DETERMINATION OF ESTIMATES OF PARAMETERS OF CALIBRATION REGRESSION STRAIGHT LINE WITH OBJECTIVE ELIMINATION OF REMOTE MEASUREMENTS

Vladimir JEHLICKA[a] and Vladimir MACH[b]

[a] *Department of Mathematics,*
University of Pardubice, 530 09 Pardubice, Czech Republic
[b] *Department of Analytical Chemistry,*
University of Pardubice, 530 09 Pardubice, Czech Republic

Received June 24, 1994
Accepted July 26, 1995

An alternative algorithm of calculation of estimates of parameters of calibration regression straight line with elimination of remote measurements has been elaborated for a function $f(x)$ composed of linear-nonlinear or nonlinear-linear parts or, as the case may be, of nonlinear-linear-nonlinear parts. This algorithm makes it possible to objectively eliminate remote measurements and determine the interval $\langle x_1;x_2\rangle$ in which is located the linear part of dependence of output measured quantity $y = f(x)$ with normal distribution $N(f(x), \sigma^2)$ on the input, i.e., independent variable x. For the procedure used for testing of remoteness of experimental points, a relation has been derived for calculation of the critical value of deviation of the point tested. The algorithm is finished by the calculation of parameters of the corresponding regression straight line and other statistical characteristics. On the basis of the algorithm suggested, a program has been assembled whose reliability was verified on a series of both model and practical examples.

Many experimental dependences between two variables have a linear course at least within a limited region $\langle x_1;x_2\rangle$ of values of the independent variable x, whereas outside this region the course of dependence is nonlinear. This behaviour is typical of dependences of absorbance values, polarographic wave heights, conductances, refractions, optical rotations etc. upon concentration of the component determined.

In practice[1–3], the term “linear dependence” is used when the ratio of increments Δy of dependent variable to the increment Δx of independent variable assumes values which are randomly spread about the mean value. If the ratio $\Delta y/\Delta x$ exhibits a certain trend, the dependence is nonlinear.

The measured experimental values loaded with random errors can be approximated by a curve characterized, e.g., by a polynomial of the n-th degree ($y = b_0 + b_1x + b_2x^2 + \ldots + b_nx^n$). The polynomial of the first degree ($y = b_0 + b_1x$, where $b_1 \neq 0$) corresponds to the equation of regression straight line[1,2] which represents the optimum fit of the linear dependence.

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:

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

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

The equation of regression straight line determined after previous elimination of remote values can be used as a calibration straight line. The task given used to be solved subjectively in graphical way[1] whereas at present it is solved numerically by regression, most often by the least squares treatment[2,4,5]. Irrespective of the model used, the regression by the least squares method involves the minimizing of vector

$$\Delta \boldsymbol{y} = \boldsymbol{e} = \boldsymbol{y} - \boldsymbol{y}_{\mathrm{p}} \ , \qquad (2)$$

for the straight line $\Delta \boldsymbol{y} = \boldsymbol{y} - (b_0 + b_1 \boldsymbol{x})$ with the distribution $N(0, (1 + 1/n + (x - \overline{x})^2 \sigma^2 / \Sigma(x_i - \overline{x}))^2$ where σ^2 is the spread, $\boldsymbol{e}$ is the vector of residua, i.e., vector $\Delta \boldsymbol{y}$ of differences between the vector $\boldsymbol{y}$ of measured values y_i of dependent variable y and vector $\boldsymbol{y}_{\mathrm{p}}$ of predicted values calculated from the regression $(b_0 + b_1 x_i)$

In Euclidean space, the length of vector $\boldsymbol{e}$ can be expressed by the relation

$$D = \sqrt{\sum_{i=1}^{n} e_i^2} \ , \qquad (3)$$

where $\Sigma e_i^2 = \Sigma \boldsymbol{e}_i \ . \ \boldsymbol{e}_i$. The square of length of vector $\boldsymbol{e}$ is numerically identical with the criterion condition

$$D^2 = U(b) = \sum_{i=1}^{n} \left[y_i - \sum_{j=0}^{m} x_i^j \, b_j \right]^2 \ . \qquad (4)$$

The vector $\boldsymbol{b}$ of estimates b_i of parameters β_i can be written in matrix form

$$\boldsymbol{b} = (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y} \qquad (5)$$

$$\boldsymbol{y}_{\mathrm{p}} = \boldsymbol{X} \ (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X})^{-1} \ \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y} = \boldsymbol{H} \boldsymbol{y} \ , \qquad (6)$$

where $\boldsymbol{X}$ is a column matrix of elements of independent variable

$$x_i \ (i \in \langle 1;n \rangle) \tag{7}$$

and $\boldsymbol{H}$ is the projection matrix projecting any vector $\boldsymbol{v}$ into plane L.

A number of methods have been suggested to impart objectivity to the procedure of verification of linearity and elimination of remote points. For a long time, predominantly the classic procedures were used, i.e., the *F*-test given by, e.g., Doerffel and Eckschlager[6,7], Grubbs' test[3,8] and other classic methods of statistical analysis of data[5,7,9]. However, these procedures often fail to reflect the real data. That is why it has been considerably advantageous to adopt the regression diagnostics involving identification of influential points and multicollinearity[10], suggestion of model (with the transformation[11]), verification of presumptions for estimate of parameters[12], and methods of preanalysis of individual variables[13] inclusive of the hidden ones[11].

Identification of Gross Errors

1. Deviating observations[10,13]. For identification of these errors the analysis of residua has been introduced. The following terms are defined for the statistical analysis of residua:

classic residua

$$e_i = y_i - \sum_{j=0}^{m} x_i^j \, b_j \tag{8}$$

normalized residua[13]

$$e_{\mathrm{N},i} = e_i/\sigma \tag{9}$$

standardized residua[13]

$$e_{\mathrm{S},i} = \frac{e_i}{\sigma \sqrt{1 - H_{ii}}} \ , \tag{10}$$

where H_{ii} are diagonal elements of projection matrix $\boldsymbol{H}$

Jacknife residua[13]

$$e_{\mathrm{J},i} = e_{\mathrm{S},i} \sqrt{\frac{n - m - 1}{n - m - e_{\mathrm{S},i}^2}} \ , \tag{11}$$

where n and m are numbers of rows and columns, respectively, in the matrix $\boldsymbol{X}$ of independent variables x

predicted residua

$$e_{\mathrm{P},i} = y_i - \sum_{j=0}^{m} x_i^j \, b_j \, (i) = \frac{e_i}{1 - H_{ii}} \ , \tag{12}$$

where $b_j(i)$ are the estimates calculated by the least squares treatment using all the points except the i-th one

recursive residua[13]

$$e_{\mathrm{R},i} = \frac{(y_i - \boldsymbol{x}_i b_{i-1})}{\sqrt{1 + \boldsymbol{x}_i \, (\boldsymbol{X}_{i-1}^{\mathrm{T}} \boldsymbol{X}_{i-1})^{-1} \boldsymbol{x}_i^{\mathrm{T}}}} \ . \tag{13}$$

Out of them the Jacknife and predicted residua have the highest information content according to some authors[13].

2. Extremes[13]. For identification of these errors the analysis of diagonal elements of the projection matrix $\boldsymbol{H}_{ii}$ has been introduced. The more remote the point x_i is from the centroid of other points, the more influential it will appear, and its H_{ii} wll increase as well:

$$H_{ii} = x_i \, (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X})^{-1} \boldsymbol{x}_i^{\mathrm{T}} \ . \tag{14}$$

Identification of Both Gross Errors and Extremes

1. The graphical identification[13] of influential points. The dependences following from the numerically evaluated characteristics have been suggested for the graphical identification[13]:

graph of predicted residua (dependence of e_i vs $e_{\mathrm{P},i}$)
Williams' graph ($e_{\mathrm{J},i} - H_{ii}$)
Pregibon's graph ($e_{\mathrm{N},i}{}^2 - H_{ii}$)
McCulloh–Metter graph ($\ln e_{\mathrm{S},i}{}^2 - \ln \{H_{ii}/[m(1 - H_{ii})]\}$)
L–R graph[14] ($e_{\mathrm{N},i}{}^2 - H_{ii}$)
index graphs (dependence of e_i, $e_{\mathrm{J},i}$, H_{ii}, b on index i)
Rankit graphs (e_i, $e_{\mathrm{J},i} - u_p$).

2. Other characteristics of deviating points. These include the so-called distances expressing the relative influence of the given point on all the estimates of the parameter:

Cook's distance[13] D_i
Atkinson's distance[13] A_i
Andrews–Pregibon's distance[15]
Cook–Weisberg's distance[16]
others[17].

A number of characteristics are used to express the fit of theoretical model to measured data (the suitability of model). The measure of relative difference between the determined regression model and the data is characterized by the multiple correlation coefficient R and/or coefficient of determination[13] R^2. Utts[18] used the tests of identity of linear model. Other tests of linearity include the mean quadratic error prediction[13] (MEP), predicted coefficient of determination[13], Akaik information criterion[13], and other specific criteria.

THE BASIS OF THE TREATMENT SUGGESTED

The aim of the present paper is to objectively determine the parameters of regression straight line with objective elimination of remote values. Using the least squares treatment a set of five neighbouring experimental points is determined with the lowest value of standard deviation $s_{x,y}$ which does not include any remote point. A set of five points is considered sufficient[1,2,6] for a reliable determination of coefficients of the regression straight line and other statistical characteristics in current analyses.

With regard to the critical value (see the below-given relation (*25*) for the critical value of the point tested) determined from the values of the set of five experimental points characterized by the lowest value of $s_{x,y}$, the absolute value of residuum (i.e., deviations of the neighbouring point from the regression straight line) is tested. If the absolute value of deviation of the point tested is smaller than the critical value caculated from the points included, the tested point is involved into the set of included values. If the absolute value of deviation of the point tested is equal to or greater than the calculated critical value, then the corresponding point is eliminated as a remote one. This procedure is continued to test gradually the deviations of all experimental points with respect to the critical value determined from the points already included, and, if possible, this is done alternately from the two sides. Hence the point tested does not distort the value of criterion used for its evaluation.

As the number of included points increases, the criterion for evaluation of remoteness of measurement becomes more strict. Therefore, one determines the point with the maximum deviation from the regression straight line given by all the included points. This point is again tested with regard to all the remaining included points. If it is remote, it is eliminated, and the testing is repeated with the remaining points. In this way are determined the resulting estimates of parameters of calibration regression straight line with objective elimination of remote measurements. The number of remaining included points must not decrease below 5.

PROGRAM TREATMENT OF ALGORITHM

The algorithm used for treating the above-mentioned problem was elaborated into a program called OK-LIN. The program is written in Turbo Pascal language with the help of Turbo Vision library and is applicable to work on any PC operating in DOS.

The input data of the program are pairs of x_i, y_i coordinates of measured points and a short text (70 characters at the most) serving for a description of the data set. The data are input from the keyboard or from a data set on disc. The program enables any deliberate editing of input data and their storing in data sets on disc.

The output of the program is the interval estimate of parameters of calibration regression straight line with other relevant statistical characteristics. The program makes it possibl to determine – from the calibration line – the value of x for a given value of y. The output also includes a graph involving all the points measured with a denotation stating whether or not they had been included in the respective linear section. Moreover the graph contains the calibration straight line with the corresponding bands of reliability.

A floppy disc with the program is available if ordered from the first author.

MODEL AND PRACTICAL EXAMPLES

Example No. 1

The calibration relation between height of polarographic peak (h) and volume of 1 . 10^{-4} mol l^{-1} Pb^{2+} (V) added to 20 ml 0.1 M NaOH. Measured with a carbon paste electrode. The eliminated points are denoted with + (see Fig. 1).

Input:

V, ml	+ 0	50	100	150	200	250	300	350	+ 400	450	500
h, cm	1.0	14.5	34.0	52.5	79.0	100.0	117.0	135.5	169.5	178.5	204.0

Output:
correlation coefficient = 0.99927
standard deviation about the regression straight line = 2.6212
equation of regression straight line: $h = -7.2 + 0.417$ V
calculation accuracy of parameters: ±4.3, ±0.014.

Example No. 2

Calibration dependence of the height of polarographic peak (h) on volume of 2.09 . 10^{-3} mol l^{-1} Ti^{4+} (V) in oxalic acid as the electrolyte with addition of Chelaton III and potassium bromate as the catalyst. Measured by means of difference pulse polarography. The eliminated points are denoted with + (see Fig. 2).

Input:

V, ml	0.000	0.020	0.040	0.060	0.080	0.100	+ 0.120
h, cm	0.00	1.10	2.10	3.20	4.16	5.20	6.40

V, ml	0.140	+ 0.160	+ 0.180	+ 0.200	+ 0.220	+ 0.240	+ 0.260
h, ml	7.30	8.20	9.20	9.90	10.60	10.90	11.50

```
Output:
correlation coefficient = 0.99992
standard deviation about the regression straight line = 0.034527
equation of the regression straight line: h = 0.033 + 51.88 V
calculation accuracy of parameters: ±0.058, ±0.75.
```

COMPARISON OF CRITERIA USED

In the references various criteria are recommended for testing the linearity of function and various criteria for testing the remoteness of experimental points. None of the diagnostics, however, is unambiguously conclusive by itself, without simultaneously considering the other statistical characteristics.

The criterion suggested in the present paper for testing of remoteness of points combines important advantages of individual other diagnostic tools used. The examples tested show a good agreement (reliability) of the criterion suggested with the diagnostic used most often at present, viz the Jacknife residuum. In addition, the algorithm involves another advantage used in the diagnostic of predicted residuum, viz the idea of

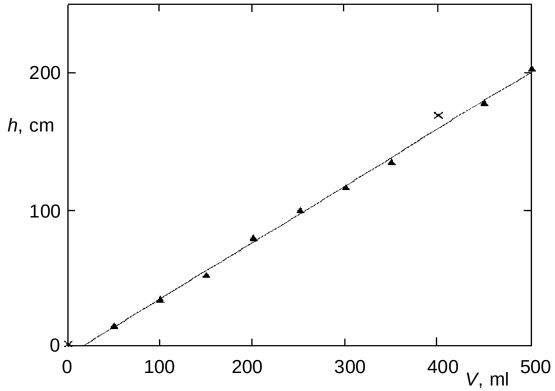

FIG. 1
Calibration dependence of height of polarographic peak (h) on volume of 1 . 10^{-4} mol l^{-1} Pb^{2+} (V) added into 20 ml 0.1 M NaOH. Measured with a carbon paste electrode. Eliminated points are denoted with ×

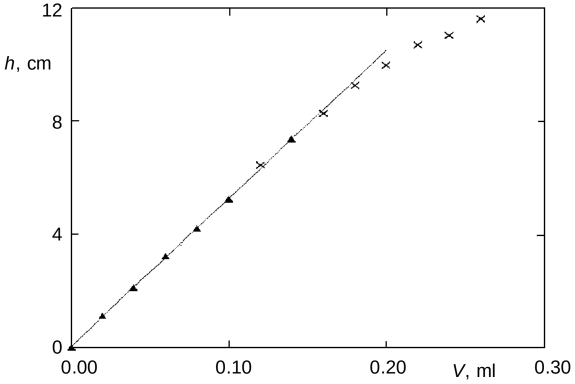

FIG. 2
Calibration dependence of height of polarographic peak (h) on volume of 2.09 . 10^{-3} mol l^{-1} Ti^{4+} (V) in oxalic acid as electrolyte with added Chelaton III and potassium bromate as catalyst. Measured by differential pulse polarography. Eliminated points are denoted with ×

testing of the pont as one not included in the set, wherefrom the critical limits are determined to decide about an inclusion. With regard to the reasons mentioned (without neglecting the monitoring of "trend", i.e., nonlinearity of data) the algorithm elaborated appears to be more versatile and unambiguously applicable than the procedures used earlier.

APPENDIX

Derivation of Relation for Critical Value of Tested Point

In all the considerations the validity of simple linear model (*1*) is presumed. 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 nonlinear section it is the value of random quantity with the spread N $(f(x); \sigma^2)$. 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 $\Delta y = e = y - (b_0 + b_1 x)$.

Impartial estimates b_0 and b_1 of unknown parameters β_0 and β_1 are obtained by the least squares treatment. For the estimate of unknown spread σ^2 we will choose the characteristic called residual spread

$$s^2_{y,x} = [\sum (y_i - b_0 - b_1 x_i)^2] \, / \, (n-2) \ . \tag{15}$$

The random quantity

$$b_1 = [n \, \Sigma x_i y_i - \Sigma x_i \, \Sigma y_i] \, / \, [n \, \Sigma x_i^2 - (\Sigma x_i)^2] = \Sigma \left\{ (x_i - \overline{x}) \, / \, [\Sigma (x_i - \overline{x})^2] \right\} y_i = \Sigma e_i y_i \tag{16}$$

represents a linear combination of independent random quantities Y_i with normal distribution (according to presumption), hence b_1 has a normal distribution. It can easily be proved that for the mean value E of parameter b and for the spread D of parameter b it is

$$E(b) = \beta$$

$$D(b) = \sigma^2 / \, \Sigma (x_i - \overline{x})^2 = \sigma^2 / \, [\Sigma x_i^2 - (\Sigma x_i)^2 / n] \ . \tag{17}$$

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

$$b_0 + b_1 x = \overline{y} + b(x - \overline{x}) \ , \qquad (18)$$

it can be seen that it is a sum of two independent normally distributed quantities, hence it has normal distribution. Numerical characteristics of this quantity are

$$E(b_0 + b_1 x) = E(\overline{y} + b(x - \overline{x})) = \beta_0 + \beta_1 \overline{x} + \beta_1 x - \beta_1 \overline{x} = \beta_0 + \beta_1 x \qquad (19)$$

$$D(b_0 + b_1 x) = D(\overline{y} + b(x - \overline{x})) = D(\overline{Y}) + (x - \overline{x})^2 \, D(b) = \sigma^2/n + [(x - \overline{x})^2 \, \sigma^2] \, / \, \Sigma(x_i - \overline{x})^2 =$$
$$= \sigma^2 \left\{ 1/n + (x - \overline{x})^2 / \, [\Sigma(x_i - \overline{x})^2] \right\} \ . \qquad (20)$$

We know now that the random quantity $Y = b_0 + b_1 x$ has normal distribution for any x

$$N(b_0 + b_1 x, \ \sigma^2 \left\{ 1/n + [(x - \overline{x})^2] / \, \Sigma(x_i - \overline{x})^2 \right\}) \ . \qquad (21)$$

The deviation $\Delta y = e = y - (b_0 + b_1 x)$ is a difference of two independent random quantities with normal distribution and, hence, has normal distribution.

By calculation it can be found that

$$E\ (\Delta y) = E(y) - E\ (b_0 + b_1 x) = \beta_0 + \beta_1 x - (\beta_0 + \beta_1 x) = 0 \qquad (22)$$

and

$$D\ (\Delta y) = D(y) + D(b_0 + b_1 x) = \sigma^2 + \sigma^2 \left\{ 1/n + [(x - \overline{x})^2] \, / \, \Sigma(x_i - \overline{x})^2 \right\} =$$
$$= \sigma^2 \left\{ 1 + 1/n + [(x - \overline{x})^2] \, / \, \Sigma(x_i - \overline{x})^2 \right\} \ . \qquad (23)$$

From a summary of the above-given results it follows that the random quantity $\Delta y = y - (b_0 + b_1 x)$ has the distribution $N(0,\sigma^2\{1 + 1/n + [(x - \overline{x})^2]/\Sigma(x_i - \overline{x})^2\})$, and the corresponding normalized quantity $\Delta y/(\sigma\{1 + 1/n + [(x - \overline{x})^2]/\Sigma(x_i - \overline{x})^2\}^{1/2})$ has normal distribution $N(0,1)$.

As the random quantity $\{(n-2)S_{y,x}{}^2\}/\sigma^2$ has χ^2 distributions with $n-2$ degrees of freedom, the random quantity (*24*), according to the definition of *t*-distribution, has a *t*-distribution of $n-2$ degrees of freedom.

$$\left\{\Delta y/(\sigma\left\{1+1/n+[(x-\overline{x})^2]\,/\,\Sigma(x_i-\overline{x})^2\right\}^{1/2})\right\}/\left\{[(n-2)\;s_{y,x}^2]\,/\,[\sigma^2(n-2)]\right\}^{1/2}=$$

$$=\Delta y/\left\{s_{y,x}\left\{1+1/n+[(x-\overline{x})^2]\,/\,\Sigma(x_i-\overline{x})^2\right\}^{1/2}\right\} \qquad (24)$$

From this relation we determine the critical value of deviation for the point tested (x_t, y_t)

$$\Delta y_{t,\mathrm{crit}}=t_{n-2,\alpha}s_{y,x}\left\{1+1/n+[(x-\overline{x})^2]\,/\,\Sigma(x_i-\overline{x})^2\right\}^{1/2}\ , \qquad (25)$$

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

The authors are indebted to Prof. Stanislav Kolda (Department of Mathematics, University of Pardubice) for his interest and valuable suggestions.

REFERENCES

1. Eckschlager K.: *Graficke metody v analyticke chemii.* SNTL, Praha 1966.
2. Eckschlager K., Horsak I., Kodejs Z.: *Vyhodnocovani analytickych vysledku a metod.* SNTL, Praha 1980.
3. Holzbecher Z., Churacek J. et al.: *Analyticka chemie.* SNTL, Praha 1987.
4. Eckschlager K.: *Chyby chemickych rozboru.* SNTL, Praha 1961.
5. Kolda S., Krajnakova D., Kimla A.: *Matematika pro chemiky II.* SNTL/ALFA, Praha 1990.
6. Doerffel K., Eckschlager K.: *Optimalni postup chemicke analyzy.* SNTL, Praha 1988.
7. Jancar L., Langova M.: Chem. Listy *86*, 20 (1992).
8. Vlacil F. et al.: *Priklady z chemicke a instrumentalni analyzy.* SNTL, Praha 1983.
9. Anscombe F. J.: Am. Statist. *27*, 17 (1973).
10. Belsey D. A., Kuh E., Welsh R. E.: *Regression Diagnostics.* Wiley, New York 1980.
11. Atkinson A. C.: *Plot, Transformation, Regression.* Clarendon Press, Oxford 1986.
12. Weisberg S.: Technometrics *25*, 219 (1983).
13. Meloun M., Militky J.: *Chemometrie – zpracovani experimentalnich dat na IBM-PC.* SNTL, Praha 1990.
14. Gray J. B.: *Prac. Stat. Comput. Sect.*, p. 159. ASA, Washington 1983.
15. Chattarje S., Hadi A. S.: Statist. Sci. *1*, 379 (1986).
16. Cook R. D., Weisberg S.: *Residuals and Influence in Regression.* Chapman and Hall, New York 1982.
17. Gorman M. A., Myers R. M.: Commun. Statist. *16*, 771 (1987).
18. Utts J.: Commun. Statist. *11*, 2801 (1982).