NAG Fortran Library Routine Document

G02EEF

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

G02EEF carries out one step of a forward selection procedure in order to enable the 'best' linear regression model to be found.

2 Specification

```
 SUBROUTINE G02EEF(ISTEP, MEAN, WEIGHT, N, M, X, LDX, NAME, ISX, MAXIP,
1                  Y, WT, FIN, ADDVAR, NEWVAR, CHRSS, F, MODEL, NTERM,
2                  RSS, IDF, IFR, FREE, EXSS, Q, LDQ, P, WK, IFAIL)
 INTEGER           ISTEP, N, M, LDX, ISX(M), MAXIP, NTERM, IDF, IFR, LDQ,
1                  IFAIL
 real              X(LDX,M), Y(N), WT(*), FIN, CHRSS, F, RSS,
1                  EXSS(MAXIP), Q(LDQ,MAXIP+2), P(M), WK(2*MAXIP)
 LOGICAL           ADDVAR
 CHARACTER*1       MEAN, WEIGHT
 CHARACTER*(*)     NAME(M), NEWVAR, MODEL(MAXIP), FREE(MAXIP)
```

3 Description

One method of selecting a linear regression model from a given set of independent variables is by forward selection. The following procedure is used:

(i) Select the best fitting independent variable, i.e., the independent variable which gives the smallest residual sum of squares. If the F-test for this variable is greater than a chosen critical value, F_c, then include the variable in the model, else stop.

(ii) Find the independent variable that leads to the greatest reduction in the residual sum of squares when added to the current model.

(iii) If the F-test for this variable is greater than a chosen critical value, F_c, then include the variable in the model and go to (b), otherwise stop.

At any step the variables not in the model are known as the free terms.

G02EEF allows the user to specify some independent variables that must be in the model, these are known as forced variables.

The computational procedure involves the use of QR decompositions, the R and the Q matrices being updated as each new variable is added to the model. In addition the matrix $Q^T X_{\text{free}}$, where X_{free} is the matrix of variables not included in the model, is updated.

G02EEF computes one step of the forward selection procedure at a call. The results produced at each step may be printed or used as inputs to G02DDF, in order to compute the regression coefficients for the model fitted at that step. Repeated calls to G02EEF should be made until $F < F_c$ is indicated.

4 References

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

Weisberg S (1985) *Applied Linear Regression* Wiley

5 Parameters

Note: after the initial call to this with ISTEP $= 0$ all parameters except FIN must not be changed by the user between calls.

1: ISTEP – INTEGER *Input/Output*

On entry: indicates which step in the forward selection process is to be carried out.

If ISTEP $= 0$, then the process is initialised.

Constraint: ISTEP ≥ 0.

On exit: ISTEP is incremented by 1.

2: 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 model.

If MEAN = 'Z' (Zero), the model will pass through the origin, zero-point.

Constraint: MEAN = 'M' or 'Z'.

3: 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'.

4: N – INTEGER *Input*

On entry: the number of observations.

Constraint: N ≥ 2.

5: M – INTEGER *Input*

On entry: the total number of independent variables in the data set.

Constraint: M ≥ 1.

6: X(LDX,M) – ***real*** array *Input*

On entry: X(i,j) must contain the ith observation for the jth independent variable, for $i = 1,2,\ldots,\mathrm{N}$; $j = 1,2,\ldots,\mathrm{M}$.

7: LDX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G02EEF is called.

Constraint: LDX $\geq$ N.

8: NAME(M) – CHARACTER*(*) array *Input*

On entry: NAME(j) must contain the name of the independent variable in column j of X, for $j = 1,2,\ldots,\mathrm{M}$.

9: ISX(M) – INTEGER array *Input*

On entry: indicates which independent variables could be considered for inclusion in the regression.

If ISX(j) ≥ 2, then the variable contained in the jth column of X is automatically included in the regression model, for $j = 1, 2, \ldots, \mathrm{M}$.

If ISX(j) $= 1$, then the variable contained in the jth column of X is considered for inclusion in the regression model, for $j = 1, 2, \ldots, \mathrm{M}$.

If ISX(j) $= 0$, the variable in the jth column is not considered for inclusion in the model, for $j = 1, 2, \ldots, \mathrm{M}$.

Constraint: ISX(j) ≥ 0 and at least one value of ISX(j) $= 1$, for $j = 1, 2, \ldots, \mathrm{M}$.

10: MAXIP – INTEGER *Input*

On entry: the maximum number of independent variables to be included in the model.

Constraints:

if MEAN = 'M', MAXIP $\geq 1+$ number of values of ISX > 0,
if MEAN = 'Z', MAXIP $\geq$ number of values of ISX > 0.

11: Y(N) – ***real*** array *Input*

On entry: the dependent variable.

12: WT(*) – ***real*** array *Input*

On entry: if WEIGHT = 'W', then WT must contain the weights to be used in the weighted regression, W.

If WT(i) $= 0.0$, then the ith 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) ≥ 0.0, for $i = 1, 2, \ldots, \mathrm{N}$.

13: FIN – ***real*** *Input*

On entry: the critical value of the F statistic for the term to be included in the model, F_{c}.

Suggested value: 2.0 is a commonly used value in exploratory modelling.

Constraint: FIN ≥ 0.0.

14: ADDVAR – LOGICAL *Output*

On exit: indicates if a variable has been added to the model.

If ADDVAR = .TRUE., then a variable has been added to the model.

If ADDVAR = .FALSE., then no variable had an F value greater than F_{c} and none were added to the model.

15: NEWVAR – CHARACTER*(*) *Output*

On exit: if ADDVAR = .TRUE., then NEWVAR contains the name of the variable added to the model.

Constraint: the declared size of NEWVAR must be greater than or equal to the declared size of NAME.

16: CHRSS – ***real*** *Output*

On exit: if ADDVAR = .TRUE., then CHRSS contains the change in the residual sum of squares due to adding variable NEWVAR.

17: F – ***real*** *Output*

On exit: if ADDVAR = .TRUE., then F contains the F statistic for the inclusion of the variable in NEWVAR.

18: MODEL(MAXIP) – CHARACTER*(*) array *Input/Output*

On entry: if ISTEP $= 0$, then MODEL need not be set.

If ISTEP $\neq 0$, then MODEL must contain the values returned by the previous call to G02EEF.

Constraint: the declared size of MODEL must be greater than or equal to the declared size of NAME.

On exit: the names of the variables in the current model.

19: NTERM – INTEGER *Input/Output*

On entry: if ISTEP $= 0$, then NTERM need not be set.

If ISTEP $\neq 0$, then NTERM must contain the value returned by the previous call to G02EEF.

On exit: the number of independent variables in the current model, not including the mean, if any.

20: RSS – ***real*** *Input/Output*

On entry: if ISTEP $= 0$, then RSS need not be set.

If ISTEP $\neq 0$, then RSS must contain the value returned by the previous call to G02EEF.

On exit: the residual sums of squares for the current model.

21: IDF – INTEGER *Input/Output*

On entry: if ISTEP $= 0$, then IDF need not be set.

If ISTEP $\neq 0$, then IDF must contain the value returned by the previous call to G02EEF.

On exit: the degrees of freedom for the residual sum of squares for the current model.

22: IFR – INTEGER *Input/Output*

On entry: if ISTEP $= 0$, then IFR need not be set.

If ISTEP $\neq 0$, then IFR must contain the value returned by the previous call to G02EEF.

On exit: the number of free independent variables, i.e., the number of variables not in the model that are still being considered for selection.

23: FREE(MAXIP) – CHARACTER*(*) array *Input/Output*

On entry: if ISTEP $= 0$, then FREE need not be set.

If ISTEP $\neq 0$, then FREE must contain the values returned by the previous call to G02EEF.

Constraint: the declared size of FREE must be greater than or equal to the declared size of NAME.

On exit: the first IFR values of FREE contain the names of the free variables.

24: EXSS(MAXIP) – ***real*** array *Output*

On exit: the first IFR values of EXSS contain what would be the change in regression sum of squares if the free variables had been added to the model, i.e., the extra sum of squares for the free variables. EXSS(i) contains what would be the change in regression sum of squares if the variable FREE(i) had been added to the model.

25: Q(LDQ,MAXIP+2) – ***real*** array *Input/Output*

On entry: if ISTEP $= 0$, then Q need not be set.

If ISTEP $\neq 0$, then Q must contain the values returned by the previous call to G02EEF.

On exit: the results of the QR decomposition for the current model:

the first column of Q contains $c = Q^T y$ (or $Q^T W^{\frac{1}{2}} y$ where W is the vector of weights if used);
the upper triangular part of columns 2 to IP + 1 contain the R matrix;
the strictly lower triangular part of columns 2 to IP + 1 contain details of the Q matrix;
the remaining IP + 1 to IP + IFR columns of contain $Q^T X_{\text{free}}$ (or $Q^T W^{\frac{1}{2}} X_{\text{free}}$).

26: LDQ – INTEGER *Input*

On entry: the first dimension of the array Q as declared in the (sub)program from which G02EEF is called.

Constraint: LDQ $\geq$ N.

27: P(M) – ***real*** array *Input/Output*

On entry: if ISTEP $= 0$, then P need not be set.

If ISTEP $\neq 0$, then P must contain the values returned by the previous call to G02EEF.

On exit: first IP elements of P must contain the zeta values for the QR decomposition (see F08AEF (SGEQRF/DGEQRF) for details).

28: WK(2∗MAXIP) – ***real*** array *Workspace*

29: 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 < 1,
or M < 1,
or LDX $<$ N,
or LDQ $<$ N,
or ISTEP < 0,
or ISTEP $\neq 0$ and NTERM $= 0$,
or ISTEP $\neq 0$ and RSS ≤ 0.0,
or FIN < 0.0,
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, the degrees of freedom will be zero if a variable is selected i.e., the number of variables in the model plus 1 is equal to the effective number of observations.

IFAIL = 4

On entry, a value of ISX < 0,
or there are no forced or free variables, i.e., no element of ISX > 0,
or the value of MAXIP is too small for number of variables indicated by ISX.

IFAIL = 5

On entry, the variables forced into the model are not of full rank, i.e., some of these variables are linear combinations of others.

IFAIL = 6

On entry, there are no free variables, i.e., no element of ISX $= 0$.

IFAIL = 7

The value of the change in the sum of squares is greater than the input value of RSS. This may occur due to rounding errors if the true residual sum of squares for the new model is small relative to the residual sum of squares for the previous model.

7 Accuracy

As G02EEF uses a QR transformation the results will often be more accurate than traditional algorithms using methods based on the cross-products of the dependent and independent variables.

8 Further Comments

None.

9 Example

The data, from an oxygen uptake experiment, is given by Weisberg (1985). The names of the variables are as given in Weisberg (1985). The independent and dependent variables are read and G02EEF is repeatedly called until ADDVAR = .FALSE.. At each step the F statistic, the free variables and their extra sum of squares are printed; also, except for when ADDVAR = .FALSE., the new variable, the change in the residual sum of squares and the terms in the model 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.

```
*     G02EEF Example Program Text
*     Mark 14 Release.  NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          NMAX, MMAX
      PARAMETER        (NMAX=20,MMAX=8)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*     .. Local Scalars ..
      real             CHRSS, F, FIN, RSS
      INTEGER          I, IDF, IFAIL, IFR, IM, ISTEP, J, M, N, NTERM
      LOGICAL          ADDVAR
      CHARACTER        MEAN, WEIGHT
      CHARACTER*3      NEWVAR
*     .. Local Arrays ..
      real             EXSS(MMAX), P(MMAX+1), Q(NMAX,MMAX+2),
```

```
     +                 WK(2*MMAX), WT(NMAX), X(NMAX,MMAX), Y(NMAX)
      INTEGER          ISX(MMAX)
      CHARACTER*3      FREE(MMAX), MODEL(MMAX), NAME(MMAX)
*     .. External Subroutines ..
      EXTERNAL         G02EEF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G02EEF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) N, M, MEAN, WEIGHT
      IF (M.LT.MMAX .AND. N.LE.NMAX) THEN
         IF (WEIGHT.EQ.'W' .OR. WEIGHT.EQ.'w') THEN
            DO 20 I = 1, N
               READ (NIN,*) (X(I,J),J=1,M), Y(I), WT(I)
   20       CONTINUE
         ELSE
            DO 40 I = 1, N
               READ (NIN,*) (X(I,J),J=1,M), Y(I)
   40       CONTINUE
         END IF
         READ (NIN,*) (ISX(J),J=1,M)
         READ (NIN,*) (NAME(J),J=1,M)
         READ (NIN,*) FIN
         IF (MEAN.EQ.'M' .OR. MEAN.EQ.'m') THEN
            IM = 1
         ELSE
            IM = 0
         END IF
         ISTEP = 0
         DO 60 I = 1, M
            IFAIL = 0
*
            CALL G02EEF(ISTEP,MEAN,WEIGHT,N,M,X,NMAX,NAME,ISX,MMAX,Y,WT,
     +                  FIN,ADDVAR,NEWVAR,CHRSS,F,MODEL,NTERM,RSS,IDF,
     +                  IFR,FREE,EXSS,Q,NMAX,P,WK,IFAIL)
*
            IF (IFAIL.NE.0) GO TO 80
            WRITE (NOUT,*)
            WRITE (NOUT,99999) 'Step ', ISTEP
            IF ( .NOT. ADDVAR) THEN
               WRITE (NOUT,99998)
     +           'No further variables added maximum F =', F
               WRITE (NOUT,99993) 'Free variables:  ', (FREE(J),J=1,IFR)
               WRITE (NOUT,*)
     +          'Change in residual sums of squares for free variables:'
               WRITE (NOUT,99992) '                 ', (EXSS(J),J=1,IFR)
               GO TO 80
            ELSE
               WRITE (NOUT,99997) 'Added variable is ', NEWVAR
               WRITE (NOUT,99996) 'Change in residual sum of squares =',
     +           CHRSS
               WRITE (NOUT,99998) 'F Statistic = ', F
               WRITE (NOUT,*)
               WRITE (NOUT,99995) 'Variables in model:',
     +           (MODEL(J),J=1,NTERM)
               WRITE (NOUT,*)
               WRITE (NOUT,99994) 'Residual sum of squares = ', RSS
               WRITE (NOUT,99999) 'Degrees of freedom = ', IDF
               WRITE (NOUT,*)
               IF (IFR.EQ.0) THEN
                  WRITE (NOUT,*) 'No free variables remaining'
                  GO TO 80
               END IF
               WRITE (NOUT,99993) 'Free variables:  ', (FREE(J),J=1,IFR)
               WRITE (NOUT,*)
     +          'Change in residual sums of squares for free variables:'
               WRITE (NOUT,99992) '                 ', (EXSS(J),J=1,IFR)
            END IF
   60    CONTINUE
   80    CONTINUE
      END IF
```

```
      STOP
*
99999 FORMAT (1X,A,I2)
99998 FORMAT (1X,A,F7.2)
99997 FORMAT (1X,2A)
99996 FORMAT (1X,A,e13.4)
99995 FORMAT (1X,A,6(1X,A))
99994 FORMAT (1X,A,e13.4)
99993 FORMAT (1X,A,6(6X,A))
99992 FORMAT (1X,A,6(F9.4))
      END
```

9.2 Program Data

```
G02EEF Example Program Data
  20 6 'M' 'U'
   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    2
 'DAY' 'BOD' 'TKN' 'TS ' 'TVS' 'COD'
 2.0
```

9.3 Program Results

```
 G02EEF Example Program Results

 Step  1
 Added variable is TS
 Change in residual sum of squares =   0.4713E+00
 F Statistic =    7.38

 Variables in model: COD TS

 Residual sum of squares =    0.1085E+01
 Degrees of freedom = 17

 Free variables:        TKN      BOD      TVS
 Change in residual sums of squares for free variables:
                     0.1175   0.0600   0.2276

 Step  2
 No further variables added maximum F =   1.59
 Free variables:        TKN      BOD      TVS
 Change in residual sums of squares for free variables:
                     0.0979   0.0207   0.0217
```