Bul. Ştiinţ. Univ. Baia Mare, Ser. B.
Matematică-Informatică, Vol. XVIII(2002), Nr. 2, 291 - 296

ANREGR - A COMPUTER PACKAGE FOR REGRESSIONAL ANALYSIS

Maria PÂRV

Abstract. This paper presents a computer package for performing regressional analysis, running under Microsoft Windows operating system. The programs in the package have a common user interface for entering data and displaying the results.

MSC Classification: 62J12 (62-04)
Keywords: general linear regression, computer package, graphical user interface.

1. Introduction

When we want to investigate causal relations between a variable (known as *dependent* or *response variable*) and several other variables (known as *independent* or *predictor variables*), we have to build a *mathematical model* in order to express how dependent variable is related to the independent variables. The process of building such a mathematical model is considered as part of *regressional analysis*.

2. Mathematical Background

2.1. Basic Notations and Terms

If we denote by y the dependent variable and by $x_1, x_2, ..., x_n$ the independent variables, then the general form of the mathematical model is:

$$y = \hat{y} + \varepsilon \,, \tag{1}$$

where $\hat{y}$ is the *computed* (*predicted*) *value* of y, and ε is the random error. The computed value $\hat{y}$ is function of independent variables:

$$\hat{y} = f(x_1, x_2, \ldots, x_n) \tag{2}$$

and the equation (2) is called *regression* or *prediction equation*.

Regressional analysis is concerned with relating a response y to a set of independent variables $x_1, x_2, ..., x_n$. Its goal is to build a good model, i.e. a prediction equation of the general form (1) that allows us to *predict* y for given values of $x_1, x_2, ..., x_n$, with a small error of prediction ε. The *quality* of the model is strong related to the absolute value of this error. Usually the sample data are collected in the form of an *observational matrix*, in which the columns represent variables and rows are observations. Considering the matrix in the form (3), with M rows and N columns, every column is considered as a variable, in the sense that its elements have a common significance and are represented using the same unit of measure. Every row is an *observation*: all its elements are related: they contain data collected the same time and space frame:

$$D = \begin{pmatrix} d_{11} & d_{12} & \ldots & d_{1N} \\ d_{21} & d_{22} & \ldots & d_{2N} \\ \ldots & \ldots & \ldots & \ldots \\ d_{M1} & d_{M2} & \ldots & d_{MN} \end{pmatrix}, \tag{3}$$

When such an observational matrix D is available, it can be used to establish several causal relations between its columns. Any column can be considered as a dependent variable; any other column from D is a possible member of the set of independent variables. Sometimes, the form of function f from (2) is known; some other times, the form of f is not known *a priori*: finding it is a trial-and-error process.

With the above assumptions, a general form of the modeling process in regression analysis has the following eight steps (see also [2]): (1) Collect the sample data – the observational matrix D; (2) Establish the dependent variable and the set of independent variables from D columns; (3) Establish the form of the model (i.e. the concrete form of function f in (2)); (4) Use the sample data to estimate the model parameters; (5) Specify the probability distribution of the random error term in (1), and estimate any unknown parameters of this distribution; (6) Statistically check the usefulness of the model; (7) Use the model for prediction, estimation, etc if its usefulness (Step 6) is proved; (8) Repeat steps 2 to 7 for a new model. The rest of this paper refers to steps 2 and 3. The steps 4 through 6 are detailed in every regression analysis textbook and are not discussed here.

2.2. Types of Regression Equations

When the number n of independent variables considered in the model is 1, the independent variable is denoted by x, and the equation (2) becomes $\hat{y} = f(x)$.

The function f in (2) expresses the computed value $\hat{y}$ of the dependent variable y using independent variables $x_1, x_2, \ldots, x_n$. The general equation (2) does not contain any information related to the form of this function. With respect to this criterium, there are additive, multiplicative, and special models. In the case of *additive* models, function f is a polynomial in the variables $x_1, x_2, \ldots, x_n$. *Multiplicative* models the form $\hat{y} = f(x) = b_0 \cdot (b_1 x_1) \cdot (b_2 x_2) \cdot \ldots \cdot (b_n x_n)$, while in the *special* ones the function f contains non-linear terms or factors with respect to the independent variables. An example is the logistic model: $\hat{y} = f(x) = \dfrac{e^{b_0 + b_1 x_1 + \ldots + b_n x_n}}{1 + e^{b_0 + b_1 x_1 + \ldots + b_n x_n}}$.

Many of the non-linear forms of f can be reduced to linear ones by using appropriate transforms, called *linearizations*. For example, the special model with one independent variable $\hat{y} = f(x) = b_0 + b_1 x + b_2 \ln(x) + b_3 e^x + b_4 \sin(x)$, by using the notations: $x_1 = x$, $x_2 = \ln(x)$, $x_3 = e^x$, $x_4 = \sin(x)$, becomes a linear one : $\hat{y} = f(x) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$.

3. The Programs

The current version of ANREGR package contains two programs: REGPART and REGMUL. Both were developed under Microsoft Visual Basic development environment.

3.1. Common User Interface Features

The user interface of the programs is using Romanian language. All programs in the package have some common features with respect to the user interface, as follows:

- input data from keyboard can be stored in a file;
- the produced results are displayed in a separate window and can be either saved in a text file or printed.

The above features are implemented by the following elements in the main window:

- the observational matrix is displayed in a spreadsheet-like way;

- five command buttons: **Încarcă (Load)** for loading data from a file, **Salvează (Save)** for saving the current data in a file; **Prelucrează (Process)** for performing the computations (using multiple linear regression algorithm) and displaying the results; **Afişează (Display)** which shows **Results Window** (see below); **Închide (Close)** which ends the execution of the program.

The buttons **Încarcă** and **Salvează** display standard Microsoft Windows dialog boxes **Open** and **Save As**. The button **Afişează** shows **Results Window** (see Fig. 1 and 2).

The **Results Window** contains a text box which fills all its client area and a menu **Fişiere (File)**. The text box shows all the results produced by the program; the user can scroll in both horizontal and vertical directions. The **Fişiere** menu has the following menu options: **Încarcă (Load)**, **Salvează (Save)**, **Listează (Print)**, and **Închide (Close)**.

Figure 1. The **Results Window** with its menu visible

3.2. What Results Are Displayed

Most of the results computed by the general multiple linear regression algorithm are presented in a tabular form. They are displayed in the client area of the **Results Window** as shown in Figure 2:

- name of the problem, observational matrix (optional), and the regression equation;
- a table containing the name, mean, and standard deviation for each variable; it also contains correlation coefficients, regression coefficient and its standard error, computed t-value for independent variables;
- intercept, multiple correlation coefficient, and the standard error of estimate;
- analysis of variance table for the multiple regression;
- residuals table.

3.3. The REGPART Program

The REGPART program performs regressional analysis computations for eight different regression equations. Its features are:

- the dependent variable and the independent variables are established at data input and cannot be modified (the observational matrix is in the form (10));
- the regression models are pre-established.

The main window of REGPART program is shown in Figure 3 and has three areas: data area, in the upper part of the window, selection area in the lower part, and command area on the right.

Data area contains the name of the problem (**Denumirea problemei**), the number of independent variables (factors) n (**Numărul de factori**), the number of observations m (**Numărul de variante**), and the observation matrix with $n+1$ columns (the dependent variable is generically called *production* - **Producţia**), and m rows. Its elements can be introduced either rowwise (**pe linii** option button from the **Culegere date** frame) or columnwise (**pe coloane** option button).

Selection area in Figure 3 displays the regression models available, while *command area* contains the three command buttons **Prelucreaza (Process)**, **Afiseaza (Display)** and **Inchide (Close)**, already described in 3.1.

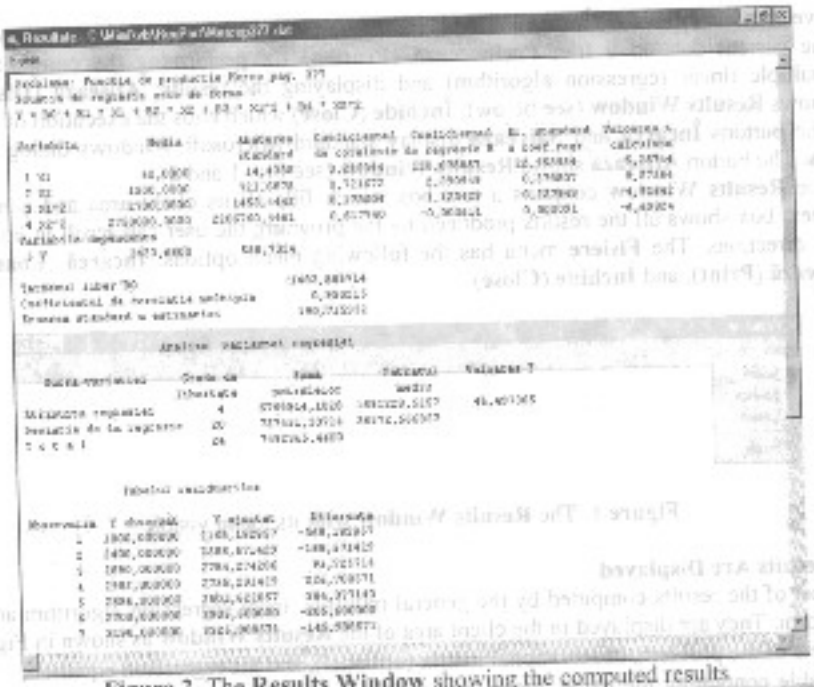

Figure 2. The **Results Window** showing the computed results.

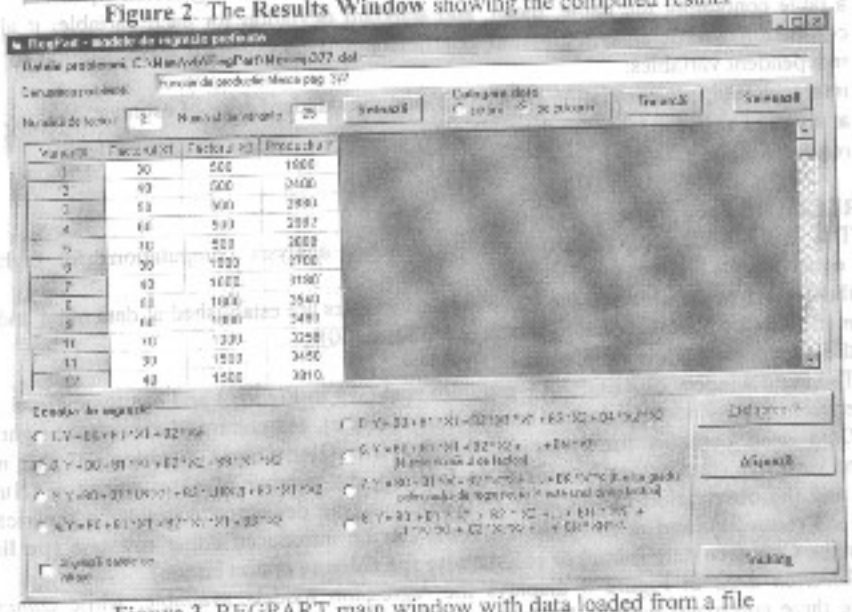

Figure 3. REGPART main window with data loaded from a file

After the model is selected, the button **Prelucrează** becomes active. By pressing it, the computation process starts and the results are displayed in the **Results Window**. Figure 2 shows some of these results for a model of type 8, applied to the problem in Figure 3.

3.4. The REGMUL Program

This program performs regressional analysis for linearizable general models. It allows the selection of the dependent variable and the independent variables during the construction of the model. Its main window, shown in Figure 4, resembles well the main window of the REGPART program. The only differences in the data area are related to the number of variables n (**Numărul de variabile**) and the number of observations m (**Numărul de observații**). The observation matrix is entered in the general form (3).

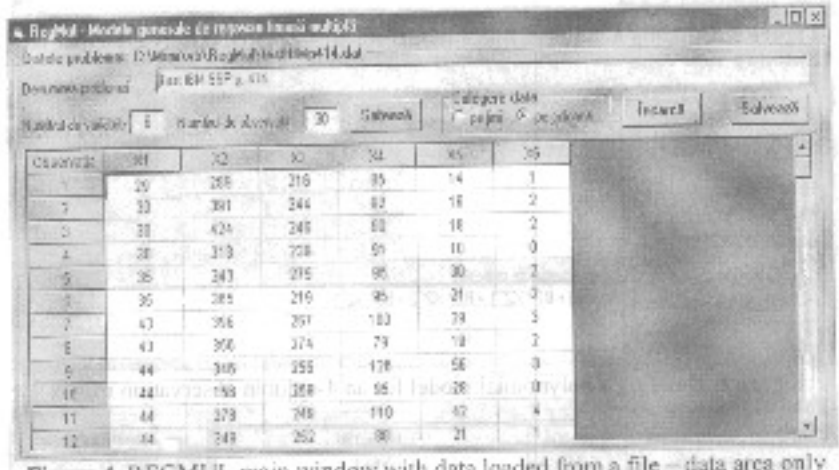

Figure 4. REGMUL main window with data loaded from a file – data area only

The selection area assists the user in building a multiple linear model of the general form $y = b_0 + b_1 z_1 + b_2 z_2 + \ldots + b_k x_k$, where the independent variables are z_i, $i = 1, 2, \ldots, k$. The three classes of regression models shown in the selection area are:

1. *multiple linear regression* – y and z_i are selected from initial variables;
2. *polynomial regression* – y and z_i are initial variables; z_i are powers of a selected initial variable;
3. *general regression* – y and z_i are defined by expressions containing functions (applied to initial variables), constants, and operators.

The content of selection area depends on the selected model, as shown in Figures 5, 6, and 7.

4. Conclusions and Future Work

Both programs discussed are based on the same general algorithm, described in [1]. The first one, REGPART, was designed to compute *production functions* in agriculture (for more details see, for example, [3]). The second program, REGMUL, can be used for every linearizable regression model. First, the user has to establish the definitions of Z and Y variables and to check if the corresponding expressions can be computed from initial variables. Next step is to complete definition table; the program does the rest of the work.

Future improvements include new predefined regression models (REGPART) and more elementary functions allowed in the definition of a variable (REGMUL).

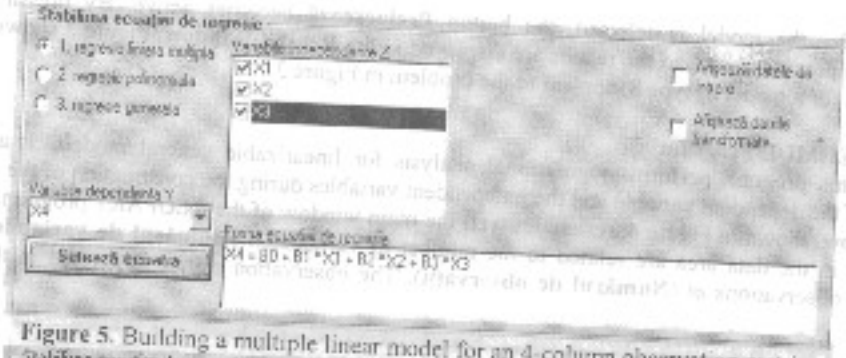

Figure 5. Building a multiple linear model for an 4-column observation matrix

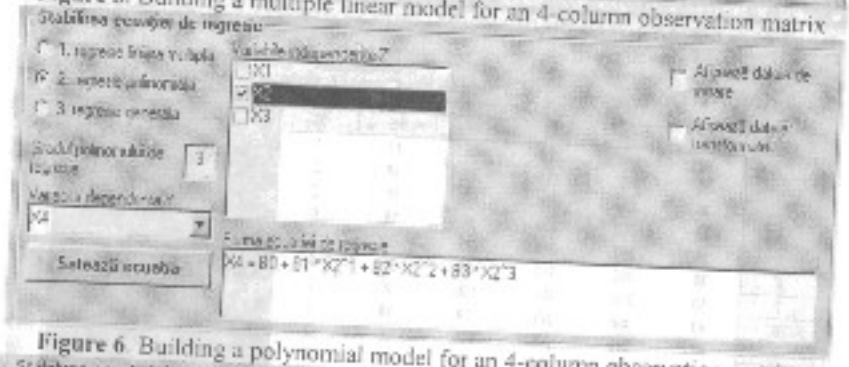

Figure 6. Building a polynomial model for an 4-column observation matrix

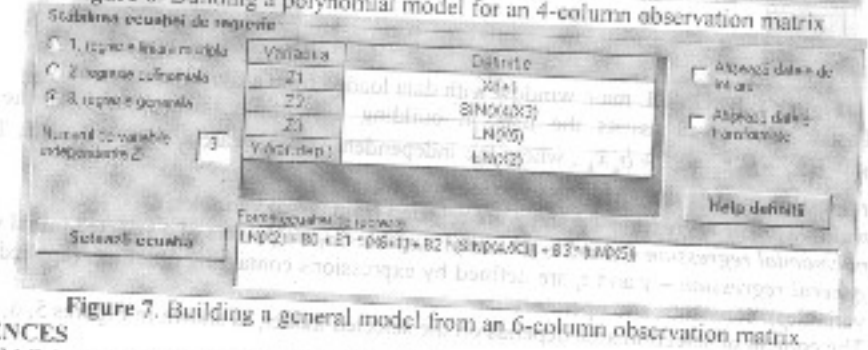

Figure 7. Building a general model from an 6-column observation matrix

REFERENCES

[1] IBM Scientific Subroutine Package (SSP), IBM Corp., 1969.

[2] Mendenhall, W., T. Sincich, *A Second Course in Statistics: Regression Analysis*, 5th ed., Prentice Hall, 1996.

[3] Merce, E., F.H. Arion, C.C. Merce, *Management general și agricol*, AcademicPres, Cluj-Napoca, 2000 (Romanian).

Received: 11.09.2002

University of Agricultural Sciences and Veterinary Medicine
Faculty of Veterinary Medicine,
Calea Mănăștur 3-5, RO 3400 Cluj-Napoca, Romania
E-mail: maria_parv@yahoo.com