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 6

Imputation of Data Using a Sequential Regression Approach

In the chapters 3, 4 and 5 we discussed imputation of data subject to different types of linear restrictions. Probability distributions are chosen such that the data can be modelled while incorporating both linear balance and inequality restrictions. The complexity of the probability models increases with the complexity of the structure of the restrictions that can be handled. The truncated singular normal appears to be the only model that can take all the linear restrictions that may occur into account. For economic data a disadvantage of this model is the dependence on the normal distribution without being able to make use of nonlinear transformations to approximate normality. In this chapter an imputation method is developed that uses the univariate full conditional distributions to model the variables separately, so that there is no need to specify a joint model.

6.1 Introduction

In the previous chapters we have developed several imputation techniques that model the distribution of the missing data conditional on the observed data, taking linear restrictions into account making use of the Dirichlet, the multivariate
singular normal or the multivariate truncated singular normal distribution. The differences between these methods are the complexity of the edit structure that can be handled and consequently the complexity of the imputation model. A property shared by all these methods is that they are multivariate and parametric. That is, the multivariate distributions of the variables are assumed to belong to a known family of probability distributions. Survey data, however, usually consist of a large number of variables, which may have several distributional forms. This means that, although ideally imputations should be draws from the conditional distribution of the missing values given the observed data, it may be difficult to find an appropriate multivariate model. In our case the search for this joint model is complicated by the fact that imputed values need to satisfy certain linear restrictions. In chapter 5 we used a truncated singular normal distribution to model data subject to restrictions. Although this method is theoretically sound, the distributional assumptions that need to be taken can be unrealistic in practice. When abandoning the dependence on the normal distribution we have, however, not been able to find an appropriate joint model that can take the linear restrictions into account.

In this chapter we will investigate an imputation method that makes use of univariate conditional distributions. The variables are modelled and imputed univariately using a sequence of regressions. This process is iterated so that the final imputed values converge to draws from the multivariate model.

First of all, in section 6.2 the different linear restrictions that may occur are discussed. Next in section 6.3 full conditional densities and problems of incompleteness are treated. The suggested imputation method is dealt with in section 6.4 and section 6.5 discusses the regression models used. Then section 6.6 deals with Box-Cox transformations to approximate normality and in section 6.7 it is explained how balance restrictions can be incorporated. Next, in section 6.8 some results for empirical data are presented and, finally, in section 6.9 some concluding remarks will be made.

### 6.2 Linear inequality and balance restrictions

Let $X$ be an $n \times k$ data matrix and let $X_i$ denote the data vector of order $k \times 1$ for respondent $i$. The inequality restrictions are represented by $1 \leq BX_i \leq u$, where the matrix $B$ is an $r \times k$ matrix containing $r$ inequality restrictions on the data. The upper and lower bounds $u$ and $l$ may equal plus or minus infinity, which means that the variables are truncated from one side only, e.g. by non-negativity constraints.
Additionally, consider a $p \times k$ restriction matrix $A$, with $p$ the number of linear balance restrictions, where it holds that $AX_i = 0$. We assume that there are no redundant balance restrictions, so $A$ is of full row rank.

We need to model the data $X_i$, $i = 1, \ldots, n$, which are subject to $1 \leq BX_i \leq u$ and $AX_i = 0$. Due to the added complexity caused by balance restrictions, we will first consider data subject to inequality restrictions only. This model will then be extended in section 6.7 to deal with balance restrictions as well.

### 6.3 Full conditional distributions

As opposed to specifying a joint model, another approach is to model the data through fully conditionally specified distributions. Let $X_j$ denote the $j$th variable and partition this into a missing, $X_j^{\text{mis}}$, and an observed part, $X_j^{\text{obs}}$. Note that this is different from $X_{i,\text{mis}}$ and $X_{i,\text{obs}}$ that was used in previous chapters, as missingness is now defined within variables and not within records. For notational convenience we assume that each variable $j$ contains at least one missing item value. Furthermore, let $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k)$. Now instead of an explicit joint model $f(x_i \mid \xi)$, the data are specified by separate conditional densities for each variable $X_j$ conditioned on the matrix $X_{-j}$ and the parameter vector $\xi_j$, $f_j(x_j \mid X_{-j}, \xi)$, which constitute a joint model. The idea is based on MCMC techniques, where Markov chains are used to generate draws from full conditionals in order to sample from multivariate, intractable probability distributions. Under certain conditions (see chapter 2) and if the Markov chain is long enough the draws stabilise to a stationary distribution, which is the distribution of interest.

In missing data situations the Bayesian point of view considers both parameters and missing data as random variables. Information about these unknown quantities is expressed in the form of a posterior distribution. MCMC methods are often applied for exploring these posterior distributions. Consider the posterior density $p(X^{\text{mis}}, \xi \mid X^{\text{obs}})$. This joint posterior can be simulated iteratively using univariate full conditional distributions, for example in combination with a Gibbs sampler. This would result in a sampler that cycles between draws for $\xi_j^{(t)}$, from

$$g_j^*(\xi_j \mid \xi_1^{(t)}, X_1^{(t)}, \ldots, \xi_{j-1}^{(t)}, X_{j-1}^{(t)}, \xi_j^{(t-1)}, X_j^{(t-1)}, \xi_{j+1}^{(t)}, X_{j+1}^{(t)}, \ldots, \xi_k^{(t-1)}, X_k^{(t-1)})$$

and draws for $X_j^{\text{mis},(t)}$ from

$$f_j^*(X_j \mid \xi_1^{(t)}, X_1^{(t)}, \ldots, \xi_{j-1}^{(t)}, X_{j-1}^{(t)}, \xi_j^{(t)}, X_j^{(t-1)}, \xi_{j+1}^{(t-1)}, X_{j+1}^{(t-1)}, \ldots, \xi_k^{(t-1)}, X_k^{(t-1)})$$
at each iteration $t$. This will constitute a Markov chain that converges to $p(X^{mis}, \xi | X^{obs})$. So, imputations for $X^{mis}$ can be obtained by using draws from this posterior distribution once the Markov chain has converged.

A major advantage of using these univariate full conditionals is that this approach is extremely flexible. Each variable can be modelled separately, which means that variable specific distributional properties can be taken into account. The downside of this is that it can be quite time-consuming to model each variable. Besides more structure is introduced by the univariate models, which means that they are more sensitive to model misspecification.

### 6.3.1 Incompatibility

Using full conditional densities without a specified joint model, unfortunately, has some drawbacks. In this instance the conditional models are specified directly, without choosing an encompassing multivariate model for the entire dataset. Due to this flexibility in the conditional distributions, there may be no implicit joint distribution underlying the imputation model. If the conditional distributions are incompatible and therefore do not constitute a joint model, there is no multivariate distribution to which the Markov chain and thus the imputed values converge.

Arnold and Press (1989) discuss necessary and sufficient conditions for the existence and uniqueness of a joint density, which cannot be easily used however, especially in multi-dimensional situations. For the bivariate case they derived that the conditional distributions $p_Y|X(y \mid x)$ and $p_X|Y(x \mid y)$ are said to be compatible if functions $u(x)$ and $v(y)$ exist, such that

$$
\frac{p_X|Y(x \mid y)}{p_Y|X(y \mid x)} = u(x)v(y),
$$

where $\int u(x) \, dx < \infty$.

The consequences of incompatibility are unclear and need to be investigated. There has been some research in this area, e.g. van Buuren et al. (2006) and Heeringa et al. (2002), and despite the lack of theory incompatibility does not appear to be a problem in practice. Gelman and Raghunathan (2001) state that if multivariate normality or other distributional assumptions are doubtful, it may make more sense to use separate regressions instead of a joint model. One argument is that having a joint distribution is less important than incorporating information from other variables and variable specific properties, such as variable type, semi-continuity, bounds and so on. But, although this approach seems
computationally attractive, caution needs to be applied and more research needs to be done in this area.

Another issue that arises due to incompatibility is the fact that the order of conditioning may play a role. If the univariate conditionals constitute a multivariate distribution the draws will converge to that stationary distribution irrespective of the order of conditioning. If the univariate conditionals are incompatible the distribution to which the draws converge changes with the different orders of conditioning. Note that this property can lead to an implicit test of (in)compatibility. For this purpose, generate a set of draws with different orders of conditioning and if the generated samples appear to be from different multivariate distributions this can be a sign of incompatibility.

Even in the case of compatibility some orders can still be better than others as they are more efficient. We suggest using the most efficient ordering when we are dealing with possible incompatibility. This means that the variables should be conditioned in order of increasing missingness. The missing variables will then be conditioned on as much observed variables as possible.

6.4 Sequential regression multivariate imputation

Based on the idea of using conditionally specified distributions, Van Buuren, Boshuizen and Knook (1999) and Raghunathan et al. (2001) have developed a general purpose imputation procedure for the imputation of multivariate missing data. In short, the algorithm starts with imputing an initial simple guess for the missing data items and next each variable is sequentially modelled and imputed by a regression model conditional on all the other variables. This regression model is specified for each variable separately and can therefore depend on the type of variable that is imputed and the appropriate restrictions that must hold. This sequence is repeated across cycles so that the final imputed values converge to draws from the multivariate distribution. In this way a multivariate problem is split into a series of univariate problems. This approach is referred to as sequential regressions, chained equations or regression switching and has been implemented in IVEware (Raghunathan et al., 2002) and MICE (Van Buuren and Oudshoorn, 1999).

In this chapter we will use a similar approach to impute economic data that are subject to multiple linear inequality and balance restrictions. The following imputation procedure is suggested, which is similar to the sequential regression

obtain \( \zeta_i^{(t)} \) by regressing \( X_i^{(t-1)} \) on \( X_2^{(t-1)}, \ldots, X_k^{(t-1)} \)

draw \( X_1^{mis,(t)} \) from \( f_1(x_1 | x_2^{(t-1)}, \ldots, x_k^{(t-1)}, \zeta_i^{(t)}) \)

obtain \( \zeta_2^{(t)} \) by regressing \( X_2^{(t-1)} \) on \( X_1^{(t)}, X_3^{(t-1)}, \ldots, X_k^{(t-1)} \)

draw \( X_2^{mis,(t)} \) from \( f_2(x_2 | x_1^{(t)}, x_3^{(t-1)}, \ldots, x_k^{(t-1)}, \zeta_2^{(t)}) \)

\vdots

obtain \( \zeta_k^{(t)} \) by regressing \( X_k^{(t-1)} \) on \( X_1^{(t)}, \ldots, X_{k-1}^{(t)} \)

draw \( X_k^{mis,(t)} \) from \( f_k(x_k | x_1^{(t)}, \ldots, x_{k-1}^{(t)}, \zeta_k^{(t)}) \).

This process is iterated for a sufficiently long time, such that the algorithm converges and the parameter estimates stabilise. Note that compatibility is not guaranteed in this case.

In order to initialise this algorithm a complete dataset, without missing values, is needed. So the first step of the algorithm is to impute the missing items with a simple imputation method. Several imputation models can be used for this purpose, but clearly the algorithm will converge faster if ‘good’ imputations are generated. Furthermore, note that it is crucial that these imputed values satisfy all linear restrictions on the data in order to avoid inconsistencies. This is not straightforward, we therefore suggest to use the EM algorithm for singular normal data, that was discussed in chapter 4, to obtain imputations after which the imputed values are adjusted with a simple procedure such that the inequality restrictions are satisfied. Although the EM algorithm converges relatively slowly and therefore will result in more computing time, using the EM algorithm seems a good approach as it is likely to generate imputations that are close to the final draws and the imputed values will immediately satisfy the linear balance restrictions on the data. Furthermore, we expect that adjusting the variables such that all inequality restrictions are satisfied can be done rather straightforwardly.

In contrary to the original sequential imputation procedure \( \zeta \) will not be drawn, but estimated. This approach ignores the uncertainty in estimating \( \zeta \) and is therefore deemed improper by Rubin (1987). It does, however, lead to acceptable approximations, when the fraction of missing data is modest (Little and Raghunathan, 1997) and when the number of observations for estimating \( \zeta \) is large.
6.5 Regression models

A major advantage of using univariate conditional distributions is the fact that variable-specific properties, such as semi-continuity, non-negativity, or linear restrictions, can be taken into account as variables are modelled separately.

Semi-continuity refers to variables that take on a single discrete value (e.g. zero) with positive probability, but are continuously distributed otherwise. In general in business surveys conducted by Statistics Netherlands variables are encountered that are either continuous or semi-continuous. Examples of semi-continuous variables are costs of hired personnel, costs of research and development and advertising costs.

In addition to this, most economic quantities are non-negative, such as for example expenses, revenues and the number of employees. Variables that can take on negative values are profit, financial benefits or provisions. It is important to realise that semi-continuity does not imply non-negativity. An example of this is other income, which is semi-continuous but not non-negative.

These different types of variables can also be subject to all sorts of linear inequality restrictions other than non-negativity constraints. For instance, the number of employees in fte (full-time equivalent) may not exceed the total number of employees, salary payments in thousands of Euros must be larger than the number of employees, and so on. In this case the domain of the variables is truncated to a certain region.

An added advantage of univariate modelling is the fact that nonlinear transformations can be employed. This is impossible for multivariate data as the edit structure will be lost after applying nonlinear transformations. Note that $X_1 + \cdots + X_k \leq X_{k+1}$ does not imply $T(X_1) + \cdots + T(X_k) \leq T(X_{k+1})$ for a nonlinear transformation $T(\cdot)$. In the univariate approach this will not be a problem, as the edit restrictions can be determined for each variable separately and transformed subsequently. For example, if we want to impute $X_1$ and model this variable using a nonlinear transformation, the resulting edit restriction will be $T(X_1) \leq T(\bar{X})$, with $\bar{X} = X_{k+1} - X_k - \ldots - X_2$. A consequence of this is that the variables can be transformed to approximate normality using Box-Cox transformations, which will be discussed in more detail in section 6.6. Once these variables are (approximately) normally distributed, the following regressions models can be used.

- **Normal linear regression model**
  The classical linear regression model will be used to handle continuous, unbounded variables.
• **Truncated normal regression model**
  The truncated regression model will be used for continuous variables that are subject to restrictions and therefore lie between certain bounds.

• **Logistic regression model**
  Semi-continuous variables can be dealt with in different ways. First of all, censored regression (or Tobit) models can be used for direct estimation of the parameters. Alternatively, a two stage procedure can be applied using a logit or a probit models, where the zero/non-zero status of the variable is determined in the first stage and the imputed value is obtained in the second stage using a (truncated) normal regression model for the non-zero records.
  All semi-continuous variables in business surveys at Statistics Netherlands have a probability mass at zero. Some of these variables are non-negative and need not satisfy any other inequality restrictions, which means that they can be straightforwardly modelled using a censored regression model. Some of the semi-continuous variables are unbounded, however, and can consequently take on both positive as well as negative values in which case censored regression models are an illogical choice. Besides, if the semi-continuous variables are bounded from above the model becomes complex as we need a combination of censored and truncated regression models then. As the two stage procedure can be applied for all instances, we prefer using this method.
  This means that a choice has to be made between logit and probit models. It has been demonstrated (e.g. Chambers and Cox, 1967) that logit and probit modelling often produce similar results. Therefore we suggest using the logit model in this instance as it was used by Raghunathan et al. (2001) as well.

Note that this imputation method is well-suited to deal with other types of variables, such as categorical or count variables and other properties such as skip patterns in the survey. As we are focusing on business surveys conducted by Statistics Netherlands, where these instances do not occur, we will not discuss these models at present. It is important to realise, however, that this model can be straightforwardly extended to handle social surveys. For a discussion of regression models for these variables we refer to Raghunathan et al. (2001). An overview of variable types, that appear in the business surveys of interest, and the accompanying suggested regression models is given in table 6.1.
6.5. Regression models

Table 6.1: An overview of variable specifics and the appropriate regression models.

<table>
<thead>
<tr>
<th>Variable type</th>
<th>Restrictions</th>
<th>Regression model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>Unbounded (e.g. financial result)</td>
<td>Normal</td>
</tr>
<tr>
<td></td>
<td>Bounded (e.g. turnover)</td>
<td>Truncated normal</td>
</tr>
<tr>
<td>Semi-continuous</td>
<td>Unbounded (e.g. other income)</td>
<td>Logistic and normal</td>
</tr>
<tr>
<td></td>
<td>Bounded (e.g. advertising costs)</td>
<td>Logistic and truncated normal</td>
</tr>
</tbody>
</table>

6.5.1 Classical linear regression model

The parameters that model $X_j$ in this case are $\beta$ and $\sigma^2$. So $\zeta$ is defined as $\zeta = (\beta', \sigma^2)$. Let $X_j$ denote the variable that will be re-imputed and let $X_{-j}$ denote the most recently updated matrix of predictors. Then

$$X_j = X_{-j}\beta_j + \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(0, \sigma_j^2 I_n).$$

The parameters are estimated by ordinary least squares

$$\hat{\beta}_j = (X'_{-j}X_{-j})^{-1}X'_{-j}X_j$$

$$\hat{\sigma}_j^2 = \frac{1}{n}(X_j - X_{-j}\hat{\beta}_j)'(X_j - X_{-j}\hat{\beta}_j).$$

Imputations are generated by drawing from the normal density using the estimated parameters. Let $m$ be the number of missing item values in variable $j$ and let $X^{mis}$ denote the matrix of predictors corresponding to the records that have missing values in variable $j$. Then

$$X^{mis}_j \sim \mathcal{N}(X^{mis}_{-j}\hat{\beta}_j, \hat{\sigma}_j^2 I_m).$$

6.5.2 Truncated regression model

In the presence of linear restrictions we cannot use the classical linear regression model, as $E[\epsilon_j | X_{-j}] \neq 0$ in this instance. The variable $X_{ij}$ conditional on $X_{i,-j}$
is now assumed to be normally distributed truncated to a region: \( 1 - \mathbf{B}_{-j} \mathbf{X}_{i,-j} \leq \mathbf{B}_j \mathbf{X}_{ij} \leq \mathbf{u} - \mathbf{B}_{-j} \mathbf{X}_{i,-j} \). This means that: \( l^*_i \leq X_{ij} \leq u^*_i \), where

\[
l^*_i = \max_{k=1,...,q} \frac{1}{b_{kj}} (1 - \mathbf{B}'_{-j} \mathbf{X}_{i,-j}), \text{ for } b_{kj} \neq 0
\]

\[
u^*_i = \min_{k=1,...,q} \frac{1}{b_{kj}} (\mathbf{u} - \mathbf{B}'_{-j} \mathbf{X}_{i,-j}), \text{ for } b_{kj} \neq 0,
\]

where \( b_{kj} \) is the \( kj \)th element of \( \mathbf{B} \). The variable \( X_{ij} \) is truncated normal: \( X_{ij} \sim \mathcal{N}(\mathbf{X}'_{i,-j} \mathbf{\beta}_j, \sigma^2_j) \), with \( l^*_i \leq X_{ij} \leq u^*_i \). The parameters of this truncated normal regression model are obtained by maximum likelihood estimation, see Amemiya (1973). The density of \( X_{ij} \) is

\[
f_{X_{ij}}(x_{ij} | \mathbf{\beta}_j, \sigma^2_j) = \frac{\psi(x_{ij} | \mathbf{\beta}_j, \sigma^2_j)}{\int_{l^*_i}^{u^*_i} \psi(x_{ij} | \mathbf{\beta}_j, \sigma^2_j) \, dx_{ij}},
\]

where \( \psi(\cdot) \) is the univariate normal density. Note that

\[
\psi(x_{ij} | \mathbf{\beta}_j, \sigma^2_j) = \frac{1}{\sigma_j} \phi\left(\frac{x_{ij} - \mathbf{X}'_{i,-j} \mathbf{\beta}_j}{\sigma_j}\right)
\]

and

\[
\int_{l^*_i}^{u^*_i} \psi(x_{ij} | \mathbf{\beta}_j, \sigma^2_j) \, dx_{ij} = \int_{l^*_i}^{u^*_i} \frac{1}{\sigma_j} \phi\left(\frac{x_{ij} - \mathbf{X}'_{i,-j} \mathbf{\beta}_j}{\sigma_j}\right) \, dx_{ij}
\]

\[
= \int_{l^*_i}^{u^*_i} \frac{1}{\sigma_j} \phi(v) \, dv
\]

\[
= \Phi(u^*_i) - \Phi(l^*_i),
\]

where \( \phi(\cdot) \) and \( \Phi(\cdot) \) respectively are the density and cumulative density function of the standard normal distribution. The upper and lower bounds are \( u^*_i = \frac{1}{\sigma_j} (x_{ij} - \mathbf{X}'_{i,-j} \mathbf{\beta}_j) \) and \( l^*_i = \frac{1}{\sigma_j} (l^*_i - \mathbf{X}'_{i,-j} \mathbf{\beta}_j) \). This means that the truncated density function for \( X_{ij} \) is

\[
f_{X_{ij}}(x_{ij} | \mathbf{\beta}_j, \sigma^2_j) = \frac{1}{\sigma_j} \phi\left(\frac{x_{ij} - \mathbf{X}'_{i,-j} \mathbf{\beta}_j}{\sigma_j}\right) \Phi(u^*_i) - \Phi(l^*_i).
\]
So the loglikelihood becomes

\[
\ell(\beta_j, \sigma_j \mid x_j) = -n \ln \sigma_j + \sum_{i=1}^{n} \ln \phi\left( \frac{x_{ij} - X_i' \beta_j}{\sigma_j} \right) - n \ln \left( \Phi(u_i^*) - \Phi(l_i^*) \right).
\]

In order to find the maximum likelihood estimates we need to differentiate the loglikelihood with respect to the parameters \( \beta_j \) and \( \sigma_j \):

\[
\frac{\partial \ell(\beta_j, \sigma_j \mid x_j)}{\partial \beta_j} = \sum_{i=1}^{n} \left( \frac{x_{ij} - X_i' \beta_j}{\sigma_j^2} + \frac{1}{\sigma_j} \phi(u_i^*) - \phi(l_i^*) \right) x_{i,-j} \\
\frac{\partial \ell(\beta_j, \sigma_j \mid x_j)}{\partial \sigma_j} = -n - \frac{1}{\sigma_j} \sum_{i=1}^{n} (x_{ij} - X_i' \beta_j)^2 + \frac{1}{\sigma_j} u_i^* \phi(u_i^*) - \phi(l_i^*) \\
+ \frac{1}{\sigma_j} l_i^* \phi(l_i^*).
\]

As these conditions cannot be solved analytically, we need to resort to iterative techniques, such as Fisher scoring. Recall that \( \zeta_j = (\beta_j, \sigma_j) \), then the maximum likelihood estimates are found by iterating, until convergence, over

\[
\zeta_j^{(t+1)} = \zeta_j^{(t)} + \nabla \ell(\zeta_j \mid x_j) \nabla \ell(\zeta_j \mid x_j) = \zeta_j^{(t)} + \nabla \ell(\zeta_j \mid x_j).
\]

where \( \nabla \ell(\zeta_j \mid x_j) = \sum_{i=1}^{n} \nabla \ell(\zeta_j \mid x_j) \) and \( I(\zeta_j \mid x_j) = \sum_{i=1}^{n} \nabla \ell(\zeta_j \mid x_j) \).

Imputations for \( X_{ij}^{\text{mis}} \) can then be obtained by drawing from the truncated normal density:

\[
X_{ij}^{\text{mis}} \sim \mathcal{N}^T(X_{i,-j} \hat{\beta}_j, \hat{\sigma}_j^2), \quad l_i^* \leq X_{ij}^{\text{mis}} \leq u_i^*.
\]

### 6.5.3 Logistic regression model

For semi-continuous data, we first need to establish whether the missing item value is zero or not. Let \( Y_{ij} \) indicate the zero/non-zero status of variable \( X_{ij} \), where 1 represents a non-zero status. Sometimes the value of \( Y_{ij} \) can be derived with certainty using information from the edit restrictions and other variables in the survey. For instance, let \( X_{ij} \) denote the costs of hired personnel. If no personnel is hired this variable equals zero, otherwise it is non-zero. This means that if this variable is missing but the number of hired personnel is observed, the zero/non-zero status can be derived with certainty. This is a form of deductive imputation and will be applied first.

For the missing values of \( Y_{ij} \) that cannot be derived in this manner we will
use a logistic regression model. Let $p_{ij} = \Pr(Y_{ij} = 1 \mid x_{i,-j})$, note that this equals $E[Y_{ij} \mid x_{i,-j}]$. The logit model is then

$$\logit p_{ij} = \delta_j' x_{i,-j}$$

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \delta_j' x_{i,-j}.$$ 

The logit function is the inverse of the cumulative logistic distribution function, which is $\Lambda(z) = e^z/(1 + e^z)$.

The coefficients of this regression model are estimated by means of maximum likelihood, where observations are treated as draws from the Bernoulli distribution. The first order conditions are

$$\frac{\partial \ell(\delta_j \mid y_j)}{\partial \delta} = X_{-j}' (y_j - \Lambda(X_{-j} \delta_j)),$$

where $\Lambda(X_{-j} \delta_j)$ is a vector of order $n \times 1$ with elements $\Lambda(\delta_j' x_{i,-j})$, $i = 1, \ldots, n$. The parameter estimates can be found using Newton-Raphson, as the Hessian can be straightforwardly derived:

$$\frac{\partial^2 \ell(\delta_j \mid y_j)}{\partial \delta_j \partial \delta_j'} = -X_{-j}' W X_{-j},$$

where $W = \text{diag}\{\Lambda(X_{-j} \delta_j) (\tau_n - \Lambda(X_{-j} \delta_j))'\}$. Note that since $Y_{ij}$ is not present in the Hessian, the Newton-Raphson algorithm is identical to Fisher scoring. Furthermore, also note that the Hessian is negative definite, which means that the loglikelihood has a global maximum.

Once $Y_{ij}$ is imputed, the variable $X_{ij}$ can be imputed by using the (truncated) linear regression model described in the previous sections, based on the records where $Y_{ij} = 1$.

### 6.6 Box-Cox transformations to normality

#### 6.6.1 Normal variables

The (truncated) regression models in this chapter depend on the assumption of a normal model. As economic variables are expected to approximate normal behaviour after an appropriate transformation this assumption seems reasonable. In
their seminal paper Box and Cox (1964) developed the Box-Cox transformation, defined by

\[
X(\lambda) = \begin{cases} 
\frac{(X+c)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\
\ln(X + c) & \text{if } \lambda = 0 
\end{cases}.
\]

The parameter \( c \) can be used to rescale \( X \) so that it is strictly positive, for now we will use \( c = 0 \).

A value for the parameter \( \lambda \) needs to be found, so that the variable \( X(\lambda) \) is as close to a normal distribution as possible. This is usually done by maximum likelihood. For a detailed description of the different estimation methods that can be used to find \( \lambda \), see Spitzer (1982). Assume that the transformed item values are normally distributed: \( X_i(\lambda) \sim \mathcal{N}(\mu, \sigma^2) \). The model parameters now are \( \mu, \sigma \) and \( \lambda \). The density for \( X_i(\lambda) \) becomes

\[
f_{X_i(\lambda)}(x_i(\lambda)) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i(\lambda) - \mu)^2\right).
\]

Using the change of variables formula, we can obtain the density for \( X_i \)

\[
f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(\frac{x_i^\lambda - 1}{\lambda} - \mu\right)^2\right) |x_i^{\lambda-1}|.
\]

This results in the following loglikelihood function

\[
\ell(\mu, \sigma, \lambda | \mathbf{x}) = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i(\lambda) - \mu)^2 + (\lambda - 1) \sum_{i=1}^{n} \ln x_i.
\]

For each fixed \( \lambda \) this likelihood equation is proportional to the likelihood equation for estimating \( \mu \) and \( \sigma \) if \( X_i(\lambda) \) was observed. This means that the maximum likelihood estimates are

\[
\hat{\mu}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} x_i(\lambda)
\]

\[
\hat{\sigma}^2(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (x_i(\lambda) - \hat{\mu}(\lambda))^2.
\]

Substitute these estimates in the likelihood equation, to obtain

\[
\ell(\lambda | \mathbf{x}) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} - n \ln \hat{\sigma}(\lambda) + (\lambda - 1) \sum_{i=1}^{n} \ln x_i,
\]
which is referred to as the profile or concentrated loglikelihood. This can be simplified by working with the alternative transformation

$$\tilde{X}(\lambda) = \begin{cases} 
\frac{(X)^{\lambda-1}}{\lambda g^{\lambda-1}} & \text{if } \lambda \neq 0 \\
\ln(X)g & \text{if } \lambda = 0 
\end{cases},$$

where $g$ is the geometric mean of the original response: $g = \prod_{i=1}^{n} x_{i}^{1/n}$, which equals $\exp(\frac{1}{n}\sum_{i=1}^{n} \ln x_{i})$. The profile loglikelihood then becomes

$$\ell(\lambda | \mathbf{x}) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} - \frac{n}{2} \ln \hat{\sigma}^{2}(\lambda),$$

where $\hat{\sigma}^{2}(\lambda)$ is the variance based on $\tilde{X}(\lambda)$. This function is maximised if $\hat{\sigma}^{2}(\lambda)$ is minimised. In practice $\hat{\sigma}^{2}(\lambda)$ is evaluated for several values of $\lambda$, e.g. from -2 to 2, with increments of 0.1, and the value of $\lambda$ is chosen where $\hat{\sigma}^{2}(\lambda)$ is minimal.

### 6.6.2 Truncated normal variables

If the original variable $X$ is bounded, e.g. $l \leq X \leq u$, the transformed variable, $X(\lambda)$, will be bounded as well: $l(\lambda) \leq X(\lambda) \leq u(\lambda)$. This means that the Box-Cox transformation obtains variables that are distributed according to a truncated normal instead of a normal distribution, see also Poirier (1978). The parameter $\lambda$ can now be obtained as follows.

Assume that $X_{i}(\lambda) \sim \mathcal{N}(\mu, \sigma^{2})$, where $l_{i}(\lambda) \leq X_{i}(\lambda) \leq u_{i}(\lambda)$. This means that the density for $X_{i}(\lambda)$ is

$$f_{X_{i}(\lambda)}(x_{i}(\lambda)) = \frac{\frac{1}{\sigma} \phi\left( \frac{x_{i}(\lambda) - \mu}{\sigma} \right)}{\Phi\left( \frac{u_{i}(\lambda) - \mu}{\sigma} \right) - \Phi\left( \frac{l_{i}(\lambda) - \mu}{\sigma} \right)}.$$

Again use the change of variables formula to obtain the density of $X_{i}$

$$f_{X_{i}}(x_{i}) = \frac{\frac{1}{\sigma} \phi\left( \frac{x_{i}(\lambda) - \mu}{\sigma} \right)}{\Phi\left( \frac{u_{i}(\lambda) - \mu}{\sigma} \right) - \Phi\left( \frac{l_{i}(\lambda) - \mu}{\sigma} \right)} \left| x_{i}^{\lambda-1} \right|,$$

which leads to the loglikelihood

$$\ell(\mu, \sigma, \lambda | \mathbf{x}) = -\sum_{i=1}^{n} \ln \left( \Phi\left( \frac{u_{i}(\lambda) - \mu}{\sigma} \right) - \Phi\left( \frac{l_{i}(\lambda) - \mu}{\sigma} \right) \right) - n \ln \sigma +$$

$$+ \sum_{i=1}^{n} \ln \phi\left( \frac{x_{i}(\lambda) - \mu}{\sigma} \right) + (\lambda - 1) \sum_{i=1}^{n} \ln x_{i}.$$
For fixed \( \lambda \) this is again proportional to the likelihood equation for estimating \( \mu \) and \( \sigma \), had we observed \( X_i(\lambda) \). Unfortunately, as we are dealing with truncated normal variables the maximum likelihood estimates for this equation are not available in closed form and need to be found with an iterative algorithm, which was described in section 6.5.2. The maximum likelihood estimates will be denoted by \( \hat{\mu}(\lambda) \) and \( \hat{\sigma}(\lambda) \). The profile loglikelihood then becomes

\[
\ell(\lambda | \mathbf{x}) = - \sum_{i=1}^{n} \ln \left( \Phi \left( \frac{u_i(\lambda) - \hat{\mu}(\lambda)}{\hat{\sigma}(\lambda)} \right) \right) - \Phi \left( \frac{L_i(\lambda) - \hat{\mu}(\lambda)}{\hat{\sigma}(\lambda)} \right) - n \ln \hat{\sigma}(\lambda) + \\
+ \sum_{i=1}^{n} \ln \phi \left( \frac{x_i(\lambda) - \hat{\mu}(\lambda)}{\hat{\sigma}(\lambda)} \right) + (\lambda - 1) \sum_{i=1}^{n} \ln x_i.
\]

Again, in practice, an estimate of \( \lambda \) can be found by evaluating several values of \( \lambda \), for example from -2 to 2 with increments of 0.1, and computing the appropriate bounds \( l(\lambda) \) and \( u(\lambda) \), the maximum likelihood estimates \( \hat{\mu}(\lambda) \), \( \hat{\sigma}(\lambda) \) and subsequently the profile loglikelihood. The optimal value for \( \lambda \) is the one for which the loglikelihood is maximal.

### 6.6.3 Conclusions

Note that since the support of the original response variable \( X \) is restricted to be positive in order to use the Box-Cox transformation, the transformed variable \( X(\lambda) \) actually always follows a truncated normal distribution for \( \lambda \neq 0 \), where \(-1/\lambda \) is the truncation point, as

\[-\infty \leq X(\lambda) \leq -\frac{1}{\lambda} \quad \text{if} \quad \lambda < 0 \]
\[-\frac{1}{\lambda} \leq X(\lambda) \leq \infty \quad \text{if} \quad \lambda > 0.\]

This fact is usually ignored when estimating \( \lambda \), as it is expected that the truncation point is in the tail of the distribution of \( X(\lambda) \). However, if the mass of the variable \( X \) is near zero, then subsequently the mass of \( X(\lambda) \) will be close to the truncation point as well. In this case it could be wiser to use the approach described in subsection 6.6.2 to obtain an estimate for \( \lambda \). If the mass of the response variable is away from zero, this method will yield the same estimates as the method for normal variables. However, due to the increased computer time we prefer to use the nontruncated estimation procedure whenever possible. To examine the effects of both estimation methods on empirical data, we will estimate \( \lambda \) for several variables in the business surveys, using both methods.
Table 6.2: Estimation of $\lambda$.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Nontruncated</th>
<th>Truncated</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{11}$</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$X_{14}$</td>
<td>-0.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>$X_{1t}$</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>$X_{2t}$</td>
<td>-0.2</td>
<td>-0.5</td>
</tr>
<tr>
<td>$X_{26}$</td>
<td>-0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>$X_{2t}$</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>$X_{t}$</td>
<td>-0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>$X_{r}$</td>
<td>-0.1</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

Example 6.6.3. Truncated versus nontruncated estimation of $\lambda$.

In this example we will investigate the effects of truncation on the estimation of $\lambda$ for empirical data. The variables used are gathered through the wholesale survey and, except for net turnover, have been used in previous chapters in the context of imputation. The following variables are examined

\[
\begin{align*}
X_{11} &= \text{gross wages and salaries} \\
X_{14} &= \text{other social costs} \\
X_{1t} &= \text{total labour costs} \\
X_{21} &= \text{costs of banking} \\
X_{26} &= \text{other costs of third party services} \\
X_{2t} &= \text{total costs of third party rendering of services} \\
X_{t}  &= \text{total operating expenses} \\
X_{r}  &= \text{net turnover}.
\end{align*}
\]

All of these variables are truncated and the following inequality restrictions need to hold

\[
\begin{align*}
0 & \leq X_{11} \cdot X_{14} \leq X_{1t} \\
0 & \leq X_{21} \cdot X_{26} \leq X_{2t} \\
0 & \leq X_{1t} \cdot X_{2t} \leq X_{t} \\
0 & \leq X_{r}.
\end{align*}
\]

The results are presented in Table 6.2, where the nontruncated column refers to $\lambda$ that is estimated using the procedure described in subsection 6.6.1 and
the truncated column refers to the estimation procedure in subsection 6.6.2. We find that the estimated values of \( \lambda \) for both methods are quite similar for most variables. The variables \( X_{21} \) and \( X_{26} \) produce the largest differences between estimation procedures. These variables are both semi-continuous, which could be the cause of this difference as \( \lambda \) is estimated based on a much smaller sample. As the number of semi-continuous variables is small and the difference between estimated \( \lambda \)'s is still acceptable, we do not expect difficulties when using the nontruncated procedure.

### 6.7 Incorporation of linear balance restrictions

A disadvantage of this univariate conditional approach is the fact that linear balance restrictions cannot be straightforwardly incorporated in the conditional densities. That is, if a variable is a component of a balance restriction, univariately its value is known with certainty based on the other variables in that balance restriction. For the algorithm this means that variables present in a balance restriction will remain stuck in the initial values, which obviously is undesirable.

We will solve this problem by leaving out the variables that are present in the balance restrictions such that the data does not contain singularities, which was inspired by section 5.7.2. If the data were completely observed this can be done without consequence as redundant information is discarded in that case. In the presence of nonresponse, however, this may lead to a loss of information because the information in the items values that are observed may not be redundant anymore as some of the variables in that balance restriction can contain missing values. Removing variables from the dataset before imputation and deriving these variables once the imputation process is completed may result in a substantial loss of predictive and distributional accuracy for that variable. To spread this loss across all variables and because the missingness is scattered across variables, we suggest to treat each record separately and to remove variables that are already missing.

For each record \( i \) partition \( X'_i = (X'_{i,\text{mis}} X'_{i,\text{obs}}) \) and \( A = (A_{i,\text{mis}} A_{i,\text{obs}}) \). Then \( \text{rank}(A_{i,\text{mis}}) = p_i \) is the number of variables that are redundant and therefore need to be left out, which was also discussed in section 5.7.2 and 5.7.3 for the multivariate truncated singular normal. Let \( A^n_{i,\text{mis}} \) be the \( p_i \times m_i \) matrix consisting of the nonredundant rows in \( A_{i,\text{mis}} \) and let \( A^n_{i,\text{mis},p_i} \) be a \( p_i \times p_i \) nonsingular submatrix of \( A^n_{i,\text{mis}} \). Note that there may be several options for \( A^n_{i,\text{mis}} \), that is if several variables in a balance restriction are miss-
ing each one of these could be discarded. Let \( m_i \) denote the number of missing items for record \( i \) and let \( q_i = m_i - p_i \). Then, at iteration \( t \), determine \( \mathbf{A}_{i,\text{mis}}^{n_i} \) for each record \( i \) and choose the appropriate \( \mathbf{A}_{i,\text{mis},p_i}^{n_i} \) and \( \mathbf{A}_{i,\text{mis},q_i}^{n_i} \), where the matrix \( \mathbf{A}_{i,\text{mis},q_i}^{n_i} \) contains the \( q_i \) columns of \( \mathbf{A}_{i,\text{mis}}^{n_i} \) that are not in \( \mathbf{A}_{i,\text{mis},p_i}^{n_i} \). Partition \( \mathbf{X}_{i,\text{mis}} = (\mathbf{X}_{i,\text{mis},p_i}, \mathbf{X}_{i,\text{mis},q_i}) \) accordingly, where \( \mathbf{X}_{i,\text{mis},p_i} \) refers to the elements in \( \mathbf{X}_{i,\text{mis}} \) that need to be left out. The variables that are removed from the dataset need to be eliminated from the set of inequality restrictions as well. This can be done by using the fact that \( \mathbf{X}_{i,\text{mis},p_i} = (\mathbf{A}_{i,\text{mis}}^{n_i})^{-1}(\mathbf{A}_{i,\text{obs}}^{n_i} \mathbf{X}_{i,\text{obs}} - \mathbf{A}_{i,\text{mis},q_i}^{n_i} \mathbf{X}_{i,\text{mis},q_i}) \) and filling this is for the \( \mathbf{X}_{i,\text{mis},p_i} \) in \( \mathbf{I} - \mathbf{B}_{i,\text{obs}} \mathbf{X}_{i,\text{obs}} \leq \mathbf{B}_{i,\text{mis}} \mathbf{X}_{i,\text{mis}} \leq \mathbf{u} - \mathbf{B}_{i,\text{obs}} \mathbf{X}_{i,\text{obs}} \), which will lead to a reduced set of inequality restrictions.

The algorithm will now be as follows. Start the algorithm at the first variable and calculate the regression parameters for this variable if it is present in \( \mathbf{X}_{i,\text{mis},q_i} \) conditioned on the other variables in \( \mathbf{X}_{i,\text{mis},q_i} \) and using the reduced set of inequality restrictions. Note that this results in several regression models as the missingness pattern varies across records. Records with similar missingness patterns, however, will obtain the same parameter estimates. Subsequently the missing items in this variable, that are present in \( \mathbf{X}_{i,\text{mis},q_i} \), are re-imputed with the appropriate regression model, the parameter estimates found and subject to the reduced set of linear inequality restrictions. After this the variables that were removed, \( \mathbf{X}_{i,\text{mis},p_i} \), need to be recalculated using the updated values for \( \mathbf{X}_{i,\text{mis},q_i} \) in order to make sure that the balance restrictions remain satisfied. Now proceed to the next variable and repeat this process until all \( k \) variables are treated.

As we stated before, leaving out data could result in a loss of information and consequently the predictive and distributional accuracy of the imputed values of \( \mathbf{X}_{i,\text{mis},p_i} \) may be reduced. As we have not found another approach to incorporate balance restrictions it appears that we cannot overcome this issue. To somewhat dampen these effects, we suggest choosing the variables that are left out for each record randomly in each cycle.

### 6.8 Imputation performance

In order to assess the performance of this imputation procedure on empirical data, we will use this method to impute data that have been gathered by Statistics Netherlands on a part of the wholesale industry for businesses with more than 10 employees. The effects of imputation on the estimation of population estimates as well as the ability of the imputation method to preserve individual
values and distributions are investigated.

6.8.1 Description of the data and generation of missing items

In this chapter we will use the dataset on labour costs that was introduced in chapter 3. The dataset consists of the variables

\[
X_{11} = \text{gross wages and salaries} \\
X_{12} = \text{social security costs} \\
X_{13} = \text{pension charges} \\
X_{14} = \text{other social costs} \\
X_{1t} = \text{total labour costs.}
\]

In chapter 3 we stated that the restrictions on this dataset are one balance restriction: \( X_{11} + X_{12} + X_{13} + X_{14} = X_{1t} \), and non-negativity restrictions on all data items. In fact this is not completely true as there are more inequality restrictions that must hold, which could not be incorporated by the imputation methods used in chapter 3. These inequality restrictions are

\[
X_{1t} \geq 1.01X_{11} \\
X_{1t} \leq 2X_{11} \\
X_{12} \geq 0.06X_{11} \\
X_{12} \leq 0.26X_{11}.
\]

The sequential imputation procedure is capable of incorporating these added inequality restrictions and therefore they will now be taken into account as well.

The missing data will be generated in the same way as was described in chapter 3. That is, assuming that the missing data are MCAR and using draws from the Bernoulli distribution to establish whether an item is missing or not. The amount of nonresponse created is also the same as in chapter 3. In this case, however, the amount of data imputed by deductive imputation may be somewhat larger as more restrictions are available. In order to be able to compare this method to other imputation procedures, we will not introduce missingness in the variable \( X_{1t} \). This means that we can impute the data with the procedures that were described in chapter 3 (disregarding the added inequality restrictions), which will provide more insight in the performance of the sequential imputation procedure. The remaining dataset with missing values is used for imputation.
6.8.2 Imputation using the sequential regression approach

The missing data will now be imputed using the sequential regression approach that was suggested in this chapter. This means that the variables $X_{11}, \ldots, X_{14}$ are sequentially modelled and imputed. The order in which this is done is based on the amount of missingness in each variable, starting with the variable with the least number of missing items and ending with the variable with the highest amount of nonresponse. This ordering is used as it has been found to be the most efficient.

As all of these variables are restricted by linear bounds, we will use a truncated regression model for each variable. Due to the skewness of the densities of these variables, they will be transformed to approximate normality using Box-Cox transformations. The value for the transformation parameter $\lambda$ is calculated based on the available cases for each variable.

As the data need to satisfy a balance restriction, one variable from the analysis needs to be eliminated as was described in section 6.7. For reasons of convenience this will not be done at random as was suggested, but using the following procedure. If a record contains missing values the first missing variable that is encountered in the order $X_{11}, \ldots, X_{14}$ is removed from the regression model. Note that this means that variable $X_{11}$ will always be discarded if it is missing and $X_{14}$ will always be used in the regression models. We therefore expect that the accuracy of the imputed variable $X_{11}$ will be somewhat lower. Further research can be aimed at choosing a variable at random.

The variable that is removed from the regression model also needs to be eliminated from the inequality restrictions. This can be done using the balance restriction. For example, if $X_{11}$ is left out, we can fill in $X_{11} = X_{14} - X_{13} - X_{12}$ in the inequality restrictions and determine the appropriate inequality restrictions that must hold for the variable that needs to be imputed. If the variable $X_{12}$ is missing this would lead to the following set of restrictions

\[
\begin{align*}
X_{12} & \geq \frac{0.01}{1.01} X_{14} - X_{14} - X_{13} \\
X_{12} & \leq \frac{1}{2} X_{14} - X_{14} - X_{13} \\
X_{12} & \geq \frac{0.06}{1.06} (X_{14} - X_{14} - X_{13}) \\
X_{12} & \leq \frac{0.26}{1.26} (X_{14} - X_{14} - X_{13})
\end{align*}
\]
and of course the non-negativity restriction: \( X_{12} \geq 0 \). Once a variable is imputed, the removed variable needs to be recalculated using the updated values of the imputed variable and the balance restriction in order to avoid inconsistencies.

A problem that arises using the sequential regression approach is the fact that the dataset contains two substantial outliers. These outliers have a large impact on the parameter estimates obtained using the truncated regression model. A result of this is that some of the expected values of the other records will fall outside the interval created by their lower and upper bounds, meaning that the algorithm may sometimes be unable to generate draws that fall inside this interval. Clearly, this is an undesirable effect and the outliers will therefore be removed from the dataset. It is a disadvantage of this method, however, that it is sensitive to outlying values, even once the data have been transformed.

If all variables have been imputed, the process has passed through one cycle. In order to assess convergence and determine the number of iterations that are needed for the process to stabilise we performed the imputation process, with 500 cycles, for a dataset with randomly assigned missing data. The estimates for the population means against the number of iterations are plotted in Figure 6.1. As was found by Van Buuren, Boshuizen and Knoek (1999) as well, the process appears to converge within a few iterations. We have repeated this procedure for several different realised sets of missing data, which resulted in similar behaviour. Therefore we will set the number of iterations used in the imputation process to 10.

### 6.8.3 The effects of imputation on parameter estimation

As the main target of our business surveys usually is to estimate the population estimates on mean and dispersion, \( \mu \) and \( \sigma \), the imputation procedure will be judged based on its ability to produce accurate point estimates. Another point of interest is the magnitude of the nonresponse variance component of the population parameters for this imputation procedure as this is an indication of the sensitivity of the procedure with respect to the realised set of missing data. Again we will conduct a simulation study to investigate the performance of the sequential regression imputation procedure.

In chapter 3 we found that incomplete data procedures, using the complete or available cases, result in both inaccurate point estimates and high nonresponse variance, due to the amount of missing data. These procedures will therefore not be used in this study. Another result obtained in chapter 3 is that the Dirichlet approach with expectations imputed (Dir) and the ratio nearest neighbour
method (RNN) produced the best results for this dataset. We will therefore use these methods in this study as well in order to be able to assess the performance of the sequential regression procedure (Seq). Note that these methods cannot take the added linear inequality restrictions into account, so the imputed values that are produced are likely to violate these restrictions.

The process of generating and imputing missing item values and subsequently estimating population parameters $\mu$ and $\sigma$ is iterated 100 times. The parameter estimates and their 95% confidence intervals are given in Figures 6.2 and 6.3. The sequential regression approach generates parameter estimates that are quite accurate for both $\mu$ and $\sigma$, with nonresponse variances that are similar to those obtained by the Dirichlet and nearest neighbour method. The sequential regres-

Figure 6.1: Plots of the population mean estimates against the number of iterations.
6.8. Imputation performance

Figure 6.2: 95% confidence intervals for the parameter estimates of $\mu$.

...ision approach even seems to somewhat outperform the Dirichlet procedure in this case, one needs to keep in mind, however, that this latter method is sensitive to outliers which have been removed from the dataset. This seriously reduces its general applicability.

Beforehand we stated that the accuracy of the parameter estimates $\mu_{11}$ and $\sigma_{11}$ will probably be lower than the accuracy of the parameters estimates of the other variables in the dataset obtained using the sequential regression imputation as the first variable is removed from the data most often. With regard to that variable we observe that the sequential regression method performs worse than the other methods when estimating $\mu_{11}$ and $\sigma_{11}$. This last quantity is strongly overestimated, which was to be expected since there will be more variability in $X_{11}$. This effect can be spread out across all variables by choosing the variable that needs to be removed at random, which will be a topic of further research.
Figure 6.3: 95% confidence intervals for the parameter estimates of $\sigma$.

6.8.4 The performance of the imputation method on item level

Other desirable properties arise if the main interest of the imputer is to obtain a general purpose dataset as well as accurate point estimates. Similar to chapter 3, the imputation methods will be judged on predictive accuracy, i.e. the ability of the method to preserve individual true values, and distributional accuracy, i.e. the ability to preserve the distribution of the true values. The average absolute deviation of the imputed data from the true data is calculated to assess the predictive accuracy and the Kolmogorov-Smirnov statistic, that measures the maximal distance between empirical distribution functions, is used to assess distributional accuracy. The results are presented in Table 6.3.

With regard to the average deviation between the imputed and the true data, the sequential regression method appears to perform slightly worse than the nearest neighbour method and quite a bit worse than the Dirichlet method,
Table 6.3: Predictive and distributional accuracy of the imputed data on labour costs.

<table>
<thead>
<tr>
<th></th>
<th>Average deviation</th>
<th>Kolmogorov-Smirnov distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dir</td>
<td>RNN</td>
</tr>
<tr>
<td>$X_{11}$</td>
<td>78</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>(10)</td>
<td>(14)</td>
</tr>
<tr>
<td>$X_{12}$</td>
<td>74</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>(11)</td>
<td>(12)</td>
</tr>
<tr>
<td>$X_{13}$</td>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>(11)</td>
<td>(12)</td>
</tr>
<tr>
<td>$X_{14}$</td>
<td>32</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>(6)</td>
<td>(7)</td>
</tr>
</tbody>
</table>

which was found to preserve individual values best in chapter 3.

With respect to distributional accuracy, the sequential regression approach performs quite well, especially for the variables $X_{11}$, $X_{12}$ and $X_{13}$. For the variable $X_{14}$ this method has used a model that developed specifically for semi-continuous data. This does result in a lower Kolmogorov-Smirnov distance for the semi-continuous variable $X_{14}$ as opposed to the Dirichlet method, but the K-S distance is still larger than the K-S distance of the nonparametric nearest neighbour method. So, it remains quite difficult to find an appropriate parametric model for semi-continuous data. The use of other models, such as probit or censored regression models, could also be investigated for this purpose, but we do not expect large differences in the outcomes.

6.9 Concluding remarks

In this chapter we have discussed the use of a sequential regression imputation procedure developed by Raghunathan et al. (2001) to impute data that are subject to inequality restrictions. This procedure models and imputes each variable separately using different regression models, based on variable-specific properties such as semi-continuity and linear bounds. Consequently, the procedure is extremely flexible and can deal with all kinds of data. We have extended this method in order to incorporate linear balance restrictions as well.
An important flaw of the sequential imputation procedure is the fact that the different univariate conditional models may be incompatible, that is they may not constitute a joint model. Although we did not encounter problems with convergence in section 6.8 on the application to empirical data, this should definitely be a topic for future research.

In section 6.8 we also found that the method appears to perform quite well with respect to preservation of population estimates and distributional accuracy of the imputed data. Therefore the sequential regression procedure appears to be a promising approach in the context of data imputation subject to linear restrictions. The huge advantage of being able to impute data while incorporating several restrictions is, however, toned down by the fact that the procedure is sensitive to outliers and can only incorporate balance restrictions by temporarily removing variables from the imputation procedure, resulting in less accurate parameter estimates. This method, however, remains the only developed imputation method that can impute data subject to both balance and inequality restrictions and that can be applied in practice.

This technique is very suitable for large datasets as the data are modelled with univariate distributions. This means that a large part of a business survey can be imputed simultaneously. So variables on company expenses, company turnover, company profits, employment and so on can be imputed all at once.

Another advantage of univariate modelling is that separate models can be used for different variables, which means that the variables are not required to be continuous anymore. This procedure can therefore be straightforwardly extended to impute all sorts of social surveys as well.