Analytica Chimica Acta, 187 (1986) 171—179
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

LEAST MEDIAN OF SQUARES: A ROBUST METHOD FOR OUTLIER AND MODEL ERROR DETECTION IN REGRESSION AND CALIBRATION

DESIRE L. MASSART* and LEONARD KAUFMAN

Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussels (Belgium)

PETER J. ROUSSEEUW and ANNICK LEROY

Department of Statistics, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels (Belgium)

(Received 13th January 1986)

SUMMARY

The least median of squares method is a robust regression method, which means that it is not sensitive to outliers or other violations of the assumptions of the usual normal model. This contrasts with the conventional regression method, which minimizes the sum of squares. It is demonstrated that the proposed method can be used to detect or correct for outliers or model errors in calibration applications and in comparing two procedures.

Linear regression is used very often by analytical and clinical chemists to obtain calibration lines, to compare two analytical procedures or to relate analytical results to some outside variable. This technique, together with the *t*-test for comparing means of series of observations, is certainly the most used statistical method for analytical purposes. It is usually done by the least-squares technique, and has become so conventional that other possibilities are rarely considered. Weighting factors may sometimes be introduced but basically the criterion (i.e., to minimize the sum of squares of the residuals) remains the same. There are, however, also robust methods; Phillips and Eyring [1] have shown that these methods can be applied in chemical analysis and that these methods possess several advantages over the more conventional least-squares technique. There are, however, many different robust techniques and they have different degrees of robustness. The aim of this paper is to propose the use of a recent technique with a very high degree of robustness and to study the possibilities of its application.

THEORY

The least-squares technique (LS) consists of minimizing the sum of squares of the residuals. For linear univariate regression, these are given by $r_i = y_i - ax_i - b$, where r_i is the residual of measurement y, and a and b are regression coefficients.

0003-2670/86/$03.50

A problem with regression techniques is the effect of outliers. Outliers may occur for three main reasons, namely, recording errors (or copying errors, misplaced decimal points, etc.), inclusion of a case with special characteristics (i.e., not a part of the population being investigated), and modelling errors caused by choosing the wrong model [e.g., use of a linear first-order (straight line) regression instead of a second-order model]. In this case, the errors usually appear at one end of the x-scale, where they are most influential.

A very extreme situation is given in Fig. 1. When the least-squares method is applied to the data shown, point A pulls the least-squares line downwards.

Breakdown point

Hampel [2] defined the breakdown point as the smallest percentage of contaminated data (outliers) that can cause the estimator to take on arbitrarily large aberrant values. The breakdown point of the LS estimator is 0%. This means that even a single outlying point can result in an entirely wrong regression line. Several suggestions have been made to remedy this situation. This has led to "robust" regression methods. A method with the theoretically highest breakdown point possible, namely 50%, has been proposed by Rousseeuw [3], who replaced the least sum of squares by the least median of squares (LMS). The LMS estimator is given by

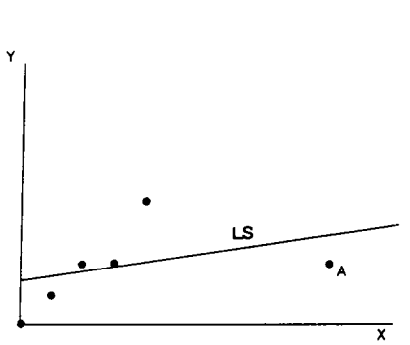
$$\text{minimize } \text{med}_i(y_i - ax_i - b)^2$$

instead of

$$\text{minimize } \text{sum}_i(y_i - ax_i - b)^2.$$

The LMS estimator provides protection from outlying x data as well as against outlying y data, making it very appropriate for situations with errors in both variables.

Phillips and Eyring [1] applied a robust method based on the iteratively



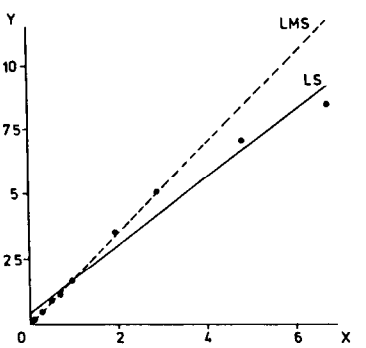


Fig. 1. Effect of outlier on a least-squares line.

Fig. 2. Detection of deviation from the linear model on a calibration line (data from [10]).

reweighted least-squares (IRLS) procedure described by Beaton and Tukey [4]. Good results were obtained on some chemical data sets. It is not clear where the breakdown of this particular IRLS method occurs. In general, IRLS methods have breakdown points between 0% and $1/(p + 1) \times 100\%$, where p is the number of coefficients in the regression equation. This implies that they soon deteriorate in multivariate situations, unlike the LMS method.

Algorithm

The algorithm has been published [3]. Briefly, the slope of the line for which the median of the residuals takes on the minimal value is evaluated by scanning over all angles and with a certain mesh size. After the line that gives the best results has been selected, this procedure is repeated with smaller angle increments around that optimum.

In order to describe an algorithm for the LMS line, the special case of n observations $z_1, z_2, \ldots, z_n$ of the same quantity is first considered. In this case, the LMS estimate is the number b given by

$$\text{minimize } \text{med}_i(z_i - b)^2.$$

The optimal b can be found explicitly. The observations are first ordered as $z_{(1)} \leqslant z_{(2)} \leqslant \ldots \leqslant z_{(n)}$. Then the smallest of the differences is sought from $z_{(h)} - z_{(1)}, z_{(h+1)} - z_{(2)}, \ldots, z_{(n)} - z_{(n-h+1)}$, where h is the largest integer $\leqslant (n/2) + 1$. If $z_{(k)} - z_{(j)}$ is the smallest difference, then $b = (1/2)z_{(j)} + (1/2)z_{(k)}$. For simple linear regression, the LMS coefficients, a and b, are given by

$$\text{minimize } \text{med}_i((y_i - ax_i) - b)^2.$$

For each value of a, the above algorithm can be applied to the $z_i = y_i - ax_i$ to evaluate the corresponding b. Therefore, it is necessary only to minimize the continuous function

$$f(a) = \min_b \text{med}_i[(y_i - ax_i) - b]^2$$

of the single variable a, which is achieved by scanning over a.

This yields the line for which the median of the squared residuals is minimal. Full details are given by Rousseeuw [3].

In the case of multiple regression, the LMS coefficients are given by

$$\text{minimize } \text{med}_i(y_i - \theta_1 x_{i1} - \ldots - \theta_{p-1} x_{ip-1} - \theta_p)^2.$$

Again the constant θ_p is computed explicitly as soon as $\theta_1, \ldots, \theta_{p-1}$ are known. To evaluate the latter, it is necessary to work with trial estimates as described in ref. 5.

EXPERIMENTAL

All literature data were taken from papers in Clinica Chimica Acta; this journal was not selected for any fundamental reason except that it is essentially analytical in nature. The data were obtained by measuring the coordi-

nates of the measurement points in centimetres. This may lead to slight discrepancies with the original figures. The units of the figures in the present paper are centimetres measured on the original plots. In the figures shown, it is sometimes concluded that the LS estimation leads to disputable regression lines. It is not implied that this invalidates other conclusions reached by the authors in the papers cited. The inductively-coupled plasma data were provided by Kornblum [6].

The LMS and LS computations were performed by a FORTRAN program [7], the portability of which was checked by submitting it to PFORT [8]. It should run on any FORTRAN-IV or FORTRAN-77 compiler. An Apple-2 version in APPLESOFT BASIC and an IBM/PC version in FORTRAN were also prepared. The times needed to run the program are given in Table 1. These times may appear rather long for the Apple version, but use of a BASIC compiler should increase considerably the speed of the program if this is considered important. The least-squares second-order polynomials and F-tests were computed with BMDP [9].

RESULTS AND DISCUSSION

Phillips and Eyring [1] demonstrated the effect of using IRLS on a calibration run for which the data are given in Table 2. Here LMS was applied to the same data and the results obtained by LS, IRLS and LMS are compared in Table 2. It can be seen that the same effect but slightly different results are obtained with LMS and IRLS. In both cases, the outlier is eliminated.

Figures 2—10 (with the exception of Fig. 8) give the results obtained with LS and LMS on some data sets from the literature. Figure 2 is a typical example of a calibration line where measurements are made outside the linear range. Solis and Codoceo [10] realized this and used only the seven lowest points; it is clear that if they had not, use of LMS would automatically have discounted the two highest points. As will be shown, however, there are less easily detectable situations of the same kind. Moreover, regression is very often done without plotting and then this situation would have passed undetected. It should be stressed here that LMS is used in these examples as an exploratory tool (i.e., to discover what information the data contain) rather than as a confirmatory technique (to prove or disprove a preconceived hypothesis).

TABLE 1

Time needed to run LMS on microcomputers

No. of points	5	10	15	20	25	27	30
Basic interpreter/ Apple II[a]	5	9.5	17	24	32	44	60
Fortran IBM-PC[b]	14	16	20	30	48	51	55

[a]Time in minutes. [b]Time in seconds.

TABLE 2

Comparison of LS, IRLS and LMS with hypothetical data

Concentration	Signal	Residual[a]		
		LS	IRLS[b]	LMS
1	1.1	0.32	−0.01	0.00
2	2.0	−0.04	−0.03	0.00
3	3.1	−0.20	0.15	0.20
4	3.8	−0.76	−0.07	0.00
5	6.5	0.68	1.71	1.80
a[c]	—	1.26	0.92	0.90
b[c]	—	−0.48	0.19	0.20

[a]Residual = observed signal − predicted signal. [b]The data and the residuals for IRLS were taken from Phillips and Eyring [1]. [c]Coefficients of signal = a (concentration) + b.

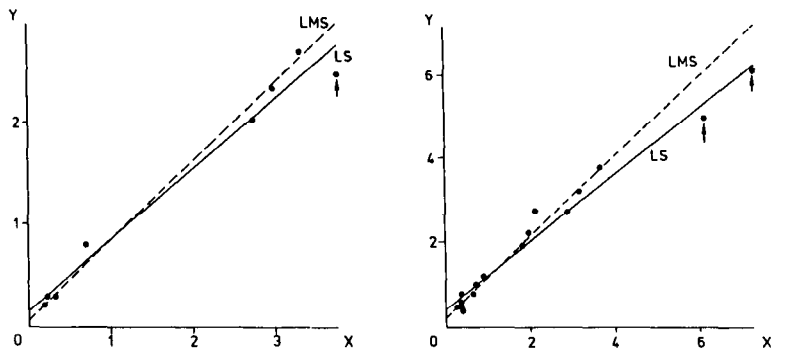


Fig. 3. Influence of outlier (arrowed) on the regression line in the comparison of two methods by LS, and detection of the outlier by LMS (data from [11]).

Fig. 4. Non-linear relationship between the results of two methods as detected by LMS (data from [12]).

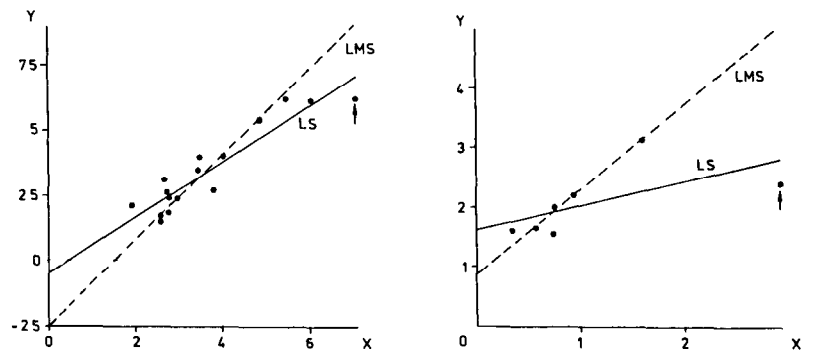


Fig. 5. Influence of outlier (arrowed) on the regression line in the comparison of two methods by LS, and detection of the outlier by LMS (data from [13]).

Fig. 6. Influence of outlier (arrowed) on the regression line in the comparison of two methods by LS, and detection of the outlier by LMS (data from [14]).

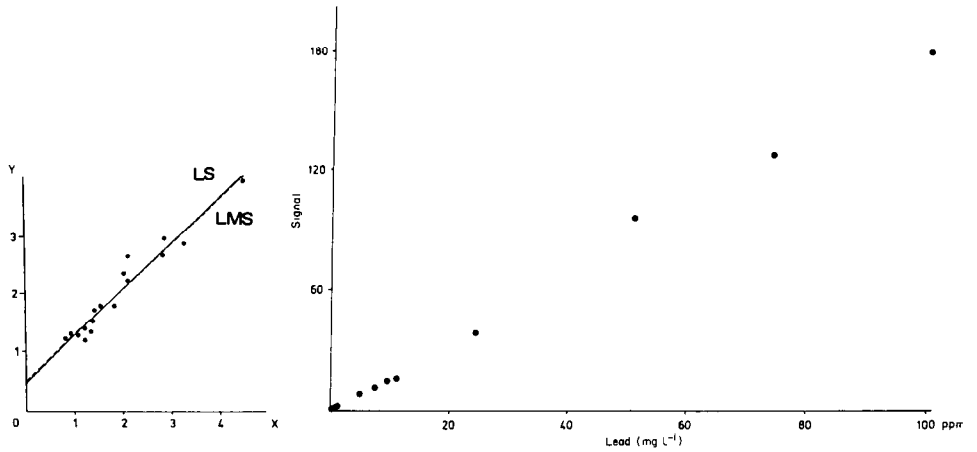


Fig. 7. When no outliers or model errors occur, LMS and LS coincide (data from [15]).

Fig. 8. Scatterplot of the data in Table 3. This plot does not reveal the model error at low concentration by visual inspection; the discrepancy between LMS and LS does show the error.

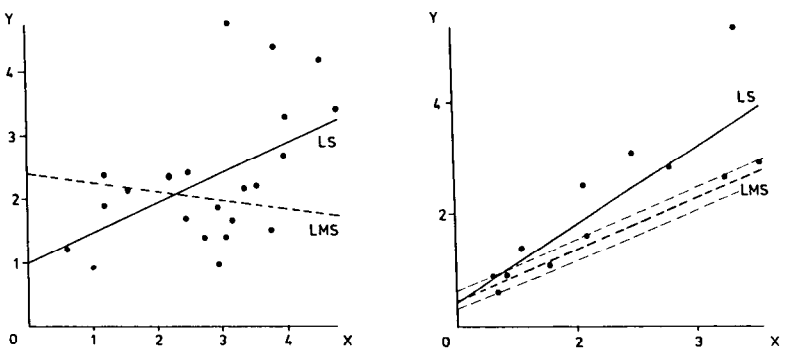


Fig. 9. The discrepancy between LMS and LS on these very diffuse data shows the regression to be meaningless (data from [16]).

Fig. 10. LMS indicates the existence of a linear cluster in the data (data from [17]) because seven points lie within a narrower range around the LMS line.

The following figures give examples where the discrepancy at the end of the scale is less easily detectable, at least if the LMS lines are removed from the figure. In each case, the original authors did not perceive the problem. Figure 3 describes the relationship between an enzymatic end-point method and a rapid kinetic method for lactate determination [11]. The LS line is affected by the measurement indicated by an arrow. To the LMS algorithm, this point is an outlier. Figure 4 [12] is also very typical; it describes two methods for determining breath H levels. The two arrowed measurement points at the end of the scale influence the slope of the LS line to a high

degree whereas the LMS rejects them as outliers. It seems probable that the relationship between the two methods is not linear over the complete measurement range. A still more pronounced effect is found in Fig. 5 [13]. This describes the relationship between the measurements obtained in the comparison of two methods for the determination of phospholipids. The LS line is forced downwards by the arrowed point. The LMS line detects about ten points which fall nearly on the line and takes only those points into account. An extreme (but not unique) example is shown in Fig. 6 [14], which describes the relationship between phosphatase activity and the reticulocyte count in sickle cell disease. The arrowed measurement point has an enormous influence on the LS line, whereas the LMS line shows clearly that the LS regression line as given in the original article is not justified.

These examples indicate one possible use of LMS. When both the LMS and LS methods are applied to the same data and when the two lines do not coincide, then the LS line is likely to be affected by outliers at the end of the scale, i.e., model errors are probable. This can then be assessed visually by inspection of the plot. Indeed, several runs on data in which there was no model error showed that LMS and LS gave very similar results. This was true even when the spread around the regression line was not negligible, as in the case of Fig. 7 [15].

A further example concerns the calibration of lead measurements by plasma emission spectrometry (Fig. 8). The data are given in Table 3. The LMS method detects a model error caused by curvature of the calibration line at the low end of the scale. Because most points were measured at this end and the value of residuals around the line is smaller at low concentrations, the lower points determine the direction of the LMS line. In contrast, the LS line minimizes the higher residuals in the higher concentration range. Neither LMS nor LS permits correct calibration over the whole range. Comparison of the results, however, immediately pinpoints the problem which would not have been possible by visual inspection.

Of course, the use of differences between LMS and LS to detect model errors does not have the character of a statistical test, but there are not many good alternatives. To detect model errors with classical parametric tech-

TABLE 3

Calibration line obtained for lead ions by inductively-coupled plasma emission spectrometry [6]

Conc. ($mg\ l^{-1}$)	Intensity	Conc. ($mg\ l^{-1}$)	Intensity	Conc. ($mg\ l^{-1}$)	Intensity
0.248	0.4738	4.921	7.7360	24.207	38.6705
0.492	0.6997	7.419	11.0610	50.820	96.9765
0.732	1.0432	9.992	15.2173	74.230	127.6312
0.983	1.1836	11.276	15.7363	99.992	180.3638
1.238	1.7150				

niques, three methods can be applied: (1) tests for lack of fit by analysis of variance, but these require replication of the measurements which is always uneconomic and often impossible; (2) analysis of the residuals, but this requires sufficient measurement points to be statistically meaningful and residuals from a least-squares fit may be unduly affected by deviating points, thereby obscuring them; and (3) comparison with higher-order models. In this third method, the fit obtained by LS with the first-order polynomial $y = b + ax$ is compared with the fit obtained by LS with the second-order polynomial $y = b + a_1x + a_2x^2$. An F-test is applied to decide whether or not the fit obtained with the second-order equation is significantly better. If it is better, then it can be concluded that the linear first-order model is not correct, i.e., the model error is detected. In the present study, this was found to be the case for the example shown in Fig. 2 and also for Figs. 4 and 6, but not for Figs. 3, 5 and 8. In the latter cases, therefore, only the LMS method detects that there may be an undue influence from measurement points at the end of the scale.

The detection of model errors seems the most important application of LMS, but LMS can also be used for other purposes. Application of linear regression to very diffuse data sets leads to meaningless regression lines. Application of both LS and LMS is a good way to show this. Again, when the LS line and the LMS line lead to divergent conclusions, the validity of the LS line should be questioned. Figure 9 is an example of this. The data come from Duvivier et al. [16], who described the relationship of estrogen receptor content in cytosol of breast tumor and arterial 17-α-hydroxyprogesterone levels. The authors concluded that there was a relationship, but the opposite (decreasing) direction of the LMS line does induce doubt about the validity of the relationship. Because LMS is not a statistical test, it does not really allow the proposed relationship to be invalidated but it does indicate the necessity of obtaining more measurements.

Finally, LMS can find some features of the data besides outliers. In Fig. 10 [17], the LMS uses only seven of the twelve points and operates very close to the breakdown point. This happens because these seven points form a linear cluster. Whether this makes sense for this particular set of data (which is concerned with the comparison of two methods for the determination of enteropeptidase activity) is impossible to verify but it suggests a possible use of LMS in clustering. Most clustering methods detect only round or ellipsoid clusters. In chemistry, as was pointed out by Wold [18], clusters are very often linear and there is no good clustering method for detecting this kind of cluster. In fact, Buydens [19], using a more complete version of LMS permitting multiple regression [5] to investigate QSAR relationships between gas chromatographic retention and some physicochemical parameters, was able to isolate different linear clusters in this way.

In conclusion, LMS seems to have some desirable properties warranting further investigation of its practical usefulness. It should be stressed that LMS does not reject 50% of the data. Rather, it finds a regression corre-

sponding to the majority of the points, which can then be used to identify the actual outliers (which may be only one or two). This diagnosis of outliers is one of the most difficult aspects of regression, especially in the multivariate situation where the data can no longer be plotted and outliers often go unnoticed. After elimination of the outliers detected by LMS, conventional or weighted least-squares can be computed (as is done by Leroy and Rousseeuw [5]). Finally, it may be noted that the LMS is not only useful for dealing with outliers, but also performs well when the y_i data are not normally distributed around their theoretical value (as in the usual linear model).

REFERENCES

1 G. R. Phillips and E. M. Eyring, Anal. Chem., 55 (1983) 1134.
2 F. R. Hampel, Ann. Math. Stat., 42 (1971) 1887.
3 P. J. Rousseeuw, J. Am. Stat. Assoc., 79 (1984) 871.
4 A. E. Beaton and J. W. Tukey, Technometrics, 24 (1974) 95.
5 P. J. Rousseeuw and A. Leroy, Robust Regression and Outlier Detection, Wiley-Interscience, New York, 1986, in press.
6 G. Kornblum, personal communication.
7 A. Leroy and P. J. Rousseeuw, Revue Belge de Statistique, d'Informatique et de Recherche Operationelle, 24 (2) (1984) 28.
8 B. G. Ryder, The PFORT Verifier, Software Practice and Experience, 4 (4) (1974) 359.
9 W. J. Dixon and M. B. Brown (Eds.), BMDP-79 Biomedical Computer Programs P-series, University of California Press, Los Angeles, CA, 1979.
10 C. Solis and R. Codoceo, Clin. Chim. Acta, 122 (1982) 433.
11 C. Cuthbert and K. G. M. M. Alberti, Clin. Chim. Acta, 90 (1978) 183.
12 T. A. Robb and G. P. Davidson, Clin. Chim. Acta, 111 (1981) 281.
13 M. Sugiwara, T. Oikawa and K. Hirano, Clin. Chim. Acta, 89 (1978) 447.
14 J. Delaunay, S. Fischer, J. P. Piau, M. Tortolero and G. Schapira, Clin. Chim. Acta, 93 (1979) 15.
15 M. Knob and I. Seidl, Clin. Chim. Acta, 106 (1980) 287.
16 J Duvivier, C. Colin, J. Hustin, A. Albert, J. Lavigne, G. Dive and F. Montfort, Clin. Chim. Acta, 112 (1981) 21.
17 I. Antonowicz, F. J. Hesford, J. R. Green, P. Grogg and B. Hadorn, Clin. Chim. Acta, 101 (1980) 69.
18 S. Wold, J. Chromatogr. Sci., 13 (1975) 525.
19 L. Buydens, personal communication.