G02CBF – NAG Fortran Library Routine Document

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

G02CBF performs a simple linear regression with no constant, with dependent variable y and independent variable x.

2 Specification

```
SUBROUTINE G02CBF(N, X, Y, RESULT, IFAIL)
INTEGER           N, IFAIL
real              X(N), Y(N), RESULT(20)
```

3 Description

The routine fits a straight line of the form

$$y = bx$$

to the data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

such that

$$y_i = bx_i + e_i; \quad i = 1, 2, \ldots, n \ (n \geq 2).$$

The routine calculates the regression coefficient, b, and the various other statistical quantities by minimizing

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

The input data consists of the n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ on the independent variable x and the dependent variable y.

The quantities calculated are:

(a) Means:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

(b) Standard deviations:

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}; \quad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

(c) Pearson product-moment correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

(d) The regression coefficient, b:

$$b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

(e) 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:

$$SST = \sum_{i=1}^{n} y_i^2; \quad SSD = \sum_{i=1}^{n}(y_i - bx_i)^2, \quad SSR = SST - SSD$$

(f) 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:

$$DFT = n; \quad DFD = n - 1, \quad DFR = 1$$

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

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

(h) The F-value for the analysis of variance:

$$F = MSR/MSD$$

(i) The standard error of the regression coefficient:

$$se(b) = \sqrt{\frac{MSD}{\sum_{i=1}^{n} x_i^2}}$$

(j) The t-value for the regression coefficient:

$$t(b) = \frac{b}{se(b)}$$

4 References

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

5 Parameters

1: N — INTEGER *Input*

On entry: the number n, of pairs of observations.

Constraint: N ≥ 2.

2: X(N) — ***real*** array *Input*

On entry: X(i) must contain x_i, for $i = 1, 2, \ldots, n$.

3: Y(N) — ***real*** array *Input*

On entry: Y(i) must contain y_i, for $i = 1, 2, \ldots, n$.

4: RESULT(20) — ***real*** array *Output*

On exit: the following information:

RESULT(1)	$\bar{x}$, the mean value of the independent variable, x;
RESULT(2)	$\bar{y}$, the mean value of the dependent variable, y;
RESULT(3)	s_x, the standard deviation of the independent variable, x;
RESULT(4)	s_y, the standard deviation of the dependent variable, y;
RESULT(5)	r, the Pearson product-moment correlation between the independent variable x and the dependent variable y;
RESULT(6)	b, the regression coefficient;
RESULT(7)	the value 0.0;
RESULT(8)	$se(b)$, the standard error of the regression coefficient;
RESULT(9)	the value 0.0;
RESULT(10)	$t(b)$, the t-value for the regression coefficient;
RESULT(11)	the value 0.0;
RESULT(12)	SSR, the sum of squares attributable to the regression;
RESULT(13)	DFR, the degrees of freedom attributable to the regression;
RESULT(14)	MSR, the mean square attributable to the regression;

RESULT(15) F, the F-value for the analysis of variance;
RESULT(16) SSD, the sum of squares of deviations about the regression;
RESULT(17) DFD, the degrees of freedom of deviations about the regression;
RESULT(18) MSD, the mean square of deviations about the regression;
RESULT(19) SST, the total sum of squares;
RESULT(20) DFT, the total degrees of freedom.

5: IFAIL — INTEGER *Input/Output*

On entry: IFAIL must be set to 0, −1 or 1. For users not familiar with this parameter (described in Chapter P01) the recommended value is 0.

On exit: IFAIL = 0 unless the routine detects an error (see Section 6).

6 Error Indicators and Warnings

Errors detected by the routine:

IFAIL = 1

On entry, N < 2.

IFAIL = 2

On entry, all N values of at least one of the variables x and y are identical.

7 Accuracy

If, in calculating F or $t(b)$ (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.

The routine does not use ***additional precision*** arithmetic in the accumulation of scalar products so there may be a loss of significant figures for large n.

8 Further Comments

Computation time depends on n.

The routine uses a two-pass algorithm.

9 Example

The following program reads in eight observations on each of two variables, and then performs a simple linear regression with no constant with the first variable as the independent variable, and the second variable as the dependent variable. Finally the results are printed.

9.1 Program Text

Note. The listing of the example program presented below uses bold italicised terms to denote precision-dependent details. Please read the Users' Note for your implementation to check the interpretation of these terms. As explained in the Essential Introduction to this manual, the results produced may not be identical for all implementations.

```
*     G02CBF Example Program Text
*     Mark 14 Revised.  NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          N
      PARAMETER        (N=8)
      INTEGER          NIN, NOUT
```

```
      PARAMETER        (NIN=5,NOUT=6)
*     .. Local Scalars ..
      INTEGER          I, IFAIL
*     .. Local Arrays ..
      real             RESULT(20), X(N), Y(N)
*     .. External Subroutines ..
      EXTERNAL         G02CBF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G02CBF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) (X(I),Y(I),I=1,N)
      WRITE (NOUT,*)
      WRITE (NOUT,*) ' Case     Independent     Dependent'
      WRITE (NOUT,*) 'number     variable       variable'
      WRITE (NOUT,*)
      WRITE (NOUT,99999) (I,X(I),Y(I),I=1,N)
      WRITE (NOUT,*)
      IFAIL = 1
*
      CALL G02CBF(N,X,Y,RESULT,IFAIL)
*
      IF (IFAIL.NE.0) THEN
         WRITE (NOUT,99998) 'Routine fails, IFAIL =', IFAIL
      ELSE
         WRITE (NOUT,99997)
     +     'Mean of independent variable               = ', RESULT(1)
         WRITE (NOUT,99997)
     +     'Mean of   dependent variable               = ', RESULT(2)
         WRITE (NOUT,99997)
     +     'Standard deviation of independent variable = ', RESULT(3)
         WRITE (NOUT,99997)
     +     'Standard deviation of   dependent variable = ', RESULT(4)
         WRITE (NOUT,99997)
     +     'Correlation coefficient                    = ', RESULT(5)
         WRITE (NOUT,*)
         WRITE (NOUT,99997)
     +     'Regression coefficient                     = ', RESULT(6)
         WRITE (NOUT,99997)
     +     'Standard error of coefficient              = ', RESULT(8)
         WRITE (NOUT,99997)
     +     't-value for coefficient                    = ', RESULT(10)
         WRITE (NOUT,*)
         WRITE (NOUT,*) 'Analysis of regression table :-'
         WRITE (NOUT,*)
         WRITE (NOUT,*)
     +'      Source        Sum of squares  D.F.    Mean square     F-val
     +ue'
         WRITE (NOUT,*)
         WRITE (NOUT,99996) 'Due to regression', (RESULT(I),I=12,15)
         WRITE (NOUT,99996) 'About  regression', (RESULT(I),I=16,18)
         WRITE (NOUT,99996) 'Total            ', (RESULT(I),I=19,20)
      END IF
      STOP
*
99999 FORMAT (1X,I4,2F15.4)
99998 FORMAT (1X,A,I2)
99997 FORMAT (1X,A,F8.4)
```

```
99996 FORMAT (1X,A,F14.4,F8.0,2F14.4)
      END
```

9.2 Program Data

```
G02CBF Example Program Data
1.0       20.0
0.0       15.5
4.0       28.3
7.5       45.0
2.5       24.5
0.0       10.0
10.0      99.0
5.0       31.2
```

9.3 Program Results

```
G02CBF Example Program Results

 Case     Independent     Dependent
number     variable       variable

   1         1.0000         20.0000
   2         0.0000         15.5000
   3         4.0000         28.3000
   4         7.5000         45.0000
   5         2.5000         24.5000
   6         0.0000         10.0000
   7        10.0000         99.0000
   8         5.0000         31.2000

Mean of independent variable               =   3.7500
Mean of   dependent variable               =  34.1875
Standard deviation of independent variable =   3.6253
Standard deviation of   dependent variable =  28.2604
Correlation coefficient                    =   0.9096

Regression coefficient                     =   8.2051
Standard error of coefficient              =   0.9052
t-value for coefficient                    =   9.0642

Analysis of regression table :-

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

Due to regression    13767.8054      1.    13767.8054       82.1591
About  regression     1173.0246      7.      167.5749
Total                14940.8300      8.
```