Imputation with Blaise and Manipula

Jelke Bethlehem and Lon Hofman, Statistics Netherlands

1. Introduction

Statistical surveys are always affected by non-response. There is item non-response, in which only the answers to some questions are missing, and there is also unit non-response, in which the answers to all questions are missing. For many years now Statistics Netherlands has been confronted with severe non-response problems. This is mainly unit non-response. Consequently, research has concentrated on correction techniques for this type of non-response. One of the products of this research is the program Bascula for adjustment weighting.

The last couple of years there has been increased attention for item non-response. This is usually taken care of by means of imputation techniques. Imputation comes down to making an estimate for the missing answer according to some model, and substituting these synthetic values in the data file. The paper presents a methodological overview of various imputation methods, and shows how they can be implemented using Blaise and Manipula.

2. The non-response problem

Survey sampling is a well-established sampling method. By choosing a proper sampling design, it is possible to compute accurate estimates based on relatively small samples. This allows for cost-effective data collection and timely publications. However, there also is another side to this coin. Not everything is under control. There are practical problems hindering a smooth execution of sample surveys. One of the most important problems survey organisations have been facing over the last decades is non-response.

Non-response is the phenomenon that elements (persons, households, companies) in the selected sample do not provide the requested information, or that the provided information is useless. The situation in which all requested information on an element is missing is called *unit non-response*. If information is missing on some items only, it is called *item non-response*. This chapter will only handle unit non-response.

Due to non-response the sample size is smaller than expected. This leads to less accurate, but still valid, estimates of population characteristics. This is not a serious problem. It can be taken care of by taking the initial sample size larger. A far more serious problem caused by non-response is that estimates of population characteristics may be biased. This situation occurs if, due to non-response, some groups in the population are over- or underrepresented, and these groups behave differently with respect to the characteristics to be investigated.

Indeed, estimators must be assumed to be biased unless very convincing evidence of the contrary is provided. Bethlehem and Kersten (1987) discuss a number of surveys of Statistics Netherlands. A follow-

up study of the Victimisation Survey showed that people who have fear when they are alone at night, are less inclined to participate in the survey. In Housing Demand Surveys it turned out that people who refuse to co-operate, have a lesser housing demand than responding people. For the Survey on the Mobility of the Population it is obvious that mobile people are relatively under-represented among the respondents.

Figure 1.1. Non-response percentages of some Statistics Netherlands surveys

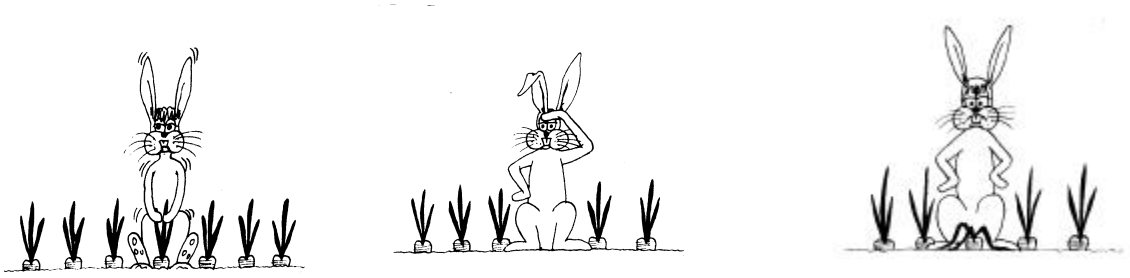
Year	Labour Force Survey	Consumer Sentiments Survey	Survey on Well-being	Mobility Survey	Holiday Survey
1972		29			
1973	12	23			
1974		25	28		
1975	14	22			14
1976		28	23 [2])		13
1977	12	31	30		19
1978		36		33	22
1979	19	37	35 [3])	31	26
1980		39	39	32	26
1981	17	35		32	26
1982		40	36 [2])	34	29
1983	19	37	42	34	26
1984		35 [1])		36	31
1985	23	31		39	32
1986		29	41	41	34
1987	40 [1])	29		41	
1988	41	32		45	
1989	39	32		42	
1990	39	32		45	
1991	40	31		43	
1992	42	31	55	43	
1993	42		54	44	
1994	41		48 [1])	45	
1995	40		46	46	
1996	42		48	48	
1997	44			50	

[1]) Revision [2]) Young people only [3]) Old people only

The magnitude of the non-response in many Dutch sample surveys is increasing to such an extent that without special adjustment techniques one has to reckon with a decrease of the quality of the results. Figure 1.1 presents non-response figures of a number of surveys carried out Statistics Netherlands.

It is difficult to compare response rates from different surveys but for each of the surveys one can study its non-response figures over the years. The magnitude of the non-response is determined by a large number of factors, including the subject of the survey, the target population, the time period, the length of the questionnaire, the quality of the interviewers, the fieldwork in general, etc. It is also clear from figure 1.1 that non-response is a considerable problem. It has an impact on the costs of a survey, since it takes more and more effort to obtain estimates with the same precision as originally specified in the sampling design.

Figure 1.2. Reasons for non-response



Non-respondents can be classified in three important groups, see figure 1.2. People in the first group refuse to co-operate. Sometimes it is possible to make an appointment for an interview at some later date. However, frequently the refusal can be considered permanent. Possible causes are fear of privacy intrusion, and interview fatigue. The second group of non-respondents is the not-contactable. No contact is made due to the fact people are not at home, due to removal or due to other circumstances like watchdogs, dangerous neighbourhoods or houses which are difficult to reach. Generally speaking, people are increasingly hard to contact. Important factors are smaller family sizes, greater mobility and a larger amount of spare time that is spent out-of-doors. The third group of non-respondents consists of people who are physically or mentally not able to co-operate during the fieldwork period. Also language problems can cause this type of non-response.

To be able to build a well-founded theory of non-response adjustment it is necessary to incorporate the phenomenon of non-response in the theory of sampling. The literature on non-response contains two different basic views on how non-response occurs. Here, these views are labelled the Random Response Model and the Fixed Response Model.

The *Random Response Model* assumes some kind of random mechanism determining whether or not a selected person will respond. Each element in the population is assigned an imaginary random number generator. This generator can only produce either the value 0 or 1. A value 1 means response, and a value 0 non-response. The probability with which the random number generator assumes the value 1, is different for each element in the population. All response probabilities are assumed to be unknown. Under the Random Response Model, there are two random processes controlling the availability of data: the sample selection mechanism defined by the researcher, and the response mechanism that is not under his control.

The *Fixed Response Model* assumes the population to consist of two mutually exclusive and exhaustive sub-populations: the response stratum and the non-response stratum. Elements in the response stratum would participate in the survey with certainty, if selected in the sample, and elements in the non-response stratum would not participate with certainty, if selected. The Fixed Response Model can be regarded as a special case of the Random Response Model in which the response probabilities are either 0 or 1. Many authors consider the Fixed Response Model a too simple and unrealistic model.

Both models can be used to study the effect of non-response on estimates of population estimates, and the conclusions from both models are the same: generally, these estimates will be biased, and the size of the bias is determined by two factors:

- The non-response rate. The higher the non-response rate, the larger the bias. For very small non-response rates, the bias may be ignored, but high non-response rates may incur a substantial bias.

- The extent to which non-respondents differ from respondents. The larger the difference between the average values of respondents and non-respondents, the larger the bias will be. If the non-respondents can be considered to be a random sub-sample, there will be no bias.

In practice, it is very difficult to asses the possible negative effects of non-response. And even if such effects can be detected, it is no simple matter to correct for them. A correction is only possible if auxiliary information is available. For example, if there is an auxiliary variable that has been measured in the sample, and for which population characteristics are know, then this variable can be used to check whether the available data show unbalancedness, i.e. they are not representative for the population.

An example illustrates this. Suppose, a sample of municipalities is selected in order to estimate the average milk production. Due to non-response not all municipalities provide information. Suppose that also the number of farms per municipality is measured as an auxiliary variable. If the average number of farms per municipality in the sample is 69, and it is known that the population average is 91, then there is something wrong. Apparently, municipalities with a small number of farms are over-represented in the survey. Taking into account that these small municipalities will also have a small milk production, it is likely that the average milk production will be under-estimated.

3. Correction for non-response

There are two approaches to non-response correction. The first approach relates to unit non-response, the situation in which all requested information on an element is missing. To correct for a possible bias due to unit non-response, often a weighting method is carried out.

The basic principle of *weighting* is that every observed element is assigned a specific weight. By processing the weighted values instead of the values themselves estimates for population characteristics are obtained. The easiest and most straightforward method used to compute weights is post-stratification. The population is divided into strata after selection of the sample. If each stratum is homogeneous with respect to the target variable of the survey, then the observed elements resemble the unobserved elements. Therefore, estimates of stratum characteristics will not be very biased, so they can be used to construct population estimates.

To carry out post-stratification, discrete auxiliary variables are needed, and preferable auxiliary variables having a strong relationship with the target variable. All observed elements within a stratum are assigned the same weight, and this weight is computed such that the weighted sample distribution of the auxiliary variables agrees with the population distribution of these variables. If the relationships are strong enough, also the weighted sample distribution of the target variable will agree with its population distribution.

There are also more advanced weighting methods, see e.g. Bethlehem and Keller (1987), and Deville and Särndal (1992). Several types of weighting have been implemented in the package Bascula. For more information about Bascula, see Bethlehem (1997).

The second approach to non-response correction relates to item non-response, the situation in which only part of the requested information about an element is missing. In this case only some questions in the questionnaire have remained unanswered, but these are usually the sensitive questions. Item non-response requires a different approach. A great deal of additional information is available for the elements involved. All available responses to other questions can be used to predict the answer to the missing questions. This computation of a 'synthetic' answer to a question is called *imputation.* The Blaise family of software packages does not (yet) contain a package of imputation. However, many imputation techniques can be implemented using Manipula. Section 4 discusses some of the theoretical background of imputation. And section 5 shows some ways of implementing imputation techniques in Blaise and Manipula.

4. Some theory of imputation

Imputation relates to a family of techniques for replacing missing values in a data set by 'synthetic values' obtained from some kind of model. Such a model describes a relationship between the variable having missing values and other variables for which the values are available.

Let Y denote the *target variable* having missing values that must be imputed. Furthermore, there are a number of *auxiliary variables* X_1, X_2, … used to predict the missing values of the target variable.The prediction model is estimated using only those records in the data file for which values of both the target variable and the set of auxiliary variables is available. Next, the model is used to predict the missing values of Y. For each missing value, the set of values of the auxiliary variables is substituted in the model, which results in a predicted value of Y.

Imputation techniques can range from simple ad hoc procedures to sophisticated prediction techniques based on complex models. Kalton and Kasprzyk present a list of some commonly used imputation techniques:

- *Deductive imputation.* Sometimes, the missing answer to a question can be deduced with certainty from the available answers to other questions. When range, consistency and route checks restrict the answer to only one possible value, deductive imputation can be applied. This is the ideal form of imputation.

- *Imputation of the mean.* This technique substitutes the mean of the available values of the target variable Y for all missing values of Y.

- *Imputation of the group mean.* The sample is divided into groups using auxiliary variables. Within each group, the mean of the available values of Y is assigned to all missing values of Y.

- *Random imputation.* For each missing value of Y, a value is chosen at random from the set of available values of this variable.

- *Random imputation within groups.* The sample is divided into groups using auxiliary variables. Within each group, a missing value is substituted by a randomly chosen value from the set of available values of Y within the group.

- *Hot-deck imputation.* This is a special implementation of random imputation within groups. For each group, a donor record is maintained. The records in the file are processed sequentially. If the field of the variable to be imputed contains a 'real' value, the value is copied to the donor record. If the value in the field is missing, the value from the donor record is copied to the field.

- *Regression imputation.* A regression model is constructed that explains the values of the target variable Y from the values of auxiliary variables X_1, X_2, … Then, the fitted regression model is used to predict the answer in missing cases. To conserve the distributional properties of the data, often a residual, drawn from some normal distribution, is added to the prediction.

At the first sight, these techniques appear to be rather diverse. Still, most of them can be put into a general framework. Let Y_i denote the value of the target variable Y in the i-th record (for i = 1,2, …, n). Suppose there are p auxiliary variables X_1, X_2, …, X_p. The values of these variables in the i-th record are denoted by X_{i1}, X_{i2}, …, X_{ip}. If the value Y_i is missing, a general expression of the imputed value can be obtained from

$$\hat{Y}_i = b_0 + \sum_{j=1}^{p} b_j X_{ij} + E_i, \qquad (4.1)$$

where $\hat{Y}_i$ denotes the imputed value, X_{ij} the value of the j-th auxiliary variable, b_0, b_1, …, b_p are regression coefficients, and E_i is the value of a random variable E obtained by drawing a value from some specific distribution depending on the chosen imputation technique.

It is clear that expression (4.1) includes regression imputation. If all values of E are set to 0, the imputed value is the value predicted by the regression model. Often a residual noise variable is added to the prediction by the regression model. This is done to conserve the distributional properties of the target variable Y. The value of the noise variable is obtained by drawing some value from a normal distribution.

If the auxiliary variables are taken to be dummy variables representing imputation groups, expression (4.1) can be used for imputation of the group mean. For this, b_0 and the E_i's must be set to 0. The remaining b_j's must be taken equal to group means of Y, i.e.

$$b_j = \overline{Y}_j,$$

for j = 1, 2, …, p. Within this framework, imputation of the mean is the special case of imputation of the group mean. It is obtained by introducing only one group, i.e. there is one auxiliary variable X_1, and it always has the value 1.

Random imputation within groups is obtained by adding a noise variable to the model for imputation of the mean. The auxiliary variables are dummy variables representing the groups, $b_0 = 0$, and b_1, b_2, …, b_p are the group means of Y. For each group j, a set of values

$$E_{ij} = \overline{Y}_j - Y_{ij}$$

(for i=1,2,…) is formed. The Y_{ij} denote the values of Y in the j-th group. The noise value E_i is obtained by selecting a random value from the proper set of values, i.e. if record i belongs to group j, E_i is selected from E_{1j}, E_{2j}, … . Random imputation is a special case. It is obtained by using only one group.

Hot-deck imputation is a special implementation of random imputation within groups. If the order of the records in the data file is completely random, both techniques are more or less equivalent.

The success of an imputation technique depends on properties of the mechanism generating item non-response. Little and Rubin (1987) consider three types of patterns leading to missing data:

1? The probability of a missing value of Y is independent of the value of Y and independent of the value of the X's. This case is called *Missing Completely At Random* (MCAR). Then the observed values of Y form a random sub-sample from the sample. The mean of the observed values is an unbiased estimate of the population mean.

2? The probability of a missing value of Y depends on the value of X but is independent of the value of Y's. This case is called *Missing At Random* (MAR). Then the observed values of Y do not form a random sub-sample of the sample. However, they are a random sub-sample within the classes defined by the values of the X's. The auxiliary variables can be used to effectively correct for a bias due to missing values.

3? The probability of a missing value of Y depends both on the value of Y and the values of X's. Then the observed values of Y do not form a random sub-sample of the sample. Also, they are not a random sub-sample within the classes defined by the values of X. Therefore, the auxiliary variable cannot be used to effectively correct for a bias due to missing values.

There are several considerations that play a role in selecting an imputation technique. One is the type of the target variable. All techniques listed above can be applied routinely on qualitative (continuous) variables. However, some of the techniques cannot be applied to quantitative (discrete or categorical) variables, because imputed values will not necessarily belong to the domain of valid values. For example, imputation of the mean or regression imputation for a variable Sex with two possible values (1 for male, and 2 for female) may easily produce a value like 1.4. So, for a qualitative variable it is better to only use random imputation (possibly within groups) or hot-deck imputation. These techniques always produce 'real' values.

Imputation techniques can be classified as random or deterministic, depending on whether a noise variable is used or not. The deterministic techniques usually work in such a way that the mean of all values (observed and imputed) is equal to the mean of the observed values. For the random imputation methods, the expected value (over the residual producing mechanism) of the mean over all values is equal to the observed mean. So both methods have the same effect on the bias of estimates. However, adding noise introduces an extra source of variation, and therefore reduces the precision of estimates. This may be a reason to prefer deterministic techniques for estimating the population mean.

Deterministic techniques have the disadvantage that they distort the properties of the distribution of the values of the variable. These techniques tend to predict values in the middle part of the distribution. The distribution of Y in the imputed data set is much more peaked and much more concentrated than the original distribution. Therefore, standard errors computed from the imputed data set are generally too small. They create a too optimistic view of the precision of estimates. Random imputation methods do not have this nasty property. They are much better able to preserve the original distribution.

A final point to take into consideration is the effect of imputation on relationships between variables. Imputation of the overall mean and random imputation causes covariances and correlations to be biased. This occurs because imputed Y values are uncorrelated with the values of other variables in the records. By applying imputation within groups, the bias is decreased, but not avoided. Also, regression imputation (with or without a residual) introduces a bias in the covariance.

It is clear that the ideal imputation technique does not exist. A researcher always has to be careful in the analysis of a data set that has been subject to imputation (unless the amount of imputation is small). Research for new imputation techniques is still in progress. An example is the multiple imputation technique proposed by Rubin(1979). This technique computes a set of, say *m* imputed values for each missing value. This results in *m* imputed data sets. Inference is based on the distribution obtained by computing the estimate for each of the *m* data sets.

Also, application of neural networks and of evolutionary algorithms looks promising. Experiments have shown that there are situations in which these techniques are useful. However, further research is necessary.

5. Imputation in Blaise

Simple deductive imputation can be carried out with the Data Entry Program (DEP) of Blaise itself. By means of compute instructions in the rules section answers to questions can be computed. Most other imputation techniques require information from other records, or from all records. Since the DEP is a form oriented utility, it is less convenient to use, or even impossible for these techniques. Manipula is much more suitable in these cases.

This section shows how a number of imputation techniques can be implemented in Manipula. A simple example is used to illustrate these techniques. A data file has been constructed for females with a job in the country of Samplonia. The file contains three variables: Province (with two values Agria and Induston), Age (in the range from 20 to 65), and Income (in the range from 0 to 1500). In five records the value of Income is missing. We will show what happens if various imputation techniques are used to replace the missing values of income by imputed values.

In Samplonia, there is a clear relationship between income, age and province. Figure 5.1 shows a scatterplot of this relationship. Two distinct clusters can be distinguished. The upper cluster relates to the province of Induston, and the lower cluster contains data from Agria. Apparently, incomes are higher in Induston, and increase with age. Incomes in Agria are low and independent from age. Of the five missing observations, two are from Induston, and three from Agria. A good imputation technique will take the structure into account, and produce imputed values that fit nicely in these clusters.

Figure 5.1. The relationship between income, age and province for females in Samplonia

Figure 5.2. contains the Manipula setup for the simplest imputation technique. It is imputation of the overall mean of the available values of income.

Figure 5.2. Manipula setup for imputation of the mean

```
SETTINGS
  AUTOREAD = NO

USES
  BlaiseData 'Samfem'

UPDATEFILE
  UpFile: BlaiseData ('Samfem', BLAISE)
SETTINGS
  CHECKRULES = YES

AUXFIELDS
  RMean: REAL
```

```
  IMean: INTEGER
  RecNum: INTEGER
  N: INTEGER

MANIPULATE
  FOR RecNum:= 1 TO UpFile.RECORDCOUNT DO
    UpFile.READNEXT
    IF Income <> EMPTY THEN
      RMean:= RMean + Income
      N:= N + 1
    ENDIF
  ENDDO

  IMean:= RMean / N
  UpFile.RESET

  FOR RecNum:= 1 TO UpFile.RECORDCOUNT DO
    UpFile.READNEXT
    IF Income = EMPTY THEN
      Income:= IMean
      Imputed:= 1
      UpFile.Write
    ENDIF
  ENDDO
```

In the Settings section *AUTOREAD* is set to *NO*. This makes it possible to make several runs through the data file in the same setup. In this example, there are two runs. In the first run, the mean is computed, and assigned to the temporary variable *IMean*. Note that the check *IF Income <> EMPTY* sees to it that the mean is computed over all 'real' values. The variable *N* counts the number of these observations. In the second run the value of the mean is assigned to the variable *Income* in all cases for which the value of Income is missing.

The Blaise model contains a field with the name Imputed. This field records whether the value of Income is 'real' or imputed. Such an imputation flag may be important for future analyse of the data.

In this example, the value of the overall mean is equal to 429. This value lies somewhere between the two clusters in the scatterplot. Therefore it is bad imputation technique in this case. The overall mean of income is not affected, but the structure of the relationship between income and age is distorted.

Figure 5.3. contains the Manipula setup for imputation of the group means, where in this case the groups are the two provinces.

Figure 5.3. Manipula setup for imputation of the mean within groups

```
SETTINGS
  AUTOREAD = NO

USES
  BlaiseData 'Samfem'

UPDATEFILE
  UpFile: BlaiseData ('Samfem', BLAISE)
SETTINGS
  CHECKRULES = YES

AUXFIELDS
  RMean: ARRAY[1..2] OF REAL
  IMean: ARRAY[1..2] OF INTEGER
  N: ARRAY[1..2] OF INTEGER
  RecNum: INTEGER
  I: INTEGER

MANIPULATE
  FOR RecNum:= 1 TO UpFile.RECORDCOUNT DO
    UpFile.READNEXT
    IF Income <> EMPTY THEN
      RMean[Province]:= RMean[Province] + Income
      N[Province]:= N[Province] + 1
    ENDIF
  ENDDO

  FOR I:= 1 TO 2 DO
    IMean[I]:= RMean[I] / N[I]
  ENDDO
  UpFile.RESET

  FOR RecNum:= 1 TO UpFile.RECORDCOUNT DO
    UpFile.READNEXT
    IF Income = EMPTY THEN
      Income:= IMean[Province]
      Imputed:= 1
```

```
      UpFile.Write
    ENDIF
  ENDDO
```

The difference with the previous setup in figure 5.2 is that now group means are computed for each province separately. So there is now an array of group means. The number of ‘real’observations in each group is also counted in an array.

The imputed group mean for the province of Induston is equal to 995. For missing values in Induston this is a better value than for imputation of the overall mean. Nevertheless, it is not ideal. The elderly people with high incomes (up to a value of 1405) would also get an imputed value of 995. The imputed value is too low. And for young people the imputed value would be too high. For the province of Agria, the imputed value is 153. This looks reasonable, since income seems to be more or less constant in this province.

Figure 5.4 contains the Manipula set for random imputation. Also for this technique, two runs must be made through the data file: one to collect all ‘real’values of income, and one to randomly assign values from this set to records having missing values.

Figure 5.4. Manipula setup for random imputation

```
SETTINGS
  AUTOREAD = NO

USES
  BlaiseData 'Samfem'
  DATAMODEL MValues
    PRIMARY
      RecNr
    FIELDS
      RecNr: INTEGER[4]
      Income: 0..6000
  ENDMODEL

UPDATEFILE
  UpFile: BlaiseData ('Samfem', BLAISE)
SETTINGS
  CHECKRULES = YES
```

```
TEMPORARYFILE
  Values: MValues
SETTINGS
  AUTOCOPY = NO

AUXFIELDS
  RecNum: INTEGER
  N: INTEGER

MANIPULATE
  FOR RecNum:= 1 TO UpFile.RECORDCOUNT DO
    UpFile.READNEXT
    IF Income <> EMPTY THEN
      N:= N + 1
      RecNr:= N
      Values.WRITE
    ENDIF
  ENDDO

  UpFile.RESET

  FOR RecNum:= 1 TO UpFile.RECORDCOUNT DO
    UpFile.READNEXT
    IF Income = EMPTY THEN
      Values.GET(RANDOM(N) + 1)       { read random record    }
      UpFile.Income:= Values.Income   { the actual imputation }
      UpFile.Imputed:= 1
      UpFile.Write
    ENDIF
  ENDDO
```

Note that the ‘real’values of income are stored in a temporary file *Values*. As primary key for this file the record number is used. The total number of ‘real’ values is counted in the variable *N*. The function call *RANDOM(N)* generates a random integer from the set 0 up to and including N-1. So, to get a value in the range from 1 to N, we have added 1 to the result of the function call. The *GET* instruction is used to retrieve a record with a ‘real’ income value from the temporary file *Values*.

Application of this imputation technique will result in assignment of random income values to records with missing values. It is possible that the missing income in the province of Agria will be replaced by an income value from the province of Induston. It will be clear that random imputation will distort the structure of the relationship between income and age within the provinces.

Figure 5.5 contains the Manipula setup for random imputation within groups. It is an extension of the setup in figure 5.4. The temporary file has been extended with the variable *Province*. This variable has also been included in the primary key of the temporary file. All 'real' income values are again written to the temporary file with the appropriate value for the primary key. For each group, missing income values are replaced by a randomly selected 'real' income from the temporary file.

Figure 5.5. Manipula setup for random imputation within groups

```
SETTINGS
  AUTOREAD = NO

USES
  BlaiseData 'Samfem'
  DATAMODEL MValues
    PRIMARY
      Province, RecNr
    FIELDS
      Province: (Agria, Induston)
      RecNr: INTEGER[4]
      Income: 0..6000, EMPTY
  ENDMODEL



UPDATEFILE
  UpFile: BlaiseData ('Samfem', BLAISE)
SETTINGS
  CHECKRULES = YES

TEMPORARYFILE
  Values: MValues
SETTINGS
  AUTOCOPY=NO

AUXFIELDS
  N: ARRAY[1..2] OF INTEGER
  RecNum: INTEGER

MANIPULATE
```

```
FOR RecNum:= 1 TO UpFile.RECORDCOUNT DO
  UpFile.READNEXT
  IF Income <> EMPTY THEN
    N[Province]:= N[Province] + 1
    RecNr:= N[Province]
    Values.WRITE
  ENDIF
ENDDO

UpFile.RESET

FOR RecNum:= 1 TO UpFile.RECORDCOUNT DO
  UpFile.READNEXT
  IF Income = EMPTY THEN
    Values.GET(Province, RANDOM(N[Province]) + 1) { read random record }
    UpFile.Income:= Values.Income                { the actual imputation }
    UpFile Imputed:= 1
    UpFile.Write
  ENDIF
ENDDO
```

Random imputation within groups is better than random imputation. Still, it suffers from the same drawbacks as imputation of the group means: elderly people (with a high income) in Induston could be assigned the value of the income of a younger person, and thus would get a too low income.

Figure 5.6 contains the Manipula setup for a hot-deck imputation technique. The big advantage of hot-deck imputation is that it requires only one run through the data file. In this example there are two groups, corresponding to the two provinces. For each group, the last encountered value of income in that group is stored in the array *Donor*. When a missing value is encountered, it is replaced by the value in the donor array.

Note that in the present form, the setup has a shortcoming. The setup assumes the donor array will be filled with 'real' values at the moment the first missing value is encountered. If the first record in the data file contains a missing value, a value of 0 will be imputed. In a real production situation, the setup must be adapted to take care of this situation. Under the assumption that the order of the records in the data file is more or less arbitrary, hot-deck imputation has more or less the same properties as random imputation within groups.

Figure 5.6. Manipula setup for hot-deck imputation

```
SETTINGS
```

```
  AUTOREAD = NO

USES
  BlaiseData 'Samfem'

UPDATEFILE
  UpFile: BlaiseData ('Samfem', BLAISE)
SETTINGS
  CHECKRULES = YES

AUXFIELDS
  Donor: ARRAY[1..2] OF INTEGER
  RecNum: INTEGER

MANIPULATE
  FOR RecNum:= 1 TO UpFile.RECORDCOUNT DO
    UpFile.READNEXT
    IF Income <> EMPTY THEN
      Donor[Province]:= Income
    ELSE
      Income:= Donor[Province]
      Imputed:= 1
      UpFile.WRITE
    ENDIF
  ENDDO
```

Figure 5.7 contains the final Manipula setup to be considered. It contains an example of regression imputation. At first sight, this also seems to be a case of one run through the data file. But this is not true. In order to compute the coefficients of the regression model, also a run through the file must be made. Manipula is not the most convenient package for regression analysis. There are many statistical analysis packages that are better equipped for this purpose. Fortunately, Blaise offers tools to use these packages in a convenient way. For the current example, Cameleon was used to export to data to SPSS. A regression analysis was carried out in SPSS, and this resulted in the following model:

$$\text{Income} = 160 \times (2 - \text{Province}) + (209 + 20 \times \text{Age}) \times (\text{Province} - 1).$$

For a missing value in the province of Agria, the value of the variable Province is equal to 1, and so the model reduces to

$$\text{Income} = 160.$$

Apparently, always a value of 160 is predicted for Income, and this is not so bad if we look at the scatterplot in figure 5.1. For a missing value in the province of Induston, the value of the variable Province is equal to 2, and so the model reduces to

$$\text{Income} = 209 + 20 \times \text{Age}.$$

The imputed value of income increases with age. The straight line corresponding to this model follows the upper cluster of points in the scatterplot. Clearly, this regression imputation is the best imputation of the techniques discusses in this section.

Figure 5.6. Manipula setup for regression imputation

```
SETTINGS
  AUTOREAD = NO

USES
  BlaiseData 'Samfem'

UPDATEFILE
  UpFile: BlaiseData ('Samfem', BLAISE)
SETTINGS
  CHECKRULES = YES

AUXFIELDS
  RecNum: INTEGER

MANIPULATE
  FOR RecNum:= 1 TO UpFile.RECORDCOUNT DO
    UpFile.READNEXT
    IF Income = EMPTY THEN
      Income:= (2 - Province) * 160 + (Province - 1) * (209 + 20 * Age)
      Imputed:= 1
      UpFile.WRITE
    ENDIF
  ENDDO
```

6. Conclusion

Imputation is a much-used technique to get rid of missing values in a data file. That is what is does and not much more. Particularly if the pattern of missing values is not random, imputation may distort the properties of distribution of the variables. So care must be taken to select a proper imputation technique.

Although the Blaise System has no dedicated tools for imputation, it is not very difficult to implement it. Manipula seems to have all that is needed for carrying out imputation.

7. References

Bethlehem, J.G. & W.J. Keller (1987): Linear weighting of sample survey data. Journal of Official Statistics, 3, pp. 141-154.

Bethlehem, J.G. & H.M.P. Kersten (1987): The Non-response Problem, Survey Methodology 7, pp. 130-156.

Bethlehem, J.G.(1997): Bascula, Current Status and Future Developments. In: INSEE, Acte de la 4e Conférence Internationale des Utilisateurs de BLAISE, Paris, 5-7 Mai 1997, pp. 21-44.

Deville J.C and C.E. Särndal (1992): Calibtation Estimators in Survey Sampling. Journal of the American Statistical Association 87, pp. 376-382.

Kalton, G. and D. Kasprzyk (1986): The Treatment of Missing Survey Data. Survey Methodology 12, pp. 1-16.

Little, R.J.A. and D.B. Rubin (1987): Statistical Analysis with Missing Data. Wiley, New York.

D.B. Rubin, D.B. (1979): Illustrating the use of multiple imputations to handle non-response in sample surveys. Bulletin of the International Statistical Institute, Book 2, pp. 517-532.