Imputation methods for a binary variable

Seppo LAAKSONEN[1], University of Helsinki, Finland

Abstract

Binary variables are common in surveys including such as employed vs unemployed, healthy vs unhealthy or poor vs non-poor. The last one is used in the examples of this paper. It is unfortunate that survey data are never complete, that is, missingness occurs, sometimes severely. In those cases missingness may violate estimates due to unit nonresponse. For this purpose, there are also sophisticated nonresponse adjustments possible to use, and these should be applied in a best way taking advantage of auxiliary variables from the individual nonrespondents as well as from the population level aggregates. In the case when the nonrespondents participate partially in the survey (item nonresponse), there exists more micro level auxiliary variables available. Such data should naturally be exploited in a best way as well. This leads to apply imputation methods. In this paper, our binary variable gives a too low poverty rate when estimating it from the completely replied respondents. So it is beneficial to improve the estimate with imputations. This is not necessarily easy since our pattern of auxiliary data is not excellent that is common in real life. Consequently, the results depend more than in a nice situation on an imputation method applied. In this case we compare several methods. The imputation process consists of the two steps, (i) imputation model and (ii) imputation task. The dependent variable of the imputation model is also binary but they are of two types: (a) the binary variable being imputed itself, or (b) the binary response indicator. The former takes advantage of the data of the respondents but the latter both from the respondents and from the nonrespondents. Since we know the true values (but so that their mechanism is not known), we can compare different methods. When using the methods with random numbers, we also can apply multiple imputation methodology. So we have both single imputation and multiple imputation methods compared in the empirical part. Our strategy is not only Bayesian for multiple imputation that is usual in software packages such as SPSS and SAS. We test their methods but concentrate on our own solutions, called non-Bayesian.

Keywords

Model-donor imputation methods, real-donor imputation methods, logistic regression, probit regression

Introduction

Imputation is for replacing missing or other incorrect or deficient values (later in the text the term ‘missing’ covers those other problematic cases as well) with plausible ones. If this procedure has performed once, it is single imputation (SI). SI is a usual tool in statistical offices or other public survey institutes, in particular. However, SI can be performed several times as well. If this procedure is repeated a number of times and ‘coordinated’ well, the outcome is ‘multiple imputation’ (MI). What such a good coordination means, it is a special question? Rubin in his books (1987, 2004) says that each imputation should be ‘proper’ (Rubin 2004, 118-119; note that since that we refer to the 2004 book that is slightly revised from the older one). Rubin also gives some rules for proper imputation but they are not necessarily easy to follow, or their implementation is not automatic. A big question here is how to repeat the imputation process well, that is, what is an appropriate Monte Carlo technique in order to get $L>1$ simulated versions for missing values?

[1] *Seppo.Laaksonen@Helsinki.Fi*

Rubin (1996, 476, 2004, 75&77) also says that a theoretically fundamental form of MI is repeated imputation. His repeated imputations are draws from the posterior predictive distribution under a specific model that is a particular Bayesian model for both the data and the missing-data mechanism.

Several proper MI implementations are given in Rubin's books and in software packages using this book. He thus recommends that imputations should be created through a Bayesian process as follows: (i) specify a parametric model for the complete, (ii) apply a prior distribution to the unknown model parameters, and (iii) simulate *L* independent draws from the conditional distribution of the missing data given the observed data by Bayes' Theorem.

These Rubin's theoretical principles are one starting point of this paper. A good point is that MI is not difficult to apply since most types of estimates can be computed in a usual way (e.g. averages, quantiles, standard deviations and regression coefficients). The Rubin's framework also serves the formulas both for point estimates and for interval estimates. The point estimates are simply averages of *L* repeated complete-data estimates, and thus very logical. His interval estimates are not indisputably accepted. Björnstad (2007) gives a modified version for the second component of Rubin's formula. This leads to a larger confidence interval, as a function of the rate of imputed values. This is logical since Rubin's formula is without any explicit term of the imputation amount.

Björnstad (2007, 433) also invents a new term, non-Bayesian MI, since his imputation is not following a Bayesian process. This term 'non-Bayesian' is not really used in imputation literature; it cannot be found 6 years alter from a book by Carpenter and Kenward (2013) that much follows Rubin's framework. We still use the term 'non-Bayesian,' although 'repeated imputation' could be a more neutral term.

Björnstad motivates his approach also from the practical points of view saying that in national statistical institutes (NSI's) the methods used for imputing for nonresponse very seldom if ever satisfy the requirement of being "proper." We agree with this view. Since a non-Bayesian approach also leads to single imputation, that is commonly used in NSI's if anything has been imputed, a conclusion could be that MI cannot be applied using a non-Bayesian framework. We do not agree with this argument. Consequently, we have over years (Laaksonen 2000, Laaksonen and Piela 2003, Laaksonen 2003, Laaksonen et al 2004, Laaksonen 2013) applied non-Bayesian tools both for single and multiple imputation, although most often for single imputation. This paper summaries our approach to imputation.

This approach first makes attempts to impute the missing values once. That is, the focus is first in single imputation. Correspondingly, the main target in imputations is to succeed in such estimates that are most important in each case. Since it is hard to impute correctly individual values, it is more relevant to try to get least unbiased estimates for some key estimates. Since we here concentrate on a categorical binary variable, one type of estimate is most important, that is, poverty rate. In the case of a continuous variable, more estimates usually are important, including distributional figures. On the other hand, interval estimates such as standard errors, are important in all cases. Both Rubin's and Björnstad's approaches offer an alternative for this purpose but both of them are rarely discussed in literature.

Rubin's approach can be implemented in various ways. We do not develop any own implementation but take advantage of the two existing implementations. These are derived from two general software packages, SAS and SPSS, respectively. We believe and assume that their MI procedures follow a Bayesian process since there are such references in their manuals. We thus use the term 'Bayesian MI' for the applications of SAS and SPSS. Respectively, our own imputation framework is called 'Non-Bayesian MI.'

In next section, we present the basics of multiple imputation, estimates in particular. These are fairly general statements but they facilitate in understanding the forthcoming sections with precise methodological points and applications. In Section 1 we first present our non-Bayesian methods of single imputations. Most of these are multiplied to MI in the same Section. Some special methods from software packages are explained in Section 3. One of these is predictive mean matching that is available both in SAS and SPSS, but our particular non-Bayesian methods have the same target to produce plausible values only. Section 4 briefly describes our test data in which 27 per cent of incomes of persons is missing and thus required to impute. Good auxiliary data are available but not extremely good since the fit for the dependent variable is fairly poor. The same section continuous to the results. The final section concludes so that we do not see any reason why Bayesian imputation methods could be preferred in the case of a binary variable,

1. Point and interval estimates of multiple imputation

The point estimates of multiply imputed complete data sets are given similarly both by Rubin and Björnstad or all others. The parameter (Rubin uses the term 'estimand') being estimated may thus be any statistic of interest. In this study, we thus have chosen one parameter, poverty rate.

In order to make the formulas simple, we denote the estimate by *Q* that is here the average. Such an estimate is calculated from a complete data set after each single imputation. Thus the estimate from a single imputed complete data set is Q_i and the respective variance is B_i, both calculated taken into account the sampling design. In our empirical data the sample is based on simple random sampling, and both estimates can be calculated using the simple formulas. Rubin even in 1996 (p. 479) says that 3 to 5 repeated imputations works well if the fraction of missing information is typical in careful surveys. In the Dacseis project of the EU (e.g. Laaksonen et al 2004) we were not fully happy with a small number imputations, in some cases even more than 10 imputations could have been needed. This number is mentioned also by Rubin (e.g. 2004, 227) but not generally recommended. For this study, we always calculate *L*=10 imputations that number seldom is exceeded in examples we have seen.

The MI point estimate is thus simply the average of the *L* imputations

(1)
$$Q_{MI} = \frac{\sum_i Q_l}{L} \quad .$$

Respectively, the variance estimate is

(2)
$$B_{MI-within} = \frac{\sum_l B_l}{L} \text{ in which } B_l \text{ is a SI variance.}$$

There are two alternatives to calculate the MI variance of the *L* complete data sets. The first term of the variance, called within-imputation variability (variance), is in both cases equal that is formula (2). But the second term, the between-imputation-variability, is larger in Björnstad's version.

(3)
$$B_{MI} = B_{MI-within} + (k + \frac{1}{L})\frac{1}{L-1}\sum_l (Q_l - Q_{MI})^2 \quad .$$

The difference is in the term k=1/(1-f) in which *f* is the fraction of missing values or the non-response rate. This increases while the fraction increases. Rubin's formula does not depend explicitly on the amount of the imputed values. Rubin's Bayesian approach possibly takes this into account implicitly but it is hard to see. Björnstad developed his formula for some most common sampling designs including simple random sampling. Thus it is allowed to use this formula in this study although we give the results using Rubin's formula as well. Of course, it is not clear what is most correct in each case. It could be considered that Rubin's formula works with Bayesian MI and Björnstad's formula with non-Bayesian MI, but we will not give any definite conclusion to this even after the empirical tests.

2. **Non-Bayesian single** and multiple **imputation for a categorical variable**

Laaksonen and Piela (2003) have called this approach ‘integrated modelling approach to imputation’ that illustrates its key points well. We here do not use this term since it is easy to recognize these two approaches without any specific term.

Before imputation, it should be decided which variables are needed to impute. The first criterion for this decision is that some advantage is expected to get due to imputation. Such an advantage may be measured both with point estimates and with interval estimates. Usually, if the bias in a point estimate is smaller after imputation, it is a good point. On the other hand, the interval estimate should be reasonably small. Without imputation, both estimates thus are too biased but this is hard to well recognize in real-life. We present also some benchmarking methods for comparisons as explained below.

In order to succeed in imputation, good auxiliary data or covariates in Rubin’s terminology are needed. In the case of lacking covariates, simple methods based on observed values only can be applied. But if there are covariates both for the respondents and for the nonrespondents, ‘real’ imputation methods can be used. In this case, our imputation framework includes the two core stages:

(i) Construction and implementing of the imputation model

(ii) Imputation itself or imputation task.

Imputation model

An imputation model can be implemented using a smart knowledge of the imputation team or it can be estimated from the same data set or from a similar data set from an earlier survey or a parallel survey of another population. If the data are not estimated from the same data set, it is expected that this replacer behaves similarly with the present data set (e.g. Chambers et al 2001). This study is ordinary, that is, we estimate the parameters of the imputation model from the same data set.

There are the two alternatives as a dependent variable in an imputation model. It is either (a) the variable being imputed or (b) the binary missingness indicator of the variable being imputed. The same auxiliary variables can be used in both models. Naturally, the estimations and the predictions respectively that are needed in the next step are derived from the different data sets, from the respondents for the model (i) and from both the respondents and the nonrespondents for the model (ii).

Imputation task

The imputed values themselves can also be determined by the two options: (i) they are calculated using the imputation model or (ii) they are borrowed from the units with the observed values using the imputation model as well. The previous imputations are called ‘model-donor’ imputation and the second ones ‘real-donor’ imputation, respectively. The latter ones are often called ‘hot deck’ but this term is not clear in all cases. Terms for the previous ones are often such that the model and the task are confused. For example, model imputation or logit imputation is not clear since these are referring to imputation model but the second step, imputation task, is not specified.

If a real-donor method is applied, an appropriate criterion and a valid technology to select a donor are needed. The natural criterion is to select an as a similar real-donor as possible. This may be based on a kind of nearness technics. If a clear criterion exists, it is good to select a nearest one or another from a neighborhood. If any valid criterion does not exist, a random selection from the neighborhood can be used. This thus means that all units with observations are as close to each other within the neighborhood that is called ‘an imputation cell,’ This method is one of our benchmarking method, called *random real-donor method*. It is expected that all good methods are better than this benchmarking method.

Imputation model and imputation task in this study

The initial variable of the data is yearly income. Using this variable a poverty indicator has been constructed so that an individual under a poverty line is poor with the code =1, and the rest is not poor and coded by 0. For the imputation model this is here specified in the two manners:

- Binary poverty indicator itself
- Binary indicator of the missingness of the poverty variable.

We thus have two binary models that may be confusing since there are different approaches to both ones. Our auxiliary variables or covariates are always the same in order to fairly compare the methods. However, we apply four link functions in all single non-Bayesian imputations, that is, linear, logit, probit and complementary log-log that are available in SAS which software package is used in our applications. For non-Bayesian multiple imputations the linear case is excluded although it could be well comparable. Bayesian methods seem to work with a logit link function. Each imputation model is estimated using a number of categorical auxiliary variables that are described in Section 3.

Our benchmarking single model-donor imputations are simply predicted values of the imputation model, however rounded to the nearest integer. This method is sometimes used although it is not expected to be good.

Figures 1 and 2 illustrate how some predicted values vary using scatter plots. The first one in Figure 1 well shows how similar are them to a logit and to probit link respectively. The 'two lines' in the middle are interesting. The other scatter plot in Figure 2 is fairly spread when comparing the models with the different dependent variables, either the binary variable of poverty or that of non-response indicator. It is expected that imputes are fairly similar of the first models but is not well seen what happens in the second case.

Fig. 1. Illustration of the two link functions, logit and probit, by their predicted values for the non-respondents.

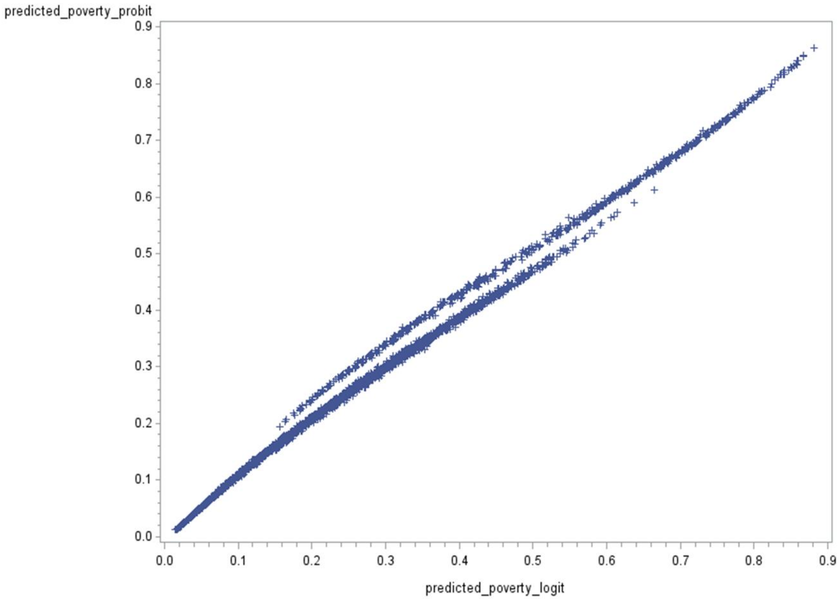
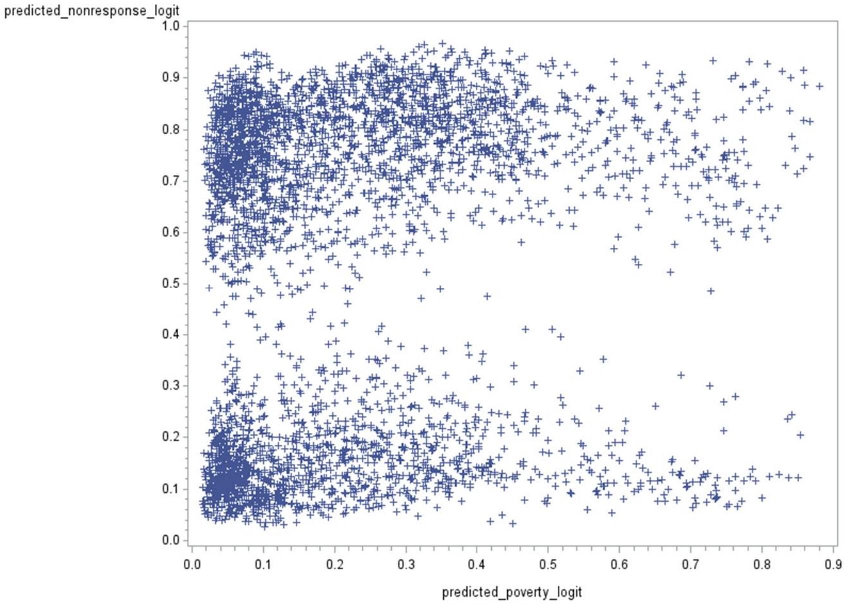


Fig. 2. Illustration of the two predicted values of the different imputation models with the same logit link function for the non-respondents.



Our model-donor method for single imputation follows a Bernoulli approach so that we first calculate the predicted values of each unit *k* with a missing value, let say p_k. On the other hand, we create a uniformly distributed random number within (0, 1) for the same units, let say u_k. The imputed values are obtained as follows:

- if $u_k > p_k$ then *y_imputed* = 1, otherwise *y_imputed* = 0;

This strategy thus gives model-donor imputed values with desired link function. It it is needed to be careful in order to get the correct codes 1 and 0 so that 1 = poor, and 0 = non-poor, for instance. When changing a random seed number 10 times, 10 non-Bayesian model-donor multiply imputed values are obtained.

The single real-donor imputations are made in the two rules, using the predicted values (propensity scores) of both types of binary regression models (above). These predicted values are first sorted so that it is possible to search for a nearest (or near) neighbor of each missing-value unit. Naturally, all these predicted values are calculated both for the respondents and the nonrespondents even though that the true values are available for the respondents. Thus the complete predicted values serve as nearest neighbor indicators for real-donor imputation. Note that a nearest neighbor may be close to the unit being imputed or quite far. In our tests, most neighbors were fairly close but a most distant neighbor was about 40th closest. It is expected that the imputations are less good in such cases.

This single real-donor imputation method is completed to multiple imputation using the following procedure:

1. The predicted values p_k are estimated as for single imputation.
2. The standard error of the predicted values is estimated and included as a constant value in the data set, let say *stderr*.
3. The normally distributed random numbers are created with the zero mean and the standard deviation equal to one, let say *u_nor*.
4. The new predicted values for searching for a nearest neighbor are = p_k + *u_nor*stderr*.

This procedure ensures that the average of the new predicted values is approximately equal to the initial one. Since there are random numbers used, the order of the units will change to some extent and leads to an additional variation in imputed values. This strategy is not obviously used before but it is logical since the order of the units vary within normally distributed limits of standard errors. The strategy in which imputation cells are constructed using predicted values and imputations are derived from random real-donor methods with cells is more commonly used (e.g. Rubin 2004). This strategy is not necessarily good since the number and the nature of imputation cells are determined subjectively.

3. **Bayesian** multiple imputation in some SAS and SPSS modules

SAS includes more MI modules than SPSS. We describe the methods we use in empirical part below.

Predictive mean matching imputation _PMM (SPSS)

The predictive mean matching method is an imputation method available for continuous variables. Our binary poverty variable is continuous at the same time and this method can be used respectively. This model can be called a binary regression model with linear link function that we examine in the empirical part and see that it works fairly well. PMM is similar to the regression method except that for each missing value, it imputes a value randomly from a set of observed values whose predicted values are closest to the predicted value for the missing value from the simulated regression model (Little 1988, Schenker and Taylor 1996). Rubin (2004, 168) also uses the term ‘predictive mean hot deck imputation’ that is one application of this method.

The PMM method requires the number of closest observations to be specified. A smaller number tends to increase the correlation among the multiple imputations for the missing observation and results in a higher variability of point estimators in repeated sampling. On the other hand, a larger number tends to lessen the effect from the imputation model and results in biased estimators.

This method is similar with our real-donor method in one respect. It gives only plausible values unlike linear regression methods cannot ensure it. The difference is in the imputation model that is basically the same as in Bayesian MI regression imputation method. Our imputation model specification for MI is simpler, that is, the sum of the predicted values and the normally distributed random term. This thus is a real-donor method in which the nearest or near neighbor indicator is derived from a regression model. But its competitive indicator could be from the binary regression model or the propensity score model that follows Bayesian rules in SAS as explained below.

Predictive mean matching imputation with MCMC (SPSS)

SPSS offers this Markov Chain Monte Carlo specification both for linear regression and predictive mean matching but we apply it for predicted mean matching only since we wish to get plausible values 1 or 0 that are not ensured with linear regression. The SPSS manual says that ‘the fully conditional specification (MCMC) is suitable for data with an arbitrary pattern of missing values.’ Allison (2005) even considers that the most popular method for multiple imputation of missing data is the MCMC algorithm. We do not here describe the MCMC methods in details, but we apply this option.

Propensity Score Method for Monotone Missing Data (SAS)

The propensity score method is another imputation method available when the data set has a monotone missing pattern. A propensity score is generally defined as the conditional probability of assignment to a particular treatment given a vector of observed covariates (Rosenbaum and Rubin 1983). In the propensity score method, for a variable with missing values, a propensity score is generated for each observation to estimate the probability that the observation is missing. The observations are then grouped based on these propensity scores, and an approximate Bayesian bootstrap (ABB) imputation (e.g. Rubin 2004, 124, 136) is applied to each group. Divide the observations into a fixed number of groups (typically assumed to be five) based on these propensity scores. See the technical details from the SAS manual and more theory about Rubin's book.

Our non-Bayesian method thus does not use Bayesian bootstrap but takes the predicted values of the binary regression model, see above. The imputation model in SAS is the logistic regression that is one option in our applications but a probit and complementary log-log link are experienced as well.

Monotone Logistic Regression Method for CLASS Variables (SAS)

This method uses logistic regression method to impute values for a binary variable in a data set with a monotone missing pattern. In this method, the imputation model is estimated for the binary variable being imputed, it is thus a Bayesian real-donor method whereas the imputation model of the previous method is for the response indicator. This method thus is familiar with our real-donor method when the imputation model is for a binary poverty indicator whereas the previous method corresponds to our method with the response indicator.

4. Test data and results

The micro data set of this study is close to real-life situation in an anonymous European country. Its initial version (Danish Labour Force Survey) has been already used in imputation tests of the Euredit project (Wagstaff 2003). For this study, a random sample of 19774 persons (10%) was drawn from the initial one. Some cleaning was made at the same time. For instance, the persons with zero income were dropped out since this is not any plausible value in real life. The poverty indicator using the income variable was constructed following a standard procedure for such a rate to some extent so that this indicator is here for individuals, not for households; this rate is fairly high but realistic.

The missingness mechanism was remained the same, since it was observed to be good in the Euredit experiments that tested both tradition and so-called new imputation methods. The main aim of these experiments was to compare methods in their biases using a number of evaluation criteria (Chambers 2003). The bias was possible to test at individual level since the true values were available after imputations. Euredit only tested single imputation methods. Also the variances estimates were left out of their tests but now we examine both point and interval estimates.

The number of missing values or the imputation size is 5315 (27%) that is fairly realistic although could be relatively larger today. The data set consists of a quite good number of covariates which all except age are categorical. The age was however categorized for tests of this study (6 categories). The full list with the number of categories that is used in all imputation models is as follows: age group (6), socio-economic status (3), education (3), gender (2), region (12), children (2), unemployed (2), civil status (2), marriage (2), web or not (2). However, we also use a smaller number of covariates in order to see how well methods work in such a case. In this case, the last five variables were excluded.

All these categorical variables are statistically significant, best being gender, education, age group and socio-economic status. Nevertheless, the fit is fairly low since any individual level strong variable is not available.

Results for single imputations are given in Table 1. The point estimate 'Mean' is the poverty rate for the nonrespondents and the interval estimate is a simple standard error assuming that imputed values are 'true.' When comparing the imputations derived from those two different imputation models, we observe that the means are always higher in the case of the full set of auxiliary variables. This in most cases leads also to less biased point estimates but real-donor methods with binary poverty models are exceptions, those giving too high values. This is not easy to explain. Fortunately, all imputations lead to less biased estimates than obtained assuming that missingness is completely random which corresponds to the true value of the respondents.

The best results are obtained using real-donor methods with response model although the difference is not substantial to model-donor methods that are obtained with the same random number. The results with different link functions vary but any clear conclusion cannot be given. Interestingly, linear link function works fairly well as well. Standard errors vary very little and are close to those of true values.

Table 1. *Single imputation results for the poverty rate with the two patterns of covariates in the model. The upper figures are based on a smaller and the lower figures for a higher number of covariates.*

Method			Mean	Standard error
	Link	Imputation Task		
Model for poverty	Linear	Model-donor	0.231 0.241	0..0058 0,0059
Model for poverty	Logit	Model-donor	0.228 0.239	0.0058 0.0059
Model for poverty	Probit	Model-donor	0.231 0.240	0.0058 0.0059
Model for poverty	Cll	Model-donor	0.228 0.242	0.0058 0.0059
Model for poverty	Linear	Real-donor	0.258 0.266	0.0060 0.0060
Model for poverty	Logit	Real-donor	0.239 0.262	0.0058 0.0060
Model for poverty	Probit	Real-donor	0.239 0.268	0.0058 0.0061
Model for poverty	Cll	Real-donor	0.238 0.258	0.0058 0.0060
Model for response	Linear	Real-donor	0.229 0.244	0.0058 0.0059
Model for response	Logit	Real-donor	0.239 0.250	0.0058 0.0058
Model for response	Probit	Real-donor	0.233 0.244	0.0058 0.0059
Model for response	Cll	Real-donor	0.238 0.236	0.0058 0.0058
Predicted value rounded to integer			0.118 0.127	0.0044 0.0046
TRUE VALUE of the respondents			0.206	0.0034
TRUE VALUE of the nonrespondents			0.249	0.0059

MI results are given in Table 2. Now all methods are derived from a maximum number of covariates. Random real-donor method is a benchmarking method and lead to the same point estimates as the true value of the respondents in Table 1. Now we can get the benchmarking interval estimates as well and their two forms, Rubin's and Björnstad's. It is expected that these standard errors are larger than those in all proper imputation methods. This is really the case but which are best, it is not clear. These estimates are lowest for model-donor methods and the bias in their point estimates is also minor. However, we can be happy with real-donor methods with response indicator as well. Real-donor methods when using poverty indicator are worst on average although they can be reasonably good for most users.

Most non-Bayesian methods work a bit better than Bayesian ones in point estimates. Standard errors of Bayesian methods are slightly higher than those of non-Bayesian methods. Now we could discuss whether the standard errors of non-Bayesian methods should be calculated using Björnstad's formula whereas these of Bayesian methods using Rubin's formula. In such a case, the standard errors would be closer to each other, but this is not any suggestion, it is a discussion point only. On the other hand, the standard errors of model-donor methods are lowest nevertheless, and this could be a value of imputation methods as well. Our study thus suggests to use this methodology for binary variables whenever it is possible.

Table 2*. Multiple imputation results for the poverty rate. The higher number of covariates are used in all methods (cf. Table 1).*

Imputation model		**Imputation task**	**Poverty Rate**	**Rubin Standard Error**	**Björnstad Standard Error**
Dependent variable	**Link function**				
Binary poverty indicator	Logit	Model-donor	**0,246**	0,0086	**0,0093**
Binary poverty indicator	Probit	Model-donor	0,244	0,0089	**0,0096**
Binary poverty indicator	CLL	Model-donor	**0,249**	0,0082	**0,0089**
Binary poverty indicator	Logit	Real-donor	0,232	0,0082	**0,0089**
Binary poverty indicator	Probit	Real-donor	0,232	0,0070	**0,0079**
Binary poverty indicator	CLL	Real-donor	0,235	0,0087	**0,0094**
Binary response indicator	Logit	Real-donor	0,243	0,0085	**0,0093**
Binary response indicator	Probit	Real-donor	0,243	0,0087	**0,0094**
Binary response indicator	CLL	Real-donor	**0,251**	0,0103	**0,0109**
SAS MI Propensity Score			0,239	**0,0097**	0,0104
SAS MI Logistic regression			**0,251**	**0,0115**	0,0121
SPSS Predictive Mean Matching			0,254	**0,0117**	0,0123
SPSS MCMC Predicted Mean Matching			0,253	**0.0092**	0,0099
Random real-donor			0.206	0,0072	0,0079

5. Conclusion

Single imputations are much used in survey institutes but they are often very simple and deterministic. The theory behind them can be called non-Bayesian or frequentist. During past 25 years multiple imputations are developed and applied. Their specific advantage is to get relatively easily standard errors if some data are imputed. Since the imputation is an additional factor of uncertainty, multiple imputation (MI) can be considered to be a useful tool for statisticians.

The initial theory behind multiple imputations is Bayesian. Consequently, MI methods using standard software packages like SAS and SPPS are implemented much following Rubin's framework (1987, 2004). A big question is whether MI methods could work under a non-Bayesian framework as well. The focus of this study has been to examine non-Bayesian techniques and tools for multiple or repeated imputations when a binary variable is attempted to impute. We have found several competitive strategies respectively so that the imputation consists of the two main stages, an imputation model and an imputation task respectively. This strategy is basically similar for continuous variables, but the implementations not.

We have presented three families of implementations so that the imputation models are of two types. The dependent variable is either the variable being imputed or the response indicator of the same variable. The imputation tasks are of the two types as well, called either real-donor or model-donor imputations. Since model-donor imputation cannot be used under the response indicator model, we will have these three methods families. Since the models can be specified with three link functions, we have nine non-Bayesian methods to compare.

The comparisons with ordinary Bayesian methods of SPSS and SAS suggest that these much used software package methods are not superior to non-Bayesian counterparties. Our results even suggest that some non-Bayesian methods are better than Bayesians. We cannot say surely why this is the case. It seems that there are in Bayesian tools some additional technical elements that do not improve anything. Hence an ordinary user have to apply them as a black box. When following appropriate non-Bayesian tools available to each situation in a better strategy, a user can see easily how each method works and revise its implementation if needed. This is a big advantage in imputations in general since the missing data replacements should be tailored to each particular case, not done automatically unless the quality of the method has not been checked well in advance.

References

(1) Allison, B.D. (2005). Imputation of Categorical Variables with PROC MI. *SUGI 30 Proceedings*. http://www2.sas.com/proceedings/sugi30/113-30.pdf

(2) Björnstad, J. (2007). Non-Bayesian Multiple Imputation. *Journal of Official Statistics*, 433–452.

(3) Carpenter, J. and Kenward, M. (2013): *Multiple Imputation and its Application*. Wiley & Sons

(4) Chambers, R. (2003). Evaluation Criteria for Statistical Editing and Imputation. *Euredit project. Papers.* http://www.cs.york.ac.uk/euredit/

(5) Chambers, R.L., Hoogland, J., Laaksonen, S., Mesa, D.M., Pannekoek, J., Piela, P., Tsai, P. and de Waal, T. (2001). *The AUTIMP-project: Evaluation of Imputation Software.* Research Paper 0122. Statistics Netherlands.

(6) (6) Laaksonen, S. (2000). Regression-Based Nearest Neighbor Hot Decking. *Computational Statistics*. 15,1, 65-71.

(7) Laaksonen, S. (2003). Alternative Imputation Techniques for Complex Metric Variables. *The Journal of Applied Statistics,* 1009-1020.

(8) Laaksonen, S. and Piela P. (2003). Integrated modelling approach to imputation. *Euredit Project Documents. Standard Methods*, D512 StatFI, ttp://www.cs.york.ac.uk/euredit/

(9) Laaksonen, S., Rässler, S. and Skinner, C. (2004). Documentation of Pseudo Code of Imputation Methods for the Simulation Study. *Dacseis Project Research Papers under Workpackage 11.2* 'Imputation and Nonresponse'. 51 pp. www.dacseis.de/deliverables.

(10) Laaksonen, S. (2013). Principles of Imputation Methods. *Baltic-Nordic-Ukrainian Workshop on Survey Statistics.* Minsk, 13-18 June.

(11) Little R.J.A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, Vol. 6, No. 3, 287-296.

(12) Rosenbaum, P. R. and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55

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

(14)Rubin, D. (2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley Classics Library Edition.

(15) Rubin, D. (1996). Multiple Imputation After 18+ Years. *Journal of American Statistical Association,* 473-489.

(16) SAS 9.3 Help and Documentation, Users' Guide for the MI Procedure. Details.

(17) Schenker, N. and Taylor, J. M. G. (1996), "Partially Parametric Techniques for Multiple Imputation," *Computational Statistics and Data Analysis*, 22, 425–446.

(18) Wagstaff, H. (2003). APPENDIX B: DATA SETS AND PERTURBATIONS. *Euredit project.* Volume 2. http://www.cs.york.ac.uk/euredit/