NAG Library Routine Document

G02BEF

Note: before using this routine, please read the Users' Note for your implementation to check the interpretation of *bold italicised* terms and other implementation-dependent details.

1 Purpose

G02BEF computes means and standard deviations of variables, sums of squares and cross-products about zero, and correlation-like coefficients for a set of data omitting completely any cases with a missing observation for any variable.

2 Specification

```
SUBROUTINE G02BEF (N, M, X, LDX, MISS, XMISS, XBAR, STD, SSPZ, LDSSPZ, RZ,      &
                   LDRZ, NCASES, IFAIL)

INTEGER           N, M, LDX, MISS(M), LDSSPZ, LDRZ, NCASES, IFAIL
REAL (KIND=nag_wp) X(LDX,M), XMISS(M), XBAR(M), STD(M), SSPZ(LDSSPZ,M),          &
                   RZ(LDRZ,M)
```

3 Description

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

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

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:

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

(b) Standard deviations:

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

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

$$\tilde{S}_{jk} = \sum_{i=1}^{n} w_i x_{ij} x_{ik}, \qquad j, k = 1, 2, \ldots, m.$$

(d) Correlation-like coefficients:

$$\tilde{R}_{jk} = \frac{\tilde{S}_{jk}}{\sqrt{\tilde{S}_{jj}\tilde{S}_{kk}}}, \qquad j, k = 1, 2, \ldots, m.$$

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

4 References

None.

5 Parameters

1:     N – INTEGER   *Input*

On entry: n, the number of observations or cases.

Constraint: N $\geq$ 2.

2:     M – INTEGER   *Input*

On entry: m, the number of variables.

Constraint: M $\geq$ 2.

3:     X(LDX,M) – REAL (KIND=nag_wp) array   *Input*

On entry: 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 $j = 1, 2, \ldots, m$.

4:     LDX – INTEGER   *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G02BEF is called.

Constraint: LDX $\geq$ N.

5:     MISS(M) – INTEGER array   *Input/Output*

On entry: 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.

On exit: the array MISS is overwritten by the routine, and the information it contained on entry is lost.

6:     XMISS(M) – REAL (KIND=nag_wp) array   *Input/Output*

On entry: 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).

On exit: the array XMISS is overwritten by the routine, and the information it contained on entry is lost.

7:     XBAR(M) – REAL (KIND=nag_wp) array   *Output*

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

8:     STD(M) – REAL (KIND=nag_wp) array   *Output*

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

9:     SSPZ(LDSSPZ,M) – REAL (KIND=nag_wp) array   *Output*

   On exit: SSPZ(j, k) is the cross-product about zero, $\tilde{S}_{jk}$, for $j = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, m$.

10:    LDSSPZ – INTEGER   *Input*

   On entry: the first dimension of the array SSPZ as declared in the (sub)program from which G02BEF is called.

   Constraint: LDSSPZ $\geq$ M.

11:    RZ(LDRZ,M) – REAL (KIND=nag_wp) array   *Output*

   On exit: RZ(j, k) is the correlation-like coefficient, $\tilde{R}_{jk}$, between the jth and kth variables, for $j = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, m$.

12:    LDRZ – INTEGER   *Input*

   On entry: the first dimension of the array RZ as declared in the (sub)program from which G02BEF is called.

   Constraint: LDRZ $\geq$ M.

13:    NCASES – INTEGER   *Output*

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

14:    IFAIL – INTEGER   *Input/Output*

   On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction for details.

   For environments where it might be inappropriate to halt program execution when an error is detected, the value -1 or 1 is recommended. If the output of error messages is undesirable, then the value 1 is recommended. Otherwise, if you are not familiar with this parameter, the recommended value is 0. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

   On exit: IFAIL $= 0$ unless the routine detects an error or a warning has been flagged (see Section 6).

6   Error Indicators and Warnings

If on entry IFAIL $= 0$ or -1, explanatory error messages are output on the current error message unit (as defined by X04AAF).

Errors or warnings detected by the routine:

IFAIL $= 1$

   On entry, N < 2.

IFAIL $= 2$

   On entry, M < 2.

$\text{IFAIL} = 3$

> On entry, $\text{LDX} < \text{N}$,
> or $\text{LDSSPZ} < \text{M}$,
> or $\text{LDRZ} < \text{M}$.

$\text{IFAIL} = 4$

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

$\text{IFAIL} = 5$

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

7 Accuracy

G02BEF 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. G02BEF treats all values in the inclusive range $\left(1 \pm 0.1^{(\text{X02BEF}-2)}\right) \times xm_j$, 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 G02BEF depends on n and m, and the occurrence of missing values.

The routine 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 about zero, and correlation-like 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

```
    Program g02befe

!   G02BEF Example Program Text

!   Mark 24 Release. NAG Copyright 2012.

!   .. Use Statements ..
    Use nag_library, Only: g02bef, nag_wp
!   .. Implicit None Statement ..
    Implicit None
!   .. Parameters ..
    Integer, Parameter              :: nin = 5, nout = 6
!   .. Local Scalars ..
    Integer                         :: i, ifail, ldrz, ldsspz, ldx, m, n,   &
                                       ncases
!   .. Local Arrays ..
    Real (Kind=nag_wp), Allocatable :: rz(:,:), sspz(:,:), std(:), x(:,:),  &
                                       xbar(:), xmiss(:)
    Integer, Allocatable            :: miss(:)
!   .. Executable Statements ..
    Write (nout,*) 'G02BEF Example Program Results'
```

```
      Write (nout,*)

!     Skip heading in data file
      Read (nin,*)

!     Read in the problem size
      Read (nin,*) n, m

      ldrz = m
      ldsspz = m
      ldx = n
      Allocate (rz(ldrz,m),sspz(ldsspz,m),std(m),x(ldx,m),xbar(m),xmiss(m), &
        miss(m))

!     Read in data
      Read (nin,*)(x(i,1:m),i=1,n)

!     Read in missing value flags
      Read (nin,*) miss(1:m)
      Read (nin,*) xmiss(1:m)

!     Display data
      Write (nout,99999) 'Number of variables (columns) =', m
      Write (nout,99999) 'Number of cases     (rows)    =', n
      Write (nout,*)
      Write (nout,*) 'Data matrix is:-'
      Write (nout,*)
      Write (nout,99998)(i,i=1,m)
      Write (nout,99997)(i,x(i,1:m),i=1,n)
      Write (nout,*)

!     Calculate summary statistics
      ifail = 0
      Call g02bef(n,m,x,ldx,miss,xmiss,xbar,std,sspz,ldsspz,rz,ldrz,ncases, &
        ifail)

!     Display results
      Write (nout,*) 'Variable   Mean    St. dev.'
      Write (nout,99996)(i,xbar(i),std(i),i=1,m)
      Write (nout,*)
      Write (nout,*) 'Sums of squares and cross-products about' // ' zero'
      Write (nout,99998)(i,i=1,m)
      Write (nout,99997)(i,sspz(i,1:m),i=1,m)
      Write (nout,*)
      Write (nout,*) 'Correlation-like coefficients'
      Write (nout,99998)(i,i=1,m)
      Write (nout,99997)(i,rz(i,1:m),i=1,m)
      Write (nout,*)
      Write (nout,99999) 'Number of cases actually used: ', ncases

99999 Format (1X,A,I5)
99998 Format (1X,6I12)
99997 Format (1X,I3,3F12.4)
99996 Format (1X,I5,2F11.4)
    End Program g02befe
```

9.2   Program Data

```
G02BEF Example Program Data
5  3                      :: N, M
 2.0    3.0    3.0
 4.0    6.0    4.0
 9.0    9.0    0.0
 0.0   12.0    2.0
12.0   -1.0    5.0       :: End of X
  1      0      1        :: MISS
 0.0    0.0    0.0       :: XMISS
```

9.3 Program Results

```
G02BEF Example Program Results

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

Data matrix is:-

           1             2             3
  1     2.0000        3.0000        3.0000
  2     4.0000        6.0000        4.0000
  3     9.0000        9.0000        0.0000
  4     0.0000       12.0000        2.0000
  5    12.0000       -1.0000        5.0000

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 about zero
           1             2             3
  1    164.0000       18.0000       82.0000
  2     18.0000       46.0000       28.0000
  3     82.0000       28.0000       50.0000

Correlation-like coefficients
           1             2             3
  1     1.0000        0.2072        0.9055
  2     0.2072        1.0000        0.5838
  3     0.9055        0.5838        1.0000

Number of cases actually used:     3
```