NAG Library Routine Document

G02BJF

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

G02BJF computes means and standard deviations, sums of squares and cross-products of deviations from means, and Pearson product-moment correlation coefficients for selected variables omitting cases with missing values from only those calculations involving the variables for which the values are missing.

2 Specification

```
SUBROUTINE G02BJF (N, M, X, LDX, MISS, XMISS, NVARS, KVAR, XBAR, STD, SSP,      &
                   LDSSP, R, LDR, NCASES, CNT, LDCNT, IFAIL)
INTEGER            N, M, LDX, MISS(M), NVARS, KVAR(NVARS), LDSSP, LDR,          &
                   NCASES, LDCNT, IFAIL
REAL (KIND=nag_wp) X(LDX,M), XMISS(M), XBAR(NVARS), STD(NVARS),                 &
                   SSP(LDSSP,NVARS), R(LDR,NVARS), CNT(LDCNT,NVARS)
```

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, 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:

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

(b) Standard deviations:

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

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

$$S_{jk} = \sum_{i=1}^{n} w_{ij} w_{ik} \big(x_{ij} - \bar{x}_{j(k)}\big)\big(x_{ik} - \bar{x}_{k(j)}\big), \qquad j,k = v_1, v_2, \ldots, v_p,$$

where

$$\bar{x}_{j(k)} = \frac{\sum_{i=1}^{n} w_{ij} w_{ik} x_{ij}}{\sum_{i=1}^{n} w_{ij} w_{ik}} \qquad \text{and} \qquad \bar{x}_{k(j)} = \frac{\sum_{i=1}^{n} w_{ik} w_{ij} x_{ik}}{\sum_{i=1}^{n} w_{ik} w_{ij}},$$

(i.e., the means used in the calculation of the sum of squares and cross-products of deviations are based on the same set of observations as are the cross-products).

(d) Pearson product-moment correlation coefficients:

$$R_{jk} = \frac{S_{jk}}{\sqrt{S_{jj(k)} S_{kk(j)}}}, \qquad j,k = v_1, v_2, \ldots, v_p,$$

where

$$S_{jj(k)} = \sum_{i=1}^{n} w_{ij} w_{ik} \big(x_{ij} - \bar{x}_{j(k)}\big)^2 \qquad \text{and} \qquad S_{kk(j)} = \sum_{i=1}^{n} w_{ik} w_{ij} \big(x_{ik} - \bar{x}_{k(j)}\big)^2,$$

(i.e., the sums of squares of deviations used in the denominator are based on the same set of observations as are used in the calculation of the numerator).If $S_{jj(k)}$ or $S_{kk(j)}$ is zero, R_{jk} is set to zero.

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

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

(The diagonal terms, c_{jj}, for $j = v_1, v_2, \ldots, v_p$, 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

1: N – INTEGER *Input*

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

Constraint: N ≥ 2.

2: M – INTEGER *Input*

On entry: m, the number of variables.

Constraint: M ≥ 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 G02BJF is called.

Constraint: LDX $\geq$ N.

5: MISS(M) – INTEGER array *Input*

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.

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

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).

7: NVARS – INTEGER *Input*

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

Constraint: $2 \leq$ NVARS $\leq$ M.

8: KVAR(NVARS) – INTEGER array *Input*

On entry: 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) $\leq$ M, for $j = 1, 2, \ldots, p$.

9: XBAR(NVARS) – REAL (KIND=nag_wp) array *Output*

On exit: the mean value, $\bar{x}_j$, of the variable specified in KVAR(j), for $j = 1, 2, \ldots, p$.

10: STD(NVARS) – REAL (KIND=nag_wp) array *Output*

On exit: the standard deviation, s_j, of the variable specified in KVAR(j), for $j = 1, 2, \ldots, p$.

11: SSP(LDSSP,NVARS) – REAL (KIND=nag_wp) array *Output*

On exit: SSP(j, k) is the cross-product of deviations, S_{jk}, for the variables specified in KVAR(j) and KVAR(k), for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, p$.

12: LDSSP – INTEGER *Input*

On entry: the first dimension of the array SSP as declared in the (sub)program from which G02BJF is called.

Constraint: LDSSP $\geq$ NVARS.

13: R(LDR,NVARS) – REAL (KIND=nag_wp) array *Output*

On exit: R(j, k) is the product-moment correlation coefficient, R_{jk}, between the variables specified in KVAR(j) and KVAR(k), for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, p$.

14: LDR – INTEGER *Input*

On entry: the first dimension of the array R as declared in the (sub)program from which G02BJF is called.

Constraint: LDR $\geq$ NVARS.

15: NCASES – INTEGER *Output*

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

16: CNT(LDCNT,NVARS) – REAL (KIND=nag_wp) array *Output*

On exit: CNT(j,k) is the number of cases, c_{jk}, actually used in the calculation of S_{jk}, and R_{jk}, the sum of cross-products and correlation coefficient for the variables specified in KVAR(j) and KVAR(k), for $j = 1,2,\ldots,p$ and $k = 1,2,\ldots,p$.

17: LDCNT – INTEGER *Input*

On entry: the first dimension of the array CNT as declared in the (sub)program from which G02BJF is called.

Constraint: LDCNT $\geq$ NVARS.

18: 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, because for this routine the values of the output parameters may be useful even if IFAIL $\neq 0$ on exit, the recommended value is -1. **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).

Note: G02BJF may return useful information for one or more of the following detected errors or warnings.

Errors or warnings detected by the routine:

IFAIL $= 1$

On entry, N < 2.

IFAIL $= 2$

On entry, NVARS < 2,
or NVARS $>$ M.

IFAIL $= 3$

On entry, LDX $<$ N,
or LDSSP $<$ NVARS,
or LDR $<$ NVARS,
or LDCNT $<$ NVARS.

IFAIL $= 4$

On entry, KVAR$(j) < 1$,
or KVAR$(j) >$ M for some $j = 1,2,\ldots,$NVARS.

IFAIL $= 5$

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 coefficients based on two or more cases are returned by the routine even if IFAIL $= 5$.

7 Accuracy

G02BJF 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. G02BJF 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 G02BJF depends on n and p, 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 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 of deviations from means, and Pearson product-moment correlation 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 variables, and cases 3 and 4 for the fourth and second variables etc.

9.1 Program Text

```
    Program g02bjfe

!     G02BJF Example Program Text

!     Mark 24 Release. NAG Copyright 2012.

!     .. Use Statements ..
      Use nag_library, Only: g02bjf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter                :: nin = 5, nout = 6
!     .. Local Scalars ..
      Integer                           :: i, ifail, ldcnt, ldr, ldssp, ldx, m, &
                                           n, ncases, nvars
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable   :: cnt(:,:), r(:,:), ssp(:,:), std(:),  &
                                           x(:,:), xbar(:), xmiss(:)
      Integer, Allocatable              :: kvar(:), miss(:)
!     .. Executable Statements ..
      Write (nout,*) 'G02BJF Example Program Results'
      Write (nout,*)

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

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

      ldcnt = nvars
      ldr = nvars
      ldssp = nvars
      ldx = n
      Allocate (cnt(ldcnt,nvars),r(ldr,nvars),ssp(ldssp,nvars),std(nvars), &
        x(ldx,m),xbar(nvars),xmiss(m),kvar(nvars),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)

!     Read in column IDs
      Read (nin,*) kvar(1:nvars)

!     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,99998)(i,i=1,m)
      Write (nout,99997)(i,x(i,1:m),i=1,n)
      Write (nout,*)

!     Calculate summary statistics
      ifail = 0
      Call g02bjf(n,m,x,ldx,miss,xmiss,nvars,kvar,xbar,std,ssp,ldssp,r,ldr, &
        ncases,cnt,ldcnt,ifail)

!     Display results
      Write (nout,*) 'Variable   Mean    St. dev.'
      Write (nout,99995)(kvar(i),xbar(i),std(i),i=1,nvars)
      Write (nout,*)
      Write (nout,*) 'Sums of squares and cross-products of deviations'
      Write (nout,99998) kvar(1:nvars)
      Write (nout,99996)(kvar(i),ssp(i,1:nvars),i=1,nvars)
      Write (nout,*)
      Write (nout,*) 'Correlation coefficients'
      Write (nout,99998) kvar(1:nvars)
      Write (nout,99996)(kvar(i),r(i,1:nvars),i=1,nvars)
      Write (nout,*)
      Write (nout,99999) &
        'Minimum number of cases used for any pair of variables:', ncases
      Write (nout,*)
      Write (nout,*) 'Numbers used for each pair are:'
      Write (nout,99998) kvar(1:nvars)
      Write (nout,99996)(kvar(i),cnt(i,1:nvars),i=1,nvars)

99999 Format (1X,A,I5)
99998 Format (1X,4I12)
99997 Format (1X,I3,4F12.4)
99996 Format (1X,I3,3F12.4)
99995 Format (1X,I5,2F11.4)
    End Program g02bjfe
```

9.2 Program Data

```
G02BJF Example Program Data
5  4  3                    :: N, M, NVARS
 3.0    3.0    1.0    2.0
 6.0    4.0   -1.0    4.0
 9.0    0.0    5.0    9.0
```

```
12.0    2.0    0.0    0.0
-1.0    5.0    4.0   12.0  :: End of X
  1      1      0      1   :: MISS
-1.0    0.0    0.0    0.0  :: XMISS
4  1  2                    :: KVAR
```

9.3 Program Results

```
 G02BJF Example Program Results

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

 Data matrix is:-
            1           2           3           4
   1      3.0000      3.0000      1.0000      2.0000
   2      6.0000      4.0000     -1.0000      4.0000
   3      9.0000      0.0000      5.0000      9.0000
   4     12.0000      2.0000      0.0000      0.0000
   5     -1.0000      5.0000      4.0000     12.0000

 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 of deviations
            4           1           2
   4     62.7500     21.0000     10.0000
   1     21.0000     45.0000     -6.0000
   2     10.0000     -6.0000      5.0000

 Correlation coefficients
            4           1           2
   4      1.0000      0.9707      0.9449
   1      0.9707      1.0000     -0.6547
   2      0.9449     -0.6547      1.0000

 Minimum number of cases used for any pair of variables:    3

 Numbers used for each pair are:
            4           1           2
   4      4.0000      3.0000      3.0000
   1      3.0000      4.0000      3.0000
   2      3.0000      3.0000      4.0000
```