TEACHING AND LEARNING WITH REGRESS

Cristian Marinoiu and Iuliana Dobre

ABSTRACT
In this paper the computer system for learning and teaching regression analysis, REGRESS, is described.
Built as a part of MIMAL (Multimedia Integrated Model for Active Learning) generated at "Petroleum and Gas" University of Ploiesti, presented at CBLIS 97, REGRESS has as its main goal to support the understanding of some topics from an introductory course about regression. The package uses simulation and interactive graphics capabilities of the modern computers.

KEYWORDS
Regression, correlation coefficient , scatterplot, simulation.

INTRODUCTION

During the development of REGRESS the authors have kept in mind the following natural ideas:

- the regression of a response variable Y on an explanatory variable X is the conditional expectation of Y given X=x, written r(x)=E(Y/X=x). The regression equation is only a model that describes the true dependence between X and Y;
- the key for a good assimilation of some regression concepts is their deep understanding both theoretically and intuitively. The simple linear regression is the first step to get to know more difficult types of regression: multilinear regression, nonlinear regression, ridge regression etc. Consequently simple linear regression has a central place in the package;
- it is a well-known fact that statistical inference in regression implies solid knowledge of mathematics. But this is not enough. Statistics is the science of learning from experience. We further need an intuitive profound understanding of the modelled phenomenon. Regarding this, the main advantages of the package, in our opinion, are:

  - the simulation routines, which can provide data for inferences very quickly;
  - graphical routines, which can visualize the resulted inference in most of the concrete situations we meet.

These tools help the students to better learn and link the regression concepts.

STRUCTURE OF REGRESS

REGRESS has its main capability the fact that it offers, at user request, a hypermedia module that contains the theory corresponding to a particular regression concept, but also illustrative examples and possibilities of simulation and/or experimentation.

Basically, REGRESS includes four major blocks:

- *INT* – the interface block, where the user can select the desired module and, inside the module, any combination of the three options: Theoretical notions, Examples, Simulation and/or Experiments.
- *SIMCOM* – the block which provides for the possibility to activate the computation and simulation algorithms.
- *GRAPH* – the block which provides for the possibility to loading the hypermedia courseware and to activate the graphical algorithms.
- I&A – the Integration and Authoring block that produces the desired module.

As shown in the Figure 1, REGRESS has an interactive link with a database containing Hypermedia Courseware, Simulation, Calculations and Graphical Algorithms.

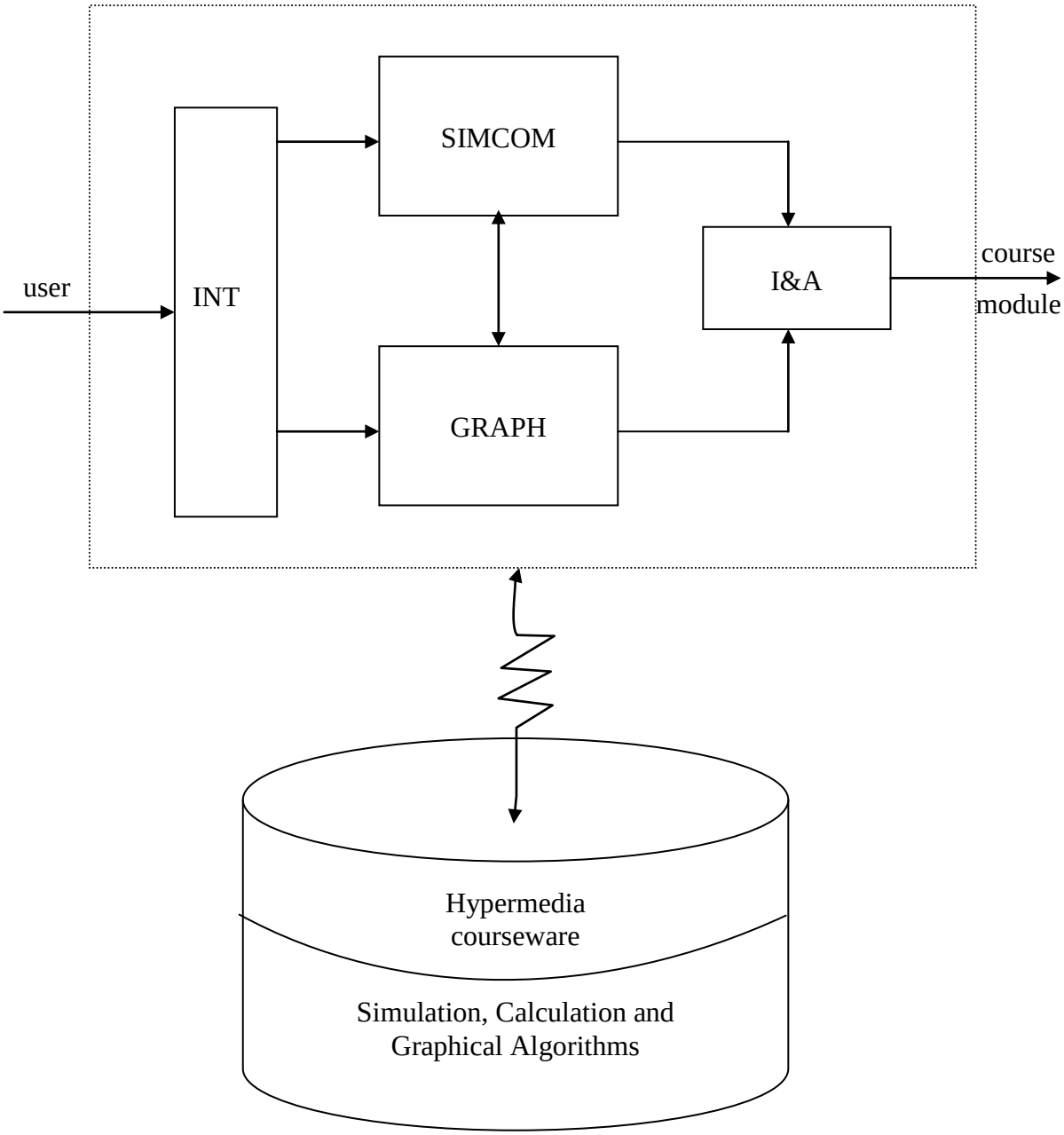

Figure 1. Structure of REGRESS

STRUCTURE OF THE COURSES

The course's table of content shows the list of the approached notions or concepts corresponding to each module.
Every course is partitioned in modules as shown in Figure 2.

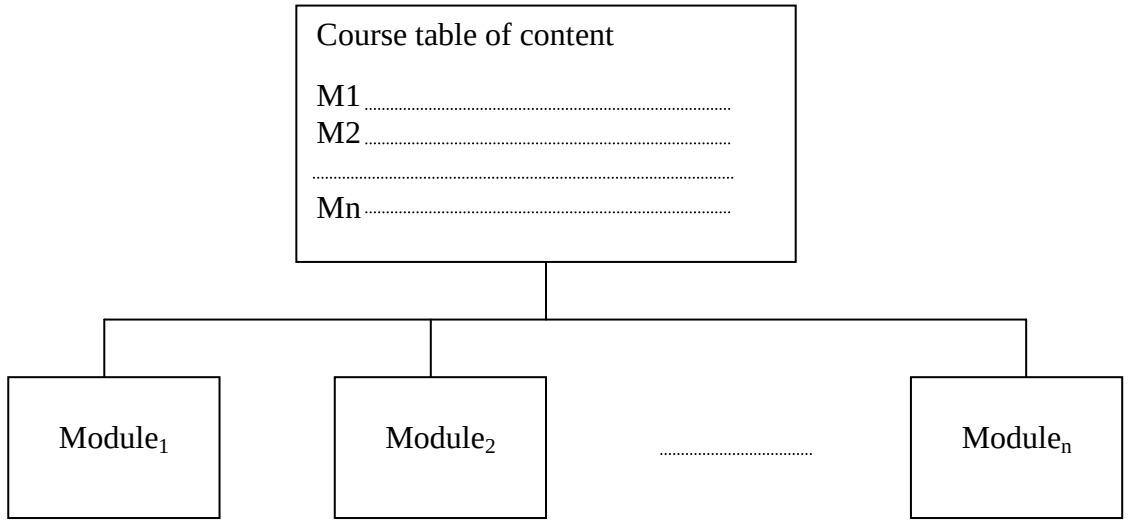

Figure 2. Structure of the course

Every course module has a standard three-section structure, as follows:

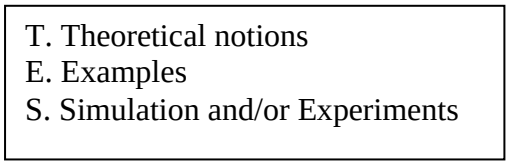

Figure 3. Structure of a module

A working session starts with user choosing from the main menu the desired module.

Usually, the study of any module implies the browsing through the three icons (Theory, Examples, Simulation and/or Experiments), but it is also possible to choose directly the subject of interest.

EXAMPLES, SIMULATIONS, EXPERIMENTS

From a pedagogical point of view the simulation and graphical capabilities of REGRESS represent the most important feature of this package. Below we describe several simulation examples and/or experiments presented in the package. The meaning of the used notation is as follows:

Ei-j is the example j within the module i;
SEi-j is a simulation or experiment j within the module i

E1-1
Goal: to obtain in the mind of the student an "image" of the sample correlation coefficient which measures the linear dependence between the random variables X and Y.
Means: for every considered pair of variables (X, Y) one shows the scatterplot of the sample correlation coefficient.

E1-2
Goal: to show that the sample of correlation coefficient is only a measure for the linear dependence between random variables.
Means: one plots in the plane the pairs of points (x_1, y_1), (x_2, y_2) … (x_n, y_n) that satisfy the equation $X^2+Y^2=1$, with the sample correlation coefficient zero. Obviously, the random variables (X, Y) are correlated but the zero value of the correlation coefficient indicates the lack of the linear dependence between X and Y only.

E1-3
Goal: to point out that the sample correlation coefficient is only an estimator of the theoretical correlation coefficient.
Means: one shows the scatterplot for the all-discrete values of the vector (X, Y) and the value of the correlation coefficient also. On the same screen one displays the scatterplots and the sample correlation coefficients for several observed samples.

SE1-1
Goal: to develop the ability to find good linear dependence between transformations of random variables X and Y.
Means: one activates the special program that is able to show the correlation and the scatterplots of any combinations of transformations of X and Y, for example, (X, log Y), (log X, Y), (log X, log Y), ($\sqrt{X}$, log Y) and so forth.

E2-1
Goal: to emphasise the following facts:

- the regression of a response variable Y on an explanatory variable X, E(Y/X) is not necessarily a line;
- the regression of Y on X, when a two-dimensional random variables (X, Y) is normally distributed is always linear;
- there is a two-dimensional random variables (X, Y), non-normally distributed and still, E (Y/X) is linear.

Means: for every presented case above, one choices an adequate example. The three examples have as a common feature the fact that E (Y/X) can be calculated analytically. Every chosen example of regression E (Y/X) has an analytical form, which is graphically represented.

E 2-2
Goal: to show the failure of the plug-in estimate formula

$$\hat{r}(x) = \frac{sum of\ y_i\ values for which\ x_i\ with\ x_i = x}{number of\ x_i\ with x_i = x}$$

for the regression of a response variable Y on an explanatory variable X, $r(x) = E(Y/x)$, where (x_i, y_i), i=1, n represents a sample. This fact justifies the use of the regression models as an alternative.

Means: for a fixed sample (x_i, y_i), i=1, n one displays on the same scaterrplot the true regression $r(x) = E(Y/x)$ and the plug-in estimate regression $\hat{r}(x)$. One can observe that the function $\hat{r}(x)$ is a rough estimation for $r(x)$.

E 3-2
Goals:

- to develop a feel for the difference between the chosen model and the true regression;
- to make aware the student of the diversity of the models which can be proposed;

Means: on the same scatterplot one displays: the ordinary least squares (OLS) regression line of Y on X, the OLS regression line of X or Y, the PC (principal components) regression line, the MAD (minimisation of the sum of absolute vertical deviations) regression line.

ES 3-1
Goal: to make aware the student of the fact that the estimated regression line depends on the chosen sample.
Means: one simulates several samples for a known distribution function of the random vector (X, Y). For every sample one shows the obtained OLS regression line.

ES 3-2
Goal: to show that the OLS estimators of the unknown parameters in the linear regression model are unbiased.
Means. one solves the linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ for N simulated samples (x_i, y_i), i=1, n. For every sample i, i=1,2,… N one obtains the estimators $\hat{\beta}_0^{(i)}$ *and* $\beta_1^{(i)}$, $i = 1, N$. For N large, one can observe, that the true regression line $E(Y/x) = \beta_0 + x\beta_1$ is almost the same with the line $y = \overline{\beta}_0 + x\overline{\beta}_1$ where $\overline{\beta}_0 = \sum_{i=1}^{N} \hat{\beta}_0^{(i)} / N$, $\overline{\beta}_1 = \sum_{i=1}^{N} \hat{\beta}_1^{(i)} / N$.

EVALUATION OF THE PACKAGE

In order to obtain a preliminary evaluation of the package, a questionnaire has been administrated to 24 students that were using REGRESS. We asked them three questions about the understandandability (Q1), the easy of use (Q2) and the usefulness of REGRESS in learning of regression concepts (Q3). Below are shown the obtained average scores and standard deviations on a 10-point scale, with 10 representing the most favourable response.

Question	Mean	Standard deviation
Q1	8.17	1.13
Q2	8.63	0.92
Q3	8.54	1.06

Figure 4. Table of scores

CONCLUSIONS AND FUTURE WORK

The scores obtained for the questions Q1 and Q2 encourage us to continue to develop REGRESS in the same initial manner. Also, based on a good score obtained for the question Q3 and on better grades the students have got on exams, we appreciate that REGRESS

improves the results of students in the assessment process. Even though there is a lot of commercial statistical software (for examples: SPSS, STATITCF, SYSTAT etc.), we have chosen to develop this package in order to focus on the pedagogical issues.

Future work directions are:

- To continue the incipient research in exploring the pedagogical merit of REGRESS;
- To extend REGRESS by enlarging the base of regression topics;
- To build a feedback module;
- To increase the reliability and extensibility of the package by using an automatic maintenance block;

Cristian Marinoiu, Department of Informatics;
PG of University Ploiesti
39 Bucuresti Blvd
2000 Ploiesti
ROMANIA
e-mail: marinoiu_c@yahoo.com

Iuliana Dobre, Department of Informatics;
PG of University Ploiesti
39 Bucuresti Blvd
2000 Ploiesti
ROMANIA
e-mail: iuliana_dobre@yahoo.com