# NAG Library Routine Document

# G02EFF

**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

G02EFF calculates a full stepwise selection from $p$ variables by using Clarke's sweep algorithm on the correlation matrix of a design and data matrix, $Z$. The (weighted) variance-covariance, (weighted) means and sum of weights of $Z$ must be supplied.

## 2 Specification

```
SUBROUTINE G02EFF (M, N, WMEAN, C, SW, ISX, FIN, FOUT, TAU, B, SE, RSQ,    &
                   RMS, DF, MONLEV, MONFUN, IUSER, RUSER, IFAIL)

INTEGER            M, N, ISX(M), DF, MONLEV, IUSER(*), IFAIL
REAL (KIND=nag_wp) WMEAN(M+1), C((M+1)*(M+2)/2), SW, FIN, FOUT, TAU,       &
                   B(M+1), SE(M+1), RSQ, RMS, RUSER(*)
EXTERNAL           MONFUN
```

## 3 Description

The general multiple linear regression model is defined by

$$y = \beta_0 + X\beta + \epsilon,$$

where

$y$ is a vector of $n$ observations on the dependent variable,

$\beta_0$ is an intercept coefficient,

$X$ is an $n$ by $p$ matrix of $p$ explanatory variables,

$\beta$ is a vector of $p$ unknown coefficients, and

$\epsilon$ is a vector of length $n$ of unknown, Normally distributed, random errors.

G02EFF employs a full stepwise regression to select a subset of explanatory variables from the $p$ available variables (the intercept is included in the model) and computes regression coefficients and their standard errors, and various other statistical quantities, by minimizing the sum of squares of residuals. The method applies repeatedly a forward selection step followed by a backward elimination step and halts when neither step updates the current model.

The criterion used to update a current model is the variance ratio of residual sum of squares. Let $s_1$ and $s_2$ be the residual sum of squares of the current model and this model after undergoing a single update, with degrees of freedom $q_1$ and $q_2$, respectively. Then the condition:

$$\frac{(s_2 - s_1)/(q_2 - q_1)}{s_1/q_1} > f_1,$$

must be satisfied if a variable $k$ will be considered for entry to the current model, and the condition:

$$\frac{(s_1 - s_2)/(q_1 - q_2)}{s_1/q_1} < f_2,$$

must be satisfied if a variable $k$ will be considered for removal from the current model, where $f_1$ and $f_2$ are user-supplied values and $f_2 \leq f_1$.

In the entry step the entry statistic is computed for each variable not in the current model. If no variable is associated with a test value that exceeds $f_1$ then this step is terminated; otherwise the variable associated with the largest value for the entry statistic is entered into the model.

In the removal step the removal statistic is computed for each variable in the current model. If no variable is associated with a test value less than $f_2$ then this step is terminated; otherwise the variable associated with the smallest value for the removal statistic is removed from the model.

The data values $X$ and $y$ are not provided as input to the routine. Instead, summary statistics of the design and data matrix $Z = (X \mid y)$ are required.

Explanatory variables are entered into and removed from the current model by using sweep operations on the correlation matrix $R$ of $Z$, given by:

$$R = \left(\begin{array}{ccc|c} 1 & \dots & r_{1p} & r_{1y} \\ \vdots & \ddots & \vdots & \vdots \\ r_{p1} & \dots & 1 & r_{py} \\ \hline r_{y1} & \dots & r_{yp} & 1 \end{array}\right),$$

where $r_{ij}$ is the correlation between the explanatory variables $i$ and $j$, for $i = 1, 2, \dots, p$ and $j = 1, 2, \dots, p$, and $r_{yi}$ (and $r_{iy}$) is the correlation between the response variable $y$ and the $i$th explanatory variable, for $i = 1, 2, \dots, p$.

A sweep operation on the $k$th row and column ($k \leq p$) of $R$ replaces:

$r_{kk}$ by $-1/r_{kk}$;
$r_{ik}$ by $r_{ik}/|r_{kk}|$, $\quad i = 1, 2, \dots, p+1$ $(i \neq k)$;
$r_{kj}$ by $r_{kj}/|r_{kk}|$, $\quad j = 1, 2, \dots, p+1$ $(j \neq k)$;
$r_{ij}$ by $r_{ij} - r_{ik}r_{kj}/|r_{kk}|$, $\quad i = 1, 2, \dots, p+1$ $(i \neq k)$; $j = 1, 2, \dots, p+1$ $(j \neq k)$.

The $k$th explanatory variable is eligible for entry into the current model if it satisfies the collinearity tests: $r_{kk} > \tau$ and

$$\left(r_{ii} - \frac{r_{ik}r_{ki}}{r_{kk}}\right)\tau \leq 1,$$

for a user-supplied value ($> 0$) of $\tau$ and where the index $i$ runs over explanatory variables in the current model. The sweep operation is its own inverse, therefore pivoting on an explanatory variable $k$ in the current model has the effect of removing it from the model.

Once the stepwise model selection procedure is finished, the routine calculates:

(a) the least squares estimate for the $i$th explanatory variable included in the fitted model;

(b) standard error estimates for each coefficient in the final model;

(c) the square root of the mean square of residuals and its degrees of freedom;

(d) the multiple correlation coefficient.

The routine makes use of the symmetry of the sweep operations and correlation matrix which reduces by almost one half the storage and computation required by the sweep algorithm, see Clarke (1981) for details.

## 4 References

Clarke M R B (1981) Algorithm AS 178: the Gauss–Jordan sweep operator with detection of collinearity *Appl. Statist.* **31** 166–169

Dempster A P (1969) *Elements of Continuous Multivariate Analysis* Addison–Wesley

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

## 5 Parameters

1: M – INTEGER *Input*

*On entry*: the number of explanatory variables available in the design matrix, $Z$.

*Constraint*: $\mathrm{M} > 1$.

2: N – INTEGER *Input*

*On entry*: the number of observations used in the calculations.

*Constraint*: $\mathrm{N} > 1$.

3: WMEAN$(\mathrm{M}+1)$ – REAL (KIND=nag_wp) array *Input*

*On entry*: the mean of the design matrix, $Z$.

4: C$((\mathrm{M}+1) \times (\mathrm{M}+2)/2)$ – REAL (KIND=nag_wp) array *Input*

*On entry*: the upper-triangular variance-covariance matrix packed by column for the design matrix, $Z$. Because the routine computes the correlation matrix $R$ from C, the variance-covariance matrix need only be supplied up to a scaling factor.

5: SW – REAL (KIND=nag_wp) *Input*

*On entry*: if weights were used to calculate C then SW is the sum of positive weight values; otherwise SW is the number of observations used to calculate C.

*Constraint*: $\mathrm{SW} > 1.0$.

6: ISX(M) – INTEGER array *Input/Output*

*On entry*: the value of ISX$(j)$ determines the set of variables used to perform full stepwise model selection, for $j = 1, 2, \ldots, \mathrm{M}$.

ISX$(j) = -1$
To exclude the variable corresponding to the $j$th column of $X$ from the final model.

ISX$(j) = 1$
To consider the variable corresponding to the $j$th column of $X$ for selection in the final model.

ISX$(j) = 2$
To force the inclusion of the variable corresponding to the $j$th column of $X$ in the final model.

*Constraint*: ISX$(j) = -1, 1$ or 2, for $j = 1, 2, \ldots, \mathrm{M}$.

*On exit*: the value of ISX$(j)$ indicates the status of the $j$th explanatory variable in the model.

ISX$(j) = -1$
Forced exclusion.

ISX$(j) = 0$
Excluded.

ISX$(j) = 1$
Selected.

ISX$(j) = 2$
Forced selection.

7: FIN – REAL (KIND=nag_wp) *Input*

*On entry*: the value of the variance ratio which an explanatory variable must exceed to be included in a model.

*Suggested value*: FIN $= 4.0$

*Constraint*: FIN $> 0.0$.

8: FOUT – REAL (KIND=nag_wp) *Input*

*On entry*: the explanatory variable in a model with the lowest variance ratio value is removed from the model if its value is less than FOUT. FOUT is usually set equal to the value of FIN; a value less than FIN is occasionally preferred.

*Suggested value*: FOUT $=$ FIN

*Constraint*: $0.0 \leq$ FOUT $\leq$ FIN.

9: TAU – REAL (KIND=nag_wp) *Input*

*On entry*: the tolerance, $\tau$, for detecting collinearities between variables when adding or removing an explanatory variable from a model. Explanatory variables deemed to be collinear are excluded from the final model.

*Suggested value*: TAU $= 1.0 \times 10^{-6}$

*Constraint*: TAU $> 0.0$.

10: B(M + 1) – REAL (KIND=nag_wp) array *Output*

*On exit*: B(1) contains the estimate for the intercept term in the fitted model. If ISX($j$) $\neq 0$ then B($j+1$) contains the estimate for the $j$th explanatory variable in the fitted model; otherwise B($j+1$) $= 0$.

11: SE(M + 1) – REAL (KIND=nag_wp) array *Output*

*On exit*: SE($j$) contains the standard error for the estimate of B($j$), for $j = 1, 2, \ldots, \mathrm{M}+1$.

12: RSQ – REAL (KIND=nag_wp) *Output*

*On exit*: the $R^2$-statistic for the fitted regression model.

13: RMS – REAL (KIND=nag_wp) *Output*

*On exit*: the mean square of residuals for the fitted regression model.

14: DF – INTEGER *Output*

*On exit*: the number of degrees of freedom for the sum of squares of residuals.

15: MONLEV – INTEGER *Input*

*On entry*: if a subroutine is provided by you to monitor the model selection process, set MONLEV to 1; otherwise set MONLEV to 0.

*Constraint*: MONLEV $= 0$ or 1.

16: MONFUN – SUBROUTINE, supplied by the NAG Library or the user. *External Procedure*

You may define your own function or specify the NAG defined default function G02EFH.

If MONLEV $= 0$, MONFUN is not referenced; otherwise its specification is:

The specification of MONFUN is:

```
SUBROUTINE MONFUN (FLAG, VAR, VAL, IUSER, RUSER)
INTEGER            VAR, IUSER(*)
REAL (KIND=nag_wp) VAL, RUSER(*)
CHARACTER(1)       FLAG
```

1: FLAG – CHARACTER(1) *Input*

*On entry*: the value of FLAG indicates the stage of the stepwise selection of explanatory variables.

FLAG = 'A'
Variable VAR was added to the current model.

FLAG = 'B'
Beginning the backward elimination step.

FLAG = 'C'
Variable VAR failed the collinearity test and is excluded from the model.

FLAG = 'D'
Variable VAR was dropped from the current model.

FLAG = 'F'
Beginning the forward selection step

FLAG = 'K'
Backward elimination did not remove any variables from the current model.

FLAG = 'S'
Starting stepwise selection procedure.

FLAG = 'V'
The variance ratio for variable VAR takes the value VAL.

FLAG = 'X'
Finished stepwise selection procedure.

2: VAR – INTEGER *Input*

*On entry*: the index of the explanatory variable in the design matrix $Z$ to which FLAG pertains.

3: VAL – REAL (KIND=nag_wp) *Input*

*On entry*: if FLAG = 'V', VAL is the variance ratio value for the coefficient associated with explanatory variable index VAR.

4: IUSER(∗) – INTEGER array *User Workspace*
5: RUSER(∗) – REAL (KIND=nag_wp) array *User Workspace*

MONFUN is called with the parameters IUSER and RUSER as supplied to G02EFF. You are free to use the arrays IUSER and RUSER to supply information to MONFUN as an alternative to using COMMON global variables.

MONFUN must either be a module subprogram USEd by, or declared as EXTERNAL in, the (sub)program from which G02EFF is called. Parameters denoted as *Input* must **not** be changed by this procedure.

17: IUSER(∗) – INTEGER array *User Workspace*
18: RUSER(∗) – REAL (KIND=nag_wp) array *User Workspace*

IUSER and RUSER are not used by G02EFF, but are passed directly to MONFUN and may be used to pass information to this routine as an alternative to using COMMON global variables.

19: 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, FIN $= \langle value \rangle$.
Constraint: FIN $> 0.0$.

On entry, FOUT $= \langle value \rangle$; FIN $= \langle value \rangle$.
Constraint: $0.0 \leq$ FOUT $\leq$ FIN.

On entry, M $= \langle value \rangle$.
Constraint: M $> 1$.

On entry, MONLEV $= \langle value \rangle$.
Constraint: MONLEV $= 0$ or 1.

On entry, N $= \langle value \rangle$.
Constraint: N $> 1$.

On entry, SW $= \langle value \rangle$.
Constraint: SW $> 1.0$.

On entry, TAU $= \langle value \rangle$.
Constraint: TAU $> 0.0$.

IFAIL $= 2$

No free variables from which to select.
At least one element of ISX should be set to 1.

On entry, invalid value for ISX($\langle value \rangle$) $= \langle value \rangle$.

On entry at least one diagonal element of C $\leq 0.0$.

IFAIL $= 3$

The design and data matrix $Z$ is not positive definite, results may be inaccurate. All output is returned as documented.

IFAIL $= 4$

All variables are collinear, no model to select.

IFAIL $= -99$

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

See Section 3.8 in the Essential Introduction for further information.

IFAIL $= -399$

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

See Section 3.7 in the Essential Introduction for further information.

IFAIL $= -999$

Dynamic memory allocation failed.

See Section 3.6 in the Essential Introduction for further information.

## 7 Accuracy

G02EFF returns a warning if the design and data matrix is not positive definite.

## 8 Parallelism and Performance

Not applicable.

## 9 Further Comments

Although the condition for removing or adding a variable to the current model is based on a ratio of variances, these values should not be interpreted as $F$-statistics with the usual interpretation of significance unless the probability levels are adjusted to account for correlations between variables under consideration and the number of possible updates (see, e.g., Draper and Smith (1985)).

G02EFF allocates internally $\mathcal{O}(4 \times \mathrm{M} + (\mathrm{M}+1) \times (\mathrm{M}+2)/2 + 2)$ of real storage.

## 10 Example

This example calculates a full stepwise model selection for the Hald data described in Dempster (1969). Means, the upper-triangular variance-covariance matrix and the sum of weights are calculated by G02BUF. The NAG defined default monitor function G02EFH is used to print information at each step of the model selection process.

### 10.1 Program Text

```
    Program g02effe

!     G02EFF Example Program Text

!     Mark 25 Release. NAG Copyright 2014.

!     .. Use Statements ..
      Use nag_library, Only: g02buf, g02eff, g02efh, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=nag_wp)               :: fin, fout, rms, rsq, sw, tau
      Integer                          :: df, i, ifail, ldz, liuser, lruser,   &
                                          m, m1, monlev, n
      Character (1)                    :: mean, weight
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: b(:), c(:), ruser(:), se(:),         &
                                          wmean(:), z(:,:)
      Real (Kind=nag_wp)               :: wt(1)
      Integer, Allocatable             :: isx(:), iuser(:)
!     .. Executable Statements ..
      Write (nout,*) 'G02EFF Example Program Results'
      Write (nout,*)
      Flush (nout)

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

!     Read in the problem size and various control parameters
      Read (nin,*) n, m, fin, fout, tau, monlev
```

```
!     Not using the user supplied arrays RUSER and IUSER
      liuser = 0
      lruser = 0

      m1 = m + 1
      ldz = n
      Allocate (wmean(m1),c(m1*(m+2)/2),isx(m),b(m1),se(m1),iuser(liuser), &
        ruser(lruser),z(ldz,m1))

!     Read in augmented design matrix Z = (X | Y)
      Read (nin,*)(z(i,1:m1),i=1,n)

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

!     No weights in this example
      weight = 'U'

!     Compute upper-triangular sums of squares and cross-products of devations
!     from the mean for the augmented matrix
      mean = 'M'
      ifail = 0
      Call g02buf(mean,weight,n,m1,z,ldz,wt,sw,wmean,c,ifail)

!     Perform stepwise selection of variables.
      ifail = 0
      Call g02eff(m,n,wmean,c,sw,isx,fin,fout,tau,b,se,rsq,rms,df,monlev, &
        g02efh,iuser,ruser,ifail)

!     Display results
      Write (nout,*)
      Write (nout,99999) 'Fitted Model Summary'
      Write (nout,99999) 'Term              Estimate   Standard Error'
      Write (nout,99998) 'Intercept:', b(1), se(1)
      Do i = 1, m
        If (isx(i)==1 .Or. isx(i)==2) Then
          Write (nout,99997) 'Variable:', i, b(i+1), se(i+1)
        End If
      End Do
      Write (nout,*)
      Write (nout,99996) 'RMS:', rms

99999 Format (1X,A)
99998 Format (1X,A,4X,1P,E12.3,5X,E12.3)
99997 Format (1X,A,1X,I3,1X,1P,E12.3,5X,E12.3)
99996 Format (1X,A,1X,1P,E12.3)
    End Program g02effe
```

## 10.2 Program Data

```
G02EFF Example Program Data
13 4 4.0 2.0 1.0D-6 1        : N,M,FIN,FOUT,TAU,MONLEV
 7.0 26.0  6.0 60.0  78.5
 1.0 29.0 15.0 52.0  74.3
11.0 56.0  8.0 20.0 104.3
11.0 31.0  8.0 47.0  87.6
 7.0 52.0  6.0 33.0  95.9
11.0 55.0  9.0 22.0 109.2
 3.0 71.0 17.0  6.0 102.7
 1.0 31.0 22.0 44.0  72.5
 2.0 54.0 18.0 22.0  93.1
21.0 47.0  4.0 26.0 115.9
 1.0 40.0 23.0 34.0  83.8
11.0 66.0  9.0 12.0 113.3
10.0 68.0  8.0 12.0 109.4    : End of augmented design matrix Z = (X | Y)
1 1 1 1                      : ISX
```

## 10.3 Program Results

```
G02EFF Example Program Results

Starting Stepwise Selection

Forward Selection
Variable    1 Variance ratio =    1.260E+01
Variable    2 Variance ratio =    2.196E+01
Variable    3 Variance ratio =    4.403E+00
Variable    4 Variance ratio =    2.280E+01

Adding variable    4 to model

Backward Selection
Variable    4 Variance ratio =    2.280E+01

Keeping all current variables

Forward Selection
Variable    1 Variance ratio =    1.082E+02
Variable    2 Variance ratio =    1.725E-01
Variable    3 Variance ratio =    4.029E+01

Adding variable    1 to model

Backward Selection
Variable    1 Variance ratio =    1.082E+02
Variable    4 Variance ratio =    1.593E+02

Keeping all current variables

Forward Selection
Variable    2 Variance ratio =    5.026E+00
Variable    3 Variance ratio =    4.236E+00

Adding variable    2 to model

Backward Selection
Variable    1 Variance ratio =    1.540E+02
Variable    2 Variance ratio =    5.026E+00
Variable    4 Variance ratio =    1.863E+00

Dropping variable    4 from model

Forward Selection
Variable    3 Variance ratio =    1.832E+00
Variable    4 Variance ratio =    1.863E+00

Finished Stepwise Selection

Fitted Model Summary
Term              Estimate   Standard Error
Intercept:       5.258E+01        2.294E+00
Variable:   1    1.468E+00        1.213E-01
Variable:   2    6.623E-01        4.585E-02

RMS:    5.790E+00
```