NAG Toolbox

nag_correg_linregm_rssq_stat (g02ec)

1 Purpose

nag_correg_linregm_rssq_stat (g02ec) calculates R^2 and C_p-values from the residual sums of squares for a series of linear regression models.

2 Syntax

```
[rsq, cp, ifail] = nag_correg_linregm_rssq_stat(mean, n, sigsq, tss, nterms, rss, 'nmod', nmod)
[rsq, cp, ifail] = g02ec(mean, n, sigsq, tss, nterms, rss, 'nmod', nmod)
```

3 Description

When selecting a linear regression model for a set of n observations a balance has to be found between the number of independent variables in the model and fit as measured by the residual sum of squares. The more variables included the smaller will be the residual sum of squares. Two statistics can help in selecting the best model.

(a) R^2 represents the proportion of variation in the dependent variable that is explained by the independent variables.

$$R^2 = \frac{\text{Regression Sum of Squares}}{\text{Total Sum of Squares}},$$

where Total Sum of Squares $=$ **tss** $= \sum (y - \bar{y})^2$ (if mean is fitted, otherwise **tss** $= \sum y^2$) and

Regression Sum of Squares $=$ RegSS $=$ **tss** $-$ **rss**, where
rss $=$ residual sum of squares $= \sum (y - \hat{y})^2$.

The R^2-values can be examined to find a model with a high R^2-value but with small number of independent variables.

(b) C_p statistic.

$$C_p = \frac{\textbf{rss}}{\hat{\sigma}^2} - (n - 2p),$$

where p is the number of arguments (including the mean) in the model and $\hat{\sigma}^2$ is an estimate of the true variance of the errors. This can often be obtained from fitting the full model.

A well fitting model will have $C_p \simeq p$. C_p is often plotted against p to see which models are closest to the $C_p = p$ line.

nag_correg_linregm_rssq_stat (g02ec) may be called after nag_correg_linregm_rssq (g02ea) which calculates the residual sums of squares for all possible linear regression models.

4 References

Draper N R and Smith H (1985) *Applied Regression Analysis* (2nd Edition) Wiley

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: **n** – INTEGER

n, the number of observations used in the regression model.

Constraint: **n** must be greater than $2 \times p_{\max}$, where $p_{\max}$ is the largest number of independent variables fitted (including the mean if fitted).

3: **sigsq** – REAL (KIND=nag_wp)

The best estimate of true variance of the errors, $\hat{\sigma}^2$.

Constraint: **sigsq** > 0.0.

4: **tss** – REAL (KIND=nag_wp)

The total sum of squares for the regression model.

Constraint: **tss** > 0.0.

5: **nterms**(**nmod**) – INTEGER array

nterms(i) must contain the number of independent variables (not counting the mean) fitted to the ith model, for $i = 1, 2, \ldots,$ **nmod**.

6: **rss**(**nmod**) – REAL (KIND=nag_wp) array

rss(i) must contain the residual sum of squares for the ith model.

Constraint: **rss**$(i) \leq$ **tss**, for $i = 1, 2, \ldots,$ **nmod**.

5.2 Optional Input Parameters

1: **nmod** – INTEGER

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

The number of regression models.

Constraint: **nmod** > 0.

5.3 Output Parameters

1: **rsq**(**nmod**) – REAL (KIND=nag_wp) array

rsq(i) contains the R^2-value for the ith model, for $i = 1, 2, \ldots,$ **nmod**.

2: **cp**(**nmod**) – REAL (KIND=nag_wp) array

cp(i) contains the C_p-value for the ith model, for $i = 1, 2, \ldots,$ **nmod**.

3: **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, **nmod** < 1,
or **sigsq** ≤ 0.0,
or **tss** ≤ 0.0.
or **mean** $\neq$ 'M' or 'Z'.

ifail $= 2$

On entry, the number of arguments for a model is too large for the number of observations, i.e., $2 \times p \geq n$.

ifail $= 3$

On entry, **rss**$(i) >$ **tss**, for some $i = 1, 2, \ldots,$ **nmod**.

ifail $= 4$

A value of C_p is less than 0.0. This may occur if **sigsq** is too large or if **rss**, **n** or IP are incorrect.

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

Accuracy is sufficient for all practical purposes.

8 Further Comments

None.

9 Example

The data, from an oxygen uptake experiment, is given by Weisberg (1985). The independent and dependent variables are read and the residual sums of squares for all possible models computed using nag_correg_linregm_rssq (g02ea). The values of R^2 and C_p are then computed and printed along with the names of variables in the models.

9.1 Program Text

```
     function g02ec_example

fprintf('g02ec 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);

tss = rss(1);
sigsq = rss(nmod)/double(n-nterms(nmod)-1);

% Calculate R^2 and Mallows Cp
[rsq, cp, ifail] = g02ec( ...
                         mean_p, nag_int(n), sigsq, tss, nterms, rss);

% Display results
fprintf(' Parameters     Cp    R^2      model\n');
for j = 1:nmod
  fprintf('%8d%11.2f%8.4f  ', nterms(j), cp(j), rsq(j));
  fprintf(' %s', model{j,:});
  fprintf('\n');
end
```

9.2 Program Results

```
     g02ec example results

 Parameters     Cp    R^2      model
       0      55.45  0.0000
       1      56.84  0.0082   TKN
       1      20.33  0.5054   TVS
       1      13.50  0.5983   BOD
       1       6.57  0.6926   COD
       1       6.29  0.6965   TS
       2      21.36  0.5185   TKN TVS
       2      11.33  0.6551   BOD TVS
       2       9.09  0.6856   BOD TKN
```

```
      2       7.70  0.7045   BOD COD
      2       7.33  0.7095   TKN TS
      2       7.16  0.7119   TS  TVS
      2       6.88  0.7157   BOD TS
      2       6.87  0.7158   TKN COD
      2       5.27  0.7376   TVS COD
      2       1.74  0.7857   TS  COD
      3       8.68  0.7184   BOD TKN TVS
      3       8.16  0.7255   TKN TS  TVS
      3       8.15  0.7256   BOD TS  TVS
      3       7.15  0.7392   BOD TVS COD
      3       6.51  0.7479   BOD TKN COD
      3       6.25  0.7515   BOD TKN TS
      3       5.67  0.7595   TKN TVS COD
      3       3.44  0.7898   BOD TS  COD
      3       3.42  0.7900   TS  TVS COD
      3       2.32  0.8050   TKN TS  COD
      4       7.70  0.7591   BOD TKN TS  TVS
      4       6.78  0.7716   BOD TKN TVS COD
      4       5.07  0.7948   BOD TS  TVS COD
      4       4.32  0.8050   BOD TKN TS  COD
      4       4.00  0.8094   TKN TS  TVS COD
      5       6.00  0.8094   BOD TKN TS  TVS COD
```