NOT FOR QUOTATION
WITHOUT PERMISSION
OF THE AUTHOR

ANALYSIS AND DESIGN OF SIMULATION EXPERIMENTS WITH LINEAR APPROXIMATION MODELS

V. Federov
A. Korostelev*
S. Leonov*

October, 1984
WP-84-74

* All-Union Institute of Systems Studies, Moscow, USSR

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
2361 Laxenburg, Austria

FOREWORD

Understanding the nature and dimensions of the world food problem and the policies available to alleviate it has been the focal point of the IIASA Food and Agriculture Program since it began in 1977.

National food systems are highly interdependent, and yet the major policy options exist at the national level. Therefore, to explore these options, it is necessary both to develop policy models for national economies and to link them together by trade and capital transfers. For greater realism the models in this scheme are kept descriptive, rather than normative.

Over the years models of some twenty countries, which together account for nearly 80 percent of important agricultural attributes such as area, production, population, exports, imports and so on, have been linked together to constitute what we call the basic linked system (BLS) of national models.

These models represent large and complex systems. Understanding the key interrelationships among the variables in such systems is not always easy. Communication of results also becomes difficult. To overcome this problem, one may consider approximating these "primary models" by more transparent "secondary models".

In this paper Valeri Federov, A. Korostelev and S. Leonov describe the package of programs for the design and analysis of simulation experiments with such secondary models. The package was prepared in the All-Union Institute of Systems Studies in Moscow. It is one of the first attempts in this field, and we hope that more experience, comments and critiques will help to improve and extend the package in a useful and practical way.

Kirit S. Parikh
Program Leader
Food and Agriculture Program.

ACKNOWLEDGEMENTS

I am very grateful to Lucy Tomsits for editing and typing the paper and to Valerie Khaborov for his help in installing the software at IIASA.

ABSTRACT

There is a necessity in a number of IIASA's researches to deal with analyzing the properties of the computerized versions of complex models. The use of simulation experiments is one of the most successful tools in solving this problem. In this paper, the package of programs for the the design and analysis of simulation experiments is described. The package was prepared in the All-Union Institute of Systems Studies in Moscow. It is one of the first attempts in this field, and the authors did not expect to have constructed a very comprehensive variant, but hope that more experience, remarks and critiques will help to improve and extend the package in a most useful and practical way.

CONTENTS

ANALYSIS AND DESIGN OF SIMULATION EXPERIMENTS WITH LINEAR APPROXIMATION MODELS

by

V. Fedorov, A. Korostelev* and S. Leonov*

1. INTRODUCTION

The construction and computer realization of mathematical models of the natural and social phenomena is nowadays one of the stable tendencies of systems analysis. Sometimes those models are so complicated that they look like "black boxes" even for their authors. That is why methods for the investigation of such models are extremely interesting. The ideas and methods of the simulation experiment are rather old (Naylor, 1971). Some aspects of design and analysis of simulation experiment were described by Fedorov (1983), and we shall follow the concepts of this paper. The main object of our study is a computer realization of a model called *primary model* which is described below.

The aim of this paper is to describe the general structure of the interactive system for design and analysis of simulation experiment, and to show its potentialities.

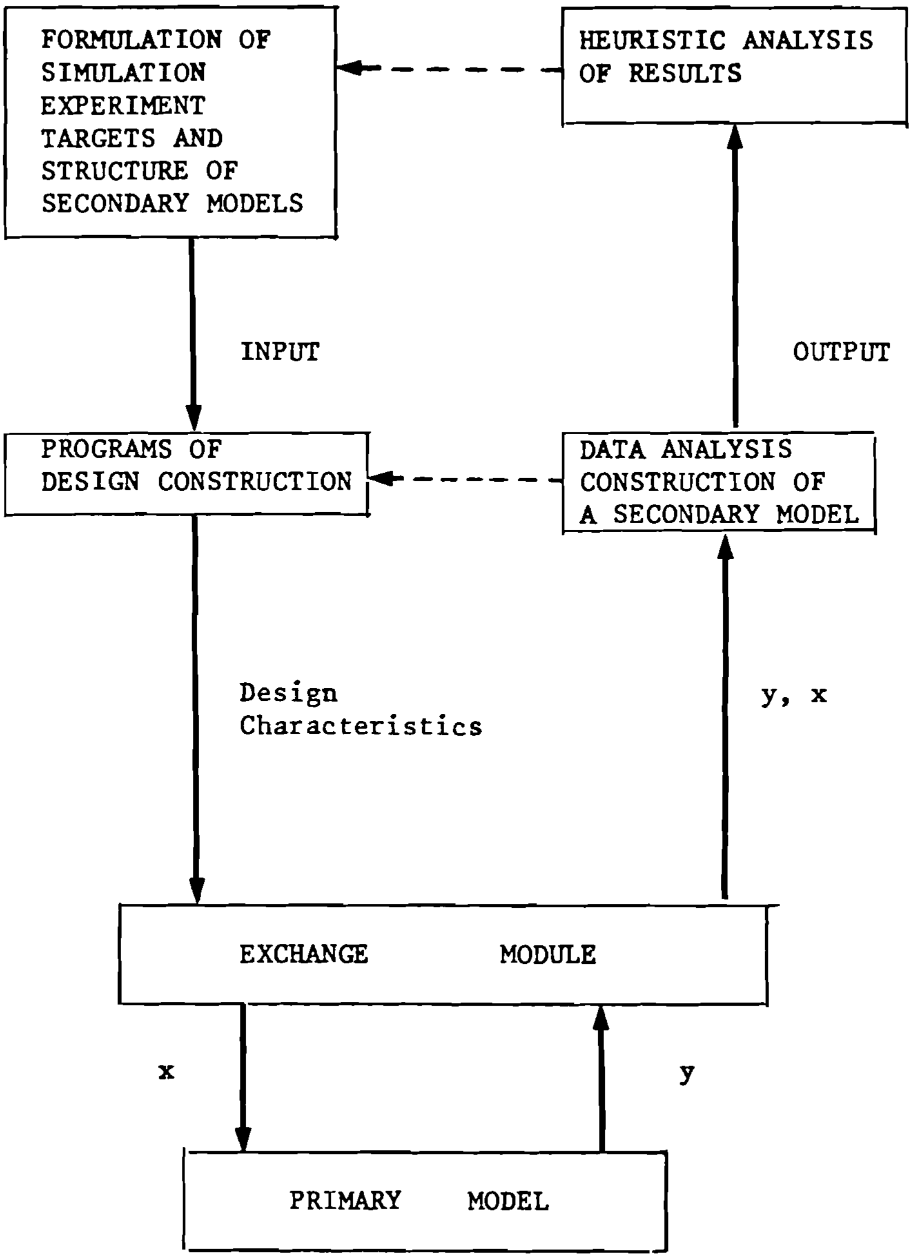
2. STRUCTURE OF THE INTERACTIVE SYSTEM

The current version of the system contains two main programs intended for construction of experimental design and data analysis. These programs are independent of each other and are linked only through input-output files of data. It is necessary to point out that the treatment of any specific model requires an exchange module. This module makes it possible to repeatedly call the primary model varying input data. The principle scheme of interactive system may be illustrated by Figure 1.

The comparatively simple approximation function, methods of optimal design, construction and statistical methods of data analysis, were deliberatedly used in the system. The choice of these simple mathematical tools can be explained as an attempt to balance between the reliability of a *secondary model*; its simplicity and lucidity taking into account the reasonability of the calculation volume. The following sections show the potentialities of programs and are illustrated by test examples. It is necessary to underline that some potentialities not foreseen in the system may be assigned to the exchange module.

3. CONSTRUCTION OF EXPERIMENTAL DESIGN

While investigating the primary model it is assumed that input variables x (factors, independent variables) are separated into groups at the heuristic level according to: Firstly, prior information on their nature; and Secondly, the expected degree of their influence on dependent



KEY: dotted lines show possible "feedbacks"

Figure 1.

variables Y. The factors are usually separated into the following groups:

a) Scenario and exogenous variables;

b) Parameters of the model of which values are obtained on the stage of identification (usually they have rather large intervals of uncertainty);

c) Variables known with "small" errors, which can often be considered as random ones.

The program for the construction of experimental design can generate designs of different types for variables from different groups. In the current version, the following types of designs are available to generate

— Orthogonal design

(i) two-level design $X = \{X_{ij}\}$; where $X_{ij} = \pm 1, i = \overline{1,N}, \; j = \overline{1,m}, i$ is a number of an observation, j is a number of a variable, X is Hadamard matrix, i.e., $X^T X = N I_N$, where I_N is identity matrix, $N = 4k, k$ is a integer number;

(ii) three-level design $X = \{X_{ij}\}, X_{ij} = -1, 0, +1$; X is conference matrix, i.e. $X^T X = (N-1) I_N$; $N = 4k + 2$.

It is recommended to use orthogonal design for group (a), if a detail investigation for the factors from (a) is required;

— Random design with two- and multilevel independent variation of factors. Usually it is used for the factors from group (b).

— Random design for simultaneous variation of all factors of the group. It may be applied for block analysis.

— Random design with continuous law of distribution: uniform and normal. It may be applied for those factors which are known up to small random error.

The criterion for design construction is the correlation coefficients of column vectors of X-matrix: the columns must have correlation which is as small as possible. The design may also be generated (for some groups) in a purely random mannner, without examination of correlations.

There exists a vast literature on the methods of constructing orthogonal design. One of the simplest approach based on the Paley concept (see Hall, 1967) is used here.

Conference matrices (C-matrices)

Let $GF(q)$ be a finite Galois field of cardinality q, $q = p^r$, where p is a prime odd number. Let $R(x)$ be a character defined on $GF(q)$:

$$R(x) = \begin{cases} 0 & x = 0, \\ 1 & \text{if there exists } \gamma \in GF \text{ such that } \gamma^2 = x, \\ -1 & \text{if such } \gamma \text{ does not exist} \end{cases}$$

If $a_0 = 0, a_1, \ldots\ldots, a_{q-1}$ are the elements of $GF(q)$ then a matrix $Q = \{R(a_i - a_j)\}$ is called Jacobsthal matrix and satisfies the equation

$$QQ^T = qI_q - J$$

and

$$QJ = JQ = 0$$

where $J_{ik} = 1$ for all $i, k = \overline{1, q}$. Let

$$C^{q+1} = \begin{bmatrix} 0 & e \\ -e^T & Q \end{bmatrix} \text{ if } q = -1 \ (\text{mod } 4),\ e = (1, 1, \cdots, 1),$$

and

$$C_{q+1} = \begin{bmatrix} 0 & e \\ e^T & Q \end{bmatrix} \text{ if } q = 1 (\text{mod } 4).$$

Then C_{q+1} is C-matrix of order $q+1$.

Hadamard matrices (H-matrices)

Hadamard matrices are constructed on the basis of the concept of Kronecker matrix product. If $q = 1$ (mod 4), then

$$H_{2n} = C_n \otimes \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} + I_n \otimes \begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix}$$

is H-matrix of order $2n$, $n = q+1$. Further, if H_n and H_m are H-matrices of orders n and m, then $H_n \otimes H_m$ is H-matrix of order nm.

It must also be noted that if $q = p^r = -1$ (mod 4), then $H_{q+1} = C_{q+1} + I_{q+1}$ - H-matrix of order $q+1$. With the use of the methods described, the program *PLAN* constructs:

- Hadamard H-matrices for all $n \leq 112$, n is 4-tuple, except $n = 92$;
- Conference C-matrices of the following orders $m (m = 2$ (mod 4)): 6, 14, 18, 26, 30, 38, 42, 54, 62, 74, 90, 98, 102

It is clear that the above methods allow the construction of saturated (number of observation equals to number of factors plus one) orthogonal designs, which are optimal for the majority of statistical criteria, only for the above enumerated dimensions of the factor space. Therefore, for intermediate dimensions, the orthogonal design for the nearest larger dimension has to be used. It will also be orthogonal (but not saturated) in these cases.

There are two variants of the application of generated design $X = \{X_{ij}\}$.

(i) X-matrix is written (row by row) into the auxiliary file (*HELP.DAT*) for application in the exchange module and further analysis of simulation experiment.

(ii) The levels of factors may be set in the real scale: in that case, mean values and scale of variation are chosen by the user. The design in the real scale are obtained with the help of the evident formula

$$FN_{ij} = F_j(1 + v_k X_{ij}),\ i = \overline{1,N};$$

here the j-th factors belongs to the chosen group k; v_k is the scale of variation for group k; F_j stands for their value of the j-th factors. Matrix $FN = \{FN_{ij}\}$ is stored (row by row) into the file *HELP.DAT.*

4. EXPERIMENTAL ANALYSIS

The aim of the simulation of analysis is the construction of secondary model of the following form:

$$y = g(x) = \vartheta_0 + \sum_{\alpha=1}^{k} \vartheta_\alpha f_\alpha(x), \qquad (1)$$

where y is a response (dependent variable); $\vartheta_0, \vartheta_1, \ldots, \vartheta_k$ are parameters to be estimated (regression coefficients); $f_1, f_2, \ldots, f_k$ are known functions depending on x-vector of input variables.

Since k is usually rather large, one of the main problems of experimental analysis is the screening of significant factors. Following is the statement of the problem: input data is set

$$\begin{array}{l} X_{11} \ X_{12} \ldots\ldots X_{1m} \\ X_{21} \ X_{22} \ldots\ldots X_{2m} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ X_{N1} \ X_{N2} \ldots\ldots X_{Nm} \end{array}$$

where N is number of an observation, m is number of variables. One variable is taken as a response and is denoted by y (sometimes y is not a variable itself, it could be some transformation -- the set of the most usable transformations are provided by the program). Then k functions $f_1, f_2, \ldots, f_k$, depending on the rest of variables, are chosen (mainly heuristically) and can be constructed with the help of the above-mentioned transformations of $x_1, x_2, \cdots, x_N$. That is the final step in the formulation of model (1); screening experiments can be carried out now.

Here we shall enumerate the possibilities of the program for the analysis of results provided by simulation experiments.

1) Input variables can be separated into groups with the help of identification vector; variables from one group only are analyzed simultaneously, but identification vector may be changed, and the groups can be rearranged easily.

2) It is possible to make transformations of factors, include their interaction and take any variable as a response.

3) The program provides the stepwise regression procedure; factors may be included into regression or deleted from the equation (Efroymson, 1962). Technically, this program for screening significant factors is based on the subroutines from SSP

package, 1970; some modifications of these subroutines are being carried out for the implementation of interactive regime and Efroymson procedure. Interpretation of input and output information in this module will cause no difficulties for a user familiar with the SSP package.

4) A user may obtain both statistics analogous to SSP subroutines and some additional information, for instance, correlation matrix of regression coefficients, and detailed analysis of residuals.

5) If a secondary model is used for interpolation or extrapolation, values of input variables (predictors) are being chosen by a user. The standard errors of the prognoses are calculated.

6) A heuristic method of random permutations for testing significance of entered variables is provided in the program. Its description is given in section 5.

Program for experimental analysis utilizes 3 files: *SYSIN.DAT* and *SYSOUT.DAT* for input and output information respectively, and an auxiliary file *SYSST.DAT* for intermediate information.

5. STEPWISE REGRESSION WITH PERMUTATIONS

It is well-known (see for instance, Pope & Webster, 1972; Draper, Guttman, and Lapczak, 1979) that the application of standard statistical criteria (F-test, for example) for testing significance of entered variables in the stepwise regression procedure is not correct by its nature. That is why a heuristic method of random permutations is used in the interactive system for testing significance of entered variables. Such an approach enables one to avoid complicated analytical methods that are necessary for calculating statistic of criterion. It must also be underlined that this method does not require the assumptions concerning the distribution of variables. Therefore it may be rather useful in practice (Devyatkina et al., 1981).

Method of random permutations is based on the following concept: two models are compared

based model $y = g(x)$

and a model $\hat{y} = \hat{g}(x)$

where response function $\hat{g}(x)$ is constructed according to permuted values of response: $y_{i_1}, y_{i_2}, \ldots, y_{i_N}$, here $i_1, i_2, \ldots, i_N$ is a random permutation of indexes $1, 2, \ldots, N$. If the first (basic) model gives an adequate approximation of the primary model, then for example, residual sum of squares for the 1st model will be significantly less than for the 2nd model. Such a comparison of statistics usually applied in stepwise procedure for testing adequacy of secondary model, underlies the method of random permutations.

Now we give a short description of the screening algorithm with permutations.

(1) *1st Step.* The most significant variable is entered into regression -- X_{NV_1}. Student's T-statistic (T_0), Fisher's F-statistic (F_0) and SS-statistic (percentage of variance explained on this step, SS_0) are computed.

Random permutation is carried out for all rows of X-matrix except the elements from column M, corresponding to the response function y: let $i_1,...,i_N$ be a random permutation of indexes $1,...,N$. For every l-th permutation, ($l = \overline{1,L}$) stepwise procedure is carried out, the most significant variable is entered into regression and corresponding values of T_l-, F_l- and SS_l-statistics are computed.

(2) *jth Step*, $j > 1$. *jth* variable, X_{NV_j}, is entered into regression; T_0, F_0 and SS_0-statistics are computed for the entered variable.

Random permutation is carried out for all rows of X-matrix except the elements from columns $NV_1, NV_2, ..., NV_{j-1}, M$ (Totally L- permutations should be done). After every permutation stepwise procedure is being carried out, variables $X_{NV_1}, X_{NV_2}, ..., X_{NV_{j-1}}$ are being forced into regression. T_l, F_l, SS_l - statistics are computed at every lth permutation for the variable entered into regression on the jth step.

After jth step ($j \geq 1$) the following information is given:

— index of entered variable, NV_j;

— value of T_0;

— mean and standard deviation of T_l-statistics, $l = \overline{1,L}$, minimal and maximal values of T-statistic after permutations; a histogram for T-statistics (after permutations); percentage of those T_l, for which $|T_l| < |T_0|$.

Analogous information is given for F- and SS-statistics.

If the null hypothesis H_0^j: "response function $y(X)$ is independent of X_{NV_j}" is not satisfied, it seems natural to expect that T_0-value (T-statistic for basic model) is greater (in absolute value) than the "significant" majority of T_l-values (analogously for F- and SS-statistics). A rule for testing null hypothesis can be formulated as follows: null hypothesis H_0^j is rejected with significance level

$$alf_i = \frac{i}{L+1},$$

if F_0-statistic is greater than $(L-i+1)$- values of F_l-statistics after permutations (the same for T- and SS-statistics).

6. EXAMPLE OF SYSTEM UTILIZATION

Let's assume that there is a model with 30 input variables, and we *suspect* that only the first 7 variables have great influence on the output variable y; the next 8 variables may be significant. It is also known that variables 16-23 can take values on three levels $-1,0,+1$; the remaining 7 variables may be continuous and will be treated as a *random noise* in the model.

A priori information concerning input variables in the *primary model* often looks like the one above. Experimental design will be chosen on the basis of this information.

The aim of the experiment is to construct the *secondary model* with a few significant variables. In the model under consideration we will try to approximate the *primary model* by the model with 5-6 variables.

Now let us assume that the true model in the "black box" has the following form:

$$\begin{aligned} y &= 5X_1 + 6X_2 + 7X_3 + 8X_4 + 9X_5 + \\ &+ 10X_6 + X_8 + 2X_9 + 3X_{10} + 4X_{11} - \\ &- X_1X_7 - X_1X_{12}/2 - X_1X_{13}/3 + \\ &+ \text{RANDOM NOISE variables.} \end{aligned}$$

The system's potentialities will be demonstrated with the help of some simple examples using this model. It should be pointed out that these illustrative examples cannot comprehend all features of the system. More detailed information on them are contained in *SYS INSTRUCTION* which are available from the IIASA computer center.

REFERENCES

Devyatkina, G.N., and A.T. Tereokhin (1981) Statistical Hypothesis Tests in Stepwise Regression Analysis Based on the Permutation Method, in *Linear and Nonlinear Parameterization in Experimental Design Problems. Problems in Cybernetics*, V. Fedorov and V. Nalimov (eds.). pp.111-121. Moscow (in Russian).

Draper, N.R., I. Guttman, and L. Lapczak (1979) Actual Rejection Levels in a Certain Stepwise Test. *Commun. Statist. - Theor. Metho.*, A8(2):99-105.

Efroymson, M.A. (1962) Multiple Regression Analysis, in *Mathematical Methods for Digital Computers*, A. Ralston, H. Wilf. (eds.). New York: Wiley.

Fedorov, V. (1983) Analysis and Design of Simulation Experiments for the Approximation of Models. WP-83-71. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Hall, M. (1967) Combinatorics. New York.

Naylor, T.H. (1971) Computer Simulation Experiments with Models of Economic Systems. New York.

Pope, P.T., and J.T. Webster (1972) The Use of an F-Statistics in Stepwise Regression Procedures. *Technometrics* 14:326-340.

System/360 Scientific Subroutine Package, Version 3, Programmer's Manual. New York: IBM, Technical Publications Department, 1970.

************************ *********************	
A NUMBER OF GROUPS AND A NUMBER OF FACTORS-?	In this example the design is generated for 4 groups,
4 30	a number of factors - 30.
************************ *********************	
FACTOR IDENTIFICATION-?	Array (IN): factor I belongs to group IN
1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3	I
3 3 3 4 4 4 4 4 4 4	
************************ *********************	
NUMBER OF GROUPS FOR CORR. ANALYSIS, INCLUDING ORT.GROUP?	A number of groups (including orthogonal group) for
2	analysis of correlations.
**************************************	Design for groups 3,4 is purely random.
THEIR INDEXES?	
1 2	
************************ *********************	
WITH ORTHOGONAL DESIGN:YES-1,NO-0	
1	
**************************************	
ORTHOGONAL GROUP-?	Orthogonal design is generated for group 1 (7 factors).
1	
**************************************	
NUMBER OF FACTORS IN ORTHOGONAL GROUP = 7	
**************************************	
A CODE FOR ORTHOGONAL DESIGN:	
1 - ORTH.DESIGN	
2 - TWO-LEVEL DESIGN (N IS 4-TUPLE)	
3 - THREE-LEVEL DESIGN (N IS 2-TUPLE)	Two-level (+ 1) design is generated with 12 experiments.
1	
NUMBER OF EXPERIMENTS-12	
************************ *********************	
GROUP CODES:1-SIMULTANEOUS VARIATION	Types of a variation which will be used.
2-TWO-LEVEL VARIATION	
3-MULTILEVEL VARIATION	
4-UNIFORM DISTRIBUTION	
5-NORMAL DISTRIBUTION	
1 2 3 4	Codes for group variation.
**************************************	
NUMBER OF GROUPS WITH CODE 3 - 1	
GROUP 3	Multilevel design (3 levels) is generated for group 3.
NUMBER OF LEVELS-?	
3	
LEVELS-?	
-1.00 0. 1.00	Corresponding probabilities and levels.
PROBABILITIES-?	
0.33 0.33 0.34	
**************************************	
NUMBER OF GROUPS WITH CODE 4(UNIFORM DISTR.)- 1	
GROUP 4	Design with uniform (-1, +1) distribution is generatd
INTERVAL OF VARIATION - [-A,A]. YOUR * A * -?	for group 4.
1.00	
************************ *********************	
NUMBER OF EXPERIMENTS N=12	Maximal correlation of factors with preceding ones.
MAXIMAL CORRELATION OF EACH FACTOR	Seven factors are from the orthogonal group, therefore
-0.169 0.333-0.447 0.333-0.333-0.354 0.354-0.354	the 1st number (-0.169) corresponds to factor 8.
************************ *********************	

```
CORRELATION ANALYSIS WAS CARRIED OUT FOR   8 FACTORS
************************    *********************
DESIGN <COLUMN BY COLUMN> FOR GROUPS WITH CORR.ANAL.
 -1.00   1.00  -1.00   1.00  -1.00  -1.00  -1.00   1.00   1.00   1.00
 -1.00   1.00
 -1.00   1.00   1.00  -1.00   1.00  -1.00  -1.00  -1.00   1.00   1.00
  1.00  -1.00
 -1.00  -1.00   1.00   1.00  -1.00   1.00  -1.00  -1.00  -1.00   1.00
  1.00   1.00
 -1.00   1.00  -1.00   1.00   1.00  -1.00   1.00  -1.00  -1.00  -1.00
  1.00   1.00
 -1.00   1.00   1.00  -1.00   1.00   1.00  -1.00   1.00  -1.00  -1.00
 -1.00   1.00
 -1.00   1.00   1.00   1.00  -1.00   1.00   1.00  -1.00   1.00  -1.00
 -1.00  -1.00
 -1.00  -1.00   1.00   1.00   1.00  -1.00   1.00   1.00  -1.00   1.00
 -1.00  -1.00
 -1.00  -1.00   1.00   1.00   1.00   1.00  -1.00   1.00   1.00  -1.00
  1.00  -1.00
 -1.00  -1.00  -1.00  -1.00   1.00  -1.00   1.00  -1.00   1.00   1.00
  1.00   1.00
  1.00  -1.00   1.00   1.00   1.00   1.00   1.00  -1.00   1.00   1.00
  1.00   1.00
 -1.00   1.00   1.00   1.00   1.00  -1.00  -1.00  -1.00  -1.00   1.00
 -1.00   1.00
 -1.00   1.00  -1.00  -1.00   1.00   1.00   1.00  -1.00   1.00  -1.00
 -1.00   1.00
  1.00  -1.00   1.00   1.00  -1.00  -1.00  -1.00  -1.00   1.00  -1.00
 -1.00  -1.00
 -1.00  -1.00   1.00  -1.00  -1.00  -1.00  -1.00   1.00  -1.00  -1.00
  1.00   1.00
 -1.00   1.00   1.00  -1.00  -1.00   1.00   1.00  -1.00  -1.00   1.00
  1.00  -1.00
************************    *********************
DO YOU NEED TO WORK WITH THE VECTOR OF MEANS:YES-1,NO-0
0
************************    *********************
ARRAY NF
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
 21  22  23  24  25  26  27  28  29  30
```

Correlation analysis was carried out for 8 factors (8 factors is the size of group 2).

Design (column by column) for groups with correlation analysis. Format: 10(X, F7.2).

Indicator of input for mean analysis which are usually used when one wish to work with factors in real scales.

Array (NF): design for factor j is the column NF_j of X-matrix. Format - 20I4. Factors from orthogonal group are always in the beginning; factors from groups with purely random design are in the end (last columns of X-matrix).

```
*************************************************************

   NUMBER OF OBSERVATIONS  ?
12

   TOTAL NUMBER OF VARIABLES  ?
31

   IDENTIFICATION VECTOR :
  1  1  1  1  1  1  1  2  2  2  2  2  2  2  2  3  3  3  3  3
  3  3  3  4  4  4  4  4  4  4  0

*************************************************************

   BLOCK  ?
1

     DEPENDENT VARIABLE  ?
31

   WITH INTERACTION  ? ( YES - 1, NO - 0)
0

*************************************************************
A NUMBER OF OBSERVATIONS        12

             TABLE OF VARIABLES
        FORMAL VARIABLE   1  REAL VARIABLE   1
        FORMAL VARIABLE   2  REAL VARIABLE   2
        FORMAL VARIABLE   3  REAL VARIABLE   3
        FORMAL VARIABLE   4  REAL VARIABLE   4
        FORMAL VARIABLE . 5  REAL VARIABLE   5
        FORMAL VARIABLE   6  REAL VARIABLE   6
        FORMAL VARIABLE   7  REAL VARIABLE   7

********** STEPWISE PROCEDURE **********
VARIABLE     MEAN       STANDARD
   NO.                  DEVIATION
    1       0.            1.04447
    2       0.            1.04447
    3       0.            1.04447
    4       0.            1.04447
    5       0.            1.04447
    6       0.            1.04447
    7       0.            1.04447
    8       2.16667      22.90561
```

COMMENTS TO THE LISTING OF PROGRAM FOR EXPERIMENTAL ANALYSIS

Number of observations is 12.

Total number of variables is 31.

Identification vector separates variables into 4 groups.

The influence of the 1st block on response function is analyzed on this stage.

Variable No.31 is the response function.

This coincidence of formal and real variables' numbers is by an incident: it is connected with numeration of variables in the 1st block.

Beginning of screening analysis.

Means and standard deviations of variables.

The response function here is the 8th variable.

```
CORRELATION MATRIX
ROW  1
    1.00000     0.          0.          0.          0.          0.          0.          0.24066
ROW  2
    0.          1.00000     0.          0.          0.          0.          0.          0.37999
ROW  3
    0.          0.          1.00000     0.          0.          0.          0.          0.41799
ROW  4
    0.          0.          0.          1.00000     0.          0.          0.          0.45852
ROW  5
    0.          0.          0.          0.          1.00000     0.          0.          0.42305
ROW  6
    0.          0.          0.          0.          0.          1.00000     0.          0.44332
ROW  7
    0.          0.          0.          0.          0.          0.          1.00000     0.08360
ROW  8
    0.24086     0.37999     0.41799     0.45852     0.42305     0.44332     0.08380     1.00000
```

Correlation matrix of input variables and response function. The part of this matrix is the identity matrix due to orthogonality of the design.

The last line is the correlation of the response with input variables.

```
NUMBER OF SELECTION    1
CODES
 0  0  0  0  0  0  0

*****************  STEP  6  *****************
VARIABLE ENTERED..... 1
SUM OF SQUARES REDUCED IN THIS STEP....       334.259
PROPORTION REDUCED IN THIS STEP........         0.056
CUMULATIVE SUM OF SQUARES REDUCED......      5556.482
CUMULATIVE PPOPORTION REDUCED..........         0.963   OF     5771.334
   MULTIPLE CORRELATION COEFFICIENT...     0.981
   F-VALUE FOR ANALYSIS OF VARIANCE...    21.552
   STANDARD ERROR OF ESTIMATE.........     6.555
 VARIABLE        REG. COEFF.       ERROR          T-VALUE
      4          10.05556          1.89232          5.314
      6           9.72222          1.89232          5.138
      5           9.27778          1.89232          4.903
      3           9.16667          1.89232          4.844
      2           8.33333          1.89232          4.404
      1           5.27778          1.89232          2.789
  INTERCEPT       2.16667
------------------------------------------------------------
 ***** END ***** STEPWISE PROCEDURE ***** END *****
```

Regression analysis after 6 steps. Six variables from the 1st block are entered.

Multiple correlation coefficient 0.981 is large enough. Student's T-statistics show possible significance of variables 1 - 6. (But the distribution of T-value for stepwise estimator does not coincide with classical Student's distribution!)

End of the first stage of the screening analysis.

```
    BLOCK  ?
 2

      DEPENDENT VARIABLE  ?
 31

    WITH INTERACTION  ? ( YES - 1, NO - 0)
 0
```

Now influence of variables 8 - 15 on the response is analyzed.

```
*************************************************************
 A NUMBER OF OBSERVATIONS       12

                TABLE OF VARIABLES
         FORMAL VARIABLE   1  REAL VARIABLE   8
         FORMAL VARIABLE   2  REAL VARIABLE   9
         FORMAL VARIABLE   3  REAL VARIABLE  10
         FORMAL VARIABLE   4  REAL VARIABLE  11
         FORMAL VARIABLE   5  REAL VARIABLE  12
         FORMAL VARIABLE   6  REAL VARIABLE  13
         FORMAL VARIABLE   7  REAL VARIABLE  14
         FORMAL VARIABLE   8  REAL VARIABLE  15

 ********** STEPWISE PROCEDURE **********
VARIABLE      MEAN       STANDARD
   NO.                   DEVIATION
    1       0.16667       1.02986
    2       0.            1.04447
    3       0.66667       0.77850
    4       0.            1.04447
    5       0.            1.04447
    6      -0.33333       0.98473
    7      -0.33333       0.98473
    8       0.            1.04447
    9       2.16667      22.90561
NUMBER OF SELECTION     1
CODES
 0  0  0  0  0  0  0  0

 *****************  STEP  1  *****************
VARIABLE ENTERED..... 4
SUM OF SQUARES REDUCED IN THIS STEP....       3050.704
PROPORTION REDUCED IN THIS STEP........          0.529
CUMULATIVE SUM OF SQUARES REDUCED......       3050.704
CUMULATIVE PPOPORTION REDUCED..........          0.529   OF      5771.334
   MULTIPLE CORRELATION COEFFICIENT...     0.727
   F-VALUE FOR ANALYSIS OF VARIANCE...    11.213
   STANDARD ERROR OF ESTIMATE.........    16.494
 VARIABLE         REG. COEFF.        ERROR            T-VALUE
     4            15.94444          4.76150             3.349
  INTERCEPT        2.16667
---------------------------------------------------------------

 *****************  STEP  2  *****************
VARIABLE ENTERED..... 5
SUM OF SQUARES REDUCED IN THIS STEP....        715.593
PROPORTION REDUCED IN THIS STEP........          0.124
CUMULATIVE SUM OF SQUARES REDUCED......       3766.297
CUMULATIVE PPOPORTION REDUCED..........          0.653   OF      5771.334
   MULTIPLE CORRELATION COEFFICIENT...     0.808
   F-VALUE FOR ANALYSIS OF VARIANCE...     8.453
   STANDARD ERROR OF ESTIMATE.........    14.926
 VARIABLE         REG. COEFF.        ERROR            T-VALUE
     4            15.94444          4.30873             3.700
     5             7.72222          4.30873             1.792
  INTERCEPT        2.16667
---------------------------------------------------------------
 ***** END ***** STEPWISE PROCEDURE ***** END *****
```

Formal and real numbers of variables of the 2nd block differ from each other.

Formal variables 4 and 5, i.e., real variables 11 and 12, explain 65.3% of variance. That's why variables 1 - 12 have to be analyzed jointly.

```
   IDENTIFICATION VECTOR :
  1  1  1  1  1  1  1  1  1  1  1  1  2  2  2  2  2  2  2  2
  2  2  2  2  2  2  2  2  2  2  0
```

After new identification they form block 1, the rest are in block 2.

```
************************************************************

   BLOCK  ?
 1

     DEPENDENT VARIABLE  ?
 31

   WITH INTERACTION  ? ( YES - 1, NO - 0)
 0

************************************************************
 A NUMBER OF OBSERVATIONS        12

              TABLE OF VARIABLES
         FORMAL VARIABLE   1  REAL VARIABLE   1
         FORMAL VARIABLE   2  REAL VARIABLE   2
         FORMAL VARIABLE   3  REAL VARIABLE   3
         FORMAL VARIABLE   4  REAL VARIABLE   4
         FORMAL VARIABLE   5  REAL VARIABLE   5
         FORMAL VARIABLE   6  REAL VARIABLE   6
         FORMAL VARIABLE   7  REAL VARIABLE   7
         FORMAL VARIABLE   8  REAL VARIABLE   8
         FORMAL VARIABLE   9  REAL VARIABLE   9
         FORMAL VARIABLE  10  REAL VARIABLE  10
         FORMAL VARIABLE  11  REAL VARIABLE  11
         FORMAL VARIABLE  12  REAL VARIABLE  12

 ********** STEPWISE PROCEDURE **********
VARIABLE     MEAN       STANDARD
   NO.                  DEVIATION
    1       0.            1.04447
    2       0.            1.04447
    3       0.            1.04447
    4       0.            1.04447
    5       0.            1.04447
    6       0.            1.04447
    7       0.            1.04447
    8       0.16667       1.02986
    9       0.            1.04447
   10       0.66667       0.77850
   11       0.            1.04447
   12       0.            1.04447
   13       2.16667      22.90561
NUMBER OF SELECTION    1
CODES
 0  0  0  0  0  0  2  2  2  2
 0  0
```

Variables may be deleted during the screening analysis.
Variables 7 - 10 are deleted here.

**************** STEP 8 ****************

VARIABLE ENTERED.....12

SUM OF SQUARES REDUCED IN THIS STEP.... 56.004

PROPORTION REDUCED IN THIS STEP........ 0.010

CUMULATIVE SUM OF SQUARES REDUCED...... 5710.720

CUMULATIVE PPOPORTION REDUCED.......... 0.989 OF 5771.334

MULTIPLE CORRELATION COEFFICIENT... 0.995

F-VALUE FOR ANALYSIS OF VARIANCE... 35.330

STANDARD ERROR OF ESTIMATE......... 4.495

VARIABLE	REG. COEFF.	ERROR	T-VALUE
11	5.03509	1.99695	2.521
6	8.73100	1.42766	6.116
4	7.38597	1.63050	4.530
5	6.60819	1.63050	4.053
3	8.47953	1.51792	5.586
2	6.65497	1.45837	4.563
1	3.59942	1.45837	2.468
12	2.97368	1.78613	1.665
INTERCEPT	2.16667		

Eight variables explain 98.9% of variance, but they essentially differ in significance (cf. T-values). Therefore it is naturally to suggest that deleting some of them will not deteriorate our approximation essentially.

********* DELETING PROCEDURE *********

**************** STEP 7 ****************

VARIABLE ENTERED..... 1

SUM OF SQUARES REDUCED IN THIS STEP.... 142.091

PROPORTION REDUCED IN THIS STEP........ 0.025

CUMULATIVE SUM OF SQUARES REDUCED...... 5654.715

CUMULATIVE PPOPORTION REDUCED.......... 0.980 OF 5771.334

MULTIPLE CORRELATION COEFFICIENT... 0.990

F-VALUE FOR ANALYSIS OF VARIANCE... 27.708

STANDARD ERROR OF ESTIMATE......... 5.400

VARIABLE	REG. COEFF.	ERROR	T-VALUE
11	4.29167	2.33806	1.636
6	9.72222	1.55871	6.237
4	8.62500	1.74269	4.949
5	7.84722	1.74269	4.503
3	7.73611	1.74269	4.439
2	6.90276	1.74269	3.961
1	3.84722	1.74269	2.208
INTERCEPT	2.16667		

Two variables are deleted by the backward procedure.

********* DELETING PROCEDURE *********

```
***************** STEP 6 *****************
VARIABLE ENTERED..... 1
SUM OF SQUARES REDUCED IN THIS STEP....      334.259
PROPORTION REDUCED IN THIS STEP........        0.058
CUMULATIVE SUM OF SQUARES REDUCED......     5556.483
CUMULATIVE PPOPORTION REDUCED..........        0.963  OF    5771.334
   MULTIPLE CORRELATION COEFFICIENT...     0.981
   F-VALUE FOR ANALYSIS OF VARIANCE...    21.552
   STANDARD ERROR OF ESTIMATE.........     6.555
 VARIABLE        REG. COEFF.       ERROR         T-VALUE
      4           10.05556        1.89232          5.314
      6            9.72222        1.89232          5.138
      5            9.27778        1.89232          4.903
      3            9.16667        1.89232          4.844
      2            8.33333        1.89232          4.404
      1            5.27778        1.89232          2.789
  INTERCEPT        2.16667
----------------------------------------------------------------
```

Variables 11 and 12 have been deleted without essential increasing of the sum of residual sequals. The most significant variables in the equation are variables 1 - 6.

Here is the secondary model:

$$\hat{y} = 2.17 + 5.28 X_1 + 8.33 X_2 + 9.17 X_3 + 10.06 X_4 + 9.28 X_5 + 9.72 X_6 .$$

*** MODEL ANALYSIS AND FORECASTING ***

REGRESSION EQUATION

NO.	SIGNIFICANT VARIABLES		RESPONSE	S. D.	CORRELATION MATRIX OF REGRESSION COEFFICIENTS (PERCENTAGE)					
			DEP. VAR. X(31)= 2.167							
1	VARIABLE	4	+ 10.056 * X(4) (5.314)	1.892	100	0	0	0	0	0
2	VARIABLE	6	+ 9.722 * X(6) (5.136)	1.892	0	100	0	0	0	0
3	VARIABLE	5	+ 9.278 * X(5) (4.903)	1.892	0	0	100	0	0	0
4	VARIABLE	3	+ 9.167 * X(3) (4.844)	1.892	0	0	0	100	0	0
5	VARIABLE	2	+ 8.333 * X(2) (4.404)	1.892	0	0	0	0	100	0
6	VARIABLE	1	+ 5.278 * X(1) (2.789)	1.892	0	0	0	0	0	100

Such correlation matrix of regression coefficients is due to orthogonality of the design.

```
TEST STATISTICS
CUMULATIVE PROPORTION REDUCED.......   0.963
MULTIPLE CORRELATION COEFFICIENT....   0.981
STANDARD ERROR ESTIMATE (SIGMA).....   6.555
F-VALUE FOR ANALYSIS OF VARIANCE....  21.552
```

*** MODEL ESTIMATION ***

TABLE OF RESIDUALS

NO.	DEPENDENT VARIABLE X(31)	RESPONSE PREDICTION	RESIDUALS ABSOLUTE	RESIDUALS PERCENTAGE	STANDARD DEVIATION	RATIO RESID./S.D.	COMMENTS
1	-50.1667	-49.6667	-0.5000	1.0	4.6352	-0.1079	
2	29.8333	35.6667	-5.8333	19.6	4.6352	-1.2585	
3	25.8333	23.3333	2.5000	9.7	4.6352	0.5393	
4	20.1667	18.7778	1.3889	6.9	4.6352	0.2996	
5	12.1667	5.6667	6.5000	53.4	4.6352	1.4023	
6	4.1667	6.6667	-2.5000	60.0	4.6352	-0.5393	
7	-7.8333	-10.1111	2.2778	29.1	4.6352	0.4914	
8	-25.1667	-20.5556	-4.6111	18.3	4.6352	-0.9948	
9	-0.8333	-3.0000	2.1667	260.0	4.6352	0.4674	
10	-1.1667	-4.1111	2.9444	252.4	4.6352	0.6352	
11	-2.8333	5.4444	-8.2778	292.2	4.6352	-1.7858	
12	21.8333	17.8889	3.9444	18.1	4.6352	0.8510	

* -OUTLIER WITH P=0.05
** -OUTLIER WITH P=0.01

Comments mark outliers (if there are present).

ATTEMPT 1

***** FORECASTING *****

ALL VARIABLES IN THE MODEL ARE ON THE AVERAGE LEVELS EXCEPT THE FOLLOWINGS:

X(1) = 0.1000
X(2) = 0.1000
X(3) = 0.1000
X(4) = 0.1000
X(5) = 0.1000
X(6) = 0.1000
X(7) = 0.1000
X(8) = 0.
X(9) = 0.
X(10) = 0.
X(11) = 0.
X(12) = 0.

FORECASTED VALUE OF DEPENDENT VARIABLE IS 2.1667 WITH S.D. 6.555

Forecasting: what is the value of the response estimator y (in the secondary model) under such values of input variables?

(X(1) = 0.1, X(2) = 0.1, etc.)

Here is the answer:

$\hat{y} = 2.1677 \pm 6.572.$

***** END ***** FORECASTING ***** END *****

****

IDENTIFICATION VECTOR :

1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2
2 2 2 3 3 3 3 3 3 3 0

STEPWISE REGRESSION WITH PERMUTATIONS

```
************************************************************

   BLOCK  ?
 1

     DEPENDENT VARIABLE  ?
 31

   WITH INTERACTION  ? ( YES - 1, NO - 0)
 0

************************************************************
 A NUMBER OF OBSERVATIONS        12

             TABLE OF VARIABLES
        FORMAL VARIABLE   1  REAL VARIABLE   1
        FORMAL VARIABLE   2  REAL VARIABLE   2
        FORMAL VARIABLE   3  REAL VARIABLE   3
        FORMAL VARIABLE   4  REAL VARIABLE   4
        FORMAL VARIABLE   5  REAL VARIABLE   5
        FORMAL VARIABLE   6  REAL VARIABLE   6
        FORMAL VARIABLE   7  REAL VARIABLE   7
        FORMAL VARIABLE   8  REAL VARIABLE   8
        FORMAL VARIABLE   9  REAL VARIABLE   9
        FORMAL VARIABLE  10  REAL VARIABLE  10
        FORMAL VARIABLE  11  REAL VARIABLE  11
        FORMAL VARIABLE  12  REAL VARIABLE  12
********  STEPWISE REGRESSION WITH PERMUTATIONS  ********
************************    ************************
NUMBER OF OBSERVATIONS   12
NUMBER OF VARIABLES      13
************************************
```

A regression model with 13 variables is analyzed,
a number of observations - 12.

```
SCALE FOR A HISTOGRAM:A POINT CORRESPONDS TO 1 VALUE/S/
NUMBER OF INTERVALS FOR A HISTOGRAM  5
```

A scale and a number of intervals for a histogram.

```
MAX.NUMBER OF STEPS 12
NUMBER OF PERMUTATIONS  30
************************    ************************
```

Maximal number of steps - 12,
a number of permutations - 30.

```
TOTAL NUMBER OF DELETED VARIABLES-?
 1
************************************
THEIR INDEXES-?
 6
************************    ************************
************************    ************************
```

Variable 6 is deleted from the regressional equation.

```
 ENTERED VARIABLES AND THEIR T-STATISTICS,STEP=  1
  11
   3.35
************************************
STEP  1
VARIABLE ENTERED 11
```

Variable 11 is entered into regression on the 1st step,
its T-statistics is $T_0 = 3.35$.

```
************************************
BASIC T-STATISTIC     3.35
T-MEAN AND ST.DEVIATION    0.5945   2.2836
MIN. AND MAX.T-STATISTICS AFTER PERMUTATION
  -2.84   4.40
HISTOGRAM FOR T-STATISTIC
   -2.8446    -1.3949  ...........*
   -1.3949     0.0549  *
    0.0549     1.5048  *
    1.5046     2.9544  ..................*
    2.9544     4.4042  .*
% OF T,FOR WHICH ABS(T) IS LESS THAN ABS(TO)   96.67
```

Mean and standard deviation of T-statistic after permutation: 0.5945 and 2.2836.

Minimal and maximal values of T-statistic after permutation: -2.84 and 4.40.

Histogram for T-statistic: one point corresponds to one value (according to assigned scale).

96.67% of T-statistics after permutation satisfy the inequality:

$$|T_1| < 3.35 = T_0$$

```
************************     ************************


************************************
BASIC F-STATISTIC      11.213
F-MEAN AND ST.DEVIATION       5.395       3.107
MIN. AND MAX.F-STATISTICS AFTER PERMUTATION
     2.588    19.397
HISTOGRAM FOR F-STATISTIC
    2.5884      5.9501  .....................*
    5.9501      9.3117  .......*
    9.3117     12.6733  *
   12.6733     16.0350  *
   16.0350     19.3966  .*
% OF F,LESS THAN FO    96.67
************************     ************************
```

Analogous information is given for F - and SS - statistics.

```
************************************
BASIC SS-PROPORTION REDUCED  0.5286
SS-MEAN AND ST.DEVIATION    0.3330   0.0934
MIN. AND MAX.SS-PROPORTIONS AFTER PERMUTATION
 0.2056 0.6598
HISTOGRAM FOR SS-STATISTIC
    0.2056      0.2965  ...........*
    0.2965      0.3873  .............*
    0.3873      0.4781  .....*
    0.4781      0.5690  *
    0.5690      0.6598  .*
% OF SS,LESS THAN SSO   96.67


************************     ************************
************************     ************************
************************     ************************
 ENTERED VARIABLES AND THEIR T-STATISTICS,STEP=  2
  11  12
   3.70   1.79
```

Variables that are entered into regression afer 2 steps.
Corresponding T-statistics.
Variable 12 is entered into regression on the 2nd step.