

NAG Toolbox

nag_correg_linregm_stat_resinf (g02fa)

1 Purpose

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

2 Syntax

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

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

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 i th diagonal elements of H , h_i , give a measure of the influence of the i th values of the independent variables on the fitted regression model. The values of r and the h_i are returned by nag_correg_linregm_fit (g02da).

nag_correg_linregm_stat_resinf (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 i th 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 i th residual is standardized by its variance when the estimate of σ^2 , s_{-i}^2 is calculated from the data excluding the i th 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

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

(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** – INTEGER

n , the number of observations included in the regression.

Constraint: **n** > **ip** + 1.

2: **ip** – INTEGER

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

Constraint: **ip** ≥ 1.

3: **res(nres)** – REAL (KIND=nag_wp) array

The residuals, r_i .

4: **h(nres)** – REAL (KIND=nag_wp) array

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

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

5: **rms** – REAL (KIND=nag_wp)

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

Constraint: **rms** > 0.0.

5.2 Optional Input Parameters

1: **nres** – INTEGER

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 Output Parameters

1: **sres(ldsres,4)** – REAL (KIND=nag_wp) array

The standardized residuals and influence statistics.

For the observation with residual, r_i , given in **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** – 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, **ip** < 1,
or **n** ≤ **ip** + 1,
or **nres** < 1,
or **nres** > **n**,
or *ldsres* < **nres**,
or **rms** ≤ 0.0.

ifail = 2

On entry, **h**(i) ≤ 0.0 or ≥ 1.0, for some $i = 1, 2, \dots, \mathbf{nres}$.

ifail = 3

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

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

A set of 24 residuals and h_i values from a 11 argument model fitted to the cloud seeding data considered in Cook and Weisberg (1982) are input and the standardized residuals etc calculated and printed for the first 10 observations.

9.1 Program Text

```
function g02fa_example

fprintf('g02fa example results\n\n');

res = [ 0.2660  -0.1387  -0.2971   0.5926  -0.4013   0.1396  ...
        -1.3173   1.1226   0.0321  -0.7111   0.3439  -0.4379  ...
         0.0633  -0.0936   0.9968   0.0209  -0.4056   0.1396  ...
         0.0327   0.2970  -0.2277   0.5180   0.5301  -1.0650 ];

h    = [ 0.5519   0.9746   0.6256   0.3144   0.4106   0.6268  ...
         0.5479   0.2325   0.4115   0.3577   0.3342   0.1673  ...
         0.3874   0.1705   0.3466   0.3743   0.7527   0.9069  ...
         0.2610   0.6256   0.2485   0.3072   0.5848   0.4794 ];

n    = nag_int(numel(res));
ip   = nag_int(11);
rms  = 0.5798;

% Calculate standardised residuals
[sres, ifail] = g02fa( ...
                  n, ip, res, h, rms);

% Display results
fprintf('          Internally   Internally\n');
fprintf('Obs.   standardized standardized   Cook''s D   Atkinson''s T\n');
fprintf('          residuals   residuals\n\n');
for j = 1:ip-1
    fprintf('%2d%13.3f%13.3f%13.3f%13.3f\n', j, sres(j,1:4))
end
```

9.2 Program Results

```
g02fa example results
```

Obs.	Internally standardized residuals	Internally standardized residuals	Cook's D	Atkinson's T
1	0.522	0.507	0.030	0.611
2	-1.143	-1.158	4.557	-7.797
3	-0.638	-0.622	0.062	-0.875
4	0.940	0.935	0.037	0.689
5	-0.686	-0.672	0.030	-0.610
6	0.300	0.289	0.014	0.408
7	-2.573	-3.529	0.729	-4.223
8	1.683	1.828	0.078	1.094
9	0.055	0.053	0.000	0.048
10	-1.165	-1.183	0.069	-0.960
