Determination of Intersection of Regression Straight Lines with Elimination of Outliers

[a]V. JEHLIČKA and [b]V. MACH

[a]*Department of Mathematics, Faculty of Economics and Administration, University of Pardubice, CZ-532 10 Pardubice*

[b]*Department of Operation Reliability, Diagnostic, and Mechanics in Transport, Jan Perner Faculty of Transport, University of Pardubice, CZ-532 10 Pardubice*

Received 27 April 1998

Algorithm of determination of coordinates of point of intersection of regression straight lines was elaborated. Regression straight lines characterize linear segments of measured dependence. Regression straight lines equations are given by simple application of the least-squares method to set of points, which with respect to given regression straight line are not outlying. Decision, whether the measured point of functional dependence is remote from the linear segment, is done in an objective way on the basis of special statistical criterion, which is a part of OK-LIN program, that allows processing of experimental data.

It is necessary to determine x-coordinate of point of intersection of linear parts of studied dependences at evaluation of some experimentally measured functional dependences. For example, determination of the end of titration of instrumental analytical determination is dealt especially at conductometric, amperometric, spectrophotometric, radiometric, and thermometric titrations.

Classical, previously used graphical processing of file of measured data [1—3] is at present replaced by data processing at personal computers with the use of various statistical programs. Therefore attention at solution of the given task was set to the possibility of utilization of present computation technique and to minimalization of experimenter's role, whose work should be finished with storage of measured data (values of independent and dependent variable) to the computer.

Some methods are discussed in literature [4], that make possible the determination of outlying values. However, the published criteria for outliers determination are not often defined in a single-value way and final resolution generally depends upon consideration of the experimentalist. Some of the quoted methods were inbuilt into programs, which are used at statistical evaluation of experimentally measured data.

M. C. Ortiz-Fernández and *A. Herrero Gutiérez* [5] advise the robust regression method of the least squares of medians (LMS) for determination of parameters of regression straight lines. But it can be used only under presupposition that at least 50 % of experimentally ascertained points suit to this dependence. LMS regression, as quoted authors publish, was originally used for calibration and detection limit calculations with anticipation of the presence of outliers. Daily exploitation of LMS regression makes easy usage of the program PROGRESS [6] . A disadvantage of this approach is, except others, the necessity of interactive work of the experimentalist, who has to appoint the final value of criterion on account of determination of deviousness of the points.

The problem of the way of elimination of remote values out of the file of measured data and determination of one linear segment of functional dependence has been already discussed by *Jehlička* and *Mach* [4] in a detailed way. The criterion for elimination of outlying values has been elaborated in [4].

In the presented work an alternative algorithm for determination of more than one linear segment of functional dependence with objective elimination of outlying values is described. The coordinates of the points of intersection of linear parts of functional dependence are always obtained by solution of the system of two equations of appropriate regression straight lines.

Principle of Designed Algorithm

A model in which only a part of the dependence studied can be replaced by the equation of regression straight line is characterized by the function $f(x) \in C$ 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 $g(x) \neq \beta_0 + \beta_1 x$. In the linear section of function

f the measured value y is a value of random quantity Y with normal spread $N(\beta_0 + \beta_1 x,\ \sigma^2)$. In the nonlinear section of function f the measured value is the value of random quantity with the spread $N(g(x),\ \sigma^2)$. The x values are measured accurately. For $x_i \neq x_j$, the corresponding random quantities $Y(x_i)$ and $Y(x_j)$ are independent.

Important for evaluation of remoteness of point is the value of residuum, *i.e.* deviation of the y value measured from the value calculated from the regression straight line $\Delta y = y - (b_0 + b_1 x)$. In paper [4] is deduced a relation for determination of the critical value of deviation for the point tested $(x_i,\ y_i)$

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

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

The algorithm of determination of coordinates of point of intersection of regression straight lines of linear parts of functional relations with objective elimination of outlying values contains six steps.

The first step: Using the least-squares treatment a set of five neighbouring experimental points is determined with the lowest value of standard deviation $s_{x,y}$. At the same time any of determined five points must not be outlying on the basis of criterion (*2*) with respect to remaining four points of examined group of five. The points that are convenient to mentioned requirements are registered to new-coined file of points that are included into linear part of the dependence studied.

The second step: With respect to the file of included points the remaining so far not included measured points are tested step by step. If the absolute value of deviation of tested point from regression straight line led through the file of still included points smaller than the critical value of criterion (*2*), the tested point is included into the file of included points. In other cases the tested point is signed as remote. All points, that are not included, are tested step by step.

The third step: With increase of number of points included into linear segment the criterion for deviousness evaluation of tested points is more strict. Therefore, if the new tested point is included into the file of included points (see the second step), then individual points from the file of included points have to be newly tested with respect to remaining included points. If any of points does not correspond with appropriate requirements, then it is eliminated out of the file of included points and it is signed as outlying.

The fourth step: The parameters of regression straight line (calculated for points accepted for the linear part) will be calculated by the method of the least squares.

The fifth step: The file of points, that are not included with respect to the given regression straight line, creates the new starting file of experimental data, in which it is possible to determine other linear parts including appropriate regression straight lines equations. The calculation is also put back to the first step and the whole cycle is repeated so long, until all the linear segments, that are in experimentally measured dependence, are determined.

The sixth step: Values of coordinates of points of intersection of considered straight lines are determined by solution of the system of two equations of appropriate regression straight lines.

Program Processing of Algorithm

The above described algorithm of evaluation of experimental data was built-in to the OK-LIN program [4], which makes possible to solve studied problems quite automatically. The user enters only input data, it means values x_i, y_i and reliability, what they were measured with.

The result of calculation is output, which is enclosed to every example presented in Chapter *4. Examples of OK-LIN Program Applications.*

The output begins with the equation of the regression straight line and interval, in which it was determined. Parameters values of the regression straight line equation are presented with accuracy resulting from their calculation [1].

Data valid for next linear segments are written in the same way.

In the very ending of the output the coordinates of points of intersection of regression straight lines are presented.

The above described numerical results of the calculation can be displayed in the computer's screen, printed at printer or saved into the data file, which can be further processed by textual editors.

The calculation results are also processed in a graphical way. The graph, in which are by default marked points included into linear segment by circle and points not included by cross is displayed in the computer's screen. Regression straight lines are depicted by continuous line, reliability strips are marked with dashed lines.

Further description of the Czech version of OK-LIN program working in DOS environment was published earlier in [7].

Examples of OK-LIN Program Applications

The following three examples demonstrate the use of OK-LIN program. Numerical result of calculation and the graph are stated at each example and everything is completed with a brief commentary.

Example 1. Conductometric CH_3COOH titration by solution of 0.2 M-NaOH (Fig. 1)

1. Regression straight line equation:

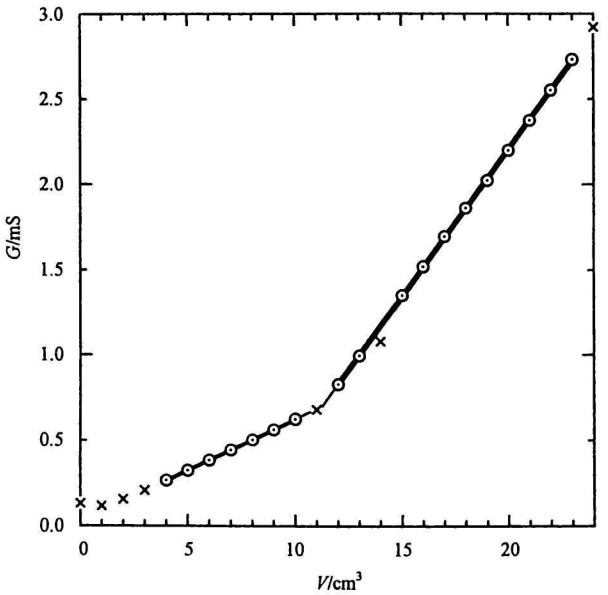

Fig. 1. Conductometric neutralization titration – Example 1. Eliminated points are denoted with × , included points are denoted with ⊙, straight lines are denoted with bold lines, confidence belts are denoted with normal lines.

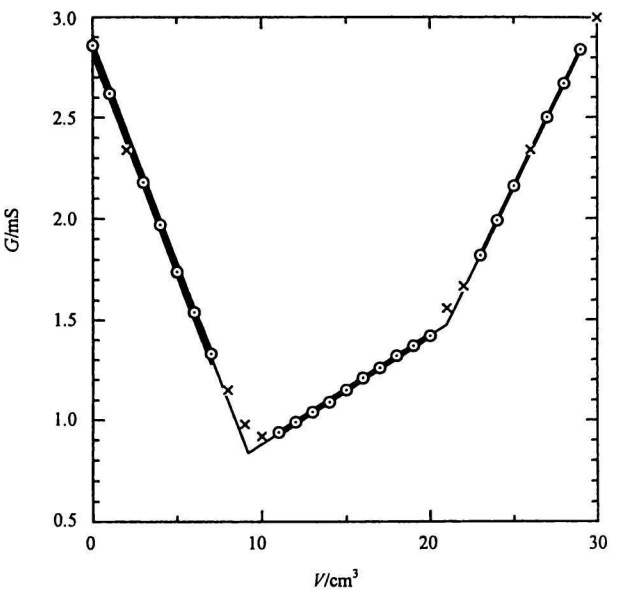

Fig. 2. Conductometric neutralization titration – Example 2. Description as in Fig. 1.

$$\{G\} = 0.0278 + 0.05964\,\{V\}$$

Calculation accuracy (±): 0.0022 0.00030

Regression straight line equation is calculated for $\{V\}$ value from the interval from: 4.00 to: 10.00

2. Regression straight line equation:

$$\{G\} = -1.239 + 0.1723\,\{V\}$$

Calculation accuracy (±): 0.024 0.0013

Regression straight line equation is calculated for $\{V\}$ value from the interval from: 12.00 to: 23.00

Coordinates of intersection of regression straight lines

$\{V\} = 11.24 \quad \{G\} = 0.698$

The first four points were eliminated as outlying values from the linear part by OK-LIN program, which is in pure harmony with the theory of conductometric titration of weak CH_3COOH by volumetric NaOH solution. Even the point close to the equivalence value is outlying. From the second linear part of volumetric curve two points were eliminated, which are outlying with respect to the remaining exactly measured values.

Example 2. Conductometric titration of the mixture of HCl and CH_3COOH by solution of 0.2 M-NaOH (Fig. 2)

1. Regression straight line equation:

$$\{G\} = 2.845 - 0.2182\,\{V\}$$

Calculation accuracy (±): 0.022 0.0050

Regression straight line equation is calculated for values $\{V\}$ from the interval from: 0.00 to: 7.00

2. Regression straight line equation:

$$\{G\} = 0.338 + 0.0542\,\{V\}$$

Calculation accuracy (±): 0.018 0.0011

Regression straight line equation is calculated for values $\{V\}$ from the interval from: 11.00 to: 20.00

3. Regression straight line equation:

$$\{G\} = -2.08999999987 + 0.16999999951\,\{V\}$$

Calculation accuracy (±):

0.00000000017 0.00000000066

Regression straight line equation is calculated for values $\{V\}$ from the interval from: 23.00 to: 29.00

Coordinates of intersections of regression straight lines

$\{V\} = 9.20 \quad \{G\} = 0.837$

$\{V\} = 20.98 \quad \{G\} = 1.476$

Conductometric titration of mixture of strong acid (HCl) with weak acid (CH_3COOH) by volumetric NaOH solution is an example of functional dependence being characterized by three linear parts, two points of intersection. $\{V\}$-Coordinate of the first point of intersection corresponds to the end of titration of strong acid, while $\{V\}$-coordinate of the second point of intersection corresponds to the end of titration of weak acid. This case was solved from entered data fully by the OK-LIN program without any experimentalist cooperation.

Example 3. Photometric titration [5] of Cu^{2+} and Bi^{3+} by standard solution of 0.1 M-EDTA; $\lambda = 745$ nm (Fig. 3)

1. Regression straight line equation:

$$\{A\} = 0.01150 + 0.00000322\,\{V\}$$

Calculation accuracy (±): 0.00021 0.00000040

Regression straight line equation is calculated for values $\{V\}$ from the interval from: 0 to: 900

2. Regression straight line equation:

$$\{A\} = -0.0190 + 0.0000345\,\{V\}$$

Calculation accuracy (±): 0.0030 0.0000021

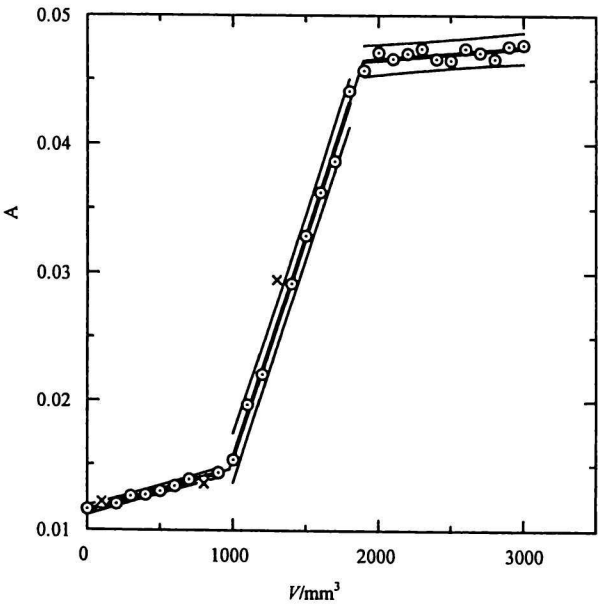

Fig. 3. Photometric titration – Example 3. Description as in Fig. 1.

Regression straight line equation is calculated for values $\{V\}$ from the interval from: 1000 to: 1800

3. Regression straight line equation:

$$\{A\} = 0.0446 + 0.00000094\,\{V\}$$

Calculation accuracy (±): 0.0022 0.00000089

Regression straight line equation is calculated for values $\{V\}$ from the interval from: 1900 to: 3000

Coordinates of intersections of regression straight lines

$\{V\} = 973 \quad \{A\} = 0.0146$

$\{V\} = 1894 \quad \{A\} = 0.0464$

At this case $\{V\}$-coordinate of the first point of intersection corresponds with the end of titration of Bi^{3+}, $\{V\}$-coordinate of the second point corresponds with the end of titration of Cu^{2+} Measured values are in comparison with the previous examples afflicted with superior points diffusion along the regression straight line. This reality is illustrated by the confidence belts in the graph, which are not with respect to the exactly measured values noticeable in previous examples.

CONCLUSION

Presented algorithm of determination of coordinates of point of intersection of linear parts of functional dependence arises thoroughly from the principles of graphical processing of measured data. This approach of experimental data evaluation even corresponds with the used method of the least squares.

Competence of applications of graphical way of data processing was authenticated during several decades as well as the least-squares method applications, because the gained results showed high-level reliability. Graphical methods put, however, higher demand on working experience at data processing. At the application of the graphical processing of measured data as well as application of the method of the least squares outlying values were determined in a subjective way.

Not even standard statistical programs (Adstat, Statgraphic, and the like) used for evaluation of experimental data so far enabled in a quite single-valued and objective way to determine which data can be included to linear segment and which data must be eliminated. In that way after processing of one file of experimental data two workers gained different results many times.

Statistical treatment of experimental data with exploitation of the OK-LIN program eliminates the above-mentioned disadvantages. Priority of the published process does not consist only in processing speed of experimental data and in obtaining numerically exact results in comparison with graphical evaluation, but above all in objectivity of determination of remote values. The proper treatment of engaged data is fully realized by the OK-LIN program and it does not require other experimentalist's intervention.

Usage of the OK-LIN program is not limited only to determination of the titration's end value, but it can be generally used in other cases as well, when the subject of evaluation is determination of coordinates of point of intersection of linear parts of functional dependence.

REFERENCES

1. Holzbecher, Z., Churáček, J. *et al.*, *Analytická chemie.* (Analytical Chemistry.) Nakladatelství technické literatury (Publishers of Technical Literature), Prague, 1987.
2. Eckschlager, K., *Grafické metody v analytické chemii.* (Graphical Methods in Analytical Chemistry.) Nakladatelství technické literatury (Publishers of Technical Literature), Prague, 1966.
3. Eckschlager, K., *Chyby chemických rozborů.* (Errors in Chemical Analyses.) Nakladatelství technické literatury (Publishers of Technical Literature), Prague, 1961.
4. Jehlička, V. and Mach, V., *Collect. Czech. Chem. Commun. 60*, 2064 (1995).
5. Ortiz-Fernández, M. C. and Herrero-Gutiérez, A., *Chemom. Intell. Lab. Syst. 27*, 231 (1995).
6. Rousseenn, P. J. and Leroy, A. M., *Robust Regression and Outlier Detection.* Wiley, New York, 1997.
7. Jehlička, V. and Mach, V., *Scientific Papers of the University of Pardubice, Series D*, Vol. *1*, 3 (1996).