SUSAM : A Teaching Tool for Multicollinearity Analysis

Gianfranco Galmacci and Maria A Pannone - Perugia, Italy

1. Introduction

The teaching of statistics requires constant integration of theoretical and practical work. In the teaching of advanced statistical methodology, this asks for strong commitment on the part of the teacher, who is expected not only to make students acquire knowledge of methods, but also to generate a *feel* for data and the competence to choose the most suitable techniques for obtaining the information needed, which are the essential abilities of an expert statistician.

Technological development makes it possible nowadays to delegate part of this work to specialised software systems to be used as tutorial aids in the teaching of specific subjects. To date, much research has been carried out in this direction, some of which has achieved extremely successful results. Because of the problems involved, however, the computer assisted teaching is still in an experimental phase and it is necessary to identify the most correct didactic approaches for devising the aids.

Our study is meant as a contribution in this direction. It aims at verifying whether a software system which operates as a tutor for practical work in a course on statistics can be used as a valid aid to the theoretical work carried out by the teacher.

In developing a system with the given tutorial features, we referred to *expert system* techniques to obtain a versatile environment. But the theoretical subject, i.e. the statistical domain, had to be quite limited, so that the expert system would have no difficulty in dealing with it; at the same time, it had to be complex enough from a statistical point of view to serve as a good test for checking the validity of this teaching approach. On the basis of these considerations, we decided to choose multicollinearity diagnostics in linear regression models.

In the following pages we shall illustrate the features of our prototype (SUSAM), which is presently being trialled within the course in linear regression run by our department.

2. Multicollinearity diagnostics in SUSAM

Systematic study of multicollinearity is quite recent and has led to controversies among researchers, first of all as regards a formal definition which could also provide guidance in practical work. Discussion has also focussed on detecting multicollinearity and the possible courses of action which should be taken to avoid its effects.

From a teaching point of view, multicollinearity poses a series of problems due to the variety of its indicators, the numerous effects it produces, and the wide range of situations in which it can arise. Consequently, teaching must include a lot of practical work to make students aware of the concepts and acquire the necessary judgement.

SUSAM was developed on the basis of these considerations. Our main objective was to create an environment that could enable students to acquire experience of the various diagnostic methods in a self-sufficient way, and give them the opportunity of having their work constantly checked.

SUSAM utilises some of the most important methods recently proposed for multicollinearity detection together with more conventional techniques, which are still quite often used. The reason for this choice is twofold: on one hand, being a teaching aid, the system should enable the user to explore various procedures freely, i.e. it should allow the user to acquire a certain amount of experience in using them and in evaluating their effectiveness; on the other hand, we believe, as do many other researchers, for example Gunst (1983), that no diagnostic method is completely satisfactory; on the contrary, every method allows the study of different aspects of the problem, and provides information with varying levels of exhaustiveness.

The methods implemented in the system are:

(i) Correlation Matrix Analysis;
(ii) t-test on regression parameters;
(iii) Variance Inflation Factors (VIF);
(iv) Condition Indexes and Variance-decomposition proportions.

Since these methods have already been fully described in numerous works, we shall not examine them here in detail; however, it is necessary to consider them briefly.

Correlation Matrix Analysis is useful when multicollinearity involves pairs of variables; it does not provide any indication as regards collinearity involving more than two variables. Therefore, when a student decides to use this method, he is made to consider that the information provided is not complete.

The second method implemented in SUSAM is based on the use of t statistics to test hypotheses about regression parameters. It is widely known that multicollinearity affects the values of the t statistics; thus, it can be one of the causes, but not the only one, of the non-significance of some parameters. Nevertheless, the t-test for eliminating variables from a model is quite often carried out without considering the possible presence of multicollinearity. We believed it was important not to underestimate this problem when devising SUSAM; consequently, we introduced an appropriate *critical* aid which would take into account this particularly delicate side of the question. When examining the values of t statistics, the user should consider the possible effects that multicollinearity can have on the test, and in particular he should analyse the factors which determine the values of t; he is also advised to use the diagnostic procedures which are more appropriate for collinearity detection.

VIF analysis is probably the most widely used approach, since Variance Inflation Factors make it possible to detect multicollinearity and to measure its effects on estimate precision. However, they do not enable the user to determine which variables are mainly responsible for variance inflation.

The last of the four methods is the one proposed by Belsley, Kuh and Welsch (1980). This method makes it possible to determine the number of quasi-linear relations on the basis of the Condition Indices and the variables involved in each of them, by means of the Variance-decomposition proportion matrix. Thus, this method provides quite an exhaustive picture of the role played by regressors. Although it is the most complete method, it is often difficult to use.

Finally, for a quantitative evaluation of multicollinearity indicators, a number of thresholds were defined on the basis of suggestions given by works on this subject which, in our opinion and experience, seemed to be the most *reasonable*.

3. The system

SUSAM was devised as an *expert system*; as we shall see in more detail, it is organised in distinct modules.

SUSAM's approach is based on interaction with the user. The system controls the whole session by means of a dialogue (*dialogue base*) supervised by a *dialogue manager*; this makes it possible for the system to examine and satisfy all the requests it receives, check the *opinions* of the user about the results, and correct wrong deductions.

At every step the user can ask for *help* to get a general idea about the characteristics of the system, to receive explanation about the theoretical aspects of the method being used, as well as about the symbols and assumptions employed in a specific context (*help screens* and *help manager*). Optionally, the system can show the most *reasonable* way to carry out the analysis.

To perform its functions, the system first arranges the data in the most suitable way for statistical computations (*FORTRAN programs*); then it gets the numerical results and interprets their statistical meaning. Once the system has acquired the necessary *knowledge of facts*, it shows the numerical results asking the student about his deductions in order to check their correctness.

During the analysis the screen is divided into two parts, one above the other: the upper part shows the statistical indicators which are being analysed, while the lower part is reserved for the dialogue with the user.

The various components of SUSAM's architecture are shown in Figure 1. SUSAM's abilities are limited to models with no more than nine regressors. Data can be provided by the user or generated automatically by simulation (see Gunst and Mason, 1980). In the latter case, an algorithm was introduced into the generation process to allow the regressors to be randomly correlated in different ways, producing various forms of multicollinearity with varying strength. To avoid the repetition of similar situations in different sessions, the seed for pseudo-random number generation is determined as a function of the time at which the simulation program is run.

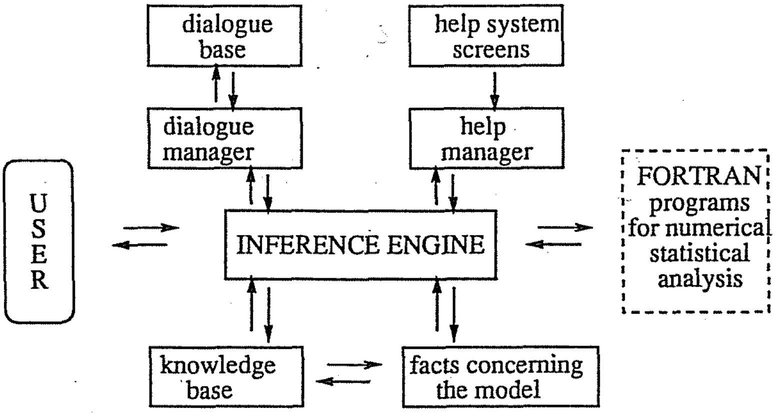


FIGURE 1
SUSAM's architecture

4. How SUSAM teaches

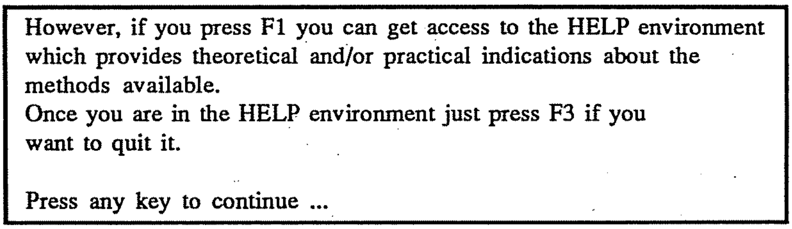
At the beginning of each session SUSAM introduces the student into the environment in a friendly way explaining the possibilities which the system provides. Panels (i)-(iv) show some phases of the initial dialogue between the system and the user.

```
Welcome in SUSAM !
I hope you'll find the work you are
about to start useful and pleasant.
Type in your name so that we
get to know each other.

===> lorena
```

Panel (i)

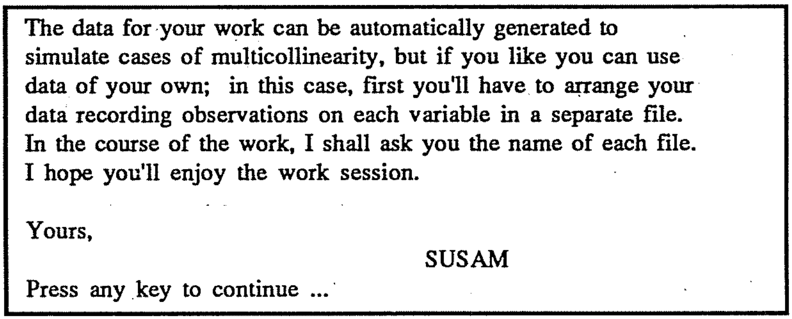
```
First of all, Lorena, remember that if you
want to quit the system you'll have to press F3.
If you wish to have some general information
about the system, please type INFO and hit RETURN;
Otherwise, just press RETURN.
```

Panel (ii)



```
However, if you press F1 you can get access to the HELP environment
which provides theoretical and/or practical indications about the
methods available.
Once you are in the HELP environment just press F3 if you
want to quit it.

Press any key to continue ...
```

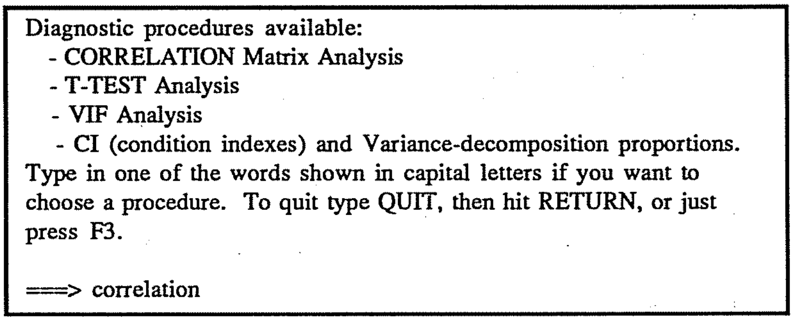
Panel (iii)



```
The data for your work can be automatically generated to
simulate cases of multicollinearity, but if you like you can use
data of your own;  in this case, first you'll have to arrange your
data recording observations on each variable in a separate file.
In the course of the work, I shall ask you the name of each file.
I hope you'll enjoy the work session.

Yours,
                                                SUSAM
Press any key to continue ...
```

Panel (iv)

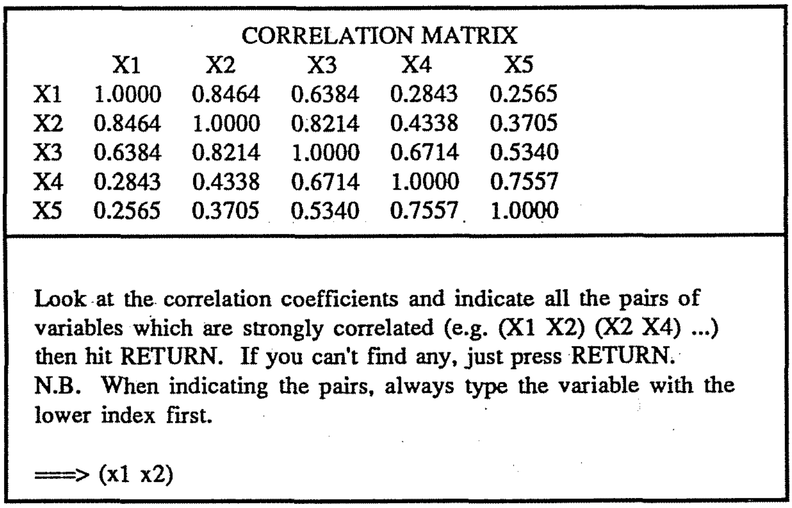
Once the parameters for the work session are established, the system proposes a list of diagnostic methods out of which the student can choose freely (Panel (v)).



```
Diagnostic procedures available:
  - CORRELATION Matrix Analysis
  - T-TEST Analysis
  - VIF Analysis
   - CI (condition indexes) and Variance-decomposition proportions.
Type in one of the words shown in capital letters if you want to
choose a procedure.  To quit type QUIT, then hit RETURN, or just
press F3.

===> correlation
```

Panel (v)

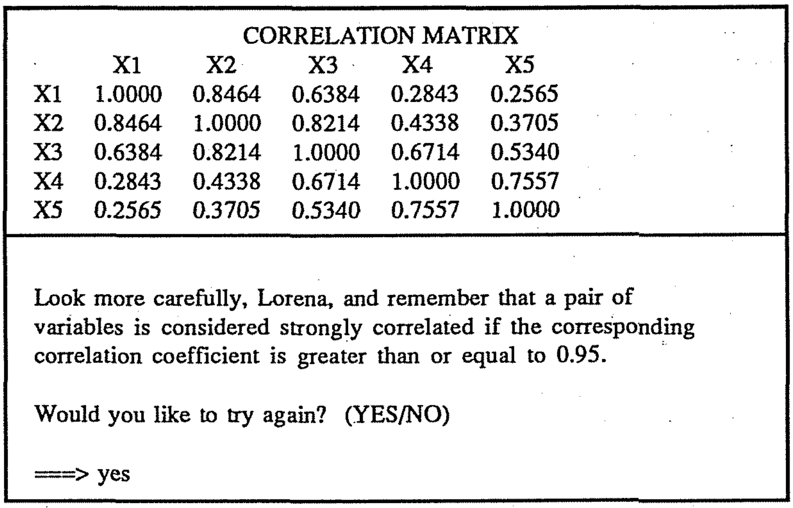
Since the Correlation Matrix Analysis is the first method requested, the system shows the matrix in the upper part of the screen, whereas in the lower part, by means of a series of questions, it asks the student to search for possible values which could reveal the existence of multicollinearity (Panel (vi)).



```
                    CORRELATION MATRIX
        X1       X2       X3       X4       X5
X1    1.0000   0.8464   0.6384   0.2843   0.2565
X2    0.8464   1.0000   0.8214   0.4338   0.3705
X3    0.6384   0.8214   1.0000   0.6714   0.5340
X4    0.2843   0.4338   0.6714   1.0000   0.7557
X5    0.2565   0.3705   0.5340   0.7557   1.0000

Look at the correlation coefficients and indicate all the pairs of
variables which are strongly correlated (e.g. (X1 X2) (X2 X4) ...)
then hit RETURN. If you can't find any, just press RETURN.
N.B. When indicating the pairs, always type the variable with the
lower index first.

===> (x1 x2)
```

Panel (vi)

Since a mistake has been made, SUSAM provides more detailed information to enable the student to provide a correct answer (Panel (vii).

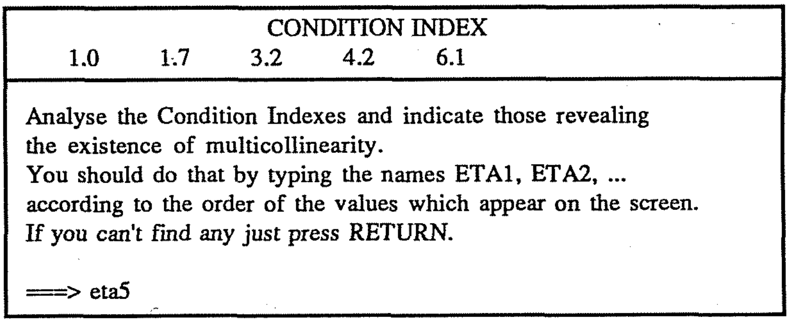


```
                    CORRELATION MATRIX
        X1       X2       X3       X4       X5
X1    1.0000   0.8464   0.6384   0.2843   0.2565
X2    0.8464   1.0000   0.8214   0.4338   0.3705
X3    0.6384   0.8214   1.0000   0.6714   0.5340
X4    0.2843   0.4338   0.6714   1.0000   0.7557
X5    0.2565   0.3705   0.5340   0.7557   1.0000

Look more carefully, Lorena, and remember that a pair of
variables is considered strongly correlated if the corresponding
correlation coefficient is greater than or equal to 0.95.

Would you like to try again?  (YES/NO)

===> yes
```

Panel (vii)

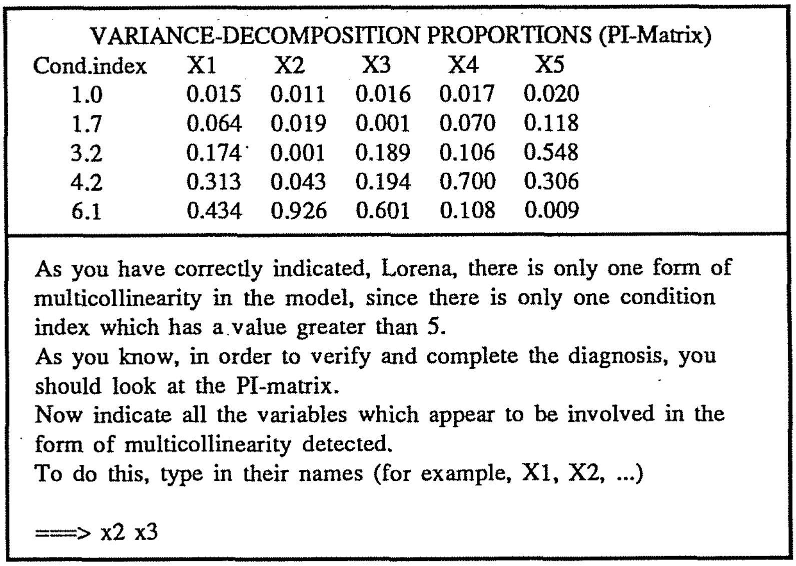
Once the values of the correlation coefficients have been analysed, SUSAM shows a summary table of all information obtained. Then the system asks a question to verify whether the student is aware of the limitations of this method as regards multicollinearity diagnosis. Once the analysis of the Correlation Matrix has been completed, since the student made some mistakes, SUSAM invites her to consider the opportunity of revising her theoretical knowledge.

Panels (viii) and (ix) show two phases of Condition Indices and Variance-decomposition proportions analysis. The student is gradually asked to find out the information about multicollinearity which can be inferred from these indicators.



```
                      CONDITION INDEX
     1.0      1.7      3.2      4.2      6.1

Analyse the Condition Indexes and indicate those revealing
the existence of multicollinearity.
You should do that by typing the names ETA1, ETA2, ...
according to the order of the values which appear on the screen.
If you can't find any just press RETURN.

===> eta5
```

Panel (viii)



```
     VARIANCE-DECOMPOSITION PROPORTIONS (PI-Matrix)
Cond.index   X1      X2      X3      X4      X5
    1.0      0.015   0.011   0.016   0.017   0.020
    1.7      0.064   0.019   0.001   0.070   0.118
    3.2      0.174   0.001   0.189   0.106   0.548
    4.2      0.313   0.043   0.194   0.700   0.306
    6.1      0.434   0.926   0.601   0.108   0.009

As you have correctly indicated, Lorena, there is only one form of
multicollinearity in the model, since there is only one condition
index which has a value greater than 5.
As you know, in order to verify and complete the diagnosis, you
should look at the PI-matrix.
Now indicate all the variables which appear to be involved in the
form of multicollinearity detected.
To do this, type in their names (for example, X1, X2, ...)

===> x2 x3
```

Panel (ix)

5. Environment and availability

At present SUSAM is operative in an MS-DOS environment; it needs a 80286/ 80386 processor with 3M bytes of extended memory and a colour monitor with a VGA graphic card. It consists of two parts: a package of FORTRAN programs for statistical computations, and a set of LISP (Common LISP dialect) programs for the part which is more closely related to the expert system. The help screens were devised using the T_EX system for document composition.

SUSAM is available freely on request from the authors (two high density clear diskettes should be sent for copy). For any information contact the authors (E-mail address: GLM@IPGUNIV.EARN).

Acknowledgement

The research reported in this article was supported by a "Consiglio Nazionale delle Richerche" grant (n. 86.02778.33). The hardware was supplied by IBM-Italia in the framework of a Study Contract.

References

Belsley, D A (1984) Demeaning conditioning diagnostics through centering (with discussion). *The American Statistician* **38(2)**, 73-77.

Belsley, D A, Kuh, E and Welsch, R E (1980) *Regression Diagnostics.* John Wiley, New York.

Gunst, R F (1983) Regression analysis with multicollinear predictor variables : definition, detection and effects. *Communications in Statistics, Theory and Models* **12(19)**, 2217-2260

Gunst, R F and Mason, R L (1980) *Regression Analysis and its Application.* Marcel Dekker, New York.