NAG Library Routine Document

G02CAF

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

G02CAF performs a simple linear regression with dependent variable y and independent variable x.

2 Specification

```
SUBROUTINE G02CAF (N, X, Y, RESULT, IFAIL)

INTEGER              N, IFAIL
REAL (KIND=nag_wp) X(N), Y(N), RESULT(20)
```

3 Description

G02CAF fits a straight line of the form

$$y = a + bx$$

to the data points

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

such that

$$y_i = a + bx_i + e_i, \qquad i = 1, 2, \ldots, n(n > 2).$$

The routine calculates the regression coefficient, b, the regression constant, a (and various other statistical quantities) by minimizing

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

The input data consist 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:

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

(b) Standard deviations:

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

(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, and the regression constant, a:

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

(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:

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

(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:

$$\mathrm{DFT} = n - 1;\ \mathrm{DFD} = n - 2;\ \mathrm{DFR} = 1.$$

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

$$\mathrm{MSR} = \mathrm{SSR}/\mathrm{DFR};\mathrm{MSD} = \mathrm{SSD}/\mathrm{DFD}.$$

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

$$F = \mathrm{MSR}/\mathrm{MSD}.$$

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

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

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

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

4 References

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

5 Parameters

1: N – INTEGER *Input*

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

Constraint: $\mathrm{N} > 2$.

2: X(N) – REAL (KIND=nag_wp) array *Input*

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

3: Y(N) – REAL (KIND=nag_wp) array *Input*

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

4: RESULT(20) – REAL (KIND=nag_wp) 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.

5: 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, $\mathrm{N} \leq 2$.

IFAIL $= 2$

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

7 Accuracy

G02CAF does not use ***additional precision*** arithmetic for the accumulation of scalar products, so there may be a loss of significant figures for large n.

If, in calculating F, $t(a)$ 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.

8 Further Comments

The time taken by G02CAF depends on n.

The routine uses a two-pass algorithm.

9 Example

This example 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. Finally the results are printed.

9.1 Program Text

```
    Program g02cafe

!     G02CAF Example Program Text

!     Mark 24 Release. NAG Copyright 2012.

!     .. Use Statements ..
      Use nag_library, Only: g02caf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Integer                          :: i, ifail, n
!     .. Local Arrays ..
      Real (Kind=nag_wp)               :: reslt(20)
      Real (Kind=nag_wp), Allocatable  :: x(:), y(:)
!     .. Executable Statements ..
      Write (nout,*) 'G02CAF Example Program Results'
      Write (nout,*)

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

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

      Allocate (x(n),y(n))

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

!     Display data
      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,*)

!     Fit linear regression model
      ifail = 0
      Call g02caf(n,x,y,reslt,ifail)

!     Display results
      Write (nout,99998) 'Mean of independent variable               = ', &
        reslt(1)
      Write (nout,99998) 'Mean of   dependent variable               = ', &
        reslt(2)
      Write (nout,99998) 'Standard deviation of independent variable = ', &
        reslt(3)
      Write (nout,99998) 'Standard deviation of   dependent variable = ', &
        reslt(4)
      Write (nout,99998) 'Correlation coefficient                    = ', &
        reslt(5)
```

```
      Write (nout,*)
      Write (nout,99998) 'Regression coefficient                          = ', &
        reslt(6)
      Write (nout,99998) 'Standard error of coefficient                   = ', &
        reslt(8)
      Write (nout,99998) 't-value for coefficient                         = ', &
        reslt(10)
      Write (nout,*)
      Write (nout,99998) 'Regression constant                             = ', &
        reslt(7)
      Write (nout,99998) 'Standard error of constant                      = ', &
        reslt(9)
      Write (nout,99998) 't-value for constant                            = ', &
        reslt(11)
      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,99997) 'Due to regression', reslt(12:15)
      Write (nout,99997) 'About  regression', reslt(16:18)
      Write (nout,99997) 'Total            ', reslt(19:20)

99999 Format (1X,I4,2F15.4)
99998 Format (1X,A,F8.4)
99997 Format (1X,A,F14.3,F8.0,2F14.3)
    End Program g02cafe
```

9.2 Program Data

```
G02CAF Example Program Data
8               :: N
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  :: End of X,Y
```

9.3 Program Results

```
 G02CAF 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                     =   7.0905
 Standard error of coefficient              =   1.3224
 t-value for coefficient                    =   5.3620

 Regression constant                        =   7.5982
 Standard error of constant                 =   6.6858
```

```
t-value for constant                          =   1.1365

Analysis of regression table :-

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

Due to regression      4625.303      1.      4625.303        28.751
About  regression       965.245      6.       160.874
Total                  5590.549      7.
```