NAG Toolbox for MATLAB

g02ch

1   Purpose

g02ch performs a multiple linear regression with no constant on a set of variables whose sums of squares and cross-products about zero and correlation-like coefficients are given.

2   Syntax

```
[result, coef, rznv, cz, ifail] = g02ch(n, k, sspz, rz, 'k1', k1)
```

3   Description

g02ch fits a curve of the form

$$y = b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

to the data points

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

such that

$$y_i = b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i, \qquad i = 1, 2, \ldots, n.$$

The function calculates the regression coefficients, $b_1, b_2, \ldots, b_k$, (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 to the function 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 $(k + 1)$ by $(k + 1)$ matrix $\left[\tilde{S}_{ij}\right]$ of sums of squares and cross-products about zero of all the variables in the regression; the terms involving the dependent variable, y, appear in the $(k + 1)$th row and column.

(v)   The $(k + 1)$ by $(k + 1)$ matrix $\left[\tilde{R}_{ij}\right]$ of correlation-like 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-like coefficients, $\left[\tilde{R}_{ij}\right]$, involving only the independent variables. The inverse is obtained using an accurate method which assumes that this sub-matrix is positive-definite (see Section 8).

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

$$c_{ij} = \frac{\tilde{R}_{ij} \tilde{r}^{ij}}{\tilde{S}_{ij}}, \qquad i, j = 1, 2, \ldots, k,$$

where $\tilde{r}^{ij}$ is the (i,j)th element of the inverse matrix of $\left[\tilde{R}_{ij}\right]$ as described in above. Each element of C is thus the corresponding element of the matrix of correlation-like 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 about zero.

(c) The regression coefficients:

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

where $\tilde{S}_{j(k+1)}$ is the sum of cross-products about zero 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:

$SST = \tilde{S}_{(k+1)(k+1)}$, the sum of squares about zero for the dependent variable, y;

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

(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:

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

(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 value 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:

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

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**

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

2: **k – int32 scalar**

the number of independent variables k in the regression.

Constraint: $\mathbf{k} = \mathbf{k1} - 1$.

3: **sspz(ldsspz,k1) – double array**

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

$\mathbf{sspz}(i,j)$ must be set to $\tilde{S}_{ij}$, the sum of cross-products about zero 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 $k + 1$.

4: **rz(ldrz,k1) – double array**

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

$\mathbf{rz}(i,j)$ must be set to $\tilde{R}_{ij}$, the correlation-like coefficient for the ith and jth variables, for $i,j = 1, 2, \ldots, k + 1$; coefficients 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 **sspz**, **irz**. (An error is raised if these dimensions are not equal.)

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

Constraint: $2 \leq \mathbf{k1} \leq \mathbf{n}$.

5.3 Input Parameters Omitted from the MATLAB Interface

ldsspz, ldrz, ldcoef, ldrznv, ldcz, 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:   **rznv**(**ldrznv,k**) – **double array**

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

4:   **cz**(**ldcz,k**) – **double array**

   The modified inverse matrix, C, where

$$\mathbf{cz}(i,j) = \frac{\mathbf{rz}(i,j) \times \mathbf{rznv}(i,j)}{\mathbf{sspz}(i,j)}, \qquad i,j = 1, 2, \ldots, k.$$

5:   **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** $<$ **k1**.

ifail $= 4$

   On entry, **ldsspz** $<$ **k1**,
   or   **ldrz** $<$ **k1**,
   or   **ldcoef** $<$ **k**,
   or   **ldrznv** $<$ **k**,
   or   **ldcz** $<$ **k**,
   or   **ldwkz** $<$ **k**.

ifail $= 5$

   This indicates that the k by k partition of the matrix RZ, which is to be inverted, is not positive-definite.

ifail $= 6$

   This indicates that the refinement following the actual inversion fails, indicating that the k by k partition of the matrix RZ, which is to be inverted, is ill-conditioned. The use of g02da, which employs a different numerical technique, may avoid the difficulty.

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 g02ch, it is the matrix of correlation-like coefficients rather than that of the sums of squares and cross-products about zero 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.

If, in calculating F 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 g02ch depends on k.

This function assumes that the matrix of correlation-like 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-like matrix and the sums of squares and cross-products (about zero) 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-like 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 regressions 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 of 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 g02bd, g02be, g02bk or g02bl, even if subsequently modified by either g02ce or g02cf, can be handled by this function.

It should 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 about zero, and correlation-like coefficients can be manipulated by using g02ce or g02cf to reorder the variables as necessary.

9   Example

```
n = int32(5);
k = int32(2);
sspz = [245, 99, 82;
        99, 271, 52;
        82, 52, 54];
rz = [1, 0.3842, 0.7129;
      0.3842, 1, 0.4299;
      0.7129, 0.4299, 1];
[result, coeff, rzinv, cz, ifail] = g02ch(n, k, sspz, rz)

result =
   28.9857
    2.0000
```

```
      14.4929
       1.7382
      25.0143
       3.0000
       8.3381
      54.0000
       5.0000
       2.8876
       0.7326
       0.5368
       0.2280
coeff =
       0.3017      0.1998      1.5098
       0.0817      0.1900      0.4299
rzinv =
       1.1732     -0.4507
      -0.4507      1.1732
cz =
       0.0048     -0.0017
      -0.0017      0.0043
ifail =
            0
```