Arbeidsnotater

STATISTISK SENTRALBYRÅ

IB 64/5
Ref.: SN/AB 19/10-64

Oslo, 1. desember 1964

Multiple Regression and Correlation Analysis

Av Svein Nordbotten og Thor Aastorp

I N N H O L D

SN/AB, 19/10-64

MULTIPLE REGRESSION AND CORRELATION ANALYSIS[1)]

1. GENERAL DESCRIPTION

a. This program is a modification and extension of BIMD 29. The primary modification is that the independent variables are listed in the order of their importance based on the reduction of sum of squares of the dependent variable attributable successively to each independent variable. Additional output includes cumulative: (1) standard errors of estimate, (2) sums of squares, (3) proportions of variance, (4) F-values, and (5) multiple correlation coefficients.

b. The maximum number of variables which can be processed by this program is 50 variables.

c. The upper limitation on sample size is 99,999, and the lower limitation on sample size is two greater than the number of variables.

d. This program can perform a different transformation of any variable and generate new variables if desired, according to the codes specified in Trans-generation cards. A transgeneration can be made conditional if desired, according to the condition codes specified in the Trans-generation cards. Any number of variables can be generated, but the total number of variables must not exceed 50.

e. Any variable, original or generated, can be named the dependent variable. There is no limit to the number of replacements.

f. The maximum number of variables which can be deleted at one time is 32. However, there is no limit to the number of deletions of different sets of variables.

g. The format for input data must be specified in the Variable Format Card(s).

h. The sums, sums of squares and cross-products, in the format produced by this program, can in a later run be used as input if additional replacements and deletions are required.

---

1) This program and its description are both modifications of the BIMD 29 program and is copied only for use within the Central Bureau of Statistics.

2. CARD PREPARATION

a. CONTROL CARDS

b. PROGRAM CONTROL CARDS

(1) PROBLEM CARD (One Problem Card for each problem)

Col. 1,2 PR (for PRoblem Card)

Col. 3-6 Problem number (May be alphabetical characters).

Col. 7,8 Number of original variables, p, ($p \leq 50$)

Col. 9-13 Sample size, n

Col. 14-16 000 no trans-generation
m m trans-generation cards ($m \leq 999$)

Col. 17,18 00 no variables added to original set after trans-generation
q q variables added to original set after trans-generation

Col. 19-21 Number of Replacement and Deletion Cards

Col. 22-26 Per cent of sum of squares to limit variables entering in a regression model. Keypunch a value with a decimal point. (See Note 1 below)

Col. 71,72 k k variable format cards ($k \leq 5$)

Col. 80 1 1 sums, sums of squares and cross-products is used as input

Note: (1) The choice of percentage value for limiting variable will depend upon the purpose of the analysis. A suggested trial value is one per cent, 0.01. If all independent variables are to be included in a regression model, keypunch 0.0.

(2) TRANS-GENERATION CARD(S)

If a non-zero number is specified in Col. 14-16 of the Problem Card, the same number of Trans-generation Card(s) must be prepared. Different types of transformation can be performed successively on the same variable, if desired. The format of Trans-generation Card is as follows:

Col. 1,2 TR (for TRans-generation Card)

Col. 3,4 Variable number to be assigned on transformed or generated variables

Col. 5,6 Transformation code. Codes 01, 02, ... 14 of the transgeneration list may be used. (See Appendix 2)

Col. 7,8 A-variable number

Col. 9-14 B-variable number or constant C (keypunch with decimal point). If the constant is a negative value, keypunch a minus sign to the left of the constant.

Col. 15-16 Condition code. Codes 01, 02 ..., 12 of the condition list may be used. (See Appendix 2)

Col. 17-18 M-variable number

Col. 19-24 N-variable number or constant K (keypunch with decimal point). If the constant is a negative value, keypunch a minus to the left of the constant

(3) VARIABLE NAME CARD(S)

The purpose of Variable Name Card(s) is to identify the variables in the output by their names. The format of Variable Name Card is as follows:

Col. 1-8 Name of 1st variable
Col. 9-16 Name of 2nd variable
Col. 17-24 Name of 3rd variable
. .
. .
Col. 73-80 Name of 10th variable

If there are more than 10 variables, continue keypunching on the second card in the same manner.

If variables are to be generated, the names of these variables must also be keypunched in addition to those of the original variables.

If the identification of the variables by their names is not desired, use blank card(s) for this purpose. Variable Name Card(s) must be present.

(4) STANDARD SCALE CARD(S)

The purpose of the Standard Scale Cards is to scale the original variables by 10^e. The exponents are specified in the Standard Scale Card(s) as follows:

Col. 1-8 : Scale factor for the 1st original variable

- -

- -

Col. 73-80 : Scale factor for the 10th original variable

If there are more than ten variables, scale factor specification is continued on a second Card in the same manner. Decimal point may be punched. The standard scale cards must be present.

(5) VARIABLE FORMAT CARD(S)

The same number of Variable Format Card(s) as specified in col. 71 and 72 of the Problem Card must be prepared. The variable format description can use all 80 columns.

(6) REPLACEMENT AND DELETION CARD(S)

This card, R-D Card, has a fourfold purpose:

1) It indicates the replacement of the dependent variable
2) It indicates the deletion of variables
3) It indicates a replacement and deletion, and
4) It controls the output.

The replacement of a dependent variable and deletion of different sets of variables can be made as many times as desired. This means that it is possible to make multiple regression and correlation analyses with selected sets of variables. The program will retain all the variables, original and generated, until all R - D Cards for one problem are processed.

The program does not give any multiple regression and correlation analysis unless R - D Cards are prepared. R - D Card(s) must be present, and the number of R - D Card(s) prepared must agree with the number specified in Col. 19-21 of the Problem Card.

The format of R - D Card is as follows:

Col. 1,2 RD (for Replacement and Deletion Card)
Col. 3,4 02 If the table of residuals and analysis of extreme

residuals are desired. (See Note 2 below).

01 if only the analysis of extreme residuals is desired.

00 if neither one is desired.

Col. 5,6 A variable to be treated as the dependent variable.

Col. 7,8 Total number of variables to be deleted.

Col. 9,10 1st variable to be deleted.

Col. 11,12 2nd variable to be deleted

. .

. .

Col. 71,72 32nd variable to be deleted

Note: (2) When the sample size is large, 02 punch should not be used.

The maximum number of variables which can be deleted at one time is 32.

Although replacements and deletions are repeated over and over again, the arramgement of variables will not change. Therefore, the number of the next dependent variable and the numbers of variables to be deleted must be stated in terms of the basic set of variables, which may be the original set of variables or the set of variables after trans-generation.

Delete low numbered variables first; namely, X_2 comes first and then X_7.

(7) FINISH CARD

This card will notify the program of the end of the entire job. This card has the following format:

Col. 1-6 FINISH

c. STANDARD INPUT DATA RECORDS

This records can be read either from cards or magnetic tape.

3. ORDER OF CARDS

More than one problem can be processed consecutively. The following example illustrates the order of cards:

...

1)	Problem Card for the first problem	Log.unit.	4
2)	Trans-generation Cards (if used)	"	4
3)	Variable Name Cards	"	4
4)	Scale Card(s)	"	4
5)	Variable Format Cards	"	4
6)	Input Data Cards or Tape	"	3
7)	R-D Card(s)	"	4
	...		
	...		
..	Problem Card for the last problem	Log.unit.	4
..	Trans-generation Cards (if used)	"	4
..	Variable Name Cards	"	4
..	Scale Card(s)	"	4
..	Variable Format Cards	"	4
..	Input Data Cards or Tape	"	3
..	R-D Card(s)	"	4
..	Finish Card	"	4

4. LOGICAL UNITS

a) Logical Unit 2 is used for output

b) Logical Unit 3 is used for input data

c) Logical Unit 4 is used for Program Control Cards

d) Logical Unit 5 is used as intermidiate store

5. REFERENCES

a. Dixon and Massey, Introduction to Statistical Analysis; pp 275-278. McGraw-Hill Book Company, Inc.; 1957.

b. Ostle, Bernard, Statistics in research; Chapter 8. The Iowa State College Press; 1954.

c. Bennett and Franklin, Statistical Analysis in Chemistry and the Chemical Industry; Appendix 6A. John Wiley and Sons, Inc.; 1954.

APPENDIX 1

Computational Procedures

Model assumed:

Y is normally distributed with mean μ and variance σ^2,
where $\mu = A + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$.

The computation steps performed by the program follow:

Step 1. Preliminary computations.

(1) Trans-generation: If desired, data are transformed and/or generated according to the codes specified in the Trans-generation Cards.

(2) Sums: ΣX_i $\quad i = 1,2 \ldots, p$
where p is the number of variables.

(3) Sums of squares: ΣX_i^2

(4) Means: $\bar{X}_i = \frac{\Sigma X_i}{n}$ where n is sample size.

(5) Standard deviations: $s_i = \sqrt{\frac{\Sigma(X_i - \bar{X}_i)^2}{n-1}}$

(6) Cross product sums:

$$P_{ij} = \Sigma X_i \cdot X_j \qquad \begin{matrix} i = 1,2,\ldots,p \\ j = 1,2,\ldots,p \end{matrix}$$

(7) Cross product of deviations:

$$D_{ij} = \Sigma(X_i - \bar{X}_i)(X_j - \bar{X}_j)$$

(8) Simple correlation coefficients:

$$r_{ij} = \frac{\Sigma(X_i - \bar{X}_i)(X_j - \bar{X}_j)}{\sqrt{\Sigma(X_i - \bar{X}_i)^2}\sqrt{\Sigma(X_j - \bar{X}_j)^2}}$$

Step 2. The program reads the Replacement and Deletion Card and performs the following computations:

(1) Rearrange the variables in the cross product of deviation matrix and construct a_{ij} in the working storage area.

(i = 1,2, ..., k; j = 1,2,..., k; where k is the number of variables after deletion.)

(2) Compute: $$b_{iy} = \frac{a_{iy}}{a_{ii}}$$

$$c_{iy} = b_{iy} \cdot a_{iy}$$

i = 1,2,...,p; where p is the number of independent variables.

Then, find the maximum c_{iy}, which is the largest sum of squares explained by the ith variable in the first step of the regression. This value is printed out under the heading of Sums of squares.

Let: $b_{1y} = b_{iy}$; $c_{1y} = c_{iy}$; $a_{1j} = a_{ij}$

where i is the variable entering in the regression.

(3) Obtain $$P = \frac{c_{1y}}{\sum (Y_i - \bar{Y})^2}$$

P is compared with the % value specified in Col. 22-26 of the Problem Card. If P is greater than the % value specified, the program goes to the section immediately following; otherwise, it jumps to (6) below. P is printed out under the heading of Proportion of Variance.

(4) Obtain f value for variable entered.

$$f = \frac{c_{1y}}{\frac{\sum (Y_i - \bar{Y})^2 - c_{1y}}{(n - 1) - 1}}$$

In general

$$f = \frac{\text{sum of squares}}{\frac{\text{residual}}{\text{D.F. residual}}}$$

(5) Compute:

$$b_{1j} = \frac{a_{1j}}{a_{11}} \qquad j = 1,2,\ldots,p$$

where p is the number of independent variables

$$a_{ij \cdot 1} = a_{ij} - a_{i1} \cdot b_{1j} \qquad i = 2,3,\ldots, k \quad j = 2,3,\ldots, k$$

where k is the number of variables including the dependent variable.

The sections (2) - (5) above are repeated to select and enter the remaining variables in the regression in the order of their contribution to the sum of squares of the dependent variable.

(6) Compute cumulative regression:

(a) Sums of squares:

$$Z_k = \sum_{j=1}^{k} c_{jy} \qquad k = 1,2,\ldots, q$$

where q is the number of the variables entered in the regression.

(b) Proportions of variances:

$$R_k^2 = \frac{Z_k}{\sum (Y_i - \bar{Y})^2}$$

(c) Multiple correlation coefficients:

$$R_k = \sqrt{R_k^2}$$

(d) Standard error of estimates:

$$E_{Y \cdot 1 \cdots k} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n - k - 1}} = \sqrt{\frac{\sum (Y_i - \bar{Y})^2 - Z_k}{n - k - 1}}$$

(e) F values:

$$F_k = \frac{\frac{Z_k}{k}}{\frac{\sum (Y_i - \hat{Y}_i)^2}{n - k - 1}} = \frac{\frac{Z_k}{k}}{\frac{\sum (Y_i - \bar{Y})^2 - Z_k}{n - k - 1}}$$

Step 3. Simple correlation coefficients of those variables entered in the regression are rearranged in the working storage area to perform the following computations:

(1) Invert the correlation coefficient matrix and obtain c'_{ij}. ($i = 1,2,\ldots,p$; $j = 1,2,\ldots,p$; where p is the number of independent variables entered.) The inversion is performed by using the subroutine INVERT, programmed at Rocketdyne, North American Aviation.

(2) c_{ij} values:

$$c_{ij} = \frac{c'_{ij}\, r_{ij}}{D_{ij}}$$

(3) Joint regression coefficients:

$$b_j = c_{1j} \Sigma\, (X_1 - \bar{X}_1)(Y - \bar{Y}) + c_{2j} \Sigma\, (X_2 - \bar{X}_2)(Y - \bar{Y}) + \ldots + c_{pj} \Sigma\, (X_p - \bar{X}_p)(Y - \bar{Y})$$

(4) Intercept (A value):

$$A = \bar{Y} - \Sigma\, b_j \bar{X}_j$$

(5) Standard deviations of the regression coefficients:

$$s_{b_j} = \sqrt{s^2_{Y.12\ldots p} \cdot c_{jj}}$$

$j = 1,2,\ldots,p$

where p is the number of independent variables entered.

(6) t values: $t_j = \dfrac{b_j}{s_{b_j}}$

(7) Partial correlation coefficients:

$$r'_{jY.\,(jY)'} = \frac{-a_{Yj}}{\sqrt{a_{YY} \cdot a_{jj}}}$$

where a_{jj} is the inverse of a simple correlation coefficient r_{jj}.

(8) Compare check on final coefficient:

The last b_{iy} computed in the section (2) of Step 2 is also the regression coefficient of the last independent variable. This coefficient is printed out in order to check the accuracy of the computing procedure explained above.

Step 4. Analysis of extreme residuals

The statistical procedure for detection of outliers explained by Dixon and Massey is used in the program. This procedure assumes that data are normally distributed and that they come from the same population. Residuals are not classified in this group. Therefore, readers are reminded that this statistical procedure is only approximate.

Step 5. The program computes the Durbin - Watson d-statistic for testing the presence of autocorrelated disturbances.

APPENDIX 2

Transgeneration codes

00 : No transformation

01 : $x' = \sqrt{x}$

02 : $x' = \sqrt{x} + \sqrt{x+1}$

03 : $x' = \ln x$

04 : $x' = e^{x}$

05 : $x' = \sin^{-1} \sqrt{x}$

06 : $x' = \sin^{-1} \sqrt{\frac{x}{n+1}} + \sin^{-1} \sqrt{\frac{x}{n+1}}$

07 : $x' = 1 / x$

08 : $x' = x + c$

09 : $x' = x \cdot c$

10 : $x' = x^{c}$

11 : $x' = x_A + x_B$

12 : $x' = x_A - x_B$

13 : $x' = x_A \cdot x_B$

14 : $x' = x_A / x_B$

Condition codes

01 : transgeneration if $x_M = x_N$

02 : " " $x_M \neq x_N$

03 : " " $x_M > x_N$

04 : " " $x_M \geq x_N$

05 : " " $x_M < x_N$

06 : " " $x_M \leq x_N$

07 : " " $x_M = K$

08 : " " $x_M \neq K$

09 : " " $x_M > K$

10 : " " $x_M \geq K$

11 : " " $x_M < K$

12 : " " $x_M \leq K$