NAG Library Routine Document

G02CHF

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

G02CHF performs a multiple linear regression with no constant on a set of variables whose sums of squares and cross-products about zero and correlation-like coefficients are given.

2 Specification

```
SUBROUTINE G02CHF (N, K1, K, SSPZ, LDSSPZ, RZ, LDRZ, RESULT, COEF,          &
                   LDCOEF, RZNV, LDRZNV, CZ, LDCZ, WKZ, LDWKZ, IFAIL)

INTEGER            N, K1, K, LDSSPZ, LDRZ, LDCOEF, LDRZNV, LDCZ, LDWKZ,     &
                   IFAIL
REAL (KIND=nag_wp) SSPZ(LDSSPZ,K1), RZ(LDRZ,K1), RESULT(13),                 &
                   COEF(LDCOEF,3), RZNV(LDRZNV,K), CZ(LDCZ,K),               &
                   WKZ(LDWKZ,K)
```

3 Description

G02CHF fits a curve of the form

$$y = b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

to the data points

$$\begin{gathered}(x_{11}, x_{21}, \ldots, x_{k1}, y_1)\\(x_{12}, x_{22}, \ldots, x_{k2}, y_2)\\\vdots\\(x_{1n}, x_{2n}, \ldots, x_{kn}, y_n)\end{gathered}$$

such that

$$y_i = b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki} + e_i, \quad i = 1, 2, \ldots, n.$$

The routine calculates the regression coefficients, $b_1, b_2, \ldots, b_k$, (and various other statistical quantities) by minimizing

$$\sum_{i=1}^{n} e_i^2.$$

The actual data values $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ are not provided as input to the routine. Instead, input to the routine consists of:

(i) The number of cases, n, on which the regression is based.

(ii) The total number of variables, dependent and independent, in the regression, $(k+1)$.

(iii) The number of independent variables in the regression, k.

(iv) The $(k+1)$ by $(k+1)$ matrix $\left[\tilde{S}_{ij}\right]$ of sums of squares and cross-products about zero of all the variables in the regression; the terms involving the dependent variable, y, appear in the $(k+1)$th row and column.

(v) The $(k+1)$ by $(k+1)$ matrix $\left[\tilde{R}_{ij}\right]$ of correlation-like coefficients for all the variables in the regression; the correlations involving the dependent variable, y, appear in the $(k+1)$th row and column.

The quantities calculated are:

(a) The inverse of the k by k partition of the matrix of correlation-like coefficients, $\left[\tilde{R}_{ij}\right]$, involving only the independent variables. The inverse is obtained using an accurate method which assumes that this sub-matrix is positive definite (see Section 9).

(b) The modified matrix, $C = \left[c_{ij}\right]$, where

$$c_{ij} = \frac{\tilde{R}_{ij}\tilde{r}^{ij}}{\tilde{S}_{ij}}, \quad i,j = 1,2,\ldots,k,$$

where $\tilde{r}^{ij}$ is the (i,j)th element of the inverse matrix of $\left[\tilde{R}_{ij}\right]$ as described in (a) above. Each element of C is thus the corresponding element of the matrix of correlation-like coefficients multiplied by the corresponding element of the inverse of this matrix, divided by the corresponding element of the matrix of sums of squares and cross-products about zero.

(c) The regression coefficients:

$$b_i = \sum_{j=1}^{k} c_{ij}\tilde{S}_{j(k+1)}, \quad i = 1,2,\ldots,k,$$

where $\tilde{S}_{j(k+1)}$ is the sum of cross-products about zero for the independent variable x_j and the dependent variable y.

(d) The sum of squares attributable to the regression, SSR, the sum of squares of deviations about the regression, SSD, and the total sum of squares, SST:

$SST = \tilde{S}_{(k+1)(k+1)}$, the sum of squares about zero for the dependent variable, y;

$$SSR = \sum_{j=1}^{k} b_j\tilde{S}_{j(k+1)}; \quad SSD = SST - SSR.$$

(e) The degrees of freedom attributable to the regression, DFR, the degrees of freedom of deviations about the regression, DFD, and the total degrees of freedom, DFT:

$$DFR = k; \quad DFD = n - k; \quad DFT = n.$$

(f) The mean square attributable to the regression, MSR, and the mean square of deviations about the regression, MSD:

$$MSR = SSR/DFR; \quad MSD = SSD/DFD.$$

(g) The F value for the analysis of variance:

$$F = MSR/MSD.$$

(h) The standard error estimate:

$$s = \sqrt{MSD}.$$

(i) The coefficient of multiple correlation, R, the coefficient of multiple determination, R^2, and the coefficient of multiple determination corrected for the degrees of freedom, $\bar{R}^2$:

$$R = \sqrt{1 - \frac{SSD}{SST}}; \quad R^2 = 1 - \frac{SSD}{SST}; \quad \bar{R}^2 = 1 - \frac{SSD \times DFT}{SST \times DFD}.$$

(j) The standard error of the regression coefficients:

$$se(b_i) = \sqrt{MSD \times c_{ii}}, \quad i = 1,2,\ldots,k.$$

(k) The t values for the regression coefficients:

$$t(b_i) = \frac{b_i}{se(b_i)}, \quad i = 1,2,\ldots,k.$$

4 References

Draper N R and Smith H (1985) *Applied Regression Analysis* (2nd Edition) Wiley

5 Arguments

1: N – INTEGER *Input*

On entry: n, the number of cases used in calculating the sums of squares and cross-products and correlation-like coefficients.

2: K1 – INTEGER *Input*

On entry: the total number of variables, independent and dependent $(k+1)$, in the regression.

Constraint: $2 \leq \text{K1} \leq \text{N}$.

3: K – INTEGER *Input*

On entry: the number of independent variables k in the regression.

Constraint: $\text{K} = \text{K1} - 1$.

4: SSPZ(LDSSPZ, K1) – REAL (KIND=nag_wp) array *Input*

On entry: SSPZ(i,j) must be set to $\tilde{S}_{ij}$, the sum of cross-products about zero for the ith and jth variables, for $i = 1,2,\ldots,k+1$ and $j = 1,2,\ldots,k+1$; terms involving the dependent variable appear in row $k+1$ and column $k+1$.

5: LDSSPZ – INTEGER *Input*

On entry: the first dimension of the array SSPZ as declared in the (sub)program from which G02CHF is called.

Constraint: $\text{LDSSPZ} \geq \text{K1}$.

6: RZ(LDRZ, K1) – REAL (KIND=nag_wp) array *Input*

On entry: RZ(i,j) must be set to $\tilde{R}_{ij}$, the correlation-like coefficient for the ith and jth variables, for $i = 1,2,\ldots,k+1$ and $j = 1,2,\ldots,k+1$; coefficients involving the dependent variable appear in row $k+1$ and column $k+1$.

7: LDRZ – INTEGER *Input*

On entry: the first dimension of the array RZ as declared in the (sub)program from which G02CHF is called.

Constraint: $\text{LDRZ} \geq \text{K1}$.

8: RESULT(13) – REAL (KIND=nag_wp) array *Output*

On exit: the following information:

RESULT(1) SSR, the sum of squares attributable to the regression;
RESULT(2) DFR, the degrees of freedom attributable to the regression;
RESULT(3) MSR, the mean square attributable to the regression;
RESULT(4) F, the F value for the analysis of variance;
RESULT(5) SSD, the sum of squares of deviations about the regression;
RESULT(6) DFD, the degrees of freedom of deviations about the regression;
RESULT(7) MSD, the mean square of deviations about the regression;
RESULT(8) SST, the total sum of squares;
RESULT(9) DFT, the total degrees of freedom;
RESULT(10) s, the standard error estimate;

RESULT(11) R, the coefficient of multiple correlation;
RESULT(12) R^2, the coefficient of multiple determination;
RESULT(13) $\bar{R}^2$, the coefficient of multiple determination corrected for the degrees of freedom.

9: COEF(LDCOEF, 3) – REAL (KIND=nag_wp) array *Output*

On exit: for $i = 1, 2, \ldots, k$, the following information:

COEF(i, 1)
b_i, the regression coefficient for the ith variable.

COEF(i, 2)
$se(b_i)$, the standard error of the regression coefficient for the ith variable.

COEF(i, 3)
$t(b_i)$, the t value of the regression coefficient for the ith variable.

10: LDCOEF – INTEGER *Input*

On entry: the first dimension of the array COEF as declared in the (sub)program from which G02CHF is called.

Constraint: LDCOEF $\geq$ K.

11: RZNV(LDRZNV, K) – REAL (KIND=nag_wp) array *Output*

On exit: the inverse of the matrix of correlation-like coefficients for the independent variables; that is, the inverse of the matrix consisting of the first k rows and columns of RZ.

12: LDRZNV – INTEGER *Input*

On entry: the first dimension of the array RZNV as declared in the (sub)program from which G02CHF is called.

Constraint: LDRZNV $\geq$ K.

13: CZ(LDCZ, K) – REAL (KIND=nag_wp) array *Output*

On exit: the modified inverse matrix, C, where

$$\mathrm{CZ}(i,j) = \frac{\mathrm{RZ}(i,j) \times \mathrm{RZNV}(i,j)}{\mathrm{SSPZ}(i,j)}, \quad i,j = 1,2,\ldots,k.$$

14: LDCZ – INTEGER *Input*

On entry: the first dimension of the array CZ as declared in the (sub)program from which G02CHF is called.

Constraint: LDCZ $\geq$ K.

15: WKZ(LDWKZ, K) – REAL (KIND=nag_wp) array *Workspace*
16: LDWKZ – INTEGER *Input*

On entry: the first dimension of the array WKZ as declared in the (sub)program from which G02CHF is called.

Constraint: LDWKZ $\geq$ K.

17: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this argument you should refer to Section 3.4 in How to Use the NAG Library and its Documentation 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 argument, 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, K1 < 2.

IFAIL $= 2$

On entry, K1 $\neq$ (K $+ 1$).

IFAIL $= 3$

On entry, N $<$ K1.

IFAIL $= 4$

On entry, LDSSPZ $<$ K1,
or LDRZ $<$ K1,
or LDCOEF $<$ K,
or LDRZNV $<$ K,
or LDCZ $<$ K,
or LDWKZ $<$ K.

IFAIL $= 5$

This indicates that the k by k partition of the matrix held in RZ, which is to be inverted, is not positive definite.

IFAIL $= 6$

This indicates that the refinement following the actual inversion fails, indicating that the k by k partition of the matrix held in RZ, which is to be inverted, is ill-conditioned. The use of G02DAF, which employs a different numerical technique, may avoid the difficulty.

IFAIL $= 7$

An unexpected error has been triggered internally whilst solving a set of linear equations. Please contact NAG.

See Section 3.9 in How to Use the NAG Library and its Documentation for further information.

IFAIL $= -99$

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

See Section 3.9 in How to Use the NAG Library and its Documentation for further information.

IFAIL $= -399$

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

See Section 3.8 in How to Use the NAG Library and its Documentation for further information.

IFAIL $= -999$

Dynamic memory allocation failed.

See Section 3.7 in How to Use the NAG Library and its Documentation for further information.

7 Accuracy

The accuracy of G02CHF is almost entirely dependent on the accuracy of the matrix inversion method used. As G02CHF works with the matrix of correlation coefficients rather than that of the sums of squares and cross-products of deviations from means all terms in the matrix being inverted are of a similar order and therefore the scope for computational error is reduced. An alternative, and potentially more numerically reliable, routine is G02DAF. G02DAF works directly with the data matrix and therefore avoids explicitly performing a matrix inversion. However, G02DAF does not handle missing values, nor does it provide the same output as this routine.

If, in calculating F or any of the $t(b_i)$ (see Section 3), the numbers involved are such that the result would be outside the range of numbers which can be stored by the machine, then the answer is set to the largest quantity which can be stored as a real variable, by means of a call to X02ALF.

8 Parallelism and Performance

G02CHF is threaded by NAG for parallel execution in multithreaded implementations of the NAG Library.

G02CHF makes calls to BLAS and/or LAPACK routines, which may be threaded within the vendor library used by this implementation. Consult the documentation for the vendor library for further information.

Please consult the X06 Chapter Introduction for information on how to control and interrogate the OpenMP environment used within this routine. Please also consult the Users' Note for your implementation for any additional implementation-specific information.

9 Further Comments

The time taken by G02CHF depends on k.

This routine assumes that the matrix of correlation-like coefficients for the independent variables in the regression is positive definite; it fails if this is not the case.

This correlation matrix will in fact be positive definite whenever the correlation-like matrix and the sums of squares and cross-products (about zero) matrix have been formed either without regard to missing values, or by eliminating **completely** any cases involving missing values for any variable. If, however, these matrices are formed by eliminating cases with missing values from only those calculations involving the variables for which the values are missing, no such statement can be made, and the correlation-like matrix may or may not be positive definite. You should be aware of the possible dangers of using correlation matrices formed in this way (see the G02 Chapter Introduction), but if they nevertheless wish to carry out regressions using such matrices, this routine is capable of handling the inversion of such matrices, provided they are positive definite.

If a matrix is positive definite, its subsequent re-organisation by either of G02CEF or G02CFF will not affect this property and the new matrix can safely be used in this routine. Thus correlation matrices produced by any of G02BDF, G02BEF, G02BKF or G02BLF, even if subsequently modified by either G02CEF or G02CFF, can be handled by this routine.

It should be noted that the routine requires the dependent variable to be the last of the $k+1$ variables whose statistics are provided as input to the routine. If this variable is not correctly positioned in the original data, the means, standard deviations, sums of squares and cross-products about zero, and correlation-like coefficients can be manipulated by using G02CEF or G02CFF to reorder the variables as necessary.

10 Example

This example reads in the sums of squares and cross-products about zero, and correlation-like coefficients for three variables. A multiple linear regression with no constant is then performed with the third and final variable as the dependent variable. Finally the results are printed.

10.1 Program Text

```
    Program g02chfe

!     G02CHF Example Program Text

!     Mark 26 Release. NAG Copyright 2016.

!     .. Use Statements ..
      Use nag_library, Only: g02chf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Integer                          :: i, ifail, k, k1, ldcoef, ldcz, ldrz, &
                                          ldrznv, ldsspz, ldwkz, n
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: coef(:,:), cz(:,:), rz(:,:),         &
                                          rznv(:,:), sspz(:,:), wkz(:,:)
      Real (Kind=nag_wp)               :: reslt(13)
!     .. Executable Statements ..
      Write (nout,*) 'G02CHF Example Program Results'
      Write (nout,*)

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

!     Read in the problem size
      Read (nin,*) n, k
      k1 = k + 1
      ldcoef = k
      ldcz = k
      ldrz = k1
      ldrznv = k
      ldsspz = k1
      ldwkz = k
      Allocate (coef(ldcoef,3),cz(ldcz,k),rz(ldrz,k1),rznv(ldrznv,k),          &
        sspz(ldsspz,k1),wkz(ldwkz,k))

!     Read in data
      Read (nin,*)(sspz(i,1:k1),i=1,k1)
      Read (nin,*)(rz(i,1:k1),i=1,k1)

!     Display data
      Write (nout,*) 'Sums of squares and cross-products about zero:'
      Write (nout,99999)(i,i=1,k1)
      Write (nout,99998)(i,sspz(i,1:k1),i=1,k1)
      Write (nout,*)
      Write (nout,*) 'Correlation-like coefficients:'
      Write (nout,99999)(i,i=1,k1)
      Write (nout,99998)(i,rz(i,1:k1),i=1,k1)
      Write (nout,*)

!     Fit multiple linear regression model
      ifail = 0
      Call g02chf(n,k1,k,sspz,ldsspz,rz,ldrz,reslt,coef,ldcoef,rznv,ldrznv,cz, &
        ldcz,wkz,ldwkz,ifail)

!     Display results
      Write (nout,*) 'Vble     Coef       Std err      t-value'
      Write (nout,99997)(i,coef(i,1:3),i=1,k)
      Write (nout,*)
```

```
      Write (nout,*) 'Analysis of regression table :-'
      Write (nout,*)
      Write (nout,*)                                                    &
        '      Source        Sum of squares  D.F.    Mean square     F-value'
      Write (nout,*)
      Write (nout,99996) 'Due to regression', reslt(1:4)
      Write (nout,99996) 'About  regression', reslt(5:7)
      Write (nout,99996) 'Total            ', reslt(8:9)
      Write (nout,*)
      Write (nout,99995) 'Standard error of estimate =', reslt(10)
      Write (nout,99995) 'Multiple correlation (R)   =', reslt(11)
      Write (nout,99995) 'Determination (R squared)  =', reslt(12)
      Write (nout,99995) 'Corrected R squared        =', reslt(13)
      Write (nout,*)
      Write (nout,*) 'Inverse of correlation matrix of independent variables:'
      Write (nout,99994)(i,i=1,k)
      Write (nout,99993)(i,rznv(i,1:k),i=1,k)
      Write (nout,*)
      Write (nout,*) 'Modified inverse matrix:'
      Write (nout,99994)(i,i=1,k)
      Write (nout,99993)(i,cz(i,1:k),i=1,k)

99999 Format (1X,3I10)
99998 Format (1X,I4,3F10.4)
99997 Format (1X,I3,3F12.4)
99996 Format (1X,A,F14.4,F8.0,2F14.4)
99995 Format (1X,A,F8.4)
99994 Format (1X,2I10)
99993 Format (1X,I4,2F10.4)
    End Program g02chfe
```

10.2 Program Data

```
G02CHF Example Program Data
5  2                            :: N, K
245.0000    99.0000    82.0000
 99.0000   271.0000    52.0000
 82.0000    52.0000    54.0000 :: End of SSPZ
  1.0000     0.3842     0.7129
  0.3842     1.0000     0.4299
  0.7129     0.4299     1.0000 :: End of RZ
```

10.3 Program Results

```
 G02CHF Example Program Results

 Sums of squares and cross-products about zero:
          1         2         3
    1  245.0000   99.0000   82.0000
    2   99.0000  271.0000   52.0000
    3   82.0000   52.0000   54.0000

 Correlation-like coefficients:
          1         2         3
    1    1.0000    0.3842    0.7129
    2    0.3842    1.0000    0.4299
    3    0.7129    0.4299    1.0000

 Vble     Coef       Std err      t-value
   1      0.3017      0.1998      1.5098
   2      0.0817      0.1900      0.4299

 Analysis of regression table :-

       Source        Sum of squares  D.F.    Mean square     F-value

 Due to regression       28.9857      2.       14.4929        1.7382
 About  regression       25.0143      3.        8.3381
 Total                   54.0000      5.
```

```
 Standard error of estimate =  2.8876
 Multiple correlation (R)   =  0.7326
 Determination (R squared)  =  0.5368
 Corrected R squared        =  0.2280

 Inverse of correlation matrix of independent variables:
          1         2
    1    1.1732   -0.4507
    2   -0.4507    1.1732

 Modified inverse matrix:
          1         2
    1    0.0048   -0.0017
    2   -0.0017    0.0043
```