

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Bonn, Germany, 25-27 September 2006)

Topic (v): New and emerging methods

**Imputation of Economic Data Subject to Multiple Linear Restrictions Using a
Sequential Regression Approach**

Supporting paper

Submitted by Statistics Netherlands¹

ABSTRACT: Economic data have to satisfy many logical linear restrictions. In general we distinguish between two types of restrictions, i.e. balance and inequality restrictions. Balance restrictions refer to equalities that must hold, such as the fact that different operating expenses need to add up to the total. Inequality restrictions refer to inequalities that must hold, such as non-negativity constraints or the fact that the number of employees in fte may not exceed the total number of employees. The aim of this paper is to develop an imputation method that can deal with several types of inequality as well as balance restrictions simultaneously.

I. INTRODUCTION

1. In previous papers, see Tempelman (2003) and (2005), 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 drawn from the conditional distribution of the missing values given the observed data, it may be difficult to find an appropriate multivariate model. In this paper we will therefore investigate an imputation method that makes use of fully 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.

2. First of all, in section II the different occurring linear restrictions are discussed. Next in sections III and IV full conditional densities and problems of incompatibility are treated. The suggested

¹Prepared by Caren Tempelman (DTMN@cbs.nl)

imputation method and its specifics is dealt with in section V and section VI discusses the regression models used. Then in section VII it is explained how balance restrictions could be incorporated. Finally, we will end with some discussion in section VIII.

II. LINEAR INEQUALITY AND BALANCE RESTRICTIONS

3. Let \mathbf{X} be an $n \times k$ data matrix and let \mathbf{X}_i denote the data vector of order $k \times 1$ for respondent i . The inequality restrictions are represented by $\mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}$, where the matrix \mathbf{B} is an $r \times k$ matrix containing r inequality restrictions on the data. The upper and lower bounds \mathbf{u} and \mathbf{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 \mathbf{A} , with p the number of linear balance restrictions, where it holds that $\mathbf{A}\mathbf{X}_i = \mathbf{0}$. We assume that there are no redundant balance restrictions. This means that \mathbf{A} is of full rank.

4. So we need to model the data \mathbf{X}_i , $i = 1, \dots, n$, which is subject to $\mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}$ and $\mathbf{A}\mathbf{X}_i = \mathbf{0}$. 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 VII to deal with balance restrictions as well.

III. FULL CONDITIONAL DISTRIBUTIONS

5. As opposed to specifying a joint model, another approach is to model the data through fully conditionally specified distributions. Let \mathbf{X}_j be the j th variable and partition this into a missing, \mathbf{X}_j^{mis} , and an observed part, \mathbf{X}_j^{obs} . Note that missingness is 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 $\mathbf{X}_{-j} = (\mathbf{X}_1, \dots, \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, \dots, \mathbf{X}_k)$. Now instead of an explicit joint model $f(\mathbf{x} | \boldsymbol{\xi})$, the data are specified by separate conditional densities for each variable \mathbf{X}_j , $f_j(\mathbf{x}_j | \mathbf{X}_{-j}, \boldsymbol{\xi})$, 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 and if the Markov chain is long enough the draws stabilise to a stationary distribution, which is the distribution of interest.

6. 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 $p(\mathbf{x}^{mis}, \boldsymbol{\xi} | \mathbf{x}^{obs})$. This joint posterior can be simulated iteratively as follows. At each iteration t , given the current value $\boldsymbol{\xi}^{(t)}$ of the parameter, $\mathbf{X}^{mis,(t)}$ is drawn from its conditional density $p(\mathbf{x}^{mis} | \mathbf{x}^{obs}, \boldsymbol{\xi}^{(t)})$. Next $\boldsymbol{\xi}^{(t+1)}$ is drawn from $p(\boldsymbol{\xi} | \mathbf{x}^{mis,(t)}, \mathbf{x}^{obs})$. This will constitute a Markov chain that converges to $p(\mathbf{x}^{mis}, \boldsymbol{\xi} | \mathbf{x}^{obs})$. So, imputations for \mathbf{X}^{mis} can be obtained by using draws from this posterior distribution once the Markov chain has converged.

7. 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 straightforwardly taken into account.

IV. INCOMPATIBILITY

8. Using full conditional densities without a specified joint model, however, also 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 distribution to which the Markov chain and thus the imputed values converge.

9. Arnold and Press (1989) discuss necessary and sufficient conditions for the existence and uniqueness of a joint density. The conditional distributions $p_{Y|X}(y | x)$ and $p_{X|Y}(x | y)$ are said to be compatible if there exists functions $u(x)$ and $v(y)$ such that

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

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

10. 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. Satisfactory theory remains to be developed however. 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 method seems computationally attractive, caution needs to be applied.

11. 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 compatible multivariate distribution the draws converge to a stationary distribution irrespective of the order of conditioning. Some orders can still be better than others, however, 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.

V. SEQUENTIAL REGRESSION MULTIVARIATE IMPUTATION

12. Based on the idea of using conditionally specified distributions, Raghunathan et al. (2001) have developed a general purpose multivariate imputation procedure in order to deal with survey data that are difficult to model by means of a multivariate density. The variables are imputed univariately through a sequence of regressions, where the regression model depends on the type of variable that is imputed and the appropriate restrictions. Imputations are drawn from the posterior predictive distribution that is specified by the regression model and an uninformative prior. 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.

13. Let \mathbf{Z} denote a complete $n \times s$ matrix with s auxiliary variables, such as size class, branch of business, item values from previous periods, other surveys or registers. Categorical auxiliary variables are represented by dummy variables in \mathbf{Z} . Again, let the $n \times k$ matrix \mathbf{X} denote the data of interest, within which without loss of generality the variables are ordered by the amount of missing values. The missing data can be modelled through a joint model: $f(\mathbf{x}^{mis}, \boldsymbol{\xi} \mid \mathbf{Z}, \mathbf{x}^{obs})$. In Tempelman (2005) we have used a truncated (singular) normal distribution to model this type of data. However, when abandoning the dependence on the normal distribution we have not been able to find an appropriate joint model that can take the linear restrictions into account. An alternative to the joint modelling is to use 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

$$g_j^*(\xi_j \mid \mathbf{Z}, \xi_1^{(t)}, \mathbf{x}_1^{(t)}, \dots, \xi_{j-1}^{(t)}, \mathbf{x}_{j-1}^{(t)}, \mathbf{x}_j^{(t-1)}, \xi_{j+1}^{(t-1)}, \mathbf{x}_{j+1}^{(t-1)}, \dots, \xi_k^{(t-1)}, \mathbf{x}_k^{(t-1)})$$

and draws for $\mathbf{X}_j^{(t)}$ from

$$f_j^*(\mathbf{x}_j \mid \mathbf{Z}, \xi_1^{(t)}, \mathbf{x}_1^{(t)}, \dots, \xi_{j-1}^{(t)}, \mathbf{x}_{j-1}^{(t)}, \xi_j^{(t)}, \xi_{j+1}^{(t-1)}, \mathbf{x}_{j+1}^{(t-1)}, \dots, \xi_k^{(t-1)}, \mathbf{x}_k^{(t-1)}).$$

It remains difficult, however, to derive these full conditional densities.

14. Raghunathan et al. (2001) therefore suggest using the following approximation

$$\begin{aligned} \text{draw } \zeta_1^{(t)} & \text{ from } g_1(\zeta_1 \mid \mathbf{Z}, \mathbf{x}_1^{(t-1)}, \dots, \mathbf{x}_k^{(t-1)}) \\ \text{draw } \mathbf{X}_1^{mis,(t)} & \text{ from } f_1(\mathbf{x}_1 \mid \mathbf{Z}, \mathbf{x}_2^{(t-1)}, \dots, \mathbf{x}_k^{(t-1)}, \zeta_1^{(t)}) \\ \text{draw } \zeta_2^{(t)} & \text{ from } g_2(\zeta_2 \mid \mathbf{Z}, \mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t-1)}, \dots, \mathbf{x}_k^{(t-1)}) \\ \text{draw } \mathbf{X}_2^{mis,(t)} & \text{ from } f_2(\mathbf{x}_2 \mid \mathbf{Z}, \mathbf{x}_1^{(t)}, \mathbf{x}_3^{(t-1)}, \dots, \mathbf{x}_k^{(t-1)}, \zeta_2^{(t)}) \\ & \vdots \\ \text{draw } \zeta_k^{(t)} & \text{ from } g_k(\zeta_k \mid \mathbf{Z}, \mathbf{x}_1^{(t)}, \dots, \mathbf{x}_{k-1}^{(t)}, \mathbf{x}_k^{(t-1)}) \\ \text{draw } \mathbf{X}_k^{mis,(t)} & \text{ from } f_k(\mathbf{x}_k \mid \mathbf{Z}, \mathbf{x}_1^{(t)}, \dots, \mathbf{x}_{k-1}^{(t)}, \zeta_k^{(t)}). \end{aligned} \tag{1}$$

This process is iterated for a sufficiently long time, such that the algorithm converges and \mathbf{X}_j^{mis} approximates draws from $f(\mathbf{x}^{mis}, \boldsymbol{\xi} \mid \mathbf{Z}, \mathbf{x}^{obs})$. Note that compatibility is not guaranteed in this case.

15. In order to start this algorithm a complete dataset is needed. So the first step of the algorithm is to impute the missing items with a simple imputation method to initialise the iterative algorithm. The proposed method is to regress \mathbf{X}_1 on \mathbf{Z} , assuming a flat prior for the regression coefficients, and subsequently imputing \mathbf{X}_1^{mis} . Next \mathbf{X}_2^{mis} is imputed by regressing \mathbf{X}_2 on $(\mathbf{Z}, \mathbf{X}_1)$, where \mathbf{X}_1 now consists of both observed and imputed values, and so on. After the initialisation step, the iterative imputation process for cycles $t = 2, \dots, c$ in model (1) is started, where the conditional densities, for $j = 1, \dots, k$, f_j are specified by a regression model that depends on the type of variable \mathbf{X}_j and the regression parameter ζ_j , which has posterior distribution g_j .

A. Some modifications with respect to the algorithm by Raghunathan et al.

16. The approach that will be used in this paper to impute missing economic data subject to linear restrictions is based on the method suggested by Raghunathan et al. (2001). Some modifications will, however, be made. First of all, in our case $\boldsymbol{\zeta}$ will not be drawn, but estimated. This approach ignores the uncertainty in estimating $\boldsymbol{\zeta}$ 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 ζ is large. Besides, it is expected that reducing the variability in ζ will reduce the variability in the imputed values and therefore result in more stable imputations.

17. Another point concerns the initialisation step. A problem with the approach suggested by Raghunathan et al. is the fact that the resulting imputed values may not satisfy all restrictions leading, to inconsistencies for subsequent variables. For instance, assume that the variable \mathbf{X}_1 contains fewer missing values than \mathbf{X}_2 , which contains fewer missing items than \mathbf{X}_3 . Furthermore, assume that it has to hold that $X_{i1} \leq X_{i3} \leq X_{i2}$. If for a specific record i , X_{i2} is observed and X_{i1} and X_{i3} are missing, then X_{i1} is imputed first, while not conditioning on \mathbf{X}_2 or \mathbf{X}_3 . Then the imputed value for X_{i1} may exceed X_{i2} . If so, once X_{i3} is imputed while conditioning on \mathbf{X}_1 and \mathbf{X}_2 , this leads to impossible bounds for X_{i3} . In this case the algorithm will not be able to obtain a valid imputation for X_{i2} . A solution would be to ignore the restrictions in this imputation phase and to adjust the imputed values afterwards using optimisation techniques. Another, faster, method is to use an existing imputation method for the missing items, after which the imputed values are adjusted so that the restrictions are satisfied. An option is to use an elaborate imputation algorithm, in order to obtain initial imputations that are close to the final draws. For this purpose a regression imputation in combination with the EM algorithm can be used, where afterwards again the variables need to be adjusted in order to satisfy the linear inequality restrictions. This method is reasonably fast and easy to implement.

VI. REGRESSION MODELS

18. 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.

19. 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 economic 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 exceptional income, which is semi-continuous but not non-negative. Furthermore, 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 exceed the number of employees, and so on. In this case the domain of the variables is truncated to a certain region.

20. Another advantage of univariate modelling is the fact that nonlinear transformations can be used. This is impossible for multivariate data as the edit structure will be lost after applying nonlinear transformations. Note that $X_1 + \dots + X_k \leq X_{k+1}$ does not imply $T(X_1) + \dots + T(X_k) \leq T(X_{k+1})$ for a nonlinear transformation $T(\cdot)$. In the univariate approach this will not be a problem, as the edits 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

TABLE 1. *An overview of variable specifics and the appropriate regression models*

Variable type	Restrictions	Regression model
Continuous	Unbounded (e.g. profit)	Normal
Continuous	Bounded (e.g. turnover)	Truncated normal
Semi-continuous	Unbounded (e.g. exceptional income)	Logistic and normal
Semi-continuous	Bounded (e.g. advertising costs)	Logistic and truncated normal

restriction will be $T(X_1) \leq T(\tilde{X})$, with $\tilde{X} = X_{k+1} - X_k - \dots - X_2$. A consequence of this is that the variables can be transformed to approximate normality using Box-Cox transformations.

21. 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 logistic regression, 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 our economic surveys have a probability mass at zero. Some of these variables are non-negative and need not satisfy other inequality restrictions, which means that they can be straightforwardly modelled using a censored regression model. However, some of the semi-continuous variables are unbounded and can 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 using logistic regression models can be applied for all instances, we prefer using this method.

22. Note that this imputation method is well-suited to deal with other types of variables, such as categorical or count variables and other specific properties such as skip patterns in the survey. As we are focusing on economic surveys conducted by Statistics Netherlands, where these instances do not occur, we will not discuss these models at present. An overview of common variable types and accompanying regression models is given in table 1.

A. Classical linear regression model

23. The parameters that model \mathbf{X}_j in this case are $\boldsymbol{\beta}$ and σ^2 . So $\boldsymbol{\zeta}$ is defined as $\boldsymbol{\zeta} = (\boldsymbol{\beta}', \sigma)^'$. Let \mathbf{X}_j denote the variable that will be re-imputed and let \mathbf{X}_{-j} denote the most recently updated matrix of predictors including the auxiliary information from the matrix \mathbf{Z} , so $\mathbf{X}_{-j} = (\mathbf{Z}, \mathbf{X}_1, \dots, \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, \dots, \mathbf{X}_k)$.

Then

$$\mathbf{X}_j = \mathbf{X}_{-j}\boldsymbol{\beta}_j + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2 \mathbf{I}_n).$$

The parameters are estimated by ordinary least squares

$$\hat{\boldsymbol{\beta}}_j = (\mathbf{X}'_{-j}\mathbf{X}_{-j})^{-1}\mathbf{X}'_{-j}\mathbf{X}_j \quad (2)$$

$$\hat{\sigma}_j^2 = \frac{1}{n}(\mathbf{X}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_j)'(\mathbf{X}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_j). \quad (3)$$

Imputations are generated by drawing from the normal density using the estimated parameters. Let m_j be the number of missing item values in variable j and let \mathbf{X}_{-j}^{mis} denote the matrix of predictors corresponding to the records that have missing values in variable j , so \mathbf{X}_{-j}^{mis} is an $m_j \times (s + k - 1)$ matrix. Then

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

B. Truncated regression model

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

$$l_i^* = \max_{k=1, \dots, q} \frac{1}{b_{kj}} (\mathbf{1} - \mathbf{B}'_{-j}\mathbf{X}_{i,-j}), \text{ for } b_{kj} \neq 0$$

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

where b_{kj} is the kj th element of \mathbf{B} . The variable X_{ij} is truncated normal: $X_{ij} \sim \mathcal{N}^T(\mathbf{X}'_{i,-j}\boldsymbol{\beta}_j, \sigma_j^2)$, 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

$$f_{X_{ij}}(x_{ij} | \boldsymbol{\beta}_j, \sigma_j^2) = \frac{\psi(x_{ij} | \boldsymbol{\beta}_j, \sigma_j^2)}{\int_{l_i^*}^{u_i^*} \psi(x_{ij} | \boldsymbol{\beta}_j, \sigma_j^2) dx_{ij}} = \frac{\frac{1}{\sigma_j} \phi\left(\frac{1}{\sigma_j}(x_{ij} - \mathbf{x}'_{i,-j}\boldsymbol{\beta}_j)\right)}{\Phi(u_i^{**}) - \Phi(l_i^{**})},$$

where $\psi(\cdot)$ is the univariate normal density and $\phi(\cdot)$ and $\Phi(\cdot)$ respectively are the density and cumulative density function of the standard normal distribution. The upper and lower bounds are $u_i^{**} = \frac{1}{\sigma_j}(u_i^* - \mathbf{x}'_{i,-j}\boldsymbol{\beta}_j)$ and $l_i^{**} = \frac{1}{\sigma_j}(l_i^* - \mathbf{x}'_{i,-j}\boldsymbol{\beta}_j)$. So the loglikelihood becomes

$$\ell(\boldsymbol{\beta}_j, \sigma_j | \mathbf{x}_j) = -n \ln \sigma_j + \sum_{i=1}^n \ln \phi\left(\frac{x_{ij} - \mathbf{x}'_{i,-j}\boldsymbol{\beta}_j}{\sigma_j}\right) - n \ln (\Phi(u_i^{**}) - \Phi(l_i^{**})).$$

25. In order to find the maximum likelihood estimates we need to differentiate the loglikelihood with respect to the parameters $\boldsymbol{\beta}_j$ and σ_j :

$$\frac{\partial \ell(\boldsymbol{\beta}_j, \sigma_j | \mathbf{x}_j)}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^n \left(\frac{x_{ij} - \mathbf{x}'_{i,-j}\boldsymbol{\beta}_j}{\sigma_j^2} + \frac{1}{\sigma_j} \frac{\phi(u_i^{**}) - \phi(l_i^{**})}{\Phi(u_i^{**}) - \Phi(l_i^{**})} \right) \mathbf{x}_{i,-j}$$

$$\frac{\partial \ell(\boldsymbol{\beta}_j, \sigma_j | \mathbf{x}_j)}{\partial \sigma_j} = \frac{-n}{\sigma_j} + \sum_{i=1}^n \frac{(x_{ij} - \mathbf{x}'_{i,-j}\boldsymbol{\beta}_j)^2}{\sigma_j^3} + \frac{1}{\sigma_j} \frac{u_i^{**} \phi(u_i^{**}) - l_i^{**} \phi(l_i^{**})}{\Phi(u_i^{**}) - \Phi(l_i^{**})}.$$

As these conditions cannot be solved analytically, we need to resort to iterative techniques, such as Fisher scoring. Recall that $\boldsymbol{\zeta}'_j = (\boldsymbol{\beta}'_j, \sigma_j)$, then the maximum likelihood estimates are found by

iterating, until convergence, over

$$\boldsymbol{\zeta}_j^{(t+1)} = \boldsymbol{\zeta}_j^{(t)} + \mathbf{I}^{-1}(\boldsymbol{\zeta}_j) |_{\boldsymbol{\zeta}_j = \boldsymbol{\zeta}_j^{(t)}} \nabla_{\boldsymbol{\zeta}_j} \ell(\boldsymbol{\zeta}_j | \mathbf{x}_j) |_{\boldsymbol{\zeta}_j = \boldsymbol{\zeta}_j^{(t)}},$$

where $\nabla_{\boldsymbol{\zeta}_j} \ell_i(\boldsymbol{\zeta}_j | \mathbf{x}_j)' = (\nabla_{\boldsymbol{\beta}_j} \ell_i(\boldsymbol{\beta}_j, \sigma_j | \mathbf{x}_j)', \nabla_{\sigma_j} \ell_i(\boldsymbol{\beta}_j, \sigma_j | \mathbf{x}_j))$, $\nabla_{\boldsymbol{\zeta}_j} \ell(\boldsymbol{\zeta}_j | \mathbf{x}_j) = \sum_{i=1}^n \nabla_{\boldsymbol{\zeta}_j} \ell_i(\boldsymbol{\zeta}_j | \mathbf{x}_j)$, and $\mathbf{I}(\boldsymbol{\zeta}_j | \mathbf{x}_j) = \sum_{i=1}^n \nabla_{\boldsymbol{\zeta}_j} \ell_i(\boldsymbol{\zeta}_j) \nabla_{\boldsymbol{\zeta}_j} \ell_i(\boldsymbol{\zeta}_j | \mathbf{x}_j)'$.

26. Imputations for X_{ij}^{mis} can then be obtained by drawing from the truncated normal density:

$$X_{ij}^{mis} \sim \mathcal{N}^T(\mathbf{X}_{i,-j} \hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2), \quad l_i^* \leq X_{ij}^{mis} \leq u_i^*.$$

C. Logistic regression model

27. 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.

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

$$\text{logit } p_{ij} = \boldsymbol{\delta}' \mathbf{x}_{i,-j}.$$

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 likelihood function becomes

$$L(\boldsymbol{\delta} | \mathbf{y}_j) = \prod_{i=1}^n p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}} = \prod_{i=1}^n \Lambda(\boldsymbol{\delta}' \mathbf{x}_{i,-j})^{y_{ij}} (1 - \Lambda(\boldsymbol{\delta}' \mathbf{x}_{i,-j}))^{1-y_{ij}}$$

and the loglikelihood is

$$\begin{aligned} \ell(\boldsymbol{\delta} | \mathbf{y}_j) &= \sum_{i=1}^n \left(y_{ij} \ln \Lambda(\boldsymbol{\delta}' \mathbf{x}_{i,-j}) + (1 - y_{ij}) \ln(1 - \Lambda(\boldsymbol{\delta}' \mathbf{x}_{i,-j})) \right) \\ &= \sum_{i=1}^n \left(y_{ij} \boldsymbol{\delta}' \mathbf{x}_{i,-j} - \ln(1 + \exp\{\boldsymbol{\delta}' \mathbf{x}_{i,-j}\}) \right). \end{aligned}$$

So, the first order conditions are

$$\frac{\partial \ell(\boldsymbol{\delta} | \mathbf{y}_j)}{\partial \boldsymbol{\delta}} = \sum_{i=1}^n (y_{ij} - \Lambda(\boldsymbol{\delta}' \mathbf{x}_{i,-j})) \mathbf{x}_{i,-j} = \mathbf{X}'_{-j} (\mathbf{y}_j - \boldsymbol{\Lambda}(\mathbf{X}_{-j} \boldsymbol{\delta})),$$

where $\boldsymbol{\Lambda}(\mathbf{X}_{-j} \boldsymbol{\delta})$ is a vector of order $n \times 1$ with elements $\Lambda(\boldsymbol{\delta}' \mathbf{x}_{i,-j})$, $i = 1, \dots, n$. The parameter estimates can be found using Newton-Raphson, as the Hessian can be straightforwardly derived:

$$\frac{\partial^2 \ell(\boldsymbol{\delta} | \mathbf{y}_j)}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}'} = - \sum_{i=1}^n \Lambda(\boldsymbol{\delta}' \mathbf{x}_{i,-j}) (1 - \Lambda(\boldsymbol{\delta}' \mathbf{x}_{i,-j})) \mathbf{x}_{i,-j} \mathbf{x}'_{i,-j} = - \mathbf{X}'_{-j} \mathbf{W} \mathbf{X}_{-j},$$

where $\mathbf{W} = \text{diag}\{\boldsymbol{\Lambda}(\mathbf{X}_{-j}\boldsymbol{\delta})(\iota_n - \boldsymbol{\Lambda}(\mathbf{X}_{-j}\boldsymbol{\delta}))'\}$. 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.

29. 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 $Y_{ij} = 1$.

VII. INCORPORATION OF LINEAR BALANCE RESTRICTIONS

30. An important disadvantage of the univariate 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. If the dataset was completely observed, this problem would be solved by leaving out the variables that are present in the balance restrictions such that the data does not contain singularities. Next these variables can be simulated using a Gibbs sampler and the variables that are left out can subsequently be derived from these simulated values. In the presence of nonresponse this is not straightforwardly applicable as the missingness is scattered across variables, which means that the information in the items values will not be redundant anymore as some of the variables in that balance restriction contain missing values. Only if a variable present in a balance restriction is completely missing, this variable can be safely removed from the analysis. Clearly, this will be highly unlikely.

31. As the missingness varies across records, the set of missing variables that can be left out differs across records as well. For each record i partition $\mathbf{X}'_i = (\mathbf{X}'_{i,mis} \ \mathbf{X}'_{i,obs})$ and partition \mathbf{A} and \mathbf{B} accordingly. So for this record the following restrictions hold: $\mathbf{A}_{i,mis}\mathbf{X}_{i,mis} = -\mathbf{A}_{i,obs}\mathbf{X}_{i,obs}$ and $\mathbf{1} - \mathbf{B}_{i,obs}\mathbf{X}_{i,obs} \leq \mathbf{B}_{i,mis}\mathbf{X}_{i,mis} \leq \mathbf{u} - \mathbf{B}_{i,obs}\mathbf{X}_{i,obs}$. Now $\text{rank}(\mathbf{A}_{i,mis}) = p_i$ is the number of variables in $\mathbf{X}_{i,mis}$ that are redundant and therefore can be left out. Let $\mathbf{A}_{i,mis}^n$ be the $p_i \times m_i$ matrix consisting of the nonredundant rows in $\mathbf{A}_{i,mis}$. Then any nonsingular $p_i \times p_i$ submatrix of $\mathbf{A}_{i,mis}^n$ corresponds to the p_i missing variables that are redundant and therefore can be left out. Now partition $\mathbf{A}_{i,mis}^n = (\mathbf{A}_{i,mis,p_i}^n \ \mathbf{A}_{i,mis,q_i}^n)$, where \mathbf{A}_{i,mis,p_i}^n refers to the variables that are left out and partition $\mathbf{X}_{i,mis}$ accordingly. Then rewrite the redundant variables in terms of the nonredundant variables

$$\mathbf{X}_{i,mis,p_i} = (\mathbf{A}_{i,mis,p_i})^{-1}(-\mathbf{A}_{i,obs}\mathbf{X}_{i,obs} - \mathbf{A}_{i,mis,q_i}\mathbf{X}_{i,mis,q_i}) \quad (4)$$

and plug these values into the linear inequality restrictions. This means that the remaining variables can be modelled using this set of linear inequality restrictions.

32. The algorithm will be as follows. At iteration t , determine $\mathbf{A}_{i,mis}^n$ for each record i and choose the appropriate \mathbf{A}_{i,mis,p_i} and \mathbf{A}_{i,mis,q_i} and partition $\mathbf{X}_{i,mis}^{(t)'} = (\mathbf{X}_{i,mis,p_i}^{(t)'} \ \mathbf{X}_{i,mis,q_i}^{(t)'})$ accordingly, where $\mathbf{X}_{i,mis,p_i}^{(t)'}$ refers to the redundant elements in $\mathbf{X}_{i,mis}^{(t)'}$. Next start the algorithm at the first variable and calculate regression parameters for this variable if it is present in \mathbf{X}_{i,mis,q_i} conditioned on the other variables in \mathbf{X}_{i,mis,q_i} . 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,mis,q_i} , are

re-imputed with the appropriate regression model and accompanying parameter estimates. Once one variable is re-imputed, the redundant variables in \mathbf{X}_{i,mis,p_i} are updated using (4) in order to make sure that the balance restrictions will hold throughout the algorithm. Now proceed to the next variable and repeat this process until all k variables are treated. After this the missing items present in \mathbf{X}_{i,mis,p_i} are re-imputed using (4), which results in $\mathbf{X}_{i,mis,p_i}^{(t+1)}$.

33. Unfortunately, leaving out data possibly results in a loss of information and consequently the quality of the imputed values of \mathbf{X}_{i,mis,p_i} may be reduced. To overcome this issue, we suggest choosing the variables that are left out for each record randomly at each iteration, as there are often several possible choices for \mathbf{A}_{i,mis,p_i}^n .

VIII. DISCUSSION

34. In this paper we discuss a method that can be used for the imputation of data subject to several constraints without specifying a joint model. Preliminary analysis shows that this method provides promising results with respect to the economic data gathered by Statistics Netherlands. The distributions of the imputed data are relatively well preserved using this method, especially with respect to semi-continuous variables. For these variables it is quite difficult to establish a joint model and this method can therefore be useful if the dataset consists of many semi-continuous variables. Further research should be done in order to establish the performance of this method with respect to the imputation methods based on joint models. Furthermore, the issue of convergence has not been treated in this paper. Preliminary tests have not indicated difficulties with convergence, but more research needs to be done in this area.

References

- [1] Amemiya, T. (1973), "Regression Analysis When The Dependent Variable Is Truncated Normal," *Econometrica*, 41, 997 -1016.
- [2] Arnold, B.C. and Press, S.J. (1989), "Compatible Conditional Distributions," *Journal of the American Statistical Association*, 84, 152-156.
- [3] Box, G.E.P. and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, Series B*, 26, 211-246.
- [4] Gelman, A. and Raghunathan, T.E. (2001), Discussion of "Conditionally Specified Distributions," by Arnold, B.C. et al, *Statistical Science*, 16, 268-269.
- [5] Heeringa, S.G., Little, R.J.A. and Raghunathan, T.E. (2002), "Multivariate Imputation of Coarsened Survey Data on Household Wealth," In *Survey Nonresponse*, eds Groves R.M. et al, New York: Wiley.
- [6] Little R.J.A and Raghunathan (1997), "Should Imputation of Missing Data Condition on All Observed Variables?," *Proceedings of the Survey Research Methods Section, American Statistical Association*, 617-622.
- [7] Poirier, D.J. (1978), "The Use of the Box-Cox Transformation in Limited Dependent Variable Models," *Journal of the American Statistical Association*, 73, 284-287.
- [8] Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using A Sequence of Regression Models," *Survey Methodology*, 27, 85-95.
- [9] Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

- [10] Spitzer, J.J. (1982), “A Primer on Box-Cox Estimation,” *Review of Economics and Statistics*, 64, 307-313.
- [11] Van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M. and Rubin, D.B. (2006), “Fully Conditional Specification in Multivariate Imputation,” *Journal of Statistical Computation and Simulation*, forthcoming.