CROATICA CHEMICA ACTA CCACAA **70** (3) 795–801 (1997)

ISSN–0011-1643
CCA–2454

Original Scientific Paper

A New Median – Based Robust Regression Method

Costel Sârbu

Department of Analytical Chemistry, »Babeş – Bolyai« University,
RO – 3400 Cluj – Napoca, Roumania

Received August 19, 1996; revised December 17, 1996; accepted December 19, 1996

Most instrumental determinations require computation of linear calibration parameters based on a small number of chemical standards. The median – based regression methods are robust methods, which means that they are 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 residuals squares. It is demonstrated that the performance of the proposed method, namely mean median exceeds that of the ordinary least squares method, and often equals that of well known robust median methods.

INTRODUCTION

Analytical calibration, identified generally as the experimental procedure giving rise to conversion of instrumental response into chemical information, is a very significant stage of analysis irrespective of the instrumental techniques applied. The use of a proper means of calibration is important especially when an interference effect is suspected, creating a source of systematic errors in determination. In many cases, correction for inteference effects is commonly accomplished by matching the composition of a set of standard solutions to the sample. In some instances, especially when the sample is completely unknown, the standard addition method is applied.[1]

Using each of the calibration methods recommended in the analytical literature,[2–5] the analytical concentration in a sample is calculated by means of either interpolation or extrapolation of a selected mathematical model, following the acquisition of a set of experimental data. A well known statistical technique commonly used for the construction of such a calibration line is that of least squares (LS). It involves the minimizing of the sum of squares of the residuals around the estimated regression line in the direction of a dependent variable (*i.e.*, instrumental response or its function). For the LS technique it is assumed that the noise in the responses has a normal distribution over the concentration range and that the noise terms are independent of each other. In analytical practice, however, the experimental data obtained may not satisfy these assumptions. An especially important case is frequently met when one or more calibration points give significantly large residuals , relative to the level of noise. The existence of such points, called outliers, violates the assumption of normality of errors. As a consequence, erroneous evaluation of the calibration line parameters is obtained using the LS technique and incorrect final analytical results are achieved.

Some methods of robust regression are known that provide protection against a non-normal distribution. One large group is based on the statistical median.

In this paper, a simpler and more rapid median – based robust regression method is discussed and compared with the ordinary least squares and single median and repeated median methods. The efficiency and practical advantages of the methods considered are shown on the basis of the data obtained in a calibration experiment for the absorptiometric determination of iron in mineral waters.

THEORETICAL CONSIDERATIONS

The ordinary least squares method (LS) consists of minimizing the sum of squares of the residuals. For linear univariate regression, these are given by $r_i = y_i - bx_i - a$, where r_i is the residual of measurement y_i, and a and b are regression parameters. The conventional least squares method assumes, among others, a normal error distribution and consequently the quality of the regression parameters is strongly influenced by the presence of outlying observations. Many different robust regression methods that safeguard against violations of the classical assumptions have been described.

In 1950, Theil[6] proposed the single – median method (SM) which is robust to outliers and estimates the slope, b, as the median of all k slopes of the C_n^2 pairs of (x_i, y_i) and $(1 \leq i < j \leq \mathrm{n})$ between two points,

$$b = \underset{1 \leq i < j \leq n}{\mathrm{med}} \; (y_j - y_i) \; / \; (x_j - x_i). \tag{1}$$

Sorting the above slopes, one can obtain the median of all slopes. The median is equal to the mean of the $k/2^{\mathrm{th}}$ and $k/2{+}1^{\mathrm{th}}$ rank if k is even, or the value of the $(k+1)/2$ if k is odd.

The estimator of the intercept, a, is calculated as the median of the intercepts, obtained with the robust slope, for all data points:

$$a = \operatorname*{med}_{1 \leq i \leq n} (y_i - bx_i). \tag{2}$$

Siegel[7] has improved the single median method by developing the repeated – median estimator (RM). The slope and intercept are obtained as:

$$b = \operatorname*{med}_{i} \operatorname*{med}_{j} (y_j - y_i) \;/\; (x_j - x_i), \tag{3}$$

$$a = \operatorname*{med}_{i} (y_i - bx_i), \qquad i = 1, 2, 3, ..., n. \tag{4}$$

Firstly the median of the slopes is calculated for $(n-1)$ pairs between a point i and all other points j $(j \neq i)$, *i.e.* $\operatorname*{med}_{j} b(i, j)$. This is carried out for all points i $(i = 1,n)$. Thus, n medians are obtained and the median of these n medians, *i.e.* $\operatorname*{med}_{i} (\operatorname*{med}_{j} b(i, j))$, is then the repeated median. The estimator of the intercept, a, is calculated in the same way as explained for the single – median method.

An alternative, and more illuminating and intuitive, expression for b is as a median of the slopes,

$$b_i = (y_i - \bar{y}) \;/\; (x_i - \bar{x}), \qquad i = 1, 2, 3, ..., n \text{ and } x_i \neq \bar{x}, \tag{5}$$

of each observation from the mean point $(\bar{x}, \bar{y})$.

Then
$$b = \operatorname*{med}_{i} b_i, \qquad i = 1, 2, 3, ..., n. \tag{6}$$

With the value of b, values a_i for the intercept are estimated for each point with the aid of the equation $y = bx + a$. Again, the estimates of a are arranged in ascending order and the median value is chosen as the best estimate of the intercept of the line. This algorithm will be named the »mean – median« robust method (MM). It is easy to observe that in this case all the points contribute to the value of the regression parameters; they are included in the mean $\bar{x}$ and $\bar{y}$, respectively.

RESULTS AND DISCUSSION

Using a spectrophotometric method by complexing iron (II) with 1,10 – phenanthroline, the Slãnic Moldova (Roumania) mineral water with an iron content of 6 ppm was analyzed.[1] The absorbances and their statistics for two instruments (Spekol 10 and Specord UV-VIS) are presented in Table I. The content of iron in the sample, C_x (ppm), was obtained using Eq. (7) derived from the standard addition method:[1]

$$C_x = \frac{a \cdot V_s \cdot C_s}{b \cdot V_x}, \tag{7}$$

where V_s = 1 mL and C_s = 0.0411 g/L are the unit volume and the concentration, respectively, of the standard iron stock solution, V_x = 25 mL being the volume of the unknown sample. The parameters of the calibration lines and the final results concerning the concentration of iron, calculated using the regression methods discussed above are shown in Table II.

To evaluate the linearity of the methods studied, this is a good opportunity to compare again the different quality coefficients (QC) used so far in

TABLE I

Calibration data for spectrophotometric determination of iron[1]

Instrument	n	Absorbance measured (y_i)	Absorbance estimated ($\hat{y}_i$)			
			LS	MM	SM	RM
SPEKOL 10	0	0.245	0.249	0.254	0.250	0.250
	1	0.340	0.334	0.337	0.335	0.335
	2	0.420	0.419	0.420	0.420	0.420
	3	0.500	0.500	0.503	0.505	0.505
	4	0.590	0.589	0.587	0.590	0.590
SPECORD UV-VIS	0	0.280	0.280	0.280	0.280	0.280
	1	0.360	0.362	0.362	0.360	0.360
	2	0.440	0.444	0.444	0.440	0.440
	3	0.520	0.526	0.526	0.520	0.520
	4	0.610	0.608	0.608	0.600	0.600

TABLE II

Final results concerning the determination of iron (ppm) in mineral water.

Instrument	Parameter	Method			
		LS	MM	SM	RM
SPEKOL 10	Intercept (a)	0.249	0.254	0.250	0.250
	Slope (b)	0.085	0.083	0.085	0.085
	Concentration (C_x)	4.80	5.02	4.83	4.83
SPECORD UV-VIS	Intercept (a)	0.278	0.280	0.280	0.280
	Slope (b)	0.082	0.082	0.080	0.080
	Concentration (C_x)	5.56	5.60	5.74	5.74

TABLE III

Values of quality coefficients for the methods in Tables I and II

Instrument	Method	QC_1	QC_2	QC_3	QC_4	NQC_5	NQC_6
SPEKOL 10	LS	0.0128	0.0128	0.0010	0.0010	0.3191	0.1369
	MM	0.0188	0.0194	0.0128	0.0128	0.1549	0.2120
	SM	0.0134	0.0135	0.0103	0.0103	0.5922	0.2354
	RM	0.0134	0.0135	0.0103	0.0103	0.5922	0.2354
SPECORD UV-VIS	LS	0.0066	0.0066	0.0071	0.0071	0.1444	0.0036
	MM	0.0079	0.0080	0.0087	0.0088	0.2354	0.1919
	SM	0.0083	0.0082	0.0114	0.0114	0.0000	1.0000
	RM	0.0083	0.0082	0.0114	0.0114	0.0000	1.0000

analytical chemistry for judging the quality of fit of a regression line. Thus, we present in Table III the values obtained for QC_1, defined as[8]

$$QC_1 = \sqrt{\frac{\sum_{i=1}^{n} ((y_i - \hat{y}_i) / \hat{y}_i)^2}{n-1}} \times 100\%, \tag{8}$$

where y_i and $\hat{y}_i$ are the responses measured at each datum and those predicted by the model in Table I, respectively, and n is the number of all data points.

We also show the values of QC_2, constructed similarly to QC_1, with the difference that the measured responses y_i replace the estimates, $\hat{y}_i$, in the denominator,[9] and also QC_3 and QC_4 respectively, referring to the mean of the estimated signal, $\hat{\bar{y}}$, and the mean signal, $\bar{y}$, rather than to the signal itself.[10]

However, taking into account the contradictory values of different quality coefficients and the diversity of the methods concerning their algorithm, we have introduced two new quality coefficients, QC_5 and QC_6, which seem to be more objective.[4,5]

The first coefficient, QC_5, refers to the maximum of absolute residuals,

$$QC_5 = \sqrt{\sum_{i=1}^{n} (r_i / \max |r_i|)^2} \tag{9}$$

and QC_6 refers to the mean of absolute residuals,

$$QC_6 = \sqrt{\sum_{i=1}^{n} (r_i / |\bar{r}|)^2}. \tag{10}$$

Moreover, in the same table, we have introduced the values of the normalized QC_5 and QC_6, namely NQC_5 Eq. (11) and NQC_6 Eq. (12), which take values within the range [0,1] and thus appear to be more practical:

$$NQC_5 = \frac{QC_5 - 1}{\sqrt{n} - 1} \tag{11}$$

and

$$NQC_6 = \frac{QC_6 - \sqrt{n}}{n - \sqrt{n}}. \tag{12}$$

The smaller are the NQC_5 and the higher the NQC_6, respectively, the better is the method.

Examining Table III, it is easy to notice that the median based methods appear to be the best. The »mean median« method proposed in this paper is much closer to the median methods than the classical least squares method.

Overall, it may be stated also that these new quality coefficients concerning the quality of fit confirm our main conclusions and are in good agreement with practical results.

CONCLUSIONS

A new median – based robust regression method was described in this paper. It was compared with the conventional ordinary least squares and other median – based robust methods. The application of these different methods to spectrophotometric data sets proved that the performance of this algorithm excceds that of LS and, in some cases, equals that of well known median robust methods, SM and RM.

In closing, the method discussed above can be successfully used in analytical chemistry and other fields, being much simpler and rather fast. Also, we have to stress that, in the case of this method, any experimental point contributes more or less to the value of the regression parameters via the mean value; whereas the other median-based regression methods practically ignore completely the contribution of outliers.

Moreover, we underline again the effectiveness and the generality of the two new criteria used for diagnosing the linearity of calibration lines.

REFERENCES

1. C. Sârbu, V. Liteanu, and R. Grecu, *Rev. Roum. Chim.* **40** (1995) 829.
2. J. C. Miller and J. N. Miller, *Statistics for Analytical Chemistry*, Horwood, New York, 1988.
3. C. Liteanu and I. Rîcă, *Statistical Theory and Methodology of Trace Analysis*, Horwood, New York, 1980.
4. C. Sârbu, *Anal. Lett.* **28** (1995) 342.
5. H. F. Pop and C. Sârbu, *Anal. Chem.* **68** (1996) 771.
6. H. Theil, *Ned. Akad. Proc. Ser.* **A53** (1950) 386, 521 and 1397.
7. A. F. Siegel, *Biometrika* **69** (1982) 212.
8. K. Knecht and G. Stork, *Fresenius'Z. Anal. Chem.* **270** (1974) 97.
9. P. Kocielniak, *Anal. Chim. Acta* **278** (1993) 177.
10. W. Xiaoning, J. Smeyers-Verbeke, and D. L. Massart, *Analusis* **20** (1992) 209.

SAŽETAK

Nova robustna regresijska metoda zasnovana na medijanu

Costel Sârbu

Većina instrumentalnih određivanja zahtijeva izračunavanje parametara linearne kalibracije uz malen broj kemijskih standarda. Regresijske metode zasnovane na medijanu robustne su, tj. nisu osjetljive na znatnija odstupanja od normalne raspodjele. To je u suprotnosti s klasičnom regresijskom metodom koja minimizira sumu kvadrata odstupanja. Predložena metoda tj. »srednji medijan« nadmašuje uobičajenu metodu najmanjih kvadrata i često je jednako uspješna kao općepoznata robusna metoda medijana.