Euredit general imputation function: gen_imp

1 Purpose

gen_imp general distance-based imputation method.

2 Specification

```
#include <euredit_sys.h>


void gen_imp (long srtype[3], long rand_select, long n_match, long match_all,
              long reuse, double min_wt, double *wts, double *R,
              long *dist_type, double *user_dist, long num_obs, long m,
              double *data, long *n_cat, long *cat_val, long maxcat,
              double miss_val, long *num_imp, long *obs_num, long *var_num,
              double *rep_val, long info[5])
```

3 Parameters

srtype

Input: controls the search strategy. If **srtype[0]** < 0 then only donors with the same value of categorical variable **srtype[1]** are searched.

If **srtype[0]** $<= 0$ then all donors are searched.

If **srtype[0]** > 0 then only donors between records **srtype[1]** and are **srtype[2]** searched.

rand_select

Input: indicates what random selection from donors is to be performed.

If **rand_select** < 0 **rand_select** is used as the repeatable seed to start the random selection.

If **rand_select** $= 0$ random selection is not used and the first donor is used.

If **rand_select** > 0 the seed random selection is taken from the system clock.

n_match

Input: the number of best donors to be retained. If **rand_select** $\neq 0$ the donor will be selected from the first **n_match** best donors encountered.

Constraint: **n_match** ≥ 1.

match_all

Input: if **match_all** $= 0$ all missing values for a record are to be imputed otherwise only the **match_all** variable will be imputed.

Constraint: **match_all** ≥ 0.

reuse

Input: if **reuse** $\neq 0$ a donor may be re-used for different records. If **reuse** $= 0$ a donor is penalised after having been used.

min_wt

Input: the minimum acceptable sum of weights for matching. That is, the sum of weights for non-missing values of the donor must be $\geq$ **min_wt**.

Constraint: **min_wt** ≥ 0.0.

wts[m]

Input: the weights to be used in calculating the distance between donors and recipients.

Constraints: $\mathbf{wts}[i] \geq 0.0$; $\sum_{i=1}^{m} \mathbf{wts}[i] \geq 0$.

R[m]

Input: the distance coefficients for continuous variables. For Euclidean and Manhattan distances $\mathbf{R}[i]$ is a multiplicative standardisation. For Threshold distances it is the threshold and for Regression distances it is the regression coefficient. If all variables are categorical, **R** is not referenced and may be set equal to NULL.

dist_type[m]

Input: the distance measure used.

For continuous variables:

1 - Manhattan

2 - Regression

3 - Threshold

Otherwise - Euclidean

For categorical variables:

1 - Rank difference

2 - User defined

Otherwise - Simple matching

user_dist[m*maxcat*maxcat]

Input: the distance tables for user-supplied distances. The **m n_cat**[i] by **n_cat**[i] tables are stored in **m maxcat*maxcat** blocks. If the user-defined distance option for categorical variables is not selected, **user_dist** may be set equal to NULL.

num_obs

Input: the number of observations in the data.

Constraint: **num_obs** ≥ 1.

m

Input: the number of variables in the data.

Constraint: **m** ≥ 1.

data[n*m]

Input: the data stored by row.

n_cat[m]

Input: if the ith variable is categorical, **n_cat**[i] must be set to the number of categories present; otherwise set **n_cat**[i] equal to zero.

cat_val[m*maxcat]

Input: the categories for the categorical variables. The categories for the ith variable are stored in **cat_val**[i***maxcat**$+j$], for $j = 1, 2, \ldots,$ **n_cat**[i].

maxcat

Input: the maximum number of categories in any categorical variable.

miss_val

Input: the missing value indicator.

num_imp

Output: the number of values replaced.

obs_num[num_imp]

Output: the observation number for a replacement value.

var_num[num_imp]

Output: the variable number for a replacement value.

rep_val[num_imp]

Output: the replacement value.

info

Output: information on the success of the function call.
info[0] = 0: the function successfully completed its task.
info[0] = i: the specification of the ith formal parameter was incorrect, $i = 1, 2, \ldots, 22$.
info[0] = 50: a category for the **srtype[1]** variable is incorrect.
info[0] = 55: a category value is incorrect.
info[0] = 99: the function failed to allocate enough memory.
info[1] contains additional information for system debugging.

info[2] contains number of records requiring imputation such that sum of weights for non-missing values is less then **min_wt**.
info[3] contains number of records requiring imputation with missing values for variables with non-zero weights.
info[4] contains number of cases for which no donors were available.