An Advisory System for Statistical Analysis

By I. Tsomokos and K. X. Karakostas
Department of Mathematics
Univ. of Ioannina - HELLAS

ABSTRACT

In this paper an advisory system for statistical analysis (ASSA) is presented for use by non-statisticians. The objective is to encourage office-workers, who are non-statisticians to make, as far as possible, a correct statistical analysis with an efficient and effective way, by exploiting some attributes of their data. ASSA is data -driven and navigates the user through QUESTIONS and ANSWERS to the selection of the statistical methodology (statistical technique or model) which is appropriate for the data under hand.

1. INTRODUCTION

It is well known that statistical methodologies are used by a lot of scientists in almost every scientific field. Therefore today the demand is increasing for systems that can implement these methodologies. More specifically the need is for systems that can give advise about the appropriate statistical methodology on formatted data.

This increasing demand has been caused, firstly by the rapid development of personal computer (pc) and workstation (ws) technology in the recent years and the continuously declining cost of them, and secondly by the large number of statistical packages available. The proliferation of pc and ws- based office systems have expanded the production of statistical analysis in a wide range of activities (i. e. civil works, social sciences, agriculture, education e.t.c.)

The rapid development of pc's has had as a consequence the implementation of more advanced statistical techniques with a lot of details in the statistical packages. This fact confuses the non-statisticians and prohibit them to select the right statistical methodology.

The ASSA system aims exactly at this point. It gives correct advice for the appropriate statistical methodology since its knowledge base is restricted by the built-in theoretical limits of existing statistical theory. This system is efficient and effective for the user.

The benefits of this system are decision quality, increasing production, time saving, cost reduction and general improved services.

In the next paragraph we discuss the steps necessary for the statistical analysis whereas in the third paragraph we give a description of the ASSA system.

2. STEPS FOR THE ANALYSIS

Roughly speaking if someone wants to statistically analyze his data he has to follow the following steps.

STEP 1. *Selection of the statistical methodology*
This is the fundamental step in the analysis. A wrong selection of statistical methodology may lead to wrong conclusions or it may not reveal the complete structure of the data.

STEP 2. *Selection of the right statistical software*
Having decided about the statistical methodology to be used we have to select a statistical package to implement it. The selection is crucial because not all packages give the same (maximum) information to the user. We will illustrate this point with the following example. Suppose that our statistical methodology, selected in the previous step, is that of regression analysis. It is well known that the various forms of residuals (e.g. raw, standardized, deleted. e.t.c.) play quite an important role in such an analysis. So a statistical package which computes only the raw residuals is not satisfactory.

STEP 3. *Main analysis*
At this step the non-statistician needs some rules to follow in order to complete the analysis. For example in regression analysis one needs to know what the assumptions are about the error term and how they can be checked.

STEP 4. *Interpretation of the results*

This means not only the final conclusions but also the intermediate ones. The last ones are sometimes quite crucial because they may lead to a different model.

For example in regression analysis if the user does not interpret correctly the various graphs of the residuals then he may end up with a wrong model and consequently with wrong conclusions.

The above are shown graphically in figure 1.

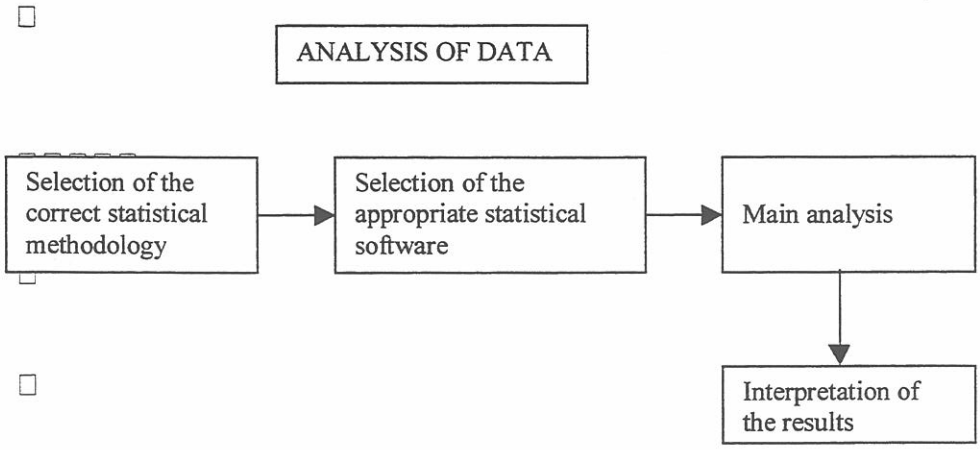


Figure 1. The steps necessary for statistically analyzing data

As we pointed out at the very beginning of this paper the user has a large number of statistical software to choose from. For a comparison between some of them see (να μπει αναφορά)

In the last few years some of the available statistical packages developed procedures along the lines of step1. (e.g SAS, S.P.S.S.). However their work in this area is not satisfactory because they mainly interpret the results and/or give some kind of a glossary and they do not suggest specific statistical methodology. To the authors knowledge there is no published work in this direction except for the work by Bill Trochim, which is available through the Internet in the address (http://trochim.human.cornell.edu/). Our system focuses on the first step although some kind of progress has been done on the third step too.

An ideal function for an expert statistical system of the kind described in this paper would be for the user to provide the necessary information in his natural language. The system, after doing a semantic analysis, recognize the user's needs and acts on the formatted data, then it response with all relevant information independently of the statistical package used by the user.

Up to now such a system does not exist. We hope that we shall be able to provide such a system in the near future.

DISCRIPTION OF THE ASSA

The ASSA, like most diagnostic systems, is data-driven. That means that the user provides some information- in the form of replies- to questions posed by the ASSA and then he receives the appropriate advice. All these questions and answers are kept on a logged window. The content in that window can be saved in a file for later use. For some questions, the user responds by selecting from a menu. In other cases it is necessary to type in the answer. The user has the option to ask the system to explain the question so that this becomes more clear. In addition the ASSA has the possibility to explain various statistical items, which exists in the question posed, in an html form. This can be done either by giving an appropriate definition or by giving an example.

A kind of such questions and possible answers are given in the following example.

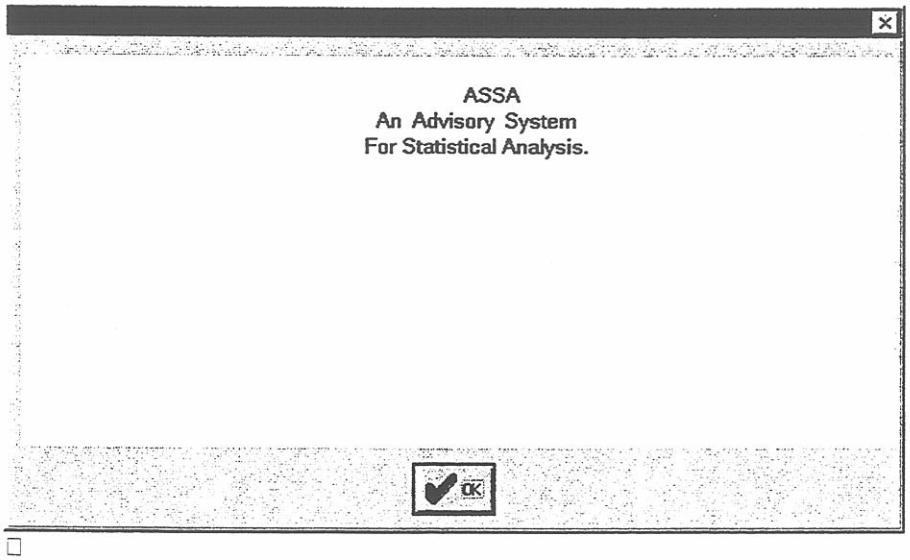
ASSA : *How many variables are involved in your analysis?*
User : Two (2)
ASSA : *Could any one of them be considered as a depended variable?*
User : Explain
ASSA : Your interest, at the moment, is focused on two variables of your data. If you are interested in studying e.g. the effect of the first variable on the second one, then the second variable is considered as dependent variable. As an example consider two variables e.g. WEIGHT and HEIGHT. If we want to see the effect of HEIGHT on WEIHGT then WEIGHT is the dependent variable. For more information you can see about that item on the HELP button.
User : Yes
ASSA : *Is that variable quantitative?*
User : Yes
ASSA : *Is the other variable (call it independent) quantitative too?*
User : Yes
ADVISE : *The suggested statistical methodology is that of the simple regression model.*

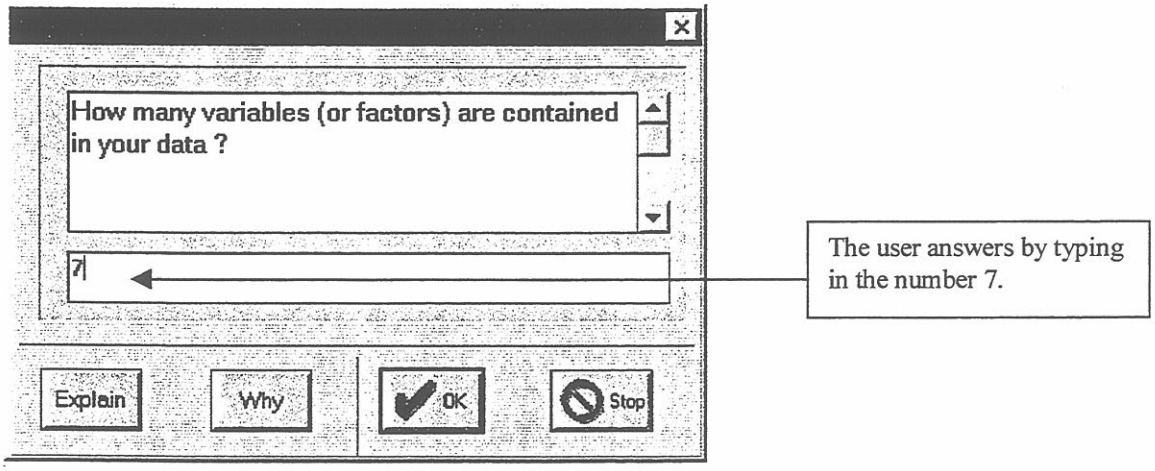
The other possibility was the goal-driven approach, which starts with a specific statistical methodology and tries to validate it. This it is not preferable since there exists great possibility for the user to pass through a large number of statistical methodologies and this may discourage him. Beside that it is also time consuming.

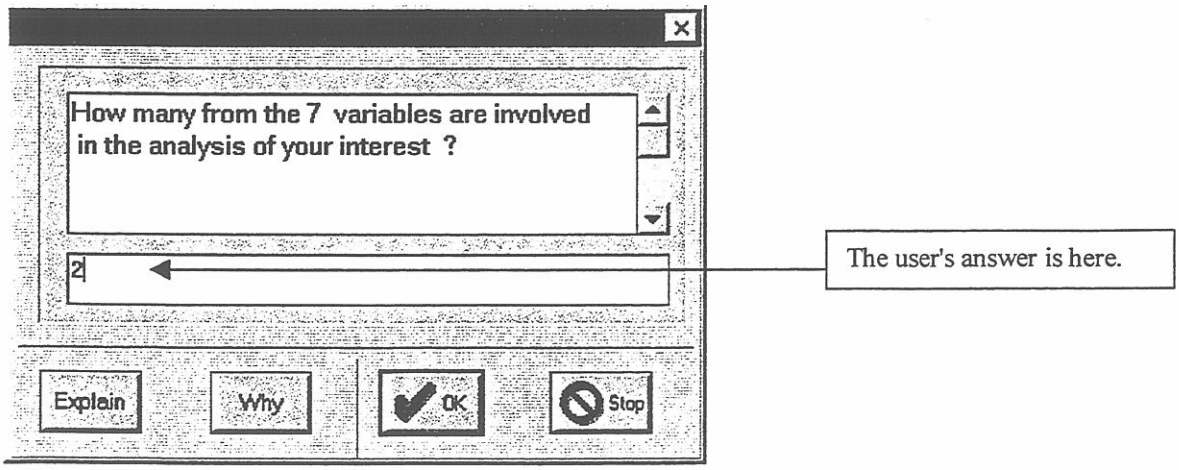
The knowledge base of ASSA contains about thirty rules. For example for the simple regression we have a rule like this:

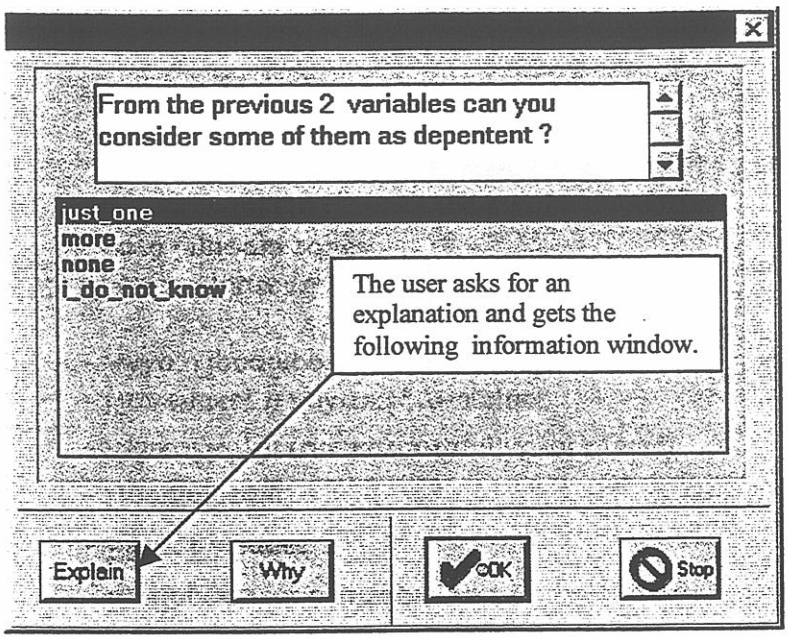
IF there are two (2) quantitative variables in the analysis
AND one of them can be thought of as a dependent variable
THEN the model is that of the simple regression.

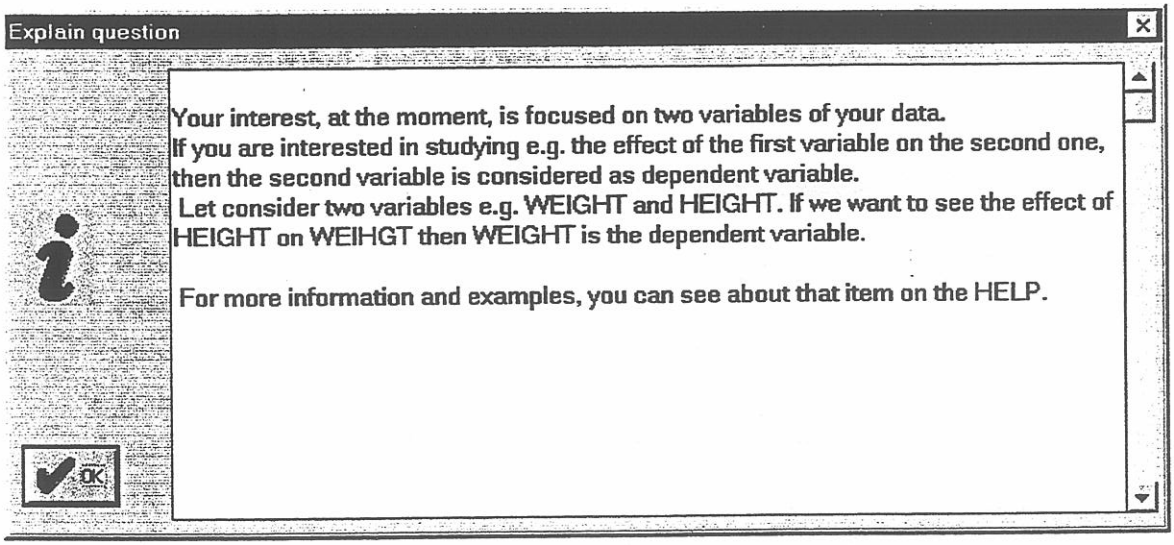
The ASSA system has been prototyped and tested with the shell ESTA
(Expert System shell for Text Animation).
An example using ASSA follows.


ASSA
An Advisory System
For Statistical Analysis.
OK

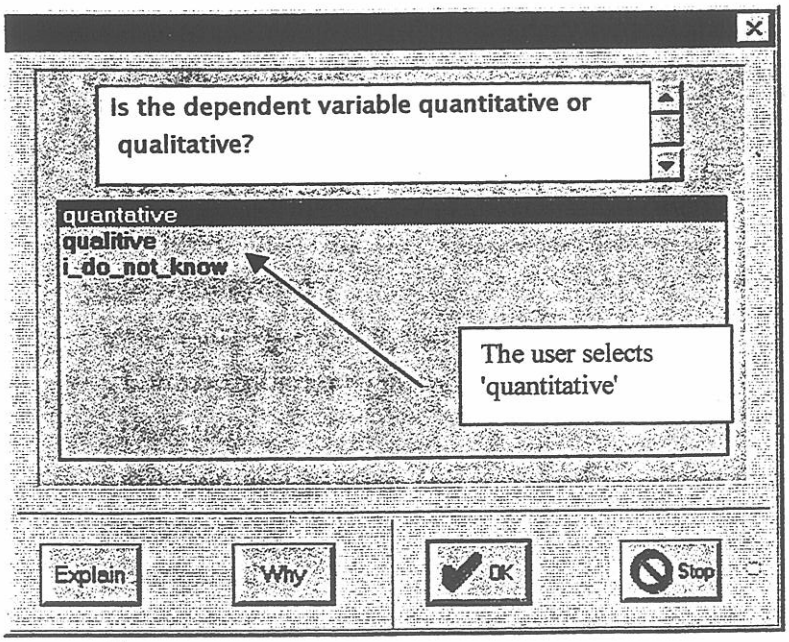

How many variables (or factors) are contained in your data ?
7
Explain
Why
OK
Stop
The user answers by typing in the number 7.

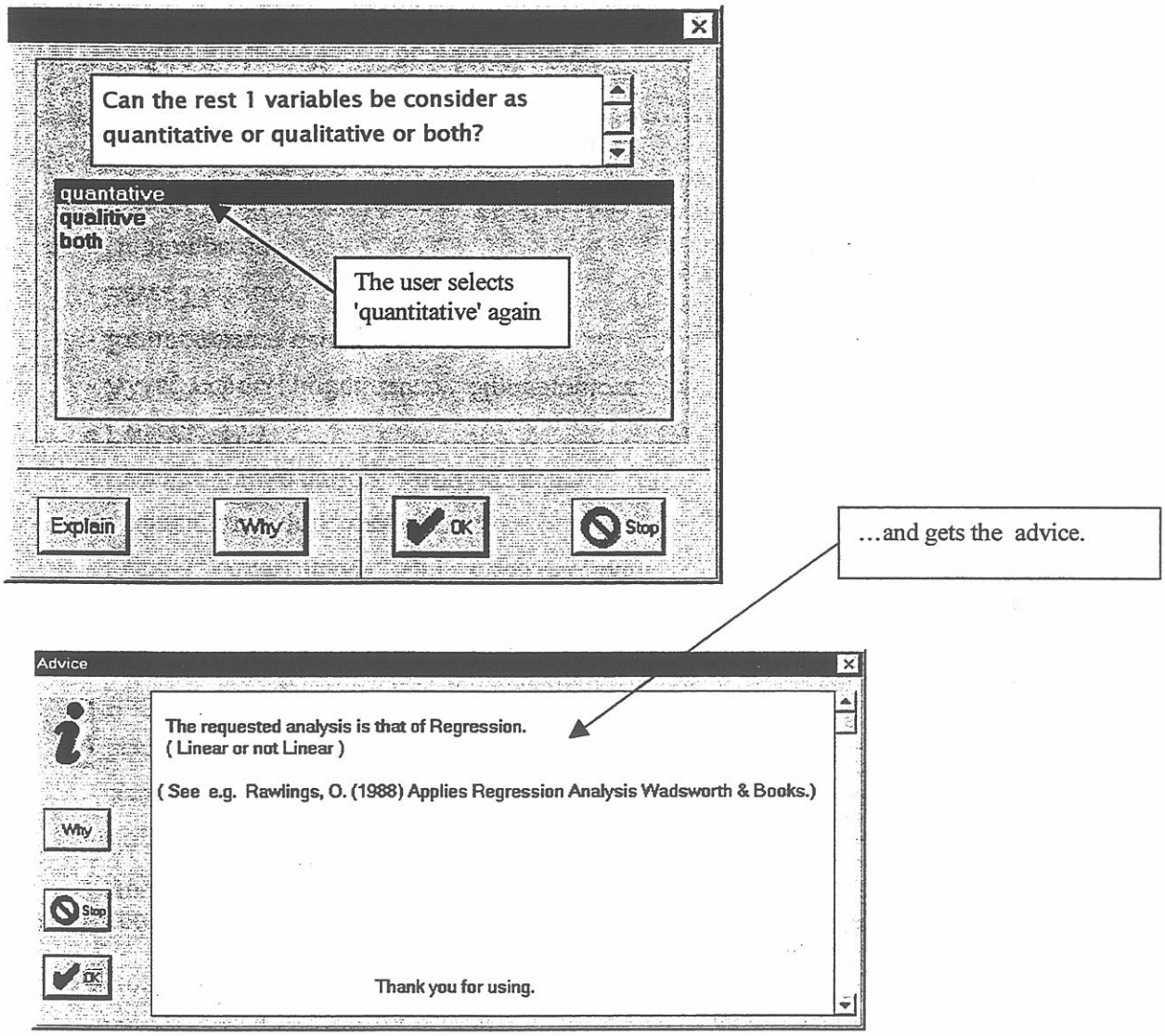

How many from the 7 variables are involved in the analysis of your interest ?
2
Explain
Why
OK
Stop
The user's answer is here.





By pressing the OK button in that window he gets the previous one where he selects the answer *'just_one'* and goes the following window.





References

[1] Morgan, W. T. (1998), A review of eight statistics software packages for general use, *The American Statistician* 52, 70.

[2] S.A.S Institute Inc. S.A.S/Lab Software: User's guide, Version 6.

[3] S.P.S.S Inc. S.P.S.S Software Version 8, Statistics coach

[4] Tabachnick, B.G., and Fidell, L.S. (1991), Software for advanced ANOVA courses A survey, *Behavior Research Methods, Instruments, & computers* 23, 208.

[5] Trochim, W.M.K Bill's Trochim center for social research methods, Cornell University.
(http://trochim.human.cornell.edu/selstat/ssstart.htm)

This research was supported by the Hellenic Minister of Education, E.P.E.A.E.K program, School of Medicine, University of Ioannina.

A modified version of this paper was presented, after being refereed, at the 5th International Conference of the Decision Sciences Institute, held at Athens, on July 4-7, 1999.