

# NAG Fortran Library Routine Document

## G02EAF

**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

G02EAF calculates the residual sums of squares for all possible linear regressions for a given set of independent variables.

### 2 Specification

```

SUBROUTINE G02EAF(MEAN, WEIGHT, N, M, X, LDX, NAME, ISX, Y, WT, NMOD,
1          MODEL, LDM, RSS, NTERMS, MRANK, WK, IFAIL)
INTEGER      N, M, LDX, ISX(M), NMOD, LDM, NTERMS(LDM), MRANK(LDM),
1          IFAIL
real       X(LDX,M), Y(N), WT(*), RSS(LDM), WK(N*(M+1))
CHARACTER*1  MEAN, WEIGHT
CHARACTER*(*) NAME(M), MODEL(LDM,M)

```

### 3 Description

For a set of  $k$  possible independent variables there are  $2^k$  linear regression models with from zero to  $k$  independent variables in each model. For example if  $k = 3$  and the variables are A, B and C then the possible models are:

- (i) null model
- (ii) A
- (iii) B
- (iv) C
- (v) A and B
- (vi) A and C
- (vii) B and C
- (viii) A, B and C.

G02EAF calculates the residual sums of squares from each of the  $2^k$  possible models. The method used involves a *QR* decomposition of the matrix of possible independent variables. Independent variables are then moved into and out of the model by a series of Givens rotations and the residual sums of squares computed for each model; see Clark (1981) and Smith and Bremner (1989).

The computed residual sums of squares are then ordered first by increasing number of terms in the model, then by decreasing size of residual sums of squares. So the first model will always have the largest residual sum of squares and the  $2^k$ th will always have the smallest. This aids the user in selecting the best possible model from the given set of independent variables.

G02EAF allows the user to specify some independent variables that must be in the model, the forced variables. The other independent variables from which the possible models are to be formed are the free variables.

### 4 References

Clark M R B (1981) A Givens algorithm for moving from one linear model to another without going back to the data *Appl. Statist.* **30** 198–203

Smith D M and Bremner J M (1989) All possible subset regressions using the *QR* decomposition *Comput. Statist. Data Anal.* 7 217–236

Weisberg S (1985) *Applied Linear Regression* Wiley

## 5 Parameters

- 1: MEAN – CHARACTER\*1 *Input*  
*On entry:* indicates if a mean term is to be included.  
 If MEAN = 'M' (Mean), a mean term, intercept, will be included in the models.  
 If MEAN = 'Z' (Zero), the models will pass through the origin, zero-point.  
*Constraint:* MEAN = 'M' or 'Z'.
- 2: WEIGHT – CHARACTER\*1 *Input*  
*On entry:* indicates if weights are to be used.  
 If WEIGHT = 'U' (Unweighted), least-squares estimation is used.  
 If WEIGHT = 'W' (Weighted), weighted least-squares is used and weights must be supplied in array WT.  
*Constraint:* WEIGHT = 'U' or 'W'.
- 3: N – INTEGER *Input*  
*On entry:* the number of observations.  
*Constraint:*  $N \geq 2$ .
- 4: M – INTEGER *Input*  
*On entry:* the maximum number of variables contained in X.  
*Constraint:*  $M \geq 2$ .
- 5: X(LDX,M) – *real* array *Input*  
*On entry:*  $X(i, j)$  must contain the  $i$ th observation for the  $j$ th independent variable, for  $i = 1, 2, \dots, N$ ;  $j = 1, 2, \dots, M$ .
- 6: LDX – INTEGER *Input*  
*On entry:* the first dimension of the array X as declared in the (sub)program from which G02EAF is called.  
*Constraint:*  $LDX \geq N$ .
- 7: NAME(M) – CHARACTER\*(\*) array *Input*  
*On entry:* NAME( $j$ ) must contain the name of the variable in column  $j$  of X, for  $j = 1, 2, \dots, M$ .
- 8: ISX(M) – INTEGER array *Input*  
*On entry:* indicates which independent variables are to be considered in the model.  
 If  $ISX(j) \geq 2$ , then the variable contained in the  $j$ th column of X is included in all regression models, i.e., is a forced variable.  
 If  $ISX(j) = 1$ , then the variable contained in the  $j$ th column of X is included in the set from which the regression models are chosen, i.e., is a free variable.  
 If  $ISX(j) = 0$ , then the variable contained in the  $j$ th column of X is not included in the models.  
*Constraint:*  $ISX(j) \geq 0$ , for  $j = 1, 2, \dots, M$  and at least one value of  $ISX = 1$ .

- 9:  $Y(N)$  – *real* array *Input*  
*On entry:*  $Y(i)$  must contain the  $i$ th observation on the dependent variable,  $y_i$ , for  $i = 1, 2, \dots, n$ .
- 10:  $WT(*)$  – *real* array *Input*  
*On entry:* if  $WEIGHT = 'W'$ , then  $WT$  must contain the weights to be used in the weighted regression.  
 If  $WT(i) = 0.0$ , then the  $i$ th observation is not included in the model, in which case the effective number of observations is the number of observations with non-zero weights.  
 If  $WEIGHT = 'U'$ , then  $WT$  is not referenced and the effective number of observations is  $N$ .  
*Constraint:* if  $WEIGHT = 'W'$ ,  $WT(i) \geq 0.0$ , for  $i = 1, 2, \dots, n$ .
- 11:  $NMOD$  – INTEGER *Output*  
*On exit:* the total number of models for which residual sums of squares have been calculated.
- 12:  $MODEL(LDM, M)$  – CHARACTER\*(\*) array *Output*  
*On exit:* the first  $NTERMS(i)$  elements of the  $i$ th row of  $MODEL$  contain the names of the independent variables, as given in  $NAME$ , that are included in the  $i$ th model.  
*Constraint:* the length of  $MODEL$  should be greater or equal to the length of  $NAME$ .
- 13:  $LDM$  – INTEGER *Input*  
*On entry:* the first dimension of the array  $MODEL$  as declared in the (sub)program from which G02EAF is called.  
*Constraint:* at a minimum  $LDM \geq M$ , but  $LDM$  also needs to be greater or equal to the number of models to be generated. If there are  $k$  free independent variables then  $LDM \geq 2^k$ .
- 14:  $RSS(LDM)$  – *real* array *Output*  
*On exit:*  $RSS(i)$  contains the residual sum of squares for the  $i$ th model, for  $i = 1, 2, \dots, NMOD$ .
- 15:  $NTERMS(LDM)$  – INTEGER array *Output*  
*On exit:*  $NTERMS(i)$  contains the number of independent variables in the  $i$ th model, not including the mean if one is fitted, for  $i = 1, 2, \dots, NMOD$ .
- 16:  $MRANK(LDM)$  – INTEGER array *Output*  
*On exit:*  $MRANK(i)$  contains the rank of the residual sum of squares for the  $i$ th model, i.e., model with smallest sum of squares has rank 1.
- 17:  $WK(N*(M+1))$  – *real* array *Workspace*
- 18:  $IFAIL$  – INTEGER *Input/Output*  
*On entry:*  $IFAIL$  must be set to 0, -1 or 1. Users who are unfamiliar with this parameter should refer to Chapter P01 for details.  
*On exit:*  $IFAIL = 0$  unless the routine detects an error (see Section 6).  
 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, for users 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.**

## 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,  $N < 2$ ,  
 or  $M < 2$ ,  
 or  $LDX < N$ ,  
 or  $LDM < M$ ,  
 or  $MEAN \neq 'M'$  or  $'Z'$ ,  
 or  $WEIGHT \neq 'U'$  or  $'W'$ .

$IFAIL = 2$

On entry,  $WEIGHT = 'W'$  and a value of  $WT < 0.0$ .

$IFAIL = 3$

On entry, a value of  $ISX < 0.0$ ,  
 or there are no free variables, i.e., no element of  $ISX = 1$ .

$IFAIL = 4$

On entry,  $LDM < \text{the number of possible models} = 2^k$ , where  $k$  is the number of free independent variables from  $ISX$ .

$IFAIL = 5$

On entry, the number of independent variables to be considered (forced plus free plus mean if included) is greater or equal to the effective number of observations.

$IFAIL = 6$

The full model is not of full rank i.e., some of the independent variables may be linear combinations of other independent variables. Variables must be excluded from the model in order to give full rank.

## 7 Accuracy

For a discussion of the improved accuracy obtained by using a method based on the  $QR$  decomposition see Smith and Bremner (1989).

## 8 Further Comments

G02ECF may be used to compute  $R^2$  and  $C_p$ -values from the results of G02EAF.

If a mean has been included in the model and no variables are forced in then  $RSS(1)$  contains the total sum of squares and in many situations a reasonable estimate of the variance of the errors is given by  $RSS(NMOD)/(N - 1 - NTERMS(NMOD))$ .

## 9 Example

The data for this example is given in Weisberg (1985). The independent variables and the dependent variable are read, as are the names of the variables. These names are as given in Weisberg (1985). The residual sums of squares computed and printed with the names of the variables in the model.

## 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.

```
*      G02EAF Example Program Text
*      Mark 14 Release.  NAG Copyright 1989.
*      .. Parameters ..
INTEGER      NMAX, MMAX, LMAX
PARAMETER    (NMAX=20,MMAX=6,LMAX=32)
INTEGER      NIN, NOUT
PARAMETER    (NIN=5,NOUT=6)
*      .. Local Scalars ..
INTEGER      I, IFAIL, II, J, M, N, NMOD
*      .. Local Arrays ..
real        RSS(LMAX), WK(NMAX*(MMAX+1)), WT(NMAX),
+           X(NMAX,MMAX), Y(NMAX)
INTEGER      ISX(MMAX), MRANK(LMAX), NTERMS(LMAX)
CHARACTER*3  MODEL(LMAX,MMAX), NAME(MMAX)
*      .. External Subroutines ..
EXTERNAL     G02EAF
*      .. Executable Statements ..
WRITE (NOUT,*) 'G02EAF Example Program Results'
*      Skip heading in data file
READ (NIN,*)
READ (NIN,*) N, M
IF (M.LE.MMAX .AND. N.LE.NMAX) THEN
  DO 20 I = 1, N
    READ (NIN,*) (X(I,J),J=1,M), Y(I)
20  CONTINUE
    READ (NIN,*) (ISX(J),J=1,M)
    READ (NIN,*) (NAME(J),J=1,M)
    IFAIL = 0
*
    CALL G02EAF('M','U',N,M,X,NMAX,NAME,ISX,Y,WT,NMOD,MODEL,LMAX,
+             RSS,NTERMS,MRANK,WK,IFAIL)
*
    WRITE (NOUT,*)
    WRITE (NOUT,*) 'Number of      RSS      RANK      MODEL'
    WRITE (NOUT,*) 'parameters'
    DO 40 I = 1, NMOD
      II = NTERMS(I)
      WRITE (NOUT,99999) II, RSS(I), MRANK(I), (MODEL(I,J),J=1,II)
40  CONTINUE
    END IF
    STOP
*
99999 FORMAT (1X,I8,F11.4,I4,3X,5(1X,A))
END
```

## 9.2 Program Data

G02EAF Example Program Data

```
20 6
 0. 1125.0 232.0 7160.0 85.9 8905.0 1.5563
 7.  920.0 268.0 8804.0 86.5 7388.0 0.8976
15.  835.0 271.0 8108.0 85.2 5348.0 0.7482
22. 1000.0 237.0 6370.0 83.8 8056.0 0.7160
29. 1150.0 192.0 6441.0 82.1 6960.0 0.3010
37.  990.0 202.0 5154.0 79.2 5690.0 0.3617
44.  840.0 184.0 5896.0 81.2 6932.0 0.1139
58.  650.0 200.0 5336.0 80.6 5400.0 0.1139
65.  640.0 180.0 5041.0 78.4 3177.0 -0.2218
72.  583.0 165.0 5012.0 79.3 4461.0 -0.1549
80.  570.0 151.0 4825.0 78.7 3901.0 0.0000
86.  570.0 171.0 4391.0 78.0 5002.0 0.0000
93.  510.0 243.0 4320.0 72.3 4665.0 -0.0969
100. 555.0 147.0 3709.0 74.9 4642.0 -0.2218
107. 460.0 286.0 3969.0 74.4 4840.0 -0.3979
122. 275.0 198.0 3558.0 72.5 4479.0 -0.1549
```

```

129. 510.0 196.0 4361.0 57.7 4200.0 -0.2218
151. 165.0 210.0 3301.0 71.8 3410.0 -0.3979
171. 244.0 327.0 2964.0 72.5 3360.0 -0.5229
220. 79.0 334.0 2777.0 71.9 2599.0 -0.0458
0    1    1    1    1    1
'DAY' 'BOD' 'TKN' 'TS' 'TVS' 'COD'

```

### 9.3 Program Results

G02EAF Example Program Results

Number of parameters	RSS	RANK	MODEL
0	5.0634	32	
1	5.0219	31	TKN
1	2.5044	30	TVS
1	2.0338	28	BOD
1	1.5563	25	COD
1	1.5370	24	TS
2	2.4381	29	TKN TVS
2	1.7462	27	BOD TVS
2	1.5921	26	BOD TKN
2	1.4963	23	BOD COD
2	1.4707	22	TKN TS
2	1.4590	21	TS TVS
2	1.4397	20	BOD TS
2	1.4388	19	TKN COD
2	1.3287	15	TVS COD
2	1.0850	8	TS COD
3	1.4257	18	BOD TKN TVS
3	1.3900	17	TKN TS TVS
3	1.3894	16	BOD TS TVS
3	1.3204	14	BOD TVS COD
3	1.2764	13	BOD TKN COD
3	1.2582	12	BOD TKN TS
3	1.2179	10	TKN TVS COD
3	1.0644	7	BOD TS COD
3	1.0634	6	TS TVS COD
3	0.9871	4	TKN TS COD
4	1.2199	11	BOD TKN TS TVS
4	1.1565	9	BOD TKN TVS COD
4	1.0388	5	BOD TS TVS COD
4	0.9871	3	BOD TKN TS COD
4	0.9653	2	TKN TS TVS COD
5	0.9652	1	BOD TKN TS TVS COD

---