NAG Fortran 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,
    1                   RMS, DF, MONLEV, MONFUN, IFAIL)
     INTEGER            M, N, ISX(M), DF, MONLEV, IFAIL
     double precision   WMEAN(M+1), C((M+1)*(M+2)/2), SW, FIN, FOUT, TAU,
    1                   B(M+1), SE(M+1), RSQ, RMS
     EXTERNAL           MONFUN
```

3 Description

The general multiple linear regression model is defined by

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

where

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

β_0 is an intercept coefficient,

X is a n by p matrix of p explanatory variables,

β is a vector of p unknown coefficients, and

ε is a vector of length n of unknown, normally distributed, random errors.

The routine 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 $i, j = 1, 2, \ldots, p$, and r_{yi} (and r_{iy}) is the correlation between the response variable y and the ith explanatory variable, for $i = 1, 2, \ldots, p$.

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

r_{kk} by $-1/r_{kk}$;

r_{ik} by $r_{ik}/|r_{kk}|$, for $i = 1, 2, \ldots, p+1$ $(i \neq k)$;

r_{kj} by $r_{kj}/|r_{kk}|$, for $j = 1, 2, \ldots, p+1$ $(j \neq k)$;

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

The kth 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 τ 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 ith 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 *Applied Statistics* **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: $\mathrm{WMEAN}(\mathrm{M}+1)$ – ***double precision*** array *Input*

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

4: $\mathrm{C}((\mathrm{M}+1) \times (\mathrm{M}+2)/2)$ – ***double precision*** array *Input*

On entry: the upper-triangular variance-covariance matrix packed by column for the design matrix, Z. The routine computes the correlation matrix R from C.

5: SW – ***double precision*** *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: $\mathrm{ISX}(\mathrm{M})$ – INTEGER array *Input/Output*

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

> $\mathrm{ISX}(j) = -1$, to exclude the variable corresponding to the jth column of X from the final model;
>
> $\mathrm{ISX}(j) = 1$, to consider the variable corresponding to the jth column of X for selection in the final model;
>
> $\mathrm{ISX}(j) = 2$, to force the inclusion of the variable corresponding to the jth column of X in the final model.

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

On exit: the value of $\mathrm{ISX}(j)$ indicates the status of the jth explanatory variable in the model:

> $\mathrm{ISX}(j) = -1$, forced exclusion;
>
> $\mathrm{ISX}(j) = 0$, excluded;
>
> $\mathrm{ISX}(j) = 1$, selected;
>
> $\mathrm{ISX}(j) = 2$, forced selection.

7: FIN – ***double precision*** *Input*

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

Constraint: $\mathrm{FIN} > 0.0$.

Suggested value: $\mathrm{FIN} = 4.0$

8: FOUT – ***double precision*** *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.

Constraint: $0.0 \leq \text{FOUT} \leq \text{FIN}$.

Suggested value: $\text{FOUT} = \text{FIN}$

9: TAU – ***double precision*** *Input*

On entry: the tolerance, τ, 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.

Constraint: $\text{TAU} > 0.0$.

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

10: $\text{B}(\text{M}+1)$ – ***double precision*** array *Output*

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

11: $\text{SE}(\text{M}+1)$ – ***double precision*** array *Output*

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

12: RSQ – ***double precision*** *Output*

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

13: RMS – ***double precision*** *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 the user 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 user. *External Procedure*

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

```
SUBROUTINE MONFUN (FLAG, VAR, VAL)
INTEGER             VAR
double precision    VAL
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 – ***double precision*** *Input*

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

MONFUN must be declared as EXTERNAL in the (sub)program from which G02EFF is called. Parameters denoted as *Input* must **not** be changed by this procedure.

17: 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, $M \leq 1$,
or $N \leq 1$,
or $SW \leq 1.0$,
or $FIN \leq 0.0$,
or $FOUT \leq 0.0$,
or $FOUT > FIN$,
or $TAU \leq 0.0$.

IFAIL = 2

On entry, at least one element of ISX was set incorrectly,
or there are no explanatory variables to select from $\mathrm{ISX}(i) \neq 1$, for $i = 1, 2, \ldots, \mathrm{M}$,
or invalid value for MONLEV.

IFAIL = 3

Warning: the design and data matrix Z is not positive-definite, results may be inaccurate.

IFAIL = 4

All variables are collinear, there is no model to select.

7 Accuracy

The routine returns a warning if the design and data matrix is not positive-definite.

8 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).

The routine allocates internally $\mathcal{O}(4 \times \mathrm{M} + (\mathrm{M} + 1) \times (\mathrm{M} + 2)/2 + 2)$ of ***double precision*** storage.

9 Example

A program that 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. An example monitor function is supplied to print information at each step of the model selection process.

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.

```
*     G02EFF Example Program Text
*     Mark 21 Release. NAG Copyright 2004.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NMAX, MMAX
      PARAMETER        (NMAX=20,MMAX=10)
*     .. Local Scalars ..
      DOUBLE PRECISION FIN, FOUT, RMS, RSQ, SW, TAU
      INTEGER          DF, I, IFAIL, J, M, MONLEV, N
*     .. Local Arrays ..
      DOUBLE PRECISION B(MMAX+1), C((MMAX+2)*(MMAX+1)/2), RUSER(1),
     +                 SE(MMAX+1), WMEAN(MMAX+1), WT(1), X(NMAX,MMAX+1)
      INTEGER          ISX(MMAX), IUSER(1)
*     .. External Subroutines ..
      EXTERNAL         G02BUF, G02EFF, MONFUN
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G02EFF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) N, M, FIN, FOUT, TAU, MONLEV
      IF (N.LE.NMAX .AND. M.LE.(MMAX)) THEN
         READ (NIN,*) ((X(I,J),J=1,M+1),I=1,N)
         READ (NIN,*) (ISX(J),J=1,M)
*
*        Compute upper-triangular correlation matrix
         IFAIL = -1
         CALL G02BUF('M','U',N,M+1,X,NMAX,WT,SW,WMEAN,C,IFAIL)
*
*        Perform stepwise selection of variables
         IFAIL = -1
         CALL G02EFF(M,N,WMEAN,C,SW,ISX,FIN,FOUT,TAU,B,SE,RSQ,RMS,DF,
     +               MONLEV,MONFUN,IUSER,RUSER,IFAIL)
```

```
*
*        Display summary information for fitted model
         WRITE (NOUT,*)
         WRITE (NOUT,99999) 'Fitted Model Summary'
         WRITE (NOUT,99999)
     +     'Term              Estimate   Standard Error'
         WRITE (NOUT,99998) 'Intercept:', B(1), SE(1)
         DO 20 I = 1, M
            IF (ISX(I).EQ.1 .OR. ISX(I).EQ.2) THEN
               WRITE (NOUT,99997) 'Variable:', I, B(I+1), SE(I+1)
            END IF
   20    CONTINUE
         WRITE (NOUT,*)
         WRITE (NOUT,99996) 'RMS:', RMS
      END IF
*
      STOP
*
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
*
*     Example monitor function for use by G02EFF
*
      SUBROUTINE MONFUN(FLAG,VAR,VAL,IUSER,RUSER)
*     .. Parameters ..
      INTEGER           NOUT
      PARAMETER         (NOUT=6)
*     .. Scalar Arguments ..
      DOUBLE PRECISION  VAL
      INTEGER           VAR
      CHARACTER         FLAG
*     .. Array Arguments ..
      DOUBLE PRECISION  RUSER(*)
      INTEGER           IUSER(*)
*     .. Executable Statements ..
      CONTINUE
      IF (FLAG.EQ.'C') THEN
         WRITE (NOUT,99999) 'Variable', VAR, 'aliased'
      ELSE IF (FLAG.EQ.'S') THEN
         WRITE (NOUT,99998) 'Starting Stepwise Selection'
      ELSE IF (FLAG.EQ.'F') THEN
         WRITE (NOUT,99997) 'Forward Selection'
      ELSE IF (FLAG.EQ.'V') THEN
         WRITE (NOUT,99996) 'Variable', VAR, 'Variance ratio =', VAL
      ELSE IF (FLAG.EQ.'A') THEN
         WRITE (NOUT,99995) 'Adding variable', VAR, 'to model'
      ELSE IF (FLAG.EQ.'B') THEN
         WRITE (NOUT,99994) 'Backward Selection'
      ELSE IF (FLAG.EQ.'D') THEN
         WRITE (NOUT,99993) 'Dropping variable', VAR, 'from model'
      ELSE IF (FLAG.EQ.'K') THEN
         WRITE (NOUT,99992) 'Keeping all current variables'
      ELSE IF (FLAG.EQ.'X') THEN
         WRITE (NOUT,99991) 'Finished Stepwise Selection'
      END IF
      RETURN
*
99999 FORMAT (1X,A,1X,I4,1X,A)
99998 FORMAT (/1X,A)
99997 FORMAT (/1X,A)
99996 FORMAT (1X,A,1X,I4,1X,A,1X,1P,E12.3)
99995 FORMAT (/1X,A,1X,I4,1X,A)
99994 FORMAT (/1X,A)
99993 FORMAT (/1X,A,1X,I4,1X,A)
99992 FORMAT (/1X,A)
99991 FORMAT (/1X,A)
      END
```

9.2 Program Data

```
G02EFF Example Program Data
13 4 4 2 1.0D-6 1    :  N,M,FIN,FOUT,TAU,MONLEV
 7 26  6 60  78.5
 1 29 15 52  74.3
11 56  8 20 104.3
11 31  8 47  87.6
 7 52  6 33  95.9
11 55  9 22 109.2
 3 71 17  6 102.7
 1 31 22 44  72.5
 2 54 18 22  93.1
21 47  4 26 115.9
 1 40 23 34  83.8
11 66  9 12 113.3
10 68  8 12 109.4    : End of X array of size N by M+1
1 1 1 1              : Array ISX
```

9.3 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
```