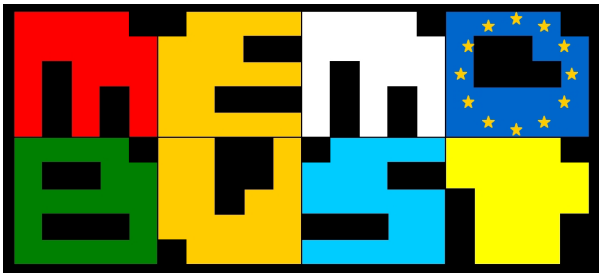

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Imputation – Main Module

Contents

General section

1. Summary

A practical problem that nearly always occurs in statistical research is that the collected data suffer from missing values. This problem occurs both for data collected in traditional surveys and for administrative data. It is usually difficult (but not impossible) to use an incomplete data set directly for inference of population parameters, such as totals or means of target variables. For this reason, statisticians often create a complete data set prior to the estimation stage, by replacing the missing values with estimated values from the available data. This process is referred to as imputation.

To impute the missing values in a data set, several methods are available. Possible imputation methods include: deductive imputation, model-based imputation (including mean, ratio, and regression imputation), and donor imputation (including cold deck, random hot deck, and nearest-neighbour imputation as well as predictive mean matching). Different methods may be useful in different contexts. This module mentions some general aspects of imputation that are not related to a particular method, such as the inclusion or exclusion of a disturbance term in the imputed values, the use of deterministic versus stochastic imputation, and the incorporation of design weights into imputation methods. We also briefly discuss multiple imputation and mass imputation.

2. General description[1]

2.1 Introduction to imputation

In surveys, respondents sometimes do not provide answers to one or more questions, even though they are required to do so. In this case, we refer to *item non-response* (or *partial non-response*) and to missing values that should have been present. Possible reasons for answers not being provided are that the respondents are not willing or able to answer a question. For example, respondents are sometimes unable to answer a question that is complicated or difficult to understand. In business surveys, some businesses may decide to skip questions for which they do not have the required information readily available in their administration. Apart from traditional surveys, administrative data sources can also have missing values.

Missing values may also be introduced when the data are processed for statistical purposes. Typically, the original data contain errors which have to be treated before the data can be used for any meaningful statistical inference. Errors are detected in a process called data editing. For some errors, it is possible to derive the correct value from the observed erroneous value. However, in many cases, the erroneous value does not provide information on the correct value. It is then common practice to initially replace a detected erroneous value by a missing value. We refer to the topic "Statistical Data Editing" for more details.

There are a number of ways to deal with missing values. One way, which we shall focus on here, is to impute valid values for the missing values in the data file. We refer to this process step as *imputing* or *imputation*, and to the resulting values as *imputed values* or *imputations*.

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

An alternative to imputation is to leave the missing values as they are. This will be done first of all for legitimately missing values. For instance, businesses that operate entirely within one country do not have to answer questions on international trade. Ideally, the routing in the questionnaire will ensure that questions are only posed when they are relevant. But even in the case of missing values that should have been present, a decision can be made not to impute, and to resolve the problem not in the data file, but instead at the estimation or analysis stage. Especially for categorical variables, there is the alternative of introducing the extra category 'unknown'. Imputation is used more often for quantitative variables than for categorical ones, and therefore also more often for business statistics than for social statistics.

Reasons to use imputation, instead of leaving missing values in the data set, are as follows:

1. It is convenient to have a 'complete' (completely filled) data file for further processing.
2. Imputation can be used to improve the quality of the microdata and/or of parameter estimates.

Sub 1. Obtaining a complete file, with complete records, makes aggregation and tabulation easier, and prevents inconsistencies when tabulating. For example, if missing values occur for the categorical variable *Legal form*, then this will cause the distribution of *Size class* in the table '*Size class* × *Legal form*' to deviate from the distribution of *Size class* in the table '*Size class* × *Economic activity*', unless the missing values are coded using the category 'unknown' and this category is included in the tables. If, in a sample survey, missing values occur for the quantitative variable *Turnover*, then we cannot directly estimate the total turnover for the entire population, but only for the rather uninteresting subpopulation of businesses who would have responded to the question on turnover when asked. Imputation can help in dealing with this problem, but it is of course only usable when the imputations are of sufficient quality.

Sub 2. If we want to use imputation to improve the quality, we should first decide more explicitly what the product is of which we want to improve the quality. Often, the primary goal is to accurately estimate population means and totals, or ratios of these. We may also want to determine the distribution of a variable, as in the first example in the previous paragraph. For some statistical processes, it is important to produce a good microdata file, which other researchers can use to perform a variety of analyses. Different objectives can have different 'optimal' imputation methods. For statistical output, however, it is generally desirable to perform imputation only once. If each researcher were allowed to use his/her own imputation method, then the results of different studies based on the same incomplete data set might not be consistent. Generally speaking, a statistical institute can provide better imputations for general use than external users, because these external parties typically have access to less background characteristics that are useful for imputation.

There exists a vast literature on imputation. The topic has been studied both at institutes for official statistics and in the academic world. References that provide an overview include Sande (1982), Kalton (1983), Kalton and Kasprzyk (1986), Rubin (1987), Little (1988), Kovar and Whitridge (1995), Schafer (1997), Little and Rubin (2002), Longford (2005), Tsiatis (2006), McKnight et al. (2007), Daniels et al. (2008), and De Waal et al. (2011).

Finally, it should be noted that we use the term 'item non-response' only to refer to missing values that occur in an otherwise observed record. A different problem that is often encountered in practice is that certain units that we would like to observe do not respond at all. This situation is referred to as *unit non-response*. Although unit non-response could, in principle, also be treated with imputation,

weighting methods are more commonly used for this. Weighting methods are described in the topic "Weighting and Estimation".

2.2 Overview of imputation methods

Sometimes, when a value is missing, it is possible to derive the true value with (near) certainty from other characteristics of the unit, including the observed values. For such cases, *deductive imputation* can be used. Often, this type of imputation is carried out by means of 'if-then rules' designed by subject-matter specialists. Deductive imputation methods may also use information from restrictions (edit rules) that have to be satisfied by the missing values. If applicable, deductive imputation has preference above all other imputation methods. However, in most cases, it is not possible to impute all missing values in this way. We refer to the module "Imputation – Deductive Imputation" for more details.

Even when a deductive imputation is not possible, there will often be extra information (in the form of auxiliary variables, or *x*-variables) that makes possible a more or less accurate prediction of the missing values (on the target variables, or *y*-variables). By searching for a suitable explanatory model, one can try to improve the quality of the file or of the population parameters to be estimated using *model-based imputation*. This method requires that the parameters of the model are estimated first based on the item respondents. The fitted model then generates the value(s) to be filled in. It is not possible to assess the exact quality of the imputations: the true values are, after all, unknown. There will usually be an imputation bias (bias in the outcomes as a result of creating erroneous imputations), because the fitted model with the parameters based on the item respondents will usually not apply exactly for the item non-respondents. This bias is acceptable so long as it is negligible compared to other sources of inaccuracy in the outcome, such as the sampling variance.

The objective in model-based imputation is to search for a model for *y* involving zero, one, or more auxiliary variables, which will predict the missing values of *y* as accurately as possible. A regression model is commonly used for this purpose, and this type of model-based imputation is referred to as *regression imputation*. The methods *mean imputation* and *ratio imputation* are special cases of regression imputation. For mean imputation, no auxiliary variables are used; all missing values are then imputed by the mean value of the item respondents (hence the name of the method). For ratio imputation, only a single quantitative auxiliary variable is used. These methods are mentioned separately because of their simplicity and frequent application in practice. We refer to the module "Imputation – Model-Based Imputation" for more details on mean, ratio, and regression imputation, as well as other model-based methods.

Another class of imputation methods is that of *donor imputation*, which includes *cold deck*, *random hot deck*, *sequential hot deck* and *nearest neighbour* (including *predictive mean matching*). Broadly speaking, donor imputation methods operate as follows: for each non-respondent *i*, a donor record *d* is found that has as many as possible of the same characteristics as the *recipient i*, insofar as these characteristics are considered to influence the target variable *y*. Subsequently, the donor score, say y_d, is used as imputation: $\tilde{y}_i = y_d$. Donor methods are somewhat easier to use than model-based methods in the case that multiple missing values must be imputed in a single record, while it is important to preserve as much as possible the correlations between the variables. We refer to the module "Imputation – Donor Imputation" for more details.

Model-based imputation and donor imputation are often used for cross-sectional data, where most of the information for imputing missing values comes from observed data for other units. In the case of longitudinal data (panels), it is also possible to use observed data from the same unit on other time points for imputation. We refer to the module "Imputation – Imputation for Longitudinal Data" for details.

Finally, we already mentioned that statistical data often have to satisfy certain restrictions or edit rules. Apart from deductive imputation, the methods that we have discussed so far do not take these edit rules into account. Hence, it is not guaranteed that the imputations made by these methods will satisfy the edit rules. There are two ways to solve this problem. The first option is to try to take the edit rules into account in the imputation method itself, for instance by choosing an appropriate model which generates imputations that automatically satisfy the edit rules. A drawback of this approach is that, in practice, it can easily lead to very complex imputation models. However, there are some exceptions where direct modeling of the edit rules is feasible.

The second option is to use a two-step approach. In the first step, the missing values are imputed using an appropriate imputation method which does not take (all) edit rules into account. Then, in a second step, the imputed values are minimally adjusted to satisfy the edit rules, according to some minimisation criterion. This approach is more commonly used in practice than the first approach, because it is easier to apply. Both approaches are discussed in the module "Imputation – Imputation under Edit Constraints".

In the remainder of this section, we briefly mention general aspects of imputation that are not related to particular imputation methods.

2.3 Imputation classes

Instead of constructing an imputation model for the entire population, we may also fit different models for different subpopulations: for instance, a different model for each stratum in the crossing *Economic activity* × *Size class*. Such strata, which are treated separately in the imputation process, are referred to as *imputation classes*. Using imputation classes can be effective if, within the classes, the variable to be imputed shows little variation, while the values between the classes vary significantly.

Since categorical x-variables can be included as dummy variables in the imputation model, for model-based imputation, distinguishing between imputation classes can also be considered as a part of the modeling, namely by selecting categorical auxiliary variables that correlate strongly with target variable y and including these variables in the model with all the interaction terms. Random hot deck donor imputation is, by definition, only intended for categorical x-variables, and consequently always performs imputation within classes. The y-variables may be categorical or quantitative for this method.

2.4 Selection of auxiliary variables

The selection of variables and interactions is not discussed in detail here. Just as regression analysis, it is a part of general multivariate analysis on which much literature is available. Basically, one will look for auxiliary variables that correlate strongly with the target variable y and, preferably, explain the occurrence of non-response as accurately as possible. It is generally a matter of trial and error and common sense, but forward or backward search procedures can also be used to automatically add or

remove *x*-variables to or from the model. There also exist automatic procedures to select homogeneous imputation classes for categorical *x*-variables, such as WAID (Chambers et al., 2001a and 2001b).

In model-based imputation and the nearest-neighbour method, both categorical and quantitative *x*-variables can be included. In hot deck donor imputation, only categorical auxiliary variables can be included. Quantitative variables can be included indirectly by first categorising them. Obviously, the quantitative aspect of the variable is then partially lost.

In hot deck donor imputation, there is a limit to the number of relevant *x*-variables that can be included, much more so than in regression imputation. This happens because, for hot deck methods, all the interactions between the categorical variables are implicitly included in the model, which means that the number of model parameters can easily become too large compared to the sample size. In practice, this problem manifests itself in the form of empty imputation classes. In a regression model, the number of parameters may be reduced by not including all interactions in the model.

Finally, it should be noted that it is also possible to include variables with missing values as explanatory variables in an imputation model. This requires a more advanced method for model-based imputation, for instance a sequential regression algorithm. In this situation, there is no clear distinction between auxiliary variables and target variables. Sequential regression imputation is briefly discussed in the module "Imputation – Model-Based Imputation".

2.5 Imputation with or without disturbance term

For a missing value on *y*, one could impute the best possible prediction according to the regression model. If this is done for all the missing values, then all the imputed records satisfy the imputation model perfectly. As a result, the imputed data are often useless for further analyses, or even sometimes for simple tabulations, because these will merely reproduce the model that was used for imputation. For this reason, it is important to 'flag' the imputed values so that they may be recognised by other users of the data, and to document the imputation model that was used.

In general, the imputation of the best possible prediction according to the regression model creates an underestimation of the variation in the scores ('regression to the mean'). This leads to distributions that are too peaked and tail areas that are too thin, especially if *y* has many missing values and the regression explains little of the variance of *y*. This effect is the strongest in mean imputation. It does not form an obstacle for the estimation of means or totals, but it does for the estimation of distributions and dispersion measures.

For an accurate estimation of a distribution it is advisable to impute values by adding a random disturbance to the best possible prediction according to the model. In regression analysis, we can choose between (1) drawing a random disturbance from a normal probability distribution, and (2) adding the residual of a randomly drawn donor. In donor imputation, a residual is always used implicitly, namely the residual of the randomly or deterministically selected donor. The dispersion in the distribution of *y* is therefore retained.

2.6 Multiple imputation

Rubin (1987) observed that, even after adding random disturbances to the imputations, the variance of an estimator can still be underestimated, because the uncertainty of the imputation model itself is not taken into account. This can be a problem, especially when the statistician producing an imputed data

set and the researcher performing analyses on this imputed data set are different persons. In this context, *multiple imputation* provides a general approach that enables a researcher to perform valid inferences (including statistical tests) and to obtain valid estimates of standard errors and confidence intervals. Multiple imputation involves, quite literally, the creation of multiple imputed values for each missing value, based on different parameter estimates, random disturbances or models.

Rubin (1996) writes: "Multiple imputations for the set of missing values are multiple sets of plausible values; these can reflect uncertainty under one model for nonresponse and across several models. Each set of imputations is used to create a completed data set, each of which is to be analyzed using standard complete-data software to yield 'completed-data' statistics (…)." Formally, the multiple imputations are obtained as draws from a posterior distribution under a posited Bayesian model for the data and the non-response mechanism (see, e.g., Rubin, 1987).

Under multiple imputation, each imputed data set produces a point estimate for a parameter of interest, as well as an associated variance estimate, which does not take the imputation variability into account. The final variance estimate, which does take the imputation variability into account, is obtained as

$$\hat{V}_{MI} = \hat{V}_{within} + \left(1 + \frac{1}{m}\right)\hat{V}_{between}, \tag{1}$$

where $\hat{V}_{within}$ denotes the mean of the separate variance estimates from the imputed data sets, $\hat{V}_{between}$ denotes the variance of the separate point estimates from the imputed data sets, and $m > 1$ denotes the number of imputations. In practice, a small value such as $m = 5$ is usually sufficient (Rubin, 1996). From a different point of view, Kim et al. (2006) showed that variance estimators based on formula (1) may be biased for finite-population sampling under a superpopulation model.

2.7 Deterministic or stochastic imputation

If a random selection is made from available donors (for donor imputation) or a random draw from a distribution of residuals is added to the predicted value (for model-based imputation), this is referred to as *stochastic imputation*. This introduces a randomness in the imputation process which makes the imputations not reproducible. In *deterministic imputation*, on the other hand, the imputations are reproducible based on the chosen imputation model. In many cases, the distinction between stochastic and deterministic imputation is identical to the distinction between using or not using a disturbance term, as discussed in Subsection 2.5. There exist some exceptions, however. For instance, nearest-neighbour imputation (including predictive mean matching) is deterministic provided that the distance function always leads to a unique donor. If a recipient has several possible donors with the same minimal distance, then an additional selection criterion is needed. This may introduce a stochastic element into the imputation procedure (e.g., when one of the nearest neighbours is chosen at random).

2.8 Weighting – yes or no?

Most imputation methods have an option to attach unequal weights to the item respondents. This option can be used, for example, to incorporate in the imputation process the design weights associated with the sampling design (i.e., the reciprocals of the inclusion probabilities), or weights that correct for selective unit non-response. In linear regression imputation, this means that weighted least squares estimation is performed to fit the model, rather than unweighted least squares estimation. In hot deck donor imputation, it means that potential donors with larger weights have a greater

probability of being selected as a donor. Weighting does not have an influence on deductive imputation and on nearest-neighbour methods.

No clear-cut recommendation can be given on the inclusion of design weights in imputation methods. From a model-based perspective, every outcome is measured equally reliably (assuming identically distributed disturbances), regardless of the inclusion probability or response probability of units. Confidence in the imputation model therefore means that weighting does not need to be used, and it is even better not to use it, because weighting increases the model-based standard errors. If we can include the variable with weights, or the variables forming the basis for the weighting, as auxiliary variables in the model, then further weighting is also unnecessary. An option therefore is to provide for this in the selection of *x*-variables.

In contrast to the above, from the perspective of sampling theory, the answers of a sample unit are 'representative' for population elements that are not selected. From this point of view, and particularly in the case of selective unit non-response, weighting is needed to obtain design-unbiased estimators.

The general problem of model estimation for sampling designs with unequal inclusion probabilities is discussed, for instance, by Skinner et al. (1989). For donor imputation, Kalton (1983) offered several methods in which the probability of being a donor is proportional to the weight. It can be useful to ensure also that the donor and recipient have a similar weight, to prevent an object with a very small weight from being the donor for a recipient with a very large weight, as a result of which the weight of the donor increases disproportionately (it receives too much weight). Again, we can try to prevent this by including the weighting variable or the auxiliary variables that form the basis for the weighting as categorical *x*-variables. See also Andridge and Little (2009) for a simulation study of hot deck imputation with and without the inclusion of sampling weights.

2.9 Mass imputation

Sometimes, it is desirable to impute values not only for the item non-respondents, but for all the units not occurring in the sample. We call this *mass imputation*, even if it concerns only one target variable *y*. Naturally, a register or sampling frame of all units is needed in this case. After mass imputation, we can easily calculate totals and means for *y* by simple aggregation of all observed or imputed values.

For weighted hot deck imputation, mass imputation corresponds to the use of the so-called post-stratification estimator. For regression imputation with weighted least squares estimation, mass imputation corresponds with the regression estimator. These estimators are discussed in the topic "Weighting and Estimation". See also De Waal et al. (2011, Section 7.3.4) for the connection between (mass) imputation and the regression estimator.

3. Design issues

4. Available software tools

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.

Chambers, R. L., Hoogland, J., Laaksonen, S., Mesa, D. M., Pannekoek, J., Piela, P., Tsai, P., and de Waal, T. (2001a), The AUTIMP-Project: Evaluation of Imputation Software. Report, Statistics Netherlands, Voorburg.

Chambers, R. L., Crespo, T., Laaksonen, S., Piela, P., Tsai, P., and de Waal, T. (2001b), The AUTIMP-Project: Evaluation of WAID. Report, Statistics Netherlands, Voorburg.

Daniels, J., Daniels, M. J., and Hogan, J. W. (2008), *Missing Data in Longitudinal Studies*. Taylor & Francis, Philadelphia.

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.

Kim, J. K., Brick, J. M., Fuller, W. A., and Kalton, G. (2006), On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling. *Journal of the Royal Statistical Society, Series B* **68**, 509–521.

Kovar, J. and Whitridge, P. (1995), Imputation of Business Survey Data. In: B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P .S. Kott (eds.), *Business Survey Methods*, John Wiley & Sons, New York, 403–423.

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.

Longford, N. T. (2005), *Missing Data and Small-Area Estimation*. Springer-Verlag, New York.

McKnight, P. E., McKnight, K. M., Sidani, S., and Figueredo, A. J. (2007), *Missing Data – A Gentle Introduction*. Guilford Publications, New York.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.

Rubin, D. B. (1996), Multiple Imputation after 18+ Years. *Journal of the American Statistical Association* **91**, 473–489.

Sande, I. G. (1982), Imputation in Surveys: Coping with Reality. *The American Statistician* **36**, 145–152.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Skinner, C. J., Holt, D., and Smith, T. M. F. (eds.) (1989), *Analysis of Complex Surveys*. John Wiley & Sons, Chichester.

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*. Springer-Verlag, New York.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Data Editing – Main Module
2. Imputation – Model-Based Imputation
3. Imputation – Donor Imputation
4. Imputation – Imputation for Longitudinal Data
5. Imputation – Imputation under Edit Constraints
6. Weighting and Estimation – Main Module

9. Methods explicitly referred to in this module

1. Imputation – Deductive Imputation

10. Mathematical techniques explicitly referred to in this module

1. n/a

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. n/a

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-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	23-12-2011	first version	Sander Scholtus	CBS (Netherlands)
0.2	29-03-2012	improvements based on Norwegian review	Sander Scholtus	CBS (Netherlands)
0.2.1	04-03-2013	adjusted to new template; minor improvements	Sander Scholtus	CBS (Netherlands)
0.3	07-10-2013	minor improvements based on Swedish 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 18:15