G02CCF – 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

G02CCF performs a simple linear regression with dependent variable y and independent variable x, omitting cases involving missing values.

2 Specification

```
SUBROUTINE G02CCF(N, X, Y, XMISS, YMISS, RESULT, IFAIL)
INTEGER           N, IFAIL
real              X(N), Y(N), XMISS, YMISS, RESULT(21)
```

3 Description

The routine fits a straight line of the form

$$y = a + bx$$

to those of the data points

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

that do not include missing values, such that

$$y_i = a + bx_i + e_i$$

for those (x_i, y_i) $i = 1, 2, \ldots, n$ $(n > 2)$ which do not include missing values.

The routine eliminates all pairs of observations (x_i, y_i) which contain a missing value for either x or y, and then calculates the regression coefficient, b, the regression constant, a (and various other statistical quantities) by minimizing the sum of the e_i^2 over those cases remaining in the calculations.

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.

In addition two values, xm and ym, are given which are considered to represent missing observations for x and y respectively. (See Section 7).

Let $w_i = 0$ if the ith observation of either x or y is missing i.e., if $x_i = xm$ and/or $y_i = ym$; and $w_i = 1$ otherwise, for $i = 1, 2, \ldots, n$.

The quantities calculated are:

(a) Means:

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

(b) Standard deviations:

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

(c) Pearson product-moment correlation coefficient:

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

(d) The regression coefficient, b, and the regression constant, a:

$$b = \frac{\sum_{i=1}^n w_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n w_i (x_i - \bar{x})^2}, \quad a = \bar{y} - b\bar{x}$$

(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:

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

(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:

$$DFT = \sum_{i=1}^{n} w_i - 1; \quad DFD = \sum_{i=1}^{n} w_i - 2; \quad DFR = 1$$

(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, $se(b)$, and the standard error of the regression constant, $se(a)$:

$$se(b) = \sqrt{\frac{MSD}{\sum_{i=1}^{n} w_i(x_i - \bar{x})^2}}; \quad se(a) = \sqrt{MSD\left(\frac{1}{\sum_{i=1}^{n} w_i} + \frac{\bar{x}^2}{\sum_{i=1}^{n} w_i(x_i - \bar{x})^2}\right)}$$

(j) The t-value for the regression coefficient, $t(b)$, and the t-value for the regression constant, $t(a)$:

$$t(b) = \frac{b}{se(b)}; \quad t(a) = \frac{a}{se(b)}$$

(k) The number of observations used in the calculations:

$$n_c = \sum_{i=1}^{n} w_i$$

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: XMISS — ***real*** *Input*

On entry: the value xm which is to be taken as the missing value for the variable x. See Section 7.

5: YMISS — ***real*** *Input*

On entry: the value ym which is to be taken as the missing value for the variable y. See Section 7.

6: RESULT(21) — ***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)	a, the regression constant;
RESULT(8)	$se(b)$, the standard error of the regression coefficient;
RESULT(9)	$se(a)$, the standard error of the regression constant;
RESULT(10)	$t(b)$, the t-value for the regression coefficient;
RESULT(11)	$t(a)$, the t-value for the regression constant;
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;
RESULT(21)	n_c, the number of observations used in the calculations.

7: 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 $\leq$ 2.

IFAIL = 2

After observations with missing values were omitted, two or fewer cases remained.

IFAIL = 3

After observations with missing values were omitted, all remaining values of at least one of the variables x and y were identical.

7 Accuracy

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

Users are warned of the need to exercise extreme care in their selection of missing values, since the routine treats as missing values for a variable, all values in the inclusive range $(1 \pm \text{ACC}) \times zm$, where zm is the missing value for that variable, specified by the user, and ACC is a machine-dependent constant, (see the Users' Note for your implementation). The user 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.

If, in calculating F, $t(b)$ or $t(a)$ (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 Further Comments

The time taken by the routine depends on n and the number of missing observations.

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 the first variable as the independent variable, and the second variable as the dependent variable, omitting cases involving missing values (0.0 for the first variable, 99.0 for the second). 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.

```
*     G02CCF 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 ..
      real             XM, YM
      INTEGER          I, IFAIL
*     .. Local Arrays ..
      real             RESULT(21), X(N), Y(N)
*     .. External Subroutines ..
      EXTERNAL         G02CCF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G02CCF 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,*)
*
*     Set up missing values
*
      XM = 0.0e0
      YM = 99.0e0
      IFAIL = 1
*
```

```
      CALL G02CCF(N,X,Y,XM,YM,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,99997)
     +     'Regression constant                       = ', RESULT(7)
         WRITE (NOUT,99997)
     +     'Standard error of constant                = ', RESULT(9)
         WRITE (NOUT,99997)
     +     't-value for constant                      = ', RESULT(11)
         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)
         WRITE (NOUT,*)
         WRITE (NOUT,99995) 'Number of cases used = ', RESULT(21)
      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)
99995 FORMAT (1X,A,F3.0)
      END
```

9.2 Program Data

```
G02CCF 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

```
G02CCF 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               =   4.0000
Mean of   dependent variable               =  29.8000
Standard deviation of independent variable =   2.4749
Standard deviation of   dependent variable =   9.4787
Correlation coefficient                    =   0.9799

Regression coefficient                     =   3.7531
Standard error of coefficient              =   0.4409
t-value for coefficient                    =   8.5128

Regression constant                        =  14.7878
Standard error of constant                 =   2.0155
t-value for constant                       =   7.3370

Analysis of regression table :-

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

Due to regression      345.0940      1.      345.0940       72.4682
About  regression       14.2860      3.        4.7620
Total                  359.3800      4.

Number of cases used =  5.
```