 1974) for the analysis of multivariate data. To see this, let us consider the (p+1)-dimensional space of the (x,y)-data. In this space, linear models are defined by $(x,y)(-\theta,1)^t=0$ for some p-dimensional vector θ . The definition of S-estimators amounts to the following: for any vector θ , we consider the projection on the y-axis in the direction orthogonal to $(-\theta,1)$, and we select the $\hat{\theta}$ for which the projected sample has minimal dispersion $s(\theta)$. (In fact, we could generalize this easily to orthogonal regression by taking any nonzero (p+1)-vector and by projecting on this vector itself instead of on the y-axis.) This means that S-estimators belong to the highly computer-intensive part of statistics, just like projection pursuit and the bootstrap (Efron 1982). In Table 4 we give a schematic overview of criteria in affine equivariant regression.

Table 4: Schematic Overview

Criterion	Method	Computation	ε^*
Best Linear Unbiased	LS	explicit	0
Minimax Variance	M	iterative	0
Bounded Influence	GM	iterative (harder)	$\leq 1/(p+1)$
High Breakdown	S	projection pursuit	constant, up to 1/2

At present, we have a portable Fortran program for the computation of S-estimators in simple regression, which is based on the LMS program of Rousseeuw (1982). We are also working on a program for multiple regression, the computation time of which is feasable up to around p = 10.

An interesting project (which hopefully will be started at the University of Washington in 1984) is to make a Monte Carlo comparison of the available methods for robust regression in different situations. This study would be the natural sequel to the work of Andrews et al (1972). For the S-estimators, this would also be a good occasion to study the behaviour of one-step improvements, both one-step M and one-step reweighted LS.

Looking at Table 3, there appears to be a tradeoff between breakdown point and efficiency for S-estimators. Probably one could obtain higher values of e for the same ε^* by using ρ-functions based on hyperbolic tangent estimators (Hampel, Rousseeuw and Ronchetti 1981). In fact, we wonder what is the maximal efficiency e, given a certain value of ε^* .

Our definition of S-estimators (1.9) could easily be generalized by allowing other types of dispersion measures s . If we consider all permutation invariant and scale equivariant measures s (that is, $s(r_{\pi(1)},\ldots,r_{\pi(n)}) = s(r_1,\ldots,r_n)$ and $s(\lambda r_1,\ldots,\lambda r_n) = |\lambda| s(r_1,\ldots,r_n)$ for all permutations π and factors λ), then also the following methods are S-estimators: least squares; least absolute deviations, least p-th power deviations (Gentleman 1965), the method of Jaeckel (1972), least median of squares (Rousseeuw 1982) and least trimmed squares (Rousseeuw 1983).

S-estimators could also be used for robust analysis of variance, even in the general linear model. Instead of comparing sums of squares (with respect to the reduced model and the full model) in order to obtain an F-statistic, we could compare the scale estimators (1.10) which are obtained from S-estimators. By (1.9), we are sure that $\hat{\sigma}$ for the reduced model is at least as large as $\hat{\sigma}$ for the full model. Making use of a similar reasoning, Doug Martin (personal communication) proposes to define robust sequential partial correlations by means of S-estimators. In both cases, it is probably not possible to find the exact finite sample distribution of these statistics in an analytical way. However, one could resort to small-sample asymptotics (Field and Hampel 1982) or simulation in order to determine critical values for hypothesis testing.

Acknowledgments The research of the first author was supported by the Belgian National Science Foundation, and that of the second author by Office of Naval Research Contract N000/4-82-K-0062. Most of the work leading to this paper was done while the authors were sharing the hospitality of Doug Martin, at the Statistics Department of the University of Washington. We also wish to thank Annick Leroy for assistance with the programming, and Frank Hampel and Werner Stahel for useful suggestions.

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 University Press.

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 2nd Berkeley Symposium on Mathematical Statistics and Probability, 159-166.

DONOHO, D.L., and HUBER, P.J. (1983), "The notion of Breakdown Point," in A Festschrift for E.L. Lehmann, Wadsworth.

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

EFRON, B. (1982), The Jackknife, the Bootstrap and other Resampling Plans, SIAM Monograph No. 38, Society for Industrial and Applied Mathematics.

FIELD, C.A., and HAMPEL, F.R. (1982), "Small-Sample Asympto-

tic Distributions of M-estimators of Location," Biometrika, 69, 29-46.
FRIEDMAN, J.H., and TUKEY, J.W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," IEEE Transactions on Computers, C-23, 881-889.
GENTLEMAN, W.M. (1965), "Robust Estimation of Multivariate Location by Minimizing p-th power Deviations," Ph.D. dissertation, Princeton University, and Bell Tel. Laboratories memorandum MM65-1215-16.
HAMPEL, F.R. (1971), "A General Qualitative Definition of Robustness," Annals of Mathematical Statistics, 42, 1887-1896.
HAMPEL, F.R. (1975), "Beyond Location Parameters: Robust Concepts and Methods," Bulletin of the International Statistical Institute, 46, 375-382.
HAMPEL, F.R. (1978), "Optimally Bounding the Gross-Error-Sensitivity and the Influence of Position in Factor Space," 1978 Proceedings of the ASA Statistical Computing Section, 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.
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 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 163-168.
HUBER, P.J. (1964), "Robust Estimation of a Location Parameter," Annals of Mathematical Statistics, 35, 73-101.
HUBER, P.J. (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," Annals of Statistics, 1, 799-821.
HUBER, P.J. (1981), Robust Statistics, New York: Wiley.
HUMPHREYS, R.M. (1978), "Studies of Luminous Stars in Nearby Galaxies. I. Supergiants and O stars in the Milky Way," The Astrophysical Journal Supplement Series, 38, 309-350.
JAECKEL, L.A. (1972), "Estimating Regression Coefficients by Minimizing the Dispersion of the 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.
MALLOWS, C.L. (1975), "On Some Topics in Robustness," unpublished memorandum, Bell Tel. Laboratories, Murray Hill.
MARONNA, R.A., and YOHAI, V.J. (1981), "Asymptotic Behaviour of General M-estimates for Regression and Scale with Random Carriers," Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 58, 7-20.
MARONNA, R.A., BUSTOS, O., 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, 91-116.
RONCHETTI, E., and ROUSSEEUW, P.J. (1982), "Change-of-Variance Sensitivities in Regression Analysis," Research Report

No. 36, Fachgruppe für Statistics, ETH Zürich.
ROUSSEEUW, P.J. (1982), "Least Median of Squares Regression," Research Report No. 178, Centre for Statistics and Operations Research, VUB Brussels.
ROUSSEEUW, P.J. (1983), "Multivariate Estimation with High Breakdown Point," Research Report No. 192, Centre for Statistics and Operations Research, VUB Brussels.
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.
THEIL, H. (1950), "A Rank-Invariant Method of Linear and Polynomial Regression Analysis," I, II and III. Nederlandsche Akademie van Wetenschappen Proceedings Serie A, 53, 386-392, 521-525, and 1397-1412.
VANSINA, F., and DE GREVE, J.P. (1982), "Close Binary Systems Before and After Mass Transfer," Astrophysics and Space Science, 87, 377-401.