NAG Toolbox

nag_correg_coeffs_zero_subset_miss_pair (g02bm)

1 Purpose

nag_correg_coeffs_zero_subset_miss_pair (g02bm) computes means and standard deviations, sums of squares and cross-products about zero, and correlation-like coefficients for selected variables omitting cases with missing values from only those calculations involving the variables for which the values are missing.

2 Syntax

```
[xbar, std, sspz, rz, ncases, cnt, ifail] =
nag_correg_coeffs_zero_subset_miss_pair(x, miss, xmiss, kvar, 'n', n, 'm', m,
'nvars', nvars)
[xbar, std, sspz, rz, ncases, cnt, ifail] = g02bm(x, miss, xmiss, kvar, 'n', n,
'm', m, 'nvars', nvars)
```

Note: the interface to this routine has changed since earlier releases of the toolbox:

At Mark 22: **n** was made optional.

3 Description

The input data consists of n observations for each of m variables, given as an array

$$\left[x_{ij}\right], \quad i = 1,2,\ldots,n(n \geq 2), j = 1,2,\ldots,m \quad (m \geq 2),$$

where x_{ij} is the ith observation on the jth variable, together with the subset of these variables, $v_1, v_2, \ldots, v_p$, for which information is required.

In addition, each of the m variables may optionally have associated with it a value which is to be considered as representing a missing observation for that variable; the missing value for the jth variable is denoted by xm_j. Missing values need not be specified for all variables.

Let $w_{ij} = 0$, if the ith observation for the jth variable is a missing value, i.e., if a missing value, xm_j, has been declared for the jth variable, and $x_{ij} = xm_j$ (see also Section 7); and $w_{ij} = 1$ otherwise, for $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,m$.

The quantities calculated are:

(a) Means:

$$\bar{x}_j = \frac{\sum_{i=1}^{n} w_{ij} x_{ij}}{\sum_{i=1}^{n} w_{ij}}, \quad j = v_1, v_2, \ldots, v_p.$$

(b) Standard deviations:

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} w_{ij}\left(x_{ij} - \bar{x}_j\right)^2}{\sum_{i=1}^{n} w_{ij} - 1}}, \quad j = v_1, v_2, \ldots, v_p.$$

(c) Sums of squares and cross-products about zero:

$$\tilde{S}_{jk} = \sum_{i=1}^{n} w_{ij} w_{ik} x_{ij} x_{ik}, \quad j,k = v_1, v_2, \ldots, v_p.$$

(d) Correlation-like coefficients:

$$\tilde{R}_{jk} = \frac{\tilde{S}_{jk}}{\sqrt{\tilde{S}_{jj(k)} \tilde{S}_{kk(j)}}}, \quad j,k = v_{1,} v_2, \ldots, v_p,$$

where $\tilde{S}_{jj(k)} = \sum_{i=1}^{n} w_{ij} w_{ik} x_{ij}^2$ and $\tilde{S}_{kk(j)} = \sum_{i=1}^{n} w_{ik} w_{ij} x_{ik}^2$

(i.e., the sums of squares about zero are based on the same set of observations as are used in the calculation of the numerator).

If $\tilde{S}_{jj(k)}$ or $\tilde{S}_{kk(j)}$ is zero, $\tilde{R}_{jk}$ is set to zero.

(e) The number of cases used in the calculation of each of the correlation-like coefficients:

$$c_{jk} = \sum_{i=1}^{n} w_{ij} w_{ik}, \quad j,k = v_1, v_2, \ldots, v_p.$$

(The diagonal terms, c_{jj}, for $j = 1, 2, \ldots, n$, also give the number of cases used in the calculation of the means $\bar{x}_j$ and the standard deviations s_j.)

4 References

None.

5 Parameters

5.1 Compulsory Input Parameters

1: **x**$(ldx, \mathbf{m})$ – REAL (KIND=nag_wp) array

ldx, the first dimension of the array, must satisfy the constraint $ldx \geq \mathbf{n}$.

x(i,j) must be set to x_{ij}, the value of the ith observation on the jth variable, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$.

2: **miss**$(\mathbf{m})$ – INTEGER array

miss(j) must be set equal to 1 if a missing value, xm_j, is to be specified for the jth variable in the array **x**, or set equal to 0 otherwise. Values of **miss** must be given for all m variables in the array **x**.

3: **xmiss**$(\mathbf{m})$ – REAL (KIND=nag_wp) array

xmiss(j) must be set to the missing value, xm_j, to be associated with the jth variable in the array **x**, for those variables for which missing values are specified by means of the array **miss** (see Section 7).

4: **kvar**$(\mathbf{nvars})$ – INTEGER array

kvar(j) must be set to the column number in **x** of the jth variable for which information is required, for $j = 1, 2, \ldots, p$.

Constraint: $1 \leq$ **kvar**$(j) \leq \mathbf{m}$, for $j = 1, 2, \ldots, p$.

5.2 Optional Input Parameters

1: **n** – INTEGER

Default: the first dimension of the array **x**.

n, the number of observations or cases.

Constraint: $\mathbf{n} \geq 2$.

2: **m** – INTEGER

Default: the dimension of the arrays **miss**, **xmiss** and the second dimension of the array **x**. (An error is raised if these dimensions are not equal.)

m, the number of variables.

Constraint: $\mathbf{m} \geq 2$.

3: **nvars** – INTEGER

Default: the dimension of the array **kvar**.

p, the number of variables for which information is required.

Constraint: $2 \leq \mathbf{nvars} \leq \mathbf{m}$.

5.3 Output Parameters

1: **xbar**(**nvars**) – REAL (KIND=nag_wp) array

The mean value, $\bar{x}_j$, of the variable specified in **kvar**(j), for $j = 1, 2, \ldots, p$.

2: **std**(**nvars**) – REAL (KIND=nag_wp) array

The standard deviation, s_j, of the variable specified in **kvar**(j), for $j = 1, 2, \ldots, p$.

3: **sspz**($ldsspz$, **nvars**) – REAL (KIND=nag_wp) array

sspz(j, k) is the cross-product about zero, $\tilde{S}_{jk}$, for the variables specified in **kvar**(j) and **kvar**(k), for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, p$.

4: **rz**($ldrz$, **nvars**) – REAL (KIND=nag_wp) array

rz(j, k) is the correlation-like coefficient, $\tilde{R}_{jk}$, between the variables specified in **kvar**(j) and **kvar**(k), for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, p$.

5: **ncases** – INTEGER

The minimum number of cases used in the calculation of any of the sums of squares and cross-products and correlation-like coefficients (when cases involving missing values have been eliminated).

6: **cnt**($ldcnt$, **nvars**) – REAL (KIND=nag_wp) array

cnt(j, k) is the number of cases, c_{jk}, actually used in the calculation of the sum of cross-product and correlation-like coefficient for the variables specified in **kvar**(j) and **kvar**(k), for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, p$.

7: **ifail** – INTEGER

ifail $= 0$ unless the function detects an error (see Section 5).

6 Error Indicators and Warnings

Note: nag_correg_coeffs_zero_subset_miss_pair (g02bm) may return useful information for one or more of the following detected errors or warnings.

Errors or warnings detected by the function:

ifail $= 1$

On entry, **n** < 2.

ifail $= 2$

On entry, **nvars** < 2,
or **nvars** $>$ **m**.

ifail $= 3$

On entry, $ldx <$ **n**,
or $ldsspz <$ **nvars**,
or $ldrz <$ **nvars**,
or $ldcnt <$ **nvars**.

ifail $= 4$

On entry, **kvar**$(j) < 1$,
or **kvar**$(j) >$ **m** for some $j = 1, 2, \ldots,$ **nvars**.

ifail $= 5$ (*warning*)

After observations with missing values were omitted, fewer than two cases remained for at least one pair of variables. (The pairs of variables involved can be determined by examination of the contents of the array **cnt**.) All means, standard deviations, sums of squares and cross-products, and correlation-like coefficients based on two or more cases are returned by the function even if **ifail** $= 5$.

ifail $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

ifail $= -399$

Your licence key may have expired or may not have been installed correctly.

ifail $= -999$

Dynamic memory allocation failed.

7 Accuracy

nag_correg_coeffs_zero_subset_miss_pair (g02bm) does not use ***additional precision*** arithmetic for the accumulation of scalar products, so there may be a loss of significant figures for large n.

You are warned of the need to exercise extreme care in your selection of missing values. nag_correg_coeffs_zero_subset_miss_pair (g02bm) treats all values in the inclusive range $\left(1 \pm 0.1^{(\text{x02be}-2)}\right) \times xm_j$, where xm_j is the missing value for variable j specified in **xmiss**.

You must therefore ensure that the missing value chosen for each variable is sufficiently different from all valid values for that variable so that none of the valid values fall within the range indicated above.

8 Further Comments

The time taken by nag_correg_coeffs_zero_subset_miss_pair (g02bm) depends on n and p, and the occurrence of missing values.

The function uses a two-pass algorithm.

9 Example

This example reads in a set of data consisting of five observations on each of four variables. Missing values of −1.0, 0.0 and 0.0 are declared for the first, second and fourth variables respectively; no missing value is specified for the third variable. The means, standard deviations, sums of squares and cross-products about zero, and correlation-like coefficients for the fourth, first and second variables are then calculated and printed, omitting cases with missing values from only those calculations involving the variables for which the values are missing. The program therefore eliminates cases 4 and 5 in calculating the correlation between the fourth and first variable, and cases 3 and 4 for the fourth and second variables, etc.

9.1 Program Text

```
     function g02bm_example

fprintf('g02bm example results\n\n');

x = [ 3,  3,  1,  2;
      6,  4, -1,  4;
      9,  0,  5,  9;
     12,  2,  0,  0;
     -1,  5,  4, 12];
[n,m] = size(x);
fprintf('Number of variables (columns) = %d\n', m);
fprintf('Number of cases     (rows)    = %d\n\n', n);
disp('Data matrix is:-');
disp(x);

miss  = [nag_int(1); 1; 0; 1];
xmiss = [       -1;  0; 0; 0];
kvar  = [nag_int(4); 1; 2];
nvar  = size(kvar,1);

[xbar, std, sspz, rz, ncases, count, ifail] = ...
  g02bm( ...
         x, miss, xmiss, kvar);

fprintf('Variable   Mean     St. dev.\n');
fprintf('%5d%11.4f%11.4f\n',[double(kvar) xbar(1:nvar) std(1:nvar)]');
fprintf('\nSums of squares and cross-products about zero\n');
disp(sspz)
fprintf('Correlation-like coefficients\n');
disp(rz);
fprintf('Number of cases used for any pair of variables = %d\n\n', ncases);
fprintf('Numbers used for each pair are:\n   ');
fprintf('%10d',kvar);
for j=1:nvar
  fprintf('\n%3d:',kvar(j));
  fprintf('%10.1f',count(j,:));
end
fprintf('\n')
```

9.2 Program Results

```
     g02bm example results

Number of variables (columns) = 4
Number of cases     (rows)    = 5

Data matrix is:-
     3     3     1     2
     6     4    -1     4
     9     0     5     9
    12     2     0     0
    -1     5     4    12
```

```
Variable   Mean     St. dev.
    4     6.7500     4.5735
    1     7.5000     3.8730
    2     3.5000     1.2910

Sums of squares and cross-products about zero
   245   111    82
   111   270    57
    82    57    54

Correlation-like coefficients
    1.0000    0.9840    0.9055
    0.9840    1.0000    0.7699
    0.9055    0.7699    1.0000

Number of cases used for any pair of variables = 3

Numbers used for each pair are:
            4         1         2
  4:       4.0       3.0       3.0
  1:       3.0       4.0       3.0
  2:       3.0       3.0       4.0
```