NAG Toolbox for MATLAB

g02ec

1 Purpose

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] = 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.

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

where p is the number of parameters (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.

g02ec may be called after 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 – 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: **n – int32 scalar**

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 – double scalar**

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

Constraint: **sigsq** > 0.0.

4: **tss – double scalar**

The total sum of squares for the regression model.

Constraint: **tss** > 0.0.

5: **nterms**(**nmod**) **– int32 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**) **– double 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 – int32 scalar**

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

the number of regression models.

Constraint: **nmod** > 0.

5.3 Input Parameters Omitted from the MATLAB Interface

None.

5.4 Output Parameters

1: **rsq**(**nmod**) **– double array**

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

2: **cp**(**nmod**) **– double array**

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

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

ifail $= 2$

On entry, the number of parameters 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.

7 Accuracy

Accuracy is sufficient for all practical purposes.

8 Further Comments

None.

9 Example

```
mean = 'M';
n = int32(20);
sigsq = 0.06894097715299177;
tss = 5.063404019999999;
nterms = [int32(0);
     int32(1);
     int32(1);
     int32(1);
     int32(1);
     int32(1);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(2);
     int32(3);
     int32(3);
     int32(3);
     int32(3);
```

```
     int32(3);
     int32(3);
     int32(3);
     int32(3);
     int32(3);
     int32(3);
     int32(4);
     int32(4);
     int32(4);
     int32(4);
     int32(4);
     int32(5)];
rss = [5.063404019999999;
     5.021872981610835;
     2.504400256872564;
     2.033792558702955;
     1.55630266252715;
     1.536980701902808;
     2.438093168251408;
     1.746202334318434;
     1.592101895299377;
     1.496267665868361;
     1.470697443158826;
     1.458986090473173;
     1.439684706601502;
     1.438803545844649;
     1.328730482955034;
     1.085046933797315;
     1.425685328311827;
     1.390030525118499;
     1.389409281541255;
     1.320363732523134;
     1.276355711100746;
     1.25824207806747;
     1.217854935490384;
     1.064355044934491;
     1.063352079567482;
     0.9871461021926983;
     1.219929564008994;
     1.156529834681246;
     1.038833684537899;
     0.9871272495792243;
     0.9652626827261778;
     0.9651736801418848];
[rsq, cp, ifail] = g02ec(mean, n, sigsq, tss, nterms, rss)
```

```
rsq =
         0
    0.0082
    0.5054
    0.5983
    0.6926
    0.6965
    0.5185
    0.6551
    0.6856
    0.7045
    0.7095
    0.7119
    0.7157
    0.7158
    0.7376
    0.7857
    0.7184
    0.7255
    0.7256
    0.7392
    0.7479
    0.7515
    0.7595
```

```
    0.7898
    0.7900
    0.8050
    0.7591
    0.7716
    0.7948
    0.8050
    0.8094
    0.8094
cp =
   55.4455
   56.8431
   20.3267
   13.5005
    6.5744
    6.2942
   21.3649
   11.3289
    9.0937
    7.7036
    7.3327
    7.1628
    6.8829
    6.8701
    5.2735
    1.7388
    8.6798
    8.1626
    8.1536
    7.1521
    6.5137
    6.2510
    5.6652
    3.4386
    3.4241
    2.3187
    7.6953
    6.7757
    5.0685
    4.3184
    4.0013
    6.0000
ifail =
           0
```