SCIENTIFIC PAPERS
OF THE UNIVERSITY OF PARDUBICE
Series A
Faculty of Chemical Technology
7 (2001)

ELIMINATION OF OUTLYING VALUES PRIOR TO EVALUATING EXPERIMENTAL DATA BY LINEAR REGRESSION ANALYSIS

Vladimír JEHLIČKA[a1] and Vladimír MACH[b]
[a]Department of Mathematics,
[b]Department of Operation Reliability, Diagnostic and Mechanic in Transport,
The University of Pardubice, CZ–532 10 Pardubice

Received April 22, 2001

In this article, an algorithm for eliminating outlying values prior to estimating linear regression parameters is described. By using such an algorithm, one can objectively exclude outlying values from experimental measurements and determine an interval (x_1, x_2) *covering a linear relation* $y = f(x)$ *between the output variable with a normal distribution* $N(f(x),\sigma^2)$ *and the independent variable x. The theoretical part concerns the derivation of a mathematical equation which allows to identify the outlying values and to determine the critical deviation of a point under testing. In the experimental section, the usefulness of the algorithm proposed is demonstrated on the evaluation of linear regression parameters of a calibration plot and for calculating the equivalence point of selected titration curves. The derived algorithm can generally be applied to evaluation of the concentration dependences of various physico-chemical*

[1] To whom correspondence should be addressed.

variables such as absorbance, polarographic wave-height, conductivity, refraction, optical rotation, etc. A method utilising this algorithm can be used e.g. for analysing the titration end-points in conductometry, amperometry, spectrophotometry, radiometry or thermometry. The algorithm proposed is especially suitable for evaluating a set with small number of experimental data and for processing such sets that comprise two linear segments with insignificantly different slopes of the corresponding regression lines.

Introduction

Numerous experimental relations of two variables are linear within a limited $\langle x_1, x_2 \rangle$ interval, where x is the independent variable, whereas out of this interval the relationship is non-linear. This is typical for concentration dependences of various physico-chemical variables: absorbance, polarographic wave-height, conductivity, refraction or optical rotation. Some functional dependences measured experimentally may also comprise more intervals of the independent x variable where the relation observed exhibits a linear character. In these cases, it is usually necessary to determine the x coordinate for the intersection of the corresponding linear regression lines. This is a case of the methods for the determination of the titration end-point in various analytical techniques such as conductometric, amperometric, spectrophotometric, radiometric, and thermometric titrations.

A model, in which only a part of the relation studied can be approximated with a regression line, is characterised by a $f(x) \in C$ function that is defined as follows

$$f(x) = \begin{cases} \beta_0 + \beta_1 x & \text{for } x \in \langle x_1, x_2 \rangle \\ g(x) & \text{for } x \notin \langle x_1, x_2 \rangle \end{cases} \qquad (1)$$

where $g(x) \neq \beta_0 + \beta_1 x$

When evaluating experimental data, it is first necessary to determine the interval $\langle x_1, x_2 \rangle$ with a linear segment of the function $f(x)$. Then, within this interval, the outlying values can be eliminated and the regression line parameters determined.

In the past, the above-stated problem was solved graphically [1]; at present, it is usually solved in a numerical way. For a long time mainly classical methods of statistic analysis [2] have been used which, however, are often inapplicable to the evaluation of real data. Some improvements were achieved *via* introducing the

regression diagnostics that comprise the identification of significant points, the analysis of multicolinearity [3], the model proposal (including the corresponding transformation [4]), and the verification of individual assumptions for the parameters evaluation [5]. (Other methods recommended for identification of gross errors are described in the literature [6–11].) Nevertheless, the criteria used so far for determination of outlying points are not often unambiguously defined and the final decision usually depends upon the user (experimenter).

For example, the article [10] reports on a robust regression analysis in combination with the method of the least squares of the medians. The authors claim that this procedure is applicable only if at least 50% of experimental points correlates with the dependence observed. Besides this, they chose a criterion for eliminating the outlying values without mentioning the reason for such a decision.

The aim of this paper is to derive an algorithm which would provide — after being used in an appropriate program — an objective and complex resolution of the above-defined task without any influence of the user. The crucial point of the whole problem can be seen in the way of determining an adequate criterion that would also allow one to define objectively whether or not a point tested is an outlying value.

Derivation of Critical Value for a Point under Test

Let us assume a simple linear model given by function (1). In the linear parts of the function f, the random variable Y has a normal distribution $N(\beta_0 + \beta_1 x, \sigma^2)$. In the non-linear part, the variable Y exhibits a distribution formulated as $N(f(x), \sigma^2)$. The measured y value is a value of random variable Y and should be called observation of the random variable Y. It is postulated that the individual x values should be measured without an error and also that, for $x_i \neq x_j$ the corresponding $Y(x_i)$ and $Y(x_j)$ random variables should be completely independent.

In order to consider the outlying character of a point, it is important to know the residuum value, i.e., the deviations between the measured value y and the value calculated from the regression line $\Delta y = y - (b_0 + b_1 x)$.

Unbiased estimates b_0 and b_1 of the unknown parameters β_0 and β_1 can be obtained by the least squares method. For an estimate of the unknown variance σ^2 one can choose the so-called residual variance

$$s_{y,x}^2 = \frac{\sum (y_i - b_0 - b_1 x_i)^2}{n - 2} \qquad (2)$$

The random variable

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \sum \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} y_i = \sum \Delta y_i y_i \qquad (3)$$

is a linear combination of independent random variables Y_i with a normal distribution (which is assumed) and hence, the variable b_1 also exhibits a normal distribution. It can easily be verified that the mean E value as well as the variance D for the parameter b are defined according to

$$\begin{aligned} Eb &= \beta \\ Db &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum \left(x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \right)} \end{aligned} \qquad (4)$$

If the $b_0 + b_1 x$ random variable is expressed in the form of

$$b_0 + b_1 x = \bar{y} + b(x - \bar{x}) \qquad (5)$$

then from the mathematical point of view, expression (5) corresponds to a sum of two independent and normally distributed variables which have a normal distribution characterised as

$$E(b_0 + b_1 x) = E(\bar{y} + b(x - \bar{x})) = \beta_0 + \beta_1 \bar{x} + \beta_1 x - \beta_1 \bar{x} = \beta_0 + \beta_1 x \qquad (6)$$

$$\begin{aligned} D(b_0 + b_1 x) &= D(\bar{y} + b(x - \bar{x})) = D\bar{Y} + (x - \bar{x})^2 Db = \\ &= \frac{\sigma^2}{n} + \frac{(x - \bar{x})^2 \sigma}{\sum (x_i - \bar{x})^2} = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] \end{aligned} \qquad (7)$$

Now, it is evident that for each x, the $Y = b_0 + b_1 x$ random variable exhibits a normal distribution

$$N\left(\beta_0 + \beta_1 x,\ \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}\right]\right) \tag{8}$$

The deviation written as $\Delta y = y - (b_0 + b_1 x)$ is a difference between two independent random variables with a normal distribution and thus it results in a normal distribution as well. By calculation, it can be found that

$$E\Delta y = Ey - E(b_0 + b_1 x) = \beta_0 + \beta_1 - (\beta_0 + \beta_1 x) = 0 \tag{9}$$

and

$$\begin{aligned} D\Delta y = Dy + D(b_0 + b_1 x) &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}\right] = \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}\right] \end{aligned} \tag{10}$$

Also, the previous results suggest that the $\Delta y = y - (b_0 + b_1 x)$ random variable can be characterised by a $N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}\right]\right)$ distribution and the corresponding $\dfrac{\Delta y}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}}}$ standardised variable has therefore a $N(0,1)$ normal distribution .

Because the $\dfrac{(n-2)s_{y,x}^2}{\sigma^2}$ variable has a χ^2- distribution with $(n-2)$ degrees of freedom, it can be written — according to the definition for a t-distribution — that the random variable

$$\frac{\Delta y}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}}\Bigg/\sqrt{\frac{(n-2)s_{y,x}^2}{\sigma^2(n-2)}} = \frac{\Delta y}{s_{y,x}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}} \tag{11}$$

also exhibits a t-distribution with $(n-2)$ degrees of freedom. Furthermore, this formula determines the critical value for the deviation of a (x_t, y_t) point tested

$$\Delta y_{t,crit} = t_{n-2,\alpha}s_{y,x}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}} \tag{12}$$

where $t_{n-2,\alpha}$ is the critical value of the t-distribution and α is the confidence level chosen.

The equation derived in this way (12) can then be applied to testing of the outlying values for individual measurements.

Experimental

Hardware, Programming, Apparatus, and Chemicals

In order to obtain the calibration dependence for spectrophotometric determination of nitrate and for chelatometric microtitration of scandium, the absorbance values were measured with a Specol spectrophotometer (Model 10, Carl-Zeiss, Jena, FRG) using ordinary glass cells.

Conductometric titration of a mixture of both strong and weak acids with the standard solution of sodium hydroxide was carried out using an OK-104 conductometer (Radelkis, Budapest, Hungary).

Proposed Algorithm and Its Function

The algorithm to solve the above-defined problem can be divided onto the following steps:

I. The algorithm selects the first ten values from the entire set of experimental

data when forming subsets that contain all possible combinations of five points. For common measurements, such a set of five points can be considered to be sufficient for a reliable estimation of the regression line coefficients. Using the least squares method, the algorithm calculates the standard deviations, $s_{y,x}$, for all subsets, with a parallel testing whether a given subset contains an outlying value, which is performed with the aid of the derived criterion (12). In this way, all other groups of ten points, i.e., $2^{nd} - 11^{th}$, $3^{rd} - 12^{th}$, $4^{th} - 13^{th}$ point etc., are analysed until all experimental points are processed. A set of five points which exhibits the smallest standard deviation, $s_{y,x}$, is then selected for further processing.

II. Afterwards, using this criterion, all the remaining experimental data are investigated against the initial set of five points. If the next point to be tested is not an outlying value, it is taken into the set of already selected points. As the number of included points gradually increases, the criterion (12) becomes more and more limiting and it is necessary — after each newly added point — to check whether the set does not include any outlying point. If a point does not fulfil the criterion, it is regarded to be outlying and the process continues by testing the subsequent point till all the data (points) are processed.

III. From the set of points belonging to the linear range of the functional dependence tested, the regression line parameters are computed by using the least squares method.

IV. If the functional dependence exhibits several linear segments, the remaining data being ascertained (in the step II) as outlying are used to form a new data set, in which another linearity is sought and approximated with the corresponding regression line. Hence, the data manipulation and calculations are repeated again starting from the step I.

V. The whole calculation is stopped at the moment when a set of five points with no outlying points in the step I is found. It means that any other linear segment has not been revealed, and the remaining points should be considered to be outliers with respect to the linear segments already ascertained.

Based on the algorithm described above, an *OK-LIN* program has been assembled. This program allows one to determine objectively the linear regression parameters for one or even more regression lines of the dependence studied. At the same time, it ensures an effective elimination of all outlying values. If the program has revealed at least two linear parts, it computes their regression lines together with the coordinates of their intersections.

The role of a user of the *OK-LIN* program is reduced to correct entering of the input data, i.e., the (x_i, y_i) coordinates of individual points. The proper process of data manipulation is then done automatically and runs practically without the user's assistance.

Results and Discussion

The following paragraphs present some selected examples of calculations obtained with the aid of the *OK-LIN* program evaluating typical experimental data.

The input data are summarised in tables where the shaded boxes denote all points which have been identified as outlying by the program. In each example, the results of linear regression analysis are provided with the individual coefficients given together with their precision. The corresponding graphs plotted directly by the program are also shown.

Example 1: *Calibration Plot for Spectrophotometric Determination of Nitrate*

Table I Experimental dat

c, mg l^{-1}	0.00	3.00	6.00	9.00	12.00	15.00	18.00	21.00	24.00	27.00	30.00
A	0.018	0.105	0.189	0.358	0.362	0.440	0.530	0.613	0.703	0.789	0.875

Results:

Linear regression analysis: $A = 0.0185 + 0.028536 * c$

Precision (±) 0.0013 0.000068

Regression equation was calculated for c (concentration) values within an interval of ⟨0; 30⟩.

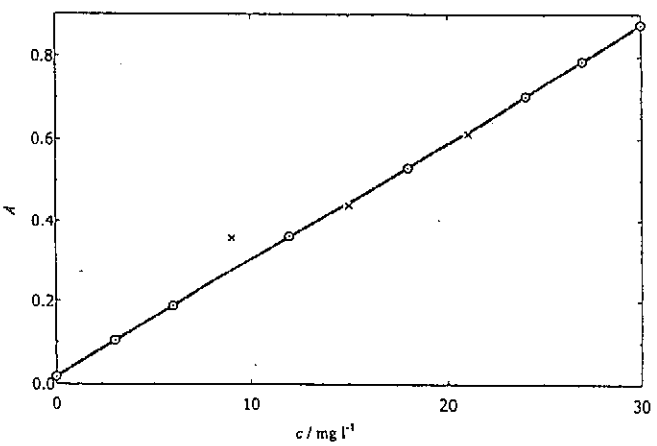

Fig. 1 Calibration plot for spectrophotometric determination of nitrate

Comments: The program eliminated a grossly outlying point with coordinates (9.00; 0.358). This point, in fact, could be excluded even by a subjective (graphical) evaluation of the data. The *OK-LIN* program eliminated also other two points, (15.00; 0.440) and (21.00; 0.613), whose outlying character might not be evident by merely subjective analysis. The reason for excluding these data by the program is that the remaining points, belonging to the linear range, were measured with a higher precision compared to that for both eliminated points.

Example 2: *Spectrophotometric Microtitration Curve for Chelatometric Determination of Scandium [12]*

Table II Experimental data

V, µl	0.0	20.0	40.0	60.0	80.0	100.0	110.0	120.0	130.0
A	0.623	0.589	0.539	0.469	0.411	0.342	0.308	0.274	0.238
V, µl	140.0	150.0	160.0	170.0	180.0	190.0	200.0	210.0	220.0
A	0.206	0.172	0.137	0.103	0.069	0.040	0.035	0.036	0.035
V, µl	230.0	240.0	250.0	260.0	280.0	300.0	320.0	340.0	
A	0.035	0.034	0.037	0.035	0.036	0.035	0.034	0.035	

Results:

Linear regression analysis, equation 1:	A =	0.6841	-	$0.0034175 * V$
Precision (±)		0.0011		0.0000080
Linear regression analysis, equation 2:	A =	0.0359	-	$0.000003 * V$
Precision (±)		0.0029		0.000011

Linear regression analysis was performed for V (volume) values within an interval of ⟨80; 180⟩ for equation 1, and of ⟨200; 340⟩ for 2.

Intersection of regression lines: $V = 189.9$; $A = 0.035$

Comments: Before and beyond the equivalence point, the titration curve has a linear character. The titration end-point is given by V-coordinate in the intersection of regression lines 1 and 2 which were calculated separately for both linear segments of the titration curve.

In this curve, the first four points from (0.0; 0.623) to (60.0; 0.469) as well as the point (190.0; 0.040) were eliminated as the values lying out of the linear part of the titration curve. Other two points, (130.0; 0.238) and (250.0; 0.037), were found outlying too; in this case, because of the lower precision of measurement in comparison with that for points lying in the linear part of the curve.

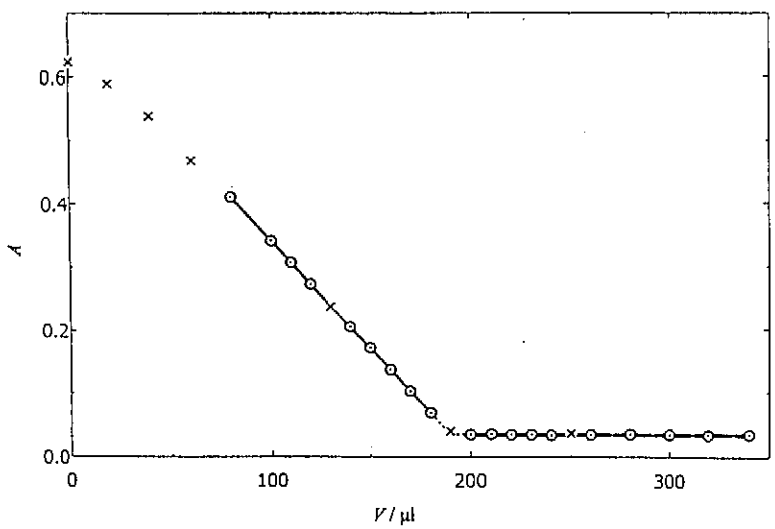

Fig. 2 Spectrophotometric microtitration curve for chelatometric determination of scandium

Example 3: *Conductometric Titration Curve for a Mixture Containing Both Strong and Weak Acid Titrated with Sodium Hydroxide Standard Solution [10]*

Table III Experimental data

V, ml	0.00	0.50	1.00	1.50	2.00	2.50	3.00	3.50	4.00
G, mS	1.85	1.72	1.59	1.48	1.35	1.23	1.10	0.97	0.85
V, ml	4.50	5.00	5.50	6.00	6.50	7.00	7.50	8.00	8.50
G, mS	0.74	0.65	0.61	0.62	0.63	0.67	0.69	0.73	0.75
V, ml	9.00	9.50	10.00	10.50	11.00	11.50	12.00	12.50	13.00
G, mS	0.79	0.82	0.85	0.88	0.94	1.02	1.10	1.19	1.28
V, ml	13.50	14.00	14.50	15.00	15.50				
G, mS	1.37	1.45	1.54	1.63	1.72				

Results:

Linear regression analysis, equation 1: $G = 1.8455 - 0.2478 * V$

Precision (±) 0.0086 0.0032

Linear regression analysis, equation 2: $G = 0.230 + 0.0620 * V$

Precision (±) 0.023 0.0027

Linear regression analysis, equation 3: $G = -1.000 + 0.1753 * V$

Precision (±) 0.027 0.0020

Linear regression analysis was performed for V values within an interval of ⟨0.00; 4.50⟩ for equation 1, of ⟨6.50; 10.50⟩ for 2, and ⟨11.50; 15.50⟩ for 3.

Intersection of regression lines 1 and 2: $V = 5.22$; $G = 0.55$
Intersection of regression lines 2 and 3: $V = 10.85$; $G = 0.90$

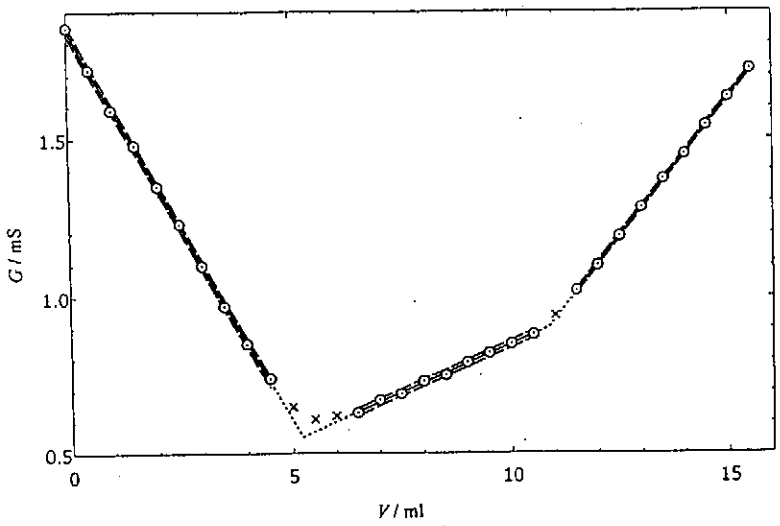

Fig. 3 Conductometric titration curve for a mixture containing both strong and weak acid titrated with sodium hydroxide standard solution

Comments: This type of alkalimetric titration is an example of a functional dependence characterised by three linear segments with two intercepts of the corresponding regression lines. The V-coordinate of the first intercept indicates the titration end-point for strong acid whereas the V-coordinate of the second intercept belongs to the end-point of weak acid.

Figure 3 showing the titration curve also illustrates the confidence intervals for individual linear segments. (Such intervals could not be evident in the previous examples whose data had been measured with substantially higher precision and hence the individual confidence intervals coincided with the corresponding regression lines.)

Even this last example did not require any involvement of the user during the data processing by the program.

Example 4: *Conductometric Titration Curve for a Mildly Strong Acid Titrated with Sodium Hydroxide Standard Solution*

Table IV Experimental data

V, ml	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00
G, mS	7.10	7.00	5.90	4.80	4.05	3.00	2.30	1.70
V, ml	9.00	10.00	11.00	12.00	13.00	14.00	15.00	16.00
G, mS	3.20	4.80	7.10	9.20	10.90	13.00	15.10	16.80

Results:

Linear regression analysis, equation 1: $G = 8.05 - 0.808 * V$

Precision (±) 0.42 0.078

Linear regression analysis, equation 2: $G = -14.38 + 1.955 * V$

Precision (±) 0.66 0.051

Linear regression analysis was performed for V (volume) values within an interval of $\langle 1.00; 8.00\rangle$ for equation 1, and of $\langle 9.00; 16.00\rangle$ for 2.

Intersection of regression lines: $V = 8.00$; $G = 1.49$

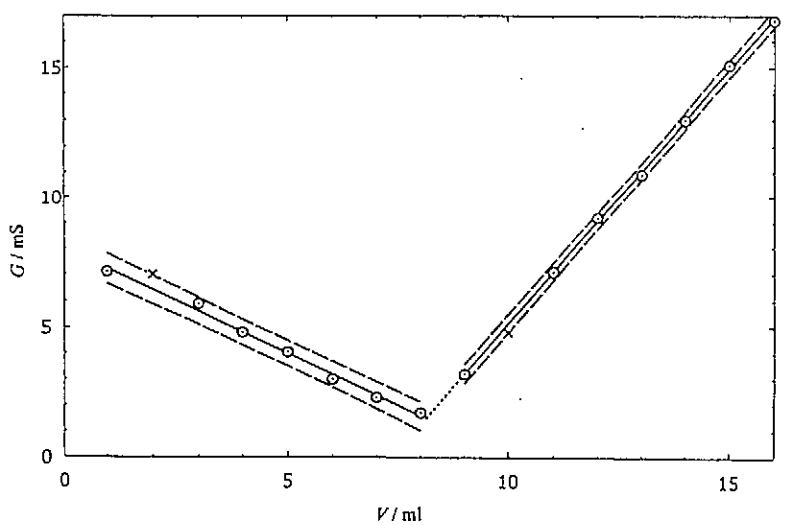

Fig. 4 Conductometric titration curve for a middle strong acid titrated with sodium hydroxide standard solution

Comments: In this case, the experimental data represent a set measured with a larger variance. Based on testing by the criterion (12), only two points, (2.00; 7.00) and (10.00; 4.80), are eliminated and all the remaining points are used for calculations. The resultant dependence is characterised by broadening of the confidence intervals, as depicted in the figure.

Example 5: *The Determination of Critical Micellar Concentrations of Surfactants [13,14]*

Table V Experimental data

c, mmol l^{-1}	1.012	1.985	3.822	5.528	7.117	8.599	9.305
γ, $\mu S\ cm^{-1}$	44	87	166	244	316	379	409
c, mmol l^{-1}	9.987	10.647	11.287	11.907	12.509	13.092	13.658
γ, $\mu S\ cm^{-1}$	439	469	497	525	545	575	597
c, mmol l^{-1}	14.208	14.742	15.262	15.766	16.257	16.735	1.199
γ, $\mu S\ cm^{-1}$	619	639	660	679	698	717	734

Results:

Linear regression analysis, equation 1: $\gamma = -0.4 + 44.08 * c$

Precision (±) 1.0 0.12

Linear regression analysis, equation 2: $\gamma = 67.4 + 38.80 * c$

Precision (±) 3.8 0.25

Linear regression analysis was performed for V (concentration) values within an interval $\langle 1.012; 11.907 \rangle$ of for equation 1, and of $\langle 13.092; 17.199 \rangle$ for 2.

Intersection of regression lines: $c = 12.830$; $\gamma = 565$

Comments: Based on evaluating the data by using the *OK-LIN* program, two linear segments were found. The case when the slope of both regression lines differs only very slightly is not solvable — as far as we know — by any present method. The resultant solution of standard evaluation allowing to find the only linear segment is illustrated in Fig. 6.

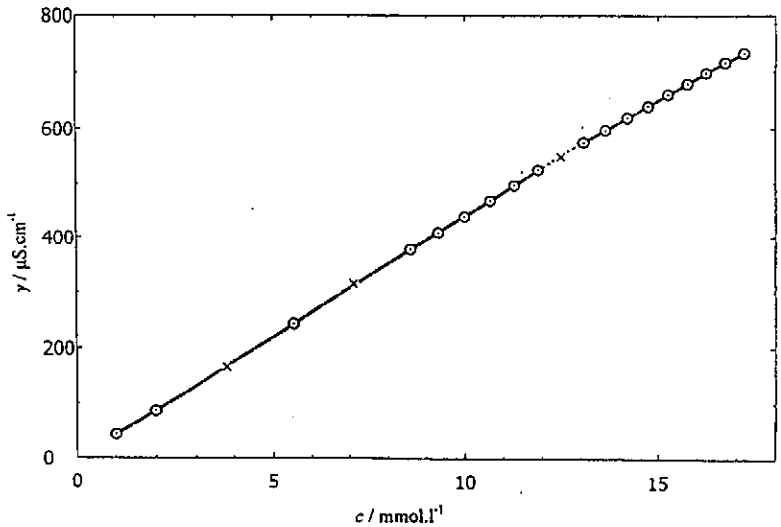

Fig. 5 The determination of critical micellar concentrations of surfactants — a graph plotted by OK-LIN program

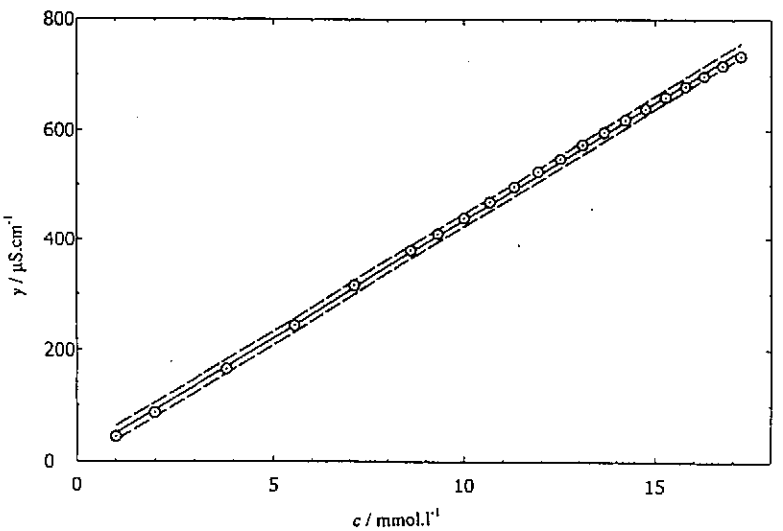

Fig. 6 The determination of critical micellar concentrations of surfactants — a dependence obtained by using method of the least squares

Conclusion

Statistical analysis of experimental data performed with the *OK-LIN* program enables to an objective elimination of outlying values when utilising a specially derived criterion. By simply applying the least squares method, the regression line parameters are calculated *via* analysing the experimental points from a linear part of the dependence studied. If the dependence tested exhibits several linear segments, the values given as the coordinates of individual intersections of regression lines can be obtained by solving the system of the corresponding linear regression equations.

The use of the *OK-LIN* program is not limited only to the analysis of calibration plots and titration curves, but, in general, it is applicable to situations when attention is to be paid to the evaluation of the intercept(s) of coordinates for linear segment(s) of the dependence studied. The proper function of the proposed algorithm can best be verified by analyses of both real and simulated experimental data.

If the algorithm proposed is compared, for example, with the LMS method [10], it can be stated that, in most cases, the results obtained are essentially the same. The difference can be found in the data evaluated for a set shown in Example 4. Only the algorithm presented herein is capable of finding two objectively existing linear segments for such a type of the data. Another advantage of the algorithm is the fact that it even allows evaluation of experiments with a small number of data. This results in the substantial reduction of the expenses necessary to carry out the corresponding analytical measurements.

In our opinion, the methodical procedure presented herein brings a highly effective processing of experimental data with reliable numerical results. However, the utmost importance can be attributed to an objective evaluation of outlying values, whose determination is not subjectively affected by the user.

Acknowledgements

The authors would like to thank Dr. I. Švancara from the Department of Analytical Chemistry at the University of Pardubice for his interest and some valuable comments.

References

[1] Gray J.B.: Prac. Stat. Comput. Sect., ASA Washington **23**, 159 (1983).
[2] Anscombe F.J.: Amer. Statist. **27**, 17 (1973).
[3] Balsey D.A., Kuh E., Welsh R.E.: *Regression Diagnostics*, Wiley, New

York, 1980.
[4] Atkinson A.C.: *Plot, Transformation, Regression*, Clarendon Press, Oxford, 1986.
[5] Weisberg S.: Technometrics **25**, 219 (1983).
[6] Chattarje S., Hadi A.S.: Statist. Sci. **1**, 379 (1986).
[7] Cook R.D., Weisberg S.: *Residuals and Influence in Regression*, Chapman and Hall, New York, 1982.
[8] Gorman M.A., Myers R.M.: Commun. Statist. **16**, 771 (1987).
[9] Utts J.: Commun. Statist. **11**, 2801 (1982).
[10] Ortiz-Fernández M.C., Herrero-Gutiérrez A.: Chemometrics and Intelligent Laboratory Systems **27**, 231 (1995).
[11] Rousseeuw P.J., Leroy A.M.: *Robust Regression and Outlier Detection*, Wiley, New York, 1987.
[12] Mach V., Kotrlý S., Vytřas K.: Chem. Papers, **43**, 377 (1989).
[13] Fischer J., Jandera P.: J. Chromatogr. B **681**, 3 (1996).
[14] Jandera P., Fischer J.: J. Chromatogr. A **728**, 279 (1996).