缺失数据的多重插补

本模块使用 R 的 MI 程序包进行多重插补。多重插补（MI，Multiple imputation，Rubin 1987，1996）是用于填补复杂数据缺失值的一种插补方法。该模块使用链式算法（A chained equation approach， Van Buuren and Oudshoorn 2000; Raghunathan, Lepkowski, Van Hoewyk, and Solenberger 2001）用户指定每个变量的条件分布，缺失值是基于数据中其它变量的分布，使用连续迭代插补方法插补缺失值。

该模块输出默认 5 套（可选）插补后完整的数据，存成一个新文件，有一个指示变量表示第几套插补。用户可以使用相应的回归模型分别对这 5 套数据进行分析，得出 5 套回归模型，再使用工具菜单下的统计工具“合并从多重插补数据得出的回归系数”功能，对 5 套回归模型的回归系数与标准误进行合并。

参考文献：Yu-Sung Su, et al. Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. Journal of Statistical Software. Volume VV, Issue II.

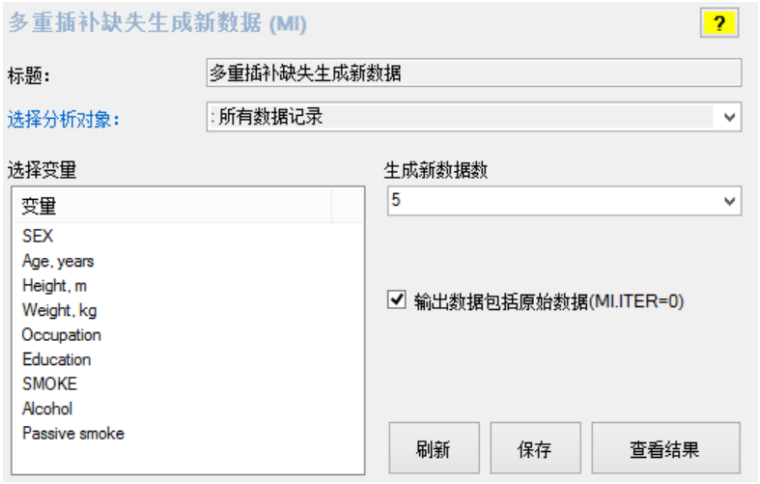
例: 练习数据 demo.xls 中，AGE、Height、Weight、Education、Smoke、Passive-smoke、Alcohol 等变量有缺失，使用“变量取值（缺失）组合”模块统计缺失情况如下：

多变量缺失组合

SEX	Age	Height	Weight	Occupation	Education	SMOKE	Passive smoke	Alcohol	Frequency
A	NA	NA	NA	NA	NA	NA	NA	NA	2
A	A	NA	NA	A	A	A	A	A	37
A	A	A	A	A	NA	NA	NA	NA	5
A	A	A	A	A	A	NA	A	NA	1
A	A	A	A	A	A	A	A	NA	3
A	A	A	A	A	A	A	A	A	784

A: 非缺失 ; NA: 缺失

现使用多重插补，对缺失数据进行插补生成 5 套完整数据，输入界面如下：



输出结果：

```
 Multiple imputation
     names include order number.mis all.mis               type collinear
1      SEX     Yes    NA          0      No             binary        No
2      AGE     Yes     1          2      No positive-continuous        No
3   HEIGHT     Yes     2         39      No positive-continuous        No
4   WEIGHT     Yes     3         39      No positive-continuous        No
5 OCCU.NEW     Yes     4          2      No             binary        No
6  EDU.NEW     Yes     5          7      No ordered-categorical        No
7    SMOKE     Yes     6          8      No             binary        No
8      ALH     Yes     7         11      No             binary        No
9     PSMK     Yes     8          7      No             binary        No

Multiply imputed data set

Call:
 .local(object = object, n.iter = ..3, R.hat = ..4, max.minutes = ..2,
    run.past.convergence = TRUE)

Number of multiple imputations:  5

Number and proportion of missing data per column:
     names               type number.mis  proportion
1      SEX             binary          0 0.000000000
2      AGE positive-continuous          2 0.002403846
3   HEIGHT positive-continuous         39 0.046875000
4   WEIGHT positive-continuous         39 0.046875000
5 OCCU.NEW             binary          2 0.002403846
6  EDU.NEW ordered-categorical          7 0.008413462
7    SMOKE             binary          8 0.009615385
8      ALH             binary         11 0.013221154
9     PSMK             binary          7 0.008413462

Total Cases: 832
Missing at least one item: 8
Complete cases: 784
          Coefs  Data
Converged FALSE FALSE
```

右击输入文件，看到新生成的 demo_XXX_mi.xls 文件（XXX 为易倆分析操作编号），该文件含有原数据及 5 套插补后无缺失的数据。其中 MI_ITER 变量表示插补号：MI_ITER=0 表示原数据；MI_ITER=1、2、3、4、5 分别表示 5 套插补后的数据。