NAG Toolbox for MATLAB

g02cg

1 Purpose

g02cg performs a multiple linear regression on a set of variables whose means, sums of squares and cross-products of deviations from means, and Pearson product-moment correlation coefficients are given.

2 Syntax

```
[result, coef, con, rinv, c, ifail] = g02cg(n, k, xbar, ssp, r, 'k1', k1)
```

3 Description

g02cg fits a curve of the form

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

to the data points

$$\begin{matrix}(x_{11}, x_{21}, \ldots, x_{k1}, y_1)\\ (x_{12}, x_{22}, \ldots, x_{k2}, y_2)\\ \vdots\\ (x_{1n}, x_{2n}, \ldots, x_{kn}, y_n)\end{matrix}$$

such that

$$y_i = a + b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki} + e_i, \qquad i = 1, 2, \ldots, n.$$

The function calculates the regression coefficients, $b_1, b_2, \ldots, b_k$, the regression constant, a, and various other statistical quantities by minimizing

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

The actual data values $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ are not provided as input to the function. Instead, input consists of:

(i) The number of cases, n, on which the regression is based.

(ii) The total number of variables, dependent and independent, in the regression, $(k + 1)$.

(iii) The number of independent variables in the regression, k.

(iv) The means of all $k + 1$ variables in the regression, both the independent variables $(x_1, x_2, \ldots, x_k)$ and the dependent variable (y), which is the $(k + 1)$th variable: i.e., $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y}$.

(v) The $(k + 1)$ by $(k + 1)$ matrix $[S_{ij}]$ of sums of squares and cross-products of deviations from means of all the variables in the regression; the terms involving the dependent variable, y, appear in the $(k + 1)$th row and column.

(vi) The $(k + 1)$ by $(k + 1)$ matrix $[R_{ij}]$ of the Pearson product-moment correlation coefficients for all the variables in the regression; the correlations involving the dependent variable, y, appear in the $(k + 1)$th row and column.

The quantities calculated are:

(a) The inverse of the k by k partition of the matrix of correlation coefficients, $[R_{ij}]$, involving only the independent variables. The inverse is obtained using an accurate method which assumes that this sub-matrix is positive-definite.

(b) The modified inverse matrix, $C = \left[c_{ij}\right]$, where

$$c_{ij} = \frac{R_{ij}r_{ij}}{S_{ij}}, \qquad i,j = 1,2,\ldots,k,$$

where r_{ij} is the (i,j)th element of the inverse matrix of $[R_{ij}]$ as described in above. Each element of C is thus the corresponding element of the matrix of correlation coefficients multiplied by the corresponding element of the inverse of this matrix, divided by the corresponding element of the matrix of sums of squares and cross-products of deviations from means.

(c) The regression coefficients:

$$b_i = \sum_{j=i}^{k} c_{ij}S_{j(k+1)}, \qquad i = 1,2,\ldots,k,$$

where $S_{j(k+1)}$ is the sum of cross-products of deviations from means for the independent variable x_j and the dependent variable y.

(d) 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:

$SST = S_{(k+1)(k+1)}$, the sum of squares of deviations from the mean for the dependent variable, y;

$$SSR = \sum_{j=1}^{k} b_j S_{j(k+1)}; \qquad SSD = SST - SSR$$

(e) 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:

$$DFR = k; \qquad DFD = n - k - 1; \qquad DFT = n - 1.$$

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

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

(g) The F values for the analysis of variance:

$$F = MSR/MSD.$$

(h) The standard error estimate:

$$s = \sqrt{MSD}.$$

(i) The coefficient of multiple correlation, R, the coefficient of multiple determination, R^2 and the coefficient of multiple determination corrected for the degrees of freedom, $\bar{R}^2$;

$$R = \sqrt{1 - \frac{SSD}{SST}}; \qquad R^2 = 1 - \frac{SSD}{SST}; \qquad \bar{R}^2 = 1 - \frac{SSD \times DFT}{SST \times DFD}.$$

(j) The standard error of the regression coefficients:

$$se(b_i) = \sqrt{MSD \times c_{ii}}, \qquad i = 1,2,\ldots,k.$$

(k) The t values for the regression coefficients:

$$t(b_i) = \frac{b_i}{se(b_i)}, \qquad i = 1, 2, \ldots, k.$$

(l) The regression constant, a, its standard error, $se(a)$, and its t value, $t(a)$:

$$a = \bar{y} - \sum_{i=1}^{k} b_i \bar{x}_i; \qquad se(a) = \sqrt{MSD \times \left(\frac{1}{n} + \sum_{i=1}^{k} \sum_{j=1}^{k} \bar{x}_i c_{ij} \bar{x}_j \right)}; \qquad t(a) = \frac{a}{se(a)}.$$

4 References

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

5 Parameters

5.1 Compulsory Input Parameters

1: **n – int32 scalar**

The number of cases n, used in calculating the sums of squares and cross-products and correlation coefficients.

2: **k – int32 scalar**

the number of independent variables k, in the regression.

Constraint: **k** = **k1** − 1.

3: **xbar(k1) – double array**

xbar(i) must be set to $\bar{x}_i$, the mean value of the ith variable, for $i = 1, 2, \ldots, k+1$; the mean of the dependent variable must be contained in **xbar**$(k+1)$.

4: **ssp(ldssp,k1) – double array**

ldssp, the first dimension of the array, must be at least **k1**.

ssp(i,j) must be set to S_{ij}, the sum of cross-products of deviations from means for the ith and jth variables, for $i,j = 1, 2, \ldots, k+1$; terms involving the dependent variable appear in row $k+1$ and column $k+1$.

5: **r(ldr,k1) – double array**

ldr, the first dimension of the array, must be at least **k1**.

r(i,j) must be set to R_{ij}, the Pearson product-moment correlation coefficient for the ith and jth variables, for $i,j = 1, 2, \ldots, k+1$; terms involving the dependent variable appear in row $k+1$ and column $k+1$.

5.2 Optional Input Parameters

1: **k1 – int32 scalar**

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

the total number of variables, independent and dependent, $(k+1)$, in the regression.

Constraint: $2 \leq$ **k1** $<$ **n**.

5.3 Input Parameters Omitted from the MATLAB Interface

ldssp, ldr, ldcoef, ldrinv, ldc, wkz, ldwkz

5.4 Output Parameters

1: **result**(**13**) – **double array**

The following information:

result(1) SSR, the sum of squares attributable to the regression;
result(2) DFR, the degrees of freedom attributable to the regression;
result(3) MSR, the mean square attributable to the regression;
result(4) F, the F value for the analysis of variance;
result(5) SSD, the sum of squares of deviations about the regression;
result(6) DFD, the degrees of freedom of deviations about the regression;
result(7) MSD, the mean square of deviations about the regression;
result(8) SST, the total sum of squares;
result(9) DFT, the total degrees of freedom;
result(10) s, the standard error estimate;
result(11) R, the coefficient of multiple correlation;
result(12) R^2, the coefficient of multiple determination;
result(13) $\bar{R}^2$, the coefficient of multiple determination corrected for the degrees of freedom.

2: **coef**(**ldcoef,3**) – **double array**

For $i = 1, 2, \ldots, k$, the following information:

coef$(i, 1)$

b_i, the regression coefficient for the ith variable.

coef$(i, 2)$

$se(b_i)$, the standard error of the regression coefficient for the ith variable.

coef$(i, 3)$

$t(b_i)$, the t value of the regression coefficient for the ith variable.

3: **con**(**3**) – **double array**

The following information:

con(1) a, the regression constant;
con(2) $se(a)$, the standard error of the regression constant;
con(3) $t(a)$, the t value for the regression constant.

4: **rinv**(**ldrinv,k**) – **double array**

The inverse of the matrix of correlation coefficients for the independent variables; that is, the inverse of the matrix consisting of the first k rows and columns of **r**.

5: **c**(**ldc,k**) – **double array**

The modified inverse matrix, where

$$\mathbf{c}(i,j) = \mathbf{r}(i,j) \times \mathbf{rinv}(i,j)/\mathbf{ssp}(i,j), \text{ for } i,j = 1, 2, \ldots, k.$$

6: **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, **k1** < 2.

ifail = 2

On entry, **k1** $\neq$ (**k** + 1).

ifail = 3

On entry, **n** $\leq$ **k1**.

ifail = 4

On entry, **ldssp** $<$ **k1**,
or **ldr** $<$ **k1**,
or **ldcoef** $<$ **k**,
or **ldrinv** $<$ **k**,
or **ldc** $<$ **k**,
or **ldwkz** $<$ **k**.

ifail = 5

The k by k partition of the matrix R which is to be inverted is not positive-definite.

ifail = 6

The refinement following the actual inversion fails, indicating that the k by k partition of the matrix R, which is to be inverted, is ill-conditioned. The use of g02da, which employs a different numerical technique, may avoid this difficulty (an extra 'variable' representing the constant term must be introduced for g02da).

ifail = 7

Unexpected error in f04ab.

7 Accuracy

The accuracy of any regression function is almost entirely dependent on the accuracy of the matrix inversion method used. In g02cg, it is the matrix of correlation coefficients rather than that of the sums of squares and cross-products of deviations from means that is inverted; this means that all terms in the matrix for inversion are of a similar order, and reduces the scope for computational error. For details on absolute accuracy, the relevant section of the document describing the inversion function used, f04ab, should be consulted. g02da uses a different method, based on f04am, and that function may well prove more reliable numerically. It does not handle missing values, nor does it provide the same output as this function. (In particular it is necessary to include explicitly the constant in the regression equation as another 'variable'.)

If, in calculating F, $t(a)$, or any of the $t(b_i)$ (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 g02cg depends on k.

This function assumes that the matrix of correlation coefficients for the independent variables in the regression is positive-definite; it fails if this is not the case.

This correlation matrix will in fact be positive-definite whenever the correlation matrix and the sums of squares and cross-products (of deviations from means) matrix have been formed either without regard to missing values, or by eliminating **completely** any cases involving missing values, for any variable. If, however, these matrices are formed by eliminating cases with missing values from only those calculations involving the variables for which the values are missing, no such statement can be made, and the correlation matrix may or may not be positive-definite. You should be aware of the possible dangers of using correlation matrices formed in this way (see the G02 Chapter Introduction), but if they nevertheless

wish to carry out regression using such matrices, this function is capable of handling the inversion of such matrices provided they are positive-definite.

If a matrix is positive-definite, its subsequent re-organisation by either g02ce or g02cf will not affect this property, and the new matrix can safely be used in this function. Thus correlation matrices produced by any of g02ba, g02bb, g02bg or g02bh, even if subsequently modified by either g02ce or g02cf, can be handled by this function.

It should be noted that in forming the sums of squares and cross-products matrix and the correlation matrix a column of constants should **not** be added to the data as an additional 'variable' in order to obtain a constant term in the regression. This function automatically calculates the regression constant, a, and any attempt to insert such a 'dummy variable' is likely to cause the function to fail.

It should also be noted that the function requires the dependent variable to be the last of the $k+1$ variables whose statistics are provided as input to the function. If this variable is not correctly positioned in the original data, the means, standard deviations, sums of squares and cross-products of deviations from means, and correlation coefficients can be manipulated by using g02ce or g02cf to reorder the variables as necessary.

9 Example

```
n = int32(5);
k = int32(2);
xbar = [5.4;
     5.8;
     2.8];
ssp = [99.2, -57.6, 6.4;
     -57.6, 102.8, -29.2;
     6.4, -29.2, 14.8];
r = [1, -0.5704, 0.167;
     -0.5704, 1, -0.7486;
     0.167, -0.7486, 1];
[result, coeff, con, rinv, c, ifail] = g02cg(n, k, xbar, ssp, r)
```

```
result =
    9.7769
    2.0000
    4.8884
    1.9464
    5.0231
    2.0000
    2.5116
   14.8000
    4.0000
    1.5848
    0.8128
    0.6606
    0.3212
coeff =
   -0.1488    0.1937   -0.7683
   -0.3674    0.1903   -1.9309
con =
    5.7350
    2.0327
    2.8213
rinv =
    1.4823    0.8455
    0.8455    1.4823
c =
    0.0149    0.0084
    0.0084    0.0144
ifail =
           0
```