NAG Toolbox for MATLAB

g02cc

1 Purpose

g02cc performs a simple linear regression with dependent variable y and independent variable x, omitting cases involving missing values.

2 Syntax

```
[result, ifail] = g02cc(x, y, xmiss, ymiss, 'n', n)
```

3 Description

g02cc fits a straight line of the form

$$y = a + bx$$

to those of the data points

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

that do not include missing values, such that

$$y_i = a + bx_i + e_i$$

for those (x_i, y_i), $i = 1, 2, \dots, n \quad (n > 2)$ which do not include missing values.

The function eliminates all pairs of observations (x_i, y_i) which contain a missing value for either x or y, and then calculates the regression coefficient, b, the regression constant, a, and various other statistical quantities, by minimizing the sum of the e_i^2 over those cases remaining in the calculations.

The input data consists of the n pairs of observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ on the independent variable x and the dependent variable y.

In addition two values, xm and ym, are given which are considered to represent missing observations for x and y respectively. (See Section 7).

Let $w_i = 0$ if the ith observation of either x or y is missing, i.e., if $x_i = xm$ and/or $y_i = ym$; and $w_i = 1$ otherwise, for $i = 1, 2, \dots, n$.

The quantities calculated are:

(a) Means:

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

(b) Standard deviations:

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

(c) Pearson product-moment correlation coefficient:

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

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

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

(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:

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

(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:

$$DFT = \sum_{i=1}^{n} w_i - 1; \qquad DFD = \sum_{i=1}^{n} w_i - 2; \qquad DFR = 1.$$

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

$$MSR = SSR/DFR; \qquad MSD = SSD/DFD.$$

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

$$F = MSR/MSD.$$

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

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

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

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

(k) The number of observations used in the calculations:

$$n_c = \sum_{i=1}^{n} w_i.$$

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 $i = 1, 2, \ldots, n$.

3: **xmiss – double scalar**

The value xm which is to be taken as the missing value for the variable x. See Section 7.

4: **ymiss – double scalar**

The value ym which is to be taken as the missing value for the variable y. See Section 7.

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(21) – 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;
result(21)	n_c, the number of observations used in the calculations.

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 $= 2$

After observations with missing values were omitted, two or fewer cases remained.

ifail $= 3$

After observations with missing values were omitted, all remaining values of at least one of the variables x and y were identical.

7 Accuracy

g02cc does not use ***additional precision*** arithmetic for the accumulation of scalar products, so there may be a loss of significant figures for large n.

You are warned of the need to exercise extreme care in your selection of missing values. g02cc treats all values in the inclusive range $(1 \pm \mathrm{ACC}) \times xm_j$, where xm_j is the missing value for variable j specified by you, and ACC is a machine-dependent constant as missing values for variable j.

You must therefore ensure that the missing value chosen for each variable is sufficiently different from all valid values for that variable so that none of the valid values fall within the range indicated above.

If, in calculating F or $t(a)$ (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 g02cc depends on n and the number of missing observations.

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];
xmiss = 0;
```

```
ymiss = 99;
[result, ifail] = g02cc(x, y, xmiss, ymiss)
```

```
result =
    4.0000
   29.8000
    2.4749
    9.4787
    0.9799
    3.7531
   14.7878
    0.4409
    2.0155
    8.5128
    7.3370
  345.0940
    1.0000
  345.0940
   72.4682
   14.2860
    3.0000
    4.7620
  359.3800
    4.0000
    5.0000
ifail =
           0
```