NAG Toolbox

nag_correg_linregm_fit_stepwise (g02ef)

1 Purpose

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

Note: the interface to this routine has changed since earlier releases of the toolbox:

At Mark 23: **monfun** was made optional; **monlevel** was removed from the interface.

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,

β_0 is an intercept coefficient,

X is an 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.

nag_correg_linregm_fit_stepwise (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:

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

where r_{ij} is the correlation between the explanatory variables i and j, for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, p$, 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 $r_{ik}/|r_{kk}|, \quad i = 1, 2, \ldots, p+1 \ (i \neq k)$;
r_{kj} by $r_{kj}/|r_{kk}|, \quad j = 1, 2, \ldots, p+1 \ (j \neq k)$;
r_{ij} by $r_{ij} - r_{ik}r_{kj}/|r_{kk}|, \quad i = 1, 2, \ldots, p+1 \ (i \neq k); \ j = 1, 2, \ldots, p+1 \ (j \neq k)$.

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 function 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 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 *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

5.1 Compulsory Input Parameters

1: **n** – INTEGER

The number of observations used in the calculations.

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

2: **wmean**$(\mathbf{m}+\mathbf{1})$ – REAL (KIND=nag_wp) array

The mean of the design matrix, Z.

3: **c**$((\mathbf{m}+\mathbf{1}) \times (\mathbf{m}+\mathbf{2})/\mathbf{2})$ – REAL (KIND=nag_wp) array

The upper-triangular variance-covariance matrix packed by column for the design matrix, Z. Because the function computes the correlation matrix R from **c**, the variance-covariance matrix need only be supplied up to a scaling factor.

4: **sw** – REAL (KIND=nag_wp)

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: $\mathbf{sw} > 1.0$.

5: **isx**$(\mathbf{m})$ – INTEGER 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 jth column of X from the final model.

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

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

Constraint: **isx**$(j) = -1, 1$ or 2, for $j = 1, 2, \ldots, \mathbf{m}$.

5.2 Optional Input Parameters

1: **m** – INTEGER

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

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

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

2: **fin** – REAL (KIND=nag_wp)

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

Default: 4.0

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

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

3: **fout** – REAL (KIND=nag_wp)

Suggested value: **fout** = **fin**.

Default: **fin**

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: $0.0 \leq$ **fout** $\leq$ **fin**.

4: **tau** – REAL (KIND=nag_wp)

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

Default: 0.000001

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: **tau** > 0.0.

5: **monfun** – SUBROUTINE, supplied by the NAG Library or the user.

You may define your own function or specify the NAG defined default function nag_correg_linregm_fit_stepwise_sample_monfun (g02efh).

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

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

Input Parameters

1: **flag** – CHARACTER(1)

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

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

3: **val** – REAL (KIND=nag_wp)

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

4: **user** – INTEGER array

monfun is called from nag_correg_linregm_fit_stepwise (g02ef) with the object supplied to nag_correg_linregm_fit_stepwise (g02ef).

Output Parameters

1: **user** – INTEGER array

6: **user** – INTEGER array

user is not used by nag_correg_linregm_fit_stepwise (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 Output Parameters

1: **isx**(**m**) – INTEGER array

The value of **isx**(j) indicates the status of the jth explanatory variable in the model.

isx(j) = -1
Forced exclusion.

isx(j) = 0
Excluded.

isx(j) = 1
Selected.

isx(j) = 2
Forced selection.

2: **b**(**m** + **1**) – REAL (KIND=nag_wp) array

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 jth explanatory variable in the fitted model; otherwise **b**($j+1$) = 0.

3: **se**(**m** + **1**) – REAL (KIND=nag_wp) array

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

4: **rsq** – REAL (KIND=nag_wp)

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

5: **rms** – REAL (KIND=nag_wp)

The mean square of residuals for the fitted regression model.

6: **df** – INTEGER

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

7: **user** – INTEGER array

8: **ifail** – INTEGER

ifail $= 0$ unless the function detects an error (see Section 5).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail $= 1$

Constraint: $0.0 \leq$ **fout** $\leq$ **fin**.

Constraint: **fin** > 0.0.

Constraint: **m** > 1.

Constraint: $monlev = 0$ or 1.

Constraint: **n** > 1.

Constraint: **sw** > 1.0.

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 .

On entry at least one diagonal element of **c** ≤ 0.0.

ifail $= 3$ (*warning*)

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.

ifail $= -399$

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

ifail $= -999$

Dynamic memory allocation failed.

7 Accuracy

nag_correg_linregm_fit_stepwise (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)).

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

9 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 nag_correg_ssqmat (g02bu). The NAG defined default monitor function nag_correg_linregm_fit_stepwise_sample_monfun (g02efh) is used to print information at each step of the model selection process.

9.1 Program Text

```
     function g02ef_example

fprintf('g02ef example results\n\n');

z   = [  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];
[n,m1] = size(z);
m = m1 - 1;

isx = ones(m,1,nag_int_name);

% Compute sums of squares and cross-products of deviations from mean for z
[sw, wmean, c, ifail] = g02bu(z);

% Perform stepwise selection of variables
fout = 2;
[isx, b, se, rsq, rms, df, user, ifail] = ...
g02ef( ...
       nag_int(n), wmean, c, sw, isx, 'fout', fout, 'monfun', @monfun);

% Display results
fprintf('\nFitted Model Summary\n');
fprintf('Term              Estimate   Standard Error\n');
fprintf('Intercept:    %12.3e     %12.3e\n', b(1), se(1));
for j = 1:m
  if isx(j)==1 || isx(j)==2
    fprintf('Variable: %3d %12.3e     %12.3e\n', j, b(j+1), se(j+1));
  end
end
fprintf('\nRMS: %12.3e\n', rms);

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;
```

9.2 Program Results

```
     g02ef example results

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

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
```