L_1-Statistical Procedures and Related Topics
IMS Lecture Notes – Monograph Series (1997) Volume 31

Recent developments in PROGRESS

Peter J. Rousseeuw and Mia Hubert

University of Antwerp, Belgium

Abstract: The least median of squares (LMS) regression method is highly robust to outliers in the data. It can be computed by means of PROGRESS (from Program for RObust reGRESSion). After ten years we have developed a new version of PROGRESS, which also computes the least trimmed squares (LTS) method. We will discuss the various new features of PROGRESS, with emphasis on the algorithmic aspects.

Key words: Algorithm, breakdown value, least median of squares, least trimmed squares, robust regression.

AMS subject classification: 62F35, 62J05.

1 Introduction

At the time when the least median of squares (LMS) regression method was introduced (Rousseeuw, 1984), a program was needed to compute it in practice. The first algorithm described in that paper was just for computing the LMS line in simple regression, based on scanning over possible slopes while adjusting the intercept each time.

However, it was clear from the start that an algorithm for LMS multiple regression was required. The first version of PROGRESS (from Program for RObust reGRESSion) was implemented in 1983. The 1984 paper already contained an example analyzed with PROGRESS and listed the program's computation times on a CDC 750, one of the fastest mainframes of that day but outperformed by today's PC's. During the next years, when people began requesting the program, it was made more user-friendly with interactive input and self-explanatory output. The use of the program was explained in detail in (Rousseeuw and Leroy, 1987). Because that book contained many sample outputs we refrained from making any substantial modifications to PROGRESS, which remained essentially unchanged from

1985 until 1995.

During that decade there were quite a few suggestions for modifications and extensions. For instance, several people asked for the inclusion of the least trimmed squares (LTS) method which had been proposed together with LMS in (Rousseeuw, 1984) but which was not built in from the start because it needed (a little) more computation time, which became less relevant with the increasing speed of hardware. Another idea was to improve the accuracy by carrying out intercept adjustments more often, and we also wanted to allow the user to replace the 'median' in LMS by another quantile. Therefore we finally gave up on the principle of keeping the outputs identical to those in the 1987 book, and created the modernized 1996 version of PROGRESS described in the present paper.

First of all, PROGRESS now allows the user to choose between two robust estimators: the least quantile of squares (LQS) method which generalizes LMS, and the least trimmed squares (LTS) method. By definition, these methods depend on a quantile h/n. In order to help the user make an appropriate choice of h, the program provides a range of h-values for which LQS and LTS have a breakdown value between 25% and 50%. (This means that the method can resist that many contaminated observations.) Section 2 describes the LQS and LTS, and Section 3 obtains their breakdown value which depends on h.

Section 4 provides an outline of the algorithm used for the LQS and LTS. Since their objective functions are difficult to minimize exactly, PROGRESS performs an approximate resampling algorithm. Whereas the 1985 version adjusted the intercept only once at the end, the intercept in now adjusted in each step. This yields a lower objective function value, and in simple regression we even find the exact minimum. The program now allows to search over all subsets, as well as over a user-defined number of random subsets.

In Section 5 we define a new version of the robust coefficient of determination (R^2) to make sure that it always takes on values in the interval [0,1]. Finally, Section 6 discusses the robust diagnostics which the program provides to identify outliers and leverage points. Section 7 explains how the program can be obtained.

2 The estimators LQS and LTS

We consider the linear multiple regression model

$$y_i = x_{i1}\theta_1 + x_{i2}\theta_2 + \ldots + x_{ip}\theta_p + \sigma e_i = \mathbf{x}_i^t\theta + \sigma e_i \tag{1}$$

for $i = 1, \ldots, n$. The p-dimensional vectors $\mathbf{x}_i$ contain the explanatory variables, y_i is the response and σe_i is the error term. The data set thus consists of n observations and will be denoted by $Z = (X, \mathbf{y})$. For a given parameter estimate $\hat{\theta}$ we denote the residuals as $r_i(\hat{\theta}) = y_i - \mathbf{x}_i^t\hat{\theta}$. In a regression model with intercept, the observations satisfy $x_{ip} \equiv 1$.

A robust regression method tries to estimate the regression parameter vector θ in such a way that it fits the bulk of the data even when there are outliers.

The new version of PROGRESS provides two such robust regression methods: the least quantile of squares (LQS) and the least trimmed squares (LTS) estimator. Whereas classical least squares (LS) minimizes the sum of the squared residuals, LQS and LTS minimize a certain quantile, resp. a trimmed sum, of the squared residuals. Their exact definition is given below. (For any numbers $u_1, \ldots, u_n$ the notation $u_{i:n}$ stands for the i-th order statistic.)

Definition 1 *Let $Z = (X, \mathbf{y})$ be a data set of n observations in $I\!R^{p+1}$. Then for all $p \leq h \leq n$, the least quantile of squares (LQS) estimate $\hat{\theta}_{LQS}(Z)$ and the least trimmed squares (LTS) estimate $\hat{\theta}_{LTS}(Z)$ are defined by*

$$\hat{\theta}_{LQS}(Z) = \underset{\theta}{argmin}(r^2(\theta))_{h:n} = \underset{\theta}{argmin}\,|r(\theta)|_{h:n} \tag{2}$$

and

$$\hat{\theta}_{LTS}(Z) = \underset{\theta}{argmin}\sum_{i=1}^{h}(r^2(\theta))_{i:n}. \tag{3}$$

It is easy to see that LQS generalizes the LMS method which minimizes the median of the squared residuals. Indeed, for n odd and $h = [n/2] + 1$ the LQS becomes the LMS.

With the parameter estimates $\hat{\theta}_{LQS}(Z)$ and $\hat{\theta}_{LTS}(Z)$ we can associate estimators of the error scale σ:

$$s_{LQS}(Z) = s_{LQS}(X, \mathbf{y}) = c_{h,n}|r(\hat{\theta}_{LQS}(Z))|_{h:n} \tag{4}$$

and

$$s_{LTS}(Z) = s_{LTS}(X, \mathbf{y}) = d_{h,n}\sqrt{\frac{1}{h}\sum_{i=1}^{h}(r^2(\hat{\theta}_{LTS}(Z)))_{i:n}}. \tag{5}$$

The constants $c_{h,n}$ and $d_{h,n}$ are chosen to make the scale estimators consistent at the gaussian model, which gives

$$\begin{aligned} c_{h,n} &= 1/\Phi^{-1}(\frac{h+n}{2n}) \\ d_{h,n} &= 1/\sqrt{1 - \frac{2n}{hc_{h,n}}\phi(1/c_{h,n})}. \end{aligned}$$

Moreover, s_{LQS} is multiplied by a finite-sample correction factor. For $h = [n/2]+1$ this factor equals $1+\frac{5}{n-p}$.

More efficient scale estimates, based on the preliminary ones, are then given by

$$\hat{\sigma} = \sqrt{\frac{\sum_i w_i r_i^2}{\sum_i w_i - p}}, \tag{6}$$

where

$$w_i = \begin{cases} 0 & \text{if } \left|r_i/s_{LQ(T)S}\right| > 2.5 \\ 1 & \text{otherwise.} \end{cases}$$

Here the notation $s_{LQ(T)S}$ stands for s_{LQS} or s_{LTS}, whichever is used.

3 Breakdown value and choice of h

In the next theorem we derive the breakdown value of LQS and LTS, which says how many of the n observations need to be replaced before the estimate is carried away. The finite-sample breakdown value (Donoho and Huber, 1983) of any regression estimator $T(Z) = T(X, \mathbf{y})$ is given by

$$\varepsilon_n^* = \varepsilon_n^*(T, Z) = \min \{\frac{m}{n}; \sup_{Z'} \|T(Z')\| = \infty\}$$

where $Z' = (X', \mathbf{y}')$ ranges over all data sets obtained by replacing any m observations of $Z = (X, \mathbf{y})$ by arbitrary points. We will assume that the original X is in *general position.* This means that no p of the $\mathbf{x}_i$ lie on a $p-1$ dimensional plane through the origin. For simple regression with intercept ($p = 2$) this says that no two x_i coincide. For simple regression without intercept ($p = 1$) it says that none of the x_i are zero.

Theorem 1 *If the $\mathbf{x}_i$ are in general position, then the finite-sample breakdown value of the LQS and the LTS is*

$$\varepsilon_n^* = \begin{cases} (h-p+1)/n & \text{if } p \le h < [\frac{n+p+1}{2}] \\ (n-h+1)/n & \text{if } [\frac{n+p+1}{2}] \le h \le n. \end{cases}$$

The proof is given in the Appendix.

Corollary 1 *If the $\mathbf{x}_i$ are in general position, then the maximal finite-sample breakdown value of the LQS and the LTS equals*

$$\max_h \varepsilon_n^* = \frac{[(n-p)/2]+1}{n}$$

and is achieved for

$$[(n+p)/2] \le h \le [(n+p+1)/2].$$

When $n+p$ is even we have $[(n+p)/2] = [(n+p+1)/2]$ hence the optimal h is unique. When $n+p$ is odd, it turns out that choosing $h = [(n+p+1)/2]$ gives the better finite-sample efficiency. Therefore, we will always define the optimal h as

$$h_{opt} = [(n+p+1)/2].$$

This is also the default value of h in PROGRESS. If the user prefers to use another quantile, the program displays a range of h-values for which a breakdown value of at least 25% is attained. The lowest h-value allowed in the program is

$$h_{min} = [n/2] + 1.$$

(This is because for each $\tilde{h} < h_{min}$ there exists some $h \geq h_{min}$ with the same breakdown value and a higher finite-sample efficiency.)

Remark 1 *If $p = 1$ and $x_{ip} = 1$ for all observations, the regression model reduces to the univariate model $y_i = \mu + \sigma e_i$. In that case Theorem 1 is still valid, whereas the LQS and LTS become much easier to compute. In the univariate setting, a fast algorithm is available to compute the exact LQS and LTS estimates of the location parameter μ and the scale parameter σ.*

1. input the data and all options
2. treatment of missing data
3. standardize the data
4. LS analysis
5. compute LQS or LTS
6. RLS analysis

Table 1: Overview of the program PROGRESS.

4 Outline of the PROGRESS algorithm

PROGRESS not only computes the LQS and LTS. First, the LS estimates and inferences about the regression parameters are obtained. And after the LQS or LTS is found, a reweighted least squares (RLS) is carried out with weights based on LQS or LTS. Table 1 gives a schematic overview of the complete program. Since the essential algorithmic changes have been made in step 5, we will focus on that part here. We refer to (Rousseeuw and Leroy, 1987) for all details about the treatment of missing data, the standardization procedure, and the LS and RLS estimates.

	ACTION	RESULT
1	draw a (random) subset of p observations	$\{z_{i_1}, \ldots, z_{i_p}\}$
2	compute hyperplane through these p observations	$\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{p-1}, \tilde{\theta}_p)$
3	if regression with intercept $\Rightarrow$ adjust intercept	$\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{p-1}, \tilde{\theta}'_p)$
4	evaluate the objective function at this estimate	$\|r(\tilde{\theta})\|_{h:n}$ or $\sum_1^h r^2(\tilde{\theta})_{i:n}$
5	repeat steps 1 until 4, and keep the estimate with lowest objective function value	$\hat{\theta}_{LQS}$ or $\hat{\theta}_{LTS}$

Table 2: Summary of the algorithm for LQS and LTS.

In general the objective functions of LQS and LTS are difficult to minimize exactly since they have several local minima. For this reason PROGRESS uses an approximate resampling algorithm (which does yield the exact solution in simple regression). Table 2 summarizes the main steps of this algorithm.

n	mechanism	number of p-subsets used
small	all subsets	C_n^p
intermediate	all subsets random	C_n^p default (Table 4) or user-defined
large	random	default or user-defined

Table 3: Subsampling mechanism in PROGRESS.

We will describe the first three steps more extensively.

1. draw a (random) subset of p observations

The drawing mechanism now implemented in PROGRESS is displayed in Table 3. According to the sample size n and the number of variables p, PROGRESS checks whether or not it is feasible to draw all subsets of p observations out of n.

For **small** values of n (see Table 4) the program automatically generates all possible subsets of p observations, of which there are $C_n^p = \binom{n}{p}$. For each of these p-subsets, steps 1 to 4 of Table 2 are carried out.

If n is **large** for the p involved, the binomial coefficient would exceed 1,000,000 and then PROGRESS switches to a random selection of p-subsets. It is possible for the user to preset the number of p-subsets to be considered. The more p-subsets you take, the lower the objective function will be, but at the cost of more computation time. On the other hand, one must select enough p-subsets for the probability of drawing at least one uncontaminated p-subset to be close to 1 (otherwise, the fit could be based on bad observations only). In (Rousseeuw and Leroy, 1987, page 198) this minimal number of p-subsets is expressed in function of the number of variables and the allowed percentage of contamination. The default number of subsets drawn in PROGRESS can be found in Table 4. For $p \leq 9$ these numbers exceed the required minimum, whereas for larger p the default is fixed at 3000 subsets so as to avoid extremely long calculations. But as already mentioned, the user can always modify the proposed number of p-subsets.

Finally, for all **intermediate** values of n the user can choose between considering all p-subsets or drawing a certain number of random p-subsets. As always, the program applies default choices unless the user explicitly asks to override them.

	p									
	1	2	3	4	5	6	7	8	9	10
n is 'small' if n $\leq$	500	50	22	17	15	14	0	0	0	0
n is 'large' if n $\geq$	10^6	1414	182	71	43	32	27	24	23	22
default number of p-subsets used	500	1000	1500	2000	2500	3000	3000	3000	3000	3000

Table 4: Sample sizes n which are considered to be small or large (for a given p). Also the default number of p-subsets used in PROGRESS is listed.

2. compute hyperplane through these p observations

If the $\mathbf{x}_i$ are in *general position* then every p-subset determines a unique hyperplane, that is found by solving the linear system formed by these p observations.

In practice also a singular p-subset can occur, and then PROGRESS draws a new p-subset. The output then reports the total number of singular p-subsets that were encountered.

3. if regression with intercept $\Rightarrow$ adjust intercept

Here, 'intercept adjustment' stands for a technique which decreases the objective value of a given fit. We will apply it to each $p-$subset. After the hyperplane through the p observations is determined, we have an initial estimate of the slope and the intercept, given by $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{p-1}, \tilde{\theta}_p)$ where $\tilde{\theta}_p$ is the intercept. The corresponding objective value for LQS then equals

$$|r(\tilde{\theta})|_{h:n} = |y_j - x_{j1}\tilde{\theta}_1 - \ldots - x_{j,p-1}\tilde{\theta}_{p-1} - \tilde{\theta}_p|_{h:n}. \tag{7}$$

For LTS we can rewrite (3) accordingly. The adjusted intercept $\tilde{\theta}'_p$ is then defined as the LQS (resp. LTS) location estimate applied to the univariate data set $\{t_i = y_i - x_{i1}\tilde{\theta}_1 - \ldots - x_{i,p-1}\tilde{\theta}_{p-1}; i = 1, \ldots, n\}$, i.e.

$$\tilde{\theta}'_p = \underset{\mu}{\operatorname{argmin}} |t_j - \mu|_{h:n} \tag{8}$$

for LQS. By construction, (8) yields a lower objective value than (7). In simple regression ($p = 2$), it follows from (Steele and Steiger, 1986) that if all 2-subsets are used and their intercept is adjusted each time, we obtain the exact LQS.

As indicated in Remark 1, the LQS and LTS location estimates can be found by an explicit algorithm. For LQS it is the midpoint of the shortest interval that contains h observations, as was proved in (Rousseeuw, 1984, page 873). We thus have to order the univariate observations $\{t_1, \ldots, t_n\}$ to $t_{1:n} \leq \ldots \leq t_{n:n}$ and then compute the length of the contiguous intervals that contain h points. When the smallest length is attained by several intervals, we take the median of the corresponding midpoints.

The univariate LTS estimator corresponds to the mean of the subset that contains h observations and that has the smallest sum of squares. This sum of squares is defined as the sum of the squared deviations from the subset mean: given an h-subset $t_{i:n}, \ldots, t_{i+h-1:n}$ with mean $\bar{t}^{(i)}$ we have

$$SQ^{(i)} = \sum_{j=i}^{i+h-1} (t_{j:n} - \bar{t}^{(i)})^2.$$

Note that the selected h-subset has to consist of successive observations, which is why we had to order the $t_1, \ldots, t_n$ first.

For a recent study of the effect of intercept adjustment on the performance of LTS regression, see Croux et al. (1996).

In order to adjust the intercepts the univariate LQS and LTS methods were included into PROGRESS, which also allows the user to analyze data sets that were univariate from the start. As in the regression situation, the preliminary scale is then defined by (4) resp. (5), both of which come out of the univariate algorithms. For LQS it is half the length of the shortest interval, whereas for LTS it is the square root of the smallest sum of squares divided by h. We then obtain the final scale estimate as in (6).

5 Coefficient of determination (R^2)

Let us first consider the regression model *with* intercept. Along with the classical least squares (LS) comes the coefficient of determination, which measures the proportion of the variance of the response variable explained by the linear model, i.e.

$$0 \leq R^2 = \frac{Var(y_i) - Var(r_i)}{Var(y_i)} = 1 - \frac{Var(r_i)}{Var(y_i)} \leq 1. \tag{9}$$

The denominator in this expression measures the variability of the response in a model without explanatory variables, which in this case is the univariate model $y_i = \mu + \sigma e_i$. If we denote the LS coefficient estimate of a sample $(X, \mathbf{y})$ by $\hat{\theta}_{LS}(X, \mathbf{y})$, and use the scale estimate given by

$$s^2_{LS}(X, \mathbf{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \hat{\theta}_{LS}(X, \mathbf{y}))^2, \tag{10}$$

we can rewrite (9) as

$$R^2_{LS} = 1 - \frac{s^2_{LS}(X, \mathbf{y})}{s^2_{LS}(\mathbf{1}, \mathbf{y})}. \tag{11}$$

By analogy, we propose the robust counterpart given by

$$R^2_{LQ(T)S} := 1 - \frac{s^2_{LQ(T)S}(X, \mathbf{y})}{s^2_{LQ(T)S}(\mathbf{1}, \mathbf{y})}. \tag{12}$$

Note that when using definition (12), the robust coefficient of determination always falls in the interval $[0, 1]$. This was not guaranteed by the earlier version of R^2 defined in (Rousseeuw and Leroy, 1987, page 44) and implemented in the first version of PROGRESS. There the denominator $(\text{mad}\mathbf{y})^2 = (\text{med}_i |y_i - \text{med}_j y_j|)^2$ was used, whereas now $s_{LQS}(\mathbf{1}, \mathbf{y}) = |y - \hat{\theta}_{LQS}(\mathbf{1}, \mathbf{y})|_{h:n}$ which is just the scale estimate of the univariate LQS applied to the response.

An analogous reasoning works for the regression model without intercept. In that case, the model without explanatory variables reduces to $y_i = \sigma e_i$ without any location parameter. For LS,

$$R^2_{LS} = 1 - \frac{s^2_{LS}(X, \mathbf{y})}{s^2_{LS}(\mathbf{0}, \mathbf{y})} = 1 - \frac{\sum_i (y_i - \mathbf{x}_i \hat{\theta}_{LS})^2}{\sum_i y_i^2} \tag{13}$$

and we propose the following robust counterparts:

$$R^2_{LQS} = 1 - \frac{s^2_{LQS}(X, \mathbf{y})}{s^2_{LQS}(\mathbf{0}, \mathbf{y})} = 1 - \frac{((y - \mathbf{x}\hat{\theta}_{LQS})^2)_{h:n}}{(y^2)_{h:n}}$$

and

$$R^2_{LTS} = 1 - \frac{s^2_{LTS}(X, \mathbf{y})}{s^2_{LTS}(\mathbf{0}, \mathbf{y})} = 1 - \frac{\sum_{i=1}^{h} ((y - \mathbf{x}\hat{\theta}_{LTS})^2)_{i:n}}{\sum_{i=1}^{h} (y^2)_{i:n}}.$$

6 Diagnostics

Observations in regression data essentially belong to four types:

regular observations with internal $\mathbf{x}_i$ and well-fitting y_i,
vertical outliers with internal $\mathbf{x}_i$ and non-fitting y_i,
good leverage points with outlying $\mathbf{x}_i$ and well-fitting y_i,
bad leverage points with outlying $\mathbf{x}_i$ and non-fitting y_i.

Figure 1 shows these four types in simple regression. Regression diagnostics aim to detect observations of one or more of these types. Here we will consider three robust diagnostics: standardized residuals, the resistant diagnostic, and the diagnostic plot.

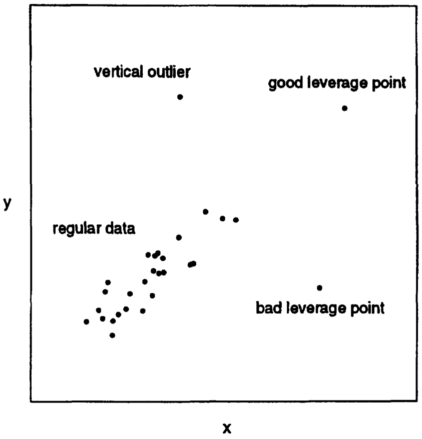

Figure 1: Simple regression data with points of all four types.

1. **Standardized residuals** are defined as $r_i(\theta)/s(\theta)$ where $s(\theta)$ denotes a robust scale estimate based on the residuals. Here we will use $s_{LQS}(\theta) = c_{h,n}|r(\theta)|_{h:n}$ or $s_{LTS}(\theta) = d_{h,n}\sqrt{\frac{1}{h}\sum_1^h r^2(\theta)_{i:n}}$. Standardized residuals help us to distinguish between well-fitting and non-fitting observations by comparing their absolute values to some yardstick, e.g.

$$\text{compare } |r_i(\theta)|/s(\theta) \text{ to } 2.5.$$

We use the yardstick 2.5 since it would determine a (roughly) 99% tolerance interval for the e_i if they had a standard gaussian distribution. Since the standardized residuals approximate the e_i, we will consider an observation as non-fitting if its standardized residual lies (far) outside this tolerance region.

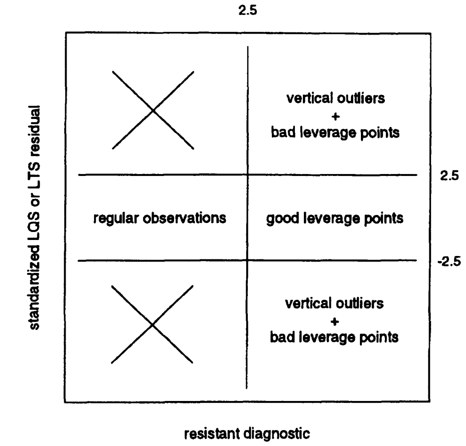

Figure 2: Classification of observations by plotting the standardized residuals versus their resistant diagnostics.

2. The resistant diagnostic. Non-regular observations have the property that they are 'far away' from some hyperplane in $I\!R^{p+1}$ (that is, further away than the *majority* of the observations). The vertical outliers and the bad leverage points are clearly far away from the ideal regression plane given by $y = \mathbf{x}^t\theta$. But also a good leverage point lies far away, relative to some other hyperplane that goes through the center of the regular observations. To define the 'distance' of an observation $(\mathbf{x}_i, y_i)$ to a plane $y = \mathbf{x}^t\theta$ we can use its absolute standardized residual. If we now define

$$U_i = \sup_\theta \frac{|r_i(\theta)|}{s_{LQS}(\theta)} \qquad \text{or} \qquad U_i = \sup_\theta \frac{|r_i(\theta)|}{s_{LTS}(\theta)}$$

we expect outliers to have a large U_i. Since the U_i are difficult to compute exactly, we approximate them by taking the maximum over all $\tilde{\theta}$ that are computed inside the LQS/LTS algorithm. For each observation, this yields the value

$$u_i = \max_{\tilde{\theta}} \frac{|r_i(\tilde{\theta})|}{s_{LQS}(\tilde{\theta})} \qquad \text{or} \qquad u_i = \max_{\tilde{\theta}} \frac{|r_i(\tilde{\theta})|}{s_{LTS}(\tilde{\theta})}. \tag{14}$$

Therefore we only need to store one array $(u_1, \ldots, u_n)$ that has to be updated at each p-subset. Finally, we define the resistant diagnostic for each observation by standardizing the u_i, yielding

$$\text{resistant diagnostic}_i = \frac{u_i}{\underset{j=1,\ldots,n}{\text{med}} u_j}. \tag{15}$$

In the new version of PROGRESS the resistant diagnostic is available for both LQS and LTS, and it is based on the trial estimates $\tilde{\theta}$ after location adjustment. From (14) it is clear that non-regular observations will have a large v_i and consequently a large resistant diagnostic. Simulations have indicated that we may consider (15) as 'large' if it exceeds 2.5. Combining the standardized residuals with the resistant diagnostic leads to the diagram in Figure 2. However, a disadvantage of Figure 2 is that it cannot distinguish between vertical outliers and bad leverage points.

3. **The diagnostic plot** makes the complete classification into the four types. Since leverage points are outlying in the space of the regressors $\mathbf{x}_i$, one can distinguish them from vertical outliers by analyzing their $\mathbf{x}$–components. For this we can run MINVOL on $X = \{\mathbf{x}_i; 1 \leq i \leq n\}$. This program computes the Minimum Volume Ellipsoid (MVE) location estimate $T(X)$ and scatter matrix $C(X)$. The MVE is a highly robust estimator of location and scatter, introduced by Rousseeuw (1985). The corresponding *robust distance* of an observation to the center is then given by

$$RD(\mathbf{x}_i) := \sqrt{(\mathbf{x}_i - T(X))^t C(X)^{-1} (\mathbf{x}_i - T(X))}.$$

Since the squares of these distances roughly have a chi-squared distribution when there are no outliers among the $\mathbf{x}_i$, we will classify an observation as a leverage point if its $RD(\mathbf{x}_i)$ exceeds the cutoff value $\sqrt{\chi^2_{p,0.975}}$. If we combine this information with the standardized LQS or LTS residual, we obtain the diagnostic plot of (Rousseeuw and van Zomeren, 1990) shown in Figure 3.

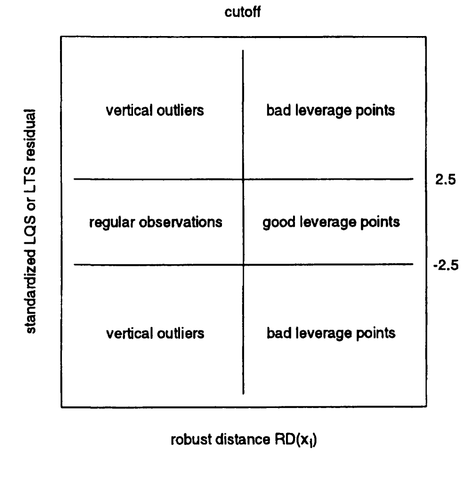

Figure 3: Diagnostic plot, obtained by plotting the standardized robust residuals versus the robust distances $RD(\mathbf{x}_i)$.

7 Software availability

The programs PROGRESS and MINVOL can be obtained from our website `http://win-www.uia.ac.be/u/statis`
Questions or remarks about the implementation can be directed to `Mia.Hubert@uia.ua.ac.be`.
The LMS, LTS and MVE methods are also available in S-PLUS as the functions `lmsreg`, `ltsreg` and `cov.mve`. Moreover, the functions LMS, LTS and MVE based on the recent versions of PROGRESS and MINVOL have been incorporated into SAS/IML (Version 6.12) in 1996. Their documentation can be obtained by writing to `sasaxs@unx.sas.com`.

Appendix: Proof of Theorem 1

We show that the proof of the breakdown value of the LMS (Rousseeuw, 1984, page 878) remains valid after making the necessary modifications. The proofs for LQS and LTS are very similar, so we will mainly consider the LQS estimator.

First suppose $h < [\frac{n+p+1}{2}]$. We obtain the lower bound on ε_n^* by replacing $h - p + 1 - 1 = h - p$ observations of Z, yielding Z'. Define $\theta = \hat{\theta}_{LQS}(Z)$ and $\theta' = \hat{\theta}_{LQS}(Z')$. The $n - h + p \geq h$ original points $(\mathbf{x}_i, y_i)$ then satisfy $|r_i(\theta)| = |y_i - \mathbf{x}_i\theta| \leq M = \max_Z |r_i(\theta)|$ such that for the corrupted data set Z', $|r_i(\theta')|_{h:n} \leq |r_i(\theta)|_{h:n} \leq M$. For $p > 1$ we refer to the geometrical construction of (Rousseeuw, 1984). In his notation, the set $Z' \setminus A$ contains at most $n - (n - h + p - (p-1)) = h - 1$ observations. If we assume $\|\theta' - \theta\| \geq 2(\|\theta\| + M/\rho)$, this implies

$$|r_i(\theta')|_{h:n} > M, \tag{16}$$

a contradiction. Therefore, $\|\theta'\|$ remains bounded. For $p = 1$ we set $C = 2M/N = 2M/min_Z|x_i|$. Now suppose $|\theta - \theta'| > C$. For all non-contaminated observations we have that $|r_i(\theta) - r_i(\theta')| = |y_i - x_i\theta - y_i - x_i\theta'| = |x_i||\theta - \theta'| > NC = 2M$, from which we get $|r_i(\theta')| \geq |r_i(\theta) - r_i(\theta')| - |r_i(\theta)| > 2M - M = M$. Again this implies (16) and thus a bounded θ'. The upper bound on ε_n^* follows from the fact that we can put $h - p + 1$ bad observations on a hyperplane that contains $p - 1$ original points. Then $(h - p + 1) + (p - 1) = h$ observations satisfy $y_i' = \mathbf{x}_i'\theta'$, and thus $\theta' = \hat{\theta}_{LQS}(Z')$. Making the hyperplane steeper will break down the estimator.

For $h \geq [\frac{n+p+1}{2}]$, we obtain the lower bound analogously to the previous case. Just observe that we now have $n - (n - h + 1) + 1 = h$ original observations, and that $Z' \setminus A$ has at most $n - (h - (p-1)) = n - h + p - 1 \leq h - 1$ points. The remaining inequality $\varepsilon_n^* \leq (n - h + 1)/n$ can be proved as follows. Take some $M > \|\theta\|$. Then we show that we can always construct a corrupted sample Z' with $n-h+1$ bad observations, such that $\|\theta'\| = \|\hat{\theta}_{LQS}(Z')\| \geq M$. Letting M go to infinity will then cause the LQS to break down. Define $M_X = \max_i \|\mathbf{x}_i\|$. Now we set all the $n-h+1$ replaced observations equal to the point $(\mathbf{x}, y) = (\mathbf{x}, 2M_X M + K)$ for which $\|\mathbf{x}\| = M_X$ and $K > 0$. These replaced observations satisfy $|\mathbf{x}_i\theta| \leq \|\mathbf{x}_i\|\|\theta\| < \|\mathbf{x}\|M = M_X M < y$ and thus $|r_i(\theta)| = |y_i - \mathbf{x}_i\theta| \geq |y| - |\mathbf{x}\theta| > M_X M + K$. As $n-h+1 > n-h$ this yields $|r_i(\theta)|_{h:n} > M_X M + K$. Since we can choose K arbitrarily large, the minimum of the objective function of LQS will not be reached for $\|\theta\| < M$. Consequently $\|\theta'\| = \|\hat{\theta}_{LQS}(Z')\|$ has to be larger than M, which ends the proof. Finally we note that using the same construction, the objective function of LTS satisfies

$$\sum_{i=1}^{h}(r^2(\theta))_{i:n} > (M_X M + K)^2$$

yielding the same result.

References

[1] Croux, C., Haesbroeck, G., and Rousseeuw, P.J. (1996). Location adjustment for the minimum volume ellipsoid estimator. Submitted for publication.

[2] Donoho, D.L., and Huber, P.J. (1983). The notion of breakdown point. In *A Festschrift for Erich Lehmann*, Eds. P. Bickel, K. Doksum, and J.L. Hodges. California: Wadsworth.

[3] Rousseeuw, P.J. (1984). Least median of squares regression. *J. Am. Statist. Assoc.* **79** 871-880.

[4] Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, Eds. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Vol. B, pp. 283-297. Dordrecht: Reidel.

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

[6] Rousseeuw, P.J., and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *J. Am. Statist. Assoc.* **85** 633-639.

[7] Steele, J.M., and Steiger, W.L. (1986). Algorithms and complexity for least median of squares regression. *Discrete Applied Mathematics* **14** 93-100.