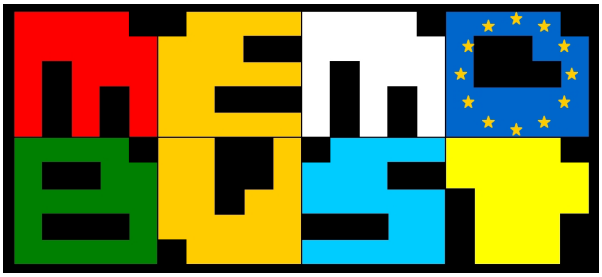

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Model-Based Imputation

Contents

General section

1. Summary

The objective in model-based imputation is to find a predictive model for each target variable in the data set that contains missing values. The model is fitted on the observed data and subsequently used to generate imputations for the missing values. Several commonly-used imputation methods are special cases of model-based imputation; this includes mean imputation, ratio imputation, and regression imputation.

2. General description[1]

2.1 Introduction to model-based imputation

The objective in model-based imputation is to find a predictive model for each target variable in the data set that contains missing values. The model is fitted on the observed data and subsequently used to generate imputations for the missing values. Many practical applications use a separate model for each variable in the data set. Some multivariate extensions will be briefly discussed in Section 2.6. Before that, we will discuss *mean imputation* (Section 2.2), *ratio imputation* (Section 2.3), and *regression imputation* (Section 2.4). Section 2.5 treats certain practical issues related to the application of these methods.

2.2 Mean imputation

In mean imputation, each missing value is replaced by the observed mean of all item respondents. That is, if y_i denotes the score of the i^{th} unit on the target variable, then each missing value is imputed by

$$\tilde{y}_i = \bar{y}_{obs} = \frac{\sum_{k \in obs} y_k}{n_{obs}}, \tag{1}$$

with obs denoting the set of n_{obs} item respondents for variable y.

Obviously, mean imputation leads to a peak in the distribution of y, because the same value is imputed for all item non-respondents. On the micro level, the quality of the imputations produced by this method is generally low. The method is potentially suitable if the intended output is limited to estimates of population means and totals. In general, mean imputation is not suitable for estimating dispersion measures such as the standard deviation, frequency distributions, or correlations between target variables, because these can all be distorted by imputing observed means. The main advantage of this method is its simplicity.

It is possible to apply mean imputation within imputation classes, i.e., groups that are more or less homogeneous with respect to the target variable. In this case, formula (1) is replaced by

$$\tilde{y}_{hi} = \bar{y}_{h;obs} = \frac{\sum_{k \in h \cap obs} y_{hk}}{n_{h;obs}},$$

[1] This section is to a large extent based on Chapters 3, 4, and 5 of Israëls et al. (2011).

where y_{hi} is the score of the i^{th} unit in imputation class h and $n_{h;obs}$ is the number of item respondents for variable y in h. This extension is sometimes referred to as 'group mean imputation'. In the context of business surveys, domain estimates by economic activity and size class are often part of the intended output. In that case, it is natural to define imputation classes based on these classifying variables, which are in fact known to correlate strongly with many economic target variables. Compared to using overall mean imputation, the use of group mean imputation should significantly improve the quality of the domain estimates and, usually, also the population estimates.

In general, group mean imputation produces a set of smaller peaks in the distribution of y (one for each imputation class). If the imputation classes are very effective in discriminating among the units, so that the variation of y between classes is much larger than the variation within classes, then this method can also be used to reasonably estimate dispersion measures. This is true because only the variation of y within classes is disregarded under this method.

2.3 Ratio imputation

For ratio imputation, we assume that there is a single auxiliary variable x that is always observed (or previously imputed) and that is more or less proportional to the target variable y. First, the unknown ratio between y and x, say R, is estimated from the units with both y and x observed:

$$\hat{R} = \sum_{k \in obs} y_k \Big/ \sum_{k \in obs} x_k \ .$$

Subsequently, the missing y_i are imputed by applying this ratio to the observed x_i:

$$\tilde{y}_i = \hat{R} x_i = \frac{\sum_{k \in obs} y_k}{\sum_{k \in obs} x_k} x_i \ . \qquad (2)$$

Thus, the imputed values are obtained by assuming that the proportion that was estimated from the respondents holds exactly for the item non-respondents.

As an illustration, suppose that y denotes *turnover* and x denotes *number of employees*. Then the ratio R represents the average turnover per employee. According to (2), multiplying the observed *number of employees* for the i^{th} unit by the estimated average turnover per employee yields an estimate of *turnover* for the i^{th} unit, and this estimate is used as an imputation.

A common application of ratio imputation occurs in repeated surveys, where the value of y measured at an earlier time (say $t-1$, with t denoting the current time) is used as auxiliary information. In this case, we can write $y = y^t$ and $x = y^{t-1}$. The imputation is then given by

$$\tilde{y}_i^t = \hat{R} y_i^{t-1} \ ,$$

with $\hat{R}$ the estimated development of the target variable between $t-1$ and t. We refer to the module "Imputation – Imputation for Longitudinal Data" for more details on imputation in this context.

As with mean imputation, ratio imputation can also be applied within imputation classes. In this case, a separate ratio R_h is estimated for each imputation class and used in formula (2). This may be called

'group ratio imputation'. In general, this extension is useful if the relationship between x and y differs strongly, or at least significantly, between the imputation classes. It should be noted that ratios of groups are usually more homogeneous than group means. Regarding domain estimates in business surveys, the same remarks apply here as for group mean imputation.

2.4 Regression imputation

Regression imputation generalises mean and ratio imputation by assuming a regression model for the prediction of y given a set of auxiliary variables $x_1,\ldots,x_q$. In many cases, a standard linear regression model is used:

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon\,, \tag{3}$$

with $\alpha,\beta_1,\ldots,\beta_q$ unknown parameters and ε a disturbance term, where it is assumed that the disturbances for all units are drawn independently from the same normal distribution with mean 0 and variance σ^2.

The parameters in model (3) are estimated – usually through ordinary least squares – from the records for which both y and the auxiliary variables are observed. This results in a prediction for y given the auxiliary variables:

$$\hat{y} = a + b_1 x_1 + \cdots + b_q x_q\,, \tag{4}$$

with $a,b_1,\ldots,b_q$ denoting the least squares estimates of $\alpha,\beta_1,\ldots,\beta_q$. Assuming that the auxiliary variables are always observed, this predicted value can be computed for both item respondents and item non-respondents on y.

There are now two generic ways to obtain an imputation $\tilde{y}_i$ from the regression model: without a disturbance term or with a disturbance term. In the first case, the predicted value from (4) is substituted directly for the missing value:

$$\tilde{y}_i = \hat{y}_i = a + b_1 x_{1i} + \cdots + b_q x_{qi}\,. \tag{5a}$$

This results in a deterministic imputation. In the second case, we add a disturbance to the predicted value, i.e., we impute:

$$\tilde{y}_i = \hat{y}_i + e_i = a + b_1 x_{1i} + \cdots + b_q x_{qi} + e_i\,. \tag{5b}$$

The disturbance e_i can be a random draw from the normal distribution with mean 0 and variance σ^2, to be in line with the posited regression model (3). (Actually, σ^2 is unknown in practice and is often estimated by the residual error of the fitted model.) Alternatively, a donor can be selected from the item respondents (either at random or according to some deterministic criterion; see the module "Imputation – Donor Imputation") and the residual of the donor with respect to the model prediction, say $e_d = y_d - \hat{y}_d$, can be substituted for e_i. In both cases, the disturbance is obtained using the regression model. Adding a disturbance results in a stochastic imputation, unless one uses a donor that is selected in a deterministic way. We refer to "Imputation – Main Module" for a discussion of the

differences between imputing with and without a disturbance term and between deterministic and stochastic imputation.

It should be noted that mean imputation can be seen as a special case of regression imputation, namely in the absence of auxiliary variables. In this case, model (3) reduces to

$$y = \alpha + \varepsilon ,$$

and the least squares estimate a is just the observed mean $\bar{y}_{obs}$, so that formula (5a) is identical to (1). Similarly, ratio imputation can be seen as a special case of regression imputation with one auxiliary variable and with the constant term fixed to 0. In this case, model (3) reduces to

$$y = \beta x + \varepsilon .$$

Under the alternative assumption that the variance of the disturbances equals $\sigma^2 x$ rather than σ^2, the weighted least squares estimate for β is just the observed ratio $\hat{R}$, and formula (5a) is identical to (2). Note that there also exist stochastic versions of mean and ratio imputation; these are obtained by taking formula (5b) instead of (5a) in the above special cases.

In practice, the standard linear regression model may not always be appropriate. More generally, a non-linear regression model could be used, i.e.,

$$y = f(\beta_1 x_1 + \cdots + \beta_q x_q)$$

for some non-linear function $f(.)$. The disturbance term ε can be added to this model, or it can be implicitly contained therein.

In the case of a binary target variable with scores 0 and 1, a logistic regression model is often used:

$$\log \frac{p}{1-p} = \alpha + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon ,$$

where p denotes the probability that y takes the score of 1, given the auxiliary variables. As before, the data of the item respondents can be used to estimate the model parameters (e.g., using maximum likelihood). Next, for each unit with y_i missing, the probability that $y_i = 1$ is estimated according to

$$\hat{p}_i = \frac{\exp(a + b_1 x_{1i} + \cdots + b_q x_{qi})}{1 + \exp(a + b_1 x_{1i} + \cdots + b_q x_{qi})} \in (0,1) .$$

Having estimated these probabilities, imputed values may be obtained either by directly imputing $\tilde{y}_i = \hat{p}_i$ (this yields a deterministic imputation) or by randomly drawing $\tilde{y}_i = 1$ with probability $\hat{p}_i$ and $\tilde{y}_i = 0$ with probability $1 - \hat{p}_i$ (this yields a stochastic imputation).

Note that if we impute $\tilde{y}_i = \hat{p}_i$ in the above case, the individual imputations are not valid scores (i.e., they are not equal to 0 or 1). More generally, regression imputation can produce imputations outside the domain of values that are theoretically possible for the target variable. For instance, an imputed number of employees may be non-integer, an imputed turnover may be negative, etc. Typically, this is not a problem for the estimation of population means, totals and many other statistics, but it may be problematic in applications where the microdata themselves are part of the output. If valid individual imputations are desired, then it may be better to turn to donor imputation (see "Imputation – Donor

Imputation"). See also the module "Imputation – Imputation under Edit Constraints" for the more general problem of imposing (multivariate) restrictions on the imputed values.

2.5 Practical issues

The regression model (3) is defined for a quantitative target variable and quantitative auxiliary variables. Categorical auxiliary variables, such as *NACE code* or *size class*, can be included in this model by defining appropriate dummy variables. In particular, group mean imputation is obtained as a special case of regression imputation by including only a dummy variable for each imputation class. For categorical target variables, other models should be used, such as (a multinomial extension of) logistic regression.

It is important to assess the quality of imputations. A direct comparison between the imputed values and the actual values is usually impossible, since the actual values are unknown. In some cases, it may be possible to obtain an impression of the quality of imputation through external validation, by comparing the imputed data to data from another source, either for the individual imputed values or at an aggregate level. Usually, however, there are conceptual differences between the various sources (different variable definitions, different target populations, etc.) so that opportunities for these types of validation are limited.

An indirect measure of the quality of a model-based imputation is provided by various indicators of model fit. For linear regression analysis with the least squares estimator, the fraction of explained variance R^2 can be used to quantify the strength of the model among the item respondents. In this way, different imputation models can be compared with one another; note that gains in R^2 for larger models should be set off against increases in degrees of freedom. For more general models, the likelihood can be used as an indicator, or a measure derived from the likelihood such as AIC or BIC. See Draper and Smith (1998) – or any other introductory book on regression analysis – for a more comprehensive discussion of model selection and ways to assess model fit. A limitation of using the model fit to assess imputation quality is that, in theory, it is possible for model *A* to have a better fit than model *B* among the item respondents, while model *B* provides better predictions than model *A* among the item non-respondents.

Another possibility to obtain an impression of the quality of different imputation methods in a particular context is to perform a simulation experiment with either the actual data set or historical data. In such an experiment, observed values are temporarily suppressed and new values are imputed for these left-out values. To the extent that the imputed values are similar to or – for categorical variables – even equal to the original values, an imputation method appears to be useful for a particular application. By defining a suitable distance function between the imputed and observed values – or, often more aptly, between target estimates based on these values –, it is possible to compare different imputation methods/models and choose the most appropriate one. This can be seen as an application of cross-validation. We refer to Schulte Nordholt (1998) and Pannekoek and De Waal (2005) for examples of such experiments. A good introduction into the design and use of simulation studies is given by Haziza (2006).

2.6 *Multivariate methods*

In the previous subsections, we have treated model-based imputation methods that impute a data set on a variable-by-variable basis. There also exist model-based methods that take a multivariate approach to imputation. Although these multivariate methods are more complex to use, they do have some theoretical advantages (De Waal et al., 2011, pp. 277-279). If y is imputed by a single-variable method, then typically the relationships between y and all other variables in the data set will be distorted *except* for those variables that were included as auxiliary variables in the imputation model for y. Thus, if the intended output includes correlations between target variables or other statistics of a multivariate nature, it is important to take this into account in the choice of the imputation model. Multivariate imputation methods provide a natural way to preserve correlations between target variables. Another advantage of multivariate methods is that there exist techniques that estimate a multivariate model by making use of all the available observed data (see below). As discussed above, for single-variable methods, the model has to be fitted using only the units with all predictors and the target variable observed.

2.6.1 *Multivariate regression imputation*

Using matrix-vector notation, a straightforward extension of the standard linear regression model (3) to the case of multiple target variables is given by:

$$\mathbf{y} = \boldsymbol{\mu}_y + \mathbf{B}_{y,x}(\mathbf{x} - \boldsymbol{\mu}_x) + \boldsymbol{\varepsilon}, \tag{6}$$

where, for simplicity, we make the assumption that each target variable in $\mathbf{y}$ is modeled using the same vector of auxiliary variables $\mathbf{x}$. In the absence of missing data, the matrix of regression coefficients $\mathbf{B}_{y,x}$ could be estimated from the data using least squares:

$$\hat{\mathbf{B}}_{y,x} = \mathbf{S}_{y,x}\mathbf{S}_{x,x}^{-1},$$

with $\mathbf{S}_{y,x}$ the matrix of observed covariances between the target variables and the auxiliary variables, and $\mathbf{S}_{x,x}$ the observed covariance matrix of the auxiliary variables. In addition, $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_x$ could be estimated by their observed means: $\hat{\boldsymbol{\mu}}_y = \bar{\mathbf{y}}$ and $\hat{\boldsymbol{\mu}}_x = \bar{\mathbf{x}}$.

In the presence of missing data, the above estimates cannot be computed, but one could base analogous estimates only on those units for which all relevant variables are observed. However, this approach has two important drawbacks. Firstly, in particular for larger models, the number of fully observed units may be very small and the resulting estimates may be unreliable. Secondly, and perhaps more importantly, the fully observed units may form a selective subset of all units. As a result, using the fitted model to impute the item non-respondents may produce a bias in the statistical output.

A more satisfactory solution may be provided by maximum likelihood estimation with incomplete data. Under certain assumptions on the mechanism that causes the missing values, the so-called *Expectation-Maximisation (EM) algorithm* provides valid estimates of the parameters in model (6). This approach uses all the available information in the observed data to estimate these parameters, including the units with partially observed records. The interested reader is referred to De Waal et al. (2011, Ch. 8) for a brief introduction and Little and Rubin (2002) for more details.

Having obtained estimates of the unknown parameters in model (6), imputations for the missing values in a record $\mathbf{y}_i$ may be obtained as before from the observed vector $\mathbf{x}_i$. That is, a deterministic imputation is obtained directly from the predicted value,

$$\tilde{\mathbf{y}}_i = \hat{\mathbf{y}}_i = \hat{\boldsymbol{\mu}}_y + \hat{\mathbf{B}}_{y,x}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x),$$

and a stochastic imputation is obtained by adding a random disturbance to this prediction:

$$\tilde{\mathbf{y}}_i = \hat{\mathbf{y}}_i + \mathbf{e}_i = \hat{\boldsymbol{\mu}}_y + \hat{\mathbf{B}}_{y,x}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x) + \mathbf{e}_i.$$

A common choice is to draw $\mathbf{e}_i$ from a multivariate normal distribution with mean vector zero and the covariance matrix of the residuals of the regression of $\mathbf{y}$ on $\mathbf{x}$ (cf. De Waal et al., 2011).

2.6.2 Sequential regression imputation

In practice, applying multivariate model-based imputation as described in the previous subsection can be complicated, particularly if the data set contains a large number of variables of different types (continuous, semi-continuous, binary, etc.). It is difficult, if not impossible, to find an explicit joint model that is appropriate for such data. Van Buuren et al. (1999) and Raghunathan et al. (2001) proposed a different method, known as *sequential regression imputation* or *multivariate imputation by chained equations*. Under this approach, one models the distribution of each target variable separately, conditional on the values of the other variables. This yields a set of single-variable regression models, which have to be estimated in an iterative manner. To do this, the following procedure can be used:

1. Initialise the procedure by imputing each missing value in the original data set by a simple method (e.g., mean imputation).
2. For each variable in turn:
    a. Estimate the parameters of the conditional regression model using all records in the current data set for which this variable was originally observed.
    b. Use the estimated conditional model to impute the originally missing values for this variable. This updates the current data set for the next iteration.
3. Repeat Step 2 until ‘convergence’.

Note that in Step 2a, the conditional regression model is estimated using the most recent imputed version of each independent variable. In Step 3, ‘convergence’ may be assessed in terms of stability across iterations of the estimated regression parameters or the imputed values. The imputations from the final iteration are to be used in subsequent processing.

As noted above, the main practical advantage of the sequential regression approach lies in the flexibility provided by the use of separate, conditional regression models. It should be noted that this approach is theoretically justified only if the conditional models imply a proper joint model for the data. (The conditional models have to be ‘compatible’.) Otherwise, the iterative estimation procedure will not converge to a stable solution. Although this assumption usually cannot be verified beforehand, experiences so far suggest that it does not pose a problem in most practical applications (Tempelman, 2007).

Sequential regression is often applied in the context of multiple imputation. (A short discussion of multiple imputation is provided in "Imputation – Main Module".) In fact, it is straightforward to repeat the above iterative procedure to generate multiple imputed data sets. Note that stochastic imputation should be used to make this procedure meaningful.

A good practical introduction into the sequential regression approach to imputation is provided by Azur et al. (2011). Applications in the context of business survey data are described by Tempelman (2007) and Drechsler (2009).

3. Design issues

4. Available software tools

Mean and ratio imputation can be implemented using almost any statistical software. Regression imputation with common types of models (e.g., linear regression, logistic regression) is provided as a standard feature in tools such as SPSS, SAS, and Stata. It is also straightforward to implement in R. Specialised packages are available for sequential regression imputation, such as IVEware (in SAS), and `mice` and `mi` (in R).

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

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011), Multiple Imputation by Chained Equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research* **20**, 40–49.

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

Draper, N. R. and Smith, H. (1998), *Applied Regression Analysis*, 3rd edition. John Wiley & Sons, New York.

Drechsler, J. (2009), Far from Normal – Multiple Imputation of Missing Values in a German Establishment Survey. Working Paper, UN/ECE Work Session on Statistical Data Editing, Neuchâtel.

Haziza, D. (2006), Simulation Studies in the Presence of Nonresponse and Imputation. *The Imputation Bulletin* **6**, 7–19.

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

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

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.

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.

Schulte Nordholt, E. (1998), Imputation: methods, simulation experiments and practical examples. *International Statistical Review* **66**, 157–180.

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

Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999), Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine* **18**, 681–694.

Interconnections with other modules

8. Related themes described in other modules

1. Imputation – Main Module
2. Imputation – Donor Imputation
3. Imputation – Imputation for Longitudinal Data
4. Imputation – Imputation under Edit Constraints

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1. Least squares estimation
2. Maximum likelihood estimation
3. EM algorithm

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. SPSS
2. SAS
3. Stata
4. 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-Model-Based Imputation

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	14-03-2013	first version	Sander Scholtus	CBS (Netherlands)
0.1.1	27-03-2013	minor changes	Sander Scholtus	CBS (Netherlands)
0.2	11-07-2013	improvements based on Norwegian and Swedish reviews	Sander Scholtus	CBS (Netherlands)
0.2.1	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