# NAG Toolbox for MATLAB

# g02ef

## 1 Purpose

g02ef 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 Syntax

```
[isx, b, se, rsq, rms, df, user, ifail] = g02ef(n, wmean, c, sw, isx,
monlev, monfun, 'm', m, 'fin', fin, 'fout', fout, 'tau', tau, 'user',
user)
```

## 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.

g02ef 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 function. 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 $i,j = 1,2,\dots,p$, 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 function 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 function 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

### 5.1 Compulsory Input Parameters

1: **n – int32 scalar**

The number of observations used in the calculations.

*Constraint*: $\mathbf{n} > 1$.

2: **wmean(m + 1) – double array**

The mean of the design matrix, $Z$.

3: **c**$((\mathbf{m}+\mathbf{1})\times(\mathbf{m}+\mathbf{2})/\mathbf{2})$ **– double array**

The upper-triangular variance-covariance matrix packed by column for the design matrix, $Z$. The function computes the correlation matrix $R$ from $C$.

4: **sw – double scalar**

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*: **sw** $> 1.0$.

5: **isx**$(\mathbf{m})$ **– int32 array**

The value of **isx**$(j)$ determines the set of variables used to perform full stepwise model selection, for $j = 1, 2, \ldots, \mathbf{m}$.

**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 $j = 1, 2, \ldots, \mathbf{m}$.

6: **monlev – int32 scalar**

If a subfunction is provided by you to monitor the model selection process, set **monlev** to 1; otherwise set **monlev** to 0.

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

7: **monfun – string containing name of m-file**

If **monlev** $= 0$, **monfun** is not referenced; otherwise its specification is:

```
[user] = monfun(flag, var, val, user)
```

**Input Parameters**

1: **flag – string**

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 – int32 scalar**

The index of the explanatory variable in the design matrix $Z$ to which **flag** pertains.

3: **val – double scalar**

If **flag** = 'V', **val** is the variance ratio value for the coefficient associated with explanatory variable index **var**.

4: **user – Any MATLAB object**

**monfun** is called from g02ef with **user** as supplied to g02ef

**Output Parameters**

1: **user – Any MATLAB object**

**monfun** is called from g02ef with **user** as supplied to g02ef

## 5.2 Optional Input Parameters

1: **m – int32 scalar**

*Default*: The dimension of the array **isx**.

the number of explanatory variables available in the design matrix, $Z$.

*Constraint*: $\mathbf{m} > 1$.

2: **fin – double scalar**

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

*Suggested value*: $\mathbf{fin} = 4.0$

*Default*: 4.0

*Constraint*: $\mathbf{fin} > 0.0$.

3: **fout – double scalar**

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*: $\mathbf{fout} = \mathbf{fin}$

*Default*: **fin**

*Constraint*: $0.0 \leq \mathbf{fout} \leq \mathbf{fin}$.

4: **tau – double scalar**

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*: $\mathbf{tau} = 1.0 \times 10^{-6}$

*Default*: 0.000001

*Constraint*: $\mathbf{tau} > 0.0$.

5: **user – Any MATLAB object**

**user** is not used by g02ef, but is passed to **monfun**. Note that for large objects it may be more efficient to use a global variable which is accessible from the m-files than to use **user**.

## 5.3 Input Parameters Omitted from the MATLAB Interface

None.

## 5.4 Output Parameters

1: **isx(m) – int32 array**

The value of $\mathbf{isx}(j)$ indicates the status of the $j$th explanatory variable in the model.

$\mathbf{isx}(j) = -1$

Forced exclusion.

$\mathbf{isx}(j) = 0$

Excluded.

$\mathbf{isx}(j) = 1$

Selected.

$\mathbf{isx}(j) = 2$

Forced selection.

2: **b(m + 1) – double array**

$\mathbf{b}(1)$ contains the estimate for the intercept term in the fitted model. If $\mathbf{isx}(j) \neq 0$ then $\mathbf{b}(j+1)$ contains the estimate for the $j$th explanatory variable in the fitted model; otherwise $\mathbf{b}(j+1) = 0$.

3: **se(m + 1) – double array**

$\mathbf{se}(j)$ contains the standard error for the estimate of $\mathbf{b}(j)$, for $j = 1, 2, \ldots, \mathbf{m} + 1$.

4: **rsq – double scalar**

The $R^2$-statistic for the fitted regression model.

5: **rms – double scalar**

The mean square of residuals for the fitted regression model.

6: **df – int32 scalar**

The number of degrees of freedom for the sum of squares of residuals.

7: **user – Any MATLAB object**

**user** is not used by g02ef, but is passed to **monfun**. Note that for large objects it may be more efficient to use a global variable which is accessible from the m-files than to use **user**.

8: **ifail – int32 scalar**

0 unless the function detects an error (see Section 6).

# 6 Error Indicators and Warnings

Errors or warnings detected by the function:

**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 **isx**$(i) \neq 1$, for $i = 1, 2, \ldots,$ **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

g02ef 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).

g02ef allocates internally $\mathcal{O}(4 \times \mathbf{m} + (\mathbf{m} + 1) \times (\mathbf{m} + 2)/2 + 2)$ of double storage.

# 9 Example

g02ef_monfun.m

```
function [user] = monfun(flag, var, val, user)

  switch flag
    case 'C'
      fprintf('\nVariable %d aliased\n', var);
    case 'S'
      fprintf('\nStarting Stepwise Selection\n');
    case 'F'
      fprintf('\nForward Selection\n');
    case 'V'
      fprintf('Variable %d  Variance ratio = %12.3f\n', var, val);
    case 'A'
```

```
      fprintf('\nAdding variable %d to model\n', var);
    case 'B'
      fprintf('\nBackward Selection\n');
    case 'D'
      fprintf('\nDropping variable %d from model\n', var);
    case 'K'
      fprintf('\nKeeping all current variables\n');
    case 'X'
      fprintf('\nFinished Stepwise Selection\n');
  end;
```

```
n = int32(13);
wmean = [7.461538461538462;
     48.15384615384615;
     11.76923076923077;
     30;
     95.42307692307693];
c = [415.2307692307692;
     251.0769230769231;
     2905.692307692308;
     -372.6153846153845;
     -166.5384615384615;
     492.3076923076923;
     -290;
     -3041;
     37.99999999999986;
     3362;
     775.9615384615385;
     2292.953846153846;
     -618.2307692307694;
     -2481.699999999999;
     2715.763076923076];
sw = 13;
isx = [int32(1);
     int32(1);
     int32(1);
     int32(1)];
monlev = int32(1);
[isxOut, b, se, rsq, rms, df, user, ifail] = ...
    g02ef(n, wmean, c, sw, isx, monlev, 'g02ef_monfun')
```

```
Starting Stepwise Selection

Forward Selection
Variable 1  Variance ratio =       12.603
Variable 2  Variance ratio =       21.961
Variable 3  Variance ratio =        4.403
Variable 4  Variance ratio =       22.799

Adding variable 4 to model

Backward Selection
Variable 4  Variance ratio =       22.799

Keeping all current variables

Forward Selection
Variable 1  Variance ratio =      108.224
Variable 2  Variance ratio =        0.172
Variable 3  Variance ratio =       40.295

Adding variable 1 to model

Backward Selection
Variable 1  Variance ratio =      108.224
Variable 4  Variance ratio =      159.295

Keeping all current variables
```

```
Forward Selection
Variable 2  Variance ratio =        5.026
Variable 3  Variance ratio =        4.236

Adding variable 2 to model

Backward Selection
Variable 1  Variance ratio =      154.008
Variable 2  Variance ratio =        5.026
Variable 4  Variance ratio =        1.863

Dropping variable 4 from model

Forward Selection
Variable 3  Variance ratio =        1.832
Variable 4  Variance ratio =        1.863

Finished Stepwise Selection
isxOut =
           1
           1
           0
           0
b =
   52.5773
    1.4683
    0.6623
         0
         0
se =
    2.2943
    0.1213
    0.0459
         0
         0
rsq =
    0.9787
rms =
    5.7904
df =
          10
user =
     0
ifail =
           0
```