Data Imputation: nagdmc_impute_simp

Purpose

nagdmc_impute_simp imputes data values based on summary statistics of variables.

nagdmc_impute_simp returns an array containing the indexes of imputed values (see 'Explanatory Code'). The memory used by this array should be returned to the operating system by the user as indicated in the Essential Introduction.

Declaration

```
#include <nagdmc.h>
long *nagdmc_impute_simp(long rec1, long nvar, long nrec, long dblk,
                         double data[], long ncat[], long maxcat, long cat[],
                         double mval, double ival[], long *nrepl, int *info);
```

Parameters

1: **rec1** – `long` *Input*

On entry: the index in the data of the first data record used in the analysis.

Constraint: **rec1** ≥ 0.

2: **nvar** – `long` *Input*

On entry: the number of variables in the data.

Constraint: **nvar** ≥ 1.

3: **nrec** – `long` *Input*

On entry: the number of consecutive records, beginning at **rec1**, used in the analysis.

Constraint: **nrec** > 1.

4: **dblk** – `long` *Input*

On entry: the total number of records in the data block.

Constraint: **dblk** $\geq$ **rec1** + **nrec**.

5: **data**[**dblk** $*$ **nvar**] – `double` *Input/Output*

On entry: data values for the jth variable (for $j = 0, 1, \ldots,$ **nvar** $- 1$) are stored in **data**[$i *$ **nvar** $+ j$], for $i = 0, 1, \ldots,$ **dblk** $- 1$.

On exit: missing values in **data** are replaced by their estimates.

6: **ncat**[**nvar**] – `long` *Input*

On entry: **ncat**[i] contains the number of categories on the ith variable, for $i = 0, 1, \ldots,$ **nvar** $- 1$. If the ith variable is continuous, **ncat**[i] must be set equal to zero. If all variables in the analysis are continuous, **ncat** must be 0.

Constraints: if **ncat** is not 0, **ncat**[i] ≥ 0, for $i = 0, 1, \ldots,$ **nvar** $- 1$.

7: **maxcat** – `long` *Input*

On entry: the maximum number of categories on any categorical variable. If all variables are continuous, **maxcat** must be 0.

Constraints: **maxcat** ≥ 0, and **maxcat** $\geq$ **ncat**[i], for $i = 0, 1, \ldots,$ **nvar** $- 1$.

8: **cat**[**nvar***maxcat] – `long` *Input*

On entry: the categories for the categorical variables. The categories for the ith variable are stored in **cat**[$i *$ **maxcat** $+ j$], for $j = 0, 1, \ldots,$ **ncat**[i] $- 1$. If all variables in the analysis are continuous, **cat** must be 0.

Constraint: if **ncat** is 0, **cat** $= 0$.

9: **mval** – `double` *Input*

On entry: all values in **data** equal within machine precision to **mval** are considered missing from the analysis.

Suggested value: a value outside the interval $[a, b]$, where a and b are the minimum and maximum value in your data, respectively.

10: **ival[ncat]** – `double` *Output*

On exit: **ival**$[i]$ contains the value used to replace missing values on the ith variable, for $i = 0, 1, \ldots,$ **nvar** $- 1$.

11: **nrepl** – `long *` *Output*

On exit: the number of **mval** values replaced by the function. The value of this parameter determines the length of the array returned by the function, see the 'Explanatory Code' for details.

12: **info** – `int *` *Output*

On exit: **info** gives information on the success of the function call:

0: the function successfully completed its task.

i; $i = 1, 2, 3, 4, 6, 7, 8$: the specification of the ith formal parameter was incorrect.

20: no missing values were found in the data; check your definition of **mval**.

99: the function failed to allocate enough memory.

100: an internal error occurred during the execution of the function.

Notation

nrec the number of data records, n.
data the data set X.
nxvar determines the number of variables, m.
mval the value of missing data values, z.

Description

Let X be a set of n data records x_i, for $i = 1, 2, \ldots, n$. The jth variable of the ith data record takes either the value x_{ij} or a dummy value, z, representing values in X missing at random, for $j = 1, 2, \ldots, m$.

Any missing values in the ith data record are replaced by the mean of the jth variable in the data:

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij},$$

if j is a continuous variable or by the category value with the highest frequency on j in X if j is a categorical variable, for $i = 1, 2, \ldots, n$.

References and Further Reading

None.

Explanatory Code

The following C function prints the index in the data and the imputed value of each of the **nrepl** replaced values after a successful call to **nagdmc_impute_simp** returning `indexes`.

```
#include <stdio.h>

void imputed_values(long nvar, double data[], long nrepl, long indexes[]) {
    long i, j, k;

#define MISSING_ROW(I) indexes[I]
#define MISSING_COL(I) indexes[I+nrepl]
#define DATA(I,J) data[(I)*nvar+J]
```

```
    printf("\n\tRow \tCol \tValue\n");
    for (i=0; i<nrepl; ++i) {
        j = MISSING_ROW(i);
        k = MISSING_COL(i);
        printf("\t%-4li\t%-4li\t%-8.4f\n",j,k,DATA(j,k));
    }
}
```

See Also

nagdmc_free_impute	returns memory allocated by **nagdmc_impute_simp** to the operating system.
impute_simp_ex.c	the example calling program.