NAG Toolbox for MATLAB

g02fa

1 Purpose

g02fa calculates two types of standardized residuals and two measures of influence for a linear regression.

2 Syntax

```
[sres, ifail] = g02fa(n, ip, res, h, rms, 'nres', nres)
```

3 Description

For the general linear regression model

$$y = X\beta + \epsilon,$$

where y is a vector of length n of the dependent variable,

X is an n by p matrix of the independent variables,

β is a vector of length p of unknown parameters,

and ϵ is a vector of length n of unknown random errors such that $\operatorname{var}\epsilon = \sigma^2 I$.

The residuals are given by

$$r = y - \hat{y} = y - X\hat{\beta}$$

and the fitted values, $\hat{y} = X\hat{\beta}$, can be written as Hy for an n by n matrix H. The ith diagonal elements of H, h_i, give a measure of the influence of the ith values of the independent variables on the fitted regression model. The values of r and the h_i are returned by g02da.

g02fa calculates statistics which help to indicate if an observation is extreme and having an undue influence on the fit of the regression model. Two types of standardized residual are calculated:

(i) The ith residual is standardized by its variance when the estimate of σ^2, s^2, is calculated from all the data; this is known as internal Studentization.

$$RI_i = \frac{r_i}{s\sqrt{1 - h_i}}.$$

(ii) The ith residual is standardized by its variance when the estimate of σ^2, s_{-i}^2 is calculated from the data excluding the ith observation; this is known as external Studentization.

$$RE_i = \frac{r_i}{s_{-i}\sqrt{1 - h_i}} = r_i\sqrt{\frac{n - p - 1}{n - p - RI_i^2}}.$$

The two measures of influence are:

(i) Cook's D

$$D_i = \tfrac{1}{p} RE_i^2 \frac{h_i}{1 - h_i}.$$

(ii) Atkinson's T

$$T_i = |RE_i|\sqrt{\left(\frac{n-p}{p}\right)\left(\frac{h_i}{1-h_i}\right)}.$$

4 References

Atkinson A C 1981 Two graphical displays for outlying and influential observations in regression *Biometrika* **68** 13–20

Cook R D and Weisberg S 1982 *Residuals and Influence in Regression* Chapman and Hall

5 Parameters

5.1 Compulsory Input Parameters

1: **n – int32 scalar**

n, the number of observations included in the regression.

Constraint: $\mathbf{n} > \mathbf{ip} + 1$.

2: **ip – int32 scalar**

p, the number of linear parameters estimated in the regression model.

Constraint: $\mathbf{ip} \geq 1$.

3: **res(nres) – double array**

The residuals, r_i.

4: **h(nres) – double array**

The diagonal elements of H, h_i, corresponding to the residuals in **res**.

Constraint: $0.0 < \mathbf{h}(i) < 1.0$, for $i = 1, 2, \ldots, \mathbf{nres}$.

5: **rms – double scalar**

The estimate of σ^2 based on all n observations, s^2, i.e., the residual mean square.

Constraint: $\mathbf{rms} > 0.0$.

5.2 Optional Input Parameters

1: **nres – int32 scalar**

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

the number of residuals.

Constraint: $1 \leq \mathbf{nres} \leq \mathbf{n}$.

5.3 Input Parameters Omitted from the MATLAB Interface

ldsres

5.4 Output Parameters

1: **sres(ldsres,4) – double array**

The standardized residuals and influence statistics.

For the observation with residual, r_i, given in $\mathbf{res}(i)$.

sres$(i,1)$

Is the internally standardized residual, RI_i.

sres$(i,2)$

Is the externally standardized residual, RE_i.

sres$(i,3)$

Is Cook's D statistic, D_i.

sres$(i,4)$

Is Atkinson's T statistic, T_i.

2: **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, **ip** < 1,
or **n** $\leq$ **ip** $+ 1$,
or **nres** < 1,
or **nres** $>$ **n**,
or **ldsres** $<$ **nres**,
or **rms** ≤ 0.0.

ifail $= 2$

On entry, **h**$(i) \leq 0.0$ or ≥ 1.0, for some $i = 1, 2, \ldots,$ **nres**.

ifail $= 3$

On entry, the value of a residual is too large for the given value of **rms**.

7 Accuracy

Accuracy is sufficient for all practical purposes.

8 Further Comments

None.

9 Example

```
n = int32(24);
ip = int32(11);
res = [0.266;
     -0.1387;
     -0.2971;
     0.5926;
     -0.4013;
     0.1396;
     -1.3173;
     1.1226;
     0.0321;
     -0.7111];
```

```
h = [0.5518999999999999;
     0.9746;
     0.6256;
     0.3144;
     0.4106;
     0.6268;
     0.5479000000000001;
     0.2325;
     0.4115;
     0.3577];
rms = 0.5798;
[sres, ifail] = g02fa(n, ip, res, h, rms)
```

```
sres =
    0.5219    0.5067    0.0305    0.6113
   -1.1429   -1.1578    4.5566   -7.7966
   -0.6377   -0.6225    0.0618   -0.8747
    0.9399    0.9354    0.0368    0.6886
   -0.6865   -0.6718    0.0298   -0.6096
    0.3001    0.2893    0.0138    0.4076
   -2.5729   -3.5286    0.7293   -4.2230
    1.6829    1.8282    0.0780    1.0939
    0.0550    0.0528    0.0002    0.0480
   -1.1653   -1.1830    0.0687   -0.9598
ifail =
           0
```