NAG Toolbox for MATLAB

g02bu

1 Purpose

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] = g02bu(mean, n, x, 'm', m, 'wt', wt)
```

3 Description

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:

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

and

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

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: **mean – string**

Indicates whether 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 – int32 scalar**

n, the number of observations in the data set.

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

3: **x(ldx,m) – double array**

ldx, the first dimension of the array, must be at least **n**.

$\mathbf{x}(i,j)$ 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: **m – int32 scalar**

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

m, the number of variables.

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

2: **wt(∗) – double array**

Note: 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', $\mathbf{wt}(i)$ must contain the weight for the ith observation.

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

5.3 Input Parameters Omitted from the MATLAB Interface

weight, ldx

5.4 Output Parameters

1: **sw – double scalar**

The sum of weights.

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

2: **wmean(m) – double array**

The sample means. $\mathbf{wmean}(j)$ contains the mean for the jth variable.

3: $\mathbf{c}((\mathbf{m} \times \mathbf{m} + \mathbf{m})/\mathbf{2})$ **– double 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 – 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, **m** < 1,
or **n** < 1,
or **ldx** < **n**.

ifail = 2

On entry, **mean** ≠ 'M' or 'Z'.

ifail = 3

On entry, **weight** ≠ 'W' or 'U'.

ifail = 4

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

7 Accuracy

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

8 Further Comments

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 Missing 'id' to give a variance-covariance matrix.

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

9 Example

```
mean = 'M';
n = int32(3);
wt = [0.13, 1.307, 0.37];
x = [9.123100000000001, 3.7011, 4.523;
     0.931, 0.09, 0.887;
     0.0009, 0.009900000000000001, 0.0999];
[sw, wmean, c, ifail] = g02bu(mean, n, x, 'wt', wt)
```

```
sw =
    1.8070
wmean =
    1.3299
    0.3334
    0.9874
c =
    8.7569
    3.6978
    1.5905
    4.0707
    1.6861
    1.9297
ifail =
           0
```