Lemma 1. The LMS estimator is regression equivariant, scale equivariant, and affine equivariant.

Proof. This follows from

$$\underset{i}{\text{med}}\,(\{y_i + \mathbf{x}_i\mathbf{v}\} - \mathbf{x}_i\{\boldsymbol{\theta} + \mathbf{v}\})^2 = \underset{i}{\text{med}}\,(y_i - \mathbf{x}_i\boldsymbol{\theta})^2\,,$$

$$\underset{i}{\text{med}}\,(cy_i - \mathbf{x}_i\{c\boldsymbol{\theta}\})^2 = c^2\,\underset{i}{\text{med}}\,(y_i - \mathbf{x}_i\boldsymbol{\theta})^2\,,$$

and

$$\underset{i}{\text{med}}\,(y_i - \{\mathbf{x}_i\mathbf{A}\}\{\mathbf{A}^{-1}\boldsymbol{\theta}\})^2 = \underset{i}{\text{med}}\,(y_i - \mathbf{x}_i\boldsymbol{\theta})^2\,,$$

respectively. □

On the other hand, it may be noted that the repeated median, defined in (2.14) of Chapter 1, is regression and scale equivariant but not affine equivariant.

In what follows we shall say the observations are in *general position* when any p of them give a unique determination of $\boldsymbol{\theta}$. For example, in case $p = 2$ 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) may not be collinear. When the observations come from continuous distributions, this event has probability one.

As promised in Chapter 1 we will now investigate the breakdown properties of the LMS, using the finite-sample version of the breakdown point introduced by Donoho and Huber (1983). Take any sample Z of n data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ and a regression estimator T. This means that applying T to Z yields a regression estimate $\hat{\boldsymbol{\theta}}$. Let bias $(m; T, Z)$ be the supremum of $\|T(Z') - T(Z)\|$ for all corrupted samples Z', where any m of the original data points are replaced by arbitrary values. Then the *breakdown point* of T at Z is

$$\varepsilon_n^*(T, Z) = \min\{m/n;\ \text{bias}\,(m; T, Z) \text{ is infinite}\}\,. \tag{4.5}$$

We prefer replacing observations to adding observations, which some authors do, because replacement contamination is simple, realistic, and generally applicable. Indeed, from an intuitive point of view, outliers are not some faulty observations that are added at the end of the sample, but they treacherously hide themselves by replacing some of the data points that should have been observed. Replacement contamination does not cause any formal problems because the contaminated sample has the same size as the original one, so we only have to consider one estimator

T_n and not several T_{n+m}. This means that replacement still applies to many situations where adding observations does not make sense (for instance, one cannot just add cells to a two-way table). Therefore, we would like to defend the standard use of the above definition.

Theorem 2. If $p>1$ and the observations are in general position, then the breakdown point of the LMS method is

$$([n/2]-p+2)/n\,.$$

Proof. 1. We first show that $\varepsilon_n^*(T, Z) \geq ([n/2]-p+2)/n$ for any sample $Z=\{(\mathbf{x}_i, y_i);\ i=1, \ldots, n\}$ consisting of n observations in general position. By the first theorem, the sample Z yields a solution $\boldsymbol{\theta}$ of (4.1). 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 $Z'=\{(\mathbf{x}_i', y_i');\ i=1, \ldots, n\}$ by retaining $n-[n/2]+p-1$ observations of Z, which will be called the "good" observations, and by replacing the others by arbitrary values. It suffices to prove that $\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|$ is bounded, where $\boldsymbol{\theta}'$ corresponds to Z'. For this purpose, some geometry is needed. We again work in the $(p+1)$-dimensional space E of the observations $(\mathbf{x}_i, y_i)$ and in its horizontal hyperplane through the origin, denoted by $(y=0)$. Put

$$\rho = \tfrac{1}{2} \inf \{\tau>0;\ \text{there exists a } (p-1)\text{-dimensional subspace } V \text{ of } (y=0) \text{ through the origin such that } V^\tau \text{ covers at least } p \text{ of the } \mathbf{x}_i\}\,,$$

where V^τ is the set of all $\mathbf{x}$ with distance to V not larger than τ. Because Z is in general position, it holds that $\rho>0$. Also, put $M:=\max_i |r_i|$, where r_i are the residuals $y_i-\mathbf{x}_i\boldsymbol{\theta}$. The rest of the proof of part 1 will be devoted to showing that

$$\|\boldsymbol{\theta}-\boldsymbol{\theta}'\| < 2(\|\boldsymbol{\theta}\|+M/\rho)\,,$$

which is sufficient because the right member is a finite constant. Denote by H the nonvertical hyperplane given by the equation $y=\mathbf{x}\boldsymbol{\theta}$, and let H' correspond in the same way to $\boldsymbol{\theta}'$. Without loss of generality assume that $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$, hence $H' \neq H$. By the dimension theorem of linear algebra, the intersection $H \cap H'$ has dimension $p-1$. If $\mathrm{pr}\,(H \cap H')$ denotes the vertical projection of $H \cap H'$ on $(y=0)$, it follows that at most $p-1$ of the good $\mathbf{x}_i$ can lie on $(\mathrm{pr}\,(H \cap H'))^\rho$. Define A as the set of remaining good observations, containing at least $n-[n/2]+p-1-(p-1)=n-$

$[n/2]$ points. Now consider any $(\mathbf{x}_a, y_a)$ belonging to A, and put $r_a = y_a - \mathbf{x}_a\boldsymbol{\theta}$ and $r'_a = y_a - \mathbf{x}_a\boldsymbol{\theta}'$. Construct the vertical two-dimensional plane P_a through $(\mathbf{x}_a, y_a)$ and orthogonal to $\text{pr}\,(H \cap H')$. It follows, as in the proof of Theorem 1, that

$$\begin{aligned} |r'_a - r_a| = |\mathbf{x}_a\boldsymbol{\theta}' - \mathbf{x}_a\boldsymbol{\theta}| &> \rho|\tan(\alpha') - \tan(\alpha)| \\ &\geq \rho\,\big|\,|\tan(\alpha')| - |\tan(\alpha)|\,\big| \\ &= \rho\,\big|\,\|\boldsymbol{\theta}'\| - \|\boldsymbol{\theta}\|\,\big|\,, \end{aligned}$$

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

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

it follows that

$$|r'_a - r_a| > \rho(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| - 2\|\boldsymbol{\theta}\|)\,.$$

Now the median of the squared residuals of the new sample Z' with respect to the old $\boldsymbol{\theta}$, 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 $\boldsymbol{\theta}'$ is a solution of (4.1) for Z', it follows that also

$$\underset{i}{\text{med}}\,(y'_i - \mathbf{x}'_i\boldsymbol{\theta}')^2 \leq M^2\,.$$

If we now assume that $\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \geq 2(\|\boldsymbol{\theta}\| + M/\rho)$, then for all a in A it holds that

$$\begin{aligned} |r'_a - r_a| &> \rho(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| - 2\|\boldsymbol{\theta}\|) \\ &> \rho(2\|\boldsymbol{\theta}\| + 2M/\rho - 2\|\boldsymbol{\theta}\|) = 2M\,, \end{aligned}$$

so

$$|r'_a| \geq |r'_a - r_a| - |r_a| > 2M - M = M$$

and finally

$$\underset{i}{\text{med}}\,(y'_i - \mathbf{x}'_i\boldsymbol{\theta}')^2 > M^2\,,$$

a contradiction. Therefore,

$$\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| < 2(\|\boldsymbol{\theta}\| + M/\rho)$$

for any Z'.

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 through zero. Now construct any nonvertical hyperplane H' through L, which determines some $\boldsymbol{\theta}'$ by means of the equation $y = \mathbf{x}\boldsymbol{\theta}'$. If all of the "bad" observations are put on H', then Z' has a total of

$$([n/2] - p + 2) + (p - 1) = [n/2] + 1$$

points that satisfy $y_i' = \mathbf{x}_i'\boldsymbol{\theta}'$ exactly; so the median squared residual of Z' with respect to $\boldsymbol{\theta}'$ is zero, hence $\boldsymbol{\theta}'$ satisfies (4.1) for Z'. By choosing H' steeper and steeper, one can make $\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|$ as large as one wants. □

Note that the breakdown point depends only slightly on n. In order to obtain a single value, one often considers the limit for $n \to \infty$ (with p fixed), so it can be said that the classical 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. Indeed, 50% is the highest possible value for the breakdown point, since for larger amounts of contamination it becomes impossible to distinguish between the good and the bad parts of the sample, as will be proved in Theorem 4 below.

Once we know that an estimator does not break down for a given fraction m/n of contamination, it is of interest just how large the bias can be. Naturally, it is hoped that bias $(m; T, Z) = \sup \|T(Z') - T(Z)\|$ does not become too big. For this purpose, Martin et al. (1987) compute the maximal asymptotic bias of several regression methods and show that the LMS minimizes this quantity among a certain class of estimators. [The same property is much more generally true for the sample median in univariate location, as proved by Huber (1981, p. 74).]

We will now investigate another aspect of robust regression, namely the *exact fit property*. If the majority of the data follow a linear relationship *exactly*, then a robust regression method should yield this equation. If it does, the regression technique is said to possess the exact fit property. (In Section 5 of Chapter 2, this property was illustated for the case of a straight line fit.) The following example provides an illustration for the multivariate case. We created a data set of 25 observations, which are listed in Table 18. The first 20 observations satisfy the equation

$$y = x_1 + 2x_2 + 3x_3 + 4x_4 , \tag{4.6}$$

and five observations fall outside this hyperplane. Applying LS regression

Table 18. Artificial Data Set Illustrating the Exact Fit Property[a]

Index	x_1	x_2	x_3	x_4	y
1	1	0	0	0	1
2	0	1	0	0	2
3	0	1	1	0	5
4	0	0	1	1	7
5	1	1	0	1	7
6	1	1	1	0	6
7	0	1	1	1	9
8	1	0	0	1	5
9	1	1	1	1	10
10	0	1	0	1	6
11	1	0	1	0	4
12	1	0	1	1	8
13	1	0	2	3	19
14	2	0	1	3	17
15	1	2	3	0	14
16	2	3	1	0	11
17	2	0	3	1	15
18	2	1	1	3	19
19	1	0	2	1	11
20	1	1	2	2	17
21	1	2	0	1	11
22	2	1	0	1	10
23	2	2	1	0	15
24	1	1	2	2	20
25	1	2	3	4	40

[a]The first 20 points lie on the hyperplane $y = x_1 + 2x_2 + 3x_3 + 4x_4$

without intercept to this data set leads to the fit

$$\hat{y} = 0.508x_1 + 3.02x_2 + 3.08x_3 + 4.65x_4 .$$

Although a large proportion of the points lie on the same hyperplane, the LS does not manage to find it. The outlying points even produce small residuals, some of them smaller than the residuals of certain good points. This is visualized in the LS index plot in Figure 17.

On the other hand, the LMS looks for the pattern followed by the majority of the data, and it yields exactly equation (4.6) as its solution. The 20 points lying on the same hyperplane now have a zero residual (see

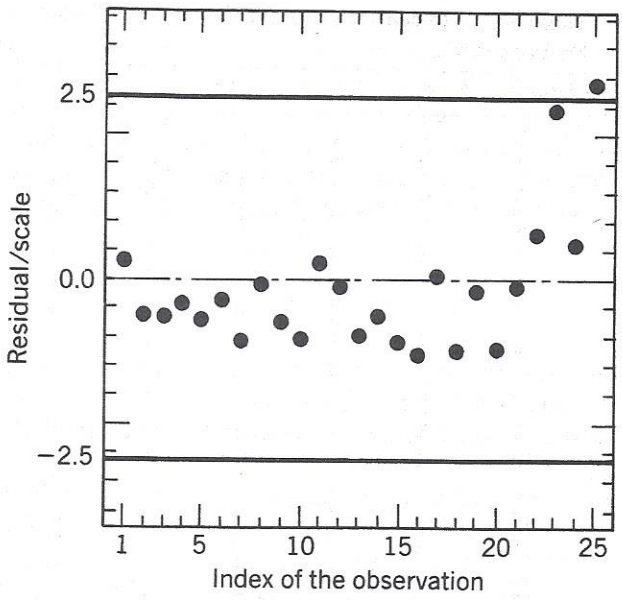

Figure 17. LS index plot for the data in Table 18.

the LMS index plot in Figure 18). Consequently, the scale estimate associated with LMS equals zero. In the index plot, the five outlying points fall far from the line through zero. The magnitude of their residual corresponds to their vertical distance from the hyperplane. In the exact fit case, reweighting on the basis of the LMS residuals is unnecessary, because the reduced data set would contain only the points that lie on the hyperplane.

The following theorem shows that the LMS satisfies the exact fit property.

Theorem 3. If $p > 1$ and there exists a $\boldsymbol{\theta}$ such that at least $n - [n/2] + p - 1$ of the observations satisfy $y_i = \mathbf{x}_i\boldsymbol{\theta}$ exactly and are in general position, then the LMS solution equals $\boldsymbol{\theta}$ whatever the other observations are.

Proof. There exists some $\boldsymbol{\theta}$ such that at least $n - [n/2] + p - 1$ of the observations lie on the hyperplane H given by the equation $y = \mathbf{x}\boldsymbol{\theta}$. Then $\boldsymbol{\theta}$ is a solution of (4.1), because $\text{med}_i\, r_i^2(\boldsymbol{\theta}) = 0$. Suppose that there is another solution $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$, corresponding to a hyperplane $H' \neq H$ and yielding residuals $r_i(\boldsymbol{\theta}')$. As in the proof of Theorem 2, $(H \cap H')$ has

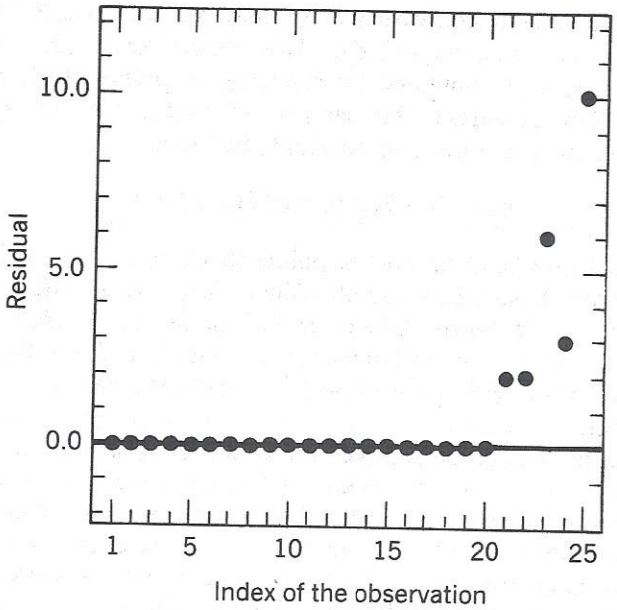

Figure 18. LMS index plot for the data in Table 18.

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

REMARK 1. It appears that Theorem 3 is a special case of a more general relation between breakdown and exact fit. (This remark was derived from joint work with D. Donoho in 1983.) Let us consider a possible formalization of this connection, by defining the *exact fit point* as

$$\delta_n^*(T, Z) = \min\{m/n;\ \text{there exists } Z' \text{ such that } T(Z') \neq \boldsymbol{\theta}\}, \quad (4.7)$$

where Z is a sample $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ such that $y_i = \mathbf{x}_i\boldsymbol{\theta}$ for all i, and Z' ranges over all corrupted samples where any m points of Z are replaced by arbitrary values. The smallest fraction of contamination capable of pulling T away from $\boldsymbol{\theta}$ is the exact fit point. If T is regression and scale equivariant, then

$$\delta_n^*(T, Z) \geq \varepsilon_n^*(T, Z). \quad (4.8)$$

Indeed, by regression equivariance we may assume that $\boldsymbol{\theta} = \mathbf{0}$, so all $y_i = 0$. Take any $m \geq n\delta_n^*(T, Z)$. Then there exists $Z' = \{(\mathbf{x}_1', y_1'), (\mathbf{x}_2', y_2'), \ldots, (\mathbf{x}_n', y_n')\}$ obtained by replacing m points of Z, such that $T(Z') \neq \mathbf{0}$. Now construct the sample $Z'' = \{(\mathbf{x}_1', cy_1'), (\mathbf{x}_2', cy_2'), \ldots, (\mathbf{x}_n', cy_n')\}$, where c is a positive constant. But then

$$\| T(Z'') - T(Z) \| = c \| T(Z') \| \neq 0$$

and Z'' differs from Z in at most m points (because at most m of the cy_i' can be different from zero). By choosing c large we see that T breaks down, so $\varepsilon_n^*(T, Z) \leq m/n$, which proves (4.8). This result becomes particularly useful if T is well behaved so that $\varepsilon_n^*(T, Z)$ is the same at most Z (say, at any Z in general position), as is the case for LMS.

Unfortunately, the reverse inequality is not generally true, because one can construct counterexamples for which ε_n^* is *strictly* smaller than δ_n^*. Consider, for instance, an estimator (for $n = 20$ and $p = 2$) that gives the right answer whenever at least 15 observations are in an exact fit situation, but which is put equal to LS in all other cases. Nevertheless, this example is clearly pathological, and it should be possible to prove $\varepsilon_n^* = \delta_n^*$ under some "reasonable" conditions.

REMARK 2. The breakdown point in Theorem 2 is slightly smaller than that of the repeated median, although they are both 50% breakdown estimators. We are indebted to A. Siegel (personal communication) for a way to overcome this. Instead of taking the median of the ordered squared residuals, consider the hth order statistic $(r^2)_{h:n}$ and

$$\underset{\hat{\boldsymbol{\theta}}}{\text{Minimize}} \, (r^2)_{h:n}, \quad \text{where } h = [n/2] + [(p+1)/2]. \tag{4.9}$$

It turns out (analogous to the proof of Theorem 2) that this variant of the LMS has breakdown point equal to $([n-p)/2] + 1)/n$, which is exactly the same value as for Siegel's repeated median. In Theorem 4, we shall show that this is the best possible result. Therefore, we actually use this version of the LMS in PROGRESS. For this variant of the LMS, Theorem 3 holds whenever strictly more than $\frac{1}{2}(n + p - 1)$ of the observations are in an exact fit situation. (This can be proven as in Theorem 3, or by making use of Remark 1 above.)

REMARK 3. The LMS estimator can be viewed as a special case of a larger family of estimators, namely the *least quantile of squares* estimators (LQS), which are defined by

$$\underset{\hat{\boldsymbol{\theta}}}{\text{Minimize}} \, (r^2)_{([(1-\alpha)n]+[\alpha(p+1)]):n}, \tag{4.10}$$

where $0 \leq \alpha \leq 50\%$. For α tending to 50%, (4.10) is asymptotically equivalent to the LMS. The breakdown point of the LQS is equal to α for $n \to \infty$. Putting $\alpha = 0\%$ in (4.10), one finds the L_∞ estimator

$$\underset{\hat{\theta}}{\text{Minimize}} \max_i r_i^2(\theta) ,$$

which is also referred to as *minimax regression* because the largest (absolute) residual is minimized. This method was already considered by Euler, Lambert, and Laplace (see Sheynin 1966 and Plackett 1972). Unfortunately, it is even *less* robust than least squares (see exercise 10).

Theorem 4. Any regression equivariant estimator T satisfies

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

at all samples Z.

Proof. Suppose that the breakdown point is strictly larger than $([(n-p)/2]+1)/n$. This would mean that there exists a finite constant b such that $T(Z')$ lies in the ball $B(T(Z), b)$ for all samples Z' containing at least $n - [(n-p)/2] - 1$ points of Z. Set $q = n - [(n-p)/2] - 1$, which also equals $[(n+p+1)/2] - 1$. Here $B(T(Z), b)$ is defined as the set of all θ for which $\|T(Z) - \theta\| \leq b$. Now construct a p-dimensional column vector $\mathbf{v} \neq \mathbf{0}$ such that $\mathbf{x}_1\mathbf{v} = 0, \ldots, \mathbf{x}_{p-1}\mathbf{v} = 0$. If $n+p+1$ is even, then $2q - (p-1) = n$; otherwise $2q - (p-1) = n - 1$. In general one can say that $2q - (p-1) \leq n$. Therefore, the first $2q - (p-1)$ points of Z can be replaced by

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{p-1}, y_{p-1}), (\mathbf{x}_p, y_p), \ldots, (\mathbf{x}_q, y_q) ,$$
$$(\mathbf{x}_p, y_p + \mathbf{x}_p\tau\mathbf{v}), \ldots, (\mathbf{x}_q, y_q + \mathbf{x}_q\tau\mathbf{v})$$

for any $\tau > 0$. For this new sample Z', the estimate $T(Z')$ belongs to $B(T(Z), b)$, since Z' contains q points of Z. But looking at Z' in another way reveals that $T(Z')$ can also be written as $T(Z'') + \tau\mathbf{v}$, where $T(Z'')$ is in $B(T(Z), b)$. Therefore, $T(Z')$ belongs to $B(T(Z) + \tau\mathbf{v}, b)$. This is a contradiction, however, because the intersection of $B(T(Z), b)$ and $B(T(Z) + \tau\mathbf{v}, b)$ is empty for large enough values of τ. □

Note that this maximal breakdown point is attained by the repeated median (Siegel 1982) and the version (4.9) of the LMS.

By putting $p = 1$ and $x_{i1} = 1$ for all i in (4.1), one obtains the special

case of one-dimensional estimation of a location parameter θ of the sample $(y_i)_{i=1,\ldots,n}$. The LMS estimator then corresponds to

$$\underset{\hat{\theta}}{\text{Minimize}} \underset{i}{\text{med}} (y_i - \theta)^2 . \tag{4.11}$$

This estimator will be investigated more fully in Chapter 4, where it will be shown that the LMS estimate T for θ corresponds to the midpoint of the "shortest half" of the sample, because one can prove that

$$T - m_T \text{ and } T + m_T \text{ are both observations in the sample} , \tag{4.12}$$

where $m_T^2 = \text{med}_i (y_i - T)^2$ equals the minimum of (4.11). This property can also be used in the regression model with intercept term, obtained by putting $x_{ip} = 1$ for all i. From (4.12) it follows that for an LMS solution $\hat{\theta}$, both hyperplanes

$$y = x_1\hat{\theta}_1 + \cdots + x_p\hat{\theta}_p - m_T$$

and

$$y = x_1\hat{\theta}_1 + \cdots + x_p\hat{\theta}_p + m_T$$

contain at least one observation. Therefore, the LMS solution corresponds to finding the thinnest "hyperstrip" (i.e., the region between two parallel hyperplanes) covering half of the observations. To be exact, the thickness of the hyperstrip is measured in the vertical direction, and it must contain at least $[n/2] + 1$ points.

From previous experience with robustness, it seems natural to replace the square in (4.1) by the absolute value, yielding

$$\underset{\hat{\theta}}{\text{Minimize}} \underset{i}{\text{med}} |r_i| . \tag{4.13}$$

However, it turns out that (4.12) no longer holds for that estimator, because there may be a whole region of solutions with the same objective function (4.13). (This can only happen when n is even, because $\text{med}_i |r_i|$ is the average of two absolute residuals. An example will be given in Section 2 of Chapter 4.) We will show in Chapter 4 that every solution of (4.1) is also a solution of (4.13), but not vice versa. Things become much more simple when the hth ordered squared residual is to be minimized, as in (4.9) and (4.10), because this is always equivalent to

$$\underset{\hat{\theta}}{\text{Minimize}} |r|_{h:n} , \tag{4.14}$$

where $|r|_{1:n} \leq |r|_{2:n} \leq \cdots \leq |r|_{n:n}$ are the ordered absolute residuals.

Steele and Steiger (1986) also investigated some properties of the estimator defined in (4.14), where they put $h = [n/2] + 1$. This reduces to the median for n odd and to the *high median*, or larger of the two middle values, if n is even. Their work is restricted to the case of simple regression $\hat{y} = \hat{\theta}_1 x + \hat{\theta}_2$. They propose necessary and sufficient conditions for a local minimum of the objective function, which yield necessary conditions for global minimizers. More details on these algorithms are given in Section 2 of Chapter 5.

In order to show that the objective functions of (4.13) and (4.14) satisfy a Lipschitz condition, we need the following lemma.

Lemma 2. Let $(a_1, a_2, \ldots, a_n)$ and $(b_1, b_2, \ldots, b_n)$ be any pair of samples with real elements. Then:

(i) For each integer $1 \le h \le n$, it holds that

$$|a_{h:n} - b_{h:n}| \le \max_k |a_k - b_k| .$$

(ii) Also $|\text{med}_i\, a_i - \text{med}_j\, b_j| \le \max_k |a_k - b_k|$.

(iii) The sharpest upper bound on (i) and (ii) is

$$\min_{f \in \mathcal{S}_n} \max_k |a_k - b_{f(k)}| = \max_h |a_{h:n} - b_{h:n}| ,$$

(where $\mathcal{S}_n$ is the set of all permutations on $\{1, \ldots, n\}$), which is a metric on the set of all samples $\{a_1, \ldots, a_n\}$ in which the sequence of the observations is disregarded.

Proof. (i) Put $c := \max_k |a_k - b_k|$. We first show that

$$|a_{1:n} - b_{1:n}| \le c$$

(otherwise assume without loss of generality that $a_{1:n} + c < b_{1:n}$, but then there can be no element b_j such that $|a_{1:n} - b_j| \le c$). Analogously, we can show that $|a_{n:n} - b_{n:n}| \le c$. For the general case, assume that there exists some $1 < h < n$ such that $a_{h:n} + c < b_{h:n}$. From the definition of c, there exists a permutation f of $\{1, \ldots, n\}$ such that $|a_{j:n} - b_{f(j):n}| \le c$ for all j. However, $j \le h$ must imply $f(j) \le h$ because otherwise $b_{f(j):n} - a_{j:n} \ge b_{h:n} - a_{h:n} > c$. Therefore, $f(\{1, \ldots, h\}) = \{1, \ldots, h\}$, but this is a contradiction because h itself cannot be attained.

(ii) If n is odd $(n = 2h - 1)$ then the median is simply the hth order

statistic. If n is even ($n = 2h$), then

$$\begin{aligned}|\operatorname{med}_i a_i - \operatorname{med}_j b_j| &= |\tfrac{1}{2}(a_{h:n} + a_{h+1:n}) - \tfrac{1}{2}(b_{h:n} + b_{h+1:n})| \\ &\le \tfrac{1}{2}|a_{h:n} - b_{h:n}| + \tfrac{1}{2}|a_{h+1:n} - b_{h+1:n}| \\ &\le c\,.\end{aligned}$$

(iii) The inequality $\le$ is immediate because this combination corresponds to a particular choice of f, and $\ge$ follows from (i). To see why this is a metric, note that $\max_h |a_{h:n} - b_{h:n}| = 0$ if and only if $a_{h:n} = b_{h:n}$ for all h. The symmetry property and triangle inequality are also straightforward. □

For continuous distribution functions F and G, this metric amounts to

$$d(F, G) = \sup_t |F^{-1}(t) - G^{-1}(t)|\,,$$

which gives the right answer at translation families but can easily become infinite.

Theorem 5. (i) For each integer $1 \le h \le n$ it holds that

$$\sup_{\theta \ne \theta'} \frac{\big||r(\theta)|_{h:n} - |r(\theta')|_{h:n}\big|}{\|\theta - \theta'\|} \le \max_i \|\mathbf{x}_i\|\,.$$

$$\text{(ii)} \quad \sup_{\theta \ne \theta'} \frac{\big|\operatorname{med}_i |r_i(\theta)| - \operatorname{med}_j |r_j(\theta)|\big|}{\|\theta - \theta'\|} \le \max_i \|\mathbf{x}_i\|\,.$$

Proof. From (i) of Lemma 2 it follows that

$$\begin{aligned}\big||r(\theta)|_{h:n} - |r(\theta')|_{h:n}\big| &\le \max_i \big||r_i(\theta)| - |r_i(\theta')|\big| \\ &= \max_i \big||y_i - x_i\theta| - |y_i - \mathbf{x}_i\theta'|\big| \\ &\le \max_i |y_i - x_i\theta - y_i + x_i\theta'| \\ &= \max_i |x_i(\theta - \theta')| \\ &\le \|\theta - \theta'\| \max_i \|\mathbf{x}_i\|\,.\end{aligned}$$

Part (ii) is completely analogous.

Note that Theorem 5 implies that $|r|_{h:n}$ and $\operatorname{med}_i |r_i|$ are continuous in $\boldsymbol{\theta}$ (but they are not everywhere differentiable).

A disadvantage of the LMS method is its lack of efficiency (because of its $n^{-1/3}$ convergence, which is proved in Section 4 of Chapter 4) when the errors would really be normally distributed. 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. After all, saying that the LS method is more efficient at the normal is merely a tautology, because Gauss actually introduced the normal distribution in order to suit that method (Huber 1972, p. 1042). However, it is not so difficult to improve the efficiency of the LMS estimator. One can use the LMS estimates as starting values for computing a one-step M-estimator (Bickel 1975) in the following way: suppose we have the LMS solution $(\theta_1^*, \ldots, \theta_p^*)^t$ and a corresponding scale estimate σ^*; then the *one-step M-estimator* (OSM) is defined as

$$\hat{\boldsymbol{\theta}}_{\mathrm{OSM}} = \boldsymbol{\theta}^* + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\psi(r_1^*/\sigma^*), \ldots, \psi(r_n^*/\sigma^*))^t \frac{\sigma^*}{B(\psi, \Phi)}, \quad (4.15)$$

where

$$B(\psi, \Phi) = \int \psi'(u)\, d\Phi(u) \quad \text{and} \quad r_i^* = y_i - \mathbf{x}_i\boldsymbol{\theta}^*$$

and $\mathbf{X}$ is an n-by-p matrix, the rows of which are the vectors $\mathbf{x}_i$. Afterwards σ^* can be replaced by $\hat{\sigma}_{OSM}$, using a M-estimator for scale.

If one uses a *redescending* ψ-function in (4.15), that is, a function for which $\psi(x) = 0$ whenever $|x| \geq c$, the large outliers will not enter the computation. One possible choice is the hyperbolic tangent estimator (Hampel et al. 1981), which possesses maximal asymptotic efficiency subject to certain robustness requirements. It is given by

$$\psi(x) = \begin{cases} x, & 0 \leq |x| \leq d \\ (A(k-1))^{1/2} \tanh\{\tfrac{1}{2}((k-1)B^2/A)^{1/2}(c - |x|)\} \operatorname{sgn}(x), & d \leq |x| \leq c \\ 0, & |x| \geq c, \end{cases} \quad (4.16)$$

where $0 < d < c$ satisfies $d = (A(k-1))^{1/2} \tanh\{\frac{1}{2}((k-1)B^2/A)^{1/2}(c - d)\}$, $A = \int \psi^2\, d\Phi$, and $B = \int \psi'\, d\Phi$. (For instance, $c = 3.0$ and $k = 5.0$ yields $A = 0.680593$, $B = 0.769313$, and $d = 1.47$.) Another possibility is the biweight ψ-function (Beaton and Tukey, 1974) corresponding to

$$\psi(x) = \begin{cases} x(1 - (x/c)^2)^2, & |x| \leq c \\ 0, & |x| \geq c. \end{cases} \quad (4.17)$$

In either case, 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 is $n^{1/2}$ consistent, but in general it even holds when the starting value is better than $n^{1/4}$ consistent (Bickel, personal communication, 1983) as is the case for LMS. Formally,

$$\mathcal{L}(n^{1/2}(\hat{\boldsymbol{\theta}}_{\mathrm{OSM}} - \boldsymbol{\theta}) \rightarrow N(\mathbf{0}, V(\psi_{\hat{\sigma}}, F)L^{-1}) ,$$

where $\boldsymbol{\theta}$ is the unknown parameter vector, $\hat{\sigma} = \hat{\sigma}_{\mathrm{OSM}}$, $L = \lim_{n\to\infty} (\mathbf{X}'\mathbf{X}/n)$, and F is the true distribution of the errors. $V(\psi_{\hat{\sigma}}, F)$ is called the *asymptotic variance*. When the errors are really normally distributed, then $F(t) = \Phi(t/\hat{\sigma})$ and the asymptotic variance can be calculated. Indeed, in that case

$$\begin{aligned} V(\psi_{\hat{\sigma}}, F) &= \frac{\int (\psi(t/\hat{\sigma}))^2 \, dF(t)}{\left[\int (\psi(t/\hat{\sigma}))' \, dF(t)\right]^2} \\ &= \hat{\sigma}^2 \frac{\int \psi^2(t/\hat{\sigma}) \, d\Phi(t/\hat{\sigma})}{\left[\int \psi'(t/\hat{\sigma}) \, d\Phi(t/\hat{\sigma})\right]^2} \\ &= \hat{\sigma}^2 \frac{\int \psi^2(u) \, d\Phi(u)}{\left[\int \psi'(u) \, d\Phi(u)\right]^2} \\ &= \hat{\sigma}^2 V(\psi, \Phi) . \end{aligned} \tag{4.18}$$

For ψ defined as in (4.16),

$$V(\psi, \Phi) = A/B^2 . \tag{4.19}$$

Consequently, one can say that the variance–covariance matrix of the estimated regression coefficients is (approximately) equal to the p-by-p matrix

$$\hat{\sigma}^2 V(\psi, \Phi)(\mathbf{X}'\mathbf{X})^{-1} . \tag{4.20}$$

For the LS estimator, $\psi(r) = r$ and hence $V(\psi, \Phi)$ equals 1. Replacing this value in (4.20) one recovers the well-known formula. Furthermore,

expression (4.20) can be used to calculate the asymptotic efficiency e for the combined procedure (LMS + one-step M) in a normal error model, namely

$$e = 1/V(\psi, \Phi)\,.$$

Hampel et al. (1981, Table 2) give a list of values of e for different constants c and k, as well as the corresponding A, B, and d. For instance, for $c = 3$ and $k = 5$ in (4.16), they obtain $e = 86.96\%$.

The diagonal elements of the matrix in (4.20) are the variances of the estimated regression coefficients $\hat{\theta}_j$. Therefore, it is possible to construct an approximate $(1-\alpha)$ confidence interval for each θ_j, namely

$$[\hat{\theta}_j - \hat{\sigma}\sqrt{V(\psi, \Phi)((\mathbf{X}'\mathbf{X})^{-1})_{jj}}t_{n-p,1-\alpha/2},$$
$$\hat{\theta}_j + \hat{\sigma}\sqrt{V(\psi, \Phi)((\mathbf{X}'\mathbf{X})^{-1})_{jj}}t_{n-p,1-\alpha/2}]\,. \qquad (4.21)$$

Another possibility for improving the efficiency is to use reweighted least squares. To each observation $(\mathbf{x}_i, y_i)$, one assigns a weight w_i that is a function of the standardized LMS residuals r_i/σ^* (in absolute value). For this purpose, one can choose several types of functions. The first kind of weight function that we will consider here is of the form

$$w_i = \begin{cases} 1 & \text{if } |r_i/\sigma^*| \le c_1 \\ 0 & \text{otherwise}\,. \end{cases} \qquad (4.22)$$

This weight function, yielding only binary weights, produces a clear distinction between "accepted" and "rejected" points.

The second type of weight function is less radical. It consists of introducing a linear part that smoothes the transition from weight 1 to weight 0. In that way, far outliers (this means cases with large LMS residuals) disappear entirely and intermediate cases are gradually down-weighted. In the general formula

$$w_i = \begin{cases} 1 & \text{if } |r_i/\sigma^*| \le c_2 \\ (c_3 - |r_i/\sigma^*|)/(c_3 - c_2) & \text{if } c_2 \le |r_i/\sigma^*| \le c_3 \\ 0 & \text{otherwise} \end{cases} \qquad (4.23)$$

the constants c_2 and c_3 have to be chosen.

A third weight function can be defined by means of the hyperbolic tangent function (4.16), namely

$$w_i = \begin{cases} 1 & \text{if } |r_i/\sigma^*| \le d \\ \dfrac{(A(k-1))^{1/2}\tanh\{\frac{1}{2}((k-1)B^2/A)^{1/2}(c - |r_i/\sigma^*|)\}}{|r_i/\sigma^*|} & \text{if } d \le |r_i/\sigma^*| \le c \\ 0 & \text{otherwise}\,, \end{cases} \qquad (4.24)$$

where the constants c, k, d, A, and B correspond to those already defined in (4.16).

Once a weight function is selected, one replaces all observations (x_i, y_i) by $(w_i^{1/2}x_i, w_i^{1/2}y_i)$. On these weighted observations, a standard least squares program may be used to obtain the final estimate. The RLS results in PROGRESS are obtained with the weight function (4.22) with $c_1 = 2.5$.

Another way to improve the slow rate of convergence of the LMS consists of using a different objective function. Instead of adding all the squared residuals as in LS, one can limit one's attention to a "trimmed" sum of squares. This quantity is defined as follows: first one orders the squared residuals from smallest to largest, denoted by

$$(r^2)_{1:n} \leq (r^2)_{2:n} \leq \cdots \leq (r^2)_{n:n} .$$

Then one adds only the first h of these terms. In this way, Rousseeuw (1983) defined the *least trimmed squares* (LTS) estimator

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i=1}^{h} (r^2)_{i:n} . \tag{4.25}$$

Putting $h = [n/2] + 1$, the LTS attains the same breakdown point as the LMS (see Theorem 2). Moreover, for $h = [n/2] + [(p+1)/2]$ the LTS reaches the maximal possible value for the breakdown point given in Theorem 4. (Note that the LTS has nothing to do with the trimmed least squares estimators described by Ruppert and Carroll 1980.) Before we investigate the robustness properties of the LTS, we will first verify its equivariance.

Lemma 3. The LTS estimator is regression, scale, and affine equivariant.

Proof. Regression equivariance follows from the identity

$$\sum_{i=1}^{h} ((y_i + \mathbf{x}_i\mathbf{v} - \mathbf{x}_i\{\mathbf{v} + \boldsymbol{\theta}\})^2)_{i:n} = \sum_{i=1}^{h} ((y_i - \mathbf{x}_i\boldsymbol{\theta})^2)_{i:n}$$

for any column vector $\mathbf{v}$. Scale and affine equivariance are analogous. □

Theorem 6. The breakdown point of the LTS method defined in (4.25) with $h = [n/2] + [(p+1)/2]$ equals

$$\varepsilon_n^*(T, Z) = ([(n-p)/2] + 1)/n .$$

Proof. In order to prove this theorem we again assume that all observations with $(x_{i1}, \ldots, x_{ip}) = \mathbf{0}$ have been deleted and that the observations are in general position.

1. We first show that $\varepsilon_n^*(T, Z) \geq ([(n-p)/2]+1)/n$. Since the sample $Z = \{(\mathbf{x}_i, y_i); i = 1, \ldots, n\}$ consists of n points in general position, it holds that

$$\rho = \tfrac{1}{2} \inf \{\tau > 0; \text{ there exists a } (p-1)\text{-dimensional subspace } V \text{ of } (y=0) \text{ such that } V^\tau \text{ covers at least } p \text{ of the } \mathbf{x}_i\}$$

is strictly positive. Suppose $\boldsymbol{\theta}$ minimizes (4.25) for Z, and denote by H the corresponding hyperplane given by the equation $y = \mathbf{x}\boldsymbol{\theta}$. We put $M = \max_i |r_i|$, where $r_i = y_i - \mathbf{x}_i\boldsymbol{\theta}$. Now construct any contaminated sample $Z' = \{(\mathbf{x}_i', y_i'); \ i = 1, \ldots, n\}$ by retaining $n - [(n-p)/2] = [(n+p+1)/2]$ observations of Z and by replacing the others by arbitrary values. It now suffices to prove that $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ is bounded, where $\boldsymbol{\theta}'$ corresponds to Z'. Without loss of generality assume $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$, so the corresponding hyperplane H' is different from H. Repeating now the reasoning of the first part of the proof of Theorem 2, it follows that

$$|r_a' - r_a| > \rho(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| - 2\|\boldsymbol{\theta}\|) ,$$

where r_a and r_a' are the residuals associated with H and H' corresponding to the point $(\mathbf{x}_a, y_a)$. Now the sum of the first h squared residuals of the new sample Z' with respect to the old $\boldsymbol{\theta}$, with at least $[(n+p+1)/2] \geq h$ of these residuals being the same as before, is less than or equal to hM^2. Because $\boldsymbol{\theta}'$ corresponds to Z' it follows that also

$$\sum_{i=1}^{h} ((y_i' - \mathbf{x}_i'\boldsymbol{\theta}')^2)_{i:n} \leq hM^2 .$$

If we now assume that

$$\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \geq 2\|\boldsymbol{\theta}\| + M(1+\sqrt{h})/\rho ,$$

then for all a in A it holds that

$$|r_a' - r_a| > \rho(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| - 2\|\boldsymbol{\theta}\|) \geq M(1+\sqrt{h}) ,$$

so

$$|r_a'| \geq |r_a' - r_a| - |r_a| > M(1+\sqrt{h}) - M = M\sqrt{h} .$$

Now note that $n - |A| \leq h - 1$. Therefore any set of h of the $(\mathbf{x}_i', y_i')$ must

contain at least one of the $(\mathbf{x}_a, y_a)$, so

$$\sum_{i=1}^{h} ((y_i' - \mathbf{x}_i'\boldsymbol{\theta}')^2)_{i:n} \geq (r_a')^2 > hM^2 ,$$

a contradiction. This implies that

$$\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| < 2\|\boldsymbol{\theta}\| + M(1+\sqrt{h})/\rho < \infty$$

for all such samples Z'.

2. The opposite inequality $\varepsilon_n^*(T, Z) \leq ([(n-p)/2]+1)/n$ immediately follows from Theorem 4 and Lemma 3. □

REMARK 1. Another way to interpret Theorem 6 is to say that T remains bounded whenever strictly more than $\frac{1}{2}(n+p-1)$ observations are uncontaminated.

REMARK 2. The value of h yielding the maximal value of the breakdown point can also be found by the following reasoning based on the proofs of Theorems 2 and 6. On the one hand, the number of bad observations $n - |A|$ must be strictly less than h; on the other hand, $|A| + p - 1$ must be at least h. The best value of h is then obtained by minimizing $|A|$ over h subject to $|A| - 1 \geq n - h$ and $|A| - 1 \geq h - p$, which yields $h = [n/2] + [(p+1)/2]$.

REMARK 3. In general, h may depend on some trimming proportion α, for instance by means of $h = [n(1-\alpha)] + [\alpha(p+1)]$ or $h = [n(1-\alpha)] + 1$. Then the breakdown point ε_n^* is roughly equal to this proportion α. For α tending to 50%, one finds again the LTS estimator, whereas for α tending to 0%, the LS estimator is obtained.

The following corollary shows that also the LTS satisfies the exact fit property.

Corollary. If there exists some $\boldsymbol{\theta}$ such that strictly more than $\frac{1}{2}(n+p-1)$ of the observations satisfy $y_i = \mathbf{x}_i\boldsymbol{\theta}$ exactly and are in general position, then the LTS solution equals $\boldsymbol{\theta}$ whatever the other observations are.

For instance, in the case of simple regression it follows that whenever 11 out of 20 observations lie on one line, this line will be obtained.

Unlike the slow convergence rate of the LMS, the LTS converges like $n^{-1/2}$, with the same asymptotic efficiency at the normal distribution as

the M-estimator defined by

$$\psi(x) = \begin{cases} x, & |x| \leq \Phi^{-1}(1-\alpha/2) \\ 0, & \text{otherwise}, \end{cases} \tag{4.26}$$

which is called a Huber-type skipped mean in the case of location (see Chapter 4 for details). 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.

REMARK. Until now we have considered the estimators obtained by substituting the sum in the definition of the LS estimator by a median, yielding LMS, and by a trimmed sum, leading to LTS. Another idea would be to replace the sum by a winsorized sum, yielding something that could be called *least winsorized squares* (LWS) regression, given by

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i=1}^{h} (r^2)_{i:n} + (n-h)(r^2)_{h:n}, \tag{4.27}$$

where h may also depend on some fraction α. Like LMS and LTS, this estimator is regression, scale, and affine equivariant, and it possesses the same breakdown point for a given value of h. However, some preliminary simulations have revealed that the LWS is inferior to the LTS.

S-estimators (Rousseeuw and Yohai, 1984) form another class of high-breakdown affine equivariant estimators with convergence rate $n^{-1/2}$. They are defined by minimization of the dispersion of the residuals:

$$\underset{\hat{\theta}}{\text{Minimize}}\ s(r_1(\boldsymbol{\theta}), \ldots, r_n(\boldsymbol{\theta})), \tag{4.28}$$

with final scale estimate

$$\hat{\sigma} = s(r_1(\hat{\boldsymbol{\theta}}), \ldots, r_n(\hat{\boldsymbol{\theta}})). \tag{4.29}$$

The dispersion $s(r_1(\boldsymbol{\theta}), \ldots, r_n(\boldsymbol{\theta}))$ is defined as the solution of

$$\frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{r_i}{s}\right) = K. \tag{4.30}$$

K is often put equal to $E_{\Phi}[\rho]$, where Φ is the standard normal. The function ρ must satisfy the following conditions:

(S1) ρ is symmetric and continuously differentiable, and $\rho(0) = 0$.

(S2) There exists $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$.

[If there happens to be more than one solution to (4.30), then put $s(r_1, \ldots, r_n)$ equal to the supremum of the set of solutions; this means $s(r_1, \ldots, r_n) = \sup\{s; (1/n)\,\Sigma\,\rho(r_i/s) = K\}$. If there exists no solution to (4.30), then put $s(r_1, \ldots, r_n) = 0$.]

The estimator in (4.28) is called an *S*-estimator because it is derived from a scale statistic in an implicit way. (Actually, s given by (4.30) is an *M*-estimator of scale.) Clearly *S*-estimators are regression, scale, and affine equivariant.

Because of condition (S2), $\psi(x) = \rho'(x)$ will always be zero from a certain value of x on, so ψ is redescending. An example is the ρ-function corresponding to

$$\rho(x) = \begin{cases} \dfrac{x^2}{2} - \dfrac{x^4}{2c^2} + \dfrac{x^6}{6c^4} & \text{for } |x| \le c \\ \dfrac{c^2}{6} & \text{for } |x| > c\,, \end{cases} \tag{4.31}$$

the derivative of which is Tukey's biweight function defined in (4.17). Another possibility is to take a ρ corresponding to the hyperbolic tangent estimator (4.16).

In order to show that the breakdown point of *S*-estimators is also 50% we need a preliminary lemma, in which an extra condition on the function ρ is needed:

(S3) $$\frac{K}{\rho(c)} = \frac{1}{2}$$

This condition is easy to fulfill. In the case of (4.31) with $K = E_\Phi[\rho]$, it is achieved by taking $c = 1.547$. Let us now look at the scale estimator $s(r_1, \ldots, r_n)$, which is defined by (4.30) for any sample $(r_1, \ldots, r_n)$.

Lemma 4. For each ρ satisfying conditions (S1)–(S3) and for each n, there exist positive constants α and β such that the estimator s given by (4.30) satisfies

$$\alpha \operatorname*{med}_i |r_i| \le s(r_1, \ldots, r_n) \le \beta \operatorname*{med}_i |r_i|\,.$$

Here $\operatorname{med}_i |r_i|$ or $s(r_1, \ldots, r_n)$ may even be zero.

Proof. 1. We first consider the case n odd ($n = 2m + 1$). Put $S = s(r_1, \ldots, r_n)$ for ease of notation. We will show that

$$\frac{\underset{i}{\text{med}}\,|r_i|}{c} \leq S \leq \frac{\underset{i}{\text{med}}\,|r_i|}{\rho^{-1}(\rho(c)/(n+1))}\,.$$

Suppose $\text{med}_i\,|r_i| > cS$. Because $\text{med}_i\,|r_i| = |r|_{m+1:n}$ it holds that at least $m+1$ of the $|r_i|/S$ are larger than c. Consequently,

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{|r_i|}{S}\right) \geq \frac{1}{n}\,(m+1)\rho(c) > \rho(c)/2 = K\,,$$

which is a contradiction. Therefore, $\text{med}_i\,|r_i| \leq cS$.

Now suppose that $\text{med}_i\,|r_i| < \rho^{-1}(\rho(c)/(n+1))S$. This would imply that the first $m+1$ of the $|r_i|/S$ are strictly smaller than $\rho^{-1}(\rho(c)/(n+1))$. Introducing this in $(1/n)\,\Sigma_{i=1}^{n}\,\rho(|r_i|/S)$, we find that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{|r_i|}{S}\right) &< \frac{m+1}{n}\,\rho\left(\rho^{-1}\left(\frac{\rho(c)}{n+1}\right)\right) + \frac{m}{n}\,\rho(c)\\ &\leq \frac{m+1}{n}\left(\frac{\rho(c)}{n+1}\right) + \frac{m}{n}\,\rho(c)\\ &= \rho(c)\left\{\frac{1}{2n} + \frac{m}{n}\right\}\\ &= \tfrac{1}{2}\rho(c) = K\,,\end{aligned}$$

which is again a contradiction, so $\text{med}_i\,|r_i| \geq \rho^{-1}\{\rho(c)/(n+1)\}S$.

2. Let us now treat the case where n is even ($n = 2m$). We will prove that

$$\frac{\underset{i}{\text{med}}\,|r_i|}{c} \leq S \leq \frac{\underset{i}{\text{med}}\,|r_i|}{\tfrac{1}{2}\rho^{-1}(2\rho(c)/(n+2))}\,.$$

Suppose first that $\text{med}_i\,|r_i| > cS$. Since n is even, $\text{med}_i\,|r_i| = \tfrac{1}{2}\{|r|_{m:n} + |r|_{m+1:n}\}$. Then, at least m of the $|r_i|/S$ are strictly larger than c, and

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{|r_i|}{S}\right) > \frac{m}{n}\,\rho(c) = \tfrac{1}{2}\rho(c) = K\,,$$

except when all other $|r_i|$ are zero, but then the set of solutions of (4.30) is the interval $(0, 2 \operatorname{med}_i |r_i|/c]$, so $S = 2 \operatorname{med}_i |r_i|/c$. In either case, $\operatorname{med}_i |r_i| \leq cS$.

Suppose now that $\operatorname{med}_i |r_i| < \frac{1}{2}\rho^{-1}\{2\rho(c)/(n+2)\}S$. Then $|r|_{m+1:n} < \rho^{-1}\{2\rho(c)/(n+2)\}S$. Hence, the first $m+1$ of the $|r_i|/S$ are less than $\rho^{-1}\{2\rho(c)/(n+2)\}$. Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{|r_i|}{S}\right) < \frac{m+1}{n}\,\frac{2\rho(c)}{n+2} + \frac{m-1}{n}\,\rho(c)$$

$$= \frac{\rho(c)}{n} + \frac{m-1}{n}\,\rho(c)$$

$$\leq (m/n)\rho(c) = \rho(c)/2 = K\,.$$

Finally, $\operatorname{med}_i |r_i| \geq \frac{1}{2}\rho^{-1}\{2\rho(c)/(n+2)\}S$.

3. We will now deal with special cases with zeroes.

Let us start with n odd ($n = 2m+1$). When $\operatorname{med}_i |r_i| = 0$, then the first $m+1$ of the $|r_i|$ are zero, and whatever the value of S, we always have $(1/n)\,\Sigma_{i=1}^{n}\,\rho(|r_i|/S) < \frac{1}{2}\rho(c)$. Therefore the set of solutions is empty, so $S = 0$ by definition.

On the other hand, when $\operatorname{med}_i |r_i| > 0$, then there are at least $m+1$ nonzero $|r_i|$, hence

$$\lim_{s \searrow 0}\left\{\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{|r_i|}{s}\right)\right\} \geq ((m+1)/n)\rho(c) > K$$

$$\lim_{s \nearrow \infty}\left\{\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{|r_i|}{s}\right)\right\} = 0 < K\,.$$

Therefore, there exists a strictly positive solution S.

Let us now consider n even ($n = 2m$). When $\operatorname{med}_i |r_i| = 0$, then the first $m+1$ of the $|r_i|$ are zero, and whatever the value of S, we have again that $(1/n)\,\Sigma_{i=1}^{n}\,\rho(|r_i|/S) < \frac{1}{2}\rho(c)$, so $S = 0$.

When $\operatorname{med}_i |r_i| > 0$, then we are certain that $|r|_{m+1:n} > 0$ too. We shall consider both the case $|r|_{m:n} > 0$ and $|r|_{m:n} = 0$. If $|r|_{m:n}$ is strictly positive, then there are at least $m+1$ nonzero $|r_i|$ and one can follow the same reasoning as in the case where n is odd. If on the other hand $|r|_{m:n} = 0$, then we are in the special case where the ordered $|r_i|$ can be written as a sequence of m zeroes and with $2 \operatorname{med}_i |r_i|$ in the $(m+1)$th position. By definition

$$S = \sup\,(0, 2 \operatorname*{med}_i |r_i|/c] = 2 \operatorname*{med}_i |r_i|/c > 0\,.$$

We may therefore conclude (for both odd and even n) that $\text{med}_i\,|r_i|$ is zero if and only if S is zero. □

REMARK. Note that Lemma 4 (as well as Theorems 7 and 8) do not rely on the assumption that $K = E_\Phi[\rho]$, which is only needed if one wants (4.30) to yield a consistent scale estimate for normally distributed residuals.

Theorem 7. For any ρ satisfying (S1) to (S3), there always exists a solution to (4.28).

Proof. Making use of the preceding lemma, this follows from the proof of Theorem 1, where the result was essentially given for the minimization of $\text{med}_i\,|r_i|$. □

Theorem 8. An S-estimator constructed from a function ρ satisfying (S1) to (S3) has breakdown point

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

at any sample $\{(\mathbf{x}_i, y_i); i = 1, \ldots, n\}$ in general position.

Proof. This follows from Theorem 2 by making use of Lemma 4. □

The breakdown point depends only slightly on n, and for $n \to \infty$ we obtain $\varepsilon^* = 50\%$, the best we can expect. The following result concerns the exact fit property for S-estimators, which again illustrates their high resistance.

Corollary. If there exists some $\boldsymbol{\theta}$ such that at least $n - [n/2] + p - 1$ of the points satisfy $y_i = \mathbf{x}_i\boldsymbol{\theta}$ *exactly* and are in general position, then the S-estimate for the regression vector will be equal to $\boldsymbol{\theta}$ whatever the other observations are.

REMARK 1. If condition (S3) is replaced by

$$\frac{K}{\rho(c)} = \alpha\,,$$

where $0 < \alpha \leq \frac{1}{2}$, then the corresponding S-estimators have a breakdown point tending to $\varepsilon^* = \alpha$ when $n \to \infty$. If it is assumed that $K = E_\Phi[\rho]$ in order to achieve a consistent scale estimate for normally distributed

residuals, one can trade a higher asymptotic efficiency against a lower breakdown point.

REMARK 2. Note that S-estimators satisfy the same first-order necessary conditions as the M-estimators discussed in Section 2 of Chapter 1. Indeed, let $\boldsymbol{\theta}$ be any p-dimensional parameter vector. By definition, we know that

$$S(\boldsymbol{\theta}) = s(r_1(\boldsymbol{\theta}), \ldots, r_n(\boldsymbol{\theta})) \geq \hat{\sigma} = S(\hat{\boldsymbol{\theta}}) .$$

Keeping in mind that $S(\boldsymbol{\theta})$ satisfies

$$(1/n) \sum_{i=1}^{n} \rho(r_i(\boldsymbol{\theta})/S(\boldsymbol{\theta})) = K$$

and that $\rho(u)$ is nondecreasing in $|u|$, it follows that always

$$(1/n) \sum_{i=1}^{n} \rho(r_i(\boldsymbol{\theta})/\hat{\sigma}) \geq K .$$

At $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, this becomes an equality. Therefore, $\hat{\boldsymbol{\theta}}$ minimizes $(1/n)\ \Sigma_{i=1}^{n}$ $\rho(r_i(\boldsymbol{\theta})/\hat{\sigma})$. (This fact cannot be used for determining $\hat{\boldsymbol{\theta}}$ in practice, because $\hat{\sigma}$ is fixed but unknown.) Differentiating with respect to $\boldsymbol{\theta}$, we find

$$(1/n) \sum_{i=1}^{n} \psi(r_i(\hat{\boldsymbol{\theta}})/\hat{\sigma})\mathbf{x}_i = \mathbf{0} .$$

If we denote $\rho - K$ by χ, we conclude that $(\hat{\boldsymbol{\theta}}, \hat{\sigma})$ is a solution of the system of equations

$$\begin{cases} \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \psi(r_i(\hat{\boldsymbol{\theta}})/\hat{\sigma})\mathbf{x}_i = \mathbf{0} \\ \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \chi(r_i(\hat{\boldsymbol{\theta}})/\hat{\sigma}) = 0 \end{cases} \tag{4.32}$$

(described in Section 2 of Chapter 1) for defining an M-estimator. Unfortunately, these equations cannot be used directly because there are infinitely many solutions (ψ is redescending) and the iteration procedures for the computation of M-estimators easily end in the wrong place if there are leverage points. [This means we still have to minimize (4.28) with brute force in order to actually compute the S-estimate in a practical

situation.] Therefore it would be wrong to say that *S*-estimators *are* *M*-estimators, because their computation and breakdown are completely different, but they do satisfy similar first-order necessary conditions.

Besides their high resistance to contaminated data, *S*-estimators also behave well when the data are not contaminated. To show this, we will look at the asymptotic behavior of *S*-estimators at the central Gaussian model, where $(\mathbf{x}_i, y_i)$ are i.i.d. random variables satisfying

$$y_i = \mathbf{x}_i \boldsymbol{\theta}_0 + e_i \,, \tag{4.33}$$

$\mathbf{x}_i$ follows some distribution H, and e_i is independent of $\mathbf{x}_i$ and distributed like $\Phi(e/\sigma_0)$ for some $\sigma_0 > 0$.

Theorem 9. Let ρ be a function satisfying (S1) and (S2), with derivative $\rho' = \psi$. Assume that:

(1) $\psi(u)/u$ is nonincreasing for $u > 0$;
(2) $E_H[\|\mathbf{x}\|] < \infty$, and H has a density.

Let $(\mathbf{x}_i, y_i)$ be i.i.d. according to the model in (4.33), and let $\hat{\boldsymbol{\theta}}_n$ be a solution of (4.28) for the first n points, and $\hat{\sigma}_n = s(r_1(\hat{\boldsymbol{\theta}}_n), \ldots, r_n(\hat{\boldsymbol{\theta}}_n))$. If $n \to \infty$ then

$$\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0 \quad \text{a.s.}$$

and

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

Proof. This follows from Theorem 2.2 and 3.1 of Maronna and Yohai (1981), because *S*-estimators satisfy the same first-order necessary conditions as *M*-estimators (according to Remark 2 above). □

Let us now show the asymptotic normality of *S*-estimators.

Theorem 10. Without loss of generality let $\boldsymbol{\theta}_0 = \mathbf{0}$ and $\sigma_0 = 1$. If the conditions of Theorem 9 hold and

(3) ψ is differentiable in all but a finite number of points, $|\psi'|$ is bounded, and $\int \psi' \, d\Phi > 0$;
(4) $E_H[\mathbf{x}^t \mathbf{x}]$ is nonsingular and $E_H[\|\mathbf{x}\|^3] < \infty$, then

$$\mathcal{L}(n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)) \to N\left(\mathbf{0}, E_H[\mathbf{x}^t\mathbf{x}]^{-1}\left\{\int \psi^2\, d\Phi\right\} \Big/ \left\{\int \psi'\, d\Phi\right\}^2\right)$$

and

$$\mathcal{L}(n^{1/2}(\hat{\sigma}_n - \sigma_0)) \to N\left(0, \frac{\int (\rho(y) - E_\Phi[\rho])^2\, d\Phi(y)}{\left\{\int y\psi(y)\, d\Phi(y)\right\}^2}\right).$$

Proof. This follows from Theorem 4.1 of Maronna and Yohai (1981) using Remark 2. □

As a consequence of Theorem 10, we can compute the asymptotic efficiency e of an S-estimator at the Gaussian model as

$$e = \frac{\left(\int \psi'\, d\Phi\right)^2}{\left(\int \psi^2\, d\Phi\right)}.$$

Table 19 gives the asymptotic efficiency of the S-estimators corresponding to the function ρ defined in (4.31) for different values of the breakdown point ε^*. From this table it is apparent that values of c larger than 1.547 yield better asymptotic efficiencies at the Gaussian central model, but yield smaller breakdown points. Furthermore we note that taking $c = 2.560$ yields a value of e which is larger than that of L_1 (for

Table 19. Asymptotic Efficiency of S-Estimators for Different Values of ε^*, Making use of Tukey's Biweight Function

ε^*	e	c	K
50%	28.7%	1.547	0.1995
45%	37.0%	1.756	0.2312
40%	46.2%	1.988	0.2634
35%	56.0%	2.251	0.2957
30%	66.1%	2.560	0.3278
25%	75.9%	2.937	0.3593
20%	84.7%	3.420	0.3899
15%	91.7%	4.096	0.4194
10%	96.6%	5.182	0.4475

which e is about 64%), and gains us a breakdown point of 30%. In practice, we do not recommend the estimators in the table with a breakdown point smaller than 25%. It appears to be better to apply the $c = 1.547$ estimator because of its high breakdown point. From this first solution, one can then compute a one-step M-estimator or a one-step reweighted least squares in order to make up for the initial low efficiency. Such a two-stage procedure inherits the 50% breakdown point from the first stage and inherits the high asymptotic efficiency from the second. An algorithm for computing S-estimators will be described in Chapter 5.

5. RELATION WITH PROJECTION PURSUIT

The goal of projection pursuit (PP) procedures is to discover structure in a multivariate data set by projecting these data in a lower-dimensional space. Such techniques were originally proposed by Roy (1953), Kruskal (1969), and Switzer (1970). The name "projection pursuit" itself was coined by Friedman and Tukey (1974), who developed a successful algorithm. The main problem is to find "good" projections, because arbitrary projections are typically not very informative. Friedman and Stuetzle (1982) give some examples of point configurations with strong structure, which possess projections in which no structure is apparent. Therefore, PP tries out many low-dimensional projections of a high-dimensional point cloud in search for a "most interesting" one, by numerically optimizing a certain objective function (which is also called a "projection index"). Some important applications are PP classification (Friedman and Stuetzle 1980), PP regression (Friedman and Stuetzle 1981), robust principal components (Ruymgaart 1981), and PP density estimation (Friedman et al. 1984). A recent survey of the field has been given by Huber (1985). The program MACSPIN (D^2 Software 1986) enables us to look at two-dimensional projections in a dynamic way.

Let us now show that there is a relation between robust regression and PP. To see this, consider the $(p+1)$-dimensional space of the $(\mathbf{x}_i, y_i)$, where the last component of $\mathbf{x}_i$ equals 1 in the case of regression with a constant. In this space, linear models are defined by

$$(\mathbf{x}, y)\begin{pmatrix} \boldsymbol{\theta} \\ -1 \end{pmatrix} = 0 \tag{5.1}$$

for some p-dimensional column vector $\boldsymbol{\theta}$. In order to find a "good" $\hat{\boldsymbol{\theta}}$, we start by projecting the point cloud on the y-axis in the direction orthogonal to $(\boldsymbol{\theta}, -1)'$ for any vector $\boldsymbol{\theta}$. This means that each $(\mathbf{x}_i, y_i)$ is projected onto $(\mathbf{0}, r_i(\boldsymbol{\theta}))$, where $r_i(\boldsymbol{\theta}) = y_i - \mathbf{x}_i\boldsymbol{\theta}$. Following Rousseeuw (1984, p.

874) and Donoho et al. (1985), we measure the "interestingness" of any such projection by its dispersion

$$s(r_1(\boldsymbol{\theta}), \ldots, r_n(\boldsymbol{\theta})), \tag{5.2}$$

where the objective s is scale equivariant

$$s(\tau r_1, \ldots, \tau r_n) = |\tau| s(r_1, \ldots, r_n) \quad \text{for all } \tau \tag{5.3}$$

but not translation invariant. The PP estimate $\hat{\boldsymbol{\theta}}$ is then obtained by minimization of the projection index (5.2). If $s(r_1, \ldots, r_n) = (\Sigma_{i=1}^n r_i^2/n)^{1/2}$, then this "most interesting" $\hat{\boldsymbol{\theta}}$ is simply the vector of least squares coefficients. Analogously, $s(r_1, \ldots, r_n) = \Sigma_{i=1}^n |r_i|/n$ yields the L_1 estimator, and minimizing $(\Sigma_{i=1}^n |r_i|^q/n)^{1/q}$ gives the L_q-estimators (Gentleman 1965, Sposito et al. 1977). Using a very robust s brings us back to our high-breakdown regression estimators: $s = (\text{med}_i\, r_i^2)^{1/2}$ yields the LMS, $s = (\Sigma_{i=1}^h (r^2)_{i:n}/n)^{1/2}$ yields the LTS, and by putting s equal to a robust M-estimator of scale we obtain S-estimators. Note that the minimization of any s satisfying (5.3) will yield a regression estimator that is regression, scale, and affine equivariant (as discussed in Section 4). So any nice s defines a type of regression estimator; by varying s one obtains an entire class of regression estimators belonging to the PP family. Thus the PP principle extends to cover the linear regression problem and to encompass both classical procedures and high-breakdown methods. This notion is best thought of as "PP Linear Regression," to distinguish it from Friedman and Stuetzle's (1981) nonlinear and nonrobust "PP Regression."

It appears that the only affine equivariant high-breakdown regression estimators known so far (the LMS, the LTS, and S-estimators) are related to PP. (GM-estimators do not have high breakdown, and the repeated median is not affine equivariant.) There is a reason for this apparent relation between high breakdown and PP. Indeed, Donoho, Rousseeuw, and Stahel have found that breakdown properties are determined by behavior near situations of *exact fit*: These are situations where most of the data lie exactly in a regression hyperplane (see Remark 1 following Theorem 3 of Section 4). Such configurations are precisely those having a projection in which most of the data collapse to a point. In other words, high breakdown appears to depend on an estimator's behavior in those situations where certain special kinds of projections occur. Since PP can, in principle, be used to search for such projections, the usefulness of PP in synthesizing high-breakdown procedures is not surprising.

Table 20. Schematic Overview of Some Affine Equivariant Regression Techniques

Criterion	Method	Computation	ε^*
Best linear unbiased	LS	Explicit	0%
Minimax variance	M	Iterative	0%
Bounded influence	GM	Iterative, with weights on $\mathbf{x}_i$ (harder)	Down to 0% if p increases
High breakdown	LMS, LTS, S	Projection pursuit techniques	Constant, up to 50%

Note, however, that PP is not necessarily the only way to obtain high-breakdown equivariant estimators, at least not in multivariate location where Rousseeuw (1983) gives the example of the minimal volume ellipsoid containing at least half the data. Also, not every PP-based affine equivariant estimator is going to have high breakdown (Fill and Johnstone 1984).

The relation of our robust regression estimators with PP also gives a clue to their computational complexity. In principle, all possible projections must be tried out (although the actual algorithm in Chapter 5 can exploit some properties to speed things up). This means that the LMS, the LTS, and S-estimators belong to the highly computer-intensive part of statistics, just like PP and the bootstrap, to which the algorithm is also related. In Table 20 we have a schematic overview of criteria in affine equivariant regression.

*6. OTHER APPROACHES TO ROBUST MULTIPLE REGRESSION

We have seen that the conditions under which the LS criterion is optimal are rarely fulfilled in realistic situations. In order to define more robust regression alternatives, many statisticians have exploited the resistance of the sample median to extreme values. For instance, the L_1 criterion can be seen as a generalization of the univariate median, because the minimization of $\Sigma_{i=1}^{n} |y_i - \hat{\theta}|$ defines the median of n observations y_i. In the regression problem, the L_1 estimator is given by

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i=1}^{n} |r_i| . \tag{6.1}$$

The substitution of the square by the absolute value leads to a considerable gain in robustness. However, in terms of the breakdown point, L_1 is not really better than LS, because the L_1 criterion is robust with respect to outliers in the y_i but is still vulnerable to leverage points. Moreover, Wilson (1978) showed that the efficiency of the L_1 estimator decreases when n increases. From a numerical point of view, the minimization in (6.1) amounts to the solution of a linear program:

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i=1}^{n} (u_i + v_i)$$

under the constraints

$$y_i = \sum_{k=1}^{p} \theta_k x_{ik} + u_i - v_i , \qquad u_i \geq 0, v_i \geq 0 .$$

Barrodale and Roberts (1974) and Sadovski (1977) described algorithms and presented FORTRAN routines for calculating L_1 regression coefficients.

The minimization of an L_q norm (for $1 \leq q \leq 2$) of the residuals has been considered by Gentleman (1965), Forsythe (1972), and Sposito et al. (1977), who presented an algorithm (with FORTRAN code) for the L_q fit of a straight line. Dodge (1984) suggested a regression estimator based on a convex combination of the L_1 and L_2 norms, resulting in

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i=1}^{n} \left((1 - \delta) \frac{r_i^2}{2} + \delta |r_i| \right) \quad \text{with } 0 \leq \delta \leq 1 .$$

Unfortunately, all these proposals possess a zero breakdown point.

In Section 7 of Chapter 2, we listed some estimators for simple regression which are also based on the median. Some of them have been generalized to multiple regression by means of a "sweep" operator (see Andrews 1974).

The idea behind Theil's (1950) estimator, which consists of looking at the median of all pairwise slopes (see Section 7 of Chapter 2), has also been the source of extensions and modifications. A recent proposal comes from Oja and Niinimaa (1984). For each subset $J = \{i_1, i_2, \ldots, i_p\}$ of $\{1, 2, \ldots, n\}$ containing p indices, they define

$$\boldsymbol{\theta}_J = \boldsymbol{\theta}(i_1, i_2, \ldots, i_p) \tag{6.2}$$

as the parameter vector corresponding to the hyperplane going exactly through the p points $(\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_p}, y_{i_p})$. They call these $\boldsymbol{\theta}_J$ *pseudo-*

observations, and there are C_n^p of them. Their idea is now to compute a multivariate location estimator (in p-dimensional space) of all these $\boldsymbol{\theta}_J$. A certain weighted average of the $\boldsymbol{\theta}_J$ yields LS, but of course they want to insert a robust multivariate estimator. If one computes

$$\hat{\theta}_j = \underset{J}{\text{med}}\,(\boldsymbol{\theta}_J)_j \quad \text{for all } j = 1, \ldots, p \tag{6.3}$$

(coordinatewise median over all subsets J), then one obtains a regression estimator that fails to be affine equivariant. For simple regression, (6.3) indeed yields the Theil–Sen estimator described in Section 7 of Chapter 2. In order to obtain an affine equivariant regression estimator, one has to apply a multivariate location estimator T which is itself affine equivariant, meaning that

$$T(\mathbf{z}_1\mathbf{A} + \mathbf{b}, \ldots, \mathbf{z}_k\mathbf{A} + \mathbf{b}) = T(\mathbf{z}_1, \ldots, \mathbf{z}_k)\mathbf{A} + \mathbf{b} \tag{6.4}$$

for any sample $\{\mathbf{z}_1, \ldots, \mathbf{z}_k\}$ of p-dimensional row vectors, for any nonsingular square matrix $\mathbf{A}$, and for any p-dimensional vector $\mathbf{b}$. For this purpose they propose to apply the *generalized median*, an ingenious construction of Oja (1983) which is indeed affine equivariant and will be discussed in Section 1 of Chapter 7. Unfortunately, the computation complexity of this generalized median is enormous (and it would have to be applied to a very large set of $\boldsymbol{\theta}_J$ vectors!). Even when the coordinatewise median (6.3) is applied, the Oja–Niinimaa regression estimator needs considerable computation time because of the C_n^p pseudo-observations. In either case the consideration of all $\boldsymbol{\theta}_J$ is impossible, so it might be useful to consider a random subpopulation of pseudo-observations.

Let us now consider the breakdown point of this technique. We can only be sure that a pseudo-observation $\boldsymbol{\theta}_J = \boldsymbol{\theta}(i_1, \ldots, i_p)$ is "good" when the p points $(\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_p}, y_{i_p})$ are all good. If there is a fraction ε of outliers in the original data, then we can only be certain of a proportion $(1-\varepsilon)^p$ of "good" pseudo-observations. Therefore, we must have that

$$(1-\varepsilon)^p \geq \tfrac{1}{2} \tag{6.5}$$

because the best possible breakdown point of a multivariate location estimator is 50%, which means that still 50% of "good" pseudo-observations are needed. Formula (6.5) yields an upper bound on the amount of contamination that is allowed in the original data, namely

$$\varepsilon^* = 1 - (\tfrac{1}{2})^{1/p}. \tag{6.6}$$

For $p = 2$, one finds again the breakdown point of Theil's estimator. The value of ε^* in (6.6) decreases very fast with p, as shown in Table 21.

M-estimators (Huber 1973) marked an important step forward in robust estimation. Much research has been concentrated on constructing functions ρ and ψ (see Chapter 1) such that the associated M-estimators were as robust as possible on the one hand, but still fairly efficient (in the case of a normal error distribution) on the other hand. Note that LS is also an M-estimator with $\psi(t) = t$ and that L_1 regression corresponds to $\psi(t) = \text{sgn}(t)$. Huber (1964) proposed the following ψ function:

$$\psi(t) = \begin{cases} t & \text{if } |t| < b \\ b \text{ sgn } (t) & \text{if } |t| \geq b \end{cases}, \tag{6.7}$$

where b is a constant. Actually, in a univariate location setting, this estimator was already constructed by the Dutch astronomer Van de Hulst in 1942 (see van Zwet 1985). Asymptotic properties of this estimator are discussed in Huber (1973). Hampel (1974) defined a function that protects the fit even more against strongly outlying observations, by means of

$$\psi(t) = \begin{cases} t & \text{if } |t| < a \\ a \text{ sgn } (t) & \text{if } a \leq |t| < b \\ \{(c - |t|)/(c - b)\} a \text{ sgn } (t) & \text{if } b \leq |t| \leq c \\ 0 & \text{otherwise} \end{cases}, \tag{6.8}$$

which is called a three-part redescending M-estimator. Figure 19 shows ψ-functions of both types. In the literature one can find many more ψ-functions (see Hampel et al. 1986 for a detailed description).

Of course, it is not sufficient to define new estimators and to study their asymptotic properties. Besides that, one has to develop a method for calculating the estimates. The solution of the system of equations (4.32) corresponding to M-estimates is usually performed by an iterative

Table 21. Value of $\varepsilon^* = 1 - (\frac{1}{2})^{1/p}$ for Some Values of p

p	ε^*	p	ε^*
1	50%	6	11%
2	29%	7	9%
3	21%	8	8%
4	16%	9	7%
5	13%	10	7%

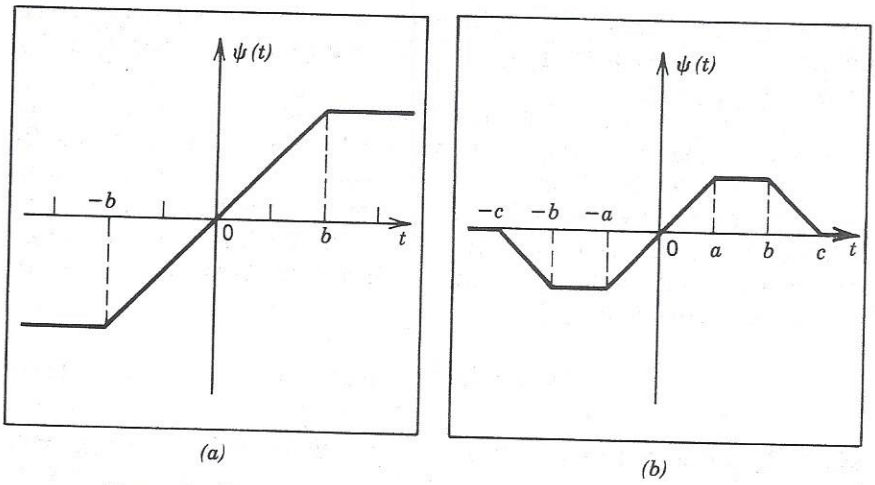

Figure 19. (*a*) Huber-type ψ-function. (*b*) Hampel-type ψ-function.

procedure. In each step, one has to estimate the coefficients and the scale simultaneously. However, it is very important to start the iteration with a "good" starting value, that is, an estimate which is already sufficiently robust. Without this precaution, one can easily end up in a local minimum that does not correspond at all to the expected robust solution. Dutter (1977) proposed some algorithms for solving the numerical problems associated with *M*-estimators. The calculation of GM-estimators or bounded-influence estimators [Chapter 1, equation (2.12)] presents similar problems. Dutter (1983a) described a user-oriented and portable computer program (BLINWDR) for calculating these estimates. Marazzi (1986) developed the subroutine library ROBETH, which computes *M*-estimators and bounded influence estimators. ROBETH also contains robust tests for linear models (Ronchetti 1982) and a variable selection procedure (see also Ronchetti 1985). The corresponding system control program is called ROBSYS (Marazzi 1987). TROLL (Samarov and Welsch 1982, Peters et al. 1982) is a large interactive system for statistical analysis which also includes the computation of bounded influence regression estimators.

Note that the breakdown point of *M*- and GM-estimators is quite different. Where ε^* is again 0% for *M*-estimators because of their vulnerability to leverage points, it becomes nonzero for GM-estimators (Maronna et al. 1979, Donoho and Huber 1983). For p tending to ∞, the breakdown point of GM-estimators drops to 0% (a small numerical study was performed by Kamber 1985). Yohai (1985) observes that some

GM-estimators have very low efficiency in the presence of good leverage points. Exact breakdown points of GM-estimators are computed by Martin et al. (1987).

Another approach to robust regression is based on the ranks of the residuals. In the framework of univariate location, these so-called *R-estimators* are due to Hodges and Lehmann (1963). The idea of using rank statistics has been extended to the domain of multiple regression by Adichie (1967), Jureckovà (1971), and Jaeckel (1972). The proposal of Jaeckel leads to the following definition: If R_i is the rank of $r_i = y_i - \mathbf{x}_i\boldsymbol{\theta}$, then the objective is to

$$\underset{\hat{\boldsymbol{\theta}}}{\text{Minimize}} \sum_{i=1}^{n} a_n(R_i) r_i \,, \tag{6.9}$$

where the scores function $a_n(i)$ is monotone and satisfies $\Sigma_{i=1}^{n} a_n(i) = 0$. Some possibilities for the scores $a_n(i)$ are:

Wilcoxon scores:	$a_n(i) = i - (n+1)/2$
Van der Waerden scores:	$a_n(i) = \Phi^{-1}(i/(n+1))$
median scores:	$a_n(i) = \operatorname{sgn}(i - (n+1)/2)$
bounded normal scores:	$a_n(i) = \min(c, \max\{\Phi^{-1}(i/(n+1)), -c\})$.

(The latter scores were proposed by Rousseeuw 1979 and Ronchetti 1979.) In the case of regression with intercept, one has to estimate the constant term separately, since the objective function is invariant with respect to the intercept. This can be done by using a robust location estimate of the residuals. An important advantage of *R*-estimators compared to *M*-estimators is that they are automatically scale equivariant, so they do not depend on a simultaneous scale estimator. Nevertheless, Jureckovà (1977) showed that (under certain conditions) *R*-estimators are asymptotically equivalent to *M*-estimators. Heiler and Willers (1979) prove the asymptotic normality of *R*-estimators under weaker conditions than those imposed by Jureckovà. Lecher (1980) developed (and implemented) an algorithm for *R*-estimators, in which the minimization (6.9) is carried out by a direct-search algorithm of Rosenbrock (1960). Cheng and Hettmansperger (1983) proposed an iteratively reweighted least squares algorithm for solving (6.9).

The class of *L-estimators* also plays a prominent role in robust univariate location. They are based on linear combinations of order statistics, and their popularity rests mainly on their simple computation. Bickel (1973) has proposed a class of one-step *L*-estimators for regression, which depend on a preliminary estimate of $\boldsymbol{\theta}$.

Koenker and Bassett (1978) formulated another proposal for L-estimators, making use of analogs of sample quantiles for linear regression. They defined the *α-regression quantile* $(0 < \alpha < 1)$ as the solution $\hat{\boldsymbol{\theta}}_\alpha$ of

$$\underset{\hat{\boldsymbol{\theta}}_\alpha}{\text{Minimize}} \sum_{i=1}^{n} \rho_\alpha(r_i) ,$$

where

$$\rho_\alpha(r_i) = \begin{cases} \alpha r_i & \text{if } r_i \geq 0 \\ (\alpha - 1) r_i & \text{if } r_i \leq 0 . \end{cases}$$

(For $\alpha = 0.5$, one obtains the L_1 estimator.) Koenker and Bassett then proposed to compute linear combinations of these $\hat{\boldsymbol{\theta}}_\alpha$. Portnoy (1983) proved some asymptotic properties of these estimators. However, note that their breakdown point is still zero.

Also the trimmed least squares estimators of Ruppert and Carroll (1980) are L-estimators. They are counterparts of trimmed means, which are well-known L-estimators of location. (Note that they are not related to the LTS discussed in Section 4.) Ruppert and Carroll proposed two ways to select observations to be trimmed: one of these uses the concept of regression quantiles, whereas the other employs residuals from a preliminary estimator.

Heiler (1981) and Kühlmeyer (1983) describe the results of a simulation study about M-, L-, and R-estimators for linear regression. The behavior of these classes of estimators for different designs and error distributions was compared. It is important to note that the generated samples were rather small, $n \leq 40$ and $p \leq 3$, and that no leverage points were constructed. The conclusions from this study can be summarized as follows: LS is very poor, even for mild deviations from normality. M-estimators with redescending ψ-function turn out to work quite well. R-estimates with Wilcoxon scores, which have the advantage of being scale equivariant (and of being simple to use because no parameter has to be fixed in advance), are a good alternative. L-estimates achieved less satisfactory results.

The first regression estimator with maximal breakdown point is the repeated median due to Siegel (1982). Like the proposal of Oja and Niinimaa (1984) discussed above, it is based on all subsets of p points. Any p observations $(\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_p}, y_{i_p})$ determine a unique parameter vector, the jth coordinate of which is denoted by $\theta_j(i_1, \ldots, i_p)$. The repeated median is then defined coordinatewise as

$$\hat{\theta}_j = \underset{i_1}{\text{med}} \left(\ldots \left(\underset{i_{p-1}}{\text{med}} \left(\underset{i_p}{\text{med}} \, \theta_j(i_1, \ldots, i_p) \right) \right) \ldots \right) . \qquad (6.10)$$

This estimator can be calculated explicitly. The fact that the medians are computed sequentially (instead of one median over all subsets) gives the estimator a 50% breakdown point, in which respect it is vastly superior to the Oja–Niinimaa proposal. (The asymptotic efficiency of the repeated median, as well as its influence function, appear to be unknown as yet.) Unfortunately, two disadvantages remain: the absence of affine equivariance and the large number of subsets. However, the second problem might be avoided by selecting some subsets at random instead of using all C_n^p of them.

Recently, Yohai (1985) introduced a new improvement toward higher efficiency for high-breakdown estimators like LMS and LTS. He called this new class *MM-estimators*. (Note that they are not related to the MM-estimators considered in Chapter 9 of Rey 1983.) Yohai's estimators are defined in three stages. In the first stage, a high-breakdown estimate $\boldsymbol{\theta}^*$ is calculated, such as LMS or LTS. For this purpose, the robust estimator does not need to be efficient. Then, an M-estimate of scale s_n with 50% breakdown is computed on the residuals $r_i(\boldsymbol{\theta}^*)$ from the robust fit. Finally, the MM-estimator $\hat{\boldsymbol{\theta}}$ is defined as any solution of

$$\sum_{i=1}^{n} \psi(r_i(\boldsymbol{\theta})/s_n)\mathbf{x}_i = \mathbf{0}\,,$$

which satisfies

$$S(\boldsymbol{\theta}) \leq S(\boldsymbol{\theta}^*)\,,$$

where

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} \rho(r_i(\boldsymbol{\theta})/s_n)\,.$$

The function ρ must be like those used in the construction of S-estimators in Section 4; in particular, it must satisfy conditions (S1) and (S2). This implies that $\psi = \rho'$ has to be properly redescending: Some possibilities are three-part redescenders (6.8), Tukey's biweight (4.17), or the hyperbolic tangent ψ (4.16). The trick is that this ρ may be quite different from that of the scale estimate s_n of the second stage, because the first and the second stage must achieve the high breakdown point whereas the third stage is allowed to aim for a high efficiency. Indeed, Yohai showed that MM-estimators inherit the 50% breakdown point of the first stage and that they also possess the exact fit property. Moreover, he proved that MM-estimators are highly efficient when the errors are normally distributed. In a small numerical study, he showed that they compare favorably with GM-estimators.

In the same spirit of combining high breakdown with high efficiency, Yohai and Zamar (1986) consider "τ-estimators" defined by

$$\underset{\hat{\theta}}{\text{Minimize}}\, s_n^2 \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{r_i}{s_n}\right),$$

where again ρ may be different from that of the scale estimate s_n, which is applied to the residuals $r_i(\boldsymbol{\theta})$. Asymptotically, a τ-estimator behaves like an M-estimator with a ρ-function that is a weighted average of the two ρ-functions used in this construction.

Another possibility, which we have not yet investigated, would be to minimize an objective function of the type

$$\left[\frac{1}{n} \sum_{i=1}^{n} r_i^2(\boldsymbol{\theta})\right] \wedge [k^2 S^2(\boldsymbol{\theta})], \tag{6.11}$$

where $\wedge$ denotes the minimum of two numbers, $k > 1$, and $S(\boldsymbol{\theta})$ is a high-breakdown estimator of scale on the residuals $r_1(\boldsymbol{\theta}), \ldots, r_n(\boldsymbol{\theta})$, which is consistent under normality. For instance, $S^2(\boldsymbol{\theta})$ may be a multiple of $\text{med}_i\, r_i^2(\boldsymbol{\theta})$ or $(1/n)\, \Sigma_{i=1}^{h} (r^2(\boldsymbol{\theta}))_{i:n}$, or $S(\boldsymbol{\theta})$ may be a suitable M-estimator of scale. It seems that the minimization (over $\hat{\boldsymbol{\theta}}$) of (6.11) would combine high asymptotic efficiency with a high breakdown point, because most often the first part (LS) would be used at "good" configurations whereas the second part protects from "bad" configurations. However, it remains to be verified whether the actual (finite-sample) behavior of this estimator would be good enough to compete with simple but effective methods like the combination of LMS with a one-step improvement.

REMARK. Suppose that we apply a weighted LS with weights given by

$$w_i = \begin{cases} 1 & \text{if } |r_i/\sigma^*| \leq c \\ 0 & \text{otherwise,} \end{cases}$$

where r_i is the LMS residual of y_i, and σ^* is the corresponding LMS scale estimate. For each constant $c \geq 1$, this estimator has breakdown point 50%, whereas for $c \to \infty$ it becomes more and more efficient, and tends to LS. This paradox can be explained by understanding that the breakdown point is only a crude qualitative notion. Indeed, the above estimator with *large* c will not become unbounded for less than 50% of contamination, but it will not be very good either. [The same is true for the univariate M-estimator of location (6.7) with large b.] One should not forget that the breakdown point is only one out of several robustness criteria, so a

high breakdown point alone is not a *sufficient* condition for a good method. We personally consider a good breakdown point as a *necessary* condition, because we do not want estimators that can become arbitrarily bad as a result of a small fraction of contamination. Indeed, Murphy's Law guarantees us that such contamination is bound to occur in practice.

EXERCISES AND PROBLEMS

Sections 1–3

1. Table 22 was taken from Gray (1985). It deals with 23 single-engine aircraft built over the years 1947–1979. The dependent variable is cost (in units of $100,000), and the explanatory variables are aspect

Table 22. Aircraft Data

Index	Aspect Ratio	Lift-to-Drag Ratio	Weight	Thrust	Cost
1	6.3	1.7	8,176	4,500	2.76
2	6.0	1.9	6,699	3,120	4.76
3	5.9	1.5	9,663	6,300	8.75
4	3.0	1.2	12,837	9,800	7.78
5	5.0	1.8	10,205	4,900	6.18
6	6.3	2.0	14,890	6,500	9.50
7	5.6	1.6	13,836	8,920	5.14
8	3.6	1.2	11,628	14,500	4.76
9	2.0	1.4	15,225	14,800	16.70
10	2.9	2.3	18,691	10,900	27.68
11	2.2	1.9	19,350	16,000	26.64
12	3.9	2.6	20,638	16,000	13.71
13	4.5	2.0	12,843	7,800	12.31
14	4.3	9.7	13,384	17,900	15.73
15	4.0	2.9	13,307	10,500	13.59
16	3.2	4.3	29,855	24,500	51.90
17	4.3	4.3	29,277	30,000	20.78
18	2.4	2.6	24,651	24,500	29.82
19	2.8	3.7	28,539	34,000	32.78
20	3.9	3.3	8,085	8,160	10.12
21	2.8	3.9	30,328	35,800	27.84
22	1.6	4.1	46,172	37,000	107.10
23	3.4	2.5	17,836	19,600	11.19

Source: Office of Naval Research.

ratio, lift-to-drag ratio, weight of the plane (in pounds), and maximal thrust. Run PROGRESS on these data. Do you find any outliers in the standardized observations? Does LS identify any regression outliers? How many outliers are identified by LMS and RLS, and of what type are they? Is there a good leverage point in the data?

2. Table 23 lists the delivery time data of Montgomery and Peck (1982, p. 116). We want to explain the time required to service a vending machine (y) by means of the number of products stocked (x_1) and the distance walked by the route driver (x_2). Run PROGRESS on these data. The standardized observations reveal two leverage points. Look at the LMS or RLS results to decide which of these is good and

Table 23. Delivery Time Data

Index (i)	Number of Products (x_1)	Distance (x_2)	Delivery Time (y)
1	7	560	16.68
2	3	220	11.50
3	3	340	12.03
4	4	80	14.88
5	6	150	13.75
6	7	330	18.11
7	2	110	8.00
8	7	210	17.83
9	30	1460	79.24
10	5	605	21.50
11	16	688	40.33
12	10	215	21.00
13	4	255	13.50
14	6	462	19.75
15	9	448	24.00
16	10	776	29.00
17	6	200	15.35
18	7	132	19.00
19	3	36	9.50
20	17	770	35.10
21	10	140	17.90
22	26	810	52.32
23	9	450	18.75
24	8	635	19.83
25	4	150	10.75

Source: Montgomery and Peck (1982).

which is bad. How does deleting the bad leverage point (as done by RLS) affect the significance of the regression coefficients?

3. Table 24 was taken from Prescott (1975), who investigated the effect of the concentration of inorganic phosphorus (x_1) and organic phosphorus (x_2) in the soil upon the phosphorus content (y) of the corn grown in this soil. Carry out a multiple regression analysis of y on x_1 and x_2 by means of PROGRESS. Are there extreme standardized observations? Which outliers are identified by LMS and RLS? In view of the fact that one of the explanatory variables has a very insignificant coefficient in both LS and RLS, it is recommended to run the analysis again without that variable. Indicate the previously discovered outliers in the scatterplot of this simple regression. Judging from the statistics and p-values in both models, do you think that switching to the smaller model is justified?
4. When fitting a multiplicative model

$$y_i = x_{i1}^{\theta_1} x_{i2}^{\theta_2} \ldots x_{ip}^{\theta_p},$$

Table 24. Phosphorus Content Data

Index (i)	Inorganic Phosphorus (x_1)	Organic Phosphorus (x_2)	Plant Phosphorus (y)
1	0.4	53	64
2	0.4	23	60
3	3.1	19	71
4	0.6	34	61
5	4.7	24	54
6	1.7	65	77
7	9.4	44	81
8	10.1	31	93
9	11.6	29	93
10	12.6	58	51
11	10.9	37	76
12	23.1	46	96
13	23.1	50	77
14	21.6	44	93
15	23.1	56	95
16	1.9	36	54
17	26.8	58	168
18	29.9	51	99

Source: Prescott (1975).

it is natural to logarithmize the variables. However, in economics it happens that some observations are zero (typically in one of the explanatory variables). It is then customary to put the transformed observation equal to a very negative value. Would you rather use LS or a robust regression method on these transformed data? Why?

5. (Research problem) Is it possible to develop collinearity diagnostics that are not so much affected by outliers?

Sections 4–6

6. Show that the repeated median estimator is regression and scale equivariant, and give a counterexample to show that it is not affine equivariant.
7. (Research problem) It would be interesting to know the asymptotic behavior of the repeated median regression estimator and to obtain its influence function.
8. Show that the variant of the LMS given by formula (4.9) has breakdown point $([(n-p)/2]+1)/n$.
9. Explain why the breakdown point of the least quantile of squares estimator (4.10) is approximately equal to α. Why doesn't the breakdown point become higher than 50% when $\alpha > \frac{1}{2}$?
10. The L_∞ fit is determined by the narrowest band covering *all* the data. Consider again some of the simple regression examples of the preceding chapters to illustrate that the L_∞ line is even less robust than LS.
11. Prove that the LWS estimator (4.27) is regression, scale, and affine equivariant, and show that its breakdown point equals the desired value.
12. (Research problem) What is the maximal asymptotic efficiency of an S-estimator defined by means of function ρ satisfying (S1), (S2), and (S3) with $K = E_\Phi[\rho]$?