NAG Toolbox

nag_correg_linregm_rssq (g02ea)

1 Purpose

nag_correg_linregm_rssq (g02ea) calculates the residual sums of squares for all possible linear regressions for a given set of independent variables.

2 Syntax

```
[nmod, modl, rss, nterms, mrank, ifail] = nag_correg_linregm_rssq(mean, x, vname, isx, y, 'n', n, 'm', m, 'wt', wt)
[nmod, modl, rss, nterms, mrank, ifail] = g02ea(mean, x, vname, isx, y, 'n', n, 'm', m, 'wt', wt)
```

3 Description

For a set of k possible independent variables there are 2^k linear regression models with from zero to k independent variables in each model. For example if $k = 3$ and the variables are A, B and C then the possible models are:

(i) null model

(ii) A

(iii) B

(iv) C

(v) A and B

(vi) A and C

(vii) B and C

(viii) A, B and C.

nag_correg_linregm_rssq (g02ea) calculates the residual sums of squares from each of the 2^k possible models. The method used involves a QR decomposition of the matrix of possible independent variables. Independent variables are then moved into and out of the model by a series of Givens rotations and the residual sums of squares computed for each model; see Clark (1981) and Smith and Bremner (1989).

The computed residual sums of squares are then ordered first by increasing number of terms in the model, then by decreasing size of residual sums of squares. So the first model will always have the largest residual sum of squares and the 2^kth will always have the smallest. This aids you in selecting the best possible model from the given set of independent variables.

nag_correg_linregm_rssq (g02ea) allows you to specify some independent variables that must be in the model, the forced variables. The other independent variables from which the possible models are to be formed are the free variables.

4 References

Clark M R B (1981) A Givens algorithm for moving from one linear model to another without going back to the data *Appl. Statist.* **30** 198–203

Smith D M and Bremner J M (1989) All possible subset regressions using the QR decomposition *Comput. Statist. Data Anal.* **7** 217–236

Weisberg S (1985) *Applied Linear Regression* Wiley

5 Parameters

5.1 Compulsory Input Parameters

1: **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'.

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

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

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

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

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

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

Indicates which independent variables are to be considered in the model.

isx$(j) \geq 2$
The variable contained in the jth column of **x** is included in all regression models, i.e., is a forced variable.

isx$(j) = 1$
The variable contained in the jth column of **x** is included in the set from which the regression models are chosen, i.e., is a free variable.

isx$(j) = 0$
The variable contained in the jth column of **x** is not included in the models.

Constraints:

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

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

y(i) must contain the ith observation on the dependent variable, y_i, for $i = 1, 2, \ldots, n$.

5.2 Optional Input Parameters

1: **n** – INTEGER

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

n, the number of observations.

Constraints:

n ≥ 2;
n $\geq m$, is the number of independent variables to be considered (forced plus free plus mean if included), as specified by **mean** and **isx**.

2: **m** – INTEGER

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

The number of variables contained in **x**.

Constraint: $\mathbf{m} \geq 2$.

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

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

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

If $\mathbf{wt}(i) = 0.0$, 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 $weight = \text{'U'}$, **wt** is not referenced and the effective number of observations is **n**.

Constraint: if $weight = \text{'W'}$, $\mathbf{wt}(i) \geq 0.0$, for $i = 1, 2, \ldots, n$.

5.3 Output Parameters

1: **nmod** – INTEGER

The total number of models for which residual sums of squares have been calculated.

2: **modl**$(ldmodl, \mathbf{m})$ – CHARACTER(*) array

The first **nterms**(i) elements of the ith row of **modl** contain the names of the independent variables, as given in **vname**, that are included in the ith model.

3: **rss**$(ldmodl)$ – REAL (KIND=nag_wp) array

rss(i) contains the residual sum of squares for the ith model, for $i = 1, 2, \ldots,$ **nmod**.

4: **nterms**$(ldmodl)$ – INTEGER array

nterms(i) contains the number of independent variables in the ith model, not including the mean if one is fitted, for $i = 1, 2, \ldots,$ **nmod**.

5: **mrank**$(ldmodl)$ – INTEGER array

mrank(i) contains the rank of the residual sum of squares for the ith model.

6: **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, $\mathbf{n} < 2$,
or $\mathbf{m} < 2$,
or $ldx < \mathbf{n}$,
or $ldmodl < \mathbf{m}$,
or **mean** $\neq$ 'M' or 'Z',
or $weight \neq$ 'U' or 'W'.

ifail $= 2$

On entry, $weight = \text{'W'}$ and a value of $\mathbf{wt} < 0.0$.

ifail $= 3$

On entry, a value of **isx** < 0,
or there are no free variables, i.e., no element of **isx** $= 1$.

ifail $= 4$

On entry, $ldmodl <$ the number of possible models $= 2^k$, where k is the number of free independent variables from **isx**.

ifail $= 5$

On entry, the number of independent variables to be considered (forced plus free plus mean if included) is greater or equal to the effective number of observations.

ifail $= 6$

The full model is not of full rank, i.e., some of the independent variables may be linear combinations of other independent variables. Variables must be excluded from the model in order to give full rank.

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

For a discussion of the improved accuracy obtained by using a method based on the QR decomposition see Smith and Bremner (1989).

8 Further Comments

nag_correg_linregm_rssq_stat (g02ec) may be used to compute R^2 and C_p-values from the results of nag_correg_linregm_rssq (g02ea).

If a mean has been included in the model and no variables are forced in then **rss**(1) contains the total sum of squares and in many situations a reasonable estimate of the variance of the errors is given by **rss**(**nmod**)/(**n** $-$ 1 $-$ **nterms**(**nmod**)).

9 Example

The data for this example is given in Weisberg (1985). The independent variables and the dependent variable are read, as are the names of the variables. These names are as given in Weisberg (1985). The residual sums of squares computed and printed with the names of the variables in the model.

9.1 Program Text

```
function g02ea_example

fprintf('g02ea 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;
vname = {'DAY'; 'BOD'; 'TKN'; 'TS '; 'TVS'; 'COD'};

% Calculate residual sums of squares for all possible models
[nmod, model, rss, nterms, mrank, ifail] = ...
  g02ea(mean_p, x, vname, isx, y);

% Display results
fprintf(' Parameters    RSS   rank   model\n');
for j = 1:nmod
  fprintf('%8d%11.4f%4d  ', nterms(j), rss(j), mrank(j));
  fprintf(' %s', model{j,:});
  fprintf('\n');
end
```

9.2 Program Results

```
     g02ea example results

 Parameters    RSS   rank   model
       0     5.0634  32
       1     5.0219  31   TKN
       1     2.5044  30   TVS
       1     2.0338  28   BOD
       1     1.5563  25   COD
       1     1.5370  24   TS
       2     2.4381  29   TKN TVS
       2     1.7462  27   BOD TVS
       2     1.5921  26   BOD TKN
       2     1.4963  23   BOD COD
       2     1.4707  22   TKN TS
       2     1.4590  21   TS  TVS
       2     1.4397  20   BOD TS
       2     1.4388  19   TKN COD
       2     1.3287  15   TVS COD
       2     1.0850   8   TS  COD
       3     1.4257  18   BOD TKN TVS
       3     1.3900  17   TKN TS  TVS
       3     1.3894  16   BOD TS  TVS
       3     1.3204  14   BOD TVS COD
       3     1.2764  13   BOD TKN COD
       3     1.2582  12   BOD TKN TS
       3     1.2179  10   TKN TVS COD
       3     1.0644   7   BOD TS  COD
       3     1.0634   6   TS  TVS COD
       3     0.9871   4   TKN TS  COD
```

```
      4     1.2199  11   BOD TKN TS  TVS
      4     1.1565   9   BOD TKN TVS COD
      4     1.0388   5   BOD TS  TVS COD
      4     0.9871   3   BOD TKN TS  COD
      4     0.9653   2   TKN TS  TVS COD
      5     0.9652   1   BOD TKN TS  TVS COD
```