NAG Toolbox

nag_correg_coeffs_pearson_miss_case (g02bb)

1 Purpose

nag_correg_coeffs_pearson_miss_case (g02bb) computes means and standard deviations of variables, sums of squares and cross-products of deviations from means, and Pearson product-moment correlation coefficients for a set of data omitting completely any cases with a missing observation for any variable.

2 Syntax

```
[xbar, std, ssp, r, ncases, ifail] = nag_correg_coeffs_pearson_miss_case(x, miss, xmiss, 'n', n, 'm', m)
[xbar, std, ssp, r, ncases, ifail] = g02bb(x, miss, xmiss, 'n', n, 'm', m)
```

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

At Mark 22: **n** was made optional; **miss** and **xmiss** are no longer output parameters.

3 Description

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

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

where x_{ij} is the ith observation on the jth variable. 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 $w_i = 0$ if observation i contains a missing value for any of those variables for which missing values have been declared, i.e., if $x_{ij} = xm_j$ for any j for which an xm_j has been assigned (see also Section 7); and $w_i = 1$ otherwise, for $i = 1,2,\ldots,n$.

The quantities calculated are:

(a) Means:

$$\bar{x}_j = \frac{\sum_{i=1}^{n} w_i x_{ij}}{\sum_{i=1}^{n} w_i}, \quad j = 1,2,\ldots,m.$$

(b) Standard deviations:

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} w_i \big(x_{ij} - \bar{x}_j\big)^2}{\sum_{i=1}^{n} w_i - 1}}, \quad j = 1,2,\ldots,m.$$

(c) Sums of squares and cross-products of deviations from means:

$$S_{jk} = \sum_{i=1}^{n} w_i \big(x_{ij} - \bar{x}_j\big)\big(x_{ik} - \bar{x}_k\big), \quad j,k = 1,2,\ldots,m.$$

(d) Pearson product-moment correlation coefficients:

$$R_{jk} = \frac{S_{jk}}{\sqrt{S_{jj}S_{kk}}}, \quad j,k = 1,2,\ldots,m.$$

If S_{jj} or S_{kk} is zero, R_{jk} is set to zero.

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).

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

5.3 Output Parameters

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

The mean value, $\bar{x}_j$, of the jth variable, for $j = 1,2,\ldots,m$.

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

The standard deviation, s_j, of the jth variable, for $j = 1,2,\ldots,m$.

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

ssp(j,k) is the cross-product of deviations S_{jk}, for $j = 1,2,\ldots,m$ and $k = 1,2,\ldots,m$.

4: **r**$(ldr, \mathbf{m})$ – REAL (KIND=nag_wp) array

r(j,k) is the product-moment correlation coefficient R_{jk} between the jth and kth variables, for $j = 1,2,\ldots,m$ and $k = 1,2,\ldots,m$.

5: **ncases** – INTEGER

The number of cases actually used in the calculations (when cases involving missing values have been eliminated).

6: **ifail** – INTEGER

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

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail $= 1$

On entry, $\mathbf{n} < 2$.

ifail $= 2$

On entry, $\mathbf{m} < 2$.

ifail $= 3$

On entry, $ldx < \mathbf{n}$,
or $ldssp < \mathbf{m}$,
or $ldr < \mathbf{m}$.

ifail $= 4$

After observations with missing values were omitted, no cases remained.

ifail $= 5$

After observations with missing values were omitted, only one case remained.

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_pearson_miss_case (g02bb) 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_pearson_miss_case (g02bb) 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_pearson_miss_case (g02bb) depends on n and m, 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 three variables. Missing values of 0.0 are declared for the first and third variables; no missing value is specified for the second variable. The means, standard deviations, sums of squares and cross-products of deviations from means, and Pearson product-moment correlation coefficients for all three variables are then calculated and printed, omitting completely all cases containing missing values; cases 3 and 4 are therefore eliminated, leaving only three cases in the calculations.

9.1 Program Text

```
     function g02bb_example

fprintf('g02bb example results\n\n');

x = [ 2,  3, 3;
      4,  6, 4;
      9,  9, 0;
      0, 12, 2;
     12, -1, 5];
[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); 0; 1];
xmiss = [0;          0; 0];
[xbar, std, ssp, r, ncases, ifail] = ...
  g02bb(x, miss, xmiss);

fprintf('Variable   Mean     St. dev.\n');
fprintf('%5d%11.4f%11.4f\n',[[1:m]' xbar std]');
fprintf('\nSums of squares and cross-products of deviations\n');
disp(ssp)
fprintf('Correlation coefficients\n');
disp(r);
fprintf('Number of cases actually used  = %d\n', ncases);
```

9.2 Program Results

```
     g02bb example results

Number of variables (columns) = 3
Number of cases     (rows)    = 5

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

Variable   Mean     St. dev.
    1     6.0000     5.2915
    2     2.6667     3.5119
```

```
    3     4.0000     1.0000

Sums of squares and cross-products of deviations
   56.0000  -30.0000   10.0000
  -30.0000   24.6667   -4.0000
   10.0000   -4.0000    2.0000

Correlation coefficients
    1.0000   -0.8072    0.9449
   -0.8072    1.0000   -0.5695
    0.9449   -0.5695    1.0000

Number of cases actually used  = 3
```