Imputation

10

Abby Israëls, Léander Kuyvenhoven, Jan van der Laan, Jeroen Pannekoek and Eric Schulte Nordholt

Statistical Methods (201112)

Statistics Netherlands

The Hague/Heerlen

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2010–2011	= 2010 to 2011 inclusive
2010/2011	= average of 2010 up to and including 2011
2010/'11	= crop year, financial year, school year etc. beginning in 2010 and ending in 2011
2008/'09–2010/'11	= crop year, financial year, etc. 2008/'09 to 2010/'11 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

60165201112 X-37

Table of Contents

1. Introduction to the theme

1.1 General description and reading guide

1.1.1 Description of the theme

In surveys, respondents sometimes do not provide answers to one or more questions, while they are required to do so. In this case, we refer to *item non-response* (or partial non-response) and to *missing values* that should have been present. Reasons for answers not being provided are that the respondents are not willing or able to answer a question. For example, people are sometimes not able to answer a question that is complicated or difficult to understand, and they frequently do not want to provide answers to sensitive questions. Registers can also have missing data that Statistics Netherlands would have liked to have.

There are a number of ways to deal with missing values. One of these is to *impute* a valid value for the missing value in the data file. We refer to this as *imputing* or *imputation* (see section 1.4 for the definition) for the process step, and an *imputed value* or *imputation* for the result.

An alternative to imputation is to leave the values unknown. This will be done first of all for legitimately missing values. People without a job do not have to answer questions about their working environment; usually the routing in the questionnaire will ensure that these questions are only posed to people who are employed. Answers such as ‘don’t know’, ‘no opinion’ or ‘unknown’ will also be left that way when they say something about the knowledge or opinion of the respondent. But even in the case of missing values that should have been present, a decision can be made not to impute, and to resolve the problem not in the data file, but instead in the estimation or analysis. Especially for qualitative variables, there is the alternative of introducing the category ‘unknown’. Imputation is used more often for quantitative variables than for qualitative ones, and therefore also more often for business statistics than for social statistics.

Reasons to impute a value, instead of leaving the field empty, are as follows:
1. To obtain a ‘complete’ (completely filled) data file;
2. To increase the quality of the micro file and/or of the parameter estimates.

Point 1. Obtaining a complete file, with complete records, makes aggregation and tabulation easier, and prevents inconsistencies when tabulating. For example, missing values for a variable Education (in classes) means that the age distribution in the table ‘Age × Education’ will deviate from the age distribution in the table ‘Age × Gender’, unless ‘unknown’ is included as a category; you could also resolve the inconsistencies by ‘consistent and repeated’ weighting (see Methods Series, theme ‘Sampling Theory’, sub-theme ‘Repeated Weighting’). If, in a sampling survey, scores are missing on the quantitative variable Income, then you can only

estimate the mean income for the population or subpopulation of people who would have responded to the questioning, and that is a parameter that is not very relevant. Imputation helps in dealing with this problem, but it is of course only usable when the imputations are of sufficient quality.

Point 2. If we want to use imputation to improve the quality, 'the quality of what' should be clear. Often, the primary goal is to accurately determine means and totals, such as for the Structural Business Statistics, where total turnovers are the main output. We may also want to determine the distribution of a variable, for instance an income distribution and the associated inequality measures. For living situation studies, it is also important to have a good micro file, which researchers can use to perform a variety of analyses. Different objectives can lead to different 'optimum' imputations. For statistical output, however, you will want to have a maximum of one imputation per missing value, because otherwise the study results will no longer be internally consistent. In general, Statistics Netherlands can provide better imputations for general use than external users, because these parties often do not have all of the background characteristics that are useful for the imputation.

1.1.2 Problem and solutions

1.1.2.1 Reading guide

Sometimes, when a score is missing, it is possible to derive the 'actual' value with 100% certainty from the other characteristics of the object. In this case, you can use *deductive imputation* (Chapter 2) to impute that value. Edit rules are used for this purpose, the same ones that are frequently used in editing. If applicable, this method has preference above all other imputation methods. This imputation method can also be used if there is slightly less than 100% certainty about the accuracy.

Even if such a derivation is not possible, there will often be extra information (auxiliary variables, *x*-variables) that makes an accurate estimation of the missing value (on the *y*-variable) possible. By searching for a suitable, effectively explanatory model, you can try to improve the quality of the file or of the population parameters to be estimated using *model-based imputation*. The selected model then generates the value(s) to be filled in. However, it is not possible to assess the exact quality of the imputations: the real values are, after all, unknown, unless it is possible to obtain information from other sources or surveys. Model estimation is only possible for the item respondents. There will also usually be an imputation bias (bias in the outcomes as a result of creating erroneous imputations), because the fitted model with the parameters will usually not apply exactly for the item non-respondents.

So if there is uncertainty with respect to the value $\tilde{y}_i$ to be imputed, you can try to estimate it using a model. You will then search for a model for *y* that will predict the missing value y_i as accurately as possible. Often, a regression model will be used for this purpose, and this is referred to as *regression imputation* (Chapter 5). This is mainly used for quantitative *y*-variables. The *mean imputation* and *ratio imputation*

to be discussed in Chapters 3 and 4 are special cases of regression imputation. For mean imputation, no auxiliary information is used, usually because this is not available; for ratio imputation, only a single quantitative auxiliary variable is used. These methods are addressed separately because of their simplicity and frequent application. There are also *donor imputation methods (hot deck)*: *random hot deck*, *sequential hot deck* and *nearest neighbour* (incl. *predictive mean matching*); see Chapter 6. In terms of their objective, these methods are comparable with regression imputation. But they are somewhat easier to use if multiple missing values must be imputed in a single record, while the relationships between the variables can accordingly be estimated more accurately. In donor imputation, for each non-respondent *i*, we look for a donor record *d* with as many as possible of the same characteristics as the *recipient i*, insofar as the characteristics are considered to influence the target variable *y*. Subsequently, the donor score, y_d, is used as imputation: $\tilde{y}_i = y_d$. Next, Chapter 7 addresses the problem of *multivariate imputation*, in which there are multiple missing values for a single object, and several solutions for this. Chapter 8 will focus attention on imputation for longitudinal data (viz. panels). Now you can use data from the same object on other time points, possibly without using data from other objects.

In the remainder of section 1.1.2, we will discuss several issues that help determine the selection of the imputation method or the way the methods are used. Incidentally, different experts can make different choices, or use different elaborations of the same method.

1.1.2.2 Imputation variable, quantitative or qualitative

Donor imputation (Chapter 6) can be used for each type of *y*-variable. Regression imputation (Chapter 5) is mainly applied if *y* is a quantitative variable. Usually, the linear regression model is used for this purpose, but there is no objection to using functions other than linear functions of *y*. Even if *y* is a qualitative variable, regression analysis can be used. However, in this case, adapted models are used, such as binary or multinomial logistic regression.

1.1.2.3 Auxiliary information available?

If, for a quantitative *y*-variable, no auxiliary information (x-variables) is used, because there is none available or because it provides virtually no benefit, regression imputation shifts to mean imputation (Chapter 3). We discuss this method separately due to its popularity.

If, for a qualitative *y*-variable, no auxiliary variables are available, we can impute the most commonly occurring value (the modus), which is normally not recommended, or we can randomly select from the categories with probabilities proportional to the observed category frequencies. This last action corresponds to imputation using a random donor (Chapter 6) from the entire population. Imputation without the use of auxiliary information can only be justified if only a few item non-

respondents are involved and the imputations have little influence on the parameters to be estimated.

1.1.2.4 Imputation per subpopulation

We can construct an imputation model for the entire population, or per subpopulation, such as per Standard Industrial Classification (NACE) × Size class (SC) for business statistics. It is useful to make a distinction between such imputation classes if, in the classes, there is little variation in the scores on imputation variable *y* (internally homogeneous) and the scores between the classes vary significantly. Because qualitative *x*-variables can also be included in the imputation model in regression analysis, distinguishing between the subpopulations can also be considered as a part of the modelling, namely the selection of auxiliary variables that correspond strongly with target variable *y* and including these variables in the model with all the interaction terms. Hot deck donor imputation (Chapter 6) is, by definition, only intended for qualitative *x*-variables, and consequently for subpopulations. The y-variables may be qualitative or quantitative.

1.1.2.5 Selection of auxiliary variables or subpopulations

The selection of variables and interactions is not discussed in detail here. Just as regression analysis, it is a part of multivariate analysis which has a lot of literature dedicated to it. You will look for auxiliary variables that correlate strongly with the target variable *y* and, preferably, explain the selection effect as accurately as possible. It is usually a question of trial and error and common sense, but forward or backward search procedures can also be used to automatically add *x*-variables to the model or remove them. There are also automatic search procedures to select homogeneous imputation classes (qualitative *x*-variables), such as WAID (co-developed by Statistics Netherlands) and the SPSS module Answer trees. Several guidelines will be provided in the final chapter.

You can set a standard for the fraction of explained variance of the model for the respondents (R^2). Usually, such a measure will be a quality standard for the strength of the linear relationship between *y* and the *x*-variables.

1.1.2.6 Imputation with or without disturbance term (y quantitative)

For a missing value on *y*, one can impute the best possible prediction according to the regression model. If this is done for all the missing values, then the imputation is "too perfect". All the imputed records then satisfy the imputation model perfectly. As a result, the imputations are often useless in further analyses of the micro data file, or even sometimes in simple tables, which is a reason to 'flag' the imputed values (section 9.1). A well-known example concerns national population statistics, where, for an unknown age of a husband or wife, the imputation rule was used stating that the husband is two years older than the wife. Such an imputation model can potentially be good for the age distribution of both men and women. But researchers using the data material made the 'surprising discovery' that there was a peak in the age difference between men and women.

In general, the imputation of the best possible prediction according to the regression model creates an underestimation of the variation in the scores ('regression to the mean'). This leads to distributions that are too peaked and tail areas that are too thin, especially if *y* has many missing values and the regression explains little of the variance of *y* (small R^2). This effect is the strongest in mean imputation. This does not form an obstacle for the estimation of means or totals, but it does for the estimation of distributions (such as an income distribution) and dispersion measures.

For an accurate determination of the distribution it is advisable to add a random disturbance to the best possible prediction. In regression analysis, we can choose between (1) sampling from a normal probability distribution, and (2) adding the residual of a randomly sampled donor. In Chapter 5, we make a distinction between regression imputation with and without the addition of such a residual. In Chapters 3 and 4, for mean imputation and ratio imputation, we only discuss the version without the disturbance term. Adding a disturbance term then falls under regression imputation.

In donor imputation, a residual is used implicitly, namely the residual of the randomly or non-randomly selected donor. The dispersion in the distribution of *y* is therefore retained.

Rubin (1987) observed that, after adding a random disturbance, the variance of *y* is still underestimated, as a result of the uncertainty of the imputation model. This underestimation can be counteracted by using *multiple imputation*. Multiple imputations are performed for each missing value by creating multiple parameter estimates, random disturbances or models. Adding the variance between the imputations per record ensures an unbiased estimation of the variance of *y*.

1.1.2.7 Deterministic or stochastic imputation

If a random selection is made from donors or from a distribution of residuals, this is referred to as *stochastic imputation*. Because of this randomness, the imputations are not reproducible. In *deterministic imputation*, the imputations are reproducible, given the chosen imputation model. In many cases, the distinction between stochastic and deterministic imputation is analogous to the distinction between using and not using a residual as discussed in the previous subsection. Nearest neighbour imputation, including predictive mean matching, however, is deterministic, because the donor is fixed using a certain distance function.

1.1.2.8 Choice between regression and donor imputation / x-variables, qualitative or quantitative

The choice between regression and donor imputation is often not self-evident. This is mainly because the actual, missing values are unknown. It is not possible to assess which model is better. But we will still provide a number of issues that can have an influence on this choice.

- In regression analysis and nearest neighbour, both qualitative and quantitative *x*-variables can be included. In hot deck donor imputation, only qualitative

variables can be included, unless the quantitative variables are discretised in advance. However, in this case, the quantitative aspect of the variable is partially lost.

- In hot deck donor imputation, there is sometimes a limitation in including important *x*-variables in the model compared to regression imputation. It is required to include all the interactions between the qualitative variables, which means the number of parameters can be large compared to the sample size. In the regression model, a smaller number of parameters can be used.
- By categorising quantitative *x*-variables, replacing them with a series of dummy variables (one per category), we lose information. But if there is a strongly non-linear relationship with y, this categorisation creates a larger explained variance.
- In donor imputation, the imputed donor score is always a valid value. If, for example, *y* can only be an integer, then the regression prediction will virtually never be an integer, while in donor imputation it is possible to impute only integers. In donor imputation, the recipient record also automatically satisfies the edit rules if the donor record satisfies them and the matching of donor and recipient is exact on the *x*-variables.
- If multiple values are missing in a record, donor imputation is easier to use; see Chapter 7 about multivariate imputation.

1.1.2.9 Weighting – yes/no

In most of the methods to be discussed, there is an option in imputation to weight the item respondents unequally, for example, by assigning them weights inversely proportional to the inclusion probabilities (probability of being in the sample), or weights that result from the reweighting for compensation of the selective unit non-response.[1] In linear regression imputation, this means that a weighted least squares estimation is performed, and in hot-deck donor imputation it means that potential donors with a low inclusion probability, and therefore a large inclusion weight, have a greater chance of being a donor than potential donors with a high inclusion probability. Weighting does not have an influence on deductive imputation and on nearest neighbour.

No clear recommendation can be provided about the use of weights. In terms of the model, every outcome is measured equally reliably, if one assumes identically distributed disturbances, regardless of the inclusion probability or response probability. Confidence in the imputation model therefore means that weighting does not need to be used, and it is even better not to use it, because weighting makes the standard errors larger. If we can include the variable with weights, or the variables forming the basis for the weighting, as explanatory variables in the model, weighting is also unnecessary. An option therefore is to provide for this in the selection of *x*-variables. More information can be found about this in Pannekoek and Israëls (2000). However, from the perspective of sampling theory, the answers of a

[1] Incidentally, the item non-respondents also have a raising weight.

sample unit are ‘representative’ for population elements that are not selected, just as if they would have given the same answers. Based on this principle (or assuming a random unit non-response), weighting is needed to obtain sample-unbiased estimators. For donor imputation, Kalton (1983) offers several methods in which the probability of being a donor is proportional to the weight. It can be useful to also ensure that the donor and recipient are given a similar weight, to prevent an object with a very small weight from being the donor for a recipient with a very large weight, as a result of which the weight of the donor increases disproportionately (it receives too much weight). We can also try to prevent this by including the weighting variable or the auxiliary variables that form the basis for the weighting as categorical *x*-variable(s).

Sometimes, the need is felt to impute a score not only for the item non-respondents, but for all the objects not occurring in the sample. We call this ‘mass imputation’, even if it concerns only one target variable *y*. Naturally, a register or sampling frame is needed. For the imputation of the non-sample units, it is also true that weighting is less necessary to the extent that the weighting variables are included as *x*-variables in the model. But it can also be the case that this is not possible, because the weighting variables are only known for the sample units. Then weighting is an option. After mass imputation, we can easily calculate totals and means for *y*. In a weighted hot-deck procedure, this corresponds to the use of the post-stratification estimator, and for the weighted least squares estimation with the regression estimator; see the theme ‘Sampling Theory’, subthemes ‘Sampling designs and Weighting methods’ (Banning et al., 2010). Such estimators are also called ‘synthetic estimators’ and are discussed in the subthemes ‘Synthetic estimation and Small area estimators’ of the theme ‘Model-based estimation’ (Boonstra and Buelens, 2011). However, there the estimators are directly calculated, without adding imputations to the data file.

1.1.2.10 Other issues

The following issues, which do not directly influence the method selection but which do deserve attention, will be discussed briefly in Chapter 9:

1. Flagging / documentation;
2. Dealing with outliers;
3. Selection of auxiliary variables;
4. Non-negative variables with many zeroes;
5. Combination of methods (hierarchy).

1.2 Scope and relationship with other themes

Item non-response is distinct from *unit non-response*, in which someone does not participate in the survey at all, or part of the objects in a register are missing. The researcher must determine whether, in the case of partial response, enough answers have been given to include the record, or to designate it as unit non-response. In this case of ‘true’ non-response, weighting is an option; see the theme ‘Weighting as correction for non-response’. As described in section 1.1.2.9, after imputation, some

total estimators correspond with certain weighting methods and can also be considered as synthetic estimators; see Boonstra and Buelens (2011).

We make a further distinction between imputation and derivation of new variables that are created as a function of variables already existing in the file. In imputation, missing values are created for an existing variable.

In the editing process (see the Methods Series, theme 'Data editing: detection and correction of errors'), errors are detected and corrected. If the original value that is considered incorrect does not play a role in the correction, we also see the correction as an imputation. Here, a missing value is actually created, by first designating the incorrect value as a missing value. However, sometimes the original value does have an influence on the value to be assigned, such as in the 'thousand-errors' in business statistics. The definition of imputation in section 1.4, makes it clear that this is not considered an imputation.

The definition of imputation does not imply that the file is internally consistent after imputation, in the sense that all the edit rules are satisfied. However, it is possible to include an extra requirement in the imputation process that the imputed values must comply with all (or some) edit rules, such that no forbidden inconsistencies or non-admissible values arise as a result of the imputation. This requirement can be satisfied by including edit rules as restrictions in the imputation, or by editing the unrestricted imputations afterwards. This second option sometimes leads to an iterative process. In large surveys with many variables and with records with multiple missing values, inconsistencies cannot always be avoided, even if multivariate imputation methods are being used.

1.3 Place in the statistical process

Imputation is a part of the statistical processing (throughput). It is not a necessary process step: one can decide to leave the fields empty and to resolve the problem by weighting or during the secondary analysis.

It important that the missing values have been clearly indicated in the file in earlier process steps. This can be done by leaving the field vacant, or by using special codes such as -1, 9 or 99 if this does not lead to confusion. It is more problematic if zeroes have been filled in for missing values, which does happen in business statistics, unfortunately. In this case, it is no longer possible to make a distinction between missing values and real zeroes. This also creates problems for the editing.

Often, imputation is a follow-up to the detection of errors, as described in the introduction to this theme report. As stated previously, we consider the correction of such errors as imputation only if the original value no longer plays a role in the correction step. After editing and imputation, the micro file is suitable for aggregation and tabulation. In sampling surveys, estimation procedures will be needed.

Later in the process, you will usually be happy to work with imputed files, and thankfully make use of the imputations. However, there are still situations in which

you would want to ignore the imputations, such as when performing secondary analyses on micro data files, but also when determining confidence intervals. These wishes can be also satisfied by 'flagging' the imputed values during the imputation process (see Chapter 9). This flagging of imputed values should be obligatory.

1.4 Definitions

Concept	Description
Item non-response	erroneously missing value(s) from a respondent
Item non-respondent	an object that erroneously did not respond on a certain variable
Imputation, imputing	determining and introducing a (new) value in a place where a value is missing or has been designated as 'unknown'
Imputed value, imputation	value that is filled in for a missing value
Imputation variable	the variable on which missing values are imputed
Imputation classes	subpopulations in which separate imputation algorithms are used
Deductive imputation (logical imputation)	imputation in which a value is imputed on a logical basis without a probability mechanism, also when it is not 100% certain the value is correct
Donor imputation	imputation in which the missing value is taken from a donor record that has as many of the same characteristics as the recipient as possible
Multivariate imputation	imputation with multiple missing values per record
Mass imputation	imputation for all the missing values in the population on a certain variable
Longitudinal imputation	imputation in which values are used for the same variable at other times/periods of the same object or other objects. This imputation can also be multivariate.

1.5 General notation

We use the following general notation in this theme:

i = index for object (record);

y = target variable, variable of interest;

y_i = score of object i on target variable y; we assume that the observed score does not contain a measurement error;

obs = set of objects for which y_i is *obs*erved;

mis = set of objects for which y_i is *not* observed (*mis*sing);

$\tilde{y}_i$ = imputed value for missing y_i.

Specific notation will be introduced for most of the methods.

2. Deductive imputation

2.1 Short description

In general, imputations are predictions for the missing values, based on a model. In some cases, however, imputations can also be derived directly from the values that were observed in the same record, using derivation rules that do not contain any parameters to be estimated, such as is the case in models.

Example 1. Marital status is unknown, but the person in question is 10 years of age. It can be derived with certainty that this person is unmarried.

Example 2. A company survey asks about the total turnover (O), turnover from the main activity (O1) and turnover from sideline activities (O2). If one of these three forms of turnover is missing, it can be calculated using the rule: O1+O2=O.

The above imputation rules are examples of *deductive* or *logical imputation*. In this imputation method, you examine whether it is possible, based on logical or mathematical relationships between the variables, to unambiguously derive the value of one or more of the missing variables from the values that were observed. For the missing variables for which this is possible, this unique value is the deductive imputation.

Imputation rules can also be applied if the rule does not necessarily always have to hold true, but only very probably holds true. Here, we also talk about deductive or logical imputation.

2.2 Applicability

For deductive imputation, it is not necessary to specify or estimate models. With only the edit rules as input, the process can be performed completely automatically. Furthermore, deductive imputations are, in a way, the best possible imputations. They are exactly equal to the actual values if the other values in the record are correct. Given this last condition, it is important to perform the method after as many as possible errors have been detected and then corrected (systematic errors), or have been designated as ‘missing’. Deductive imputation is then the most logical subsequent step. Model-based and donor methods can be used afterwards. For estimating the parameters, these methods can profit from the values already filled in deductively.

In view of the advantages of the method, it will always have to be determined what options there are for deductive imputation.

2.3 Detailed description

2.3.1. Simple imputation rules

Many deductive imputations can be performed using simple rules in 'if-then' form, for example:

if *marital status = unknown and age < 15* **then** *marital status = unmarried.* Or
if *total labour costs = unknown* and *employees on the payroll* = 0 **then** *total labour costs* = 0.

These rules are compiled by specialists familiar with the content, and can each be applied with many different types of software.

2.3.2 The use of equality restrictions

A particularly rich source for deductive imputations is formed by the extensive systems of equations that should apply for Structural Business Statistics. This can amount to around 100 variables with 30 equality restrictions. Most of these equality restrictions are in the form 'Total variable' = 'sum of the Subtotals (or sub-items or specifications)'. If, in such a case, one of the subtotals or the total is missing, it is immediately clear with which value the missing variable should be imputed. There is a single equation with a single unknown. In practice, many variables occur in many equations. This means we have a system of equations, usually with multiple missing variables, for which it is not immediately clear whether the values of some missing variables can be uniquely determined for this system, and what these unique values would be. Below we describe a method to automatically generate the deductive imputations for such systems of equations.

Suppose that a record consists of p variables and that q linear equality restrictions apply to these p variables. These restrictions can be represented in the form

$$\mathbf{R}\mathbf{y} = 0 \qquad (2.3.1)$$

where y is the p-vector with variables, and $\mathbf{R}$ is a q×p matrix in which each row represents one restriction. For example, the operating income block consists of the following five variables:

Table 1. Five variables from the operating income block

Net turnover from main activity	y_1
Net turnover from other activities	y_2
Total net turnover	y_3
Total other operating income	y_4
Total operating income	y_5

Two restrictions apply to these variables: $y_3 = y_1 + y_2$ and $y_5 = y_4 + y_3$. These restrictions can be formulated in the form (2.3.1) where

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 \end{bmatrix}.$$

If the vector with variables *y* consists of *o* observed values and *m* missing values, then, after a permutation of elements, this vector can be partitioned as $\mathbf{y} = (\mathbf{y}_o', \mathbf{y}_m')'$, in which $\mathbf{y}_o$ is the *o*-vector with the observed values of *y* and $\mathbf{y}_m$ the *m*-vector with the missing values. If we partition **R** in accordance with the partitioning of **y**, we can write

$$[\mathbf{R}_o \quad \mathbf{R}_m]\begin{bmatrix}\mathbf{y}_o \\ \mathbf{y}_m\end{bmatrix} = \mathbf{0}, \tag{2.3.2}$$

such that, say,

$$\mathbf{R}_m\mathbf{y}_m = -\mathbf{R}_o\mathbf{y}_o = \mathbf{a}. \tag{2.3.3}$$

This last expression is a system of linear equations in the missing values $\mathbf{y}_m$. The intention of deductive imputation is to resolve as many as possible missing values from this system.

For a system of linear equations, it is common practice to make a distinction between three cases: I) there are no solutions (the system is inconsistent), II) there is exactly one solution, and III) there are an infinite number of solutions.

Case I occurs if the rank of $\mathbf{R}_m$ is not equal to the rank of [$\mathbf{R}_m$ **a**]. If the restrictions are formulated in such a way that no contradictions arise as a result, then case I can only occur if there are errors in the data. These types of errors, which cause violations of the restrictions, are however detected first in economic statistics. Next, a number of values are characterised as incorrect and then designated as 'missing'. The new missing values are indicated in such a way that there are imputations for the missing values that satisfy the restrictions. If we deal with the violation of restrictions as described above, case I can therefore no longer occur.

Case II occurs if the rank of $\mathbf{R}_m$ is equal to the number of missing values *m*. All missing values can then be deductively imputed; there is only one value for $\mathbf{y}_m$ that satisfies the restrictions.

In general, however, we will encounter case III; there are an infinite number of solutions for $\mathbf{y}_m$. In this last case, however, it is possible that some elements of $\mathbf{y}_m$ have the same values in all possible solutions. These elements can be deductively imputed.

The set of solutions for $\mathbf{y}_m$, say $\tilde{\mathbf{y}}_m$, is given by (see, for example, Rao (1973), page 24)

$$\tilde{\mathbf{y}}_m = \mathbf{R}_m^-\mathbf{a} + (\mathbf{R}_m^-\mathbf{R}_m - \mathbf{I})\mathbf{z} = \mathbf{b} + \mathbf{C}\mathbf{z} \tag{2.3.4}$$

where $\mathbf{R}_m^-$ is a generalised inverse of $\mathbf{R}_m$ (in other words, an m × q matrix for which $\mathbf{R}_m\,\mathbf{R}_m^-\,\mathbf{R}_m = \mathbf{R}_m$), and **z** an arbitrary m-vector. Because **z** can be selected arbitrarily, (2.3.4) generally generates an infinite number of solutions for $\mathbf{y}_m$, there

is only a unique solution if $\mathbf{R}_m$ is of full rank is and $\mathbf{R}_m^-$ is therefore the regular inverse. If some elements of $\mathbf{y}_m$ are the same for all possible solutions, i.e. for each arbitrary value of **z**, then the corresponding rows of **C** must contain only zeroes**.** These elements can thus be easily detected, and they can be deductively imputed with the corresponding values of **b**.

2.3.3 The use of non-negativity

Another possibility to perform deductive imputation is to use the non-negativity of many variables. Suppose, for example, that only two sub-items of an addition of eight items were observed, but that these do add up to the reported total. If the missing sub-items are not allowed to be negative, then they all can be imputed with zero because their sum must be zero.

To find these types of solutions, we again consider the equality $\mathbf{R}_m\mathbf{y}_m = \mathbf{a}$. Suppose that there is an element a_j of **a** that is equal to zero. For the corresponding row, $\mathbf{r}_{m.j}$, of $\mathbf{R}_m$, it is then true that $\mathbf{r}'_{m.j}\mathbf{y}_m = 0$. Now, if, for all elements of $\mathbf{y}_m$ that correspond with the non-zero elements of $\mathbf{r}_{m.j}$, it is true that

i) these elements cannot be negative,

ii) the corresponding non-zero elements of $\mathbf{r}_{m.j}$ are all negative or all positive,

then these elements of $\mathbf{y}_m$ are equal to zero.

The deductive 0 imputations derived in this way for the missing values $\mathbf{y}_m$ are therefore given by

$$\tilde{y}_{mj} = 0 \quad \text{if} \quad a_j = 0 \quad \text{and conditions i and ii are satisfied.} \qquad (2.3.5)$$

2.4 Example

An example where deductive imputation was used in business statistics is described in Pannekoek and Tempelman (2005). This example concerns data from Structural Business Statistics that relates to the Wholesale Sector and the Retail Sector. The data concerning the Wholesale Sector consists of 875 companies (in size classes 4 to 9) and 102 variables. There are 30 equality restrictions that apply to these variables, and there are also 26 simple imputation rules formulated by using a relationship in the form 'if $y_1 = 0$ then $y_2 = 0$', and use is made of the non-negativity of almost all these variables. The data for Retail Sector consists of 1242 records (in size classes 0 to 3) and 54 variables to which 15 equality restrictions apply, and there are also 21 simple imputation rules formulated in the same form as for the Wholesale Sector. The non-negativity was used in this case too.

This Structural Business Statistics data has already undergone several processing steps, in which very obvious errors were corrected. This includes, for example, uniform thousand-errors or observations which were erroneously negative.

Furthermore, during this step, empty totals and subtotals were also filled in if the related sub-items were filled in. This last step is an initial deductive imputation step. In addition, the error localisation algorithm of the programme CherryPi checked all the edit rules and, if the edit rules were violated, the necessary values were characterised as incorrect and then designated as 'missing'. The missing values in these files were the result of both partial non-response and detected errors.

All possible deductive imputations were performed on this data using the equality restrictions and the simple imputation rules. The results are shown in Table 2.

Table 2. Numbers of deductive imputations in the Wholesale Sector and the Retail Sector

	Wholesale Sector		Retail Sector	
Number of missing values	35068		27693	
Number of deductive imputations	24048 (69%)		12927 (47%)	
Of which equal to zero		22647 (94%)		11708 (91%)
Of which not equal to zero		1401 (6%)		1219 (9%)
Remaining missing values	11020		14766	

This table shows that deductive imputation is highly effective. In this way, for a large part of the missing values (69% and 47%), imputation can be performed – without an imputation model and without adaptations of imputations – using the only possible value that satisfies all the edit rules.

The deductive imputations in Table 2 are mostly (more than 90%) equal to zero. We should point out that these are not the only deductive imputations. In the T040 step, a number of deductive imputations have already taken place that are not zero: the filling in of empty subtotals. Many of the zero imputations are due to the fact that reporters left the questions about specific costs items where they did not have any expenses as empty fields, instead of answering with 0. The same is true for income from specific components of the operating income. Using deductive imputation, a large number of these zero values not filled in can be recovered. Incidentally, always imputing a zero in a field that was not filled in is not recommended, even if this is not in conflict with the edit rules. Pannekoek and Tempelman (2005) demonstrate that this can sometimes result in significant bias in the publication totals.

3. Mean imputation / Group mean imputation

3.1 Short description

In *mean imputation*, a missing value is replaced by the mean score on the variable concerned for objects that have a valid score.

In *group mean imputation*, a missing value is replaced by the mean score on the variable concerned for objects that have a valid score *and* are in the same subpopulation as the item non-respondent.

Mean imputation leads to a peak in the distribution, because the same mean is imputed for each missing value. In group mean imputation, there are a number of smaller peaks.

3.2 Applicability

No auxiliary information is used in pure mean imputation. This method is therefore only recommended if no auxiliary information is available or when the available auxiliary variables are only marginally associated with the imputation variable y. If the fraction of missing values on a variable is very small, and the imputations will have a marginal effect on the parameter to be estimated (such as the population total), mean imputation may be permissible due to efficiency considerations. However, using this rather overly simplistic method should be an exception.

Auxiliary information is used in group mean imputation, and this involves a classification into groups (subpopulations, imputation classes) based on one or more qualitative variables. The more homogeneous the subpopulations are with respect to the variable to be imputed, the better the imputations, based on the assumption that the classification into subpopulations not only effectively discriminates among the respondents, but also among the item non-respondents (see section 1.1.2.8).

As stated above, pure mean imputation results in peaked distribution. The method is therefore potentially suitable if the output is limited to estimation of population means and totals. The fact that a complete data file is obtained because of the imputation guarantees the consistency of the aggregated outcomes. Pure mean imputation, however, is not suitable for estimating an income (or other) distribution or for estimating a dispersion measure such as the standard deviation. It does not generally lead to high-quality individual imputations, but no imputation method offers this type of guarantee.

In group mean imputation, the peak of the distribution is usually much smaller, because the variation between the groups is included in the imputation; only the variation within the groups is disregarded. If the ratio between this interclass and intra-class variation is large, this method can also be used to reasonably estimate the dispersion measures, given the validity of the imputation model.

3.3 Detailed description

In accordance with the notation from section 1.5, the imputed value $\tilde{y}_i$ for a missing score y_i in mean imputation is equal to the observed mean

$$\tilde{y}_i = \bar{y}_{obs} = \frac{\sum_{k \in obs} y_k}{n_{obs}}, \qquad (3.3.1)$$

where y_k is the observed score of the k^{th} respondent and n_{obs} the number of item respondents for variable *y*.

If desired, the objects can be weighted unequally, for example, due to differences in the inclusion probability; see subsection 1.1.2.9 and attribute point 3 in section 3.5. In this case, raising to population figures does not take place using a fixed raising factor N/n (where N is the population size, and n the sample size or the number of respondents), but using individual weights w_i that vary. The resulting imputation

$$\tilde{y}_i = \bar{y}_{obs}^{(w)} = \frac{\sum_{obs} w_k y_k}{\sum_{obs} w_k} \qquad (3.3.2)$$

is then usually a better, less biased estimator of the population mean.

Mean imputation can be used for the non-response in the sample or for the missing values in the population. For each missing value, the same mean is imputed. In most cases, you can apply this method more effectively after first having determined imputation classes. In this group mean imputation, (3.3.1) is replaced by

$$\tilde{y}_{hi} = \bar{y}_{h;obs} = \frac{\sum_{h \in obs} y_{hk}}{n_{h;obs}}, \qquad (3.3.3)$$

where y_{hk} is the observed score of the k^{th} respondent in class *h* and $n_{h;obs}$ the number of item respondents for variable *y* in *h*.

No complex software is required to impute the mean or the group mean. Using SPSS14.0, mean imputation or group mean imputation can easily be applied via Transform \ Replace Missing values \ Method Series mean. The procedure Replace Missing values is intended for time series, and is therefore usable for missing values in longitudinal imputation (Chapter 8).

So mean imputation can be used:

- with the mean of the entire sample or population, or per imputation class;
- unweighted, or weighted with weights w_i.

We will discuss the option of applying the method with a disturbance term in regression imputation in Chapter 5.

3.4 Example

Example 1. *Energy statistics-1*[2]

Until recently, the survey 'Energy use in companies' was used to estimate the energy consumption of companies in the Netherlands. As this survey is no longer being used, efforts are being made to set up a secondary observation process, where the 'usage data per company' from the power companies is used to estimate the total energy use. For this purpose, usage data based on the name, address and city/town details are matched with business units in the General Business Register (ABR).

An example of group mean imputation is to use the mean electricity use per company for each company sector (NACE), such as greenhouse farming.

Example 2. *Structural Business Statistics-1*

In Structural Business Statistics (SBS), subpopulations are formed based on the Standard Industrial Classification (NACE) and the size class (SC). The sample size is too small to distinguish between all cells of NACE × SC. The imputation procedure differs slightly between large and smaller companies.

If auxiliary information about a company with incomplete response is available, for example, in the form of turnover from the previous year or from the Short Term Statistics (STS), then this should clearly be used. Example 2 in section 4 demonstrates how this is done. However, if such information is not available, then group mean imputation can be used. If the turnover is missing, the mean turnover in the imputation class can be imputed. This will often be used for new companies, for which no data from a previous period is available.

3.5 Characteristics

1. After applying mean imputation according to (3.3.1) for all item non-respondents, the unweighted sample mean is equal to the unweighted response mean. If we apply mass imputation by using the response mean not only as the imputed value for the possible item non-respondents, but also for those who are not in the sample, then the population mean estimated in this way is equal to the response mean, and also equal to the direct estimator for the population mean (with raising weights N/n).
2. Likewise, group mean imputation leads to the same overall totals and means as the stratification or post-stratification estimator, if the strata are used as imputation classes.
3. After applying weighted mean imputation according to (3.3.2) for all item non-respondents, the sample mean weighted with inclusion probabilities is equal to the response mean weighted with inclusion probabilities, regardless of the weights of the item non-respondents. Here, weighting (raising) also ensures that

[2] With thanks to Edgar Soufan.

the population estimate is not influenced by the imputations. Likewise, after mass imputation, the population mean is equal to the weighted response mean.

3.6 Quality indicators

Mean imputation results in an underestimation of the variance S_y^2 of imputation variable *y*,

$$\hat{S}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \ , \tag{3.6.1}$$

because, for the item non-respondents, the contribution in the numerator is a zero. If, for $V(\bar{y})$, the variance of the sample mean $\bar{y}$, we use the naive estimator

$$\hat{V}(\bar{y}) = (1-\frac{n}{N})\frac{1}{n}\hat{S}_y^2 = (1-\frac{n}{N})\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.6.2}$$

where y_i is either known, or imputed with $\bar{y}$, then this variance (square of the standard error) would also be underestimated, and therefore also the confidence margin. Using this estimator incorrectly suggests that there is data available for all n objects, rather than only for those who have responded to *y*. The correct estimator is obtained for $V(\bar{y})$ by replacing the sample size *n* in formula (3.6.2) by the number of item respondents n_{obs}, and by only determining S_y^2 over the item respondents. Obviously, the sample mean $\bar{y}$ is equal to the response mean. For group mean imputation, the above applies per group.

See section 5.6 for further quality indicators.

4. Ratio imputation

4.1 Short description

In *ratio imputation*, for variable *y*, a single auxiliary variable *x* is used that is associated strongly with *y*, in the sense that *x* proportional is with *y* in (reasonable) approximation. If *R* represents the relationship between *y* and *x*, the missing value y_i is replaced by

$$\tilde{y}_i = Rx_i \ . \qquad (4.1.1)$$

An example is the determination of an unknown company turnover (y) from the number of employed people (x). For *R*, you will use the mean company turnover per employed person. The most common situation is that *x* measures the same thing as *y*, but in an earlier observation period. We then notate the variables *y* and *x* as y^t and y^{t-1} respectively. Formula (4.1.1) then changes to

$$\tilde{y}_i^t = Ry_i^{t-1} \ , \qquad (4.1.2)$$

where *R* is the relative increase of the variable from period *t-1* to *t*. Generally, *R* is estimated from the data.

4.2 Applicability

Ratio imputation can be applied for missing values on a quantitative variable *y*, if a quantitative (auxiliary) variable *x* can be found which has a more or less fixed ratio with the target variable *y*. You can see formula (4.1.1) as a simple regression equation in which the regression line passes through the origin. This means that no constant term is used. Ratio imputation is therefore a special case of regression analysis (estimated with weighted least squares). If a model with a constant term fits better, or if we want to add extra variables to model (4.1.1), the general regression imputation *may* be more appropriate.

Generally, at Statistics Netherlands, no residual is added to (4.1.1). In many statistics where ratio imputation is used, means and totals are the main output. In the past, as an exception in some turnover statistics, a table was produced with the number of companies that had a higher vs. lower turnover than the previous year. If imputation is applied according to (4.1.2) and R is estimated to be 1.01, then it is assumed for all item non-respondents that they had turnover growth from period *t-1* to *t*, which is unlikely in this situation. For this table, it is therefore necessary to add a residual to (4.1.2). We discuss this addition of a residual further in Chapter 5; see also subsection 1.1.2.6.

Just as in mean imputation, ratio imputation can be applied separately per subpopulation (imputation class). This is done mainly if the ratios between the subpopulations vary strongly. This option is discussed in the next section.

4.3 Detailed description

Often, you will have an auxiliary variable *x* that is more or less proportional to y. If y_i is missing but x_i is known, you can use (4.1.1) as imputation, where *R* is the proportional constant. Generally, *R* is not known and is estimated from the records where *x and y* are known:

$$\hat{R} = \sum_{obs} y_i \Big/ \sum_{obs} x_i \quad . \tag{4.3.1}$$

Substituting this in (4.1.1) gives us

$$\tilde{y}_i = \hat{R} x_i = \frac{\sum_{obs} y_i}{\sum_{obs} x_i} x_i \quad . \tag{4.3.2}$$

So the proportional constant is equal to the quotient (ratio) of the means of *y* and *x* for the item respondents of variable *y*.

In the case that *x* and *y* only differ in the period, formula (4.1.2) changes to

$$\tilde{y}_i^t = \hat{R} y_i^{t-1} = \frac{\sum_{obs} y_i^t}{\sum_{obs} y_i^{t-1}} y_i^{t-1} \quad . \tag{4.3.3}$$

The parameter to be estimated, R, is now the relative increase of the variable from *t-1* to *t*.

Model (4.3.2) can also be applied separately for different subpopulations. Each subpopulation *h* therefore has its own ratio R_h. This may be called *group ratio imputation*. The application of this method is only useful if the linear relationship between *x* and *y* differs strongly, and at least significantly, between the subpopulations. The subpopulations can also not be too small, because this can lead to bias and possibly large standard errors for total estimators. Working with groups usually offers less of a benefit in ratio imputation than in group mean imputation; ratios of groups are usually more homogeneous than group means.

To determine the ratio *R*, there is again the option of weighting item respondents with inclusion weights.

No complex software is needed for ratio imputation. Formulas (4.3.2) and (4.3.3) are easy to calculate after estimating the ratio *R*.

4.4 Example

Example 1. *Energy statistics-2* [3]
For ratio imputation, the total number of employed people or the turnover per company seems to be a good indicator for the level of energy use. It could be

[3] With thanks to Edgar Soufan.

investigated whether different ratio factors exist for different business sectors. It could also be investigated whether expanding to a more general regression model provides a benefit.

Example 2. *Structural Business Statistics-2*
For missing values in Structural Business Statistics, there is an automatic imputation procedure for the smaller (non-crucial) companies, which mainly uses ratio imputation. A fixed order is used for the availability of auxiliary information. This hierarchy, decreasing in quality of the auxiliary information, is:

1. Observation at the same company in year *t*-1 (for all variables);
2. Observation at the same company from Short Term Statistics (STS) of year *t* (only for *y* = turnover);
3. Observation of others companies in the same class (SC × NACE) in year *t*.

If a company has item non-response, we first look to see whether this company had a valid score on that variable in the previous year. If yes, then formula (4.3.3) is applied, where y^t is the variable concerned in year *t*, y^{t-1} in the previous year and $\hat{R}$ a trend correction. For the turnover variables, the trend correction represents the turnover development. This all takes place within a combination of SC and NACE (3-digit) with a minimum number of 15 companies contributing to the cell.

If, however, y_i^{t-1} is unknown, for example because the company was not in the sample the previous year, the second or third option is selected, depending on the target variable. However, these options are not ratio imputations. In the second option, for companies who also participated in the STS for year *t*, the totalised annual turnover is copied exactly; imputed turnovers are not allowed here either. This copying of the value from another file is called 'cold deck'; see Chapter 6. Option 3 is a group mean imputation, with a combination of SC and NACE as the imputation class. Option 3 will usually be used for new companies.

4.5 Characteristics

- A special case of ratio imputation is obtained by using *R=1*. This means that the imputation $\tilde{y}_i$ is equal to x_i. Variable *x* is then a 'proxy variable' for y. If *x* originates from an external source, this is called 'cold deck imputation' (see Chapter 6). An example is that, for a missing value y_i^t, the value from a previous period, y_i^{t-1}, is used. With variables that are stable over time, this can be considered, but often the preference will be to estimate *R*, instead of supposing it equal to 1.
- The ratio $\sum y_i / \sum x_i$ does not change because of ratio imputation. If the ratio estimator (see Banning et al., 2010) is used for raising from sample to population where *x* is the auxiliary variable for y, then the population estimate does not change by including the imputed values.

4.6 Quality indicators

Ratio imputation results in an underestimation of the dispersion of the values of $y_i - Rx_i$ if no disturbance term is included in the model. If, for the variance of the estimated population mean using the ratio estimator, the naïve estimator

$$\hat{V}(\hat{\bar{Y}}_R) \equiv \hat{V}(\hat{R}\bar{X}) = (1-\frac{n}{N})\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(y_i - \hat{R}x_i\right)^2 \qquad (4.6.1)$$

would be used, where y_i is either known or imputed, then the variance would also be underestimated, and therefore also the confidence margins. This incorrectly gives the impression that there are *y*-scores for all n objects, instead of only those that responded to *y*. The correct estimator for $V(\hat{\bar{Y}}_R)$ is obtained by replacing the sample size *n* in the formulas for the ratio estimator by the number of item respondents n_{obs} and only summing over the item respondents.

See section 5.6 for additional quality indicators.

5. Regression imputation

5.1 Short description

In regression imputation, for a missing value y_i , the optimum prediction is imputed that follows from a suitably selected regression model that predicts y from one or more x-variables. The parameters of the model are estimated using the objects with a valid score on y and on most of the x-variables.

Sometimes, a random disturbance term is added to this optimum prediction, to prevent the imputed data set from satisfying the regression model too well.

5.2 Applicability

In regression imputation, the target variable y is quantitative. The explanatory auxiliary variables of the regression model are quantitative, but due to the use of dummy variables, qualitative variables can also be included in the model. In this case, linear regression analysis is also called 'analysis of variance'. Such regressions can lead to values that cannot occur theoretically, such as non-integers if the value range of y only contains integers. Donor imputation – which can to some extent be understood as a form of regression analysis – prevents this problem.

Regression imputation is also applicable for a binary (dichotomous) target variable. For example, a logistic regression model can then be used; see example 3 in section 5.4.

Subsection 1.1.2.6 already explained that, for each item non-respondent to y, either the best prediction can be imputed, or a random disturbance term can be added to this. This choice depends on the goal of the imputation. To estimate means and totals, this type of residual is not necessary, but if you want the dispersion in y to also remain after imputation, then the preference is to add a residual.

In subsection 1.1.2.9, we pointed out the possibility of performing a weighted regression analysis, if the respondents with a higher sample weight should count more. Heterogeneity of the disturbances can be another reason for such an estimate with weighted least squares.

5.3 Detailed description

In the Methods Series, we do not discuss the theory of regression analysis, but rather consider this as general knowledge. There is enough literature available about linear and other types of regression. For model selection, we limit our comments to those in subsection 1.1.2.5 and section 9.3.

In regression imputation, a regression model is assumed for the prediction of y by means of a set of auxiliary variables $x_1,\ldots,x_p$. The regression model is as follows

$$y = \alpha + \beta_1 x_1 + ... + \beta_p x_p + \varepsilon = \alpha + \beta' x + \varepsilon \ , \qquad (5.3.1)$$

where x is a p-vector with variables $x_1,\ldots,x_p$, α a scalar parameter, β is a p-vector with parameters and $\varepsilon \sim N(0,\sigma^2 I)$ is a vector with n_{obs} independent, normally distributed disturbances with variance σ^2; I is the identity matrix. We can also consider the model without a constant term, by leaving out α.

The parameters α and $\beta_1,\ldots,\beta_p$ are estimated using the records for which both y and the auxiliary variables are observed. This results in parameter estimators a, $b_1,\ldots, b_p$. In most cases, the least squares method is used as the estimation method. This results in a predictor variable

$$\hat{y} = a + b' x \ , \qquad (5.3.2)$$

with the least squares estimators a and b for α and β respectively. This predictor variable is defined for both item respondents and item non-respondents.

There are now two ways to determine an imputation $\tilde{y}_i$ for the item non-respondents:

1. Without a disturbance term:

$$\tilde{y}_i = \hat{y}_i = a + b' x_i \ , \qquad (5.3.3)$$

2. With a disturbance term:

$$\tilde{y}_i = \hat{y}_i + e_i = a + b' x_i + e_i \ . \qquad (5.3.4)$$

In accordance with subsection 1.1.2.6, there are two ways to determine the disturbance term e_i:

a. $e_i = e_d$ where e_d is the residual of an arbitrary or specially selected donor.

b. e_i is a selection from the normal distribution with the expectation 0 and variance σ^2.

In both cases, the residual is determined using the regression model.

Non-linear models have a more general form:

$$y = f(\beta' x) \ . \qquad (5.3.5)$$

The disturbance term ε can be added to this model, or it can be implicitly contained therein.

In the case of a binary y-variable with scores 0 and 1, a logistic regression model can be used:

$$\ln \frac{p}{1-p} = \alpha + \beta_1 x_1 + ... + \beta_p x_p \equiv \alpha + \beta' x \ , \qquad (5.3.6)$$

where p is the probability that y takes the score of 1, given the x-variables and the model. In the case of a missing y-value, the β-parameters can be estimated, for

example, using the maximum likelihood, and subsequently the imputation probability p on the score 1 by means of

$$\hat{p} = \frac{e^{\hat{\alpha}+\hat{\beta}'x}}{1+e^{\hat{\alpha}+\hat{\beta}'x}} = \frac{1}{e^{-(\hat{\alpha}+\hat{\beta}'x)}+1} \ . \qquad (5.3.7)$$

In SPSS14.0 \ Analyze \ Regression, the predicted values according to (5.3.2) can be saved using SAVE \ Unstandardized predicted values, both for linear and non-linear regression. In this case, a variable is created with the default name PRE_1, which contains the value $\hat{y}_i$ for both the item non-respondents and the item respondents. The variable y after imputation is then obtained by replacing the $\hat{y}$-score for each item respondent by the true score y_i. (Of course, in the original y-variable, we can also replace the missing values by the model scores $\hat{y}_i$, even though the imputed values will have to be flagged.) In 'binary logistic' and 'multinomial logistic', the predicted category probabilities can be saved. This can be done for both the item respondents and the item non-respondents, so that the imputations according to (5.3.7) are obtained immediately.

5.4 Examples

Example 1. *Energy statistics*[4]

To use regression imputation to determine missing energy usage figures, consider a regression model with the number of employed people, turnover and NACE as *x*-variables. It is possible that it is not useful to use both the number of employed people and the turnover for this purpose; we could make the energy use dependent on only the turnover in each business sector. This is more general than in the example 1 in section 4.4 for the ratio imputation per business sector, because the constant term can also be included in the regression equation and because non-linear relationships are also possible.

Example 2. *Structural Business Statistics-2*

In the raising of the Structural Business Statistics, for each 'basic cell' (combination of SC × NACE), the regression estimator is used for companies for which the VAT turnover (x) is known in addition to the reported turnover (y). This raising from response to population makes imputation for missing y-turnovers unnecessary. However, the same results are obtained if we first use a regression imputation according to formula (5.3.3) with the same regression from y to x, and then raise the sample (including the item non-response) to the population, at least if this is dealt with in the same way as the sample weights.

[4] With thanks to Edgar Soufan.

Example 3. *Household statistics*

Each year, Statistics Netherlands receives a copy of a part of the Municipal Personal Records Database (*Gemeentelijke Basisadministratie* – GBA) on 1 January. The GBA has information about the residents at each address, including their family relationships. However, the household composition is missing. For the Annual Household Statistics, it is essential to know which people living at a particular address form a single household according to the current definition. Until the 1999 statistics year, this statistic was based on the 'household box' of the Labour Force Survey (LFS). Starting in 1999, the GBA became the basis and the variables 'number of households' and 'household composition' were derived from the family structure (Harmsen and Israëls, 2000). For more than 90% of the GBA addresses, the data is known based on these derived variables. For the other addresses, however, neither the number of households nor the exact composition is known. Imputations are performed for these addresses, with separate imputation models for different situations.

We discuss here the simplest type of addresses with an unknown household composition: addresses with two people living there who are not in a family relationship (in short: addresses with two 'separate' people). For these addresses, it is not known whether the two residents together form one household, or whether they are both single. First, deductive imputation (see section 2.1) is used, using a derivation rule: if, according to the GBA, both people moved to the same address on the same date, then 'one household' is imputed. This will produce a slight underestimation of the number of households. The remaining addresses are matched with the LFS sample. For 1999, this produces a matching sample LFS × 'GBA with two separate people' for 1662 addresses. An imputation model was made based on these sample addresses.

Using visit accounts and the LFS household box, it was determined whether each sample address contained one or two households. This was sometimes complicated due to non-response or due to deviations between the actual and registered residence. The probability of there being two households corresponded strongly with the age of both people (especially the difference in age), whether or not they were of the same gender, the degree of 'urbanness' and the number of unmarried people at the address. For 1999, the logistic regression model (5.3.6) where p is the probability of two households was:

$$\ln[\hat{p}/(1-\hat{p})] = \textit{(.1470 * DIFAGE) + (.0527 * AVGAGE) – (.3916 * URB) + (.7513 * NONMARR) + (.0888 * MM) – (6.4201* MW) – (5.7154 * WM) – (DIFAGE * SAMEGEN) - (.0631 AVGAGE * SAMEGEN) – (.9184 * NONMARR * SAMEGEN) + constant.}$$

The terms above are defined as follows

- DIFAGE = abs. age difference;
- AVGAGE = average age
- URB = degree of 'urbanness' (scores 1-5, with 1 = high and 5 = low);
- NONMARR = number of unmarried people (0, 1 or 2);
- SAMEGEN = 2 if two people of the same gender, otherwise 1.

The combination of the gender of the oldest and youngest person contains four categories entered as dummy variables MM, MW, WM and WW with scores 1 (belonging to the category concerned) and 0. As a reference category, WW was not included in the equation.

The plan was, for these and other groups, to perform the weighted logistic regression using inclusion weights (including oversampling of people registered as unemployed) or the LFS raising weights. This was not done for the addresses with two separate people, because these weight variables did not make a significant contribution to the model. For some other matching groups, weighted regressions were performed, which is more in line with the sampling theory, because in this imputation, the sample was supplemented up to the population; see subsection 1.1.2.9.

Finally, formula (5.3.7) was used, for each non-sample address from the matching sample, to estimate the probability of there being two households, after which either a '1' or '2' was imputed for each record using a random selection mechanism. Cumulative rounding was used to prevent rounding up or down from occurring too frequently.

The above example is a case of register imputation (mass imputation): an imputation is performed for all addresses with a missing score on 'number of households'. Moreover, the missing scores are very selective. 'Number of households' is a variable that can be derived from the GBA, but only for specific groups. Only by matching with an external sample file did information about the number of households become available for those groups.

5.5 Characteristics

1. If, in formula (5.3.1), no auxiliary variables *x* are used, this formula changes to $y = \mu + \varepsilon$ where μ is the expected value of *y*, and formula (5.3.3) changes to $\tilde{y}_i = \hat{\mu} = \bar{y}$. This is mean imputation (Chapter 3).

2. If no constant term is used and only the quantitative auxiliary variable x_1, then formula (5.3.1) changes to $y = Rx + \varepsilon$, and (5.3.3) to formula (4.1.1). Under certain heterogeneity assumptions, the weighted least squares estimator leads to ratio imputation according to formula (4.3.2).

3. If $\tilde{y}_i$ is imputed according to formula (5.3.3), then the inclusion of the imputations does not affect the estimate of the population total, if, for this, the regression estimator is used with the same model as the imputation model; see the 'Sampling Theory' theme (Banning et al., 2010). As discussed in subsection 1.1.2.9, such estimators are also called 'synthetic estimators' (Boonstra and Buelens, 2007).

4. If the imputation is repeated periodically, the individual mutations/changes are strongly overestimated (see Chapter 8).

5.6 Quality indicators

It is important to be aware of the quality of an imputation. A problem in this regard is that the actual value is usually unknown. Often, means differ before and after imputation. This is not necessarily a cause of concern because the item non-response could have been selective. If there is an overlap with other surveys, external validations can be performed to obtain an impression of the quality of the imputation produced. Usually, however, there are definition and population differences between the various studies such that opportunities for these types of validations are limited.

Because, generally, no real assessment of the quality of imputations is possible, the quality indicators below for regression imputation are based only on the model as fitted for the item respondents.

- Fit measures. For linear regression analysis with the least squares estimator, R^2 can be used to quantify the strength of the model among the respondents, and therefore to compare different imputation models with one another. The gains in R^2 must be set off against the extra number of degrees of freedom. This fit measure also applies for donor imputation (Chapter 6), which can be viewed as imputation based in regression on dummy variables. For some non-linear models, the likelihood can be used as an indicator, or a variable derived from the likelihood, such as AIC or Nagelkerke's R^2. Incidentally, it is theoretically possible that model A, despite being a better fit than model B among the item respondents, is a poorer fit among item non-respondents, in other words, it has larger residuals on average.

- Validation/simulation. Another possibility to obtain an impression of the quality of an imputation method is to perform a simulation experiment. Valid values are temporarily left out, and subsequently new valid values are imputed for these left-out values. All the item respondents can be left out one by one or in small groups, but it is also possible to limit the values left out to a part of the item respondents. If the subsequently imputed values $\tilde{y}_i$ are similar to, or, for qualitative *y*-variables, are even equal to the original values y_i, then this inspires confidence in the imputation method. By defining a suitable distance function, it is possible to choose the most appropriate method or the most appropriate model. An example of a distance function is the mean absolute deviation of the imputed from the actual values, $\frac{1}{I}\sum_{i=1}^{I}|\tilde{y}_i - y_i|$, where I is the number of imputations considered. At aggregate level, you can use as a distance function the mean absolute deviation – over the simulations – between the aggregate values with and without imputation, $\frac{1}{T}\sum_{t=1}^{T}|\tilde{Y}_t - Y_t|$. A similar experiment was conducted in Schulte Nordholt (1998).

- Calibration with external data is generally not possible or difficult to use, both for the individual imputed values and at aggregate level. Obtaining the missing data by approaching item non-respondents is also not easy to achieve.

- The calculation of variance and bias is generally complicated. One may have to deal with sampling errors, selective non-response, systematic errors in the imputation model and uncertainty in the imputation model (due to the addition of residuals or the random designation of donors). More information about variance calculation can be found, for example, in Rao (1996). Sometimes, exact variances can only be calculated using *multiple imputation* (Rubin, 1987). For each missing value, different values are imputed. Adding the variance between the imputations of the same record ensures an unbiased estimate of the variance of the population mean. There are practical problems with multiple imputation, such as data storage, more complicated calculations of simple population parameters and more complex analyses of the data. Furthermore, the underestimation with 'single' imputation is often not so large. It is possible that multiple imputation will be used more often in the future.

6. Donor imputation (hot deck imputation)

6.1 Short description

In *donor imputation* (hot deck imputation), for each item non-respondent i , you look for a donor record d in the file with as many of the same characteristics as possible, insofar as these are considered to influence the imputation variable(s) y. For this donor, the score, y_d, is used as imputation:

$$\tilde{y}_i = y_d \,. \tag{6.1.1}$$

The item non-respondent is called the 'recipient'.
There are different ways of finding a donor. These can be broken down into:

1. Methods that utilise imputation classes;
2. Methods that look for a donor by minimising a distance function (nearest neighbour hot deck).

Examples of the first class of methods are *random hot deck* and *sequential hot deck* imputation. In random hot deck imputation, imputation classes are formed based on categorical auxiliary variables (background characteristics). From the remaining group of potential donors with the same characteristics (x-variables) as the item non-respondent, one is chosen randomly as donor for the imputation concerned. In sequential hot deck imputation, groups are not actively formed, but for each item non-respondent, the score on the target variable is imputed from the next record in the data file with the same scores on certain background characteristics.

A special case of the second class is *predictive mean matching*, in which the nearest neighbour donor is determined using the predicted y-value for a chosen regression model.

Besides hot deck imputation, there is also *cold deck imputation*. Here, the value to be imputed is taken from *another* file, for example, a value of the same object on the same variable at a previous point in time. In this sense, cold deck is not true donor imputation. We will not consider cold deck imputation as a validated method. The method is used infrequently nowadays. If the imputation from another file is a correct value, we can view this as a deductive or logical imputation (Chapter 2). If it concerns a value from an earlier period, then just copying this value as is can seldom be properly justified. A trend factor is usually added to the value, which means ratio imputation comes into play (Chapter 4).

6.2 Applicability

Random and sequential hot deck imputation are used if the auxiliary variables are categorical. If most of the variables are qualitative in nature, then the other, quantitative variables will be divided into classes in advance. For very large files on which hot deck imputation is applied, the sequential hot deck method is sometimes

used based on practical considerations. The processing time would otherwise increase substantially, while the quality of the imputation (see section 5.6) would not change appreciably. To obtain a random donor, the records will first have to be placed in a random order in the file, but using a random selection mechanism is no longer needed.

Nearest neighbour imputation is used especially in the imputation with the help of quantitative *x*-variables, if information would be lost if these variables were temporarily divided into classes. However, it is also possible to include qualitative auxiliary variables, as long as the distance function deals with this in a prudent manner. Because, in nearest neighbour, a distance function between the potential donor and recipient is minimised, it is essential that the importance of every *x*-variable is quantified in the form of a weighting factor; see section 6.3 for more information on this.

Donor imputation is also used if, per record, multiple values are missing on related variables. By designating a single donor for this, inconsistency between the imputations is prevented. This can be seen as a specific solution for the problem of multivariate imputation (Chapter 7).

6.3 Detailed description

6.3.1 Random and sequential hot deck imputation

The intention in hot deck imputation is to find an object in the same file with similar background characteristics, for example, an individual of the same gender, in the same age class, residing in the same province and working in the same sector. The idea is, once again, that if a number of background characteristics of two individuals correspond, the values of the variable to be imputed will better correspond with each other. In random and sequential hot deck, the donors must have the exact same values on the background characteristics, in other words, they must be in the same imputation class. In nearest neighbour (section 6.3.2), no imputation classes are formed, and some discrepancy in the scores on the *x*-variables between donor and recipient is allowed.

So in random and sequential hot deck, the scores on the background characteristics must be identical. If, in the above example, no respondent can be found with the same four characteristics as the item non-respondent, then the imputation class is evidently too limited. For the imputation for this item respondent, we will therefore have to eliminate at least one of the four characteristics, or combine classes. If, however, there is more than one potential donor in the relevant imputation class, then one should be selected randomly. Instead of random selection, a characteristic can be added, in the hope of retaining a single donor. The situation should be prevented, however, where a single object becomes the donor of many recipients. This type of multiple donorship increases the standard errors of means and totals of *y*, due to the risk of outliers being ‘magnified’. This can be prevented, for example,

by only allowing multiple donors in an imputation class after the majority of the objects have had a turn.

In section 6.1 we already explained that, in sequential hot deck, for each item non-respondent, the score on *y* is imputed from the next respondent record in the data file with the same background characteristics. Of course, it is also possible to use the previous record with those background characteristics. If a number of item non-respondents from the same imputation class occur close to one another in the file, there is a risk that they will all be given the same donor. To prevent this, you can adapt the sequential hot deck method by not repeatedly selecting a single record, but instead the first *m* records, and then choosing one of these randomly. Sequential hot deck can be applied after a random sorting of the records, in which case the method is called the 'random sequential hot deck method'. Sequential hot deck can also be performed without advance sorting or only after sorting based on the selected background characteristics. The composition of the file may then lead to bias. In all cases, the imputations depend on the order of the records.

The selection of the auxiliary variables is a difficult process. Both content-based and statistical arguments play a role in this process. Refer to sections 1.1.2.6 and 9.3 for more information about this.

Up to now, we have not taken account of any possible sample weights. Random hot deck and random sequential hot deck, however, are also often performed with weights; see Kalton (1983) and section 1.1.2.10.

6.3.2 Nearest neighbour imputation

In nearest neighbour (hot deck) imputation, a distance *d(i,j)* is defined between two objects *i* and *j*, where *i* is the item non-respondent and *j* an arbitrary item respondent. The distance function *d* can be defined in many ways. A frequently used function is the Minkowski distance $d(i,j) = (\sum_k | x_{ki} - x_{kj} |^z)^{1/z}$, in which the x-variables are quantitative. The respondent *j* with the smallest value of *d(i,j)* is the nearest neighbour of item non-respondent *i* and becomes its donor. For $z = 2$, the Minkowski distance changes to the Euclidian distance, and for $z = 1$ in the so-called city block distance. The larger *z* is, the higher 'penalties' are imposed on large distances between x_{ki} and x_{kj}.

A better, more general distance function is the weighted distance function

$$d_v(i,j) = (\sum_k v_k \, | x_{ki} - x_{kj} |^z)^{1/z} \quad . \qquad (6.3.2.1)$$

The extra factor v_k represents the weight (importance) of variable x_k. Because only the relative weight is relevant, without loss of generality, we can assume that $\sum_k v_k = 1$. It is essential that the weight of each *x*-variable is determined in advance. In fact, this weight cannot be viewed separately from the value range or the

dispersion of the *x*-variables. In practice, the weights are often easier to determine if the *x*-variables have first been normalised to a variance of 1.

It is also possible, when defining *d(i,j),* to take account of covariances between the variables, but this generally makes the determination of the weights more difficult. Another possible distance function is $\max_k v_k \mid x_{ki} - x_{kj} \mid$ or, somewhat more general, $\max_k v_k d(x_{ki}, x_{kj})$. This involves looking for a donor that does not vary strongly from the recipient on any *x*-variable. Incidentally, this distance function results from formula (6.3.2.1) with z infinite.

A special case of nearest neighbour is the *predictive mean matching* method described in Little (1988). In this imputation method, a linear regression is first performed of the imputation variable *y* on different quantitative explanatory *x*-variables, based on the records without item non-response on the variables used in the regression. Next, the resulting regression equation is used to predict values for imputation variable *y* for all the records, in accordance with formula (5.3.2). Item non-respondent *i* is then given the item respondent *j* as donor for which the predicted value $\hat{y}_j$ is as close as possible to the predicted value $\hat{y}_i$ of the item non-respondent. Finally, the *observed* value y_j of donor *j* is imputed, in other words, $\tilde{y}_i = y_d \equiv y_j$ in accordance with formula (6.1.1). The fact that predictive mean matching is a special case of nearest neighbour imputation follows from the distance function:

$$d(i, j) = \mid \hat{y}(x_i) - \hat{y}(x_j) \mid . \qquad (6.3.2.2)$$

In nearest neighbour, including predictive mean matching, you can also select the closest *m* records and then randomly select one of them, exactly as described for sequential hot deck; it is also possible to give donors with a smaller score on the distance function a greater chance of being selected. Including sample weights, as in the weighted random hot deck method, does not have an influence on the nearest neighbour, if this is limited to a single neighbour. In predictive mean matching, the weighting in the regression analysis will also not have much of an influence.

The random and nearest neighbour hot deck methods can be combined by first forming classes based on one or more background characteristics, and then applying the nearest neighbour method in these ‘blocks’. This is one of the ways to use nearest neighbour with both qualitative and quantitative variables. In this case, the qualitative variables have a greater weight (infinitely greater) than quantitative variables. More generally, we can add a distance function for qualitative variables to distance function (6.3.2.1), and use a weighted sum of both as a combined distance function. In this context, the qualitative variables can also be assigned weights among themselves.

In section 1.2, we made a distinction between ‘imputation’ and the broader concept of ‘correction’. In imputation, a missing value is replaced by a valid value; the correction of an incorrect value by a valid value is only considered as imputation if

the original incorrect value does not play a role in the correction. Nearest neighbour can easily be extended to correction, in which the original value does have an influence. The distance function to be selected is then expanded using a restriction that the new value may differ very little from the original incorrect value. See the theme report 'Data editing' in the Methods Series (Hoogland et al., 2011) and Scholtus (2008).

6.4 Example

Example. *Housing Demand Survey (Woningbehoeftenonderzoek - WBO)*
In the past at Statistics Netherlands, donor imputation was used frequently in the WBO. This involved, for example, imputing income variables and variables concerning a dwelling, such as the market value of the dwelling. Many missing values occur in these variables. Various personal characteristics, and also the number of rooms in the dwelling and whether it had a garden, could be used as background characteristics. Due to the qualitative character of most of the *x*-variables, use was mainly made of donor imputation (random hot deck and predictive mean matching), but regression imputation could also have been used. The programme SURFOX from ABF Research in Delft was used in this context.

6.5 Characteristics

Sequential hot deck and cold deck are deterministic imputation methods (section 1.1.2.7). But after random sorting of the file, the sequential hot deck method becomes a stochastic method. As the name indicates, the random hot deck method is also a stochastic method. And also if a disturbance term is added (in most cases, $\varepsilon_i \sim N(0,\sigma^2)$ is selected), deterministic imputation methods become stochastic methods.

6.6 Quality indicators

See section 5.6.

7. Multivariate imputation

7.1 Short description

Until now, there was always just one target variable which had missing values. Often, in a single record, there are missing values on multiple variables, and there is a connection between these variables. In this case, the imputation of all the missing variables is a multivariate problem. This chapter discusses various ways to deal with multivariate imputation.

Donor imputation (Chapter 6) is easy to use in the case of multiple missing variables. A single donor record then provides all the missing values for the recipient. In such a case, you must create imputation classes that are homogeneous for multiple target variables or, in the case of nearest neighbour imputation, ensure that there are auxiliary variables in the distance function that are associated with multiple target variables. Taking all missing target variables from the same donor record also ensures that the imputed values are consistent among themselves. Consistency between the imputed values and the original values of the recipient is, in general, not guaranteed. However, it is possible to obtain consistency between imputed and original values by taking account of this in the selection of the donor. This form of donor imputation is described in Chapter 6 of the Methods Series theme 'Data editing' (Hoogland et al., 2011). Applications of this method to data from the Municipal Personal Records Database (GBA) are described in Pannekoek et al. (2008) and Scholtus (2008).

If there are multiple variables with missing values, in regression imputation (and, as a special case, ratio imputation), the predictor(s) will often contain missing values. Statistics Netherlands has two solutions that it uses frequently for this problem. One solution is based on an order of the target variables determined in advance. The first target variable is imputed using a model that contains only predictors without missing values. For the next target variable, predictors can be selected from the variables without missing values *and* the imputed variable in the previous step, and so forth. The second solution does not use imputed values in the predictors, but a number of optional models with different predictors are specified for each target variable. The selection of the model to be used for a certain target variable in a certain record is determined by going through the models in an order determined in advance. If the predictors from the first model do not contain any missing values, then that model is used, otherwise the second model is used if the predictors of this do not contain any missing values, and so forth. These methods are explained further in subsection 7.3.1.

In business statistics, there is often a situation where restrictions apply to different target variables. For example, the total turnover and the turnovers of a number of sub-items may be known, but other sub-items are not filled in. A simultaneous form of ratio imputation can impute the missing sub-items in such a way that a consistent

record is created in which the imputed and other sub-items add up to the total. In general, separate ratio imputations lead to an inconsistent record. This method is not yet used at Statistics Netherlands, but it is being discussed here because it is a simple and useful expansion of ratio imputation.

In this chapter, it is assumed that the auxiliary variables for imputing a target variable can also contain missing values themselves, and therefore can also be target variables. Because the distinction between auxiliary variables (x-variables) and target variables (y-variables) therefore no longer applies, *y* is used for all variables in this chapter.

7.2 Applicability

With respect to the use of donor and regression imputation techniques for multivariate problems, the same applies for the measurement level of the variables as what is stated for the univariate use of these techniques in Chapters 5 and 6.

7.3 Detailed description

7.3.1 Sequential imputation; order of variables and order of models

In section 5.3, regression imputation is discussed for a single target variable. Now we assume that multiple target variables must be imputed using regression imputation. The simplest method to solve the problem is the repeated application of the method for a single target variable. This is an unambiguous method if the auxiliary variables for each target variable do not contain any missing values, but if the auxiliary variables themselves also contain missing values, various choices must be made to come to a feasible solution.

One option is to impute the variables in a certain order, so that the predictors for each target variable are imputed first. In this case, values for predictors are always available. This method is used, for example, in Structural Business Statistics.

Another option is to specify models with different predictors for each target variable. In the imputation, a model can be selected for which the predictors have been observed in the record concerned. In this case, no imputed values are used in the predictors. This method is used, for example, in the imputation for the statistic Building Objects in Preparation (*Bouwobjecten In Voorbereiding* - BIV), (see Van der Loo and Pannekoek, 2007).

The method in which the predictors are imputed first is described below using a simplified description of the imputation procedure used for Structural Business Statistics. Ratio imputation is used in Structural Business Statistics, the same as for many other economic statistics. Table 3 indicates for a number of target variables which auxiliary variable is used to impute missing values using ratio imputation.

Table 3. Imputation diagram for variables from Structural Business Statistics

Variable	*Auxiliary variable*
y_1: Turnover	-
y_2: Total operating expenses	Turnover
y_3: Total staff costs	Total operating expenses
y_4: Accommodation costs	Total operating expenses
y_5: Energy costs	Total operating expenses
y_6: Other costs	Total operating expenses
y_7: Permanent staff costs	Total staff costs
y_8: Other staff costs	Total staff costs

The variable *Turnover* is not imputed. Records in which this central variable is missing are considered as non-response. *Turnover* has therefore always been observed for the records to be imputed. The other variables are imputed using the ratio method as described in Chapter 4. The imputed value $\tilde{y}_{ij}$ for a target variable y_j in a record *i* can then be represented as:

$$\tilde{y}_{ij} = y_{ik}^{*}\hat{R}_{jk},$$

where y_{ik}^{*} is the value of the auxiliary variable y_k for the target variable y_j if this is observed, and the imputed value $\tilde{y}_{ik}$ otherwise, and $\hat{R}_{jk}$ is the estimate for the proportional constant R_{jk} pertaining to the variables y_j and y_k. This imputation method is used within classes formed by combinations of size class and industry sector (*group ratio imputation*, see section 4.3).

The order in which the target variables are imputed is as follows: first, y_2 is imputed using y_1; next, y_3-y_6 using y_2; and finally, y_7 and y_8 using y_3. Each variable that is used as an auxiliary variable is first imputed before it is used as an auxiliary variable. In this way, there is always a value available for the auxiliary variable: either an observed value or an imputed value.

Ratio imputation using imputed values for the auxiliary variable is comparable with a method in which different models are used for imputation but imputed values are not used for the auxiliary variable. This relationship is described below. If the auxiliary variable is imputed, the following applies for the imputation of the target variable:

$$\tilde{y}_{ij} = y_{ik}^{*}\hat{R}_{jk} = \tilde{y}_{ik}\hat{R}_{jk} = y_{il}\hat{R}_{kl}\hat{R}_{jk},$$

where y_l is the auxiliary variable for y_k. Here, it is assumed that y_{il} has been observed. This demonstrates that, for the records for which y_k is imputed, the imputations do not vary with y_k, but they do with y_l. The product $\hat{R}_{kl}\hat{R}_{jk}$ can be understood as an estimator for the ratio R_{jl}. If the estimates for the ratios R_{jl}, R_{kl} and R_{jk} are based on the same records, then $\hat{R}_{jl} = \hat{R}_{kl}\hat{R}_{jk}$ applies exactly, and the imputed value is equal to a ratio imputation where y_l is the auxiliary variable. The

method described above is comparable with: impute y_j using the auxiliary variable y_k if this is observed, and otherwise using the auxiliary variable y_l. This is an example of the specification of different models for a single target variable.

Specifying different models for each target variable and then selecting a model for which the predictors are observed is applicable to regression imputation in general. The drawback of this method is that more models must be specified than in the imputation of the predictors. An advantage, however, is that there are more options to specify the best possible predictive models. If, for example, in the Structural Business Statistics for a certain branch, the variable *total staff costs* is strongly associated with the variable *total operating expenses*, a decision can be made to impute missing values in *total staff costs* using *total operating expenses* as the auxiliary variable and to impute missing values in *total operating expenses* using *total staff costs* as the auxiliary variable. If both variables are missing, it is still possible to fall back on *turnover* as the auxiliary variable for each of these variables.

7.3.2 Ratio imputation of sub-items

In the example in the previous subsection, ratio imputation was applied for sub-variables where the total concerned was the auxiliary variable. This situation occurs frequently in economic statistics.

In general, this relates to variables y_j; $j = 0, \cdots, J$, for which the restriction (or edit rules) applies: $y_0 = \sum_{j=1}^{J} y_j$.

If one of the sub-variables y_j is missing, then this one missing value can easily be imputed using a deductive method (see Chapter 2). Furthermore, if the sum of the observed variables is equal to the 'total variable', deductive imputation is possible, namely with the value of zero for each of the missing variables. If, however, the sum of the observed sub-variables is smaller than the value of the total variable and there are multiple sub-variables with missing values, then there is still part of the total remaining that must be divided among the missing values.

A method to determine this distribution is by using the ratios of the sub-variables to the total, rescaling these in such a way that the sum of the imputed values is equal to the difference between the total and the sum of the observed sub-variables. If we index the observed sub-variables in record *i* with $j = 1, \cdots, J_{i,obs}$ and the missing sub-variables with $j = J_{i,obs} + 1, \cdots, J$, then the sum of the observed sub-variables in record *i* is

$$S_{i,obs} = \sum_{j=1}^{J_{i,obs}} y_{ij}$$

and the sum of the missing sub-variables in that record is

$$S_{i,mis} = y_{i0} - S_{i,obs}.$$

The imputations for the sub-variables using the rescaled ratios to the total are then indicated by

$$\tilde{y}_{ij} = S_{i,mis} \frac{\hat{R}_j}{\sum_{j=J_{i,obs}+1}^{J} \hat{R}_j}.$$

Because the rescaled ratios add up to 1, the sum of the imputed values is equal to $S_{i,mis}$, and the imputed record satisfies the edit rule.

This form of ratio imputation, in which use is made of the extra information that the sum of the missing values is known, will lead to better results than the usual ratio imputation that does not use the known total of the missing values. A hot deck variant of this method is discussed in Pannekoek and De Waal (2005). In this variant, the ratios are not estimated using estimations of the totals of auxiliary and target variables (as described in section 4.3), but they are estimated using the corresponding ratios as these are observed in a donor record (the ratio hot deck method).

7.3.3 Simultaneous regression imputation

A general multivariate regression method that is described in much of the literature about imputation methods is a method based on the assumption that the simultaneous distribution of the target and auxiliary variables concerned is multivariate normal. Using this method, it is possible to generate stochastic imputations in which not only the variances of imputed variables, but also the correlations between all the variables, are retained as accurately as possible.

The basic principle in this method is that each missing variable is imputed using a regression model with all observed variables as predictors. If, for example, the first three variables in a record have missing values, we perform imputation using the three regression models (analogous to formula 5.3.1)

$$\begin{aligned} y_{i1} &= a_1 + b_1 y_{i,obs} + e_{i1} \\ y_{i2} &= a_2 + b_2 y_{i,obs} + e_{i2} \\ y_{i3} &= a_3 + b_3 y_{i,obs} + e_{i3} \end{aligned},$$

where $y_{i,obs}$ is the vector with the values of the variables observed in record *i*.

More generally, the regression equations for the missing values in a record *i* can be summarised in the form

$$y_{i,mis} = a_{i,mis} + b_{m.o.(i)} y_{i,obs} + e_{i,mis} \qquad (7.3.1)$$

where $y_{i,mis}$ is the vector with missing values in record *i* and $a_{i,mis}$ is the vector with the constants for the regressions, $b_{m.o.(i)}$ is the $q_i \times p_i$-matrix with the regression coefficients for the regression of the q_i variables that are missing for

record *i* on the p_i (predictor) variables that are observed for record *i* and $\varepsilon_{i,mis}$ is the vector with disturbances for the q_i regressions. The matrix of regression coefficients is dependent on *i*, but only because the variables that are missing can differ per record. For records in which the same variables are missing, the matrix $\beta_{m.o.(i)}$ is the same. The disturbances will in general be correlated, so that we assume that the disturbances are normally distributed with the expectation 0 and a non-diagonal covariance matrix: $\varepsilon_{i,mis} \sim N(0, \Sigma_{\varepsilon_{i,mis}})$.

If there are no missing values, the parameters of a multivariate regression model such as (7.3.1) can be obtained using the least squares method, analogously to the estimation procedure for univariate regression models. If there are missing values, parameters could be estimated based on the records in which all variables are observed. The number of complete records, however, may be limited, especially if there are a lot of variables. An alternative in such cases is to calculate the estimations using the so-called EM algorithm. This is an iterative procedure in which the parameters can be estimated based on incomplete data; all the data (also from the records with non-response) is used for this (see Little and Rubin, 1987).

Using the estimates $a_{i,mis}$ and $b_{m.o(i)}$ for the parameters $\alpha_{i,mis}$ and $\beta_{m.o.(i)}$, the missing values in record *i* can be imputed according to

$$\tilde{y}_{i,mis} = a_{i,mis} + b_{m.o.(i)} y_{i,obs} . \qquad (7.3.2)$$

This is an imputation without disturbances, aimed only at reproducing the means but not the variances or covariances. If we also want to retain the variances and covariances of the variables after imputation as accurately as possible, we can use a vector with disturbances $e_{i,mis}$ that is selected from the multivariate normal distribution with the expectation 0 and covariance matrix $\Sigma_{\varepsilon_{i,mis}}$. The EM algorithm also produces an estimation for this covariance matrix.

8. Methods for longitudinal imputation

8.1 Brief description

We refer to longitudinal data when the same variables are measured multiple times for the same objects. Panels, in which objects selected by a sample are followed for a longer period of time, are a special case of this. However, the methods for longitudinal imputation described in this chapter also apply to other types of longitudinal data, such as registers that become available with some degree of regularity. Examples at Statistics Netherlands of rotating panels are the STS (Short Term Statistics for company turnovers) and the Labour Force Survey (LFS). The Municipal Personal Records Database (GBA) is a longitudinal register that is updated annually, while data is also obtained about people moving house in the interim and about changes in, for example, people's marital status. Most registers produce longitudinal information when data from different dates are matched – for example, files with jobs, benefit payments and incomes. In particular, longitudinal files can be compiled from the Social Statistics Database (SSB). However, these files will often have to be imputed longitudinally. These files allow us to follow, for example, the course of an individual's life. Within the framework of EU-SILC, the Netherlands is required to submit panel data to Eurostat. This is done based on a large number of files, including the panel component of the LFS.

Longitudinal imputation is distinct from other methods described in this report because, during the imputation, use is made of data from the same object at different times, often without using data from other objects. So for each object, there is a time series with one or more missing values, for which imputation must be performed.

Missing values in longitudinal data come in two forms:

1. Scattered missing values, because objects are not observed during one or more periods, or because not all of the variables are observed for the objects.
2. Panel dropout; at a certain point, objects no longer wish to participate and, consequently, there are no more observations of the object from a certain point in time.

It should be noted that death and migration do not produce missing values. These people or companies are no longer part of the target population and therefore must not be imputed. Lepkowski (1989) offers a more detailed explanation of various forms of missing values in longitudinal data.

8.2 Applicability

Longitudinal imputation can be used if there are missing values in longitudinal data. Let y_{it} be a missing value of object *i* at period *t* on variable *y*. Then *y*-values of object *i* at previous and subsequent periods can be used to create an imputed value $\tilde{y}_{it}$. Often, the information about *y* is limited to earlier periods, in other words,

$y_{it-1}, y_{it-2}, \ldots$ This information can be used for dealing with both dropout and scattered missing values. Information about later periods is only useable if there is time to wait for the results concerned or if imputation is performed for a number of periods at the same time, with the goal of obtaining the best and most complete longitudinal data file possible.

There are two main reasons to use longitudinal imputation techniques instead of the cross-sectional methods discussed in previous chapters.

1. First, earlier or later observations of the same object are very good predictors for the missing value. This means that the quality of the imputation can be strongly improved. To achieve this, each of the methods discussed above can in fact be used, in which the previous and future observations are used as auxiliary variable.
2. Second, we generally look at longitudinal data not only cross-sectionally (such as the number of people cohabitating at a certain point in time), but we are also interested in changes over time (such as the number of people who have started cohabitating). To correctly estimate these changes, it is important that the imputation takes into account previous and future values.

Panel dropout in samples can usually also be resolved by weighting. If we want to estimate a population parameter at a certain time, then we can consider the recent dropout as unit non-response and add it to the non-response of previous points in time. Panel dropout in registers is usually justifiable: it occurs due to death and emigration. For literature about panel dropout, see, for example, Fitzmaurice et al. (2004, Chapter 14).

Many methods for longitudinal data can deal with missing data. See, for example, Van der Laan and Kuijvenhoven (2008) for several of these methods and a literature list for longitudinal analysis methods. Furthermore, it is not always necessary to impute missing data. Depending on the objective of the analyses, the preference is sometimes to not perform imputation.

8.3 Detailed description

In view of the fact that longitudinal imputation methods do not constitute one single method, each of the methods is discussed separately in the following sections. For this reason, we discuss only a few characteristics of longitudinal imputation methods here.

The different methods are characterised by a number of features.

- Use of information from other objects. Several methods use only previous and future observations of an object in the imputation. The advantage of this is that the imputation method if often relatively simple and also easy to apply to large datasets. A drawback of this, however, is that the additional information from other objects is not included, which means that information loss can occur. For example, income can be taken from the previous period, with a correction for

the average income increase. The use of this information, if available, will generally lead to a better imputation.

- Suitability for continuous and/or categorical data. All the methods discussed are suitable for continuous data. However, not all of the methods are appropriate for categorical data.
- Multivariate/univariate. In longitudinal data, it will regularly occur that, for a single object, multiple observations of *y* are missing. Some methods impute multiple missing values all at the same time and, as a result, will often be better able to retain the correlation between the observations at the different periods. Other methods can only impute one missing value at a time. In the case of multiple missing values, these methods must be applied several times. These methods does not guarantee in advance that the correlation between the observations will be retained at different periods.

Table 4 shows the abovementioned characteristics for each of the methods. These methods are explained in more detail in the sections below.

Table 4. Characteristics of the imputation methods

Method	Continuous	Categorical	Use of information from other objects	Multivariate
Interpolation	+	-	-	-
Last observation carried forward	+	+	-	-
Ratio imputation	+	-	+	-
Regression imputation	+	+	+	+
Cold deck	+	+	-	+/-
Hot deck	+	+	+	+
Little and Su	+	-	+	+

8.4 Interpolation

8.4.1 Brief description

In interpolation, missing observations are estimated from previous and future observations. In this context, no use is made of information from other individuals or from auxiliary variables. For individual *i*, $\tilde{y}_{it}$ is determined by

$$\tilde{y}_{it} = f(y_{it-1}, y_{it-2}, \ldots, y_{it-K}, y_{it+1}, y_{it+2}, \ldots, y_{it+L}) . \qquad (8.4.1)$$

Here, K observations from the past and L observations from the future are used.

8.4.2 Applicability

Interpolation can be used for quantitative variables in a situation where it is difficult to make model assumptions and where the other objects do not provide any information about the value to be imputed. If the other objects do contain information about the object to be imputed, then using methods that utilise this

information (such as regression imputation, ratio imputation and the Little and Su method) is recommended.

8.4.3 Detailed description

For quantitative *y*-variables, the following rather general interpolation formula exists for $\tilde{y}_t$ based on the observations $y_{t-K,}\ldots, y_{t-1}, y_{t+1}, \ldots, y_{t+L}$:

$$\tilde{y}_t = \frac{\sum_{k=1}^{K} w_{-k} y_{t-k} + \sum_{\ell=1}^{L} w_{\ell} y_{t+\ell}}{\sum_{k=1}^{K} w_{-k} + \sum_{\ell=1}^{L} w_{\ell}}, \qquad (8.4.2)$$

with weights $w_{-1} \geq w_{-2} \geq \ldots \geq w_{-K}$ and $w_1 \geq w_2 \geq \ldots \geq w_L$; this means that y_T has a smaller weight in both directions from period *t*, as period *T* is further away from period *t*. The weights can be freely selected. For example, it is possible to choose $w_k = w_{-k} = 1/k$.

Formula (8.4.2) can also be used if multiple scores of object *i* are missing. If, for example, y_{t+k} is not known and we want to determine $\tilde{y}_t$, we define $w_k = 0$. The formula can also be used if only information from the past is known ($w_1 = w_2 = \ldots = w_L = 0$), which is the case for panel dropout.

Special cases of formula (8.4.2) are:

1. *Linear interpolation between the nearest preceding and subsequent observation.*

   If y_{t-1} and y_{t+1} are both available, then formula (8.4.2) changes to the arithmetic mean of the two:

   $$\tilde{y}_t = \frac{w_1(y_{t-1} + y_{t+1})}{2w_1} = \frac{y_{t-1} + y_{t+1}}{2}, \qquad (8.4.3)$$

   if $w_{-1} = w_1$. If y_{t-1} or y_{t+1} is missing, then the observations y_{t-k} and $y_{t+\ell}$ closest to period *t* should be used, with the weights $w_{-k} = 1/k$ and $w_{\ell} = 1/\ell$ respectively. Formula (8.4.2) then changes to

   $$\tilde{y}_t = \frac{\ell y_{t-k} + k y_{t+l}}{k + \ell}. \qquad (8.4.4)$$

   Suppose, for example, that y_t and y_{t+1} are unknown and that $y_{t-1} = 3$ and $y_{t+2} = 4$. In this case, it follows from formula (8.4.4) that $\tilde{y}_t = \left((2 \times 3) + (1 \times 4)\right)/\left(1+2\right) = 10/3 = 3.333$. The interpolation can also be used on y_{t+1}, with the help of formula (8.4.4):

$\tilde{y}_{t+1} = \big((1\times 3)+(2\times 4)\big)/\big(2+1\big) = 11/3 = 3.667$. For $\tilde{y}_{t+1}$, we obtain the same value if we assume the previously imputed value $\tilde{y}_t$ is known: $\tilde{y}_{t+1} = \big((1\times \tilde{y}_t)+(1\times y_{t+2})\big)/\big(1+1\big) = \big(10/3+4\big)/2 = 3.667.$

2. *Mean of the nearest preceding and subsequent p observations.*

   We can determine an unweighted mean of the nearest preceding and subsequent *p* observations, or give the observations unequal weights as in formula (8.4.2). Linear interpolation is then a special case where *p*=1, $w_{-k} = 1/k$ and $w_{\ell} = 1/\ell$.

3. *Linear trend (regression of y on T)*

   The regression equation $y = \alpha + \beta T + \varepsilon$ can be estimated using the y-observations that we want to include. For period *T*=*t*, when observations were not made, we thus obtain the regression imputation

   $$\tilde{y}_t = \hat{y}_t = a + bt \tag{8.4.5}$$

   where *a* and *b* are the least squares estimators, in accordance with formula (5.3.3). From the regression analysis theory, it is known that $\hat{y}_t$ is a linear combination of the observations $y_{t'}$. The linear trend changes to linear interpolation if the auxiliary information is only based on the nearest preceding and subsequent period. Here, we have not weighted the observations.

   Of course, the parameters from formula (8.4.5) can be estimated using other loss functions, or a form of non-linear regression can be used. In this case, the imputation is not necessarily in the form of (8.4.2) anymore.

Of the above three methods, the simple linear interpolation (between the nearest preceding and subsequent observation) is generally preferred. This is certainly true if the data follows a memory-free process. The observations at the nearest preceding and subsequent period then contain all the information; the information at the other periods is irrelevant. If, however, large measurement errors occur in the data, the scores at other periods will also be important.

SPSS includes the module RMV for estimating missing values in a time series. This module contains the following methods, with how the method follows from our formulas in parentheses:

- Linear interpolation (formula (8.4.4));
- Mean of p nearest preceding and p subsequent values (method 2);
- Median of p nearest preceding and p subsequent values (variant of method 2);
- Series mean; in other words, the mean of all values from a time series (specification of formula (8.4.2));

- Linear trend (method 3).

8.4.4 Characteristics

- Interpolation is easy to use on large datasets, because interpolation only utilises information from a single object. Objects can therefore be processed one by one.
- Because no information from other objects is used, this method may produce less accurate estimations than methods that do use information from other objects.
- Because no disturbance term is used in the imputation, the series can be 'too perfect'. The significance of correlations between the different periods can be overestimated. This can be prevented by adding a disturbance term; see section 1.1.2.6.

8.5 Last observation carried forward/backward

8.5.1 Brief description

Last observation carried forward (LOCF) is a method that is often used in practice outside of Statistics Netherlands. The method is not without problems, but it is frequently applied because it is very easy to use. In this method, the last observed value of an individual is used for the values of all later periods that must be imputed. Variations of this method are discussed in the detailed description.

8.5.2 Applicability

This method is mainly applicable to categorical variables for which it is known that they change very little or not at all over time. An example of such a variable is gender. For other categorical and quantitative variables, this method often mistakenly produces an overly stable picture of the actual situation. For example, for index figures, this method can lead to the observation of a non-existent price stability.

8.5.3 Detailed description

In LOCF, the last observed value y_{it-1} is used to impute the missing value y_{it}. Another variant is last observation carried backward (LOCB), in which the next observed value y_{it+1} is replaced for the value y_{it} to be imputed. As for LOCF, this value can be used for multiple successive missing values.

In random carry-over (Williams and Bailey, 1996), a missing intermediate value y_{it} is imputed by using y_{it-1} or y_{it+1}. This means, incidentally, that the method cannot be used if values are missing for two or more consecutive periods. Moreover, this method cannot be applied if the first observation and/or last observation is missing. In these cases, other imputation methods should be used.

8.5.4 Characteristics

The problem with LOCF is that it is often not realistic to assume that the last value will no longer change over time. This assumption must be investigated. Normally, the data for an individual has some variation over time due to random fluctuations (or measurement errors). LOCF does not acknowledge this variance. In this way, however, the imputation uncertainty is also not adequately taken into account, which leads to incorrect statistical conclusions. A simple solution is to add a disturbance term (see section 1.1.2.6). The same applies for the LOCB method, for which it must also be investigated whether or not an overly stable time series deviates from the actual situation.

8.6 Ratio imputation

This method was already discussed in Chapter 4. As also indicated there, this method is frequently used for longitudinal data for which it is often reasonable to assume that the observation at period *t* is proportional to the observation at period *t-1*. This method can be considered as a refinement of last observation carried forward, in which corrections are also made for general changes over time. It should be noted that a different disturbance term for each time period can be selected every time. For a further discussion of this method, please refer to Chapter 4. This form of imputation is frequently used in economic statistics.

8.7 Regression imputation

8.7.1 Brief description

Regression imputation was already discussed in Chapter 5. What was explained in that chapter is generally also true for the longitudinal situation. In this section, we will therefore only address issues that have to do with the longitudinal character of the data. Because longitudinal data is in fact multivariate, the analysis thereof is often more complex. However, an advantage of longitudinal data is that, in general, the past and/or future observed values of a variable are very good predictors of missing values.

In Chapter 5, the situation is discussed where we want to predict the value of a single variable y using a number of variables x_j. In this context, we are primarily interested in the variable y. In the case of longitudinal data, we have multiple observations y_{it} for each individual i, where t runs from 1 to M. Multiple y_{it} can be missing for a single individual. In analyses of longitudinal data, we are generally interested in the correlation between the observations at the different periods – for example, we want to study change. It is therefore important to retain the correlation between the observations in the imputation. This means the imputation is multivariate; multivariate imputation is discussed in Chapter 7.

An option that is not multivariate is to set up a separate univariate model for each missing y_{it}, where y_{it} depends on both a set of covariates x_{ij} and the previous and future observations of y_{it}:

$$\mathrm{E}[y_{it}] = f(x_{i1},\ldots,x_{ip},y_{it-1},y_{it-2},\ldots,y_{it+1},y_{it+2},\ldots)\,. \qquad (8.7.1)$$

A model must therefore be created for each missing observation, and separate models must be set up in the case of multiple missing observations and missing covariates. This can be extremely complex, and it is very difficult to retain the correlations between the observations.

Another option is to use a multivariate model (see, for example, Verbeke and Molenberghs (2000)). Here, a single model is set up that describes all the observations. The different observations of individual *i* are written as vector $\mathbf{y_i}$ and a model is created that describes this vector. For example, a linear model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i\,, \qquad (8.7.2)$$

where the vector $\boldsymbol{\varepsilon}_i$ follows a multivariate normal distribution. In the case of longitudinal data, it is important to model the correlation that exists between the different observations (such as the fact that someone with a high income for a certain observation will probably also have a high income for the next observation).

The multivariate modelling of longitudinal data falls outside the scope of the theme Imputation and will therefore not be discussed further here. For more information, see, for example, Van der Laan and Kuijvenhoven (2008), Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005) for discrete longitudinal data. These articles also provide an overview of the literature on this subject. Gelman and Hill (2006) and Longford (2005) provide more detailed descriptions of hierarchical or multi-level models.

8.7.2 Applicability

- Regression imputation can be used for both quantitative variables and categorical variables. In the second case, however, no use can be made of multivariate linear regression; logistic regression, for example, must be used instead.
- The multivariate regression models discussed above can often deal with different observation times for the various individuals. Most other methods discussed in this chapter assume that all individuals are observed at fixed observations times (such as each year or each quarter).

8.7.3 Characteristics

In the analysis of longitudinal data, we are generally interested in changes over time. As discussed in section 1.1.2.6, a decision can be made as to whether or not to use a disturbance term in the imputation. If, in the case of longitudinal data, the disturbance term is not used, the significance of the changes will be strongly overestimated.

8.8 Cold deck

Cold deck imputation was already discussed in Chapter 6, where the method was considered as a non-validated method. We will not discuss this method further. It should be noted that we do not consider the last observation carried forward/backward method to be a cold deck method. In cold deck imputation, use is made of information from an external source. In last observation carried forward/backward, however, use is made of a file from an earlier or later time period respectively. This file is not viewed as an external source.

8.9 Hot deck

Hot deck imputation or donor imputation was already discussed in Chapter 6. There it was already stated that donor imputation is used if multiple values are missing per record. This makes donor imputation especially suited for use with longitudinal data. In the hot deck method, multiple values for a single individual can be imputed. As a rule, one donor is designated for this purpose to ensure consistency among the imputations. In longitudinal data, the correlation between consecutive values over time is better retained in this way. Chapter 7, which discusses multivariate imputation, addresses this subject in more detail.

8.10 Little and Su method

8.10.1 Brief description

The Little and Su method (Little and Su, 1989) includes both the individual level and the mean trend over time in the imputation. The following model is used in this context:

$$\text{(imputation)} = \text{(row effect)} \times \text{(column effect)} \times \text{(residual)}. \qquad (8.10.1)$$

The column effect describes the mean change over time and is therefore also called the ‘period effect’, while the row effect describes the individual level corrected for the period effect. In the Little and Su method, the residual is taken from another individual which, in terms of the row effect, is most similar to the individual that is imputed. The assumption is that individuals that are similar with respect to the row effect are also similar with respect to residuals.

If an individual is missing multiple related values (such as gross and net salary; in SURFOX, this is called record matching), these values are imputed all at one time, and a single donor is used for the residuals.

8.10.2 Applicability

The Little and Su method can be used for missing values in a quantitative positive variable y, which can be modelled as a period effect multiplied by an individual effect, and for which stochastic imputation is desired. This method is reasonably

easy to use and can deal with different patterns of missing data, including multiple missing values per individual.

The method has problems dealing with individuals for which the observed values are all equal to zero. These individuals cannot be imputed using the Little and Su method.

8.10.3 Detailed description

The column effect c_t gives the mean change of the objects over time and is estimated by

$$c_t = \frac{\bar{y}^t}{\frac{1}{M}\sum_{t=1}^{M}\bar{y}^t}, \tag{8.10.2}$$

where $\bar{y}^t$ is the mean of the observed y_i^t at period *t*, *M* is the number of periods. The row effect r_i for individual *i* is represented by

$$r_i = \frac{1}{m_i}\sum_t \frac{y_i^t}{c_t}, \tag{8.10.3}$$

where the sum is calculated over the m_i available y_i^t for individual *i*.

The residual is taken from another individual *j* for which the periods missing for individual *i* are observed. Individual *j* is selected by first sorting all individuals based on the row effect and then selecting the individual for which the row effect is closest to that of *i*. The residual of individual *j* is represented by

$$e_j^t = \frac{y_j^t}{r_j c_t}. \tag{8.10.4}$$

Substituting this in equation (8.10.1), we obtain

$$\tilde{y}_i^t = r_i c_t e_j^t = r_i c_t \frac{y_j^t}{r_j c_t} = \frac{r_i}{r_j} y_j^t. \tag{8.10.5}$$

In the ideal case, the donor (of the residuals) has as many as the same attributes as the recipient as possible. The standard method as discussed above tries to achieve this by matching the donor and recipient with each other using the row effect. However, it is also possible to expand the method, by applying the standard method in the strata. As a result, the column effects are also allowed to differ between the strata; therefore, the mean progress over time may also differ between the strata. This method is also called ‘extended Little and Su’, and is used, for example, in HILDA (Starick and Watson, 2006).

8.10.4 Characteristics

- Due to the way that the residuals are determined, this method can also impute the value zero, even if the observed values are not equal to zero. The frequency with which zero will be imputed will be of the same order as the fraction of zeros in the complete data. Many other methods, such as regression imputation and ratio imputation, lack this characteristic.
- This method assumes implicitly that the row effect is greater than zero. For an individual for which the observed values are equal to zero, a value not equal to zero can never be imputed. In general, this is not realistic.
- Donors for which the row effect is equal to zero present a problem, because formula (8.10.5) divides by this. In general, these donors will mainly be matched with recipients for which the row effect is also equal to zero, which, as discussed in the previous point, is also problematic. This method can therefore not be used for individuals with a row effect equal to zero.

8.10.5 Quality indicators

- The residuals can be calculated relatively easily using formula (8.10.4). If the model fits well, then the residuals are approximately equal to one. In this context, however, account must be taken of the fact that the residuals are not symmetrically distributed, for the residuals are always greater than zero.
- Validation/simulation. See section 5.6.
- There are no known formulas to determine the variance and the inaccuracy. Multiple imputation (see 5.6 and Rao, 1996) cannot be performed with the methods as described here, because the imputation of multiple different values is required for this purpose. The methods discussed here, however, always impute the same value. Perhaps adapting the donor selection would make multiple imputation possible.

8.11 Conclusion

One point that must be taken into account for longitudinal data is the way in which new information must be dealt with. In a longitudinal data file, the best possible imputation at micro level is obtained if as much information as possible from the past and the future is included. If, therefore, new information comes in, such as a new wave of data for a panel, then this new information can be used to revise or improve the values already imputed. A decision must be taken as to how far back we want to incorporate information:

- It can be decided not to use the new information to improve imputations performed earlier. The earlier imputations are, in this case, not as good as they possibly could be, but this prevents a situation where we have different versions of the same file. A drawback is that the comparability of the data over time becomes an issue. The new information could, for example, be in conflict with

the values already imputed. This makes it difficult to perform longitudinal analysis.

- If new information is indeed used to revise earlier imputations, then we will have to deal with different versions of the data. For example, we will have a file for 2008 with the information that was available in 2008, and a file for 2008 with the information that was available to 2009 inclusive.

9. Conclusion

9.1 Flagging / documentation

It is necessary to document which values are imputed and which methods were used for this purpose, and this includes the auxiliary variables and parameters used in the model. This is needed to make the process reproducible. There are various options for identifying imputed values in the file:

- 'Flagging' the imputed values;
- Working with unimputed and imputed files;
- Making a distinction between variables before and after imputation.

Such documentation is also necessary for researchers who wish to conduct further analyses on the micro file. For them, using the imputations may be undesirable, because this could lead to the wrong conclusions. In addition, when determining standard errors, it is necessary to know which scores are real and which are imputed, and also which imputation method was used.

A working method used in some statistics is to immediately assume that a file is imputed, even if no data has been received yet. The imputed values can then initially be based on the values of period *t-1*. And each time new data is received, this data replaces the imputations, after which the remaining imputations are updated. Such a working method only deviates in terms of the process (it is possible to quickly produce estimations at any time), but not in terms of the method.

9.2 Dealing with outliers

If, among the respondents, outliers occur on the variable *y*, one could consider limiting the influence of this in the imputation. For example, a robust form of regression analysis can be performed; or a potential donor with an extreme value on *y*, given the auxiliary variables, can be given a lower probability of acting as donor. Taking account of outliers in the imputation in this way reduces the confidence margins, but introduces extra bias. You must therefore be very careful with this and have a good understanding of what parameter estimators the study should generate. The tendency will be to use such robust methods for smaller populations or subpopulations rather than for very large populations, because otherwise standard errors become too large. Knowledge about the content must contribute to the decision of how to deal with the outliers. If, for example, a person visits his or her garden allotment 400 times a year, this is not necessarily a reason to not include this person as a donor. However, suppose that this is a 50-year-old man from Assen; then there is little reason to strengthen this outlier by designating him as donor for a man of about the same age from Assen.

9.3 Selection of auxiliary variables

As a supplement to subsection 1.1.2.5, we provide several guidelines here about the selection of auxiliary variables.

- Select *x*-variables if you expect that they are also relevant for the item non-respondents. As a rule, you will still check whether the variables have significant explanatory value for the item respondents, because assessing the model for the item non-respondents is not possible.
- Do not include too many variables in a regression model. This will cause the parameters to be poorly estimated. For good predictions (imputations), choose a reasonably economical model.
- In donor imputation, however, it is not a problem if a distinction is made between many subpopulations (many variables with many categories). Even the addition of nonsense variables with the goal of being left with a unique donor is not a problem, but at the most, an alternative way of ultimately selecting a single random donor from a subpopulation. However, you must watch out for multiple donors; see section 6.3.
- The order of entering the variables in the model is a question of model selection. Use quality measures to quantify the benefit of adding a variable; for example, the increase in R^2, F test, AIC, BIC.

9.4 Non-negative variables with many zeroes

For activities in which some people do not participate, a distribution is created in which part of the population, the non-participants, scores zero and the other part, the participants, have a variety of positive values. Examples are the amount in euros spent by people on their vacation, the number of kilometres driven in their car, and turnover from a certain sideline activity. Hot deck methods work well with this type of variables, in the sense that they retain the distribution. However, if mean imputation is used, then no zeroes will be imputed. Regression imputation also creates problems. Negative imputations can occur for such non-negative variables. If the goal is only to estimate the population means, then this is not a big problem. But if you also want to keep the dispersion of the variables ‘reasonable’, or if you want to properly estimate the fraction of participants, then these techniques cannot be used. An option is then to perform the imputation in two steps. For example, you can first use a logistic regression to decide (impute) whether item non-respondents participate or not, and then determine the score for the assumed participants using a linear regression model.

9.5 Combination of methods (hierarchy)

If you want to perform imputation for missing values on a variable *y*, you can sometimes use a strategy with different methods or models, depending on the available auxiliary information for the record; see example 2 in section 4.4. In that example, information about the same variable in a previous period is first examined, then information from another source, and finally information about the same variable from the item respondents.

10. References

Banning, R., Camstra, A. and Knottnerus, P. (2010), *Methodenreeks: Thema: Steekproeftheorie, Deelthema's Steekproefontwerp en Ophoogmethoden [Theme: Sampling Theory, Subthemes: Sample design and Weighting methods]*. Methods Series document, Statistics Netherlands, The Hague [in Dutch; to be translated into English].

Boonstra H.J. and Buelens, B. (2007), *Theme: Model-based estimation, Subthemes: Synthetic estimators and Small area estimators*. Methods Series document, Statistics Netherlands, Heerlen [English translation from Dutch in 2011].

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004), *Applied Longitudinal Analysis*. Wiley, New York.

Gelman, A. and Hill, J. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press, Cambridge.

Harmsen, C.N. and Israëls, A.Z. (2000), *Nieuwe Huishoudensstatistiek: Nieuw versus Oud*. Internal report. Statistics Netherlands, Voorburg.

Hoogland, J., Loo, M.P.J. van der, Pannekoek, J. and Scholtus, S. (2010), *Data editing: detection and correction of errors*. Methods Series document, Statistic Netherlands, The Hague [English translation from Dutch in 2011].

Kalton, G. (1983), *Compensating for Missing Survey Data*. Survey Research Center Institute for Social Research, The University of Michigan.

Laan, D.J. van der and Kuijvenhoven, L. (2008), *Longitudinale analyse: Multilevel-modellen voor paneldata.* Internal report, Statistics Netherlands, Voorburg.

Lepkowski, J.M. (1989), *Treatment of wave nonresponse in panel surveys* In: Kasprzyk, D., Duncan, G.J., Kalton, G., Singh, M.P., eds., *Panel Surveys*. Wiley, New York.

Little, R.J.A. (1988), Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics* 6, 287-296.

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

Little, R. J. A. and Su, H.L. (1989), Item Non-response in Panel Surveys. In: Kasprzyk, D., Duncan, G.J., Kalton, G., Singh, M.P., eds., *Panel Surveys*. Wiley, New York.

Longford, N. (2005), *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician.* Springer-Verlag, New York.

Loo, M. van der and Pannekoek, J. (2007), *Advies gaafmaken en imputeren van de statistiek Bouwobjecten in Voorbereiding.* Internal report, Statistics Netherlands, Voorburg.

Molenberghs, G. and Verbeke, G. (2005), *Models for Discrete Longitudinal Data*. Springer-Verlag, New York.

Pannekoek, J., Harmsen, C., Huis, M. van and Prins, K. (2008) *Automatisch gaafmaken van GBA-gegevens met de "Nearest-neighbour Imputation Methodology"*. Internal report, Statistics Netherlands, The Hague.

Pannekoek, J. and Israëls, A.Z. (2000), Effecten van steekproefontwerp op (regressie) analyses. *Kwantitatieve Methoden* 65, 113-131.

Pannekoek, J. and D.C.G. Tempelman (2005). *Imputatiemethoden voor Impect-statistieken: deductieve imputatie en correctie voor overduidelijke fouten*. Internal report. Statistics Netherlands, Voorburg.

Pannekoek, J. and Waal, T. de (2005). Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics* 21, 257-286.

Rao, C.R. (1973*), Linear statistical inference and its applications 92nd ed*. Wiley, New York.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association* 91, 499-506.

Rubin, D.B. (1987), *Multiple imputation for nonresponse in surveys*. Wiley, New York.

Scholtus, S. (2008), *Automatisch gaafmaken van GBA-gegevens met CANCEIS*. Internal report, Statistics Netherlands, The Hague.

Schulte Nordholt, E. (1998), Imputation: methods, simulation experiments and practical examples. *International Statistical Review* 66, 157-180.

Starick, R. and Watson, N. (2006), *An Evaluation of Alternative Income Imputation Methods in the HILDA Survey*, HILDA Project Technical Paper Series, Melbourne Institute of Applied Economic and Social Research, The University of Melbourne.

Verbeke, G. and Molenberghs, G. (2000), *Linear mixed models for longitudinal data*. Springer-Verlag, New York.

Williams, T.R. and Bailey, L. (1996), *Compensating for Missing Wave Data in the Survey of Income and Program Participation (SIPP).* Proceedings of the American Statistical Association, Survey Research Methods Section, pp. 305–310.

Version history

Version	Date	Description	Authors	Reviewers
Dutch version: Imputatie				
1.0	18-12-2007	First Dutch version	Abby Israëls Jeroen Pannekoek Eric Schulte Nordholt	Eric Schulte Nordholt Edgar Soufan
1.1	23-01 2008	Changes to layout	Abby Israëls Jeroen Pannekoek Eric Schulte Nordholt	
1.2	02-03-2010	Additions to Chapter 7 'Multivariate imputatie'	Abby Israëls Jeroen Pannekoek Eric Schulte Nordholt	Eric Schulte Nordholt Edgar Soufan
1.3	30-11-2010	Additions to Chapter 8 'Methoden voor Longitudinale imputatie'	Abby Israëls Léander Kuijvenhoven Jan van der Laan Jeroen Pannekoek Eric Schulte Nordholt	Eric Schulte Nordholt Edgar Soufan
English version: Imputation				
1.3E	17-02-2011	First English version	Abby Israëls Léander Kuijvenhoven Jan van der Laan Jeroen Pannekoek Eric Schulte Nordholt	