NAG Toolbox for MATLAB

g02ea

1 Purpose

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] = 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.

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.

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 – string**

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,m) – double array**

ldx, the first dimension of the array, must be at least **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(m) – string array**

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

4: **isx(m) – int32 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.

We denote the total number of free variables a k, where k is the number of free variables in the model, see **nmod** to **mrank**.

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

5: **y(n) – double 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 – int32 scalar**

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

the number of observations.

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

2: **m – int32 scalar**

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

the maximum number of variables contained in **x**.

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

3: **wt**($*$) – **double array**

Note: the dimension of the array **wt** must be at least **n**.

If **weight** = '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** = 'U', **wt** is not referenced and the effective number of observations is **n**.

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

5.3 Input Parameters Omitted from the MATLAB Interface

weight, ldx, ldmodl, wk

5.4 Output Parameters

1: **nmod – int32 scalar**

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

2: **modl**(**ldmodl**,**m**) – **string 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**) – **double array**

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

4: **nterms**(**ldmodl**) – **int32 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**) – **int32 array**

mrank(i) contains the rank of the residual sum of squares for the ith model, i.e., model with smallest sum of squares has rank 1.

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

ifail $= 2$

On entry, **weight** = 'W' and a value of **wt** < 0.0.

ifail $= 3$

On entry, a value of **isx** < 0.0,
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.

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

g02ec may be used to compute R^2 and C_p-values from the results of 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

```
mean = ’M’;
x = [0, 1125, 232, 7160, 85.90000000000001, 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, 82.09999999999999, 6960;
     37, 990, 202, 5154, 79.2, 5690;
     44, 840, 184, 5896, 81.2, 6932;
     58, 650, 200, 5336, 80.59999999999999, 5400;
     65, 640, 180, 5041, 78.40000000000001, 3177;
     72, 583, 165, 5012, 79.3, 4461;
     80, 570, 151, 4825, 78.7, 3901;
     86, 570, 171, 4391, 78, 5002;
     93, 510, 243, 4320, 72.3, 4665;
     100, 555, 147, 3709, 74.90000000000001, 4642;
     107, 460, 286, 3969, 74.40000000000001, 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, 71.90000000000001, 2599];
vname = {’DAY’; ’BOD’; ’TKN’; ’TS ’; ’TVS’; ’COD’};
```

```
isx = [int32(0);
     int32(1);
     int32(1);
     int32(1);
     int32(1);
     int32(1)];
y = [1.5563;
     0.8976;
     0.7482;
     0.716;
     0.301;
     0.3617;
     0.1139;
     0.1139;
     -0.2218;
     -0.1549;
     0;
     0;
     -0.0969;
     -0.2218;
     -0.3979;
     -0.1549;
     -0.2218;
     -0.3979;
     -0.5229;
     -0.0458];
[nmod, model, rss, nterms, mrank, ifail] = g02ea(mean, x, vname, isx, y)
```

```
nmod =
          32
model =
       ''       ''       ''       ''       ''     ''
    'TKN'       ''       ''       ''       ''     ''
    'TVS'       ''       ''       ''       ''     ''
    'BOD'       ''       ''       ''       ''     ''
    'COD'       ''       ''       ''       ''     ''
    'TS '       ''       ''       ''       ''     ''
    'TKN'    'TVS'       ''       ''       ''     ''
    'BOD'    'TVS'       ''       ''       ''     ''
    'BOD'    'TKN'       ''       ''       ''     ''
    'BOD'    'COD'       ''       ''       ''     ''
    'TKN'    'TS '       ''       ''       ''     ''
    'TS '    'TVS'       ''       ''       ''     ''
    'BOD'    'TS '       ''       ''       ''     ''
    'TKN'    'COD'       ''       ''       ''     ''
    'TVS'    'COD'       ''       ''       ''     ''
    'TS '    'COD'       ''       ''       ''     ''
    'BOD'    'TKN'    'TVS'       ''       ''     ''
    'TKN'    'TS '    'TVS'       ''       ''     ''
    'BOD'    'TS '    'TVS'       ''       ''     ''
    'BOD'    'TVS'    'COD'       ''       ''     ''
    'BOD'    'TKN'    'COD'       ''       ''     ''
    'BOD'    'TKN'    'TS '       ''       ''     ''
    'TKN'    'TVS'    'COD'       ''       ''     ''
    'BOD'    'TS '    'COD'       ''       ''     ''
    'TS '    'TVS'    'COD'       ''       ''     ''
    'TKN'    'TS '    'COD'       ''       ''     ''
    'BOD'    'TKN'    'TS '    'TVS'       ''     ''
    'BOD'    'TKN'    'TVS'    'COD'       ''     ''
    'BOD'    'TS '    'TVS'    'COD'       ''     ''
    'BOD'    'TKN'    'TS '    'COD'       ''     ''
    'TKN'    'TS '    'TVS'    'COD'       ''     ''
    'BOD'    'TKN'    'TS '    'TVS'    'COD'     ''
rss =
    5.0634
    5.0219
    2.5044
    2.0338
    1.5563
    1.5370
    2.4381
    1.7462
    1.5921
    1.4963
    1.4707
    1.4590
    1.4397
    1.4388
    1.3287
    1.0850
    1.4257
    1.3900
    1.3894
    1.3204
    1.2764
    1.2582
    1.2179
    1.0644
    1.0634
    0.9871
    1.2199
    1.1565
    1.0388
    0.9871
    0.9653
    0.9652
nterms =
           0
           1
           1
           1
           1
           1
           2
           2
           2
           2
           2
           2
           2
           2
           2
           2
           3
           3
           3
           3
           3
           3
           3
           3
           3
           3
           4
           4
           4
           4
           4
           5
mrank =
          32
          31
          30
          28
          25
          24
          29
          27
          26
          23
          22
          21
          20
          19
          15
           8
          18
          17
          16
          14
          13
          12
          10
           7
           6
           4
          11
           9
           5
           3
           2
           1
ifail =
           0
```