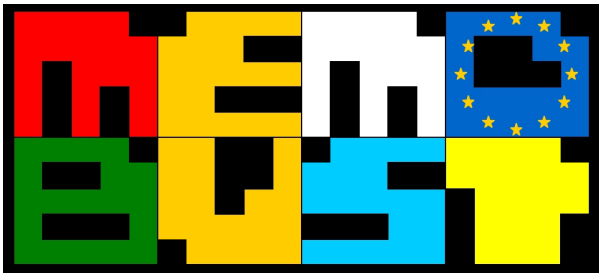

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Donor Imputation

Contents

General section

1. Summary

The objective in donor imputation is to fill in the missing values for a given unit by copying observed values of another unit, the donor. Typically, the donor is chosen in such a way that it resembles the imputed unit as much as possible on one or more background characteristics. The rationale behind this is that if the two units match (exactly or approximately) on a number of relevant auxiliary variables, it is likely that their scores on the target variable will also be similar.

2. General description[1]

2.1 Introduction to donor imputation

The objective in donor imputation is to fill in the missing values for a given unit (the *recipient*) by copying the corresponding observed values of another unit (the *donor*). The term *hot deck* donor imputation applies when the donor comes from the same data set as the recipient. In the context of business statistics, this is the most commonly encountered form of donor imputation. If the donor is taken from another data set, this is known as *cold deck* donor imputation. Most applications of cold deck imputation use data that were collected at a previous point in time. Often, the donor record is then simply an earlier observation of the recipient unit itself. This type of donor imputation is only valid for variables that can be considered more or less constant between observation times; its applicability in the context of business statistics is therefore limited. In the remainder of this module, we shall focus on hot deck imputation.

Letting y_i denote the score of the i^{th} unit on the target variable y and using the index d for a donor, we can write the generic formula for hot deck donor imputation as:

$$\tilde{y}_i = y_d \,. \tag{1}$$

Typically, one searches for a donor that resembles the recipient as much as possible on one or more auxiliary variables. There exist different ways to select a donor, leading to different variants of hot deck imputation. In this module, we shall describe *random* and *sequential hot deck imputation* (Section 2.2), *nearest-neighbour imputation* (Section 2.3), and *predictive mean matching* (Section 2.4). Some practical issues are discussed in Section 2.5.

In formula (1) and in the description below, we focus on imputing one target variable at a time. In practice, one often encounters records with several missing values. In that case, the standard approach is to impute all missing values in a record from the same donor. This helps to preserve the multivariate relations between the imputed variables. In fact, an important practical advantage of donor imputation compared to model-based imputation is that it can be extended to multivariate imputation in this natural way.

[1] This section is to a large extent based on Chapter 6 of Israëls et al. (2011).

2.2 *Random and sequential hot deck imputation*

In random hot deck imputation, imputation classes are formed based on categorical auxiliary variables. For each recipient unit i in a given imputation class, the group of potential donors consists of the units within the same class with y observed. Of these potential donors, one is selected at random – typically through equal-probability sampling – and used to impute the recipient. Note that this procedure implies that the donor and the recipient have exactly the same values on all auxiliary variables that are used to define the imputation classes. Conditional on these auxiliary variables, the donor is selected completely at random.

Sequential hot deck imputation also requires that the donor and the recipient have identical values on the auxiliary variables, but here the data set is not explicitly split into groups. Instead, one goes over the records in the data set in order and imputes each missing value by the last previously encountered observed value for a unit with the same scores on the auxiliary variables. Thus, the recipient is imputed using as a donor the last unit with y observed that belongs to the same imputation class and that comes before the recipient in the data file. Historically, the sequential hot deck method had the advantage that it can be carried out by a computer in a very efficient manner. The algorithm requires just one pass over the data set (Kalton and Kasprzyk, 1986). With the rise of computing power, this is no longer considered a real advantage for most practical applications.

For the sequential hot deck method, the imputations obviously depend on the order of the records in the data set. The method can be applied after a random sorting of the records; this yields stochastic imputations and is sometimes called 'random sequential hot deck'. Alternatively, deterministic imputations may be obtained by sorting the records on one or more background characteristics. Either way, it is recommended to perform some form of explicit sorting before applying this method, because otherwise the results may be biased due to an implicit and unforeseen ordering of the units in the file.

Typically, the standard errors of means and totals of y will be inflated by random (sequential) hot deck imputation (Little and Rubin, 2002). In part, this may be due to the risk of outliers being 'magnified', which can be avoided by excluding outliers from the group of potential donors. More generally, it is desirable to avoid that the same unit can be used as a donor for many different recipients. In random hot deck imputation, this can be achieved by using a more elaborate selection mechanism, so that a repeated use of the same donor is only allowed once all or most of the potential donors within an imputation class have had a turn. In sequential hot deck imputation, a repeated use of the same donor may occur whenever there are several item non-respondents close together in the data file. One way to prevent this is to consider an extension of sequential hot deck imputation. Under this extension, one stores the last K observed values within an imputation class (for some $K>1$). Whenever an item non-respondent is encountered, it is imputed by choosing at random one of the K potential donor values.

2.3 *Nearest-neighbour imputation*

In nearest-neighbour imputation, we drop the restriction that the donor and the recipient have identical scores on all auxiliary variables. Instead, the auxiliary variables are used to define a distance function $D(i,k)$ between units i and k, where i is the recipient and k is a potential donor. The *nearest neighbour* of unit i is defined as the respondent d that minimises this distance function. Formally,

$$d = \arg\min_{k \in obs} D(i,k), \tag{2}$$

where obs denotes the set of units with y observed, i.e., the set of potential donors.

Before going into the imputation method itself, we will briefly discuss possible choices of the distance function in formula (2). Assuming for now that the auxiliary variables $(x_1,\ldots,x_q)$ are all quantitative (but see Section 2.5), a frequently used family of distance functions is given by:

$$D_z(i,k) = \left(\sum_{j=1}^{q} \left| x_{ji} - x_{jk} \right|^z \right)^{1/z} \tag{3}$$

with $z > 0$. For $z = 2$, formula (3) yields the well-known Euclidean distance. For $z = 1$, it is just the sum of the absolute differences $|x_{ji} - x_{jk}|$; this is sometimes called the 'city-block' or 'Manhattan' distance. As z becomes larger, formula (3) places a higher penalty on large differences for individual auxiliary variables. In fact, by letting z tend to infinity in (3), we obtain the so-called 'minimax' distance given by

$$D_\infty(i,k) = \max_{j=1,\ldots,q} \left| x_{ji} - x_{jk} \right|. \tag{4}$$

According to distance (4), the nearest neighbour should not deviate strongly from the recipient on any auxiliary variable x_j. Practical applications of nearest-neighbour imputation that involve distance function (3) with choices other than $z = 1$, $z = 2$, or $z \to \infty$ are rare.

A generalisation of (3) is obtained by including weight factors γ_j that express the importance of each auxiliary variable for the purpose of finding accurate imputations:

$$D_{z,\gamma}(i,k) = \left(\sum_{j=1}^{q} \gamma_j \left| x_{ji} - x_{jk} \right|^z \right)^{1/z}. \tag{5}$$

In addition, note that the contributions of the auxiliary variables to (3) or (5) are implicitly weighted if these variables are measured on different scales. For instance, if x_1 represents last year's turnover in Euros and x_2 represents the number of employees, then the value of $D_1(i,k) = |x_{1i} - x_{1k}| + |x_{2i} - x_{2k}|$ will depend almost exclusively on the first term in practice. To prevent this, one should first standardise the auxiliary variables so that their variances are equal to 1. Alternatively, the so-called Mahalanobis distance could be used which also takes correlations between variables into account (see, e.g., Little and Rubin, 2002); this can be seen as a generalisation of the Euclidean distance $D_2(i,k)$.

In its basic form, the nearest-neighbour method imputes an item non-respondent by using its nearest neighbour as donor. This yields a deterministic imputation. As before, the underlying idea is that two units that are closely matched on relevant background characteristics [i.e., for which $D(i,k)$ has a small value] are likely to also have a similar score on the target variable.

A stochastic generalisation of nearest-neighbour imputation first selects the K units that are closest to unit i in terms of $D(i,k)$ – i.e., the K nearest neighbours – as potential donors and then draws one of these units at random. In some applications, unequal drawing probabilities are assigned to the K nearest neighbours so that within this group the units with smaller values of $D(i,k)$ are more likely to

be selected as donor. Following Bankier et al. (2000), an appropriate choice of drawing probability for the k^{th} potential donor is then given by:

$$p(k) \propto \left(\frac{D_{\min}}{D(i,k)} \right)^t, \quad (k = 1, \ldots, K), \tag{6}$$

where $D_{\min} = \min_{k \in obs} D(i,k)$ denotes the distance of the nearest neighbour and $t \geq 0$ is a parameter determining the selection mechanism. Equal-probability selection is obtained as a special case of (6) with $t = 0$. The method coincides with ordinary deterministic nearest-neighbour imputation in the limit $t \to \infty$.

2.4 Predictive mean matching

Little (1988) described a variant of donor imputation known as predictive mean matching. In this imputation method, a linear regression is first performed of the target variable y on some auxiliary variables $x_1, \ldots, x_q$. The regression model is fitted on the data of units without item non-response. Next, the resulting regression equation is used to obtain predicted values $\hat{y}$ for all records, in accordance with formula (4) in the module "Imputation – Model-Based Imputation". For item non-respondent i with predicted value $\hat{y}_i$, we select as donor the item respondent d for which the predicted value $\hat{y}_d$ is as close as possible to $\hat{y}_i$. Finally, the *observed* value y_d of the donor is imputed, in accordance with formula (1) above. The latter feature makes this method a form of donor imputation rather than model-based imputation.

It should be noted that predictive mean matching is actually a special case of nearest-neighbour imputation. This is easily seen by considering the distance function

$$D_{pmm}(i,k) = \left| \hat{y}_i - \hat{y}_k \right|$$

and choosing the donor according to formula (2). Alternatively, this distance function can be expressed as a weighted sum of differences between the auxiliary variables used in the regression (De Waal et al., 2011, p. 253).

2.5 Practical issues

Random and sequential hot deck imputation require that the auxiliary variables are categorical, because these variables are used to construct imputation classes. Quantitative auxiliary variables can be included by first deriving 'categorised' versions of them (e.g., a size class variable based on the number of employees).

Nearest-neighbour imputation is used mainly with quantitative auxiliary variables. It is also possible to include categorical auxiliary variables, but this requires an appropriate extension of the distance function. One way to do this is to assign, for each categorical variable separately, a distance to each possible pair of values. For an auxiliary variable x_j with m categories, this 'local' distance function can be summarised in the form of an $m \times m$ matrix A_j. Next, we can define a 'global' distance function of the form (3) or (5), by replacing the absolute difference $|x_{ji} - x_{jk}|$ by the value

$A_j(x_{ji}, x_{jk})$ in these expressions. Similarly, a combination of quantitative and qualitative auxiliary variables can also be handled in nearest-neighbour imputation.

An alternative way to handle a combination of quantitative and qualitative auxiliary variables is to combine the random and nearest-neighbour hot deck methods. That is, we first use the categorical variables to construct imputation classes. Next, within each imputation class, we apply the nearest-neighbour method using a distance function of quantitative variables. In this case, the donor has to match the recipient exactly on the categorical variables but their scores on the quantitative variables may be different. The approach in the previous paragraph offers more flexibility.

It is possible to take sampling weights into account in the selection of the donor; see Kalton (1983) and Andridge and Little (2009). As discussed in "Imputation – Main Module", there is no consensus of opinion on the necessity in general of incorporating sampling weights into imputation procedures. However, it is often useful to ensure that recipients are imputed from donors with similarly-sized weights. Effectively, donor imputation increases the weight of a donor by adding the weights of its recipients (Kalton, 1983). Therefore, if a donor with a small weight is used to impute a recipient with a much larger weight, the influence of that donor on the survey estimates increases disproportionally; as a result, the variances of these estimates will be inflated. To prevent this, the weighting variable – or the design variables that constitute the weighting model – may be included as auxiliary variables in the donor selection. Andridge and Little (2009) compared the performance of hot deck imputation with and without the inclusion of sampling weights in a simulation study.

3. Design issues

4. Available software tools

Several R packages are available that can perform hot deck donor imputation, including `StatMatch` and `mice`. The Banff system by Statistics Canada performs nearest-neighbour imputation for quantitative data. CANCEIS, another tool by Statistics Canada, offers more advanced nearest-neighbour imputation functionality for quantitative and qualitative data. It should be noted that CANCEIS is mainly aimed at social statistics, in particular the population census.

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

Andridge, R. R. and Little, R. J. (2009), The Use of Sampling Weights in Hot Deck Imputation. *Journal of Official Statistics* **25**, 21–36.

Bankier, M., Lachance, M., and Poirier, P. (2000), 2001 Canadian Census Minimum Change Donor Imputation Methodology. Working Paper, UN/ECE Work Session on Statistical Data Editing, Cardiff.

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

Israëls, A., Kuijvenhoven, L., van der Laan, J., Pannekoek, J., and Schulte Nordholt, E. (2011), *Imputation*. Methods Series Theme, Statistics Netherlands, The Hague.

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

Kalton, G. and Kasprzyk, D. (1986), The Treatment of Missing Survey Data. *Survey Methodology* **12**, 1–16.

Little, R. J. A. (1988), Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics* **6**, 287–296.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, second edition. John Wiley & Sons, New York.

Interconnections with other modules

8. Related themes described in other modules

1. Imputation – Main Module
2. Imputation – Model-Based Imputation

9. Methods explicitly referred to in this module

1.

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. Banff
2. CANCEIS
3. R

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-Donor Imputation

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	28-03-2013	first version	Sander Scholtus	CBS (Netherlands)
0.2	15-07-2013	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.3	07-10-2013	improvements based on Norwegian review	Sander Scholtus	CBS (Netherlands)
0.3.1	21-10-2013	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 18:16