NAG Toolbox

nag_correg_ssqmat (g02bu)

1 Purpose

nag_correg_ssqmat (g02bu) calculates the sample means and sums of squares and cross-products, or sums of squares and cross-products of deviations from the mean, in a single pass for a set of data. The data may be weighted.

2 Syntax

```
[sw, wmean, c, ifail] = nag_correg_ssqmat(x, 'mean', mean, 'n', n, 'm', m, 'wt', wt)
[sw, wmean, c, ifail] = g02bu(x, 'mean', mean, 'n', n, 'm', m, 'wt', wt)
```

Note: the interface to this routine has changed since earlier releases of the toolbox:

At Mark 24: **mean** was made optional

At Mark 22: **n** was made optional.

3 Description

nag_correg_ssqmat (g02bu) is an adaptation of West's WV2 algorithm; see West (1979). This function calculates the (optionally weighted) sample means and (optionally weighted) sums of squares and cross-products or sums of squares and cross-products of deviations from the (weighted) mean for a sample of n observations on m variables X_j, for $j = 1, 2, \ldots, m$. The algorithm makes a single pass through the data.

For the first $i-1$ observations let the mean of the jth variable be $\bar{x}_j(i-1)$, the cross-product about the mean for the jth and kth variables be $c_{jk}(i-1)$ and the sum of weights be W_{i-1}. These are updated by the ith observation, x_{ij}, for $j = 1, 2, \ldots, m$, with weight w_i as follows:

$$\begin{aligned} W_i &= W_{i-1} + w_i \\ \bar{x}_j(i) &= \bar{x}_j(i-1) + \tfrac{w_i}{W_i}\big(x_j - \bar{x}_j(i-1)\big), \quad j = 1, 2, \ldots, m \end{aligned}$$

and

$$c_{jk}(i) = c_{jk}(i-1) + \frac{w_i}{W_i}\big(x_j - \bar{x}_j(i-1)\big)\big(x_k - \bar{x}_k(i-1)\big)W_{i-1}, \quad j = 1, 2, \ldots, m \text{ and } k = j, j+1, \ldots, m.$$

The algorithm is initialized by taking $\bar{x}_j(1) = x_{1j}$, the first observation, and $c_{ij}(1) = 0.0$.

For the unweighted case $w_i = 1$ and $W_i = i$ for all i.

Note that only the upper triangle of the matrix is calculated and returned packed by column.

4 References

Chan T F, Golub G H and Leveque R J (1982) *Updating Formulae and a Pairwise Algorithm for Computing Sample Variances* Compstat, Physica-Verlag

West D H D (1979) Updating mean and variance estimates: An improved method *Comm. ACM* **22** 532–555

5 Parameters

5.1 Compulsory Input Parameters

1: **x**$(ldx, \mathbf{m})$ – REAL (KIND=nag_wp) array

ldx, the first dimension of the array, must satisfy the constraint $ldx \geq \mathbf{n}$.

x(i, j) must contain the ith observation on the jth variable, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$.

5.2 Optional Input Parameters

1: **mean_p** – CHARACTER(1)

Default: 'M'

Indicates whether nag_correg_ssqmat (g02bu) is to calculate sums of squares and cross-products, or sums of squares and cross-products of deviations about the mean.

mean = 'M'
The sums of squares and cross-products of deviations about the mean are calculated.

mean = 'Z'
The sums of squares and cross-products are calculated.

Constraint: **mean** = 'M' or 'Z'.

2: **n** – INTEGER

Default: the first dimension of the array **x**.

n, the number of observations in the dataset.

Constraint: $\mathbf{n} \geq 1$.

3: **m** – INTEGER

Default: the second dimension of the array **x**.

m, the number of variables.

Constraint: $\mathbf{m} \geq 1$.

4: **wt**(:) – REAL (KIND=nag_wp) array

The dimension of the array **wt** must be at least **n** if $weight$ = 'W', and at least 1 otherwise

The optional weights of each observation.

If $weight$ = 'U', **wt** is not referenced.

If $weight$ = 'W', **wt**(i) must contain the weight for the ith observation.

Constraint: if $weight$ = 'W', **wt**$(i) \geq 0.0$, for $i = 1, 2, \ldots, n$.

5.3 Output Parameters

1: **sw** – REAL (KIND=nag_wp)

The sum of weights.

If $weight$ = 'U', **sw** contains the number of observations, n.

2: **wmean**$(\mathbf{m})$ – REAL (KIND=nag_wp) array

The sample means. **wmean**(j) contains the mean for the jth variable.

3: **c**$((\mathbf{m} \times \mathbf{m} + \mathbf{m})/\mathbf{2})$ – REAL (KIND=nag_wp) array

The cross-products.

If **mean** = 'M', **c** contains the upper triangular part of the matrix of (weighted) sums of squares and cross-products of deviations about the mean.

If **mean** = 'Z', **c** contains the upper triangular part of the matrix of (weighted) sums of squares and cross-products.

These are stored packed by columns, i.e., the cross-product between the jth and kth variable, $k \geq j$, is stored in $\mathbf{c}(k \times (k-1)/2 + j)$.

4: **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, **m** < 1,
or **n** < 1,
or $ldx <$ **n**.

ifail = 2

On entry, **mean** ≠ 'M' or 'Z'.

ifail = 3

On entry, $weight \neq$ 'W' or 'U'.

ifail = 4

On entry, $weight =$ 'W', and a value of **wt** < 0.0.

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

For a detailed discussion of the accuracy of this algorithm see Chan *et al.* (1982) or West (1979).

8 Further Comments

nag_correg_ssqmat_to_corrmat (g02bw) may be used to calculate the correlation coefficients from the cross-products of deviations about the mean. The cross-products of deviations about the mean may be scaled using to give a variance-covariance matrix.

The means and cross-products produced by nag_correg_ssqmat (g02bu) may be updated by adding or removing observations using nag_correg_ssqmat_update (g02bt).

Two sets of means and cross-products, as produced by nag_correg_ssqmat (g02bu), can be combined using nag_correg_ssqmat_combine (g02bz).

9 Example

A program to calculate the means, the required sums of squares and cross-products matrix, and the variance matrix for a set of 3 observations of 3 variables.

9.1 Program Text

```
     function g02bu_example

fprintf('g02bu example results\n\n');

wt = [0.1300  1.3070  0.3700];
x  = [9.1231  0.9310  0.0009;
      3.7011  0.0900  0.0099;
      4.5230  0.8870  0.0999];
[m,n] = size(x);
cn = (m*(m+1))/2;
m = nag_int(m);

[sw, wmean, c, ifail] = g02bu(x', 'wt', wt);

disp('Means');
disp(wmean');
disp('Weights');
disp(wt);

mtitle = 'Sums of squares and cross-products:';
uplo   = 'Upper';
diag   = 'Non-unit';
[ifail] = x04cc( ...
                 uplo, diag, m, c, mtitle);

% Convert the sums of squares and cross-products to a variance matrix
v = c/(sw-1);
fprintf('\n');
mtitle = 'Variance matrix:';
[ifail] = x04cc( ...
                 uplo, diag, m, v, mtitle);
```

9.2 Program Results

```
     g02bu example results

Means
    1.3299    0.3334    0.9874

Weights
    0.1300    1.3070    0.3700

 Sums of squares and cross-products:
             1          2          3
 1      8.7569     3.6978     4.0707
 2                 1.5905     1.6861
 3                            1.9297

 Variance matrix:
             1          2          3
 1     10.8512     4.5822     5.0443
 2                 1.9709     2.0893
 3                            2.3912
```