A&QT-R 2002 (THETA 13)
International Conference on Automation, Quality and Testing, Robotics
May 23-25, 2002, Cluj-Napoca, Romania

CORRELATIONS AND REGRESSIONS WITH MATHLAB

Lorentz Jäntschi **Mihaela Ungureşan**

Technical University of Cluj-Napoca, Romania
Department of Chemistry; Tel/Fax: 40-64-415051

Abstract. MATHLab is very useful tool for data processing, especially for numerical processing of experimental data. Statistical processing of data is first one step for finding links between measured values and/or theoretical results. Most frequent used regression models are linear-based. Value of link strength is, most frequently given by correlation coefficient r.

A MATHLab computer program is implemented. The program presented demonstrates the power of MATHLab in working with statistical processing, and can be considered as an example for future applications. The program draws an exportable figure of fitted data and regression curve and calculates the correlation coefficient r and sum of residues.

Key-words: MATHLab, regression, correlation coefficient.

1. Introduction

The term correlation has introduced by Galton in 1888 [1]. The correlation coefficient has introduced for the first time by Pearson in 1896 [2]. Since then these notions are the basis of experimental statistics.

Correlation is a measure of the relation between 2 or more variables. The correlation coefficients have values between –1.00 and +1.00. The value –1.00 represents a perfect negative correlation and the value +1.00 represents a perfect positive value. A value of 0.00 represents a lack of correlation.

The mostly used type of correlation coefficient is Pearson r, also called linear correlation or product-moment. The value of correlation (that is the value of the correlation coefficient) does not depend on the measurements units used in measuring the two variables. Let there be the series of data X (X_1,..., X_n) and Y (Y_1,...,Y_n). Analytically, they are defined:

$$\nu_{X,Y} = M(XY) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \,,\ \mu_{X,Y} = \nu_{X,Y} - M(X)M(Y)$$

where ν is the moment of order 2 and so it is the average of products $X_i \cdot Y_i$ and μ is the centered moment of order 2 or a correlation of the two sizes given on the numerical rows considered and M is the average operator.

Also, the correlation coefficient is given by the equation:

$$r(X,Y) = \frac{\mu_{X,Y}}{\sqrt{D^2(X)D^2(Y)}} = \frac{\mu_{X,Y}}{\sigma(X)\sigma(Y)}$$

where D is the dispersion operator ($D^2(X) = M(X^2)-M^2(X)$).

If between the characteristics Y, X_1, X_2,..., X_n studied simultaneously for a certain type of phenomenon's we can notice a very tight connection, close to a functional one, we can apply the regression analyses (similar to empirical modeling, curve fitting or forecasting). This allows us to find a regression equation, a function that makes the calculus for one of the mentioned characteristics easier.

Due to the fortuitous errors, which practically appear always, the connection between the factors that affect the analytical signal is statistical (stochastically). That's why we try to detect through interpolation procedures of the values Y from the distribution $Y(X_1,X_2,..., X_p)$, obtaining a proximity to the functional connection (ideal) to the statistical one (real).

Taking into consideration the mathematical shape of the model we can distinguish linear models and nonlinear models. After the number of independent variables involved we have models monovariable (Y=Y(X)) and models multivariable ($Y=Y(X_1,X_2,...,X_n)$). Even in the case of linear regression there is a large development of the concept of linear dependency, this one evolving trough a linearisable dependency. According to this concept, an equation of regression is linear if the functional dependency between the considered variables can be brought to a linear form [3]. According to this principle, the equations of regressions:

$$y = a\cdot\log(x)+b;\ y = a\cdot\log(\log(x))+b;\ y = a\cdot(1/x)+b;\ y = a\cdot e^{x}+b$$

are linearised dependencies and can be associated to the linear model of regression: $y = a\cdot z+b$ where the new independent variable z is obtained $z = \log(x)$ or $z = \log(\log(x))$ or $z = 1/x$ or $z = e^{x}$. Also another extension of the linear model of regression is obtained when the error factor takes action on both variables involved in regress in regression. [4]

Linear one-dimensional regression frequently applied in analytical practical work considers as valid, for the experimental data, the model:

$$y = \hat{y} + \varepsilon;\ \hat{y} = b_0 + b_1 \cdot x$$

where x, y are the characteristics measured by the annalist, $\hat{y}$ is the characteristic estimated as model for y, b_0 and b_1 are coefficients which are estimated with the help of the model and ε is the error. The most well known model of estimation of the parameters is fundamental by Kolmogorov "minimizing the risk defined as the average of the losing squared function" [5] known as the method of the smallest squares:

$$K(X,Y,B) = \sum(\hat{y} - y)^2 = \sum(b_0 + b_1 \cdot x - y)^2$$

where X,Y,B are the vectors column of the independent variable, of the dependent variable and of the coefficients.

A series of other papers [6-8] it approach, different sides of the estimation way based on the method of the smallest squares. Due to the fact that acquisition of data is now present in the laboratories of experimental sciences, the methods based on linear algebra and multi linear statistics are applied in quantitative analyses multicomponent [9-11]. The method has been also applied in determining the functional dependencies of the chromatographically capacity factors (log k') from the molecular parameters of the substances separated [12]. The method has a great future because of the diversification and improving the detectors like the quality of resolution and enlarging the sensibility.

Once established the coefficients and errors that affect the results of the signals based on the equations of regression, we take the way back, the equations of regression becoming equations of calibration (the multidimensional correspondent of the calibration curve in two dimensions). Between those two aspects there are elements that need to be clarified in practice. This method of data analyses has been lately the object of a significant number of books [13-15] and reviews [16-18].

The development of software industry brought an explosion on the market of programs specialized on statistic manufacturing. Most of these programs have implemented routines for the calculus of regressions of different natures. We can mention [19-24].

2. Using MATHLab in correlations and regressions

MATHLab has specialized functions on different categories of operations. The version that we used (4.0) has categories for color control, the analyses of data and functions of transformation Fourier, mathematical basic functions, basic matrix and the manipulation of matrix, functions of functions at numerical nonlinear methods, general graphical functions, functions of input and output, program language and repairs, functions on matrix of linear algebra, operators and special symbols, graphics 2D, graphics 3D, polynomial type and interpolation functions, functions for processing the sound, additional functions on matrix, specialized mathematical functions, functions on rows of characters.

MATHLab has the possibility of direct work with the help of the commands or making scripts, which allow the grouping of instructions in a similar way to the program languages. The syntax in writing is closed to the syntax of the language C. There is the possibility of making scripts that contain a code, which is loaded and executed in the moment of the script's executing.

The implemented program makes correlations and regressions on a asset of data, and has the possibility of defining the equation of regression interactively, using the facility of transmission of code at the up mentioned execution. The source code of the program, with comments, is:

```
% M-File: RegLin_4.m; All items UI-Control are in normal units so that
%the user can resize the window; create and memorize the object, the figure
% Create the object axe in the left up corner of the figure
h_f = figure; axes('position',[.07 .5 .86 .4],'box','on');
% Create two frames; the first looks like UI-Objects; the second one: setting the properties
h_frame_1 = uicontrol(h_f, 'Position',[ 0 0 1 0.4 ],'Style','frame', 'Units','normalized');
h_frame_2 = uicontrol(h_f, 'Position',[0.08 0.05 0.84 0.11 ],'Style','frame',... 'Units','normalized');
% Create a callback for "Box"; it determines the value checkbox from h_box;
% it sets in accordance to the current properties for axes; it post a message
% by setting the static row of characters UI-Control memorized in h_status
box_clbk_str = ['boxstatus = get(h_box,"value"); if boxstatus == 0;'...
'set(gca,"box","off"); else; set(gca,"box","on"); end; boxstatus = get(gca,"box");'...
'set(h_status,"string", ["Box is " boxstatus]);'];
% I create check-box-ul, sett the value on 1; initialize the axes.
h_box = uicontrol(h_fig, 'CallBack',box_clbk_str, 'Position',[ 0.84 0.2 0.16 0.07 ], ...
'String','Box', 'Style','checkbox', 'Units','normalized', 'Value',[ 1 ]);
% I create the callback for the check-box "Grid"; this callback determines the value
% of the checkbox of which's handle is stocked in h_grid, and then it uses the function
% of the corresponding grid;  at the end it posts a message setting the row
% of statically characters UI-Control of which's handle is memorized in h_status (created after words)
grid_clbk_str = ['gridstatus = get(h_grid,"value"); if gridstatus == 0; grid off; else;'...
'  grid on; end; gstatus = get(gca,"xgrid"); set(h_status,"string",["Grid e " gstatus]);'];
% I create the check-box grid.
h_grid = uicontrol(h_fig,'CallBack',grid_clbk_str,'Position',[ 0.84 0.3 0.16 0.07 ],...
'String','Grid','Style','checkbox','Units','normalized');
% I define the title and initialize calques, data
title ='Measured data= "o" and Regression = "--" ';
x = [1.1 2.2 8 31 58 71 92 100]; y = [-1.7 -1.95 -3.8 -5.7 -7.3 -8.4 -9.9 -10.3];
% I create the call-back that could make the function anytime the values x and y would be altered
% by the user; there can be errors if the defined functions by the user cannot be fitted.
plot_clbk_str = ['err_ind = 0; eval(["x1 = " get(h_xdata,"string") ";"],"err_ind=1;");'...
```

```
'if err_ind == 0; eval(["p1 = " get(h_ydata,"string") ";"],"err_ind=2;"); end;'...
'if err_ind == 0; y1 = polyval(p1,x1);'...
'xminc = sprintf("cor(x,y)=%2.3f",mean(min(corrcoef(x,y))));'...
'yminc = sprintf("cor(P,y)=%2.3f",mean(min(corrcoef(polyval(p1,x),y))));'...
'zminc = sprintf("err(P,y)=%2.3f",sum(abs(y-polyval(p1,x))));'...
'plot(x,y,"og",x1,y1,"--c"); title(titlu); set(h_rdata ,"string",xminc);'...
'set(h_rdata1,"string",yminc); set(h_rdata2,"string",zminc);'...
'box status = get(h_box,"value"); if box status == 0; set(gca,"box","off");'...
'else; set(gca,"box","on"); end;'...
'gridstatus = get(h_grid,"value"); if gridstatus == 0; grid off; else; grid on; end;'...
'set(h_status,"string","Regression represented");'...
'elseif err_ind == 1; set(h_status,"string","Error in defining x");'...
'elseif err_ind == 2; set(h_status,"string","error in defining y(x)"); end'];
%I create the edit-boxes for the values of x and y; these are used by the previous call -back-ul;
% More, it's been initialized with valid values.
h_ydata = uicontrol(h_f,'CallBack',plot_clbk_str,'Position',[ 0.25 0.2 0.30 0.07 ],...
'String','polyfit(x,y,2)','Style','edit','Units','normalized');
h_xdata = uicontrol(h_f,'CallBack',plot_clbk_str,'Position',[ 0.25 0.3 0.30 0.07 ],...
'String','[0:10:100]','Style','edit','Units','normalized');
h_rdata = uicontrol(h_f,'CallBack',plot_clbk_str,'Position',[ 0.56 0.26 0.28 0.07 ],...
'Style','text','Units','normalized');
h_rdata2 = uicontrol(h_f,'CallBack',plot_clbk_str,'Position',[ 0.56 0.32 0.28 0.07 ],...
'Style','text','Units','normalized');
h_rdata1 = uicontrol(h_f,'CallBack',plot_clbk_str,'Position',[ 0.56 0.2 0.28 0.07 ],...
'Style','text','Units','normalized');
%I create a static object for posting the messages for the user
h_status = uicontrol(h_f,'CallBack','guiplot1("h_uic_12");',...
'Position',[ 0.1 0.07 0.8 0.07 ],'String','State Window','Style','text','Units','normalized');
%Creating the objects text static "x = " and "y(x)="; there is no need of stocking handles as long as
%these objects are never manipulated or interrogated
uicontrol(h_f,'Position',[ 0.08 0.3 0.15 0.07 ],'String','x =',...
'Style','text','Units','normalized');
uicontrol(h_f,'Position',[ 0.08 0.2 0.15 0.07 ],'String',...
'y = P(x) =','Style','text','Units','normalized');
% Initializes the drawing with the initial values x and y through
%The evaluation of the string callback which is the dependency of x from y.
eval(plot_clbk_str);
```

3. Conclusion

The execution of the program makes the posting of a window like in the figures attached, in which we can modify the expression of the function (y = P(x)) or the form of posting (box or grid).

For exemplifying let's consider a correlation study of efficiency time executing of numerical methods that are using a Runge-Kuta-Niström [25] predictor – corrector (Tab.1):

Tab. 1

	1	2	3	4	5	6	7	8
time(s)	1.1	2.2	8.0	31	58	71	92	100
log(err)	-1.7	-1.95	-3.8	-5.7	-7.3	-8.4	-9.9	-10.3

There is an automatic calculus for every defined function the values of the correlation between $\hat{Y} = P(X)$ and X and the value of residues sum given by the formula:

$$\mathrm{err}(\hat{y}, y) = \sum_{i=1}^{N} \left| \hat{y}_i - y_i \right|$$

The resulted image can be after words, with the help of MATHLab, copied in clipboard as image BMP or WMF.

In Fig. 1 are presented fitted dates from Tab.1 with a parabolic regression and program makes the corresponding correlations and sums of residues. In Fig. 2, in same execution, with same dates from Tab. 1 are fitted with a linear regression model and program makes rebuild corresponding correlations and sums of residues.

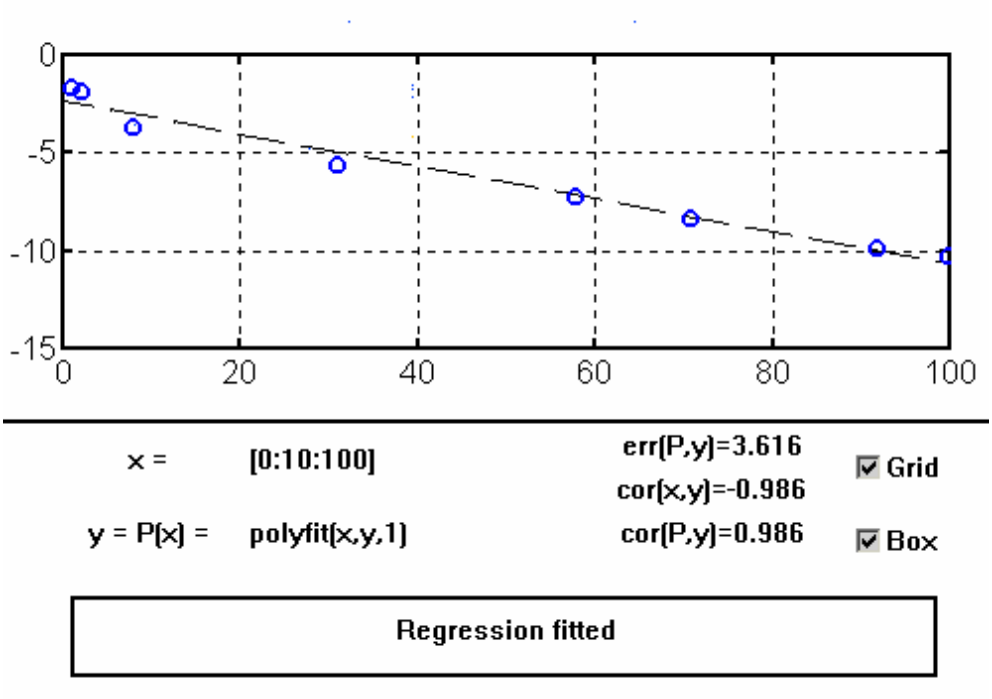

Fig. 1 Linear regression demo of the program

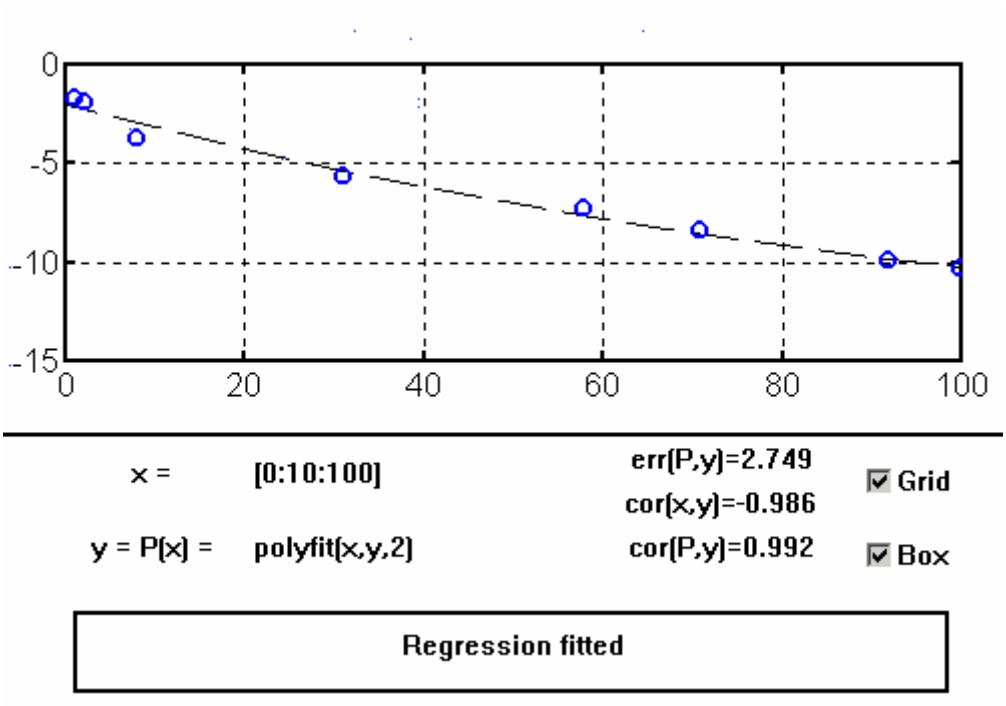

Fig. 2 Parabolic regression demo of the program

References

[1] Galton, F. (1888). Co-relations and their measurement. *Proceedings of the Royal Society of London*, 45, 135-145.
[2] Pearson, K. (1896). Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London*, Ser. A, 187, 253-318.
[3] Naşcu H., Jäntschi L., Hodişan T., Cimpoiu C., Câmpan G (1999). Some Applications of Statistics in Analytical Chemistry, *Reviews in Analytical Chemistry*, 18, 409-456.
[4] Costel Sârbu, Lorentz Jäntschi (1998). Validarea şi evaluarea statisticã a metodelor analitice prin studii comparative (I)Validarea metodelor analitice folosind analiza de regresie, *Revista de chimie*, 49(1), 19-24.
[5] Moritz H. (1980). Advanced Physical Geodesy, H. Wichman Verlag.
[6] Bjerhammar A. (1973). Theory of errors on generalized matrix inverses, Elsevier, Amsterdam-London-N.Y.
[7] Brandin N.V. ş.a. (1974). Osnovî eksperimentalnoi kosmiceskoi balistiki, Moscow.
[8] Tiron M. (1972). Teoria erorilor şi metoda celor mai mici pătrate, Editura Tehnică, Bucureşti.
[9] A. Lorbes, K. Faber, R. Kowalski (1997). Net Analyte Figural Calculation in Multivariate Calibration, *Anal. Chem.*, 69, 1620-1626.
[10] L. Yu, I. Schercter (1997). A Calibration Method Free of Optimum Foelor Number. Selection of a Analnucled Multivariate Analitical Experimental and Theoretical Study, *Anal. Chem.*, 69, 3722-3730.
[11] M.P. Nelson, J.F. Aust, J.A. Dobrowolski, P.G. Verly, M.L. Myrick (1998). Multivariate Optical Computation for Predictive Spectroscopy, *Anal. Chem.*, 70, 73-82.
[12] P.J. Jackson, M.R. Schure, T.P. Weber, P.W. Carr (1997). Intermolecular Interaction Involved in Solute Retention on Carbon Media in Reversed Phase HPLC, *Anal. Chem.*, 69, 416-425.
[13] H. Markus, T. Naes (1989). Multivariate Calibrations, Wiley, N.Y.
[14] Jalliffe I. (1986). Principal Component Analysis, Springer-Verlag, N.Y.
[15] Miller J.C.; Miller J.N.; Ellis Horwood (1984). Statistics for Analytical Chemistry, Limited Publishers, London.
[16] Balke S.T. (1989). *J. Appl. Polym. Sci.*, 43, 5-38.
[17] Kaliszan R. (1994). *Chemom. Intell. Lab. Syst.*, 24, 89-97.
[18] Berridge J.C., Jones P., Roberts A.S. (1991). *J. Pharm. Biomed. Anal.*, 9, 597-604.
[19] Slide; Slide Write Plus for Windows; Advanced Graphics Software Inc.
[20] MathCad; MathSoft Inc.; Collabra Software Inc.
[21] Excell; Microsoft Corporation; Soft Art Dictionary and Program.
[22] Statistics; Statistics for Windows; StatSoft Inc.
[23] Surfer for Windows; Software Package; Golden Software.
[24] MathLab; MathWorks Inc.
[25] Nguyen Huu Cong, Karl Stremhel, Rüdiger Weiner, Helmut Podhaisky, (1999). Runge-Kuta-Niström-type Paralel Block Predictor-Corector Methods, *Advances in Computational Mathematics*, 10, 115-133.