REGREX: A JAVA-BASED PACKAGE FOR TEACHING AND ACTIVE LEARNING IN REGRESSION

Cristian Marinoiu, Iuliana Dobre

ABSTRACT

REGREX is a software system designed to improve teaching and learning some concepts in regression. Even though there is a lot of commercial statistical software we have chosen to develop this computerized package, in order to focus on the pedagogical issues. In order to do this we have developed REGRESS (Marinoiu and Dobre, 2001), presented at CBLIS'2001, in Brno, Cehia. REGREX is an improved version of REGRESS. Obviously, in REGREX we enlarge the base of regression topics. However the main gain with this new version is the possibility to offer to our students the support for active and discovery-based learning. Simulated experiments using Java applets are the main means for that.

KEYWORDS

Regression, simulation, experiment

INTRODUCTION

To beginners in statistics, many mathematical results appear quite abstract. Despite the teachers' effort, the real signification of many statistical formulas represents just a desideratum. The consequence is a superficial understanding of fundamental, statistical concepts, reflected in the inability to solve practical problems.

The authors of this paper thought that the understanding, assimilation and retention of these concepts can be improved using the opportunity provided by modern computing technologies. Making experiments with REGREX forces the student to actively engage in the training process.

REGREX SOFTWARE

The interface of REGREX is written in HTML that can be easily interpreted using, for example, the well-known browsers Internet Explorer or Netscape Navigator.

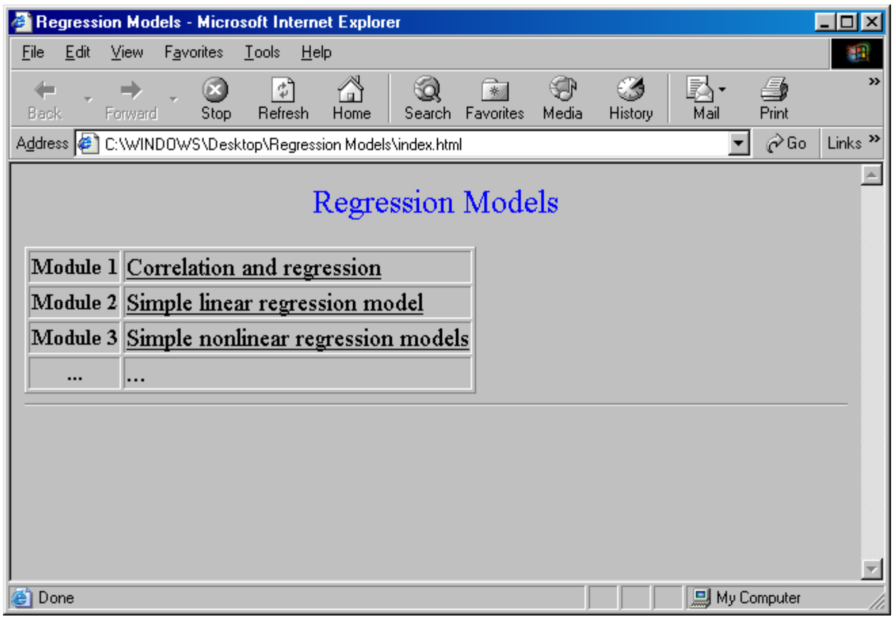
Besides the advantage of the accessibility, HTML also offers two major facilities for our goal:

- easy access to the presented course contents, due to the existence of the hyperlinks;
- execution of some illustrative routine in Java language (applets). Without the hyperlinks and the Java applets, our courses in regression would be just a common transcription in electronic format of a classical course;

By using applets, dynamic aspects of some concepts intuitively acquired are quickly introduced into the theoretical framework of the course.

USING REGREX

A normal working session starts with the presentation of the course's table of contents structured in modules. Below (Figure 1) one can see the beginning of this table of contents.

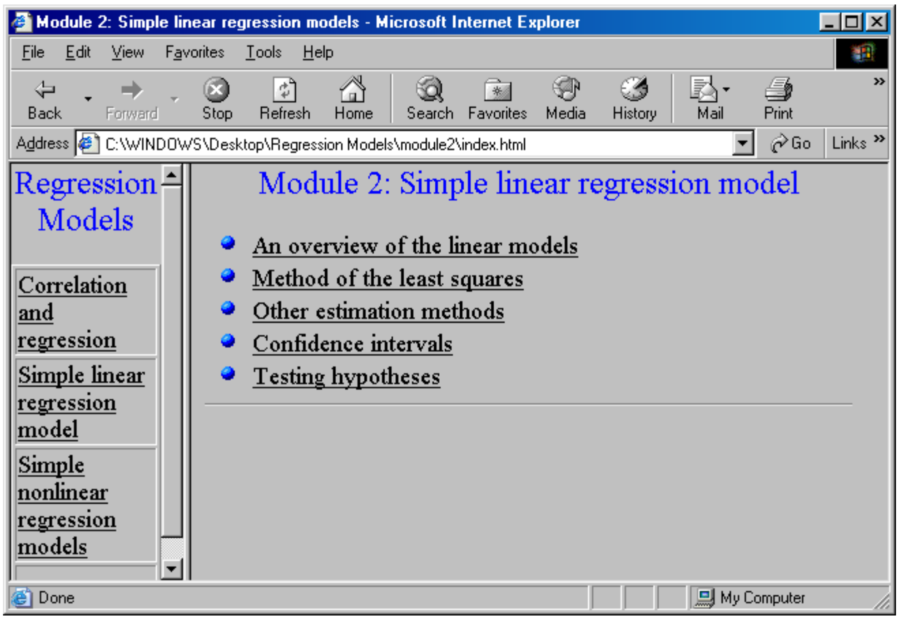

Regression Models - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print
Address C:\WINDOWS\Desktop\Regression Models\index.html
Go Links
Done
My Computer

Regression Models

Module 1	Correlation and regression
Module 2	Simple linear regression model
Module 3	Simple nonlinear regression models
...	...

Figure 1. Course table of contents

Every course module refers to a specific table of regression topics, each of them containing: theoretical notions, examples, simulations and/or experiments. For example, the second module contains typical topics as in Figure 2.


Module 2: Simple linear regression models - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print
Address C:\WINDOWS\Desktop\Regression Models\module2\index.html
Go Links
Done
My Computer

Regression Models

Correlation and regression
Simple linear regression model
Simple nonlinear regression models

Module 2: Simple linear regression model

- An overview of the linear models
- Method of the least squares
- Other estimation methods
- Confidence intervals
- Testing hypotheses

Figure 2. The contents of the second module

From a main menu (the course's table of contents), by a simple click the student "jumps" inside the chosen module. Usually, the study implies the browsing through the listed topics, but it is also possible to choose directly the subject of interest. The examples, simulations and experiments are accessed by activating the suitable applet. Below, in Figure 3, we give a graphical image as a result of the execution of such an applet.

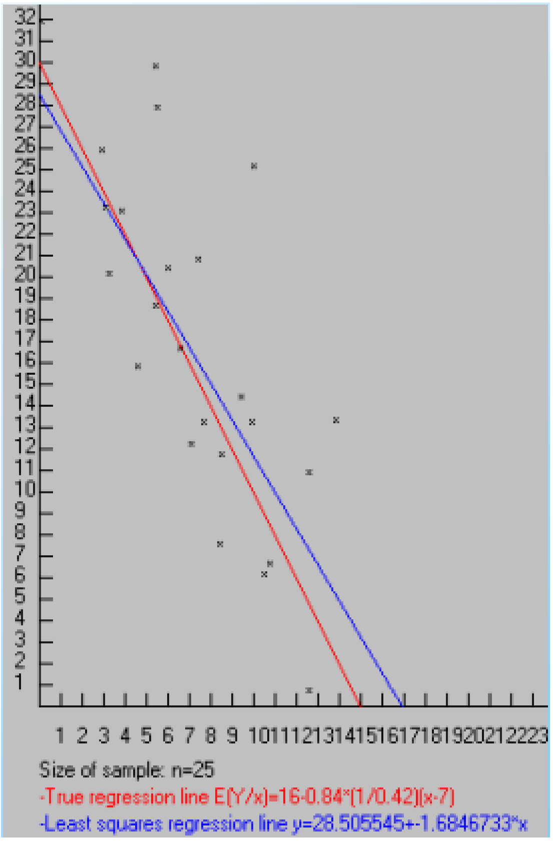


Figure 3. The true regression line and the least squares regression line

EXPERIMENTS

The main feature of REGREX software is its capability to support several statistical experiments. Such experiments allow the students to implement the scenarios suggested by the teacher; the feedback that is promptly provided by REGRESS reinforces some concepts acquired in the classical static manner.

We describe below three such experiments presented in the package.

Experiment 1

Objective

This experiment highlights to the students the limit of the plug-in principle in estimating the regression curve; it also justifies the use of the regression models as alternatives to the use of the plug-in principle in the regression estimation.

Preliminary framework

In an introductory course in regression, the students learn that the regression of a response variable Y on an explanatory variable X is, by definition, the conditional expectation of Y given X, namely

$$r(x) = E(Y / X = x) .$$

Usually, in practice only a random sample $(x_i, y_i)\ i = 1, n$ is available from the unknown distribution of the random vector (X, Y). The problem raised to the students is how they suggest to estimate $r(x)$ from the available sample.

The plug-in principle is a simple method to estimate some interesting aspects of an unknown probability distribution *F*, using the corresponding aspects of $\hat{F}$, the empirical distribution function built on the basis of a sample drawn from *F* (Efron and Tibshirani, 1993).

Since the plug-in principle was familiar to our students, their responses were not a surprise: 92% of the respondents indicated the plug-in principle as a method to estimate $r(x)$, namely

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

Is $\hat{r}(x)$ really a good estimator for $r(x)$?

An intuitive response to this question is available to the students themselves by experimenting with REGREX.

The suggested experiment

1. Each student activates on his/her computer a special routine of REGREX that will display the true regression line $r(x) = E(Y / X = x)$ for a normally distributed vector (X, Y). Notice that the entrance parameters of the normal distribution can differ from a student to another student.

2. For a pre-established entrance number *n*, REGREX plots on the previous scatterplots the coordinates $(x_i, y_i)\ i = 1, n$ of a random sample from the above considered normal distribution and traces the graph of the $\hat{r}(x)$ curve [Figure 4].

Comments

The significant discrepancy observed between the smoothness of the true regression curve $r(x)$ and the rough aspect of its plug-in estimate curve $\hat{r}(x)$, was a surprise for the students. The majority of them indicated as the cause for the rough aspect of $\hat{r}(x)$ the fact than n was perhaps too small. Because the latest version of REGREX permits different entrance values for n, the teacher invited the students to repeat step 2 for arbitrary values of n. The results obtained indicated that generally the problems of variability still remain. Therefore, in this situation the plug-in principle doesn't work very well. The experiment pointed out to the students the necessity to use alternatives to the plug-in principle in order to obtain a reasonable smooth estimate of $r(x)$.

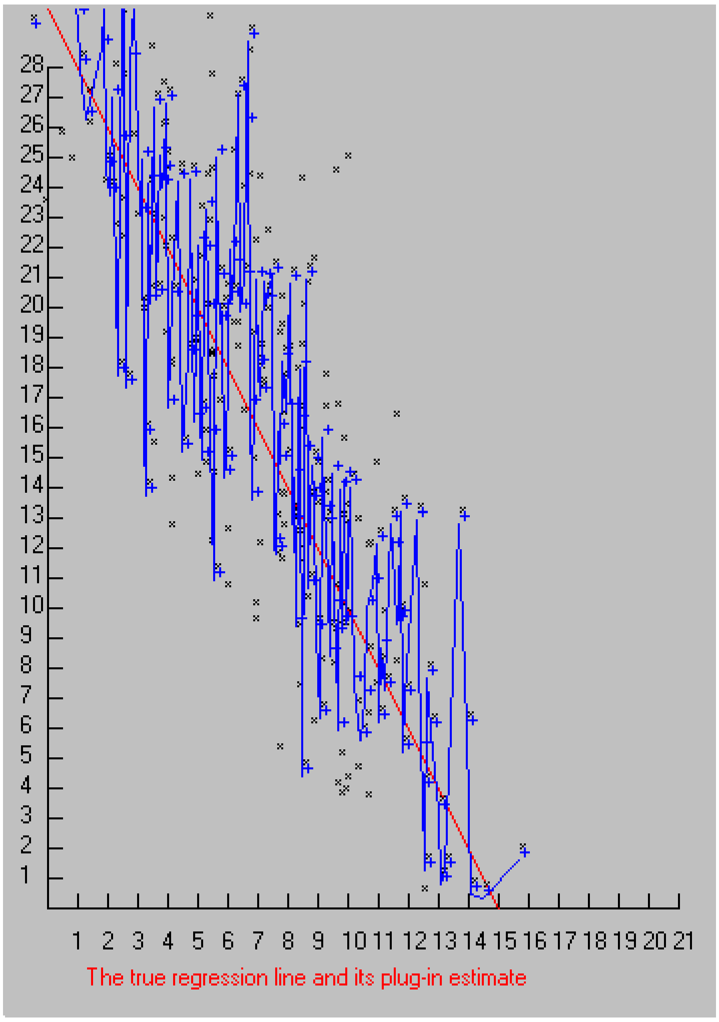


Figure 4. The true regression line and its plug-in estimate

Experiment 2

Objective

This experiment has the main goal of showing to the students that in the choice of a family of smooth functions in the regression models, the use of the residuals sum of squares as an absolute criterion might provide a poor estimate of the true regression.

Preliminary framework

The students know that the least squares method is an elegant alternative of the plug-in principle in the estimation of the true regression $r(x) = E(Y / X = x)$.

This technique particularly offers the possibility to find a good estimate of $r(x)$ considered as

$$r_{\beta,p}(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p,$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ is an unknown parameter.

The unknown parameter β can be estimated by minimizing the residual squared error

$$RSE(\beta, p) = \sum_{i=1}^{n} \left(y_i - r_{\beta,p}(x_i)\right)^2$$

where (x_i, y_i) $i = \overline{1,n}$ is a sample from the unknown distribution of the vector (X,Y). Let us denote by $\hat{\beta}$ the obtained estimator of β and by $r_{\hat{\beta},p}$ the estimate of the true regression $r(x)$.

Therefore

$$RSE(\hat{\beta}, p) = \min_{\beta} RSE(\beta, p)$$

It is known that (Efron and Tibshirani, 1993) $RSE(\hat{\beta}, p)$ is a non-increasing function of *p*. In other words the larger *p* is, the smaller $RSE(\hat{\beta}, p)$.

The question asked to the students is the following: knowing that $RSE(\hat{\beta}, p)$ decreases when *p* increases, is $r_{\hat{\beta},p}(x)$, for *p* large, always a better estimate for the true regression $r(x)$?

To the previous question the students responded as follows: 83% - *yes*, 10% - *no*, 7% - *I do not know*.

In order to obtain an intuitive response to the question raised, the students were invited to make the following experiment.

The suggested experiment

1. Each student activates on his/her computer a special routine of REGREX that will display on the scatterplots the true regression line $r(x) = E(Y / X = x)$ for a normally distributed vector (X,Y).

2. Another special routine of REGREX plots on the previous scatterplots successive graphs of the curves $r(\hat{\beta}, p)$, for different increasing values of *p*.

Comments

The result of this experiment was also a surprise for the students. Indeed, when *p* increases the values of *RSE* decrease but, at the same time on the scatterplots REGREX displays an increasingly rougher estimate of the true regression $r(x)$.

For large values of *p* this estimate will more and more resemble the plug-in estimate $\hat{r}(x)$, that is a poor estimate of $r(x)$. Therefore, our choice of a certain degree of smooth polynomial function is implicitly a choice of how smooth we believe the true regression to be.

Experiment 3

Objective

The aim of this experiment is to point out the lack of robustness of the least squares method and to get used to alternative robust procedures.

Preliminary framework

Despite its well-known capability to provide good estimators, the least squares method is very sensitive to the outliers. It is sufficient for a single outlier to occur, to compromise the result completely.

So far, the robust regression has not been presented in our course. However, by experimenting with REGRESS our students can get a first idea about outliers and their impact on the estimated regression curve.

The suggested experiment

1. REGREX plots on the scatterplots a special configuration of points and the corresponding regression line (*LS*). We say that this configuration is “special” because it has one point (say *P*, called outlier point) with y_i far away from the others.

2. The teacher asks the students to delete the point *P* (using a special routine of REGREX); REGREX provides a new regression line without P (*LS-P*).

3. The teacher asks the students to move the point *P* arbitrarily on the scatterplot. For every position of the point *P*, REGREX provides the *LS* and *LS-P* lines.

4. The same scenario presented in step 3. Furthermore, after each movement of the point *P*, REGREX displays the *LSM* line, based on the robust Huber's M-estimation method.

Comments

The high discrepancy between *LS-P* and *LS* [Figure 5] gives the students a first intuitive image about the sensitivity of the least squares procedure to the outliers (steps 1 and 2). Moreover, step 3 of the experiment allows them to observe the behavior of *LS-P* line against *LS* line when *P* is taken arbitrarily on the scatterplot. In this case, the discrepancy observed between *LS-P* and *LS* can be thought as a guide to labeling *P* as an "outlier" or as a "normal" point (Cook and Weisberg, 1982).

Step 4 highlights to the students the power and the vulnerability of the Huber's M-estimation method. When *P* is an ordinary outlier (with y_i far away from the others) *LSM* line is closer to *LS-P* line, namely the M-estimator works well (it rejects the outlier). When *P* is an outlying leverage point (with x_i far away from others) *LSM* line is closer to *LS* line, namely the M-estimator does not work well (Hampel, Ronchetti, Rousseuw, Stahel, 1986).

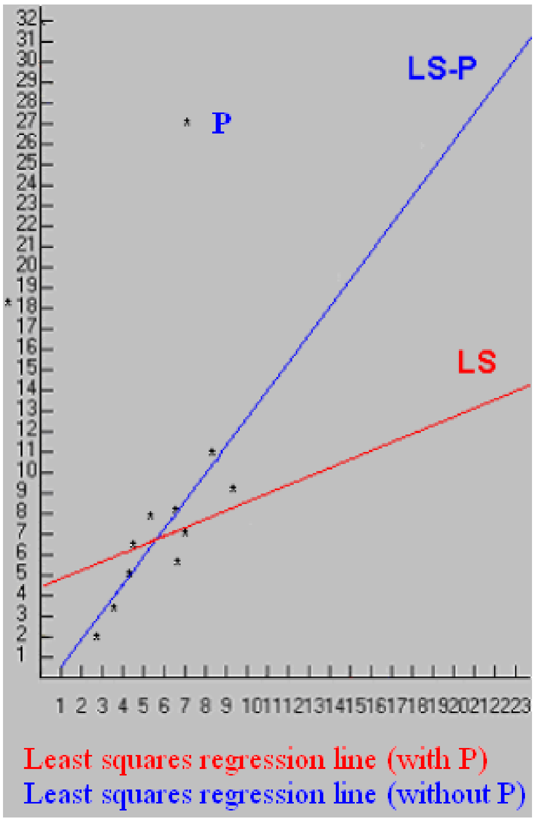


Figure 5. The regression lines LS (with P) and LS-P (without P)

EVALUATION OF THE PACKAGE

REGRESS – the old version of REGREX – had as a main goal supporting the understanding of some topics from an introductory course in regression. It was well received by students.

A new questionnaire has been administrated to the students that used REGREX. Besides the old questions about the understandability (Q1), the easiness of use (Q2), and the usefulness of REGREX a new question (Q4) was introduced.

The latter refers to how the students appreciated the introduction of the experiments in REGREX. The obtained average scores and standards deviation on a 10-point scale, with 10 representing the most favorable response are listed below [Table 1].

Table 1. Table of scores

Question	Mean	Standard deviation
Q1	8.91	0.89
Q2	8.53	1.12
Q3	8.62	1.15
Q4	8.83	0.56

CONCLUSIONS AND FUTURE WORK

We interpreted the good score obtained by the last question (Q4) as a recommendation to continue to improve REGREX, by projecting and implementing new experiments.

We have observed that the active participation in the learning process increases the students' interest in regression. Therefore, we would like to extend this procedure of training in a future version, using the experiments on the computer, in a tridimensional space.

REFERENCES

Cook R. D. and Weisberg S., (1982). Residuals and Influence in Regression, Chapman and Hall, New York.

Efron B., Tibshirani R. J., (1993). An introduction to the Bootstrap, Chapman & Hall,
London.

Hampel F. R., Ronchetti E. M., Rousseuw P. J., Stahel W. A., (1986). Robust Statistics, John Wiley & Sons, New York.

Marinoiu Cr., Dobre I., (2001). Teaching and learning with REGRESS, Paper C5 in Proceedings of the 5rd International Conference in Computer Based Learning In Science, CBLIS 7-10 July 2001, Brno, Edited by G.M. Chapman, University of Ostrava Press, Czech Republic.

Marinoiu Cristian Department of Informatics
PG of University Ploiesti
39 Bucuresti Bd
2000 Ploiesti
Romania
Email: marinoiu_c@yahoo.com

Iuliana Dobre, Department of Informatics
PG of University Ploiesti
39 Bucuresti Bd
2000 Ploiesti
Romania
Email: iulianadobre@yahoo.com