NAG Toolbox

nag_correg_linregs_const (g02ca)

1 Purpose

nag_correg_linregs_const (g02ca) performs a simple linear regression with dependent variable y and independent variable x.

2 Syntax

```
[result, ifail] = nag_correg_linregs_const(x, y, 'n', n)
[result, ifail] = g02ca(x, y, 'n', n)
```

3 Description

nag_correg_linregs_const (g02ca) fits a straight line of the form

$$y = a + bx$$

to the data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

such that

$$y_i = a + bx_i + e_i, \quad i = 1, 2, \ldots, n (n > 2).$$

The function calculates the regression coefficient, b, the regression constant, a (and various other statistical quantities) by minimizing

$$\sum_{i=1}^{n} e_i^2.$$

The input data consist of the n pairs of observations

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

on the independent variable x and the dependent variable y.

The quantities calculated are:

(a) Means:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

(b) Standard deviations:

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}; \quad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

(c) Pearson product-moment correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

(d) The regression coefficient, b, and the regression constant, a:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; a = \bar{y} - b\bar{x}.$$

(e) The sum of squares attributable to the regression, SSR, the sum of squares of deviations about the regression, SSD, and the total sum of squares, SST:

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2; \text{SSD} = \sum_{i=1}^{n}(y_i - a - bx_i)^2; \text{SSR} = \text{SST} - \text{SSD}.$$

(f) The degrees of freedom attributable to the regression, DFR, the degrees of freedom of deviations about the regression, DFD, and the total degrees of freedom, DFT:

$$\text{DFT} = n - 1;\ \text{DFD} = n - 2;\ \text{DFR} = 1.$$

(g) The mean square attributable to the regression, MSR, and the mean square of deviations about the regression, MSD:

$$\text{MSR} = \text{SSR}/\text{DFR}; \text{MSD} = \text{SSD}/\text{DFD}.$$

(h) The F value for the analysis of variance:

$$F = \text{MSR}/\text{MSD}.$$

(i) The standard error of the regression coefficient, $se(b)$, and the standard error of the regression constant, $se(a)$:

$$se(b) = \sqrt{\frac{\text{MSD}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}; \quad se(a) = \sqrt{\text{MSD}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}.$$

(j) The t value for the regression coefficient, $t(b)$, and the t value for the regression constant, $t(a)$:

$$t(b) = \frac{b}{se(b)}; \quad t(a) = \frac{a}{se(a)}.$$

4 References

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

5 Parameters

5.1 Compulsory Input Parameters

1: **x**(**n**) – REAL (KIND=nag_wp) array

x(i) must contain x_i, for $i = 1, 2, \ldots, n$.

2: **y**(**n**) – REAL (KIND=nag_wp) array

y(i) must contain y_i, for $i = 1, 2, \ldots, n$.

5.2 Optional Input Parameters

1: **n** – INTEGER

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

n, the number of pairs of observations.

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

5.3 Output Parameters

1: **result**(**20**) – REAL (KIND=nag_wp) array

The following information:

result(1)	$\bar{x}$, the mean value of the independent variable, x;
result(2)	$\bar{y}$, the mean value of the dependent variable, y;
result(3)	s_x the standard deviation of the independent variable, x;
result(4)	s_y the standard deviation of the dependent variable, y;
result(5)	r, the Pearson product-moment correlation between the independent variable x and the dependent variable y;
result(6)	b, the regression coefficient;
result(7)	a, the regression constant;
result(8)	$se(b)$, the standard error of the regression coefficient;
result(9)	$se(a)$, the standard error of the regression constant;
result(10)	$t(b)$, the t value for the regression coefficient;
result(11)	$t(a)$, the t value for the regression constant;
result(12)	SSR, the sum of squares attributable to the regression;
result(13)	DFR, the degrees of freedom attributable to the regression;
result(14)	MSR, the mean square attributable to the regression;
result(15)	F, the F value for the analysis of variance;
result(16)	SSD, the sum of squares of deviations about the regression;
result(17)	DFD, the degrees of freedom of deviations about the regression
result(18)	MSD, the mean square of deviations about the regression;
result(19)	SST, the total sum of squares;
result(20)	DFT, the total degrees of freedom.

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, $\mathbf{n} \leq 2$.

ifail $= 2$

On entry, all **n** values of at least one of the variables x and y are identical.

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

nag_correg_linregs_const (g02ca) does not use ***additional precision*** arithmetic for the accumulation of scalar products, so there may be a loss of significant figures for large n.

If, in calculating F, $t(a)$ or $t(b)$ (see Section 3), the numbers involved are such that the result would be outside the range of numbers which can be stored by the machine, then the answer is set to the largest quantity which can be stored as a double variable, by means of a call to nag_machine_real_largest (x02al).

8 Further Comments

The time taken by nag_correg_linregs_const (g02ca) depends on n.

The function uses a two-pass algorithm.

9 Example

This example reads in eight observations on each of two variables, and then performs a simple linear regression with the first variable as the independent variable, and the second variable as the dependent variable. Finally the results are printed.

9.1 Program Text

```
function g02ca_example

fprintf('g02ca example results\n\n');

x = [ 1.0   0.0    4.0    7.5   2.5   0.0  10.0   5.0];
y = [20.0  15.5   28.3   45.0  24.5  10.0  99.0  31.2];

n = numel(x);
fprintf('  i    independent(x)   dependent(y)\n');
fprintf('%3d%14.4f%14.4f\n',[1:n; x; y]);

[result, ifail] = g02ca(x, y);

fprintf('\n');
fprintf('Mean of independent variable              = %8.4f\n', result(1));
fprintf('Mean of   dependent variable              = %8.4f\n', result(2));
fprintf('Standard deviation of independent variable = %8.4f\n', result(3));
fprintf('Standard deviation of   dependent variable = %8.4f\n', result(4));
fprintf('Correlation coefficient                   = %8.4f\n', result(5));
fprintf('\n');
fprintf('Regression coefficient                    = %8.4f\n', result(6));
fprintf('Standard error of coefficient             = %8.4f\n', result(8));
fprintf('t-value for coefficient                   = %8.4f\n', result(10));
fprintf('\n');
fprintf('Regression constant                       = %8.4f\n', result(7));
fprintf('Standard error of constant                = %8.4f\n', result(9));
fprintf('t-value for constant                      = %8.4f\n', result(11));

fprintf('\nAnalysis of regression table :-\n\n');

fprintf('     Source       Sum of squares  D.F.    Mean square     F-value\n');
fprintf('Due to regression %11.3f%8d%14.3f%14.3f\n', result(12:15));
fprintf('About  regression %11.3f%8d%14.3f\n', result(16:18));
fprintf('Total             %11.3f%8d\n', result(19:20));
```

9.2 Program Results

```
     g02ca example results

  i    independent(x)   dependent(y)
  1         1.0000        20.0000
  2         0.0000        15.5000
  3         4.0000        28.3000
  4         7.5000        45.0000
  5         2.5000        24.5000
  6         0.0000        10.0000
  7        10.0000        99.0000
  8         5.0000        31.2000

Mean of independent variable               =   3.7500
Mean of   dependent variable               =  34.1875
Standard deviation of independent variable =   3.6253
Standard deviation of   dependent variable =  28.2604
Correlation coefficient                    =   0.9096

Regression coefficient                     =   7.0905
Standard error of coefficient              =   1.3224
t-value for coefficient                    =   5.3620

Regression constant                        =   7.5982
Standard error of constant                 =   6.6858
t-value for constant                       =   1.1365

Analysis of regression table :-

     Source       Sum of squares  D.F.    Mean square     F-value
Due to regression    4625.303       1      4625.303        28.751
About  regression     965.245       6       160.874
Total                5590.549       7
```