

A MULTIPLE REGRESSION TECHNIQUE FOR DETECTING OUTLIERS

A.Leroy * and P.Rousseeuw **

ABSTRACT

The ordinary least squares regression method is not a reliable tool in regression analysis without first diagnosing possible outliers present in the data set. The least median of squares regression technique (Rousseeuw 1984), which is designed to lessen the impact of outlying observations, is presented and some alternatives are given. The output of a Fortran implementation of this regression technique, called PROGRES (Leroy and Rousseeuw 1984), is illustrated with an example. The results can be interpreted by means of a graphical representation of the standardized residuals. It is showed how PROGRES can be used as a diagnostic tool in regression analysis. Furthermore, conclusions are drawn from a small simulation study which compares some robust and non-robust regression estimators in different design situations.

1. INTRODUCTION

In a linear model, an output variable y is written as a linear combination of p input variables x_1, \dots, x_p

$$y_i = x_{i1}\theta_1 + \dots + x_{ip}\theta_p + e_i, \quad i=1, \dots, n$$

* Vrije Universiteit Brussel, Centrum voor Statistiek & Operationeel Onderzoek, Pleinlaan 2, B-1050 Brussels, Belgium. (Tel: 02/ 641 20 48)

** Technische Hogeschool Delft, Onderafdeling Wiskunde & Informatica, Julianalaan 132, 2628 BL Delft, The Netherlands. (Tel: 015/ 78 25 47)

where e_i is often assumed to be normally distributed with mean zero and standard deviation σ . Until recently, most people have been estimating the coefficients $\theta_1, \dots, \theta_p$ almost exclusively by means of the least squares (LS) method, defined by

$$\underset{\hat{\theta}}{\text{minimize}} \quad \sum_{i=1}^n r_i^2 \quad (1.1)$$

where the residuals r_i equal $y_i - \hat{x}_{i1}\hat{\theta}_1 - \dots - \hat{x}_{ip}\hat{\theta}_p$.

The main advantage of this method lies in the fact that explicit formulas exist for the estimates, making it the only feasible method in the pre-computer age. For the same reason, nowadays many computer programs for LS are available, which explains why this method has been used so often. Moreover, many mathematicians adore the LS estimator because of its nice optimality properties under the condition of a normal error structure. In practical situations however, this condition is hardly fulfilled, and the LS regression technique is quite sensitive to the presence of outlying points. Therefore, it is important to have a diagnostic tool for identifying such points. In the last decades, several statisticians have given consideration to robust regression (see Rousseeuw 1984 for an overview) on the one hand and to regression diagnostics on the other hand. Both approaches are closely related by two important common aims, namely identifying outliers and pointing out inadequacies of the model. The books of Belsley, Kuh and Welsch (1980) and Cook and Weisberg (1982) are dedicated to regression diagnostics. However, most of these methods deal with the effects of deleting a single point, and often do not succeed in identifying multiple outliers. On the other hand, the robust regression technique described in this paper does manage to solve this problem. When the robust and the LS fit differ substantially, this indicates that the data require a thoughtful analysis.

In order to express in a statistical way the robustness of a regression technique against outlying observations, Hampel (1971, 1975) proposed a general asymptotic definition of the breakdown point ϵ^* . We will use the finite sample version

of this notion given by Donoho and Huber (1982), namely

$$\epsilon^*(X, T) = \min \{m/n ; \sup \|T(X')\| = \infty\}$$

where the X' are obtained by replacing m points of the sample X (containing n data points) by arbitrary ones. T stands for a regression estimator. In words, ϵ^* is the smallest fraction of contamination that can cause the estimates to take on arbitrarily large values. For LS regression ϵ^* equals $1/n$ because one bad point is sufficient to carry the LS estimator over all bounds. Considering the limit for n going to infinity (p fixed), one can establish that LS has ϵ^* equal to 0%. The best possible value for the breakdown point is 50%, because for larger amounts of contaminated data in a sample, one cannot tell the 'good' and the 'bad' observations apart. The first regression estimator which is equivariant for linear transformations on the x_i and which attains a breakdown point of 50% is the least median of squares (LMS) estimator (Rousseeuw 1984).

The LMS estimate of θ corresponds to

$$\underset{\hat{\theta}}{\text{minimize}} \quad \underset{i}{\text{median}} \quad r_i^2 \quad (1.2)$$

Compared to LS (1.1), the sum has been replaced by the median. Preceding improvements towards robustness consisted of substituting the square by something else, but none of these led to a high breakdown point.

In the following two sections we will outline the algorithm we use for computing the LMS estimator as well as some other robust regression estimates derived from it. Section 4 is devoted to an example. In section 5, the results of a simulation study show the finite sample performance of some robust and non-robust regression techniques in three different situations.

2. ALGORITHM FOR COMPUTING THE LMS ESTIMATES

The special case of one-dimensional estimation of location is obtained by putting $p=1$ and $x_i=1$ for all i in (1.2). Then the minimization becomes

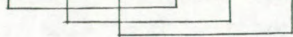
$$\underset{\hat{\theta}}{\text{minimize}} \quad \underset{i}{\text{median}} \quad (y_i - \theta)^2 \quad (2.1)$$

and the sample reduces to $(y_i)_{i=1, \dots, n}$.

The LMS estimate is then equal to the midpoint of the shortest half of the sample $(y_i)_{i=1, \dots, n}$. The shortest half is given by the smallest of the differences $y_{(h)} - y_{(1)}$, $y_{(h+1)} - y_{(2)}$, \dots , $y_{(n)} - y_{(n-h+1)}$, where $h = \lceil n/2 \rceil + 1$ ($\lceil x \rceil$ means integer part of x), and $y_{(1)} < \dots < y_{(n)}$ are the ordered observations.

The following simple example will illustrate the LMS estimate. Consider the one-dimensional sample consisting of the observations:

21, 23, 25, 26, 26, 299.



The halves of this sample are indicated by the lines below the values. The LMS estimate of location is 24.5, because it is the midpoint of the shortest half. The least squares estimate of location is the mean, which equals 70 in this sample. Comparing both estimates, it appears that 24.5 is a better parameter of location for the majority of the data. The aberrant value 299 has badly affected the mean, whereas the LMS has completely neglected its presence.

In the general regression model, it is probably not possible to write down a straightforward formula for the LMS estimate. For this case we have therefore constructed a heuristic algorithm which can be outlined in the following way: Choose at random p observations out of the n and determine the unique regression surface through these p points. This solution gives a trial estimate $(\theta_1^0, \dots, \theta_p^0)$. This procedure is repeated m times and the trial estimate for which the objective function is minimal is retained. The number of replications (m) is determined by requiring that the probability that at least one of the m subsamples is 'good' is at least 95%. A subsample is 'good' if it consists of p good observations of the sample, which may contain (in the most extreme case) up to 50% of bad observations. The expression for this probability is

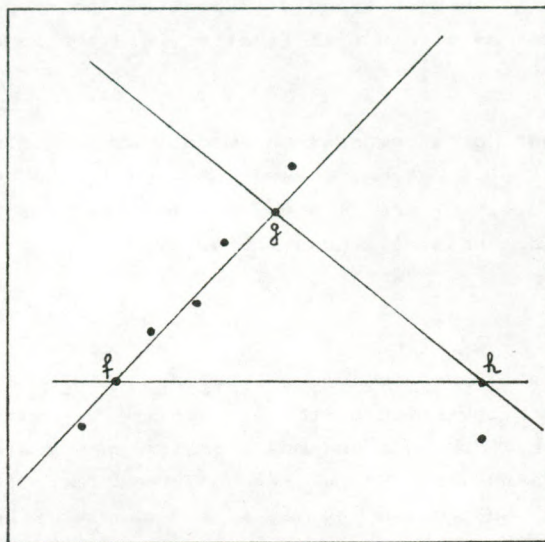
$$1 - (1 - (1/2)^p)^m \quad \text{if } n/p \text{ is large.}$$

(This idea was already used by Stahel in 1981 for multivariate location.)

When n and p are rather small, all possible combinations of p

points out of n are considered instead of the repeated random subsamples.

The basic idea of this algorithm is illustrated in the artificial two-dimensional example below:



For this case n equals 9 and p equals 2. The algorithm will handle all pairs of points out of the 9. We will restrict the explanation for only three such combinations, namely (f,g) , (f,h) and (g,h) . Let us start with the points f and h . The regression surface (which is a line here) passing through the points f and h is found by solving the system of equations

$$y' = \theta_1^{\circ} x' + \theta_2^{\circ}$$

$$y'' = \theta_1^{\circ} x'' + \theta_2^{\circ}$$

where (x', y') and (x'', y'') are the coordinates of respectively f and h . The trial estimate θ_1° and θ_2° are the unknowns. Then, the residuals $y_i - \theta_1^{\circ} x_i - \theta_2^{\circ}$ corresponding with this line are determined for each point i in the sample. The median of the squared residuals (which is the objective function) is calculated and compared with the best value eventually found for previous pairs of points. As a minimization of the squared residuals has to be performed,

the trial estimate corresponding with the points f and h will be retained only when it leads to a lower objective function value. Examining the scatterplot above, it is easy to find out that the pair of points (f,g) will be the 'best' out of the three combinations considered. Indeed, the majority of the observations have a small residual with respect to the line passing through f and g . Repeating this procedure for each pair of points will finally yield the lowest objective function value.

When handling a regression model with intercept, the estimator of location is used for finding the constant term. Once $\theta_1, \dots, \theta_{p-1}$ are found, θ_p is the LMS estimate of location of the sample constituted by

$$z_i = y_i - x_{1,i}\hat{\theta}_1 - \dots - x_{p-1,i}\hat{\theta}_{p-1}, \quad i=1, \dots, n.$$

Apart from the regression coefficients, the scale parameter σ (σ =standard deviation of the e_i) has to be estimated in a robust way. For that purpose a preliminary scale estimate s° is calculated. This s° is based on the value of the objective function, multiplied by a finite-sample correction factor (which depends on n and on p) for the case of normal errors:

$$s^\circ = [\min_i \text{median}_i r_i^2]^{1/2} \times 1.4826 \times (1 + 5/(n-p)) \quad (2.2).$$

The factor $1.4826 = 1/\bar{\sigma}^{-1}(3/4)$ was introduced because $\text{med}_i |\xi_i| / \bar{\sigma}^{-1}(3/4)$ is a consistent estimator of σ when the random variables ξ_i are distributed like $N(0, \sigma)$. From an empirical study, it appeared that the scale estimator was too small in normal error situations, especially for small samples. Therefore the multiplication with the factor $1 + 5/(n-p)$, which has been derived from a simulation study, was necessary.

With this scale estimate the standardized residuals r_i/s° are computed and used to determine a weight w_i for the i -th observation as follows:

$$\begin{aligned}
 w_i &= 1 && \text{if } |r_i/s^0| < 2.5 \\
 &= 0 && \text{elsewhere}
 \end{aligned}
 \tag{2.3}$$

Then the final scale estimate for the LMS regression is given by

$$\sigma^* = [(\sum_{i=1}^n w_i r_i^2) / (\sum_{i=1}^n w_i - p)]^{1/2}.$$

At the classical model, σ^* would be a consistent estimator of σ if the weights w_i were independent of the data (x_i, y_i) .

This algorithm has been implemented in FORTRAN. We called it PROGRES: program for robust regression.

The output of PROGRES consists of results concerning LS and concerning reweighted LS based on the LMS, which is described below. For both methods, PROGRES gives the regression coefficients with their standard deviations and T-values, their variance-covariance matrix, an estimate for the scale parameter σ , the determination coefficient (R squared), the standardized residuals, and residual plots of two types. PROGRES provides also two different options for handling data sets with missing values.

In order to have other classical regression results, like F-tests and options for variable selection, one could run PROGRES first and then use the weights provided by the LMS in a standard package (for example BMDP or SAS). Pursuing this course, one is safeguarded against outliers which may disturb the ordinary LS regression analysis.

The program has been written in a very portable way. It should run without problems on any FORTRAN IV or FORTRAN 77 compiler, as it passed the PFORT portability verifier completely. The program length of PROGRES is about 1800 lines.

A skilful study of the residuals is an important task of applied regression analysis. Therefore PROGRES has a plot option which permits to obtain for both regression techniques a plot of the standardized residuals versus the estimated

value of y , or a plot of the standardized residuals versus the index of the observation i (this is called an index plot). A point in the scattergram is represented by a digit. This digit corresponds to the number of points having approximately the same coordinates. When more than 9 points coincide, an asterisk '*' is printed on that position. In problems with several variables, the residual plots corresponding to the reweighted LS estimator are very useful for spotting the outlying observations. If the residual plot of both the robust and non-robust regression method agree closely, the LS result can be trusted.

In the residual plot a dotted line is drawn through zero and a horizontal band on the interval $[-2.5, 2.5]$ is marked. These lines facilitate the interpretation of the results. When the observed y_i value equals the estimated \hat{y}_i value, then the resulting residual becomes zero. Points in the neighbourhood of this zero line are best fitted by the model.

If the residuals are normally distributed, then one can expect that roughly 98% of the standardized residuals will lie in the interval $[-2.5, 2.5]$. In the residual plots of the reweighted LS, the outliers are far away from this zone. So observations for which the standardized residual is situated far from the horizontal confidence band can be identified as outlying. A warning must be given for this interpretation on the residual plots corresponding to the LS estimator. A true outlier does not necessarily possess a large LS residual. The distortion produced by the outlier(s) pushes the otherwise 'good' observations away from the regression hyperplane. This effect makes it nearly impossible to identify the 'bad' observation(s). This phenomenon is also illustrated by the example in section 4.

Besides the identification of outliers, the residual plots contain also very important information for detecting common types of model inadequacies. A pattern showing that the variance of the residuals increases or decreases with increasing estimated y , points out that it could be favourable to apply a suitable transformation to either an input variable or the output variable. A pattern resembling a horse-shoe may be caused by nonlinearity. In this case a transformation on an input or on the output variable, or an

additional squared term in the model, or the addition of another input variable may be required.

3. ROBUST REGRESSION ESTIMATES DERIVED FROM THE LMS REGRESSION

Several methods exist for improving the efficiency of the LMS. Some of these are presented in this section.

3.1 The reweighted least squares regression

The reweighted least squares regression (RLS) technique consists of minimizing the sum of the squared residuals multiplied by a weight w_i

$$\underset{\hat{\theta}}{\text{minimize}} \sum_{i=1}^n w_i r_i^2 \quad (3.1).$$

The weights w_i are determined from the LMS solution as in equation (2.3) but with the final scale estimate σ^* instead of s^0 . In this way, the result is protected against the presence of outlying points by means of the weights based on the robust LMS estimator.

3.2 The one-step M-estimator

An M-estimate is defined as a solution $\theta = (\theta_1, \dots, \theta_p)^t$ of the system of equations

$$\sum_{i=1}^n x_{ji} \Upsilon(r_i / \sigma) = 0$$

The function Υ is absolutely continuous with derivative Υ' . We use the tangens-hyperbolicus function as defined by Hampel, Rousseeuw and Ronchetti (1981):

$$\begin{aligned} \Upsilon(x) &= x && \text{for } 0 < |x| < p \\ &= (A(k-1))^{1/2} \tanh[\frac{1}{2}((k-1)B^2/A)^{1/2} \\ &\quad \cdot (c - |x|)] \text{sign}(x) && \text{for } p < |x| < c \\ &= 0 && \text{for } c < |x| \end{aligned}$$

where $p=1.470089$, $c=3.0$, $k=5.0$, $A=.680593$ and $B=.769313$.

Let $\theta^* = (\theta_1^*, \dots, \theta_p^*)^t$ be the vector of an initial solution (we will take here the LMS estimates of $\theta = (\theta_1, \dots, \theta_p)^t$ and σ^* the corresponding estimate for the scale parameter σ .)

Bickel (1975) defined a one-step M-estimate as

$$\hat{\theta} = \theta^* + \frac{\sigma^*}{B(\gamma, \bar{\sigma})} (\gamma(r_1^*/\sigma^*), \dots, \gamma(r_n^*/\sigma^*)) X (X'X)^{-1}$$

where $B(\gamma, \bar{\sigma}) = \int \gamma'(u) d\bar{\sigma}(u)$ and X is the $p \times n$ matrix containing the input variables.

4. AN EXAMPLE

In order to illustrate the output provided by PROGRES we have chosen for the famous 'stackloss data' set presented by Brownlee (1965). The data describe the operation of a plant for the oxidation of ammonia to nitric acid. The 3 input variables and the output variable can be described as follows:

- x₁ rate of operation
- x₂ cooling water inlet temperature
- x₃ acid concentration
- y stack loss

We will use a linear regression model with constant term (this is obtained by creating a fourth input variable which takes on the value one for all cases).

We have selected this example because it is a set of real data and it has been examined by a great number of statisticians (Draper and Smith (1966), Daniel and Wood (1971), Andrews (1974), Atkinson (1980) and many others) with the help of several methods. Summarizing their findings, it can be said that most people concluded that observations 1,3,4,21 were outliers. According to some people, observation 2 is reported as an outlier too. Running PROGRES on this data set gives rise to the following output:

 * ROBUST MULTIPLE LINEAR REGRESSION WITH A CONSTANT. *

NUMBER OF CASES = 21
 NUMBER OF COEFFICIENTS (INCLUDING CONSTANT TERM) = 4

THE EXTENSIVE SEARCH ALGORITHM WILL BE USED.

DATA SET = BROWNLEE STACK LOSS DATA

THIS ROBUST MULTIPLE REGRESSION ALGORITHM IS BASED ON
 THE LEAST MEDIAN OF SQUARES (LMS) METHOD.
 (SEE F. ROUSSEEUW, (1984), LEAST MEDIAN OF SQUARES REGRESSION,
 JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, 79, 871-880).
 THIS PROGRAM HAS BEEN WRITTEN BY A. LEROY AND F. ROUSSEEUW.
 FOR FURTHER INFORMATION OR COMMENTS, PLEASE CONTACT

A. LEROY
 VRIJE UNIVERSITEIT BRUSSEL
 C.S.O.D. (M203)
 PLEINLAAN 2
 B-1050 BRUSSELS (BELGIUM)

PRINT OPTION = 2
 PLOT OPTION = 2
 THERE ARE NO MISSING VALUES.

YOUR DATA RESIDE ON FILE : B:GSTACK.DAT
 THE OBSERVATIONS

	OPERATION	TEMPERATUR	ACID CONC.	STACKLOSS
1	80.0000	27.0000	89.0000	42.0000
2	80.0000	27.0000	89.0000	37.0000
3	75.0000	25.0000	89.0000	37.0000
4	62.0000	24.0000	87.0000	28.0000
5	62.0000	22.0000	87.0000	18.0000
6	62.0000	23.0000	87.0000	18.0000
7	62.0000	24.0000	93.0000	19.0000
8	62.0000	24.0000	93.0000	20.0000
9	58.0000	23.0000	87.0000	15.0000
10	58.0000	18.0000	80.0000	14.0000
11	58.0000	19.0000	89.0000	14.0000
12	58.0000	19.0000	88.0000	13.0000
13	58.0000	18.0000	82.0000	11.0000
14	58.0000	19.0000	93.0000	12.0000
15	50.0000	18.0000	89.0000	8.0000
16	50.0000	18.0000	86.0000	7.0000
17	50.0000	19.0000	72.0000	8.0000
18	50.0000	19.0000	75.0000	8.0000
19	50.0000	20.0000	80.0000	9.0000
20	56.0000	20.0000	82.0000	15.0000
21	70.0000	20.0000	91.0000	15.0000

MEDIANS =

OPERATION	TEMPERATUR	ACID CONC.	STACKLOSS
58.0000	20.0000	87.0000	15.0000

DISPERSIONS =

OPERATION	TEMPERATUR	ACID CONC.	STACKLOSS
5.9304	2.9652	4.4478	5.9304

THE STANDARDIZED OBSERVATIONS

	OPERATION	TEMPERATUR	ACID CONC.	STACKLOSS
1	3.7097	2.3607	.4497	4.5528
2	3.7097	2.3607	.2248	3.7097
3	2.8666	1.6862	.6745	3.7097
4	.6745	1.3490	.0000	2.1921
5	.6745	.6745	.0000	.5059
6	.6745	1.0117	.0000	.5059
7	.6745	1.3490	1.3490	.6745
8	.6745	1.3490	1.3490	.8431
9	.0000	1.0117	.0000	.0000
10	.0000	-.6745	-1.5738	-.1686
11	.0000	-.6745	.4497	-.1686
12	.0000	-1.0117	.2248	-.1772
13	.0000	-.6745	-1.1242	-.6745
14	.0000	-.3372	1.3490	-.5059
15	-1.3490	-.6745	.4497	-1.1804
16	-1.3490	-.6745	-.2248	-1.3490
17	-1.3490	-.3372	-.3372	-1.1804
18	-1.3490	-.3372	-1.7986	-1.1804
19	-1.3490	.0000	-1.5738	-1.0117
20	-.2248	.0000	-1.1242	.0000
21	2.0235	.0000	.8993	.0000

SPERMAN RANK CORRELATION COEFFICIENTS BETWEEN THE VARIABLES
 (STACKLOSS IS THE OUTPUT VARIABLE)

OPERATION	1.00			
TEMPERATUR	.74	1.00		
ACID CONC.	.61	.36	1.00	
STACKLOSS	.92	.85	.50	1.00

PEARSON CORRELATION COEFFICIENTS BETWEEN THE VARIABLES
 (STACKLOSS IS THE OUTPUT VARIABLE)

OPERATION	1.00			
TEMPERATUR	.78	1.00		
ACID CONC.	.50	.39	1.00	
STACKLOSS	.92	.88	.40	1.00

LEAST SQUARES REGRESSION

VARIABLE	COEFFICIENT	STAND. ERROR	T - VALUE
OPERATION	.71564	.13486	5.2961
TEMPERATUR	1.25522	.36802	3.41157
ACID CONC.	-.15212	.15629	-0.97331
CONSTANT	-39.91968	11.89600	-3.35572

SUM OF SQUARES = 178.83000

SCALE ESTIMATE = 3.24336

VARIANCE - COVARIANCE MATRIX =

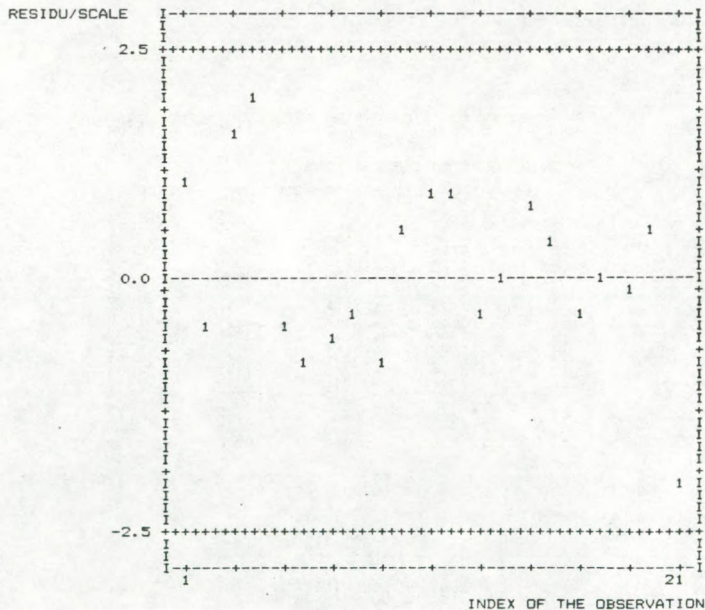
.1819D-01			
-.3651D-01	-.1354D+00		
-.7144D-02	-.1048D-04	.2443D-01	
.2876D+00	-.6518D+00	-.1676D+01	.1415D+03

COEFFICIENT OF DETERMINATION (R SQUARED) = .91358

OBSERVED STACKLOSS	ESTIMATED STACKLOSS	RESIDUAL	NO RES/SC
42.00000	38.76536	3.23464	1 1.00
37.00000	38.91748	-1.91748	2 -.59
37.00000	32.44447	4.55553	3 1.40
28.00000	22.30222	5.69778	4 1.76
18.00000	19.71165	-1.71165	5 -.53
18.00000	21.00694	-3.00694	6 -.93
19.00000	21.38949	-2.38949	7 -.74
20.00000	21.38949	-1.38949	8 -.43
15.00000	18.14438	-3.14438	9 -.97
14.00000	12.73280	1.26720	10 .39
14.00000	11.36370	2.63630	11 .81
13.00000	10.22054	2.77946	12 .86
11.00000	12.42856	-1.42856	13 -.44
12.00000	12.05050	-.05050	14 -.02
6.00000	5.63858	2.36142	15 .73
7.00000	6.09495	-.09505	16 -.28
8.00000	9.51995	-1.51995	17 -.47
8.00000	8.45509	-.45509	18 -.14
9.00000	9.59826	-.59826	19 -.18
15.00000	13.58785	1.41215	20 .44
15.00000	22.23771	-7.23771	21 -2.23

BROWNLEE STACK LOSS DATA

--- LEAST SQUARES ---



LEAST MEDIAN OF SQUARES REGRESSION

THE MINIMIZATION OF THE 12TH QUANTILE OF THE SQUARED RESIDUALS IS PERFORMED.

ON A TOTAL OF 2092 SUBSAMPLES (OF 4 POINTS OUT OF 21)

92 SUBSAMPLES LED TO A SINGULAR SYSTEM OF EQUATIONS.
THE SOLUTION IS ONLY BASED ON THE GOOD SAMPLES.

MULTIPLE LMS SOLUTION

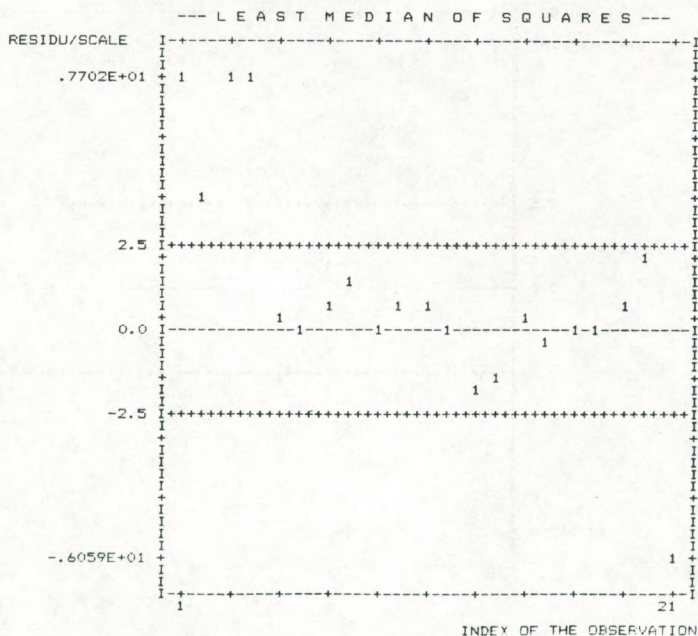
VARIABLE	COEFFICIENT
OPERATION	.71429
TEMPERATUR	.35714
ACID CONC.	.00000
CONSTANT	-34.50000

FINAL SCALE ESTIMATE = 1.26134

COEFFICIENT OF DETERMINATION = .97105

OBSERVED STACKLOSS	ESTIMATED STACKLOSS	RESIDUAL	NO	RES/SC
42.00000	32.28572	9.71428	1	7.70
37.00000	32.28572	4.71428	2	3.74
37.00000	28.00000	9.00000	3	7.14
28.00000	18.35714	9.64286	4	7.64
18.00000	17.64286	.35714	5	.28
18.00000	18.00000	.00000	6	.00
19.00000	18.35714	.64286	7	.51
20.00000	18.35714	1.64286	8	1.30
15.00000	15.14286	-.14286	9	-.11
14.00000	13.35714	.64286	10	.51
14.00000	13.35714	.64286	11	.51
13.00000	13.00000	.00000	12	.00
11.00000	13.35714	-2.35714	13	-1.87
12.00000	13.71429	-1.71429	14	-1.36
8.00000	7.64286	.35714	15	.28
7.00000	7.64286	-.64286	16	-.51
8.00000	8.00000	.00000	17	.00
8.00000	8.00000	.00000	18	.00
9.00000	8.35714	.64286	19	.51
15.00000	12.64286	2.35714	20	1.87
15.00000	22.64286	-7.64286	21	-6.06

BROWNLEE STACK LOSS DATA



[illegible]

VARIABLE	COEFFICIENT	STAND. ERROR	T - VALUE
OPERATION	.68609	.07358	9.32433
TEMPERATUR	.56710	.12872	4.40576
ACID CONC.	-.01725	.05305	-1.32519
CONSTANT	-35.48420	3.80302	-9.33053

WEIGHTED SUM OF SQUARES = 16.02457

CORRESPONDING SCALE ESTIMATE = 1.11229

VARIANCE - COVARIANCE MATRIX =

```

.5414D-02
-.4512D-02      .1657D-01
-.1917D-02      -.3982D-03      .2814D-02
-.5118D-01      -.4384D-01      -.1246D+00      .1446D+02

```

COEFFICIENT OF DETERMINATION (R SQUARED) = .96288

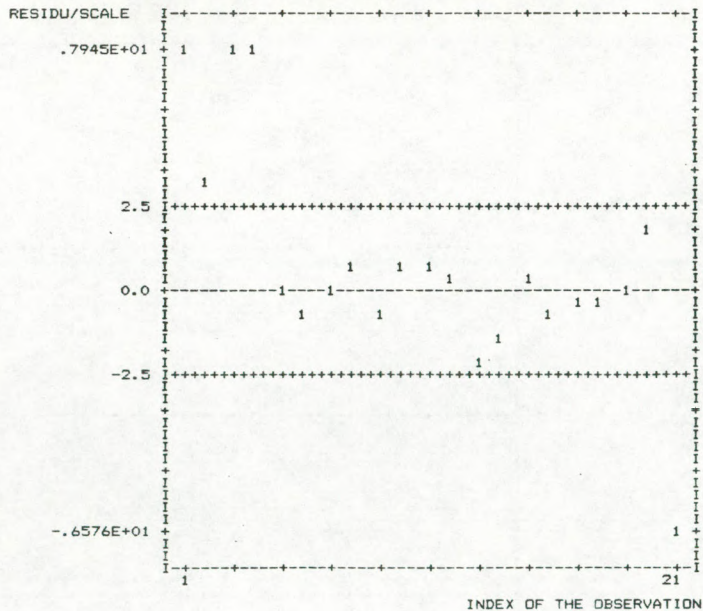
THERE ARE 16 POINTS WITH NON-ZERO WEIGHT.

AVERAGE WEIGHT = .76190

OBSERVED STACKLOSS	ESTIMATED STACKLOSS	RESIDUAL	NO RES/SC	WEIGHT
42.00000	33.17970	8.82030	1	7.93 .0
37.00000	35.19695	8.80305	2	7.42 .0
37.00000	29.59778	8.40222	2	7.25 .0
28.00000	19.16323	8.83677	4	7.94 .0
18.00000	18.02902	-.02902	5	-.03 1.0
18.00000	18.59612	-.59612	6	-.54 1.0
19.00000	19.05973	-.05973	8	-.05 1.0
20.00000	19.05973	-.94027	9	-.97 1.0
15.00000	15.85175	-.85175	9	-.77 1.0
14.00000	13.13700	1.86300	10	.78 1.0
14.00000	12.98174	1.01826	11	.92 1.0
13.00000	12.43189	1.56811	12	1.51 1.0
11.00000	13.10249	-2.10249	14	-1.89 1.0
12.00000	13.47984	-1.47984	14	-1.33 1.0
8.00000	7.49300	1.50700	15	.46 1.0
7.00000	7.54475	-.54475	16	-.49 1.0
8.00000	8.35336	-.35336	17	-.32 1.0
8.00000	8.23260	-.23260	18	-.21 1.0
9.00000	8.78246	-.21754	19	-.20 1.0
15.00000	12.86451	2.13549	20	1.92 1.0
15.00000	22.31456	-7.31456	21	-6.58 .0

BROWNLEE STACK LOSS DATA

--- R E W E I G H T E D L S (B A S E D O N L M S) ---



Examining the residual plot of the reweighted LS confirms that the observations 1,3,4 and 21 are outliers, as their residuals lie far from the confidence band. Observation number 2 is an intermediate case because it is just on the verge of the area containing the outliers. However, the residual plot corresponding to the LS fit masks the bad points.

Concluding this example we would like to emphasize that it is necessary to compare the standardized residuals of both the LS and the robust method in each regression analysis. Only the robust technique can be used as a reliable tool for diagnosing the outliers.

In the following section we will compare the behaviour of the LS, the LMS, the RLS based on the LMS and the one-step M-estimator based on the LMS for different situations by means of a simulation study.

5. A SIMULATION STUDY

In order to have an accurate insight into the performance of the regression estimators mentioned above, we have resorted to a simulation study under three different situations. Samples with size 50 and dimension 10 were simulated and repeated 200 times in each case.

FIRST SITUATION: the normal error case, without outliers

All the input variables are normally distributed $x_{ji} \approx N(0,10)$ and the $y_i = x_{1,i} + \dots + x_{p-1,i} + 1 + e_i$ where the error $e_i \approx N(0,1)$.

SECOND SITUATION: outliers in the y-direction

We considered a contaminated distribution, where 80% of the cases are generated as in the first situation and 20% of the cases are the result of $x_{ji} \approx N(0,10)$ and $y_i = x_{1,i} + \dots + x_{p-1,i} + 1 + e_i$ with $e_i \approx N(10,1)$.

THIRD SITUATION: outliers in the x-direction

As in the first situation the observations have a normal error structure, but in 20% of the observations the original $x_{1,i}$ are replaced by some numbers which are distributed like $N(100,10)$ while the other x_{ji} and the y_i remain unchanged.

The following table shows some summary values resulting from the simulation runs.

ESTI- MATOR	FIRST SITUATION				SECOND SITUATION				THIRD SITUATION			
	LS	LMS	RLS	OSM	LS	LMS	RLS	OSM	LS	LMS	RLS	OSM
VAR 1	1.0006	1.0018	1.0014	1.0025	0.9924	1.0018	1.0008	1.0011	0.0407	1.0018	1.0007	1.0017
	0.0003	0.0012	0.0004	0.0006	0.0059	0.0018	0.0004	0.0006	0.9215	0.0018	0.0004	0.0018
	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.9203	0.0000	0.0000	0.0000
VAR 2	1.0008	1.0023	1.0011	1.0009	0.9985	0.9969	1.0002	0.9997	0.9814	0.9969	1.0000	0.9995
	0.0003	0.0015	0.0005	0.0006	0.0053	0.0015	0.0005	0.0006	0.0222	0.0015	0.0004	0.0007
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0000	0.0000	0.0000
VAR 3	0.9997	0.9990	0.9989	0.9997	1.0101	0.9943	0.9998	0.9978	1.0049	0.9943	0.9996	0.9978
	0.0003	0.0013	0.0005	0.0007	0.0054	0.0015	0.0004	0.0006	0.0292	0.0015	0.0004	0.0006
	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
VAR 4	1.0012	1.0074	1.0029	1.0038	0.9960	1.0018	1.0017	1.0021	0.9853	1.0018	1.0016	1.0016
	0.0003	0.0015	0.0005	0.0006	0.0050	0.0014	0.0004	0.0005	0.0208	0.0014	0.0004	0.0006
	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000
VAR 5	0.9990	0.9960	0.9988	0.9995	0.9946	1.0004	0.9983	0.9989	0.9986	1.0004	0.9982	0.9987
	0.0003	0.0012	0.0005	0.0006	0.0064	0.0017	0.0004	0.0006	0.0261	0.0017	0.0004	0.0006
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
VAR 6	0.9999	0.9992	0.9998	0.9991	0.9982	0.9993	1.0001	0.9996	1.0070	0.9993	1.0001	1.0001
	0.0003	0.0013	0.0004	0.0005	0.0063	0.0015	0.0004	0.0005	0.0258	0.0015	0.0004	0.0006
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000
VAR 7	0.9990	1.0026	0.9990	0.9993	0.9922	1.0007	0.9993	0.9998	0.9931	1.0007	0.9989	0.9998
	0.0003	0.0016	0.0004	0.0006	0.0066	0.0014	0.0004	0.0005	0.0231	0.0014	0.0003	0.0005
	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000
VAR 8	0.9981	0.9987	0.9979	0.9972	1.0058	0.9933	0.9957	0.9943	0.9850	0.9933	0.9957	0.9943
	0.0002	0.0015	0.0004	0.0006	0.0065	0.0017	0.0004	0.0006	0.0256	0.0017	0.0004	0.0007
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000
VAR 9	0.9986	1.0002	0.9980	0.9979	0.9971	1.0013	0.9985	0.9992	1.0151	1.0013	0.9986	0.9992
	0.0003	0.0013	0.0004	0.0006	0.0057	0.0015	0.0004	0.0006	0.0211	0.0015	0.0004	0.0006
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000
CON- STANT	0.9999	0.9996	1.0008	1.0012	0.9977	0.9997	0.9990	0.9985	1.0163	0.9997	0.9990	0.9991
	0.0002	0.0015	0.0005	0.0006	0.0047	0.0014	0.0003	0.0005	0.0284	0.0014	0.0003	0.0005
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0000	0.0000	0.0000
SCALE	0.9892	1.0941	0.8617	0.8929	4.5415	1.4165	0.9657	1.2824	9.6902	1.4085	0.9619	1.3140
	0.0115	0.0015	0.0400	0.0342	12.6127	0.2441	0.0227	0.1192	76.7364	0.2307	0.0205	0.1414
	0.0001	0.0000	0.0191	0.0005	12.5419	0.1735	0.0012	0.0797	75.5203	0.1669	0.0015	0.0986

The first value in each part of the table is the mean estimated value

$$\bar{\theta}_j = 1/M \sum_{k=1}^M \hat{\theta}_{j(k)}, \text{ where } M \text{ is equal to } 200.$$

As the theoretical value for the regression coefficients and the scale parameter is known (in all our situations the theoretical value is 1) the mean squared error (MSE) defined by

$$MSE(\theta_j) = \text{squared bias} + \text{'variance'}$$

$$\sum_{k=1}^M (\hat{\theta}_{j(k)} - \theta_j)^2 = (\theta_j - \theta_j)^2 + 1/M \sum_{k=1}^M (\hat{\theta}_{j(k)} - \theta_j)^2.$$

can be computed. (Analogous summary values can be defined for the scale parameter σ .) The MSE and the squared bias of each estimate are the second and the third value in the above table.

From these results one can deduce that in the normal error design, the mean of the estimates produced by the robust techniques are no worse than the optimal LS estimates, except for the constant term, for which the MSE is slightly larger than for the other coefficients. In the designs where outliers are apparent, the LS shows its dramatic lack of robustness. This conclusion can be drawn from the fact that the mean estimated values differ completely from the theoretical values. Accordingly the MSE values are too large. The robust methods however give rise to mean estimated values which are very close to the theoretical values, with a moderate bias. The value of the estimated scale parameter is also an important criterion to judge the good behaviour of the regression fits. In both designs with outliers the LS scale estimate lead to values which are far away from the theoretical expected one. The LMS and the RLS give a mean estimated scale parameter close to one with a rather small bias, whereas greater values for MSE and bias appear at the OSM scale estimates.

Summarizing, one can say that the alternative robust techniques give rise to very good results for error distributions which depart from the normal distribution,

while they still yield acceptable results in the classical situation where LS is optimal.

References

-
- ANDREWS, D.F. (1974), "A Robust Method for Multiple Linear Regression," *Technometrics*, 16, 523-531.
- ATKINSON, A.C. (1980), "Examples showing the use of two graphical displays for the detection of influential and outlying observation in regression," *Compstat 1980*, 276-282.
- BELSLEY, D.A., KUH, E., and WELSH, R.E. (1980), "Regression Diagnostics," Wiley, New York.
- BICKEL, P.J. (1975), "One-Step Huber Estimates in the Linear Model," *J.A.S.A.*, 70, 428-434.
- BROWNLEE, K.A. (1965), "Statistical Theory and Methodology in Science and Engineering," John Wiley, 2nd ed., New York.
- COOK, R.D. and WEISBERG, S. (1982), *Residuals and Influence in Regression*, Chapman and Hall, London.
- DANIEL, C. and WOOD, F.S. (1971), "Fitting Equations to Data," John Wiley, New York.
- DONOHU, D.L. and HUBER, P.J. (1983), "The Notion of Breakdown Point," in *A Festschrift for Erich Lehmann*, edited by P. Bickel, K. Doksum, and J.L. Hodges, Jr., Wadsworth International Group, Belmont, California.
- DRAPER, N.R. and SMITH, H. (1966), "Applied Regression Analysis," John Wiley, New York.
- HAMPEL, F.R. (1971), "A General Qualitative Definition of Robustness," *Annals of Mathematical Statistics*, 42, 1887-1896.
- HAMPEL, F.R. (1975), "Beyond Location Parameters: Robust Concepts and Methods," *Bulletin of the International Statistical Institute*, 46, 375-382.
- HAMPEL, F.R., ROUSSEEuw, P.J. and RONCHETTI, E. (1981), "The Change-of-Variance Curve and Optimal Redescending M-Estimators," *Journal of the American Statistical Association*, 76, 643-648.
- LEROY, A. and ROUSSEEuw, P.J. (1984), "PROGRES: A Program for Robust Regression," Research Report no. 201, Centrum voor Statistiek en O.R., University of Brussels.
- ROUSSEEuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871-880.