# NAG 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, VNAME, ISX, MAXIP,    &
                   Y, WT, FIN, ADDVAR, NEWVAR, CHRSS, F, MODEL, NTERM,      &
                   RSS, IDF, IFR, FREE, EXSS, Q, LDQ, P, WK, IFAIL)

INTEGER            ISTEP, N, M, LDX, ISX(M), MAXIP, NTERM, IDF, IFR,        &
                   LDQ, IFAIL
REAL (KIND=nag_wp) X(LDX,M), Y(N), WT(*), FIN, CHRSS, F, RSS,               &
                   EXSS(MAXIP), Q(LDQ,MAXIP+2), P(MAXIP+1),                 &
                   WK(2*MAXIP)
LOGICAL            ADDVAR
CHARACTER(*)       VNAME(M), NEWVAR, MODEL(MAXIP), FREE(MAXIP)
CHARACTER(1)       MEAN, WEIGHT
```

## 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 (ii), otherwise stop.

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

G02EEF allows you 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 $Q^{\mathrm{T}}X_{\mathrm{free}}$, where $X_{\mathrm{free}}$ 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 Arguments

**Note:** after the initial call to G02EEF with ISTEP $= 0$ all arguments except FIN must not be changed by you between calls.

1: ISTEP – INTEGER *Input/Output*

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

ISTEP $= 0$
The process is initialized.

*Constraint*: ISTEP $\geq 0$.

*On exit*: is incremented by 1.

2: MEAN – CHARACTER(1) *Input*

*On entry*: indicates if a mean term is to be included.

MEAN = 'M'
A mean term, intercept, will be included in the model.

MEAN = 'Z'
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.

WEIGHT = 'U'
Least squares estimation is used.

WEIGHT = 'W'
Weighted least squares is used and weights must be supplied in array WT.

*Constraint*: WEIGHT = 'U' or 'W'.

4: N – INTEGER *Input*

*On entry*: $n$, the number of observations.

*Constraint*: N $\geq 2$.

5: M – INTEGER *Input*

*On entry*: $m$, the total number of independent variables in the dataset.

*Constraint*: M $\geq 1$.

6: X(LDX, M) – REAL (KIND=nag_wp) array *Input*

*On entry*: X$(i, j)$ must contain the $i$th observation for the $j$th independent variable, for $i = 1, 2, \ldots,$ N and $j = 1, 2, \ldots,$ 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: VNAME(M) – CHARACTER(*) array *Input*

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

9: ISX(M) – INTEGER array *Input*

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

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

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

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

*Constraint*: ISX($j$) $\geq 0$ 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 (KIND=nag_wp) array *Input*

*On entry*: the dependent variable.

12: WT($*$) – REAL (KIND=nag_wp) array *Input*

**Note**: the dimension of the array WT must be at least N if WEIGHT = 'W'.

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

If WT($i$) $= 0.0$, the $i$th observation is not included in the model, in which case the effective number of observations is the number of observations with nonzero weights.

If WEIGHT = 'U', WT is not referenced and the effective number of observations is N.

*Constraint*: if WEIGHT = 'W', WT($i$) $\geq 0.0$, for $i = 1, 2, \ldots, \mathrm{N}$.

13: FIN – REAL (KIND=nag_wp) *Input*

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

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

*Constraint*: FIN $\geq 0.0$.

14: ADDVAR – LOGICAL *Output*

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

ADDVAR = .TRUE.
A variable has been added to the model.

ADDVAR = .FALSE.
No variable had an $F$ value greater than $F_{\mathrm{c}}$ and none were added to the model.

15: NEWVAR – CHARACTER(*) *Output*

*On exit*: if ADDVAR = .TRUE., 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 VNAME.

16: CHRSS – REAL (KIND=nag_wp) *Output*

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

17: F – REAL (KIND=nag_wp) *Output*

*On exit*: if ADDVAR = .TRUE., 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, MODEL need not be set.

If ISTEP $\neq 0$, 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 VNAME.

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

19: NTERM – INTEGER *Input/Output*

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

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

*Constraint*: if ISTEP $\neq 0$, NTERM $> 0$.

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

20: RSS – REAL (KIND=nag_wp) *Input/Output*

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

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

*Constraint*: if ISTEP $\neq 0$, RSS $> 0.0$.

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

21: IDF – INTEGER *Input/Output*

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

If ISTEP $\neq 0$, 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, IFR need not be set.

If ISTEP $\neq 0$, 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 $= 0$, FREE need not be set.

If ISTEP $\neq 0$, 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 VNAME.

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

24: EXSS(MAXIP) – REAL (KIND=nag_wp) 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 (KIND=nag_wp) array *Input/Output*

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

If ISTEP $\neq 0$, 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 $c = Q^{\mathrm{T}}y$ (or $Q^{\mathrm{T}}W^{\frac{1}{2}}y$ where $W$ is the vector of weights if used);

the upper triangular part of columns 2 to $p+1$ contain the $R$ matrix;

the strictly lower triangular part of columns 2 to $p+1$ contain details of the $Q$ matrix;

the remaining $p+1$ to $p+$ IFR columns of contain $Q^{\mathrm{T}}X_{free}$ (or $Q^{\mathrm{T}}W^{\frac{1}{2}}X_{free}$),

where $p =$ NTERM, or $p =$ NTERM $+ 1$ if MEAN = 'M'.

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(MAXIP $+ 1$) – REAL (KIND=nag_wp) array *Input/Output*

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

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

*On exit*: the first $p$ elements of P contain details of the $QR$ decomposition, where $p =$ NTERM, or $p =$ NTERM $+ 1$ if MEAN = 'M'.

28: WK($2 \times$ MAXIP) – REAL (KIND=nag_wp) array *Workspace*

29: IFAIL – INTEGER *Input/Output*

*On entry*: IFAIL must be set to $0$, $-1$ or $1$. If you are unfamiliar with this argument you should refer to Section 3.4 in How to Use the NAG Library and its Documentation 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 argument, 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, 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 $\leq 0.0$,
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.

IFAIL $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

See Section 3.9 in How to Use the NAG Library and its Documentation for further information.

IFAIL $= -399$

Your licence key may have expired or may not have been installed correctly.

See Section 3.8 in How to Use the NAG Library and its Documentation for further information.

IFAIL $= -999$

Dynamic memory allocation failed.

See Section 3.7 in How to Use the NAG Library and its Documentation for further information.

## 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 Parallelism and Performance

G02EEF is threaded by NAG for parallel execution in multithreaded implementations of the NAG Library.

G02EEF makes calls to BLAS and/or LAPACK routines, which may be threaded within the vendor library used by this implementation. Consult the documentation for the vendor library for further information.

Please consult the X06 Chapter Introduction for information on how to control and interrogate the OpenMP environment used within this routine. Please also consult the Users' Note for your implementation for any additional implementation-specific information.

## 9 Further Comments

None.

## 10 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.

### 10.1 Program Text

```
    Program g02eefe

!     G02EEF Example Program Text

!     Mark 26 Release. NAG Copyright 2016.

!     .. Use Statements ..
      Use nag_library, Only: g02eef, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6, vnlen = 3
!     .. Local Scalars ..
      Real (Kind=nag_wp)               :: chrss, f, fin, rss
      Integer                          :: i, idf, ifail, ifr, istep, ldq, ldx, &
                                          lwt, m, maxip, n, nterm
      Logical                          :: addvar
      Character (1)                    :: mean, weight
      Character (3)                    :: newvar
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: exss(:), p(:), q(:,:), wk(:), wt(:), &
                                          x(:,:), y(:)
      Integer, Allocatable             :: isx(:)
      Character (vnlen), Allocatable   :: free(:), model(:), vname(:)
!     .. Intrinsic Procedures ..
      Intrinsic                        :: count
```

```
!     .. Executable Statements ..
      Write (nout,*) 'G02EEF Example Program Results'
      Write (nout,*)

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

!     Read in the problem size and various control parameters
      Read (nin,*) n, m, mean, weight, fin

      If (weight=='W' .Or. weight=='w') Then
        lwt = n
      Else
        lwt = 0
      End If
      ldx = n
      Allocate (x(ldx,m),y(n),wt(lwt),isx(m),vname(m))

!     Read in data
      If (lwt>0) Then
        Read (nin,*)(x(i,1:m),y(i),wt(i),i=1,n)
      Else
        Read (nin,*)(x(i,1:m),y(i),i=1,n)
      End If

!     Read in variable inclusion flags
      Read (nin,*) isx(1:m)

!     Read in first VNLEN characters of the variable names
      Read (nin,*) vname(1:m)

!     Calculate the maximum number of parameters in the model
      maxip = count(isx(1:m)>0)
      If (mean=='M' .Or. mean=='m') Then
        maxip = maxip + 1
      End If

      ldq = n
      Allocate (model(maxip),free(maxip),exss(maxip),q(ldq,maxip+2),           &
        p(maxip+1),wk(2*maxip))

!     Loop over each variable, attempting to add each in turn
      istep = 0
      Do i = 1, m
!       Fit the linear regression model
        ifail = 0
        Call g02eef(istep,mean,weight,n,m,x,ldx,vname,isx,maxip,y,wt,fin,      &
          addvar,newvar,chrss,f,model,nterm,rss,idf,ifr,free,exss,q,ldq,p,wk,  &
          ifail)

!       Display the results at each step
        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(1:ifr)
          Write (nout,*)                                                       &
            'Change in residual sums of squares for free variables:'
          Write (nout,99992) '                 ', exss(1:ifr)
          Go To 100
        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(1:nterm)
          Write (nout,*)
          Write (nout,99994) 'Residual sum of squares = ', rss
          Write (nout,99999) 'Degrees of freedom = ', idf
          Write (nout,*)
          If (ifr==0) Then
            Write (nout,*) 'No free variables remaining'
```

```
            Go To 100
          End If
          Write (nout,99993) 'Free variables:  ', free(1:ifr)
          Write (nout,*)                                                &
            'Change in residual sums of squares for free variables:'
          Write (nout,99992) '                 ', exss(1:ifr)
          Write (nout,*)
        End If
      End Do

100   Continue

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 Program g02eefe
```

## 10.2 Program Data

```
G02EEF Example Program Data
  20 6 'M' 'U' 2.0                                :: N,M,MEAN,WEIGHT,FIN
   0.0 1125.0 232.0 7160.0 85.9 8905.0  1.5563
   7.0  920.0 268.0 8804.0 86.5 7388.0  0.8976
  15.0  835.0 271.0 8108.0 85.2 5348.0  0.7482
  22.0 1000.0 237.0 6370.0 83.8 8056.0  0.7160
  29.0 1150.0 192.0 6441.0 82.1 6960.0  0.3010
  37.0  990.0 202.0 5154.0 79.2 5690.0  0.3617
  44.0  840.0 184.0 5896.0 81.2 6932.0  0.1139
  58.0  650.0 200.0 5336.0 80.6 5400.0  0.1139
  65.0  640.0 180.0 5041.0 78.4 3177.0 -0.2218
  72.0  583.0 165.0 5012.0 79.3 4461.0 -0.1549
  80.0  570.0 151.0 4825.0 78.7 3901.0  0.0000
  86.0  570.0 171.0 4391.0 78.0 5002.0  0.0000
  93.0  510.0 243.0 4320.0 72.3 4665.0 -0.0969
 100.0  555.0 147.0 3709.0 74.9 4642.0 -0.2218
 107.0  460.0 286.0 3969.0 74.4 4840.0 -0.3979
 122.0  275.0 198.0 3558.0 72.5 4479.0 -0.1549
 129.0  510.0 196.0 4361.0 57.7 4200.0 -0.2218
 151.0  165.0 210.0 3301.0 71.8 3410.0 -0.3979
 171.0  244.0 327.0 2964.0 72.5 3360.0 -0.5229
 220.0   79.0 334.0 2777.0 71.9 2599.0 -0.0458   :: End of X,Y
  0     1    1      1     1    2                 :: ISX
 'DAY' 'BOD' 'TKN' 'TS ' 'TVS' 'COD'             :: VNAME
```

## 10.3 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
```