NAG Toolbox

nag_correg_linregm_fit_onestep (g02ee)

1 Purpose

nag_correg_linregm_fit_onestep (g02ee) carries out one step of a forward selection procedure in order to enable the 'best' linear regression model to be found.

2 Syntax

```
[istep, addvar, newvar, chrss, f, model, nterm, rss, idf, ifr, free, exss, q,
p, ifail] = nag_correg_linregm_fit_onestep(istep, mean, x, vname, isx, y, model,
nterm, rss, idf, ifr, free, q, p, 'n', n, 'm', m, 'maxip', maxip, 'wt', wt,
'fin', fin)
[istep, addvar, newvar, chrss, f, model, nterm, rss, idf, ifr, free, exss, q,
p, ifail] = g02ee(istep, mean, x, vname, isx, y, model, nterm, rss, idf, ifr,
free, q, p, 'n', n, 'm', m, 'maxip', maxip, 'wt', wt, 'fin', fin)
```

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.

nag_correg_linregm_fit_onestep (g02ee) 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_{free} is the matrix of variables not included in the model, is updated.

nag_correg_linregm_fit_onestep (g02ee) 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 nag_correg_linregm_update (g02dd), in order to compute the regression coefficients for the model fitted at that step. Repeated calls to nag_correg_linregm_fit_onestep (g02ee) 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 Parameters

Note: after the initial call to nag_correg_linregm_fit_onestep (g02ee) with **istep** $= 0$ all arguments except **fin** must not be changed by you between calls.

5.1 Compulsory Input Parameters

1: **istep** – INTEGER

Indicates which step in the forward selection process is to be carried out.

istep $= 0$
The process is initialized.

Constraint: **istep** ≥ 0.

2: **mean_p** – CHARACTER(1)

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: $\mathbf{x}(ldx, \mathbf{m})$ – REAL (KIND=nag_wp) array

ldx, the first dimension of the array, must satisfy the constraint $ldx \geq \mathbf{n}$.

$\mathbf{x}(i,j)$ must contain the ith observation for the jth independent variable, for $i = 1, 2, \ldots, \mathbf{n}$ and $j = 1, 2, \ldots, \mathbf{m}$.

4: $\mathbf{vname}(\mathbf{m})$ – CHARACTER(*) array

$\mathbf{vname}(j)$ must contain the name of the independent variable in column j of **x**, for $j = 1, 2, \ldots, \mathbf{m}$.

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

Indicates which independent variables could be considered for inclusion in the regression.

$\mathbf{isx}(j) \geq 2$
The variable contained in the jth column of **x** is automatically included in the regression model, for $j = 1, 2, \ldots, \mathbf{m}$.

$\mathbf{isx}(j) = 1$
The variable contained in the jth column of **x** is considered for inclusion in the regression model, for $j = 1, 2, \ldots, \mathbf{m}$.

$\mathbf{isx}(j) = 0$
The variable in the jth column is not considered for inclusion in the model, for $j = 1, 2, \ldots, \mathbf{m}$.

Constraint: $\mathbf{isx}(j) \geq 0$ and at least one value of $\mathbf{isx}(j) = 1$, for $j = 1, 2, \ldots, \mathbf{m}$.

6: $\mathbf{y}(\mathbf{n})$ – REAL (KIND=nag_wp) array

The dependent variable.

7: $\mathbf{model}(\mathbf{maxip})$ – CHARACTER(*) array

If **istep** $= 0$, **model** need not be set.

If **istep** $\neq 0$, **model** must contain the values returned by the previous call to nag_correg_linregm_fit_onestep (g02ee).

Constraint: the declared size of **model** must be greater than or equal to the declared size of **vname**.

8: **nterm** – INTEGER

If **istep** $= 0$, **nterm** need not be set.

If **istep** $\neq 0$, **nterm** must contain the value returned by the previous call to nag_correg_linregm_fit_onestep (g02ee).

Constraint: if **istep** $\neq 0$, **nterm** > 0.

9: **rss** – REAL (KIND=nag_wp)

If **istep** $= 0$, **rss** need not be set.

If **istep** $\neq 0$, **rss** must contain the value returned by the previous call to nag_correg_linregm_fit_onestep (g02ee).

Constraint: if **istep** $\neq 0$, **rss** > 0.0.

10: **idf** – INTEGER

If **istep** $= 0$, **idf** need not be set.

If **istep** $\neq 0$, **idf** must contain the value returned by the previous call to nag_correg_linregm_fit_onestep (g02ee).

11: **ifr** – INTEGER

If **istep** $= 0$, **ifr** need not be set.

If **istep** $\neq 0$, **ifr** must contain the value returned by the previous call to nag_correg_linregm_fit_onestep (g02ee).

12: **free**(**maxip**) – CHARACTER(*) array

If **istep** $= 0$, **free** need not be set.

If **istep** $\neq 0$, **free** must contain the values returned by the previous call to nag_correg_linregm_fit_onestep (g02ee).

Constraint: the declared size of **free** must be greater than or equal to the declared size of **vname**.

13: **q**(ldq, **maxip** $+ 2$) – REAL (KIND=nag_wp) array

ldq, the first dimension of the array, must satisfy the constraint $ldq \geq$ **n**.

If **istep** $= 0$, **q** need not be set.

If **istep** $\neq 0$, **q** must contain the values returned by the previous call to nag_correg_linregm_fit_onestep (g02ee).

14: **p**(**maxip** $+ 1$) – REAL (KIND=nag_wp) array

If **istep** $= 0$, **p** need not be set.

If **istep** $\neq 0$, **p** must contain the values returned by the previous call to nag_correg_linregm_fit_onestep (g02ee).

5.2 Optional Input Parameters

1: **n** – INTEGER

Default: the dimension of the array **y** and the first dimension of the arrays **x**, **q**. (An error is raised if these dimensions are not equal.)

n, the number of observations.

Constraint: **n** ≥ 2.

2: **m** – INTEGER

Default: the second dimension of the array **x** and the dimension of the arrays **vname**, **isx**. (An error is raised if these dimensions are not equal.)

m, the total number of independent variables in the dataset.

Constraint: **m** ≥ 1.

3: **maxip** – INTEGER

Default: the dimension of the arrays **model**, **free**. (An error is raised if these dimensions are not equal.)

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.

4: **wt**(:) – REAL (KIND=nag_wp) array

The dimension of the array **wt** must be at least **n** if $weight =$ 'W'

If $weight =$ 'W', **wt** must contain the weights to be used in the weighted regression, W.

If **wt**$(i) = 0.0$, the ith 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 $i = 1, 2, \ldots,$ **n**.

5: **fin** – REAL (KIND=nag_wp)

Default: 2.0 is a commonly used value in exploratory modelling.

The critical value of the F statistic for the term to be included in the model, F_{c}.

Constraint: **fin** ≥ 0.0.

5.3 Output Parameters

1: **istep** – INTEGER

Is incremented by 1.

2: **addvar** – LOGICAL

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_{c} and none were added to the model.

3: **newvar** – CHARACTER(*)

If **addvar** $= true$, **newvar** contains the name of the variable added to the model.

4: **chrss** – REAL (KIND=nag_wp)

If **addvar** $= true$, **chrss** contains the change in the residual sum of squares due to adding variable **newvar**.

5: **f** – REAL (KIND=nag_wp)

If **addvar** $= true$, **f** contains the F statistic for the inclusion of the variable in **newvar**.

6: **model**(**maxip**) – CHARACTER(*) array

The names of the variables in the current model.

7: **nterm** – INTEGER

The number of independent variables in the current model, not including the mean, if any.

8: **rss** – REAL (KIND=nag_wp)

The residual sums of squares for the current model.

9: **idf** – INTEGER

The degrees of freedom for the residual sum of squares for the current model.

10: **ifr** – INTEGER

The number of free independent variables, i.e., the number of variables not in the model that are still being considered for selection.

11: **free**(**maxip**) – CHARACTER(*) array

The first **ifr** values of **free** contain the names of the free variables.

12: **exss**(**maxip**) – REAL (KIND=nag_wp) array

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.

13: **q**$(ldq,$ **maxip** $+ 2)$ – REAL (KIND=nag_wp) array

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'.

14: **p**(**maxip** $+ 1$) – REAL (KIND=nag_wp) array

The first p elements of **p** contain details of the QR decomposition, where $p =$ **nterm**, or $p =$ **nterm** $+ 1$ if **mean** = 'M'.

15: **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 $= 1$

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** ≤ 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.

ifail $= -399$

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

ifail $= -999$

Dynamic memory allocation failed.

7 Accuracy

As nag_correg_linregm_fit_onestep (g02ee) 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 Further Comments

None.

9 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 nag_correg_linregm_fit_onestep (g02ee) 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.

9.1 Program Text

```
     function g02ee_example

fprintf('g02ee example results\n\n');

x = [  0, 1125, 232, 7160, 85.9, 8905;
       7,  920, 268, 8804, 86.5, 7388;
      15,  835, 271, 8108, 85.2, 5348;
      22, 1000, 237, 6370, 83.8, 8056;
      29, 1150, 192, 6441, 82.1, 6960;
      37,  990, 202, 5154, 79.2, 5690;
      44,  840, 184, 5896, 81.2, 6932;
      58,  650, 200, 5336, 80.6, 5400;
      65,  640, 180, 5041, 78.4, 3177;
      72,  583, 165, 5012, 79.3, 4461;
      80,  570, 151, 4825, 78.7, 3901;
      86,  570, 171, 4391, 78.0, 5002;
      93,  510, 243, 4320, 72.3, 4665;
     100,  555, 147, 3709, 74.9, 4642;
     107,  460, 286, 3969, 74.4, 4840;
     122,  275, 198, 3558, 72.5, 4479;
     129,  510, 196, 4361, 57.7, 4200;
     151,  165, 210, 3301, 71.8, 3410;
     171,  244, 327, 2964, 72.5, 3360;
     220,   79, 334, 2777, 71.9, 2599];
y = [ 1.5563;  0.8976;  0.7482;  0.7160;  0.3010;
      0.3617;  0.1139;  0.1139; -0.2218; -0.1549;
      0.0000;  0.0000; -0.0969; -0.2218; -0.3979;
     -0.1549; -0.2218; -0.3979; -0.5229; -0.0458];
[n,m] = size(x);

mean_p = 'M';
isx = ones(m,1,nag_int_name);
isx(1) = 0;
isx(m) = 2;
vname = {'DAY'; 'BOD'; 'TKN'; 'TS '; 'TVS'; 'COD'};

nzero = nag_int(0);
model = {'   '; '   '; '   '; '   '; '   '; '   '};
nterm = nzero;
rss = 0;
idf = nzero;
ifr = nzero;
free = model;
q = zeros(n,m+2);
p = zeros(m+1,1);

% Loop attempting to add each variable in turn
istep = nzero;
addvar = true;
while addvar
  [istep, addvar, newvar, chrss, f, model, nterm, ...
   rss, idf, ifr, free, exss, q, p, ifail] = ...
  g02ee( ...
```

```
        istep, mean_p, x, vname, isx, y, model, nterm, ...
        rss, idf, ifr, free, q, p);

  % Display the results at each step
  fprintf('Step %3d\n', istep);
  if ~addvar
    fprintf('No further variables added max  F = %7.2f\n', f);
  else
    fprintf('Added variable is %s\n', newvar);
    fprintf('Change in residual sum of squares = %13.4e\n', chrss);
    fprintf('F Statistic                       = %7.2f\n\n', f);
    fprintf('Variables in model              :');
    fprintf('     %s', model{1:nterm,1});
    fprintf('\n\nResidual sum of squares           = %13.4e\n', rss);
    fprintf('Degrees of freedom                = %2d\n\n', idf);
  end
  if ifr==0
    fprintf('No free variables remaining\n');
    addvar = false;
  else
    fprintf('Free variables                  :')
    fprintf('     %s', free{1:ifr,1});
    fprintf('\nChange in RSS for free variables:\n%33s',' ');
    fprintf('%8.4f', exss(1:ifr));
    fprintf('\n\n');
  end
end
```

9.2 Program Results

```
     g02ee example results

Step   1
Added variable is TS
Change in residual sum of squares =    4.7126e-01
F Statistic                       =    7.38

Variables in model              :     COD     TS

Residual sum of squares           =    1.0850e+00
Degrees of freedom                = 17

Free variables                  :     TKN     BOD     TVS
Change in RSS for free variables:
                                   0.1175  0.0600  0.2276

Step   2
No further variables added max  F =    1.59
Free variables                  :     TKN     BOD     TVS
Change in RSS for free variables:
                                   0.0979  0.0207  0.0217
```