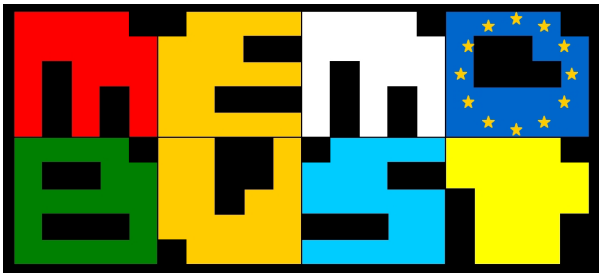

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Imputation under Edit Constraints

Contents

General section

1. Summary

In the context of business surveys at National Statistical Institutes (NSIs), imputation of missing values is often complicated by the fact that the data should conform to a large number of edit rules. In this module, we consider two basic approaches to obtain imputations that satisfy edit rules. Under the first approach, the edits are incorporated directly in the imputation model, so that all imputations are automatically consistent. Unfortunately, this can lead to a very complex model. Therefore, in practice, another approach is often used, in which the missing values are first imputed without taking the edits into account. In a subsequent step, the initial imputations are then minimally adjusted to become consistent with the edits.

2. General description

2.1 Introduction

In the context of business surveys at NSIs, the imputation of missing values is often complicated by the fact that the data should conform to a large number of restrictions, known as *edit rules*, *edit constraints*, or *edits* (see also "Statistical Data Editing – Main Module"). For instance, if a survey includes the variables *turnover*, *costs*, and *profit*, then the edit rule

profit = *turnover* – *costs*

is supposed to hold for the corresponding values. In addition, there are edits stating that the values of *turnover* and *costs* should be non-negative. It is desirable to avoid imputations that are inconsistent with the edit rules, because data with obvious inconsistencies are likely to be rejected by most users, even if they could in fact be used to make valid statistical inferences (Pannekoek and De Waal, 2005). Särndal and Lundström (2005, p. 176) wrote: "Whatever the imputation method used, the completed data set should be subjected to the usual checks for internal consistency. All imputed values should undergo the editing checks normally carried out for the survey."

Obviously, if a standard imputation method such as regression imputation (see "Imputation – Model-Based Imputation") or random hot deck imputation (see "Imputation – Donor Imputation") is applied without taking the edit rules into account, then one should generally not expect the resulting imputations to satisfy the edits. Unfortunately, taking edit rules into account directly in the imputations tends to introduce complications. De Waal et al. (2011) give the following simple example. Suppose that we are given a record with missing values on the variables x and y, and suppose that the following edit rules have been defined for these variables:

$$x \geq 50; \tag{1}$$

$$y \leq 100; \tag{2}$$

$$y \geq x. \tag{3}$$

If we first impute x, the only edit which can be evaluated at this stage is (1). Taking this edit into account, we might impute the value $\tilde{x} = 150$. The resulting edit rules for y given by (2) and (3) cannot be satisfied simultaneously: $y \leq 100$ and $y \geq 150$. Furthermore, if we start by imputing y,

taking edit (2) into account, we might impute the value $\tilde{y} = 40$ and encounter a similar problem with the resulting edit rules for x. Thus, consistency with the edit rules is not guaranteed under this sequential procedure. The point is that if the variables are imputed sequentially, in general, edit rules involving variables that will be imputed later cannot be ignored.

There are two general approaches to imputation under edit constraints. The first approach is to, somehow, include the edit rules in the (implicit or explicit) model used for imputation, so that the imputed values automatically satisfy all constraints. The second approach is to apply a two-step procedure. In the first step, the missing values are imputed without taking (all) constraints into account. In the second step, the initially imputed values are minimally adjusted to satisfy all edits. These two approaches will be discussed further in Sections 2.2 and 2.3, respectively. Finally, it should be noted that values derived by deductive imputation methods (see "Imputation – Deductive Imputation") trivially satisfy the edits that were used in the derivation. We will return to this point in Section 2.3.

2.2 Imputation under edit constraints by direct modeling

2.2.1 Ratio hot deck imputation

In general, imputation methods that take edit constraints into account directly tend to be complex. One exception is the ratio hot deck method. This is an extension of the ordinary hot deck donor imputation method (see "Imputation – Donor Imputation") that is appropriate to impute missing values among a set of non-negative variables $y_1, \ldots, y_m$ that should satisfy a linear balance edit of the form:

$$y_1 + \cdots + y_m = y_{tot}, \tag{4}$$

where it is assumed that the total value y_{tot} is always observed (or previously imputed). Basically, instead of imputing the donor values directly, we use the donor to distribute the total missing amount over the missing variables.

Consider the i^{th} record that requires imputation and suppose for notational convenience that the first t variables are observed (with values $y_{i,1}, \ldots, y_{i,t}$) and the last $m - t$ values are missing. We first compute the total missing amount, $r_i = y_{i,tot} - y_{i,1} - \cdots - y_{i,t}$. Next, using any of the ordinary donor imputation methods, we choose a donor from the completely observed records. The donor record should be consistent with the edits. We compute the sum of the donor values of the variables to impute, say, $r_d = y_{d,t+1} + \cdots + y_{d,m}$. The ratio hot deck imputations are given by:

$$\tilde{y}_{i,j} = \frac{r_i}{r_d} y_{d,j}, \quad (j = t+1, \ldots, m).$$

By construction, the imputed values are non-negative and consistent with edit (4):

$$y_{i,1} + \cdots + y_{i,t} + \tilde{y}_{i,t+1} + \cdots + \tilde{y}_{i,m} = y_{i,1} + \cdots + y_{i,t} + \frac{r_i}{r_d} r_d = y_{i,tot}.$$

For an application of the ratio hot deck method in practice, see Pannekoek and Van Veller (2004) or Pannekoek and De Waal (2005). A straightforward generalisation of the method can be applied if the

balance edit contains coefficients unequal to 1 (De Waal et al., 2011). Unfortunately, the method cannot be used to obtain consistent imputations if there are multiple, inter-related restrictions.

2.2.2 Parametric imputation models

To introduce the direct modeling of edit constraints in a parametric model, it is useful to consider a small univariate example. Suppose that a certain variable y is to be imputed using the normal distribution $N(\mu,\sigma^2)$, and suppose in addition that we require the imputations to be non-negative; i.e., the edit rule $y \geq 0$ should hold. To make the example interesting, consider the case that μ and σ are such that the distribution $N(\mu,\sigma^2)$ has a significant probability of generating negative values (e.g., $\mu = 1$ and $\sigma = 2$). The edit would be failed quite often if we imputed values directly from $N(\mu,\sigma^2)$. An intuitively sensible approach to obtain consistent imputations in this case works as follows: obtain a random draw z from $N(\mu,\sigma^2)$. If it holds that $z \geq 0$, then impute $\tilde{y} = z$. Otherwise, repeat the procedure until a draw with $z \geq 0$ is obtained. By construction, all resulting imputations will satisfy the non-negativity edit. Technically, these imputations follow a so-called *truncated* normal distribution (Geweke, 1991).[1] The above iterative procedure for obtaining values from this distribution is known as *Acceptance/Rejection sampling* (Tempelman, 2007).

The univariate truncated normal distribution is a relatively simple example of a model that incorporates constraints on the modeled variables (in this case: one inequality constraint and one variable). The general idea of imputation under edit constraints by direct modeling is to find a model that incorporates all the relevant constraints on the variables to impute. The main advantage of this approach is that it avoids having to adjust the imputations later on to satisfy the edit rules. Two important disadvantages of the direct modeling approach are: (*i*) in most practical applications, the resulting imputation methods are mathematically complex and require heavy computational work; and (*ii*) as this methodology is relatively new, only a limited number of models have been developed.

Tempelman (2007) developed imputation models that can incorporate particular types of constraints:

- If all edits are linear inequalities (i.e., the restrictions can be written as $\mathbf{Qy} \geq \mathbf{b}$ for a given matrix $\mathbf{Q}$ and vector $\mathbf{b}$ of constants), then the multivariate truncated normal distribution can be used. The distribution is truncated to the region defined by the constraints $\mathbf{Qy} \geq \mathbf{b}$. This is a multivariate extension of the univariate example given above.
- If all edits are linear equalities (i.e., the restrictions can be written as $\mathbf{Ry} = \mathbf{a}$ for a given matrix $\mathbf{R}$ and vector $\mathbf{a}$ of constants), then the multivariate singular normal distribution can be used. This is a generalisation of the ordinary multivariate normal distribution $N_d(\boldsymbol{\mu},\boldsymbol{\Sigma})$ for the

[1] In general, a random variable with density function $f(x|\theta)$ can be truncated to any subdomain G of its original support by defining the truncated density function:

$$f(x|\theta;G) = \begin{cases} \dfrac{f(x|\theta)}{\int_G f(x|\theta)dx} & \text{if } x \in G \\ 0 & \text{if } x \notin G \end{cases}$$

case that the covariance matrix $\boldsymbol{\Sigma}$ is singular (Khatri, 1968). In fact, the covariance matrix is singular here because the constraints $\mathbf{Ry} = \mathbf{a}$ induce a linear dependence in this matrix.

- If both linear equalities and inequalities occur, then the multivariate truncated singular normal distribution can be used. This distribution combines the features of the two previous cases.
- For the special case of one linear equality with non-negativity edits for all variables involved – i.e., the case that can be handled by the ratio hot deck method –, an alternative model is given by the Dirichlet distribution (Wilks, 1962).

A full treatment of these models is beyond the scope of this module. We refer to Tempelman (2007) and De Waal et al. (2011, Ch. 9) for more details. An important theoretical limitation of the first three models is that they are only appropriate for data that are approximately normally distributed. Moreover, it is not useful here to apply a standard (non-linear) transformation to the data to obtain a closer resemblance to a normal distribution, because the edits for the transformed data would not have the linear structure ($\mathbf{Qy} \geq \mathbf{b}$ and/or $\mathbf{Ry} = \mathbf{a}$) of the original edits.

2.2.3 The elimination approach

In the above approaches, a joint model is used to impute all variables with missing values in a record at once. A somewhat different, less complex approach was proposed by Coutinho et al. (2007). They used a technique called *Fourier-Motzkin elimination* (Williams, 1986; De Waal et al., 2011) to reduce the problem of consistent imputation to a sequence of univariate problems. This elimination technique is used more traditionally in algorithms for automatic error localisation. We refer to the module "Statistical Data Editing – Automatic Editing" for a brief description of Fourier-Motzkin elimination.

A full discussion of the elimination approach is beyond the scope of this module. Here, we will only give a small example. Consider again the example from Section 2.1, where the objective is to impute the variables x and y in such a way that the edits (1), (2), and (3) are satisfied. Before we can start imputing, we have to posit and estimate a joint model for the data. In contrast to Section 2.2.2, this model need not incorporate the edit constraints, which makes the modeling task much easier. Following Coutinho et al. (2007), we will use an ordinary bivariate normal distribution in this example for simplicity:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N_2\left(\begin{pmatrix} 60 \\ 55 \end{pmatrix}, \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}\right). \tag{5}$$

We begin by applying Fourier-Motzkin elimination to the original edits (1)–(3) to eliminate x from these edits. In this particular example, this yields two implied edits for the remaining variable y:

$$y \geq 50; \tag{6}$$

$$y \leq 100. \tag{7}$$

We would now like to impute y from its posited $N(55,100)$ distribution, in such a way that the imputed value satisfies the inequalities (6) and (7). This can be achieved, as in the example from Section 2.2.2, by drawing from a truncated normal distribution by means of Acceptance/Rejection

sampling. That is, we draw random values from the $N(55,100)$ distribution until we obtain a value that lies between 50 and 100. Suppose that we obtain the value $\tilde{y} = 70$.

In the next step, we substitute the imputed value $\tilde{y} = 70$ for y in the original edits (1)–(3). This yields two reduced edit rules that involve only x:

$$x \geq 50; \tag{8}$$

$$70 \geq x. \tag{9}$$

Finally, x is imputed by drawing from the posited $N(60,100)$ distribution until we obtain a value that complies with edits (8) and (9). (In general, we would use the conditional distribution of x, given the previously imputed value for y, but the two variables are uncorrelated in this example.) This might yield the value $\tilde{x} = 52$. In this manner, we obtain the imputed record $(\tilde{x}, \tilde{y}) = (52,70)$ which is consistent with the original edits (1)–(3).

By a fundamental property of the Fourier-Motzkin elimination technique, the above method always yields imputations that are consistent with the edit rules (Coutinho et al., 2007). Note that according to model (5), the mean of x is larger than the mean of y. In this sense, the posited model does not comply with edit rule (3). Nevertheless, the elimination approach yields consistent imputations, as was illustrated by the example. However, it should be noted that if the model strongly disagrees with the edit rules, the procedure of Acceptance/Rejection sampling from a truncated distribution may become very inefficient. In fact, an appropriate model for the data should not strongly disagree with the edit rules, provided that these rules are substantively meaningful.

For a general description of the elimination approach to consistent imputation, we refer to Coutinho et al. (2007) and De Waal et al. (2011, Ch. 9). Extensions of this method have been considered by Pannekoek et al. (2008, 2013) and Coutinho et al. (2013).

2.3 Imputation under edit constraints by adjustment methods

Since most of the methods discussed in Section 2.2 have limited practical applicability, a less complex approach is often applied in practice. Under this approach, the variables with missing values are first imputed by any method that produces a complete data set with good statistical properties, without taking (all) edit constraints into account. That is to say, any appropriate method discussed in the other modules on imputation can be used. Denote the initial imputed record by $\hat{\mathbf{y}}$. Next, an adjusted imputed record $\tilde{\mathbf{y}}$ is obtained from $\hat{\mathbf{y}}$ as the solution to a constrained minimisation problem:

$$\begin{aligned} &\text{Minimise } D(\hat{\mathbf{y}}, \tilde{\mathbf{y}}), \\ &\text{so that } \tilde{\mathbf{y}} \text{ satisfies all edit constraints.} \end{aligned} \tag{10}$$

Here, D is a function that measures the distance between the initial imputed record $\hat{\mathbf{y}}$ and the adjusted record $\tilde{\mathbf{y}}$. It is customary to demand that only the imputed values may be adjusted under this minimisation problem, i.e., the variables that were originally observed retain their original values.

Adjusting the imputed values for consistency with the edit constraints is a special case of the general problem of *data reconciliation*. Methods for this more general problem are treated in "Micro-Fusion – Reconciling Conflicting Microdata" and in particular the underlying method module "Micro-Fusion –

Minimum Adjustment Methods". The reader is referred to these modules and to De Waal et al. (2011, Ch. 10) for more details.

In the special case that all edits are linear equalities (written as a linear system of the form $\mathbf{Ry} = \mathbf{a}$), one could also apply the methodology discussed in the module "Imputation – Deductive Imputation" to obtain a consistent record in the second step above. Suppose that the initial imputed record $\hat{\mathbf{y}}$ is partitioned as $\hat{\mathbf{y}} = (\hat{\mathbf{y}}_o', \hat{\mathbf{y}}_m')'$ and the imputed values in $\hat{\mathbf{y}}_m$ are suppressed (i.e., replaced by missing values). The matrix $\mathbf{R}$ is partitioned accordingly as $\mathbf{R} = [\mathbf{R}_o \quad \mathbf{R}_m]$. If $\mathbf{R}_m$ has full rank, it follows that the missing values are imputed consistently by $\tilde{\mathbf{y}}_m = \mathbf{R}_m^{-1}(\mathbf{a} - \mathbf{R}_o\hat{\mathbf{y}}_o)$; see "Imputation – Deductive Imputation" for more details. Thus, we should choose $\hat{\mathbf{y}}_m$ in such a way that $\mathbf{R}_m$ has full rank.[2] Since this choice is not unique in practice, we may randomly vary the selection of $\hat{\mathbf{y}}_m$ for each imputed record; thereby, we avoid the introduction of a systematic effect in some variables. The resulting approach may be seen as a heuristic approximation to minimisation problem (10). However, if appropriate software is available, finding the optimal solution to (10) directly should be relatively straightforward and there is little to be gained from a heuristic approach.

3. Design issues

4. Available software tools

There are no generally available tools that have the imputation methods described in this module as standard functionality. Some NSIs have developed dedicated tools for particular applications. On the other hand, the methods are relatively easy to implement in statistical computing environments such as R and SAS, using the existing functionality available in these environments. Some standard tools do exist for solving problem (10) in the adjustment step of Section 2.3; e.g., the R package `rspa`, as well as commercial solvers such as CPLEX and Xpress.

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Coutinho, W., de Waal, T., and Remmerswaal, M. (2007), Imputation of Numerical Data under Linear Edit Restrictions. Discussion Paper 07012, Statistics Netherlands, The Hague.

[2] It seems undesirable to suppress and impute values that were originally observed. To avoid this, one should restrict the system $\mathbf{Ry} = \mathbf{a}$ to those edits that involve at least one imputed value (the other edits should already be satisfied by the observed values). The partitioning can and should then be made in such a way that $\hat{\mathbf{y}}_m$ contains only variables that were initially imputed.

Coutinho, W., de Waal, T., and Shlomo, N. (2013), Calibrated Hot Deck Imputation Subject to Edit Restrictions. *Journal of Official Statistics* **29**, 299–321.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.

Geweke, J. (1991), Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities. Report, University of Minnesota.

Khatri, C. G. (1968), Some Results for the Singular Normal Multivariate Regression Models. *Sankhyā Series A* **30**, 267–280.

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

Pannekoek, J. and van Veller, M. G. P. (2004), Regression and Hot-Deck Imputation Strategies for Continuous and Semi-Continuous Variables. In: J. R. H. Charlton (ed.), *Methods and Experimental Results from the EUREDIT Project*. (http://www.cs.york.ac.uk/euredit/)

Pannekoek, J., Shlomo, N., and de Waal, T. (2008), Calibrated Imputation of Numerical Data under Linear Edit Restrictions. Working Paper, UN/ECE Work Session on Statistical Data Editing, Vienna.

Pannekoek, J., Shlomo, N., and de Waal, T. (2013), Calibrated Imputation of Numerical Data under Linear Edit Restrictions. *Annals of Applied Statistics* **7**, 1983–2006.

Särndal, C.-E. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Chichester.

Tempelman, D. C. G. (2007), *Imputation of Restricted Data*. PhD Thesis, University of Groningen.

Wilks, S. S. (1962), *Mathematical Statistics*. John Wiley & Sons, New York.

Williams, H. P. (1986), Fourier's Method of Linear Programming and Its Dual. *The American Mathematical Monthly* **93**, 681–695.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Data Editing – Main Module
2. Imputation – Main Module
3. Imputation – Model-Based Imputation
4. Imputation – Donor Imputation

9. Methods explicitly referred to in this module

1. Micro-Fusion – Reconciling Conflicting Microdata
2. Micro-Fusion – Minimum Adjustment Methods
3. Statistical Data Editing – Automatic Editing
4. Imputation – Deductive Imputation

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. GSBPM Sub-process 5.4: Impute

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1. Imputation, i.e., determining and filling in new values for occurrences of missing or discarded values in a data file

Administrative section

14. Module code

Imputation-T-Imputation under Edit Constraints

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	05-03-2013	first version	Sander Scholtus	CBS (Netherlands)
0.2	10-07-2013	improvements based on Norwegian review	Sander Scholtus	CBS (Netherlands)
0.2.1	19-11-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:17