NAG Toolbox for MATLAB

g02ca

1 Purpose

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

2 Syntax

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

3 Description

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

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

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:

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

(b) Standard deviations:

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

(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, $se(a)$:

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

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

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

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) – double array**

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

2: **y(n) – double array**

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

5.2 Optional Input Parameters

1: **n – int32 scalar**

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 Input Parameters Omitted from the MATLAB Interface

None.

5.4 Output Parameters

1: **result(20) – double 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 – 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, $\mathbf{n} \leq 2$.

ifail $= 2$

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

7 Accuracy

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 x02al.

8 Further Comments

The time taken by g02ca depends on n.

The function uses a two-pass algorithm.

9 Example

```
x = [1;
     0;
     4;
     7.5;
     2.5;
     0;
     10;
     5];
y = [20;
     15.5;
     28.3;
     45;
     24.5;
     10;
     99;
     31.2];
[result, ifail] = g02ca(x, y)
```

```
result =
   1.0e+03 *
    0.0037
    0.0342
    0.0036
    0.0283
    0.0009
    0.0071
    0.0076
    0.0013
    0.0067
    0.0054
    0.0011
    4.6253
    0.0010
    4.6253
    0.0288
    0.9652
    0.0060
    0.1609
    5.5905
    0.0070
ifail =
           0
```