Imputation of restricted data
Tempelman, Caren

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2007

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 1

An Introduction to Imputation and Editing

Statistics Netherlands collects and publishes information about all kinds of aspects of Dutch society. This information is based on data provided by persons, households, businesses, and so on. The economic figures published are of great importance to policy makers, researchers and businesses as they can be used to forecast or monitor economic quantities and to make informed business decisions. Some examples of a variety of economic statistics are the measurement of gross domestic product, use of energy, employment statistics, and investments of financial institutions. These statistics are based on survey estimates, which unfortunately are potentially subject to error. In this chapter we will discuss different sources of error and strategies that have been developed to deal with them.

1.1 Potential sources of error in survey estimates

Differences between survey estimates and the actual population parameters are caused by sampling as well as nonsampling errors. Sampling error is the error that arises due to surveying only a subset of the population rather than conducting a census on all businesses. Sampling errors would vanish if all businesses were sampled, and can be controlled by using a correct sampling design.
Nonsampling errors, however, can still occur when all businesses are surveyed. These errors arise during the data collection process and can be subdivided into the following categories, see for example Lessler and Kalsbeek (1992).

- **Frame error**
  The businesses that are to be surveyed are drawn from a sample frame, which is usually a business register. Frame errors arise due to discrepancies between the frame and the actual population. A large part of frame errors are caused by over- and undercoverage of businesses. Overcoverage occurs when businesses, that do not belong to the target population, are in the frame and undercoverage refers to businesses, belonging to the target population, that are not in the frame and therefore will never be sampled. Since business populations change rapidly due to births, deaths, and organisational changes, the updating of business registers is crucial in order to prevent large frame errors. Errors occurring in auxiliary variables, that are recorded in the business register such as size class or branch of business, are another type of frame error.

- **Nonresponse error**
  If the sampled business does not provide answers, we are confronted with nonresponse. In general two types of nonresponse are distinguished: unit and item nonresponse. Unit nonresponse refers to the case where the contacted business does not respond at all and item nonresponse refers to businesses that do respond, but omit answering some of the questions. There are several causes for nonresponse. First of all, some respondents refuse or are unable to provide data for a particular item or items, or sometimes they simply miss a particular question. Secondly, the interviewers may fail to record data items or data entry clerks may omit keying the data item. Besides, questionnaires may get lost or sent to the wrong address.
  If nonrespondents differ significantly from respondents in their answers, the error due to nonresponse can become large which may lead to biased estimates.

- **Measurement error**
  Measurement error is defined as the difference between the reported value and the true value. In general we distinguish between three sources of measurement error. First of all the true value may be unknown or difficult
1.2. Missing data

to obtain. Businesses may keep their accounting data according to different definitions than those used by the statistical agency. For example, the reference period used at businesses may differ from the reference period requested, e.g. financial year versus calendar year. Moreover, businesses may not even keep the information requested by the agency. In this case the effort required to obtain an answer may cause the respondent to guess the value or leave it blank. Note that this implies that high item nonresponse rates may be indicative of measurement errors. Secondly, questions may be misunderstood or misread by the respondent. The respondent is, for example, asked to report in thousands of Euros, but he or she actually reports in Euros, leading to a huge bias in the survey estimates if not corrected. Finally, the firms accounting system itself may contain errors. When measurement errors are present the difficulty is how to determine whether data are erroneous or not. Some mistakes are obvious, such as negative reported values for variables that should be non-negative, but others, such as differences in definitions that are used by respondents and statistical agencies, may be much harder to find.

• Processing error

After the data have been collected they pass several processes, such as keying, coding, editing, weighting and tabulation. Errors that arise during these processes are referred to as processing errors.

At Statistics Netherlands all business surveys are mail surveys, which means that the data need to be keyed onto a computer. Obviously this is subject to error as responses need to be interpreted. The numbers 1 and 7, for example, are easily mistaken for each other.

Other processing errors that may occur are mistakes in the software or adjustments of item values that appear to be in error but actually are correct.

In this thesis we will focus our attention on how to deal with missing data, this means that we will first elaborate on (item) nonresponse. As there is a strong relationship between dealing with missing data and handling erroneous data, the latter will also be discussed in more detail in this chapter.

1.2 Missing data

Missing values not only mean less efficient estimates because of the reduced size of the data set, but also that standard complete data methods cannot
straightforwardly be used to analyse the data. Moreover, possible bias exists because respondents are often systematically different from nonrespondents. If they do not differ, analysis of the responding items is sufficient to obtain valid inference. However, if they are systematically different analysis restricted to the respondents may result in serious bias. Assumptions about the selectivity of nonresponse are formalised by the missing data mechanism.

1.2.1 The missing data mechanism

The missing data mechanism concerns the reasons why values are missing, and in particular whether these reasons relate to values in the data set. Any analysis of data involving item nonresponse requires some assumptions about the missing data mechanism.

This mechanism can be formalised as follows, see Rubin (1976) and Little and Rubin (2002). Consider an \( n \times k \) data matrix \( X \), where \( X_{ij} \) is the value of variable \( j \) for respondent \( i \) and the joint distribution of \( X \) depends on a parameter vector \( \theta \). The main interest of the researcher is usually to obtain inference about the parameter vector \( \theta \) that determines the joint distribution of \( X \). We introduce a missing data indicator matrix \( M \) corresponding to the data matrix \( X \), where \( M_{ij} = 1 \) if \( X_{ij} \) is observed and zero otherwise (so \( M \) and \( X \) are of the same dimensions). Furthermore, let \( X_i \) denote the \( i \)th row and \( X_j \) the \( j \)th column of the data matrix \( X \), \( M_i \) and \( M_j \) are defined similarly. The data vector \( X_i \) can be partitioned in a missing and an observed part: \( X_i = (X_\text{mis}, X_\text{obs})' \). Partition \( M_i \) accordingly. This notation is somewhat sloppy as the missingness pattern may vary across respondents. However, as this notation is convenient and generally used and accepted, we will make use of it throughout this thesis as well.

In general the density of the missing data indicator is written as

\[
f(m_i | x_{i,\text{mis}}, x_{i,\text{obs}}, \phi),
\]

where we assume the parameter vector \( \phi \) to be distinct from \( \theta \). When the density of \( M_i \) can be simplified to depend solely on the parameter \( \phi \) and not the data, i.e. \( f(m_i | x_{i,\text{mis}}, x_{i,\text{obs}}, \phi) = f(m_i | \phi) \), then \( M_i \) is independent of \( X_j \). This means that the distribution of the indicator does not depend on the observed nor the missing data and the data are said to be missing completely at random (MCAR).

If the data are not MCAR it is important to establish whether the differences between nonrespondents and respondents can be explained by other reported or
auxiliary variables, such as business size or branch of business. If the conditional distribution of $M_i$ given $X_j$ does depend on the observed but not on the missing values of $X_j$, such that $f(m_i | x_{j,mis}, x_{j,obs}, \phi) = f(m_i | x_{j,obs}, \phi)$, then the missing values are missing at random (MAR). An example of this is when the missingness in certain variables depends on business size. For instance when small companies have a higher nonresponse rate for a certain variable than large companies.

Finally, if missingness depends on both the observed and the missing values of $X_j$, that is if the distribution of $M_i$ given in (1.1) cannot be simplified, the data are said to be not missing at random (NMAR). In other words the fact that the respondent answers a certain question depends on the actual item value for that question. In social surveys this is quite common, especially when sensitive information, such as income specifics or sexual preferences, is gathered. Respondents with either a relatively low or a relatively high income, for example, are more likely not to provide an answer to questions about their income than respondents with an income closer to the average. In business surveys we do not expect this kind of effect as the requested information is likely not to be perceived as sensitive by the person who is filling in the questionnaire. Besides, businesses are obliged by law to respond to questionnaires sent out by Statistic Netherlands.

If the data are MCAR or MAR the missing data mechanism is said to be ignorable and valid estimates can be obtained without explicitly modelling the missing data mechanism. If the data are NMAR, however, the missing data mechanism is nonignorable and needs to be modelled when estimating $\theta$. In general most methods for handling nonresponse in the survey literature assume that the missing data are MAR. Throughout this thesis we will also make this assumption.

### 1.2.2 Strategies for handling nonresponse

Three general strategies can be distinguished for dealing with missing data.

1. **Direct analysis of the incomplete data.**
   Most of the time cases with missing values are simply discarded. This is also referred to as complete case analysis. An advantage of complete case analysis is the ease of implementation, but a serious drawback is the rejection of information in the incomplete cases. Besides, most users are unaware of the fact that complete case analysis is only valid when the data are MCAR, which is not very likely.
For univariate analysis all cases where the variable of interest is observed can be included, which is referred to as available case analysis. An advantage is that all available information is used, but a disadvantage is that the sample base changes from variable to variable. There are also more elaborate methods that model the incomplete data, such as the Expectation Maximisation (EM) algorithm, developed by Dempster, Laird and Rubin (1977). The EM algorithm obtains maximum likelihood estimates in the presence of nonresponse. We will discuss the EM algorithm and its underlying theory in detail in chapter 2.

2. **Weighting.**
   In this case the nonrespondents are removed from the data set and weights are assigned to the respondents, based on auxiliary information from the sample frame. The survey estimates are calculated based on the respondents and their assigned weights. Weighting is often used to deal with unit nonresponse, where the sample frame contains the only available information. Weighting leads to valid inference if the missing data are MAR with respect to the auxiliary variables in the frame.

3. **Imputation.**
   A third strategy is to replace the missing values by an estimated value that can be constructed from the sample frame and the observed responses. This is referred to as imputation. More information is available when faced with item nonresponse as opposed to unit nonresponse, as now both information from the sample frame as well as information from the responses to other survey items is present. Therefore imputation is mostly used when dealing with item nonresponse. As opposed to weighting, imputation can also lead to valid inference if the missing data are MAR with respect to other variables in the frame.

As we are dealing with item nonresponse in this thesis, our focus will be on imputation of the missing data items. Imputation has several desirable features, see for example Kalton (1983). First of all if used correctly, the nonresponse bias will be reduced. In many surveys no compensation is made for missing data. If the nonrespondents are significantly different from the respondents, this will lead to seriously biased estimates; see Bethlehem and Kersten (1986). Secondly, imputation of the missing items will lead to a complete data set, which will make it easier to carry out statistical analyses. Finally, the results obtained from different analyses are bound to be consistent, a feature which need not apply with
an incomplete data set. Leaving the imputation to the users may lead to conflicting estimates due to different imputation methods employed. Moreover, at the statistical agency a wealth of external information about businesses is available in administrative records such as tax records, which are not accessible to the data users for reasons of confidentiality.

Imputation also has some serious drawbacks, however. First of all, imputation does not necessarily lead to estimates that are less biased than those obtained from the incomplete data set. Biases could arise depending on the imputation procedure, the actual missing data mechanism and the form of the estimate. Besides, there is also the risk that the analysts treat the completed data set as if they were all actual responses, and thereby underestimating the uncertainty caused by imputation. We will elaborate on this in section 1.2.4. First an overview of the most common imputation methods will be given.

1.2.3 An overview of imputation methods

There is considerable literature discussing imputation methods; see for example Kalton and Kasprzyk (1982, 1986), Kalton (1983), Sande (1982), Rubin (1987), Little (1988) and Kovar and Whitridge (1995). We will give an overview of the most common methods. Note that most imputation strategies make use of several imputation methods, depending on for example the type of variable or the available auxiliary information.

- **Logical (or deductive) imputation.**
  In this case the missing item value can be established with certainty from the other items. For example, only one component of a total is missing and can be easily deduced by subtraction, or some logical restrictions that must hold constrain the value to one possibility. That is, if the total operating expenses equal zero, and the costs of energy use are a part of total operating expenses and non-negative, then the energy costs must be equal to zero. Clearly this is the best imputation method possible, as the information needed can be derived with certainty from the observed data. This method should therefore always be applied first, before using any of the other imputation methods.

- **Mean imputation.**
  This method replaces the missing value by the mean of the responding units in a certain class for that item. The units are divided into classes according to their auxiliary variables, such as business size or type. Although respondent means are preserved, distributions of variables and
relationships between variables are seriously distorted if mean imputation is used. This is due to the fact that a peak is created at the average value of the variable and therefore the variation will be strongly underestimated.

- **Cold deck imputation.**
  This method substitutes the missing item value with a value taken from an external source. In business surveys, often data from a previous period are used. Obviously, trend and change will not be accounted for by using the value of the previous period, so extensions of this method adjust the values by modelling some sort of trend based on the reported items for that unit or other units in the sample. Other external sources could be administrative data such as tax records, which can be matched to the survey data.

- **Hot deck imputation.**
  In this instance the missing item values will be replaced with actual values from respondents in the present sample, which will be referred to as donors. This method is referred to as hot deck as the imputations are drawn from the present sample and not external data sources, which was the case for cold deck imputation. By using these donors the distribution of the population as represented by the sample is preserved. Another attractive feature of this method is that it is nonparametric, so no strong model assumptions need to be made in order to estimate the individual values. The only assumption that is made, is that the data are MAR with regard to the auxiliary variables. In general there are three common hot deck procedures:

  - **Sequential hot deck imputation.**
    The missing item value, that is filled in, is based on the value from the last responding unit preceding it in the data file. Usually, the data file is sorted according to business size or geographical location or in the case of qualitative auxiliary variables divided into subclasses based on these auxiliary variables. A problem that may occur is the multiple use of donors when there are a lot of missing values for a certain item, a feature that contributes to lowering the precision of survey estimates and underestimating variance. Another problem is that the imputed values depend completely on the order of the data set.

  - **Random hot deck imputation.**
    With this method respondents are divided into imputation classes
1.2. Missing data

based on auxiliary data, so that elements in the same class are considered similar. An item value of a randomly selected respondent within an imputation class will be assigned to the missing item value. Due to the stochastic nature of this method the variances will be better preserved.

- **Nearest neighbour imputation.**

  If there are several quantitative auxiliary variables, then the use of imputation classes may have undesirable effects. First, one has to choose the boundaries for the imputation classes. Secondly, two units may be matched whereas the first unit may be near the upper bound of a certain class and the second unit may be near the lower bound of that class. To overcome this problem some sort of distance function can be used to find the nearest record. The missing item value will then be replaced by the item value of this respondent. To avoid one single respondent being used as a donor several times, a component can be included in the distance function that reduces the multiple use of donors.

  This is a deterministic procedure, but can be randomized by finding several nearest neighbours and randomly choosing a donor from these. Furthermore, the variables need to be standardized before the distance between units is calculated in order to assign equal weights to each variable.

  For further reading on hot deck procedures see Ford (1983) and Sande (1983).

- **Regression imputation.**

  This method replaces the missing values of a certain item by predicted values from a regression of that variable on (some of) the other variables in the survey and the auxiliary variables in the sample frame. The regressors may be both quantitative and qualitative. The latter can be incorporated by dummy variables.

  An advantage of regression imputation is that information from a previous period can be easily exploited as well by simply adding predictor variables, hereby increasing the available information for imputation. The cold deck method, mentioned previously, that uses a trend to adjust missing values can be seen as a special case of regression imputation. Furthermore, unlike hot deck methods, where it may be difficult to find a suitable donor when imputing a somewhat outlying record, the regression approach will always
produce a replacement value. A disadvantage is that the conditional mean is imputed in this case and therefore this method corresponds to mean imputation mentioned earlier, which means it has the same undesirable properties with regard to variance estimation. This can be overcome to some extent by adding a random component.

- **Random regression imputation.**
  This procedure is the stochastic version of regression imputation. The missing item value will be replaced by the predicted value plus a random residual term. There are several possibilities for choosing residuals, such as assuming that the residuals are normally distributed with mean zero and variance equal to the residual variance of the regression. Another method would be to find a respondent with a similar predicted value and add this residual to the predicted value of the missing item. See Kalton (1983) for more suggestions. Using random regression imputation will correct somewhat for the underestimation of variance in the case of regular regression imputation.
  A problem that may occur with regression imputation (both deterministic and random) is that infeasible item values are imputed, such as for example negative predicted values for non-negative variables. One simple solution to this would be to set the imputed value equal to zero, but this could seriously distort the distribution of the residuals. A better solution would be to make use of (nonlinear) transformations, for example by taking the natural logarithm.
  Another disadvantage is the fact that regression imputation is strongly dependent on a model and consequently may be sensitive to model misspecification.

- **Predictive mean matching.**
  A method that combines both regression and hot deck imputation is predictive mean matching. This method was developed by Little (1988). First perform a regression of the variable that needs to be imputed on a set of predictor variables. Match the predictive mean of the missing item to the closest predictive mean of the responding records and then impute the actual item value of that respondent. This method is actually similar to nearest neighbour imputation, using the differences between predicted mean values as a distance measure. Its advantage over regression imputation is that only feasible values are imputed.
• Imputation using the Expectation-Maximisation (EM) algorithm.
  The EM algorithm was originally designed to obtain maximum likelihood
  estimates in the presence of nonresponse, but it can also be used for im-
  putation purposes. The EM algorithm consists of two steps. In the E-step
  the expected complete data loglikelihood is calculated based on the ob-
  served data and the current parameter values. Subsequently in the M-step,
  the parameter values are updated by maximising the expected complete
  data loglikelihood. This process is iterated until convergence. Imputa-
  tions can be generated deterministically by using the expected values for the
  missing data as imputations, which are often, but not always, a by-product
  of the E-step. Stochastic imputations are obtained by randomly drawing
  values from the specified distribution using the maximum likelihood es-
  timates as parameters.

**Example 1.2.3. The EM algorithm in SPSS**
The EM algorithm is being used extensively. A result of this is that well-known
statistical software, such as SPSS, provide procedures to apply the EM al-
gorithm. Care should be taken, however, as SPSS does not allow the user to
specify starting values for the algorithm, it automatically uses the available
cases estimates. But if the fraction of missing data is high, one may want to
evaluate the behaviour of the loglikelihood by starting from different values.

The fact that every variable in the survey is potentially subject to missing
data complicates the imputation task, as the variables that will be conditioned
on may contain missing values. In this case the EM algorithm is an attractive
and straightforward option, especially for high-dimensional data, as all missing
items are imputed simultaneously. The donor methods, however, will now prob-
ably lead to different donors for each missing item, which distorts associations
between items. If these associations are important, for example in multivariate
analysis, it may be wise to use one donor to fill in all the missing items of a
responding unit. The problem with this strategy is that it makes the imputation
process more complex. First of all, there will be fewer donors available as the
donors are not allowed to have any missing items for all variables that need to
be imputed. So the danger arises that one particular donor will be used multiple
times. Secondly, it will be difficult to create imputation classes that are homo-
genous with respect to all missing items. Similar problems arise for regression
imputation as standard regression techniques cannot deal with covariates that
contain missing values. In this case one could use the EM algorithm for normally
distributed data, which is an iterative form of regression imputation. Addition-
ally, adjustments are made to the covariance matrix in each iteration to correct for the fact that predicted means are imputed.

A serious defect of all imputation methods is that they invent data. More specifically, a single imputed value cannot represent all of the uncertainty about which value to impute, so analyses that treat imputed values just like observed values generally underestimate uncertainty. Moreover the usual estimates of variance are inadequate since they do not include the error due to estimation.

1.2.4 Variance estimation in the presence of imputation

Imputing the average value of a variable $X$ will lead to reasonably accurate estimates for the population mean or total of that variable if the data are MCAR or MAR. If the data are MCAR, the observed units simply are a random subset of the sample. Imputing the average does have an effect on the estimation of the population variance, however. Let $\mu$ denote the population mean and $\sigma^2$ the population variance. We will draw a sample of size $n$, which contains $r$ respondents and $m$ nonrespondents. The set of respondents is denoted by $R$ and the set of nonrespondents by $M$. The estimated population mean will be

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{i \in R} x_i + \sum_{i \in M} x_i.$$ 

Imputing the respondent mean $\bar{x}_r$ for all missing items leads to: $\hat{\mu} = \bar{x}_r$. Then

$$E[\hat{\mu}] = E[\bar{X}_r] = E\left[ \frac{1}{r} \sum_{i \in R} X_i \right] = E[X] = \mu,$$

if the data are MCAR, and $\hat{\mu}$ is an unbiased estimator of the population mean. Note that if the data are MAR imputing the respondent mean within imputation classes will lead to the same results.

Now the estimated population variance is

$$\hat{\sigma}^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x}_r)^2 = \frac{1}{n - 1} \sum_{i \in R} (x_i - \bar{x}_r)^2.$$

So

$$E[\hat{\sigma}^2] = \frac{r}{n - 1} E[(X - \bar{X}_r)^2] = \frac{r}{n - 1} \frac{r - 1}{r} \sigma^2 = \frac{r - 1}{n - 1} \sigma^2 < \sigma^2.$$

This means that the estimate of the population variance is calculated solely based on the variance of the respondents. The missing item values that were
imputed are not accounted for, which leads to an underestimation of the true population variance. Note that inflating the estimated variance by the factor \((n - 1)/(r - 1)\) will provide an unbiased estimate of the population variance.

Another important issue is the precision of these estimates, which is overstated if the imputed values are treated as actual observed values. As a result the confidence intervals of the estimates will be too short and consequently this may lead to erroneous conclusions. Stochastic imputation methods, such as random hot deck or random regression imputation, were introduced to somewhat correct for this. Although the variance of the estimate will be closer to the true variance, this still does not completely account for the extra component of variation introduced by the imputation process.

In general the total true variance consists of two components: the ordinary sampling variance and the nonresponse variance. Sampling variance is the variance that occurs due to surveying only a subset of the population and would vanish if the whole population was sampled. Nonresponse variance is introduced by the nonresponse in the sample, so if the sample would be completely observed no nonresponse variance would arise. In the literature this is mostly referred to as imputation variance, as imputation is used to deal with nonresponse most of the time. Since we are also employing incomplete data procedures, such as the EM algorithm and available case analysis, we prefer to use the term nonresponse variance.

The main purpose of the methods treated in this section is to obtain a variance estimate for the total variance of a parameter. This will produce a measure of data quality and will make valid inference possible. Additionally, knowledge of the nonresponse portion of the total variance reveals the impact of nonresponse (and consequently imputation) on precision. In the context of repeated surveys, resources can then be better allocated between a larger sample or increased follow-up and more advanced imputation procedures. In general three types of methods have been developed to calculate variance in the presence of nonresponse and imputation.

1. The analytical method

According to Särndal (1992) the total variance can be represented as

\[ V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}} + V_{\text{mix}} \]

where \(V_{\text{sam}}\) is the sampling variance, \(V_{\text{imp}}\) the imputation variance and \(V_{\text{mix}}\) a covariance between the sampled and the imputed variables, which is often negligible. \(V_{\text{sam}}\) is the sampling variance in case of a complete sample. This means that a correction needs to be made for the imputed
part of the data. These variances are now estimated separately. The problem with this approach is that Särndal only derived these quantities for simple random sampling with mean or ratio imputation. For more complex survey designs or imputation algorithms this method becomes analytically difficult.

2. Replication and resampling methods
Rao and Shao (1992) developed the adjusted jackknife technique for variance estimation after imputation. The adjusted jackknife variance estimate of an estimator is calculated as follows. Each sampled element \(i\), \(i = 1, \ldots, n\) is removed from the sample once and the missing items are imputed using the \(n - 1\) remaining respondents. Next the estimator of interest is calculated and the adjusted jackknife variance estimate is based on the \(n\) values for the estimator.

Another well-known variance estimation procedure that falls in this category is the bootstrap technique (Shao and Sitter, 1996). In this case a random sample with replacement of size \(n\) is drawn from the observed sample \(n\). This means that in the bootstrap sample some respondents may be present more than once and others may not be present at all. Next the bootstrap sample is imputed and the estimator of interest is calculated. The bootstrap variance estimate is based on the different values for the estimator for some bootstrap samples.

A major advantage of these methods is the fact that the variance of complicated estimators can be calculated relatively easy without the theoretical derivation of variance formulas as is the case with the analytical method.

3. Multiple imputation
Another way to estimate the variance due to imputation is to impute several, say \(m\), times and calculate the variance based on a combination of the within and the between variance of these \(m\) datasets. This is referred to as multiple imputation, which was developed by Rubin (1978, 1987). This method extremely simplifies variance estimation in the presence of imputation. However, the drawback is that the data user needs to incorporate \(m\) datasets in his analyses.

The main difference between the first two methods and multiple imputation, is the fact that in the first case missing data items are imputed only once, referred to as single value imputation, whereas with multiple imputation \(m\) imputations are generated for one missing data item. The actual consideration
between single value and multiple imputation is whether the benefit of an imme-
diate relatively simple variance estimation outweighs the simplicity of imputing
only once.

The obvious appeal of single value imputation is that it allows a straightforward
use of standard complete data methods and software. Besides it is easy to
implement and understand for data analysts. However, users may perceive the
imputed dataset as a truly observed dataset, calculating variances using regular
variance estimators. Unfortunately, the true variance of estimators can only be
calculated using more advanced procedures, such as the ones described above.

Multiple imputation is intuitively attractive since it incorporates the idea
that imputations have a certain variability. Although theoretically appealing,
multiple imputation is not used very often in (large) surveys or at statistical
agencies because of the practical implications. Multiple imputation requires
maintaining and storing multiple complete data sets, which is operationally
difficult. Besides, data dissemination such as the tabulation of data is seriously
complicated by multiple imputation. Moreover, in order to obtain valid inference
with multiple imputation, the imputation method used needs to be proper. That
is, the imputations should satisfy conditions 1-3 in Rubin (1987, p 118-119). In
words, a proper method is a method that has enough variability between rep-
lies to provide appropriate variance estimates. As noted by Rao (1996) some
commonly used imputation methods, including random hot deck and random
regression imputation, are improper because these draws do not represent the
full uncertainty in estimating the data for purposes of variance estimation with
multiple imputation. Another consequence of this is that multiple imputation
can only be used for random imputation methods, as there will be no variability
with deterministic procedures. Note that this means that the popular nearest
neighbour method cannot be used. Taking all this into account, statistical agen-
cies prefer the use of single value imputation and as valid variance estimates can
be obtained based on single imputation we will not recommend using multiple
imputation.

1.2.5 Concluding remarks on imputation

Although imputation is a commonly used and convenient method to deal with
missing data it is crucial that information about imputation is provided along
with each dataset, in order to inform data users about data quality. First of all
imputed values should always be flagged, so that the user will be able to distingui-
sh between observed and constructed data. Furthermore, the user should be
informed about what imputation techniques were used and additional informa-
tion should be given for certain techniques. For instance, if regression imputation was used the regressors should be specified, if nearest neighbour imputation was used the distance function and the number of times certain donors were used should be given, and if cold deck imputation was used the external sources should be described. Additionally nonresponse percentages and counts of the number of records that required at least one imputation are also informative.

In addition to this the statistical agency should evaluate the effects of imputation and if possible provide an estimate of the variance after imputation in order to give the user an insight in the precision of the data. Furthermore it should also be made clear to the user that relations between variables may be attenuated through imputation.

1.3 Erroneous data

As we mentioned in section 1.1, besides errors due to nonresponse, survey data can also be subject to measurement and processing error. In the case of nonresponse a missing data mechanism was developed, this can be done similarly for erroneous data in order to distinguish between stochastic and systematic errors.

1.3.1 The error mechanism

Stochastic errors are errors that are randomly introduced in the sample, for instance by a writing error of the respondent, by the respondent accidentally misreading or misunderstanding a question, or by a keying error made at the statistical agency. A distinction can be made between random errors throughout the sample and random errors within classes based on auxiliary data such as branch of business or size class. One may expect small companies to make more errors than large companies as their accounting system may be less well-kept.

On the other hand systematic errors are those errors that arise because respondents do not understand or misinterpret concepts, definitions or questions being asked, or because of faults in concepts or procedures of data collection and processing. Systematic errors have a greater potential to affect the quality of survey estimates than random errors; if a large number of respondents misinterpret a question in the same way, a bias will be introduced in the estimates. Examples are when gross values are reported instead of net values, and when values are reported in Euros instead of the requested thousands of Euros. Most of the time systematic errors are hard to locate as records will not seem suspicious or implausible compared to other respondents, who made the same
mistake. Besides, if a respondent reports all variables in gross values instead of net values, accounting identities will not be violated and the record will seem correct. In order to get some insight in systematic errors questionnaires should be tested and the observed data should be compared to several external sources, such as values from previous periods and tax records.

The most difficult issue with regard to erroneous data is how to determine whether data (items) are incorrect or not and subsequently how to correct data items once they are found to be in error. Several approaches have been used to deal with data that contain errors. These approaches will be discussed in the next section.

1.3.2 Strategies for dealing with errors

In general there are three ways to deal with errors in survey data.

1. Direct analysis of the erroneous data
   One option is to do nothing. This means that one assumes that the data do not contain errors. In the case of large systematic errors this approach may lead to a large bias in the estimates. If only small random errors are present in the data, the estimates will probably be reasonably accurate.

2. Measurement error models
   Another way to deal with erroneous data is to use measurement error models, also referred to as errors-in-variables models, see for example Wansbeck and Meijer (2000). A simple form of a classical measurement model is
   \[ X = X^* + \epsilon, \]
   where \( X \) is the measured value and \( X^* \) the true value, \( \epsilon \) represents the measurement error and is mostly assumed to be normally distributed with mean zero and covariance \( \Sigma \). Measurement error models also assume that the errors are independent of the actual value, which means that they are mostly only suitable for data with stochastic errors that are approximately normally distributed. The aim of measurement error modelling is to obtain survey estimates that are free from error. It, however, does not attempt to provide a complete correct dataset. Besides, the assumption that errors are normally distributed may be quite unlikely even for random errors.

3. Data editing
   A third option is to use data editing. Data editing is a procedure that loc-
ates individual errors using pre-specified rules, these errors will be removed
and subsequently the data item will be imputed. The main advantage of
data editing as opposed to measurement error models is the fact that data
editing will provide us with a corrected and consistent dataset, which can
then be analysed or released for publication. If the data published by a
statistical agency still contain obvious mistakes, even if they are small, the
agency will surely lose credibility.

As our main interest is to obtain a complete multi-purpose dataset our focus
will be on data editing. Editing is becoming increasingly important. As we said
before, without editing data intended for tabulation, publication and research
may be spurious, which could harm the perceived reliability of the statistical
agency.

1.3.3 Data editing

Editing is the localisation of erroneous or suspicious values in the data, which is
done based on checks. These checks are referred to as edit rules, edit restrictions
or simply edits.

In general two types of editing are distinguished. First of all editing can be
applied as a validating procedure. In this case editing is used to detect inco-
sistencies and errors within a certain record. Examples are checking whether
the parts of a sum add up to the reported total, and whether a ratio of values,
such as wages divided by the number of employees, is within certain bounds.
Other consistency edits are of the 'if-then' type, for example if a business is of a
certain size then the costs of wages need to be at a certain level. For economic
data most of the consistency edits are accounting definitions which must hold.
However, some of the consistency edits are constructed by subject-matter spe-
cialists based on previous experience or specific knowledge.

Secondly, editing can be applied as a statistical procedure. In this case edit-
ing is used to detect errors or inconsistencies across records. The edits are mostly
based on statistical analysis of the data. For example, by detecting outliers based
on the sample, or by using models to obtain edit limits for a certain variable
using a trend based on previous values of that record and the other responding
units. In business surveys outliers are common and difficult to treat because the
outlier may very well be a legitimate value and therefore should not be changed.
However, outliers should always be identified and excluded from the imputation
model as they may have a large effect on the imputed values.

Another classification of edits is in fatal and query edits. The first refers to
item values that do not satisfy edits, and therefore are erroneous with certainty, such as accounting identities that do not hold. Examples are the fact that profit has to equal turnover minus operating expenses and the fact that the purchasing price of goods should not exceed total operating expenses. If the record does not satisfy these restrictions it is certain that some items are incorrect.

Query edits concern item values that are highly unlikely, but may in fact be true. An example of a query edit is the fact that a value is expected to be within certain bounds based on other variables in the survey, but it need not be. Caution needs to be taken in the case of query edits, as the danger arises that the data are over-edited by using too many query edits or bounds that are too strict. Over-editing refers to the fact that the impact of editing on survey estimates is negligible and therefore the editing is unnecessary and thus needlessly expensive. Over-editing can, however, also refer to the practice that the data are edited too much in order to fit the expectations of the editor, resulting in possible bias. For further reading on this subject see Grauqust (1997).

A last important distinction that is often used in this thesis is the difference between balance and inequality edit restrictions. Balance edits refer to equality restrictions on the data and inequality edits refer to inequality restrictions. Note that balance edits are always fatal edits, but that inequality edits can be both fatal and query edits. Fatal inequality edits are, for instance, the fact that turnover should be non-negative or that the number of employees should exceed the number of employees in full time equivalent. A query inequality edit is the fact that turnover is expected to be larger than the number of employees multiplied by a constant.

### 1.3.4 The editing process at Statistics Netherlands

The traditional editing process at statistical agencies was generally as follows. Erroneous or suspicious data were located by means of edit rules and then subject-matter specialists corrected these values by recontacting the respondent or using their expert knowledge and external information to obtain imputations. In order to speed up the editing process and lower the associated costs, nowadays editing is also done automatically using some sort of model and a computer.

Currently at Statistics Netherlands a combination of manual and automatic editing is used, which is referred to as selective editing. In this case only the most influential and implausible records are reviewed manually, the other records are edited automatically. Each reporting unit is given a certain score based on several aspects, such as business size, sampling weight, relative importance of erro-
neas items and so on. Records that have a score above a certain predetermined value are reviewed manually. Clearly this results in a far more efficient editing process, especially since it has been recognised that it is not necessary to correct all data in every detail in order to obtain valid survey estimates, see Granquist (1997) and Granquist and Kovar (1997). Results published by a statistical office are usually aggregated data, such as totals or means, so small (random) errors will often cancel out when aggregated. Moreover, in this instance the measurement and processing error will become negligible with respect to the sampling error.

The automatic editing at Statistics Netherlands is done using validity editing. This means that if a record fails the stated validity edit restrictions it is considered to be in error. Once an erroneous record is encountered, the incorrect item values within that record need to be found. The widely used Fellegi-Holt method is employed for this purpose. Fellegi and Holt (1976) developed a procedure that locates the erroneous items within a failed record by assuming that the number of errors made are as few as possible. This means that the number of errors occurring in a record equals the minimum number of values that have to be changed within this record such that it satisfies all edit restrictions.

Once an item is found to be erroneous it is set to missing and needs to be imputed, which leads to the main subject of this thesis as these items need to be imputed taking the linear edit restrictions into account.

1.3.5 Concluding remarks on editing

As we have emphasized before, data users should always be aware of the fact that the data have been edited. This means that imputed items should be flagged. We would recommend different flags for items that were truly missing and items that were imputed because they were found to be in error, such that the user will be able to distinguish between them. Additionally, other numbers on editing should be provided along with the dataset. The number of edits that were violated for each record and the number of times a certain variable is present in violated edits will give the user insight in data quality. Furthermore, the number of times a variable has been changed and the relative change that has been made is also informative.

Currently editing is only used as a procedure to clean up data. It can, however, be much more useful. Editing provides crucial information about the quality of the data and the survey process and can therefore suggest valuable changes for future surveys, for example, by identifying nonsampling error sources.
1.4 The relationship between editing and imputation

There is a strong relationship between editing and imputation as the data that were found to be incorrect need to be imputed and additionally the imputed values need to satisfy the edit constraints. This seriously complicates the imputation process, especially for economic data which are subject to a large number of linear edit restrictions. The general imputation procedures that were described in section 1.2.3 do not take edit restrictions into account, which means that imputed values are likely not to satisfy the restrictions on the data.

Consider hot deck imputation as an example. In this instance the missing data items are replaced by responses from a similar donor record. If some of the operating expenses, which need to add up to the total operating expenses reported, are missing the imputed donor values will almost certainly not be consistent with this balance edit. This means that the imputed values need to be adjusted afterwards, in order to satisfy the restrictions. Although this can be done quite straightforwardly using an optimisation algorithm and requiring minimal change in the imputed values, the effects can be harmful as changing the imputed values will distort the distribution of the imputed values and consequently the distribution of the final completed dataset.

Besides, the edits provide the imputer with valuable information as they restrict the possible outcomes of the imputed values. For instance, if the total operating expenses is reported but some of the components that constitute this total are missing, the imputer does know the sum of the variables that need to be imputed. This information should definitely be incorporated in the imputation model. Therefore it is desirable to be able to generate imputations that are properly distributed while simultaneously satisfying all edit restrictions on the data.

Despite the importance of economic data no general procedures have been developed for the imputation of missing data items such that the edit constraints are utilised and satisfied. The main focus of this thesis will therefore be on the development of imputation methods that generate imputations satisfying all edit restrictions and preserving the distribution of the data simultaneously.

Example 1.2.3 continued. The EM algorithm in SPSS

The fact that the starting values of the EM algorithm cannot be specified in SPSS does not lead to serious problems in common well-behaved missing data situations. If we are dealing with economic data that are subject to balance
Table 1.1: Maximum likelihood estimates obtained by the EM algorithm.

<table>
<thead>
<tr>
<th></th>
<th>EM using AC estimates (SPSS)</th>
<th>EM using CC estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Costs of sales</td>
<td>44344 (1)</td>
<td>44335 (0)</td>
</tr>
<tr>
<td>Labour costs</td>
<td>5037 (15)</td>
<td>5086 (0)</td>
</tr>
<tr>
<td>Other personnel costs</td>
<td>839 (0)</td>
<td>860 (0)</td>
</tr>
<tr>
<td>Costs of transportation</td>
<td>509 (0)</td>
<td>532 (0)</td>
</tr>
<tr>
<td>Costs of energy</td>
<td>169 (0)</td>
<td>179 (0)</td>
</tr>
<tr>
<td>Costs of housing</td>
<td>630 (-1)</td>
<td>623 (0)</td>
</tr>
<tr>
<td>Costs of machinery and equipment</td>
<td>240 (0)</td>
<td>251 (0)</td>
</tr>
<tr>
<td>Sales expenses</td>
<td>1081 (2)</td>
<td>1077 (0)</td>
</tr>
<tr>
<td>Communication expenses</td>
<td>98 (n/a)</td>
<td>96 (n/a)</td>
</tr>
<tr>
<td>Costs of rendering of services</td>
<td>434 (0)</td>
<td>433 (0)</td>
</tr>
<tr>
<td>Other company expenses</td>
<td>596 (n/a)</td>
<td>234 (n/a)</td>
</tr>
<tr>
<td>Depreciation on fixed assets</td>
<td>770 (n/a)</td>
<td>770 (n/a)</td>
</tr>
<tr>
<td>Total operating expenses</td>
<td>54483 (-264)</td>
<td>54476 (0)</td>
</tr>
</tbody>
</table>

×1000 Euro

restrictions, however, the available cases (AC) starting values supplied by SPSS lead to imputations that do not satisfy the balance restrictions on the data. Whereas using starting values that do satisfy the balance restrictions, such as the complete cases (CC) estimates, does provide imputations that satisfy these restrictions.

This is illustrated by Table 1.1, where we have calculated the maximum likelihood estimates of the population means of different operating expenses in the trade industry by means of the EM algorithm, assuming normality. The data items are subject to a considerable amount of balance edits. First of all, the different operating expenses need to add up to the total operating expenses. Secondly, most of these operating expenses represent a total of an underlying balance edit restriction. The amount of violation of the underlying balance edit is shown between brackets. If there is no underlying balance edit, this is indicated by n/a (not applicable).

A formal proof of the fact that the EM estimates satisfy the balance restrictions if the starting values are chosen such that the restrictions are satisfied as
well will be given in chapter 4. Note that for now we have ignored the fact that the variables also need to satisfy non-negativity restrictions.

1.4.1 Linear balance and inequality restrictions

As we mentioned earlier a distinction often is made between balance and inequality restrictions. As these two types of restrictions have different implications for the imputation process, we will treat these restrictions separately as well. In this section we will give a more formal definition of the restrictions that need to hold throughout this thesis.

Consider an $n \times k$ data matrix $X$ and let $X_i$ denote the $i$th row of $X$, corresponding to the $i$th respondent. Now define a $p \times k$ restriction matrix $A$ that contains all balance restrictions on the data items, such that $AX_i = 0$. We will assume that there are no redundant balance restrictions, which means that the number of balance restrictions cannot exceed the number of variables and $A$ is of full row rank. Some examples of balance restrictions that occur are the fact that different company expenses add up to the total operating expenses, that financial result equals financial income minus financial expenses, or that the total number of employees equals the number of employees on the company’s payroll plus personnel lent out to other businesses.

Now define an $r \times k$ restriction matrix $B$ that contains all inequality restrictions on the data and the set $G := \{X_i \in \mathbb{R}^k : 1 \leq BX_i \leq u\}$, which defines all possible outcomes of the data vector $X_i$. The upper and lower bounds $u$ and $1$ may equal plus or minus infinity respectively. Note that in the case of non-negativity restrictions $B$ will be the identity matrix, $1 = 0$ and $u = \infty$. Also note that $B$ need not be of full row rank as the data may be subject to non-negativity restrictions as well as all kinds of other inequality restrictions, so the number of restrictions $r$ may exceed the number of variables $k$. Some examples of inequality restrictions are the fact the total number of employees is larger than the total number of employees in full time equivalent or the fact that the number of employees on a company’s payroll is smaller than the total labour costs (in thousands of Euros) and larger than 0.001 times the total labour costs.

So, the data completed by imputation need to satisfy $AX_i = 0$ and $1 \leq BX_i \leq u$. Due to the linear nature of these restrictions it is not possible to use nonlinear transformations, such as Box-Cox transformations, in a multivariate model as the edit structure would be lost in that case. For example, taking natural logarithms of the variables in the balance restriction $X_1 + X_2 + \cdots + X_{k-1} = X_k$ does not imply $\ln X_1 + \ln X_2 + \cdots + \ln X_{k-1} = \ln X_k$. This means that we cannot establish restrictions for the transformed data and consequently the im-
puted data are still likely not to satisfy the linear restrictions.

1.5 Overview of this thesis

The focus of this thesis is on the imputation of (economic) data that are subject to different types of linear restrictions. A strong need for imputation models that can incorporate restrictions has arisen at statistical agencies as this results in data for publication that do not contain obvious mistakes, i.e. violations of the linear restrictions, which could seriously harm the credibility of a statistical agency. In this study several imputation procedures are developed and analysed in order to provide the imputer with a set of models that can be used for varying types of restriction structures and datasets.

In chapters 3, 4, 5 and 6 we develop imputation methods that deal with restriction structures of varying complexity. For an overview of which restriction structures can be handled by the imputation models that are discussed in these chapters see Table 1.2, where A denotes the p × k matrix containing the balance restrictions and B the r × k matrix containing the inequality restrictions. Note that B can represent non-negativity as well as other inequality constraints.

Every chapter on imputation models is written according to the following structure. First, the different types of restrictions that can be incorporated in the imputation model are discussed. Next the suggested imputation model is treated and, subsequently, estimation procedures that provide parameter estimates for the model of interest are derived. Finally, deterministic and stochastic imputation procedures are discussed and applications to empirical data are presented.

First, in chapter 2 we discuss maximum likelihood estimation as throughout this thesis this procedure is used to obtain parameter estimates. Furthermore, the EM algorithm that was developed to obtain maximum likelihood estimates in the presence of nonresponse is treated extensively as well and as we may need to estimate high-dimensional integrals in the EM algorithm, Markov chain Monte Carlo methods are also be dealt with. The aim of this chapter is to provide a theoretical background that is needed for a complete understanding of the material of the later chapters in this thesis.

In chapter 3 we develop an imputation method that uses the Dirichlet distribution to model the data. This method is convenient because of its flexibility. This procedure can impute data items that are non-negative and subject to one linear balance restriction.

The Dirichlet method cannot incorporate multiple balance restrictions, however. Therefore, in chapter 4, we suggest the use of the multivariate singular
normal distribution to deal with multiple balance restrictions instead of only one. It is found that the EM algorithm for multivariate normally distributed data can be extended such that singular normal data can be managed as well. This leads to an imputation procedure that is easy to implement and whose properties are well-known.

As inequality restrictions are not incorporated in the singular normal model, there is still the need for a general purpose method that can handle all sorts of balance and inequality restrictions. With this objective, the multivariate singular normal density is truncated to the region defined by the inequality restrictions in chapter 5.

This truncated singular normal distribution consists of high-dimensional integrals and consequently leads to complex modelling issues, therefore a completely different approach is investigated in chapter 6. In this chapter the joint model is split into a sequence of univariate conditional distributions. These univariate conditional models are used to sequentially impute each variable. This method can incorporate both balance and inequality restrictions simultaneously as well.

Finally, in chapter 7 conclusions, directions for future research and examples of possible applications in other fields of interest are given.