# Imputation of missing values

In PRELIS there are two ways of handling missing values: pairwise and listwise deletion. In many situations, particularly when values are missing not completely at random, these procedures are far from satisfactory (see, for example, Little & Rubin, 1987, and Rubin, 1987). PRELIS offers yet another possibility of handling missing values, namely by imputation, *i.e.*, by substitution of real values for the missing values. The value to be substituted for the missing value for a case is obtained from another case that has a similar response pattern over a set of matching variables. To do this, include a line

```
IM (Ivarlist) (Mvarlist) VR = n XN XL
```

in the input file, where Ivarlist is a set of variables whose missing values should be imputed and Mvarlist is a set of matching variables. VR, XN, and XL are explained below.

The imputation scheme is as follows. Let $y_1, y_2, ..., y_p$ denote the variables in Ivarlist and let $x_1, x_2, ..., x_q$ denote the variables in Mvarlist. To begin, let us assume that there is only a single variable *y* in Ivarlist whose missing values are to be imputed and that *y* is not included in Mvarlist. Let $z_1, z_2, ..., z_q$ be the standardized $x_1, x_2, ..., x_q$, *i.e.*, for each case *c*

$$z_{cj} = (x_{cj} - \bar{x}_j) / s_j \quad j = 1, 2, ..., q,$$

where $\bar{x}_j$ and $s_j$ are the estimated mean and standard deviation of $x_j$. These are estimated from all complete data on $x_j$.

The imputation procedure is as follows.

1. Find the first case $\alpha$ with a missing value on *y* and no missing values on $x_1, x_2, ..., x_q$. If no such case exists, imputation of *y* is impossible. Otherwise, proceed to impute the value $y_\alpha$ as follows.
2. Find all cases *b* which have no missing value on *y* and no missing values on $x_1, x_2, ..., x_q$, and which minimizes

$$\sum_{j=1}^{q} \left( z_{bj} - z_{aj} \right)^2 \qquad \text{(B.1)}$$

3. Two cases will occur
   - If there is a single case *b* satisfying 2, $y_a$ is replaced by $y_b$.
   - Otherwise, if there are *n* > 1 matching cases *b with the same minimum value* of (B.1), denote their *y*-values by $y_1^{(m)}, y_2^{(m)}, ..., y_n^{(m)}$. Let

$$\bar{y}_m = (1/n) \sum_{i=1}^{n} y_i^{(m)}, \quad s_m^2 = [1/(n-1)] \sum_{i=1}^{n} (y_i^{(m)} - \bar{y}_m)^2,$$

   be the mean and variance of the *y*-values of the matching cases. Then imputation will be done only if

$$\frac{s_m^2}{s_y^2} < v, \tag{B.2}$$

where $s_y^2$ is he total variance of $y$ estimated from all complete data on $y$ and $v$ is the value VR specified on the MI command. This may be interpreted to mean that the matching cases predict the missing value with a reliability of at least $1 - v$. The default value of VR is VR = 0.5, *i.e.*, $v = 0.5$. Larger values than this is not recommended. Smaller values may be used if one requires high precision in the imputation. For each value imputed, PRELIS gives the value of the variance ratio and the number of cases on which $s_m^2$ is based.

If condition (B.2) is satisfied, then $y_a$ is replaced with the mean $\bar{y}_m$ if $y$ is continuous or censored, or with the value on the scale of $y$ closest to $\bar{y}_m$ if $y$ is ordinal. Otherwise, no imputation is done and $y_\alpha$ is left as a missing value.

4. This procedure is repeated for the next case *a* for which $y_\alpha$ is missing, and so on, until all possible missing values on $y$ have been imputed.

This procedure has the advantage that it gives the same results under linear transformation of the matching variables. Thus, if age is a matching variable, age can be in years or months, or represented by the year of birth, and the resulting imputed data will be the same in each case. Another advantage is that the results of the imputation will be the same regardless of the order of cases in the data.

If Ivarlist contains several variables, they will be imputed in the order they are listed. This is of no consequence if no variables in Ivarlist is included in Mvarlist. Ideally, Ivarlist contains the variables with missing values and Mvarlist contains variables without missing values. However, PRELIS can also handle the case when some variables are included in both varlists, it is automatically excluded from Mvarlist when its values are imputed. In this case, the order of the variables in Ivarlist can make a difference, since a variable already imputed can be used as matching variable when another variable is imputed.

Imputation of missing values should be done with utmost care and control, since missing values will be replaced by other values that will be treated as real observed values. If possible, use matching variables which are *not* to be used in the LISREL modeling. Otherwise, if the matching variables are included in the LISREL model, it is likely that the imputation will affect the result of the analysis. This should be checked by comparing with the result obtained without imputation.

For each variable to be imputed, PRELIS lists all the cases with missing values. If imputation is successful, it gives the value imputed, the number of matching cases and the variance ratio. If the imputation is not successful, it gives the reason for the failure. This can be either that no matching case was found or that the variance ratio was too large. The XN option on the IM command will make PRELIS list only successful imputations, and the XL option makes PRELIS skip the entire listing of cases. PRELIS always gives the number of missing values per variable, both before and after imputation.

## Example

The following input file (**Ex7d.prl** in the **PRELIS Examples** folder) is used to illustrate. The missing values of each variable are imputed using all the other variables as matching variables. Cases with missing values are eliminated after imputation. Category labels and then assigned to the data values that remain after listwise deletion and the data screening is done on this subsample.

```
EXAMPLE 7D
Imputing Missing Values
DA NI=6 MI=8,9
LA
NOSAY VOTING COMPLEX NOCARE TOUCH INTEREST
RA FI=EX7.RAW FO;(T142,6F2.0)
IM (NOSAY - INTEREST) (NOSAY - INTEREST)
```

```
CL  NOSAY - INTEREST 1=AS 2=A 3=D 4=DS
OU
```

The output file gives the following information concerning missing values and imputation.

```
Number of Missing Values per Variable

     NOSAY    VOTING   COMPLEX    NOCARE     TOUCH  INTEREST
  --------  --------  --------  --------  --------  --------
         5         8         3         7        14        14

 Imputations for    NOSAY

 Case   56 not imputed because of missing values for matching variables
 Case   88 imputed with value     3 (Variance Ratio = 0.393), NM=    4
 Case   99 not imputed because of missing values for matching variables
 Case  229 not imputed because of missing values for matching variables
 Case  274 imputed with value     3 (Variance Ratio = 0.315), NM=   11

 Imputations for   VOTING

 Case   13 not imputed because of Variance Ratio = 2.312 (NM=    6)
 Case   18 not imputed because of missing values for matching variables
 Case   62 not imputed because of missing values for matching variables
 Case   99 not imputed because of missing values for matching variables
 Case  138 imputed with value     1 (Variance Ratio = 0.000), NM=    1
 Case  180 not imputed because of missing values for matching variables
 Case  188 not imputed because of missing values for matching variables
 Case  257 imputed with value     2 (Variance Ratio = 0.324), NM=   13

 Imputations for  COMPLEX

 Case  143 not imputed because of missing values for matching variables
 Case  188 not imputed because of missing values for matching variables
 Case  240 imputed with value     2 (Variance Ratio = 0.394), NM=   18

 Imputations for   NOCARE

 Case   40 not imputed because of missing values for matching variables
 Case  143 not imputed because of missing values for matching variables
 Case  144 imputed with value     3 (Variance Ratio = 0.000), NM=    1
 Case  206 not imputed because of missing values for matching variables
 Case  229 not imputed because of missing values for matching variables
 Case  233 imputed with value     3 (Variance Ratio = 0.000), NM=    1
 Case  270 imputed with value     3 (Variance Ratio = 0.000), NM=    7

 Imputations for    TOUCH

 Case   18 not imputed because of missing values for matching variables
 Case   28 not imputed because of missing values for matching variables
 Case   29 imputed with value     2 (Variance Ratio = 0.000), NM=    1
 Case   37 imputed with value     2 (Variance Ratio = 0.000), NM=    2
 Case   40 not imputed because of missing values for matching variables
 Case   56 not imputed because of missing values for matching variables
 Case   62 not imputed because of missing values for matching variables
 Case   99 not imputed because of missing values for matching variables
 Case  104 imputed with value     2 (Variance Ratio = 0.000), NM=    1
 Case  143 not imputed because of missing values for matching variables
```

```
 Case  188 not imputed because of missing values for matching variables
 Case  203 not imputed because of missing values for matching variables
 Case  209 not imputed because of Variance Ratio = 0.618 (NM=    5)
 Case  238 imputed with value     1 (Variance Ratio = 0.000), NM=    1

 Imputations for INTEREST

 Case   12 not imputed because of Variance Ratio = 0.611 (NM=    3)
 Case   18 not imputed because of missing values for matching variables
 Case   28 not imputed because of missing values for matching variables
 Case   48 imputed with value     2 (Variance Ratio = 0.000), NM=    2
 Case   56 not imputed because of missing values for matching variables
 Case   62 not imputed because of missing values for matching variables
 Case   64 imputed with value     3 (Variance Ratio = 0.000), NM=    1
 Case   67 imputed with value     2 (Variance Ratio = 0.000), NM=    1
 Case   99 not imputed because of missing values for matching variables
 Case  180 not imputed because of missing values for matching variables
 Case  188 not imputed because of missing values for matching variables
 Case  203 not imputed because of missing values for matching variables
 Case  206 not imputed because of missing values for matching variables
 Case  229 not imputed because of missing values for matching variables

Number of Missing Values per Variable After Imputation

     NOSAY    VOTING   COMPLEX    NOCARE     TOUCH  INTEREST
  --------  --------  --------  --------  --------  --------
         3         6         2         4        10        11

 Distribution of Missing Values

 Total Sample Size(N) =    312

 Number of Missing Values     0     1     2     3     4
          Number of Cases   297     3     5     5     2
```

Fifteen data values were successfully imputed, two in NOSAY, two in VOTING, one in COMPLEX, three in NOCARE, four in TOUCH, and three in INTEREST. The listwise sample was increased from 282 to 297. Many cases could not be imputed because of missing values in the matching variables. Only three cases could not be imputed because of a variance ratio being too large. For more successful examples of imputation, see Aish, A.M. & Jöreskog, K.G. (1990).

## References

Aish, A.M. & Jöreskog, K.G. (1990). A panel model for political efficacy and responsiveness: An application of LISREL 7 with weighted least squares. *Quality and Quantity*, **24**, 405-426.

Little, R.J.A. & Rubin, D.B. (1987). Statistical analysis with missing data. New York: Wiley.
Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.